Smoothing with Kernel Regression and Related to Principal Component Analysis

Sena Ilgaz

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
February 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Assoc. Prof. Dr. Ali Hakan Ulusoy
Acting Director

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

Prof. Dr. Nazım 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. Assoc. Prof. Dr. Tolgay Karanfiller
3. Asst. Prof. Dr. Yücel Tandoğdu

ABSTRACT

In any process that produces useful output, more than one and in many cases tens or hundreds of variables are involved. With the advancement of technology the number of observations has also dramatically increased, to the point that without using a computer software it is impossible to process such data. For processing multivariate big data sets, there are many different techniques available.

In this thesis Kernel Regression which is a non-parametric regression method is used for estimating various dependent variables. In chapter 3 basic theory related with kernel regression is given, supported by the proof of various theorems and application data.

For large number of variables the Principal Component Analysis (PCA) technique is used to reduce the number of variables to manageable level. Basic theory related with PCA is given under chapter 4. In this thesis a logical link between kernel regression and PCA is established for the estimation of the variables governing a process. The variables governing the process are taken as dependent X_i, and Principal Components (PC) as independent variables, using kernel regression.

In chapter 5, a data set consisting of 14 variables was used to determine the necessary number of PCs, using both covariance and correlation matrices separately. Then, variables that exhibited high correlation with PCs, and variables with high contribution to a PC were taken as dependent variables, while PCs were used as independent variables in kernel regression.

For obtaining optimal bandwidth simulations were carried out. Mean Squared Error (MSE) and the ratio of MSE to the average of the variance of estimated values (AVE) were used as criteria, in obtaining the optimal bandwidth. It is determined that the linear correlation between the PC and the variable, and the contribution of a variable to the PC has significant effect on the error levels.

Keywords: Kernel Regression, Bandwidth, Principal Component Analysis (PCA), Principal Components (PCs), Mean Squared Error (MSE), Covariance, Correlation.

ÖZ

Kullanışlı çıktı üreten herhangi bir işlemde, birden fazla ve çoğu zaman onlarca veya yüzlerce değişken söz konusudur. Teknolojinin gelişmesiyle birlikte elde edilebilen gözlem sayısı ciddi şekilde artarken, bilgisayar yazılımlarını kullanmadan bunların analiz edilmesi imkansızdır. Çok değişkenli büyük verilerin işlenmesi için, birçok farklı teknik mevcuttur. Bu tezde parametrik olmayan bir regresyon yöntemi olan Kernel Regresyonu, çeşitli bağımlı değişkenleri tahmin etmek için kullanılmıştır. Bölüm 3'te kernel regresyonu ile ilgili temel teori, çeşitli teoremlerin ispatı ve bir uygulama örneği ile desteklenerek verilmiştir.

Çok sayıda değişken için, değişken sayısını yönetilebilir seviyeye düşürmek için Temel Bileşen Analizi (TBA) tekniği kullanılır. TBA ile ilgili temel teori bölüm 4'te verilmiştir. Bu tezde, süreçte geçerli olan değişkenlerin tahmini için kernel regresyonu ile TBA arasında mantıksal bağlantı kurulmuştur. Bu mantıkta değişkenler (X_i) bağımlı olarak, Temel Bileşenler (TB) bağımsız değişkenler olarak alınarak kernel regresyonu uygulanmıştır.

Beşinci bölümde 14 değişkenden oluşan bir veri setinin kovaryans ve korelasyon matrisleri ayrı ayrı kullanılarak gerekli TB sayısı belirlemiştir. Daha sonra, TB'lerle yüksek korelasyon gösteren değişkenler ve TB'ne yüksek katkısı olan değişkenler, bağımlı TB'ler ise bağımsız değişkenler olarak alınarak kernel regresyonu uygulanmıştır.

Optimal bant genişliği elde etmek için simülasyonlar yapıldı. Hata Karelerinin Ortalaması (HKO) ve HKO'nin tahmin edilen değerlerin varyans ortalamasına oranı, optimal bant genişliğinin elde edilmesinde ölçüt olarak kullanılmıştır. TB ile değişken arasındaki doğrusal korelasyonun ve bir değişkenin TB'ye katkısının hata seviyeleri üzerinde önemli bir etkiye sahip olduğu tespit edilmiştir.

Anahtar Kelimeler: Kernel Regresyonu, Bant Genişliği, Temel Bileşenler Analizi (TBA), Temel Bileşenler (TB), Hata Karelerinin Ortalaması (HKO), Kovaryans, Korelasyon.

DEDICATION

Thanks to; my only hero in my life is my father Tuncay Ilgaz for his supporting me, my mother Hanife Ilgaz for her eternal love and my twin sister Hande Ilgaz for not leaving me alone even when I was born.

ACKNOWLEDGMENT

I would like to thank my supervisor, Asst. Prof. Dr. Yücel Tandoğdu for his support and guidance throughout this study. He always guided me through this research, made me safe and motivated to share his knowledge and experience on the subject, encouraged, motivated and trusted me while doing this research.

A special thanks to my parents Tuncay Ilgaz, Hanife Ilgaz and my twin sister Hande Ilgaz for the love and support they have given to me.

I am grateful to my aunt Nur Kayanselçuk, who I have taken as a model to study mathematics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

λ	Eigenvalue
PCA	Principal Component Analysis
PC	Principal Components
MSE	Mean Squared Error
$\mathbf{e}$	Eigenvector
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathbf{X}'$	Transpose of a vector $\mathbf{X}$
$N(0,1)$	Standard normal distribution
σ	Standard deviation
$\bar{\mathbf{x}}$	Sample mean vector
$\bar{x}$	Sample mean
$\mathbf{S}$	Sample covarince matrix
$\mathbf{R}$	Sample correlaiton coefficient matrix
Y	Response variable
X	Predictor variable
$\boldsymbol{\mu}$	Population mean vector
μ	Population mean
$\boldsymbol{\Sigma}$	Population covariance matrix
$\boldsymbol{\rho}$	Population correlation coefficient matrix
$\hat{h}_0$	Optimum *h* value
$\hat{m}_h(x)$	Nadaraya - Watson Estimator
f	Kernel function

K	Kernel
$\hat{R}$	Interquartile range
$\hat{f}$	Estimator of f
ε	Error variable
$\mathbf{\Lambda}$	Diagonal matrix of eigenvalues
ρ_{Y_i,X_j}	Correlation between the i^{th} PC Y_i and its j^{th} variable X_j
h	Bandwidth
$\mathbf{X}$	Sampling data vector

Chapter 1

INTRODUCTION

All natural processes are governed by many different variables interacting with each other in highly complicated ways. Modelling such variables by means of some function based on parameters is a very demanding task or in many cases almost impossible. This is primarily due to the fact that for any random variable, the exact distribution function and its parameter values are unknown. Recent developments in modelling a process without the need of the population parameters, and techniques that enable the reduction in the number of variables without a big loss of the true nature of the process has become possible. Kernel regression is a well-known nonparametric method that enables the estimation of a variable without the need for the population parameters. Principal Component Analysis (PCA) is the technique that enables the reduction of the number of variables, in a multivariate process without the loss of main variation inherent to the process. Recent advancements in science and technology resulted in a dramatic increase in the volume of data generated, to the point that without using a computer software it is impossible to process such data. For processing of multivariate and big data sets, there are many different techniques available.

In this thesis in addition to the explanation and application of Kernel regression and PCA technique, a relationship between the two methods is proposed. In this proposal, the principal components (PC) were taken as the independent (predictor) variables to

estimate some variables from the process that satisfy high correlation and high contribution to the PCs.

Following literature review in Chapter 2, theoretical background regarding the kernel regression, and some important theorems with their proofs are given in Chapter 3. Nadaraya Watson kernel estimator is explained together with associated theoretical background. Bandwidth to be used in kernel regression is explained in fair detail, since it is the most important variable that determines the amount of smoothing, and hence the bias - error variance balance relationship. A data set using 63 observations on temperature – humidity taken from a meteorological database was used to highlight the important points explained theoretically in the chapter.

In Chapter 4 the basic concepts of dimension reduction technique, as part of the principal components analysis is explained. Theoretical background behind the PCA that leads to the formation of the PCs as a linear combination of the variables governing the process under study is explained. Situations under which the covariance or correlation matrices to be used in the formation of PCs are also given. Various criteria in determining the number of PCs to represent the process under study are studied.

Kernel regression and PCA theory explained in Chapter 3 and Chapter 4 respectively are applied to a data set consisting of 14 variables representing various properties of leaf shapes. Emphasis was given to the correlation between the optimum number of PCs and 14 independent variables, and the variables with high contribution to the PCs. Some of the variables governing the process were selected as dependent variables, to be estimated using a PC as independent variable. Selection criteria for these variables were high correlation with the PC, and/or high contribution of a variable to the PC.

Obtained results are summarized using tables and graphs, and interpreted. In all computations Matlab and Microsoft Excel were used.

Chapter 2

LITERATURE REVIEW

Principal Component Analysis (PCA) is a dimension reduction process. When the number of variables governing a process is very large, statistical manipulation of such data is difficult. Using PCA the number of variables can be reduced to manageable level, without loss of information carried by the original variables. An English mathematician, Karl Pearson introduced the first ideas on how to reduce the number of variables in a multivariate problem [3].

In 1931, Hotelling H. contributed by focusing on confidence intervals and regression slopes and the issue of from univariate to multivariate distributions [2].

Girschick A. M. (1939) worked on the topic of PCA, and produced useful results regarding the distribution of the roots and characteristic vectors associated with certain determinantal equations [15].

Anderson T. W. worked on the characteristics of multivariate distributions, principal components, canonical correlation and asymptotic properties of the characteristic roots [1].

Rao C. R. (1964) contributed about theories in multivariate data analysis, and characteristics of probability distributions. He also proposed graphical representation of multi-dimensional data in reduced dimensions which is closely related with PCA [20].

Jeffers J. (1967) proposed new methods of enhancing PCA using graphical approach to facilitate the clear understanding of the role of PCA in application [13].

In 1982, the regression method was introduced by Jolliffe in the field of main components analysis with principal component analysis [16].

In 2002, Fotheringham and his colleagues introduced the concept of local weighted principal components and the concept of geographic weighted components [27].

A review of developments in Kernel regression is as follows.

Fix E., Joe L. & Hodges J. L. (1951) in a technical report presented at the USAF Texas Base mainly focused on the discrimination problem of two populations, ways of freeing discriminant analysis from rigid distributional assumptions [8].

Rosenblatt M. (1956) discussed some aspects of the estimation of a univariate density function, where he classified his arguments under three headings [22].

1. Estimation of a density function,
2. The difference quotient of the sample distribution function
3. A class of estimates of the density function.

An important contribution was made by Farrell in 1972 on lower limits on the convergence rates of core estimators [7].

In 1979, the first MISE analysis of the histogram was performed by Scott [24].

Cline (1988) defined the notion of admissibility for kernels and showed that asymmetric and multimodal kernels are inadmissible [5].

Morron J.S. & Hall P. (1994) studied the aspect of determining the band width to be used in a Kernel Density Estimation process [23].

Chapter 3

KERNEL REGRESSION

3.1 Introduction

Kernel smoothing or kernel regression is a statistical technique that uses the local weights of a real-valued function to predict the weighted average of neighboring data. There are two main reasons for using kernel smoothing in the univariate density estimator. The first of which is an effective way to show that estimating non-parametric density in analytical data is important. The second reason is that the kernel estimators are simple in terms of mathematical traceability. Kernel smoothing provides simple, reliable and useful answers to major problems, which enables drilling down into the data features.

3.2 Estimation of Density and Histogram

The estimation of the probability density of random variables in the absence of population parameters, becomes a challenging problem. Let $X_1, \dots, X_n$ be continuous random variables with common density *f*. Parametric regression does not provide any flexibility in modelling due to the rigidity of the parametrs. A non-parametric density estimator does not depend on predetermined parameters of functional form of *f*. The oldest and most widely used nonparametric density estimator is the histogram. The histogram is constructed by dividing the actual line representing the range of the data into equally spaced intervals, called bins. The histogram is a step function where the height of each rectangle or the value of the smooth function *f* is a ratio of the number of samples in the bin in which *x* lies, to the product of *b* and size of sample data *n*.

Assume that, b is bin width, for predicting the histogram at point x [25]. Then the histogram value at point x is given by

$$f_H(x;b) = \frac{\text{number of observation in bin contaning } x}{nb}$$

Two things must be considered when creating a histogram: bin width and positioning of the dividing edges. Binwidth b is also called the smoothing parameter which controls the amount of smoothing. One of the main problems of the histograms is the bin edge. One way of solving this problem is the average shifted histogram (Scott,1985). That is the average of several histograms obtained by shifting the bin edges. This method has some similarities with the kernel density estimator [27].

3.3 Kernel Theory

Parametric estimators are not necessarily ideal tools for identifying the true characteristic of a process. On the other hand a nonparametric model such as kernel estimators can give more accurate estimation of the true trend exhibited by the process under study. This is possible by obtaining a density estimator that does not assume that the density has a particular functional form. In this thesis a univariate kernel density estimator is studied due to its simplicity and its concepts being readily amenable for extension into upper dimensional cases. Regression estimators based on the kernel functions are often referred to as kernel smoothers [9].

The basic idea behind kernel estimation in the univariate case is the assumption that there is a random sample $X_1, X_2, \ldots X_n$ of independent and identically distributed (i.i.d) observations from a continuous univariate distribution having probability density function (p.d.f) f that is to be estimated. Let $\hat{f}$ be the kernel estimator of the unknown (p.d.f) f. The estimator $\hat{f}$ is obviously depending on available data and the

kernel function *K* to be used. $\hat{f}(x)$ is considered as a random variable, due to its dependence on the sample $X_1, X_2, \ldots X_n$. Discrepancy of $\hat{f}$ from *f* can be measured via the Mean Squared Error (MSE) or the Mean Integrated Squared Error (MISE). Here

$$\text{MSE}(\hat{f}) = E\{[\hat{f}(x) - f(x)]^2\}$$

This can be decomposed into Bias and Variance components. This decomposition is important since in an estimation process a delicate balance between bias and variance needs to be maintained. This balance is a function of the bandwidth used in the computation of kernel values. The decomposition can be performed as follows.

Suppose that a data set $x_1, x_2, \ldots x_n$ is given. Using this data the function $y = f(x) + \varepsilon$; ε being the random error component or noise with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon)=\sigma^2$ to be estimated using the $\hat{f}(x)$. There is no doubt, the closer the $\hat{f}(x)$ is to $f(x)$, the better the estimation will be. That is to obtain the minimum MSE by employing the sample $x_1, x_2, \ldots x_n$, which will also be valid for points not included in the sample. Then the expected error can be decomposed into 3 components as follows.

$$E\left\{\left[y - \hat{f}(x)\right]^2\right\} = \text{Bias}\left[\hat{f}(x)\right]^2 + \text{Var}\left[\hat{f}(x)\right] + \sigma^2 \qquad (3.1)$$

Where $\text{Bias}\left[\hat{f}(x)\right] = E\left[\hat{f}(x) - f(x)\right]$, and $\text{Var}\left[\hat{f}(x)\right] = E\left[\hat{f}(x)^2\right] - f\left[(x)\right]^2$

Square of the bias is a function of the assumption made in approximating the unknown $f(x)$ by the estimator $\hat{f}(x)$. Variance of $\hat{f}(x)$ is a measure of how much it varies

around its mean. Therefore, the more sophistication put into $\hat{f}(x)$ towards reducing the bias, will lead to higher $\mathrm{Var}\left[\hat{f}(x)\right]$.

Decomposition given in equation (3.1) can be obtained as follows

Note the following. The variance of a random variable is.

1. $\sigma^2 = E\left[\left(X-\mu\right)^2\right] \Rightarrow \mathrm{Var}(X)=E(X^2)-E(X)^2 \Rightarrow E(X^2)=\mathrm{Var}(X)+E(X)^2$.

   Then

2. $E\left[f(x)+\varepsilon\right]=E\left[f(x)\right]=f(x)$ as $f(x)$ is deterministic and $E(\varepsilon)=0$.

3. $\mathrm{Var}(y)=E\left[\left(y-E(y)\right)^2\right]=E\left[\left(y-f(x)\right)^2\right]=E\left[\left(f(x)+\varepsilon-f(x)\right)^2\right]$

$$=E\left(\varepsilon^2\right)=\mathrm{Var}(\varepsilon)+E(\varepsilon)^2=\sigma^2$$

Since ε and $\hat{f}(x)$ are independent

$$E\left[\left(y-\hat{f}(x)\right)^2\right]=E\left[y^2+\hat{f}(x)^2+2y\hat{f}(x)\right]$$

$$=\mathrm{Var}(y)+E(y)^2+\mathrm{Var}\left(\hat{f}(x)\right)+E\left(\hat{f}\right)^2-2f(x)E\left(\hat{f}(x)\right)$$

$$=\mathrm{Var}(y)+\mathrm{Var}(\hat{f}(x))+(f(x)^2-2f(x)E(\hat{f}(x))+E(\hat{f}(x))^2)$$

$$=\mathrm{Var}(y)+\mathrm{Var}(\hat{f}(x))+(f(x)-E(\hat{f}(x)))^2$$

$$=\sigma^2+\mathrm{Var}(\hat{f}(x))+\mathrm{Bias}(\hat{f}(x))^2 \quad \textit{Q.E.D.}$$

3.4 Kernel Density Estimator

The kernel density estimator based on a random sample $\{X_1, X_2, \ldots, X_n\}$ is given as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left\{\frac{(x - X_i)}{h}\right\} \qquad (3.2)$$

where K is a function satisfying $\int K(x)dx = 1$, $\int x^2 K(x)dx < \infty$, and $\int K^2(x)dx < \infty$.

K is called the kernel, and $h>0$, called the bandwidth which is the smoothing parameter.

Let $K_h(u) = h^{-1}K\left(\frac{u}{h}\right)$. Then equation (3.2) becomes

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

The weight is defined by the Kernel $\left(K\left(\frac{x - x_i}{h}\right)\right)$, such that closer points are given higher weights. Since K is non negative, so is $\hat{f}(x)$. Then

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)dx$$

$$= \frac{1}{nh}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{x - X_i}{h}\right)dx$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K(u)du = 1.$$

Hence $\hat{f}(x)$ is a probability density.

Lemma 3.1 [12]**:** Assume *w(y)* is bounded and integrable function satisfying $\lim_{y\to\infty}|yw(y)| = 0$, and g be an integrable function. Also $h_n \to 0$ as $n \to \infty$,

$$\lim_{n\to\infty} \frac{1}{h_n}\int w\left(\frac{u - x}{h_n}\right)g(u)du = g(x)\int w(u)du,$$

for every continutiy point x of y.

Theorem 3.1: The estimator $\hat{f}(x)$ of the kernel probability density converges $f(x)$ in probability ($\hat{f}(x) \xrightarrow{P} f(x)$).

Proof: The Markov's inequality states that if X is a nonnegative random variable and $a > 0$, then the probability that X is at least a, is at most $P(X \geq a) \leq E(X)/a$. Based on this inequality it is sufficient to show $E(\hat{f}(x) - f(x))^2 \to 0$. Starting with,

$$E\left(\hat{f}(x)\right) = E\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right) = \frac{1}{nh}\sum_{i=1}^{n} E\left(K\left(\frac{x - X_i}{h}\right)\right)$$

$$= \frac{1}{h} E\left(K\left(\frac{x - X_i}{h}\right)\right)$$

$$= \frac{1}{h}\int K\left(\frac{x - x_1}{h}\right) f(x_1)dx_1 = \frac{1}{h}\int K\left(\frac{x_1 - x}{h}\right) f(x_1)dx_1$$

$$\to f(x)\int K(u)du \quad \text{(by Lemma 3.1)}$$

$$= f(x)$$

Then the bias term

$$E\left(\hat{f}(x) - f(x)\right) \to 0. \qquad (3.3)$$

This is followed by

$$\mathrm{Var}\left(\hat{f}(x)\right) = \mathrm{Var}\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right)$$

$$= \frac{1}{n^2h^2}\sum_{i=1}^{n} \mathrm{Var}\left(K\left(\frac{x - X_i}{h}\right)\right) \quad (X_i\text{'s are independent})$$

$$= \frac{1}{nh^2}\mathrm{Var}\left(K\left(\frac{x - X_1}{h}\right)\right)$$

$$= \frac{1}{nh^2}\left(E\left(K^2\left(\frac{x - X_1}{h}\right)\right) - \left(E\left(K\left(\frac{x - X_1}{h}\right)\right)\right)^2\right)$$

$$= \frac{1}{nh} \cdot \frac{1}{h} \int K^2\left(\frac{x - x_1}{h}\right) f(x_1) dx_1 - \frac{1}{n}\left(\frac{1}{h} E\left(K\left(\frac{x - X_1}{h}\right)\right)\right)^2$$

$$= \frac{1}{nh} I_1 - \frac{1}{n} I_2 .$$

By Lemma 3.1 $I_1 \to f(x)\int K^2(u)du$. From the computation for $E\left(\hat{f}(x)\right)$, $I_2 \to f^2(x)$. Therefore as $n \to \infty$.

$$\text{Var}\left(\hat{f}(x)\right) \to 0 \qquad (3.4)$$

From (3.3) and (3.4) $E\left(\hat{f}(x) - f(x)\right)^2 \to 0$ is obtained.

It is preferred to set *K* as a unimodal probability density function and symmetric about zero. This results in $\hat{f}(x)$ also being a density. A kernel density estimate constructed where 5 observations were used is shown in Figure 3.1. Kernel has $N(0,1)$ with density, $K(x) = g(x)$. Five observations were used to highlight the concept, where as in a real world problem number of observations tend to be much larger.

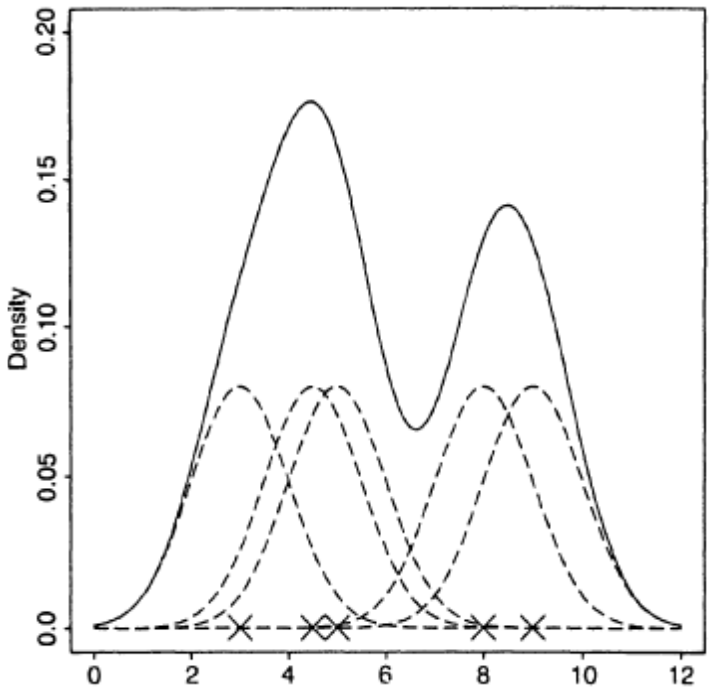

Figure 3.1: A kernel density estimate to highlight the effect of the density of observations

To start with at each point a scaled kernel is computed. The estimated value of the kernel at point *x* is obtained by averaging the *n* kernel estimates at that point.

Since a kernel estimate is the average of the contributions of observations in close proximity, a fairly large estimate is obtained where observations are dense, and kernel estimates will be relatively low where data values are sparse.

Bandwidth *h* has a significant effect on the level of smoothing achieved in kernel estimation. Figure 3.2 shows the density estimates where a sample of size 1000 is used, with different bandwidths. The density used for this purpose is

$$f_1(x) = \frac{3}{4} g(x) + \frac{1}{4} g_{1/3}(x - 3/2)$$

$f_1(x)$ is a combination of standard normal observations with probability ¾, and normally distributed observations with probability ¼. As seen in Figure 3.2 amount of kernel smoothing

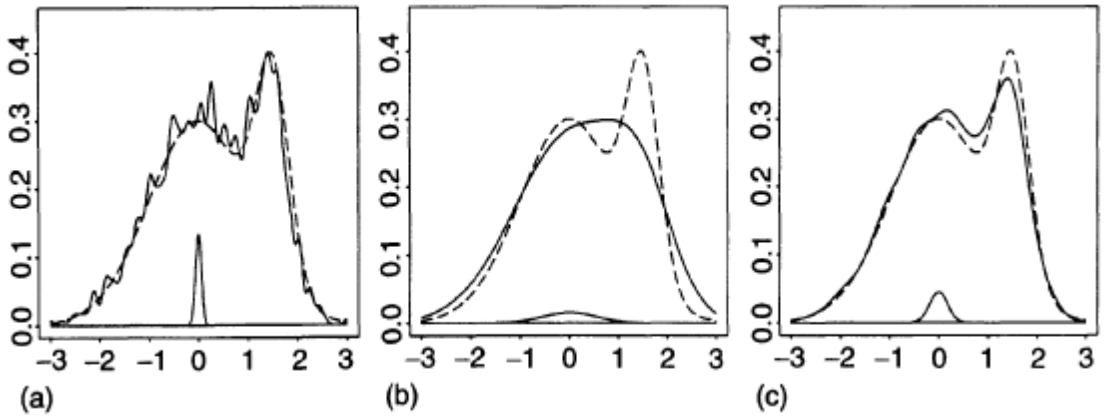

Figure 3.2: Amount of smoothing as a function of bandwidth.

is a function of bandwidth. Larger bandwidth results an increase in smoothing. This is clearly visible in Figure 3.2.

Computation of kernel values can be undertaken by various kernel functions.

Two most commonly used Kernel function are the Gaussian and Epanechnikov kernel functions.

The Epanechnikov's formula is, $\frac{3}{4}\left(1-u^2\right) I\left(|u| \le 1\right)$

Gaussian's formula is, $\frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}u^2}$ [10]

where, $u = \frac{x - x_i}{h}$

3.5 Selection of Bandwidth

The bandwidth *h*, that is known as Kernel smoothers, is called the smoothing parameter to regulate the degree of smoothness. There are different ways of determining *h*, but most of these formulae do not give the desired result in practice. The main problem in estimating the Kernel density is the choice of *h*. The Gaussian Kernel is $K(u)=\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$. The reference density is normal density of $N(\mu,\sigma^2)\rightarrow f(x)=\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$. To obtain the estimator of the optimum bandwidth h_o, let

$$\int K^2(x)dz=\int\frac{1}{2\pi}e^{-u^2}du=\int\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du.\frac{1}{2\sqrt{\pi}}=\frac{1}{2\sqrt{\pi}};$$

τ^2 is the second moment of $N(0,1)$. So, $\tau^2=1$. $f''(x)=\frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right)$. Hence,

$$\int(f''(x))^2dx=\frac{1}{\sigma^6}\int\left(\varphi''\left(\frac{x-\mu}{\sigma}\right)\right)^2dx=\frac{1}{\sigma^5}\int(\varphi''(y))^2dy=\frac{3}{8\sqrt{\pi}\sigma^5}$$

Hence,

$$h_0=\left(\frac{4\hat{\sigma}}{3n}\right)^{1/5}\cong 1.06\hat{\sigma}n^{-1/5} \qquad (3.5)$$

Equation (3.5) tends to detected outliers easily, which is not desired. The interquartile range of the data can be considered in place of $\hat{\sigma}^2$, which is defined as

$$\hat{R}=X_{0.75}-X_{0.25}$$

Then equation (3.5) is modified into,

$$\hat{h}_0=0.79\hat{\mathrm{R}}\mathrm{n}^{-1/5} \qquad (3.6)$$

Combining equation (3.5) and (3.6) gives a better estimate for band width

$$h_0 = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-1/5}$$

Theoretically computed *h* value may not give the desired smoothing. In application choosing a very small *h* value will reduce the bias, while a large *h* value will result in an increase of the variance. Ballancing between bias – variance relation becomes a matter of trial and error. Therefore, finding the optimum *h* value by simulation, and checking for the minimization of mean square error (MSE) or average of the sum squared error (ASSE) can be one way of tackling this problem [11].

3.6 Kernel Regression Smoothing with Nadaraya–Watson Estimator

In any nonparametric regression smoothing process the smooth or average function $m_h(x)$ is estimated by the estimator $\hat{m}_h(x)$ given by

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} w_{hi}(x)\, \mathrm{y}_i$$

where y is the response variable and w_h is the weight function depending on the distance between *x* and X_i the i^{th} observed value of *X* and on the bandwidth *h*.

Nadaraya – Watson estimator is also using a weighting system as explained below.

Given n observations of i.i.d. random variables $\{(X_i, Y_i)\}_{i=1}^{n}$, $X_i \in \Re$, $Y_i \in \Re$, the conditional expectation

$$m(x) = E(Y \mid X = x) = \int y f(x, y) dy \,/\, f(x) \qquad (3.7)$$

can be written.

$f(x)$ can be estimated using a kernel density estimator. The joint density $f(x, y)$ in the numerator of equation (3.7) can be estimated using the multiplicative kernel

$$\hat{f}_{h_1, h_2}(x, y) = n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i) K_{h_2}(y - Y_i)$$

Then an estimator for the expression in the numerator of equation (3.7) is obtained as follows,

$$\int y\hat{f}_{h_1,h_2}(x,y)dy = n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)\int yK_{h_2}(y-Y_i)dy$$

$$= n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)\int \frac{y}{h_2}K_{h_2}\left(\frac{y-Y_i}{h_2}\right)dy$$

$$= n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)\int (sh_2+Y_i)K(s)ds$$

$$= n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)Y_i \qquad (3.8)$$

The estimate of the conditional expectation $m(x)$ given in equation (3.7) can be expressed as the ratio of the result obtained in equation (3.8) and the kernel estimate of *f(x)*. This ratio is the Nadaraya-Watson estimator expressed as;

$$m_h(x) = n^{-1}\frac{\sum_{i=1}^{n} K_h(x-X_i)Y_i}{n^{-1}\sum_{j=1}^{n} K_h(x-X_j)} \qquad (3.9)$$

In general the non-parametric regression smoother can be written as $\hat{m}_h(x) = n^{-1}\sum_{i=1}^{n} W_{hi}(x)Y_i$. The weights $W_{hi}(x)$ can be expressed as

$$W_{hi}(x) = \frac{h^{-1}K\left(\frac{x-X_i}{h}\right)}{\hat{f}_h(x)} \qquad (3.10)$$

The weights in equation (3.10) depends on the whole sample $\{X_j\}_{j=1}^{n}$ via $\hat{f}_h(x)$.

Higher weights are assigned to Y_i where X_i is sparse.

For situations where denominator is zero, the numerator is also zero. Meaning the estimate is zero.

When $h \to 0$, $W_{hi}(x) \to n$ if $x = X_i$. It means the estimate in X_i converges to Y_i.

When $h \to \infty$, $W_{hi}(x)$ converges to $1 \forall x$. Therefore, the estimate of $m(x)$ converges to $\bar{Y}$.

Bandwidth *h* determines the level of smoothness of the estimate.

3.7 Mean and Variance of the Nadaraya – Watson Estimator

Since the numerator and denominator of this statistic are both random variables, they can be dealt with separately. Starting with the numerator, let us define

$$r(x) = \int yf(x, y)dy = m(x) f(x) \text{ and } \hat{r}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i .$$

Then the regression curve estimate becomes

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)}$$

Theorem 3.2: The numerator $\hat{r}_h(x)$ of the Nadaraya-Watson smother is asymtotically unbiased.

Proof: Expectation of $\hat{r}_h$ is

$$E[\hat{r}_h(X)] = E[n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i] = E[K_h(x - X)]\mathrm{Y}$$

$$= \int\int yK_h(x-u) f(y|u) f(u)dydu = \int K_h(x-u) f(u)\left(\int yf(y|u)dy\right)du$$

$$= \int K_h(x-u) f(u)(E[Y|X = u])du = \int K_h(x-u) f(u)m(u)du = \int K_h(x-u)\,\mathrm{r}(u)du \quad (3.11)$$

Similar to density estimation with kernels, if $r \in C^2$, then

$$E[\hat{r}_h(x)] = r(x) + \frac{h^2}{2} r''(x)\mu_2(YK) + o(h^2)$$

indicating $r_h(x)$ is asymptotically unbiased as $h \to 0$.

Theorem 3.3: Using the variance of the denominator $\hat{r}_h(x)$ of the Nadaraya-Watson smother, it can be shown that $\hat{r}_h(x)$ is asymptotically consistent.

Proof: To find the variance of $\hat{r}_h(x)$ let $s^2(x) = E\left[Y^2 \middle| X = x\right]$, then

$$Var\left[\hat{r}_h(x)\right] = Var\left[n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i\right]$$

$$= n^{-1}Var\left[K_h(x - X)Y\right]$$

$$= n^{-1}\left\{\int K_h^2(x-u)s^2(u)f(u)du - \left(\int K_h(x-u)r(u)du\right)^2\right\}$$

$$= n^{-1}h^{-1}\int K_h^2 s^2(x+uh)f(x+uh)du + o((nh)^{-1})$$

$$= n^{-1}h^{-1}f(x)s^2(x)\|K\|_2^2 + o((nh)^{-1}) \qquad (nh \to \infty) \qquad (3.12)$$

Combining equations (3.11), (3.12) when $h \to 0, nh \to \infty$ the mean square error of $\hat{r}_h(x)$ becomes

$$MSE\left[\hat{r}_h(x)\right] = \frac{1}{nh} f(x)s^2(x)\|K\|_2^2 + \frac{h^4}{4}\left(r''(x)\mu_2(K)\right)^2 + o(h^4) + o\left((nh)^{-1}\right).$$

If $nh \to \infty$, $\text{MSE}\left[\hat{r}_h(x)\right] \to 0$. This means $\hat{r}_h(x)$ is a consistent estimator of $r(x)$. That is for any $c > 0$ and $c \to 0$, $\lim_{\substack{h\to 0 \\ nh\to\infty}} P\left[\left|\hat{r}_h(x) - r(x)\right| < c\right] = 1$. Shortly, $\hat{r}_h(x) \xrightarrow{P} r(x)$.

Theorem 3.4: The denominator $\hat{f}_h(x)$ of the Nadaraya-Watson smother is asymtotically unbiased.

Proof: Since $X_i;\ i = 1, \ldots, n$ are i.i.d. we have

$$E\left[\hat{f}_h(x)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[K_h(x - X_i)\right]$$

$$= E\left[K_h(x - X)\right]$$

$$= \int K_h(x-u)f(u)du$$

$$= \int K(s) f(x+sh) ds$$

Letting $h \to 0$ results in

$$E\left[\hat{f}_h(x) \right] \to f(x) \int K(s) ds = f(x) \ .$$

When bandwidth *h* convergence to 0 the $E\left[\hat{f}_h(x) \right]$ is asymptotically unbiased *Q.E.D.*

In general bias can be analyzed using the Taylor expansion of $f(x+sh)$ in *x* assuming *f* is twice continuously differentiable ($f \in C^2$).

$$\text{Bias}\left[\hat{f}_h(x) \right] = \int K(s) f(x+sh) ds - f(x)$$

$$= \int K(s) \left[f(x) + shf'(x) + \frac{h^2 s^2}{2} f''(x) + o(h^2) \right] ds - f(x)$$

$$= f(x) + \frac{h^2}{2} f''(x) \mu_2(K) + o(h^2) - f(x) \qquad (3.13)$$

Proof of equation (3.13) see [14].

Due to symmetric property of *K* around 0, the term $\int sK(s) hf'(x) ds = 0$. Then the bias of kernel density becomes

$$\text{Bias}\left[\hat{f}_h(x) \right] = \frac{h^2}{2} f''(x) \mu_2(K) + o(h^2) \ , \ h \to 0 \qquad (3.14)$$

Since equation (3.14) contains h^2 , it means the bandwidth should not be very big to avoid large bias values. Also, the bias is proportional to f'' in *x*. It means $E\left[\hat{f}_h(\bullet) \right]$ will be greater than the true value $f(x)$ when estimating points around a local minimum ($f''(x) > 0$), and it will be greater than $f(x)$ when estimating points around a local maximum ($f''(x) < 0$).

Theorem 3.5: Using the variance of the denominator $\hat{f}_h(x)$ of the Nadaraya-Watson smother, it can be shown that $\hat{f}_h(x)$ is asymptotically consistent.

Proof: Since $X_i;\ i=1,\ldots,n$ are i.i.d.

$$\mathrm{Var}\left[\hat{f}_h(x)\right] = n^{-2}\mathrm{Var}\left[\sum_{i=1}^{n} K_h(x-X_i)\right]$$

$$= n^{-2}\sum_{i=1}^{n}\mathrm{Var}\left[K_h(x-X_i)\right]$$

$$= n^{-1}\mathrm{Var}\left[K_h(x-X)\right]$$

$$= n^{-1}\left\{E\left[K_h^2(x-X)\right]-\left(E\left[K_h(x-X)\right]\right)^2\right\}$$

$$= n^{-1}\left\{h^{-2}\int K^2\left(\frac{x-u}{h}\right)f(u)du-\left(f(x)+o(h)\right)^2\right\}$$

$$= n^{-1}\left\{h^{-1}\int K^2(s)\,f(x+sh)ds-\left(f(x)+o(h)\right)^2\right\}$$

$$= n^{-1}\left\{h^{-1}\|K\|_2^2\left(f(x)+o(h)\right)-\left(f(x)+o(h)\right)^2\right\}.$$

From equation (3.13) we have $E\left[K_h(x-X)\right] = f(x)+o(h)$ and

$$\int K^2(s)\,f(x+sh)ds = \int K^2(s)ds\left(f(x)+o(h)\right) = \|K\|_2^2\left(f(x)+o(h)\right)$$

From here

$$\mathrm{Var}\left[\hat{f}_h(x)\right] = (nh)^{-1}\|K\|_2^2\ f(x)+o\left((nh)^{-1}\right),\ nh\to\infty. \qquad (3.15)$$

It is obvious that $(nh)^{-1}$ has a strong influence on variance, leading to larger values of *h* to reduce the variance. On the other hand, small *h* value is desired for decrease in bias. If we consider the MSE combining the variance and square bias of $\hat{f}_h(x)$, as $h\to 0$ and $nh\to\infty$ we have

$$\mathrm{MSE}\left[\hat{f}_h(x)\right] = \frac{1}{nh}f(x)\|K\|_2^2+\frac{h^4}{4}\left(f''(x)\mu_2(K)\right)^2+o\left((nh)^{-1}\right)+o(h^4)$$

Then the kernel density estimate is consistent satisfying $\hat{f}_h(x) \xrightarrow{P} f(x)$ *Q.E.D.*

MSE establishes balance between variance and bias such that

i. Decreasing variance results in under smoothing.

ii. Decreasing bias results in over smoothing.

Note that based on MSE the optimal bandwidth kernel density can be determined as

$$h_0 = \left(\frac{f(x)\|K\|_2^2}{\left(f''(x)\right)^2 \left(\mu_2(K)\right)^2 n} \right)^{1/5}.$$

See reference [10], page 59.

3.8 Scatter Plots

Usually, called scatter diagram, but which have many names, such as scatter plot, scatter graph and correlation chart etc., are important steps in the study of the bivariate data set. In this diagram the independent variable is shown on the *x*-axis while the dependent one goes on the *y*-axis. A scatter plot is a good tool that visualy shows the relationship between the dependent and independent variables. For the airquality data the scatter diagram is given in Figure 3.3 [26]. A visiaul inspection of scatter plot shows some kind of linear relation between the two variables. The linear correlation coefficient $r =$ -0.6469 confirm this.

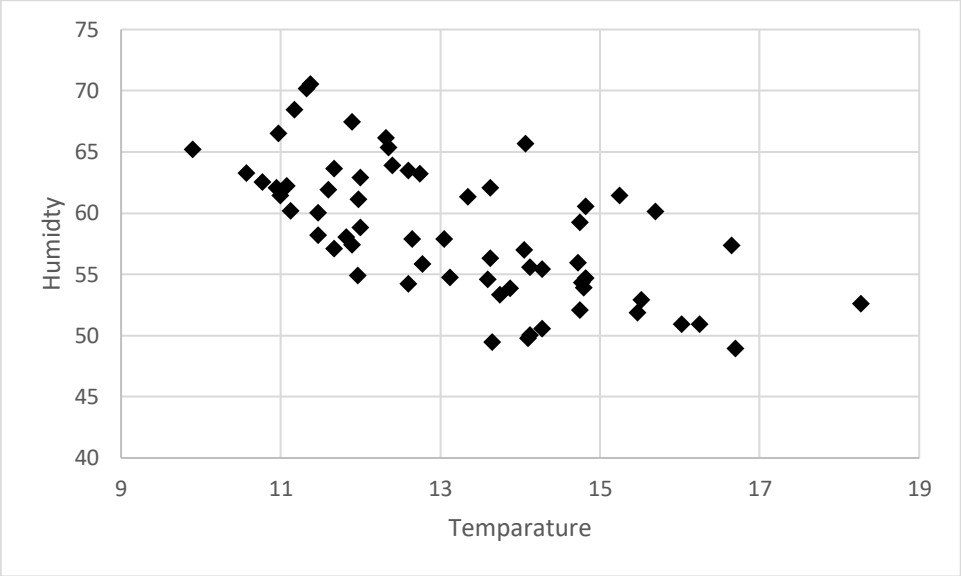

Figure 3.3: Scatter diagram where temparature is independent, humidty is dependent variable

3.9 Application of Kernel Smoothing

In the previous section some of the important aspects of kernel smoothing are given. Application of kernel smoothing nececiates the use of computer support in order to handle the huge amount of number crunching to reach at the required results. Especially the theoretically proposed band width *h* may not always produce the desired results, in terms of minimizing the error term involved in the smoothing process. A simulation approach based on the use of different band width values and comparing associated errors tends to be more productive in determining the value of *h*.

A data set consisting of 63 observations with 2 variables, where air temperature in ^{0}C taken as the independent variable *X*, and relative humidity is the dependent variable *Y*.

A scatter diagram of the data given in Figure 3.3 indicated some kind of linear relation between the two variables. The degree of the relationship is computed as a linear

correlation coefficient value of 0.65. This was considered a reasonable degree of correlation between the two variables to continue for further analysis of the data.

For kernel regression the Gaussian kernel $e^{\left(-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2\right)}$ was used to compute the kernel values. $X_{\min} = 9.9000001$, $X_{\max} = 18.275$, giving a range of $8.3749999\,^0C$. Following the careful examination of the data values for the variable *X*, it was considered adequate to use an increment value *dx=0.03* to be used in the computation of the kernel values. Computations were carried out using a set of 5 different band widths $h = (0.05,\ 0.1,\ 0.3,\ 0.5,\ 2.3)$. For each band width the estimated values $\hat{Y}$ were computed using the Nadaraya – Watson estimator given in equation (3.9).

For different bandwidths the computed kernel smoothers are given in Figures 3.4 to 3.8. It is clearly visible that the smaller band width h=0.05 produced very little smoothing, the kernel graph producing wild fluctuations, and increasing band width reduced the fluctuations resulting in smoother curves as *h* approaches 0.5 as seen Figure 3.7. Band width computed using equation (3.5) h=2.3, is also used for comparison and as seen in Figure 3.8 has produced very smooth curve almost equivalent to linear regression line. This emphasizes our argument that the determination of the optimal band width can be obtained through simulation and checking for resulting MSE, bias and variance values. Statistically $\text{MSE} = \frac{\sum\left(Y - Y\right)^2}{n}$. From table 3.1 an examination of MSE, bias and variance values suggests h=0.1 can be good candidate for optimal band width.

This was followed by the comparison of Mean Square Error (MSE), bias, and variance values. It can easily be observed that an increase in the band width results in an increase in MSE (Figure 3.9), while a decrease is observed in the variance of estimates (Figure 3.10). Behavior of the bias in relation to band widh is shown in Figure 3.11. Bias is computed as given below

$$\text{Bias} = \text{E}\left[\left|\hat{Y} - Y\right|\right].$$

A clear increases as h increase.

The results we obtained from these calculations are as follows;

Table 3.1: MSE, Bias and Variance values obtained from the kernel analysis of the data

	MSE	Bias	$\text{Var}(\hat{Y})$
h=0.05	8.258801	2.0428818	27.71789
h=0.1	10.221633	2.4617001	24.13091
h=0.3	14.56699	3.1498393	18.705991
h=0.5	15.42330	3.2387629	17.137332
h=2.3	20.97414	3.8808349	3.9560030

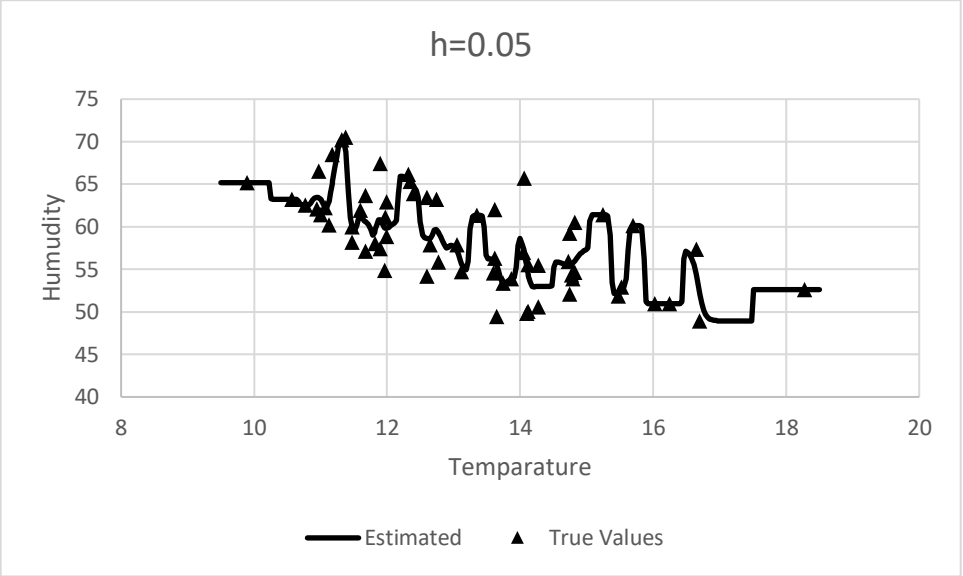

Figure 3.4: Kernel estimator for band width h=0.05

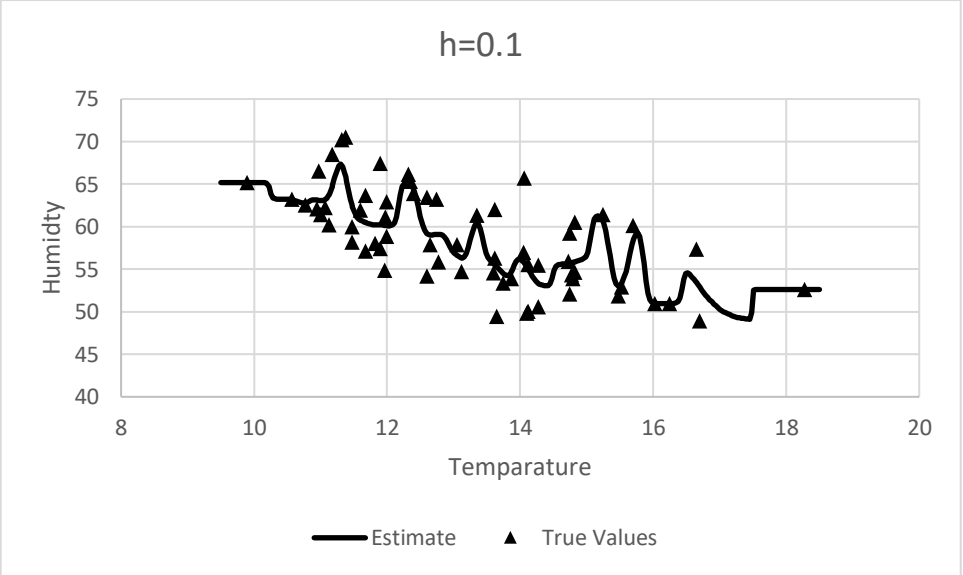

Figure 3.5: Kernel estimator for band width h=0.1

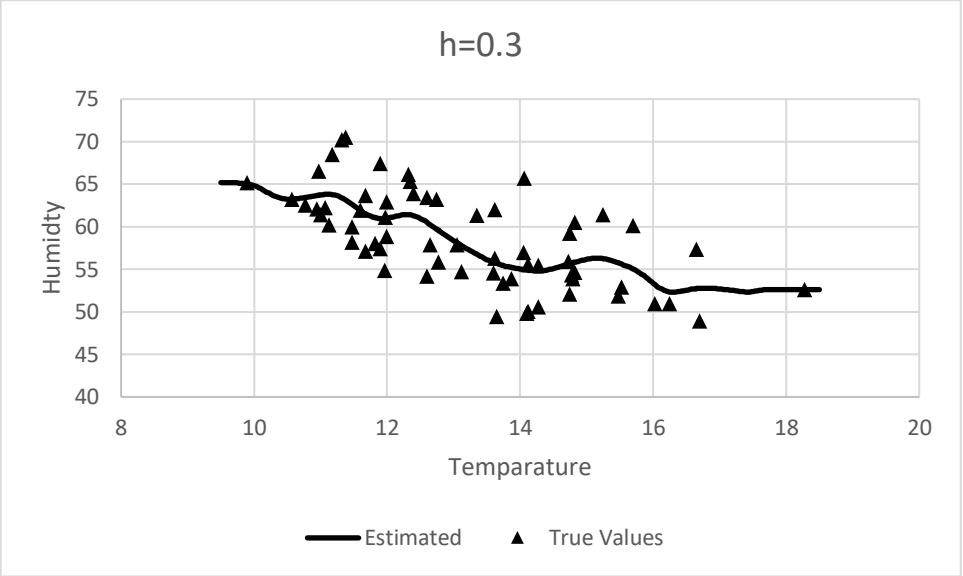

Figure 3.6: Kernel estimator for band width h=0.3

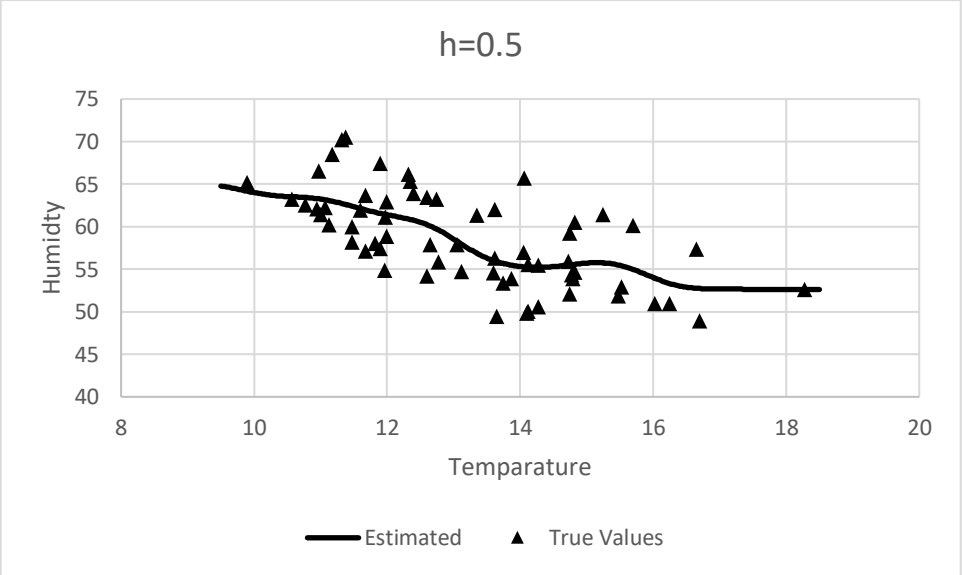

Figure 3.7: Kernel estimator for band width h=0.5

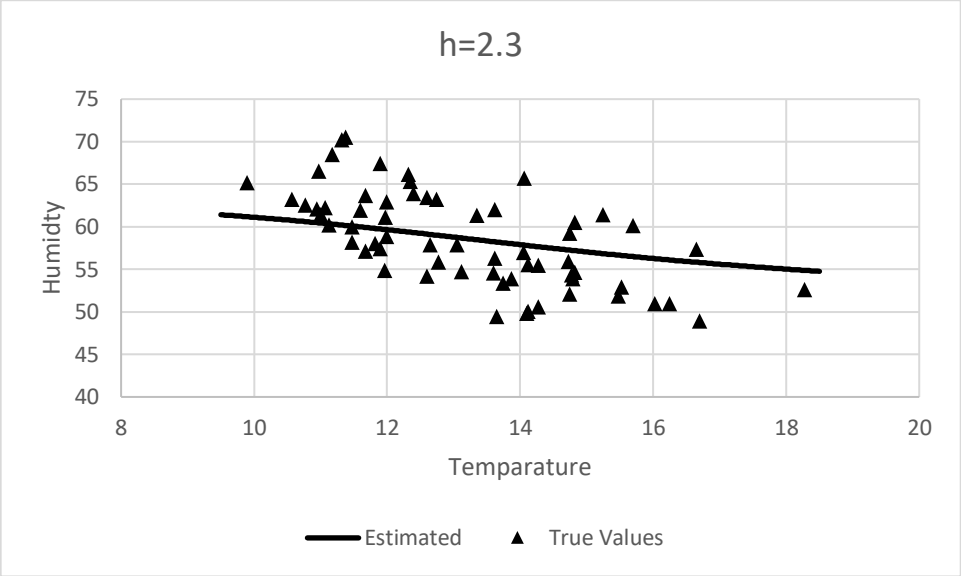

Figure 3.8: Kernel estimator for band width h=2.3

We repeated process for 5 different band widths in order to see the relationship between h, error, variance and bias. The graphics we obtained from these operations are as follows;

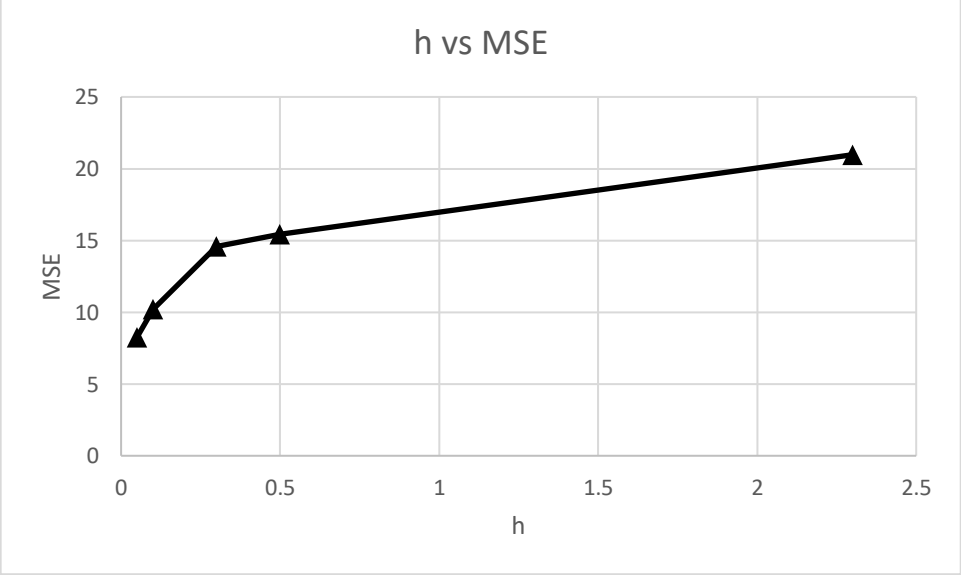

Figure 3.9: MSE as a function of a band width

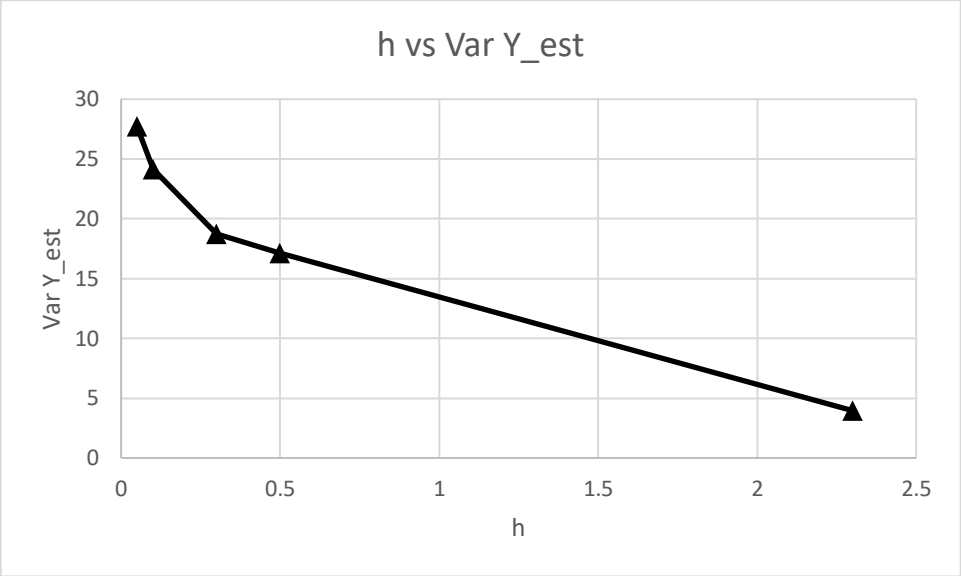

Figure 3.10: Variance as a function of a band width

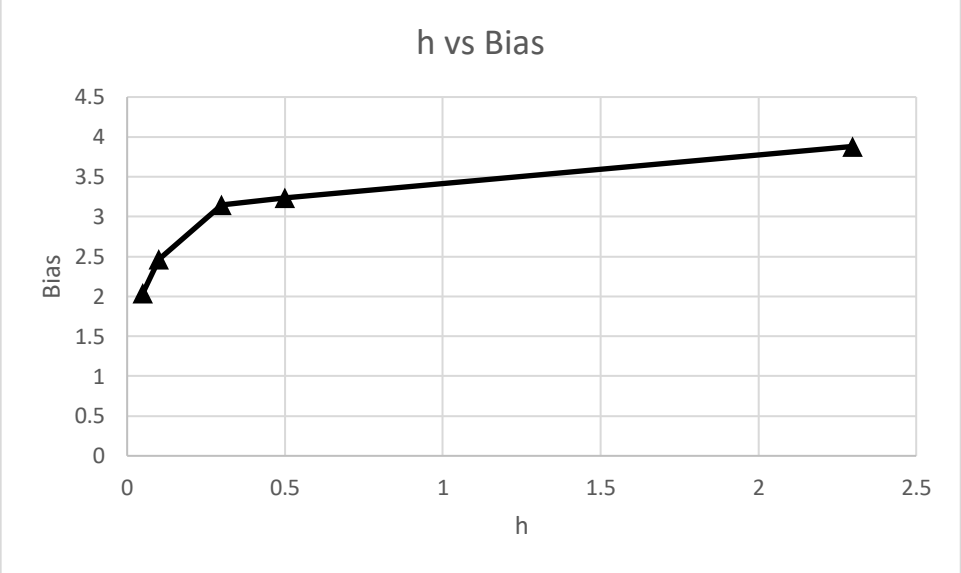

Figure 3.11: Bias as a function of a band width

Chapter 4

PRINCIPAL COMPONENT ANALYSIS

4.1 What is Principal Component Analysis

Multivariate statistics deals with situations where there are more than one variables. When the number of variables (*p*) is too large, the analysis of multivariate data becomes difficult as *p* becomes larger. For this reason, a method was developed to reduce the number of variables. This method, derives linear combinations of actual variables, which are called principal components (PC) [14,28]. The number of variables and the number of principal component is the same. Nevertheless, we can represent more than %90 of the total variation in the data with only the first few principal components. The systematic reduction of a large number of independent variables to smaller dimension is done by the principal component analysis (PCA). Principal component analysis, is a statistical technique that transforms data represented by a large number of variables into a smaller number of uncorelated variables or PCs. It requires a lot more effort to interpreat the uncorrelated PCs than to understand a large set of correlated variables [6,21].

4.2 Concept of PCA

Algebraically, the PCs are certain linear combinations of *p* random variables. The principal components are only dependent on the covariance $\mathbf{\Sigma}$ or correlation $\boldsymbol{\rho}$ matrices of the data matrix $\mathbf{X}$. The eigenvalues obtained from $\mathbf{\Sigma}$ or $\boldsymbol{\rho}$ are listed as, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. The corresponding eigenvectors are $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p$. Then the PCs

are written as linear combinations of the p variables where the coefficients are the elements of the eigenvectors as given below.

$$\begin{aligned} Y_1 &= \mathbf{e}_1'\mathbf{X} = e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_p \\ &\vdots \\ Y_p &= \mathbf{e}_p'\mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + \ldots + e_{pp}X_p \end{aligned} \tag{4.1}$$

Understanding the theory of PCA requires a clear understanding of the covariance and correlation concepts. Hence, before whenturing further into PCA analysis, it is consider necessary to explain the covarince, correlation concepts in detail [17].

4.2.1 Abstract Definition of Covariance and Correlation

Definition 4.1: Let random variable X with probability density function $f(x)$ and mean μ, then the variance of X is given by

$$\sigma^2 = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 f(x) \qquad [18]$$

When p random variables with joint probability density $f(x_1, x_2, \ldots, x_p)$ is given, then the covariance matrix becomes $\boldsymbol{\Sigma} = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'$

$$= E\left(\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_p-\mu_p \end{bmatrix} \left[X_1-\mu_1, X_2-\mu_2, \ldots, X_p-\mu_p \right] \right)$$

$$= E\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \cdots & (X_p-\mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E\left(X_1-\mu_1\right)^2 & E\left(X_1-\mu_1\right)(X_2-\mu_2) & \cdots & E\left(X_1-\mu_1\right)\left(X_p-\mu_p\right) \\ E\left(X_2-\mu_2\right)\left(X_1-\mu_1\right) & E\left(X_2-\mu_2\right)^2 & \cdots & E\left(X_2-\mu_2\right)\left(X_p-\mu_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(X_p-\mu_p\right)\left(X_1-\mu_1\right) & E\left(X_p-\mu_p\right)\left(X_2-\mu_2\right) & \cdots & E\left(X_p-\mu_p\right)^2 \end{bmatrix}$$

or

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \tag{4.2}$$

$\boldsymbol{\mu}$ is the population mean vector and $\boldsymbol{\Sigma}$ is the population variance-covariance matrix. However, when the *p* variables have different units or the magnitute of data values are significantly different, then it is wise to standardize the data and use the correlation matrix for the computation of PCs.

Theorem 4.1: Given random variables X_i and X_j with joint probability density function $f(x_i, x_j)$ and if $f(x_i, x_j) = f(x_i) f(x_j)$ indicating the independence of the random variables X_i and X_j, then the covariance $\sigma_{X_i X_j} = 0$.

Proof: For the discrete case we have,

$$E\left(X_i X_j\right) = \sum_{x_i} \sum_{x_j} x_i x_j \cdot f(x_i, x_j)$$

Since X_i and X_j are independent, we can write $f(x_i, x_j) = f_{x_i}(x_i) \cdot f_{x_j}(x_j)$, where $f_{x_i}(x_i)$ and $f_{x_j}(x_j)$ are the marginal distributions of X_i and X_j, then

$$E\left(X_i X_j\right) = \sum_{x_i} \sum_{x_j} x_i x_j \cdot f(x_i) f(x_j)$$

$$= \left[\sum_{x_i} x_i \cdot f(x_i)\right]\left[\sum_{x_j} x_j \cdot f(x_j)\right]$$

$$= E\left(X_i\right) \cdot E\left(X_j\right)$$

Hence,

$$\sigma_{X_i X_j} = E(X_i X_j) - E(X_i) \cdot E(X_j)$$

$$= E(X_i) \cdot E(X_j) - E(X_i) \cdot E(X_j)$$

$$= 0$$

Example: Let X_1 and X_2 be two random variables, with joint distrubition;

$$f(x_1, x_2) = \begin{cases} e^{-x_1 - x_2} & \text{for } x_1 > 0,\ x_2 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$f_{x_1}(x_1) = \int_0^\infty e^{-x_1 - x_2} dx_1 = \left(-e^{-x_1 - x_2}\right)_0^\infty = \left(-\frac{1}{e^{x_1 + x_2}}\right)_0^\infty = e^{-x_1}$$

$$f_{x_2}(x_2) = \int_0^\infty e^{-x_1 - x_2} dx_1 = \left(-e^{-x_1 - x_2}\right)_0^\infty = \left(-\frac{1}{e^{x_1 + x_2}}\right)_0^\infty = e^{-x_2}$$

Random variables X_1, X_2 are independent iff $f(x_1, x_2) = f_{x_1}(x_1) \cdot f_{x_2}(x_2)$. Applying this condition to the example

$$e^{-x_1 - x_2} \stackrel{?}{=} e^{-x_1} \cdot e^{-x_2}$$

it is seen that the independence condition is satisfied, leading to $\sigma_{X_1 X_2} = 0$, as shown

$$E\left[X_1\right] = \int_0^\infty x_1 e^{-x_1} dx_1 = \left(\frac{-x_1 - 1}{e^{x_1}}\right)_0^\infty = 1$$

$$E\left[X_2\right] = \int_0^\infty x_2 e^{-x_2} dx_2 = \left(\frac{-x_2 - 1}{e^{x_2}}\right)_0^\infty = 1$$

$$E\left[X_1 X_2\right] = \int_0^\infty \int_0^\infty x_1 x_2 e^{-x_1 - x_2} dx_2 dx_1 = \int_0^\infty x_1 e^{-x_1} \left(\int_0^\infty x_2 e^{-x_2} dx_2\right) dx_1 = \int_0^\infty x_1 e^{-x_1} \left(\left(-x_2 e^{-x_2}\right)_0^\infty\right) dx_1 = \int_0^\infty x_1 e^{-x_1} dx_1 = 1$$

Hence,

$$Cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2] = 1 - 1 \cdot 1 = 0$$

Theorem 4.2: When $\sigma_{X_i X_j} = 0$ does not always indicate the independence of X_i and X_j.

Example: As an example, take $f(x_i, x_j) = x_i x_j$ for discreate random variables X_i and X_j with the joint probability distribution given in the table.

$f(x_i, x_j)$		x_i		
		0	1	2
x_j	0	0	0.25	0
	1	0.25	0	0.25
	2	0	0.25	0

The expectation of X_i X_j is computed as follows:

$$E[X_i] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

$$E[X_j] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

$$E[X_i X_j] = (0 \cdot 0) \cdot 0 + (1 \cdot 0) \cdot \frac{1}{4} + (2 \cdot 0) \cdot 0 + (0 \cdot 1) \cdot \frac{1}{4} + (1 \cdot 1) \cdot 0 + (2 \cdot 1) \cdot \frac{1}{4} + (0 \cdot 2) \cdot 0 + (1 \cdot 2) \cdot \frac{1}{4} + (2 \cdot 2) \cdot 0 = 1$$

Hence,

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = 1 - 1 \cdot 1 = 0$$

$$f(x_i, x_j) \stackrel{?}{=} f(x_i) \cdot f(x_j)$$

For example $f(1,0) \stackrel{?}{=} f(1) f(0) \rightarrow \frac{1}{4} \neq \frac{1}{4} \cdot \frac{1}{2}$

Hence, while $\sigma_{X_i X_j} = 0$, the random variables X_i and X_j are not indepedent.

Definition 4.2: Correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The population correlation coefficient for any two variables $X_i,\ X_j\ ;\ i, j = 1,\ldots, p$ is denoted by ρ , and defined by

$$\rho_{X_i,X_j} = \frac{\text{cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

To express ρ in terms of expectations, the covariance between the variables $X_i,\ X_j$, and standard deviations of the variables $X_i,\ X_j$ have to be expressed in terms of expectations.

Note:

First moment about the origin: $\mu_{X_i} = E\left[X_i\right]$

Second moment about the mean: $\mu_2 = E\left[\left(X - E\left[X\right]\right)^2\right] = \int_{\Re}(x-\mu)^2 f(x)dx$

$$\sigma_{X_i}^2 = E\left[\left(X_i - E\left[X_i\right]\right)^2\right] = E\left[X_i^2\right] - \left[E\left[X_i\right]\right]^2$$

$$E\left[\left(X_i - \mu_{x_i}\right)\left(X_j - \mu_{X_j}\right)\right] = E\left[\left(X_i - E\left[X_i\right]\right)\left(X_j - E\left[X_j\right]\right)\right] = E\left[X_i X_j\right] - E\left[X_i\right]E\left[X_j\right]$$

$$\text{cov}\left(X_i, X_j\right) = \sigma_{X_iX_j} = E\left[\left(X_i - \mu_{x_i}\right)\left(X_j - \mu_{X_j}\right)\right]$$

the formula for ρ becomes

$$\rho_{X_i,X_j} = \frac{E\left[\left(X_i - \mu_{x_i}\right)\left(X_j - \mu_{X_j}\right)\right]}{\sigma_{X_i}\sigma_{X_j}}$$

Expressing the covariance in terms of moments about the origin, yields

$$\rho_{X_i,X_j} = \frac{E\left[X_i X_j\right] - E\left[X_i\right]E\left[X_j\right]}{\sqrt{E\left[X_i^2\right] - \left[E\left[X_i\right]\right]^2}\sqrt{E\left[X_j^2\right] - \left[E\left[X_j\right]\right]^2}}$$

Considering the fact that, $\sigma_{ik} = \sigma_{ii};\ when\ i = k$, the correlation coefficient can be written as

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}} = \frac{\sigma_{ii}}{\sigma_{ii}} = 1$$

To show that $-1 \le \rho \le 1$ we utilize the Cauchy – Schwarz Inequality

$$\left|Cov(X_i, X_j)\right|^2 \le Var(X_i)Var(X_j)$$

$$\therefore \ \left|Cov(X_i, X_j)\right| \le \sqrt{Var(X_i)Var(X_j)}$$

substitute this from Cauchy-Schwarz inequality into the equation formul

$$\left|\rho\right| = \left|\frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}\right| \le \frac{\sqrt{Var(X_i)Var(X_j)}}{\sqrt{Var(X_i)Var(X_j)}} = 1 \ \rightarrow \ \text{-}1 \le \rho \le 1 \qquad [11]$$

ρ is a measure of a linear correlation coefficient between the random variables X_i and X_j.

Given p random variables the correlation matrix can be written as

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{1p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

Let the $p \times p$ standard deviation matrix be

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{bmatrix} \tag{4.3}$$

Then it can easily be verified that

$$\boldsymbol{\rho} = \left(\mathbf{V}^{1/2}\right)^{-1} \boldsymbol{\Sigma} \left(\mathbf{V}^{1/2}\right)^{-1}$$

Theorem 4.3: Consider the following q linear combinations of p random variables

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

The above equation can be written in matrix format as follows.

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

It follows that;

a) $\boldsymbol{\mu}_Z = E(\mathbf{Z}) = E(\mathbf{CX}) = \mathbf{C}\mu_x$

b) $\boldsymbol{\Sigma}_Z = \mathrm{Cov}(\mathbf{Z}) = \mathrm{Cov}(\mathbf{CX}) = \mathbf{C}\boldsymbol{\Sigma}_X\mathbf{C}'$

Proof for part a : For a simple random variable X_1 and constant c;

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

$$\mathrm{Var}(cX_1) = E(cX_1 - c\mu_1)^2 = c^2\mathrm{Var}(X_1)$$

can be computed.

For two random variables X_1 and X_2 and constants a, b, the covariance matrix can be computed as follows

$$\text{Cov}(aX_1, bX_2) = E(aX_1 - a\mu_1)E(bX_2 - b\mu_2)$$

$$= abE(X_1 - \mu_1)(X_2 - \mu_2)$$

$$= ab\text{Cov}(X_1, X_2) = ab\sigma_{12}$$

If we have combination of two variables $aX_1 + bX_2$, then

$$Z = E(aX_1 + bX_2) = aE(X_1) + bE(X_2) = a\mu_1 + b\mu_2$$

$$\text{Var}(aX_1 + bX_2) = E\left[(aX_1 + bX_2) - (a\mu_1 - b\mu_2)\right]^2$$

$$= E\left[a(X_1 - \mu_1) + b(X_2 - \mu_2)\right]^2$$

$$= E\left[a^2(X_1 - \mu_1)^2 + b^2(X_2 - \mu_2)^2 + 2ab(X_1 - \mu_1)(X_2 - \mu_2)\right]$$

$$= a^2\text{Var}(X_1) + b^2\text{Var}(X_2) + 2ab\text{Cov}(X_1, X_2)$$

$$= a^2\sigma_{11} + b^2\sigma_{22} + 2ab\sigma_{12}$$

with $\mathbf{c}' = [a, b],$ and $aX_1 + bX_2$ can be written as

$$\begin{bmatrix} a & b \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mathbf{c}'\mathbf{X}.$$

Also, $E(aX_1 + bX_2) = a\mu_1 + b\mu_2$ in matrix format will be

$$\begin{bmatrix} a & b \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mathbf{c}'\boldsymbol{\mu}$$

If $\boldsymbol{\Sigma}$ is expressed as in equation 4.2 for bivariate case we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

then the variance-covariance matrix of $\mathbf{X}$,can be written as,

$$\text{Var}(aX_1 + bX_2) = \text{Var}(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$$

since,

$$\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} = [a \quad b]\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix} = a^2\sigma_{11} + 2ab\sigma_{12} + b^2\sigma_{22}$$

In the case of a linear combination with *p* random variables: $\mathbf{c}'\mathbf{X} = c_1X_1 + \ldots + c_pX_p$ has mean $E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'\boldsymbol{\mu}$, where $\boldsymbol{\mu} = E(\mathbf{X})$.

Proof for part b can similarly be done.

4.2.2 Statistical Definition of Covariance and Correlation

In order to understand and compute the covariance, the sample mean and sample variance concepts has to be understood.

4.2.2.1 Sample Mean

Sample mean is the arithmetic average of *n* observations $(x_1,\ldots,x_n)$ taken at random from the population represented by the random variable *X*. Computing the mean of *p* variables involved in a process will obviously result in a vector of average values $\bar{\mathbf{x}}' = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$

Definition 4.3: If *X* represents a set of *n* observations, $(x_1,\ldots,x_n)$, then the mean of data is the scaler $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Definition 4.4: Consider the $n \times p$ data matrix **X=**
$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1j} & \ldots & x_{1p}\\ x_{21} & x_{22} & \ldots & x_{2j} & \ldots & x_{2p}\\ \vdots & & & & & \\ x_{i1} & x_{i2} & \ldots & x_{ij} & \ldots & x_{ip}\\ \vdots & & & & & \\ x_{n1} & x_{n2} & \ldots & x_{nj} & \ldots & x_{np}\end{pmatrix}$$

where *p* variables involved with *n* observations in each variable.

The sample vector mean is then defined as $\bar{\mathbf{x}}' = \frac{1}{n}\mathbf{1}_n'\mathbf{X}$.

Let $\mathbf{1}_n$ be the column vector of 1s. The column vector representing p sample averages is denoted $\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$.

The expectation of the sample mean vector is given by

$$E(\bar{\mathbf{x}}) = E\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} E(\bar{x}_1) \\ E(\bar{x}_2) \\ \vdots \\ E(\bar{x}_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

Definition 4.5: Given random variable X_1 with mean μ_1 the $E(cX_1) = cE(X_1) = c\mu_1$.

Definition 4.6: X is the random variable consisting of n observations $(x_1, \ldots, x_n)$, with mean $\bar{x}$. Then the variance is $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.

The nature of the binary linear relationship between the variables needs to be determined. The covariance and linear correlation coefficients are used this purpose. The strength of the linear relationship between the variables becomes evident in the correlation coefficient.

When there are two random variables, X_1 and X_2 the relationship between them can be determined by the covariance given in definition 4.7.

Definition 4.7: Given random variables X_i and X_j with joint probability $f(x_i, x_j)$ then the covariance between X_i and X_j is given by (4.4).

$$S_{X_i X_j} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}_i\right)\left(x_j - \bar{x}_j\right)}{n-1} \qquad (4.4)$$

When the number of variables is more than 2, then each element of the covariance matrix (***S***) can be computed using equation (4.4). Diagonal elements of ***S*** matrix will be the variance of individual variables, off diagonal elements will be the covariances between the variables X_i, X_j ; $i, j = 1, \ldots, p$. Covariance matrix is symmetric.

Definition 4.8: Consider two random variables X_i and X_j with *n* observations then the correlation coefficient between X_i, X_j is given by $r_{X_i X_j} = \dfrac{\mathrm{Cov}\left(X_i, X_j\right)}{\sqrt{\mathrm{Var}\left(X_i\right)\mathrm{Var}\left(X_j\right)}}$

Note that $-1 \le r \le 1$.

4.3 Theory of PCA

In order to understanding the theory of PCA the following theorems are given.

Theorem 4.4: Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X}' = \left[X_1, X_2, \ldots, X_p\right]$. Let $\boldsymbol{\Sigma}$ have the eigenvalue-eigenvector pairs $\left(\lambda_1, \mathbf{e_1}\right), \ldots \left(\lambda_p, \mathbf{e_p}\right)$ where $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_p \ge 0$. Then the i^{th} principal component is given by

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p, \quad i = 1, 2, \ldots, p$$

it can be shown that,

$$\mathrm{Var}\left(Y_i\right) = \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i, \quad i = 1, 2, \ldots, p$$

$$\mathrm{Cov}\left(Y_i, Y_k\right) = \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_i = 0, \quad i \ne k$$

Proof: Algebraically it can be shown that the first eienvalue of a square symmetric matrix is

$$\max_{a\neq 0} \frac{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_1 \qquad (4.5)$$

when $\mathbf{a} = \mathbf{e}_1$. However $\mathbf{e}_1'\mathbf{e}_1 = 1$as the eigenvectors are normalized. Then,

$$\max_{a\neq 0} \frac{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_1 = \frac{\mathbf{e}_1'\mathbf{\Sigma}\mathbf{e}_1}{\mathbf{e}_1'\mathbf{e}_1} = \mathbf{e}_1'\mathbf{\Sigma}\mathbf{e}_1 = \mathrm{Var}(Y_1)$$

In a similar fashion

$$\max_{a\perp e_1, e_2, \ldots, e_k} \frac{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_{k+1} \qquad k = 1, 2, \ldots, p-1 \text{ can be written.}$$

If we chose $\mathbf{a} = \mathbf{e}_{k+1}$, with $\mathbf{e}_{k+1}'\mathbf{e}_i = 0, \quad \text{for } i = 1, 2, \ldots, k \text{ and } k = 1, 2, \ldots, p-1,$

$\mathbf{e}_{k+1}'\mathbf{\Sigma}\mathbf{e}_{k+1} / \mathbf{e}_{k+1}'\mathbf{e}_{k+1} = \mathbf{e}_{k+1}'\mathbf{\Sigma}\mathbf{e}_{k+1} = \mathrm{Var}(Y_{k+1})$

On the other hand

$\mathbf{e}_{k+1}'\left(\mathbf{\Sigma}\mathbf{e}_{k+1}\right) = \lambda_{k+1}\mathbf{e}_{k+1}'\mathbf{e}_{k+1} = \lambda_{k+1}$ so $\mathrm{Var}\left(Y_{k+1}\right) = \lambda_{k+1}$.

That is $\mathbf{e}_i$ and $\mathbf{e}_k$ are perpendicular (orthogonal) which means $\mathbf{e}_i'\mathbf{e}_k = 0, \ i \neq k,$ resulting in

$\mathrm{Cov}(Y_i, Y_k) = 0$. Then,

1. If all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are distinct, the eigenvectors of $\mathbf{\Sigma}$ will be orthogonal.
2. If the eigenvalues are not all distinct, the eigenvectors corresponding to common eigenvalues may be selected to be orthogonal. So, for any two eigenvectors $\mathbf{e}_i, \mathbf{e}_k$ where $\mathbf{e}_i'\mathbf{e}_k = 0, \ i \neq k.$

$\mathbf{\Sigma}\mathbf{e}_k = \boldsymbol{\lambda}_k\mathbf{e}_k$ Multiplying by $\mathbf{e_i}'$ gives

$\mathrm{Cov}(Y_i, Y_k) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_k = \mathbf{e}'\boldsymbol{\lambda}_k\mathbf{e}_k = \boldsymbol{\lambda}_k\mathbf{e}'\mathbf{e}_k = 0$

For any $i \neq k$.

Note: It must be stressed that some of the coefficient vectors $\mathbf{e}_i$ and corresponding Y_i will not be unique, if some eigenvalues λ_i are equal. *Q.E.D.*

Theorem 4.5: Consider a vector of p random variables $\mathbf{X} = \left[X_1, \ldots, X_p \right]$ with its associated covariance matrix $\mathbf{\Sigma}$, computed from *n* observations of *p* random variables. Then the eigenvalue – eigenvector pairs $(\lambda_1, e_1), \ldots (\lambda_p, e_p)$ of the covariance matrix $\mathbf{\Sigma}$, such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ and the PCs are

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

It can be shown that the following relationship holds

$$\sigma_{X_1 X_1} + \sigma_{X_2 X_2} + \ldots + \sigma_{X_p X_p} = \sum_{i=1}^{p} \text{var}\left(X_i \right) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} \text{var}\left(Y_i \right) \qquad [16]$$

Proof: It is known that $tr(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$ where $\mathbf{A} = \{a_{ij}\}$ indicates a $k \times k$ square matrix. Applying this to the covariance matrix $\mathbf{\Sigma}$ yields

$\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = tr(\mathbf{\Sigma})$, then by using $\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_{i_{(k\times 1)}} \mathbf{e}'_{i_{(1\times k)}} = \mathbf{P}_{(k\times k)} \mathbf{\Lambda}_{(k\times k)} \mathbf{P}'_{(k\times k)}$ with

$\mathbf{A} = \mathbf{\Sigma}$

$\mathbf{\Sigma} = \mathbf{P \Lambda P'}$ where $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues and, $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p]$

Then $\mathbf{P'P = PP' = I}$.

$tr(\mathbf{\Sigma}) = tr(\mathbf{P \Lambda P'}) = tr(\mathbf{\Lambda PP'})$, Since $tr(\mathbf{AB}) = tr(\mathbf{BA})$.

Then, $tr(\mathbf{\Lambda PP'}) = tr(\mathbf{\Lambda}) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$.

Then, $\sum_{i=1}^{p} \text{Var}(\mathbf{X}_i) = tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda}) = \sum_{i=1}^{p} \text{Var}(\mathbf{Y}_i)$. *Q.E.D.*

It is worth noting total population variance = $\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p$

Proportion of total population variance due to k^{th} principal component = $\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}$;

$k = 1, 2, \ldots, p$. (4.6)

The number of PCs to be used in the analysis of data (*k*) is determined by the ratio

$$\tau = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \qquad (4.7)$$

In general $\tau > 0.9$ is preferred.

Another method that helps determine the most suitable number of component is the scree plot method.

Scree Plot, a simple line segment graph, represents the fraction of the total variance in the data described or represented by each PC. The *y* axis contains the eigenvalues sorted by decreasing order of total variance explained. The point of separation is often called the 'elbow', that location might indicate a good number of principal components (PCs) to retain.

Theorem 4.6: Consider the PCs

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

computed from the covariance matrix $\boldsymbol{\Sigma}$, with the associated eigenvalue - eigenvector tuples $(\lambda_1, \mathbf{e}_1), \ldots (\lambda_p, \mathbf{e}_p)$. The correlation coefficient between the i^{th} PC Y_i and the k^{th} variable X_k is given by

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{X_k X_k}}}, \text{ where } 1 \le i,k \le p \qquad [22]$$

Proof: Consider the vector $\mathbf{a}_k' = [0,\ldots,0,1,0,\ldots,0]$ such that $X_k = \mathbf{a}_k'\mathbf{X}$ and $\text{Cov}(X_k, Y_i) = \text{Cov}(\mathbf{a}_k'\mathbf{X}, \mathbf{e}_i'\mathbf{X})$. Also $\text{Cov}(\mathbf{a}_k'\mathbf{X}, \mathbf{e}_i'\mathbf{X}) = \mathbf{a}_k'\boldsymbol{\Sigma}\mathbf{e}_i$ based on the maximization of quadratic forms on the unit sphere concept and remembering that $\text{Cov}(X_k, Y_i) = \mathbf{a}_k'\boldsymbol{\lambda}\mathbf{e}_i = \lambda_i e_{ik}$, $\boldsymbol{\Sigma}\mathbf{e}_i = \lambda_i\mathbf{e}_i$, then $\text{Var}(Y_i) = \mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_k = \lambda_i$ yields $\text{Var}(Y_i) = \lambda_i$ and $\text{Var}(\mathbf{X}_k) = \sigma_{kk}$ which gives

$$\rho_{Y_i,X_k} = \frac{\text{Cov}(Y_i, X_k)}{\sqrt{\text{Var}(Y_i)}\sqrt{\text{Var}(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i,k = 1,2,\ldots,p \quad \textit{Q.E.D.}$$

4.3.1 Standardized Variables

Generally standardzation is necessary when the units of data for different variables is variable, or there is significant difference between the magnituteds of data values for different variables. Then PCs may also be obtained from the standardized variables $\left(Z_1, Z_2,\ldots, Z_p\right)$, where

$$Z_i = \frac{\left(X_i - \mu_i\right)}{\sqrt{\sigma_{ii}}}$$

In matrix notation

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

here $\mathbf{V}^{1/2}$ is the diagonal standard deviation matrix as given in 4.3. Since $\mathbf{Z}$ is the vector of standardized random variables $Z_1, Z_2,\ldots, Z_p$, then $E\left(\mathbf{Z}\right) = \mathbf{0}$ and

$$\text{Cov}\left(\mathbf{Z}\right) = \left(\mathbf{V}^{1/2}\right)^{-1}\boldsymbol{\Sigma}\left(\mathbf{V}^{1/2}\right)^{-1} = \boldsymbol{\rho}.$$

The PCs of $\mathbf{Z}$ are obtainable from the eigenvectors of the correlation matrix $\boldsymbol{\rho}$. Results pertaining to PCA mentioned so far are valid for the case of standardized data.

An important point is that PCs obtained from $\boldsymbol{\Sigma}$ matrix and from the $\boldsymbol{\rho}$ matrix will not in general be the same [4].

Theorem 4.7: The standardized variables $\mathbf{Z}' = \left[Z_1, Z_2, \ldots, Z_p \right]$ with $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$ yields PCs as given by

$$Y_i = e_i' \mathbf{Z} = e_i' \left(\mathbf{V}^{1/2} \right)^{-1} (\mathbf{X} - \boldsymbol{\mu}), \quad i = 1, 2, \ldots, p$$

Furhter,

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p \qquad (4.8)$$

and

$$\rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i} \qquad i, k = 1, 2, \ldots, p$$

Then $(\lambda_1, \mathbf{e_1}), \ldots (\lambda_p, \mathbf{e_p})$ are the eigenvalue - eigenvector pairs for $\boldsymbol{\rho}$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Proof: Theorem 4.7, is proven using 4.4, 4.5 and 4.6. Instead of $X_1, \ldots, X_p$, $Z_1, \ldots, Z_p$ comes and $\boldsymbol{\rho}$ instead of $\boldsymbol{\Sigma}$.

We see from 4.8, the sum of the diagonal elements of matrix $\boldsymbol{\rho}$ is equal to the total population variance *p*.

Proportion of standardized population variance due to k^{th} principal component = $\frac{\lambda_k}{p}$; k=1,2,…,p where the λ_k '*s* are the eigenvalues of $\boldsymbol{\rho}$.

Chapter 5

APPLICATIONS

5.1 Application for PCA

A data set consisting of 30 different plants and a total of 340 data observations is considered. Plant type Primula Vulgaris had 16 observations with no missing values, the largest number of observations out of the 30 plant types, therefore choosen for the study. See Table 5.1. Although 16 observations may not be a large data set, but it was not possible to mix the data from diffrent plant types, as it would result in incosistent reults in subsequent computations. The data set consists of 14 attributes based on the analysis of leaf shapes [19]. Attributes are; X1**:** Eccentricity, X2: Aspect Ratio, X3: Elongation, X4: Solidity , X5: Stochastic Convexity, X6: Isoperimetric Factor, X7: Maximal Indentation Depth, X8: Lobedness, X9: Average Intensity, X10: Average Contrast, X11: Smoothness, X12: Third moment, X13: Uniformity, X14: Entropy.

The mean vector of the data is computed as

$$\bar{\mathbf{X}} = \begin{pmatrix} 0.417701 & 1.087388 & 0.661612 & 0.529682 & 0.652521 & 0.158526 & 0.128979 & 3.04685 & 0.025704 & 0.08644 & 0.007714 & 0.002377 & 0.000144 & 0.808387 \end{pmatrix}$$

This gives a preliminary idea about the magnitude of the values for each variable. Highest values observed in X_8, and lowest ones in X_{13}.

Table 5.1: Leaf data with 14 variables

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
0.51247	1.1116	0.65626	0.57724	0.59298	0.16867	0.11187	2.2776	0.016001	0.061238	0.003736	0.000879	0.000103	0.6508
0.54893	1.1111	0.63983	0.56623	0.6	0.15743	0.13081	3.1143	0.021231	0.079722	0.006316	0.001912	7.39E-05	0.71949
0.43425	1.095	0.68828	0.52398	0.54211	0.15396	0.13761	3.4462	0.021147	0.073202	0.00533	0.00134	0.000164	0.75496
0.38501	1.0656	0.63042	0.51223	0.59123	0.13705	0.12292	2.7498	0.033373	0.098907	0.009688	0.002787	0.000214	1.0015
0.26758	1.1316	0.60128	0.54301	0.77368	0.20311	0.15354	4.2904	0.017945	0.07145	0.005079	0.001465	7.25E-05	0.63273
0.24465	1.047	0.60511	0.56524	0.79474	0.21788	0.12522	2.854	0.037595	0.127	0.015874	0.006587	0.000108	0.8331
0.39092	1.087	0.68174	0.50961	0.6614	0.15361	0.14082	3.6093	0.028638	0.089135	0.007882	0.002118	0.00021	0.90082
0.4042	1.0965	0.65899	0.52833	0.68421	0.1771	0.13017	3.0837	0.029057	0.096836	0.00929	0.002893	0.000127	0.82383
0.50692	1.127	0.67203	0.53024	0.75263	0.16792	0.13006	3.0788	0.015279	0.057592	0.003306	0.000728	0.00011	0.67289
0.47565	1.0656	0.69172	0.5233	0.49649	0.14133	0.12987	3.0697	0.023977	0.0842	0.00704	0.002085	0.000113	0.77399
0.52382	1.1117	0.67175	0.54701	0.62982	0.15157	0.13674	3.4028	0.026434	0.085792	0.007306	0.002137	0.000166	0.90513
0.36462	1.0811	0.67755	0.49042	0.68772	0.14118	0.1243	2.8118	0.037866	0.11692	0.013485	0.004648	0.000177	0.9229
0.52212	1.1191	0.70988	0.50678	0.64912	0.1412	0.13192	3.1674	0.025478	0.085964	0.007336	0.002179	0.000149	0.82809
0.38203	1.0405	0.6901	0.48549	0.63684	0.13165	0.11852	2.5565	0.027997	0.093312	0.008632	0.002659	0.000125	0.84994
0.27123	1.096	0.68075	0.49446	0.53684	0.13088	0.12815	2.9887	0.021943	0.072882	0.005284	0.001339	0.000221	0.80402
0.44882	1.0118	0.6301	0.57134	0.81053	0.16187	0.11115	2.2486	0.027309	0.088889	0.007839	0.002273	0.000175	0.86

The covariance (**S**) and correlation (**R**) matrices were computed from the data set, and both used for the computation of the principal components. Since the eigenvectors are used in the formation of the PCs, then the set of PCs obtained from the covariance matrix will be different to those obtained from the correlation matrix.

The computed covariance matrix using raw data from Table 5.1 is given below.

$$\mathbf{S} = \begin{pmatrix}
0.0094 & 0.0009 & 0.0013 & 0.0008 & -0.0025 & -0.0008 & -0.0002 & -0.0099 & -0.0003 & -0.0008 & -0.0001 & -0.0001 & -0.0000 & -0.0014 \\
0.0009 & 0.0011 & 0.0002 & 0.0000 & -0.0005 & 0.0001 & 0.0002 & 0.0102 & -0.0001 & -0.0004 & -0.0001 & -0.0000 & -0.0000 & -0.0018 \\
0.0013 & 0.0002 & 0.0010 & -0.0006 & -0.0017 & -0.0006 & -0.0000 & -0.0009 & -0.0000 & -0.0001 & -0.0000 & -0.0000 & 0.0000 & 0.0005 \\
0.0008 & 0.0000 & -0.0006 & 0.0009 & 0.0009 & 0.0005 & -0.0000 & -0.0020 & -0.0001 & -0.0001 & -0.0000 & -0.0000 & -0.0000 & -0.0013 \\
-0.0025 & -0.0005 & -0.0017 & 0.0009 & 0.0089 & 0.0015 & -0.0000 & 0.0011 & 0.0001 & 0.0005 & 0.0001 & 0.0001 & -0.0000 & -0.0005 \\
-0.0008 & 0.0001 & -0.0006 & 0.0005 & 0.0015 & 0.0006 & 0.0001 & 0.0033 & -0.0000 & 0.0001 & 0.0000 & 0.0000 & -0.0000 & -0.0011 \\
-0.0002 & 0.0002 & -0.0000 & -0.0000 & -0.0000 & 0.0001 & 0.0001 & 0.0053 & -0.0000 & -0.0000 & -0.0000 & -0.0000 & -0.0000 & -0.0002 \\
-0.0099 & 0.0102 & -0.0009 & -0.0020 & 0.0011 & 0.0033 & 0.0053 & 0.2550 & -0.0008 & -0.0018 & -0.0004 & -0.0001 & -0.0000 & -0.0127 \\
-0.0003 & -0.0001 & -0.0000 & -0.0001 & 0.0001 & -0.0000 & -0.0000 & -0.0008 & 0.000012 & 0.0001 & 0.0000 & 0.0000 & 0.0000 & 0.0006 \\
-0.0008 & -0.0004 & -0.0001 & -0.0001 & 0.0005 & 0.0001 & -0.0000 & -0.0018 & 0.0001 & 0.0003 & 0.0001 & 0.0000 & 0.0000 & 0.0013 \\
-0.0001 & -0.0001 & -0.0000 & -0.0000 & 0.0001 & 0.0000 & -0.0000 & -0.0004 & 0.0000 & 0.0001 & 0.000016 & 0.0000 & 0.0000 & 0.0002 \\
-0.0001 & -0.0000 & -0.0000 & -0.0000 & 0.0001 & 0.0000 & -0.0000 & -0.0001 & 0.0000 & 0.0000 & 0.0000 & 0.000021 & 0.0000 & 0.0001 \\
-0.0000 & -0.0000 & 0.0000 & -0.0000 & -0.0000 & -0.0000 & -0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.00001 & 0.0000 \\
-0.0014 & -0.0018 & 0.0005 & -0.0013 & -0.0005 & -0.0011 & -0.0002 & -0.0127 & 0.0006 & 0.0013 & 0.0002 & 0.0001 & 0.0000 & 0.0107
\end{pmatrix}$$

S is a symmetric matrix, its diagonal elements being the variances of each variable. Off diagonal elements are the covariance values for all possible pairs of variables,

indicating whether the relation is positive or negative between the pairs. However, the absolute value of the covariance is not an indication of the strength of the relation between the variables in each pair.

Another drawback of the use of covariance matrix becomes evident when there is significant difference between the observations belonging to different variables, or units of variables are not consistent. Under such circumstances, the use of the correlation matrix for the determination of the PCs becomes more desirable.

Eigenvalues ($\mathbf{\Lambda}_s$) vector obtained from the covariance matrix is

$$\Lambda_S = \{0.000002, 0.000001, 0.00008, 0.00004, 0.00002, 0.00001, 0.0001, 0.0002, 0.0004, 0.0010, 0.0058, 0.0109, 0.0130, 0.2566\}$$

The corresponding eigenvectors matrix $\mathbf{E}_S$ is

$$\mathbf{E}_S = \begin{pmatrix}
-0.0002 & -0.0008 & 0.0027 & -0.0007 & -0.0026 & 0.0772 & -0.1428 & -0.0065 & -0.0789 & -0.0531 & 0.6975 & 0.1046 & -0.6823 & 0.0396 \\
0.00002 & 0.0001 & -0.0026 & 0.0115 & 0.0007 & -0.1065 & 0.0532 & 0.0984 & 0.9611 & 0.1594 & 0.0040 & -0.0443 & -0.1528 & -0.0400 \\
0.0003 & 0.0016 & -0.0019 & 0.0047 & -0.0049 & -0.1227 & 0.5069 & 0.4886 & -0.2090 & 0.6331 & -0.0034 & 0.1692 & -0.1269 & 0.0038 \\
0.0007 & 0.0026 & -0.0090 & -0.0120 & -0.0015 & -0.4578 & 0.5735 & 0.1493 & -0.0097 & -0.6303 & 0.0931 & -0.1731 & -0.0540 & 0.0076 \\
0.0001 & 0.0001 & -0.0017 & -0.0015 & 0.0024 & -0.0576 & 0.0279 & -0.0536 & -0.0020 & 0.2655 & 0.5825 & -0.6014 & 0.4707 & -0.0050 \\
-0.0008 & -0.0011 & 0.0005 & 0.0237 & -0.0064 & 0.7398 & 0.0838 & 0.5704 & 0.0667 & -0.2819 & 0.0160 & -0.1773 & 0.0634 & -0.0134 \\
0.0111 & 0.0170 & -0.4973 & -0.5563 & 0.6646 & 0.0116 & -0.0092 & 0.0191 & 0.0035 & 0.0026 & 0.0013 & 0.0044 & -0.0006 & -0.0209 \\
-0.0002 & -0.0003 & 0.0107 & 0.0107 & -0.0142 & -0.0067 & -0.0008 & -0.0050 & -0.0369 & -0.0155 & 0.0462 & 0.0488 & 0.0019 & -0.9967 \\
0.0876 & -0.1582 & -0.2430 & -0.6036 & -0.6897 & -0.1046 & -0.1597 & 0.1685 & -0.0008 & -0.0222 & 0.0156 & 0.0271 & 0.0446 & 0.0033 \\
-0.0047 & -0.0109 & -0.0832 & 0.2727 & 0.1493 & -0.4311 & -0.5732 & 0.5957 & -0.0448 & -0.0890 & 0.0330 & 0.0467 & 0.1178 & 0.0073 \\
-0.2339 & 0.6178 & 0.5870 & -0.4306 & 0.0635 & -0.0698 & -0.1020 & 0.1165 & -0.0011 & -0.0153 & 0.0048 & 0.0060 & 0.0215 & 0.0014 \\
0.2478 & -0.6856 & 0.5843 & -0.2544 & 0.2360 & -0.0346 & -0.0436 & 0.0583 & -0.00001 & -0.0081 & 0.0013 & 0.0008 & 0.0091 & 0.0006 \\
-0.9360 & -0.3505 & -0.0202 & -0.0243 & -0.0108 & -0.0003 & 0.0023 & -0.0009 & 0.0003 & 0.0002 & 0.00004 & 0.0003 & 0.0002 & 0.00003 \\
-0.0009 & 0.0024 & 0.0077 & 0.0128 & 0.0139 & 0.0672 & 0.1295 & -0.0218 & -0.1360 & -0.1214 & 0.4022 & 0.7280 & 0.5001 & 0.0517
\end{pmatrix}$$

Clearly, 14 PCs can be written using the $\mathbf{E}_S$ matrix. However, according to equation (4.7) the first 3 eigenvalues represents 97.39% of total variation in the data, computed as $\tau_3 = (0.2566 + 0.0130 + 0.0109) / 0.288 = 0.9739$**.** It is also observed from Figure 5.1 that the elbow point is occurring around the third eigenvalue suggesting three PC will be adequate.

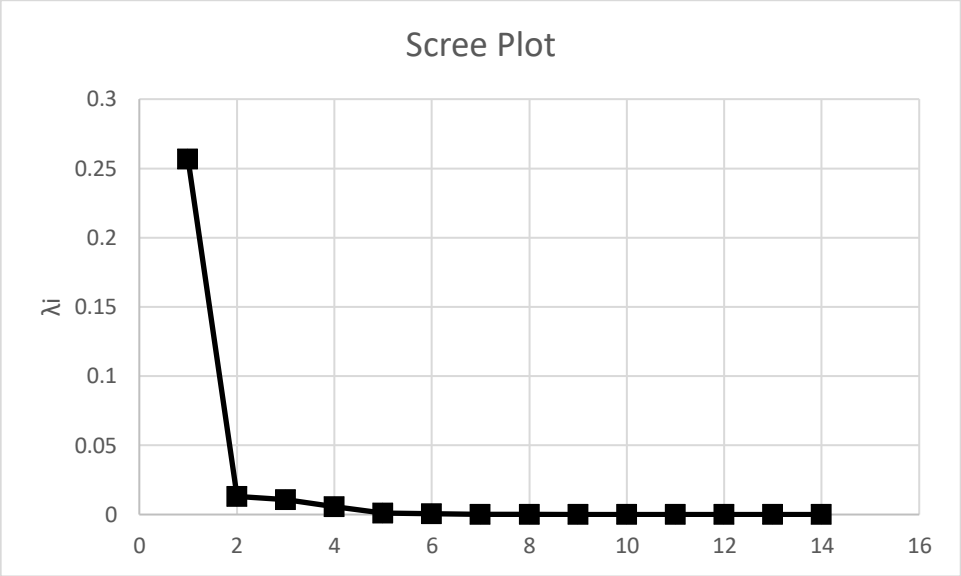

Figure 5.1: The scree plot of the eigenvalues computed from the covariance matrix

The first 3 PCs obtained from the data are given in Appendix C, Table C1.

Coefficients used in each variable making up a PC indicates the contribution of each variable to the formation of a PC. This is an important point as variables with high coefficients will carry more importance in the analysis of multivariate data. For example in the first PC (Y_1) which represents $(\lambda_1 / \Sigma\lambda)100 = (0.2566 / 0.288)100 = 89.1\%$ of total variation in the data, X_8 has the highest influence in absolute terms its coefficient provides $(e_{81} / \Sigma e_{8i})100 = (0.9967 / 1.1916)100 = 83.6\%$ of the contribution in the formation of PC1. Therefore X_8 can be considered as the most important variable in the whole data set.

Another important factor is the correlation coefficient between each PC and the random variables. Correlation coefficients r_{Y_i, X_j}, and eigenvectors e_i; $i = 1, 2, 3$ for this example are given in Table 5.2.

Table 5.2: Correlation coefficient between variables X_i and each PC, and contribution values e_i and each PC

r_{Y_i,X_j}	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
Y_1	0.2069	-0.6109	0.0609	0.1283	-0.0268	-0.2771	-0.9999	-0.9998	0.4826	0.2135	0.1773	0.0663	0.0481	0.2532
e_1	0.0396	-0.04	0.0038	0.0076	-0.005	-0.0134	-0.0209	-0.9967	0.0033	0.0073	0.0014	0.0006	0.00003	0.0517
Y_2	-0.8024	-0.5253	-0.475	-0.2052	0.5689	0.2951	0.0068	0.000004	0.9999	0.7755	0.6128	0.2264	0.0072	0.5512
e_2	-0.6823	-0.1528	-0.1269	-0.054	0.4707	0.0634	-0.0006	0.0019	0.0446	0.1178	0.0215	0.0091	0.0002	0.5001
Y_3	0.1126	-0.1395	0.5586	-0.6024	-0.6656	-0.7557	0.0459	0.0101	0.8168	0.2815	0.1566	0.0182	0.0099	0.7348
e_3	0.1046	-0.0443	0.1692	-0.1731	-0.6014	-0.1773	0.0044	0.0488	0.0271	0.0467	0.006	0.0008	0.0003	0.728

According to given in table 5.2 the variable X_7 has the highest correlation (-0.9999) with the first PC, Y_1. In other word there is almost perfect negative correlation between X_7 and Y_1. On the other hand contribution of X_7 in determining the value of Y_1 is only 1.75%. It means a high correlation between a variable and the PC does not necessarily mean it will have a high influence on the computation of the PC value. This is due to the coefficient of that variable representing its contribution to the value of the PC may be very low. In this data set the highest contribution comes from X_8, while the remaining variables contribution does not exceed 4.3% (Figure 5.2). As the contribution of X_8 is much higher than other variables, it was excluded from the graph in Figure 5.2 to avoid distortion in the graph.

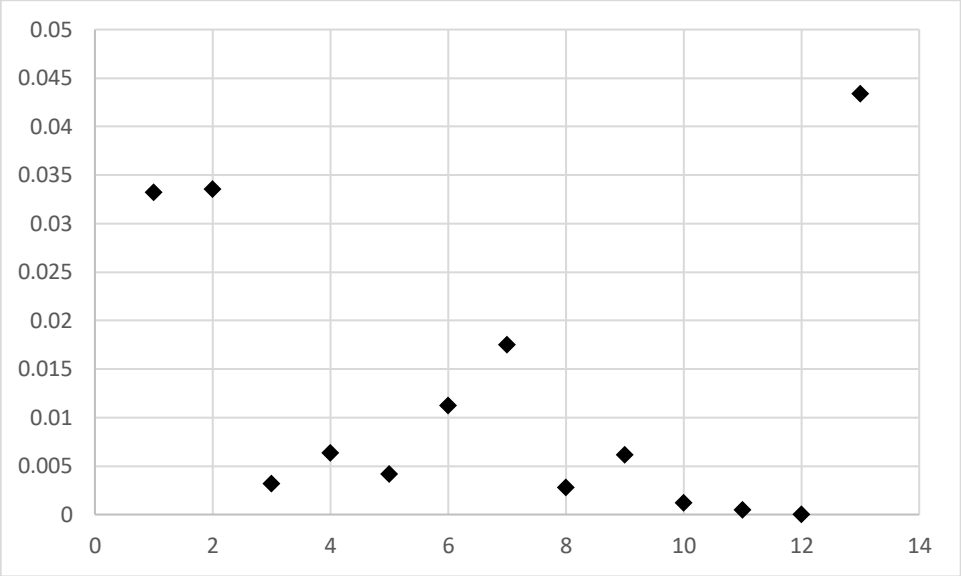

Figure 5.2: Contribution of each variable to the value of PC1

Similar interpretations can be made for Y_2 and Y_3 that are seen in Figure 5.3 and 5.4.

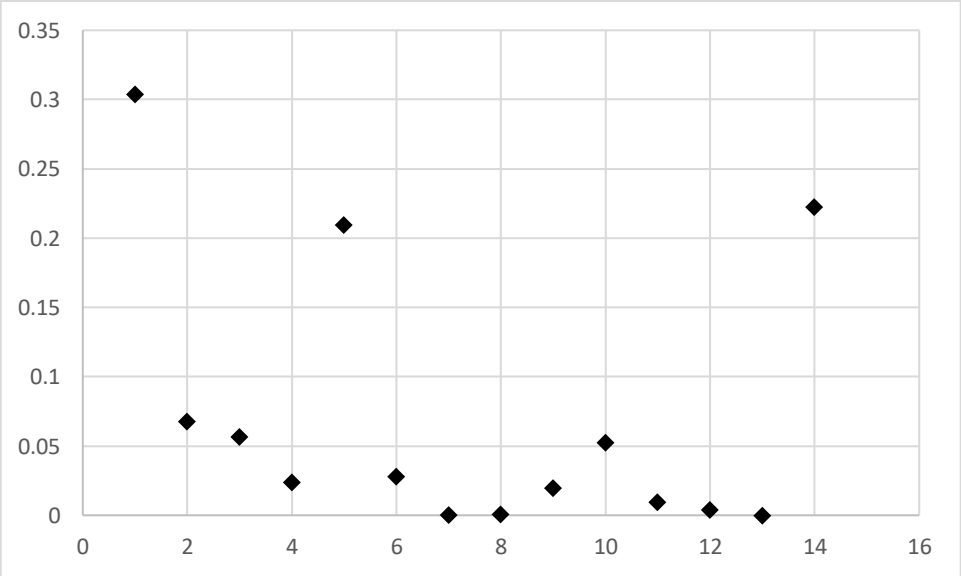

Figure 5.3: Contribution of each variable to the value of PC2

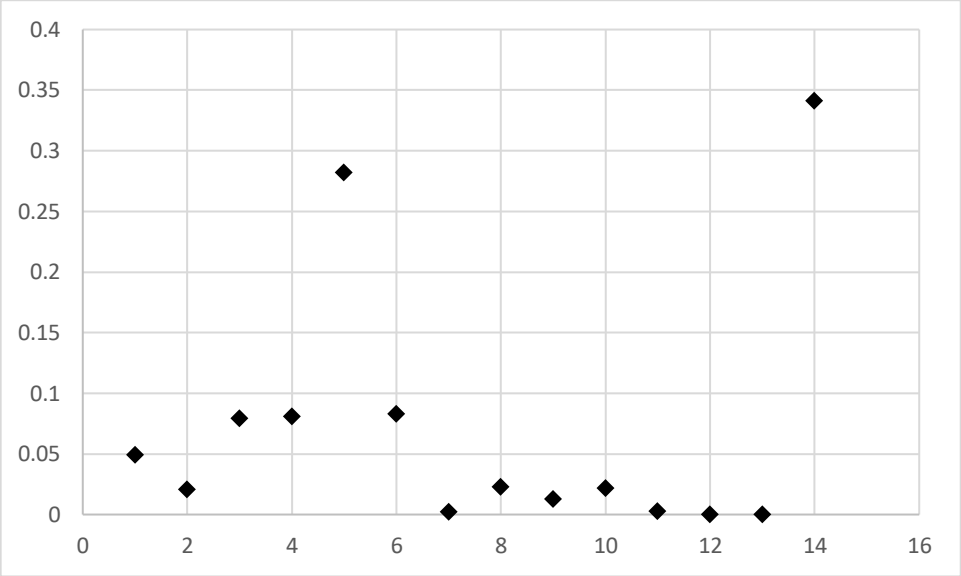

Figure 5.4: Contribution of each variable to the value of PC3

The relationship between the random variable's contributions to a PC (e_{ij}; $i, j = 1,2,\ldots,14$), and the correlation r_{Y_i,X_j} between random variables and the PC are given in Figure 5.5.

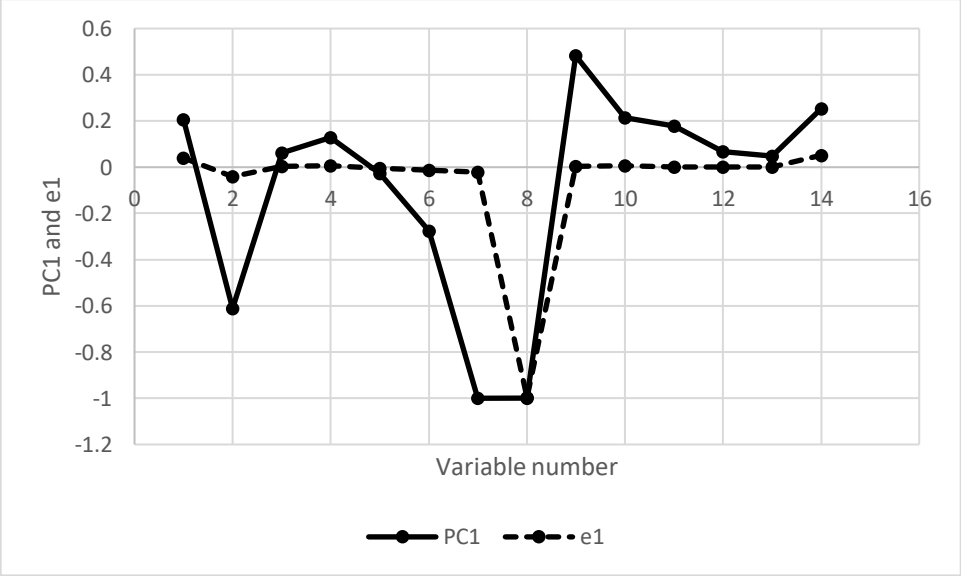

Figure 5.5: Relationship between the random variables X_i and the PC1

The dotted line shows the contribution values for the variables X_i. The solid line shows the PC values for each variable.

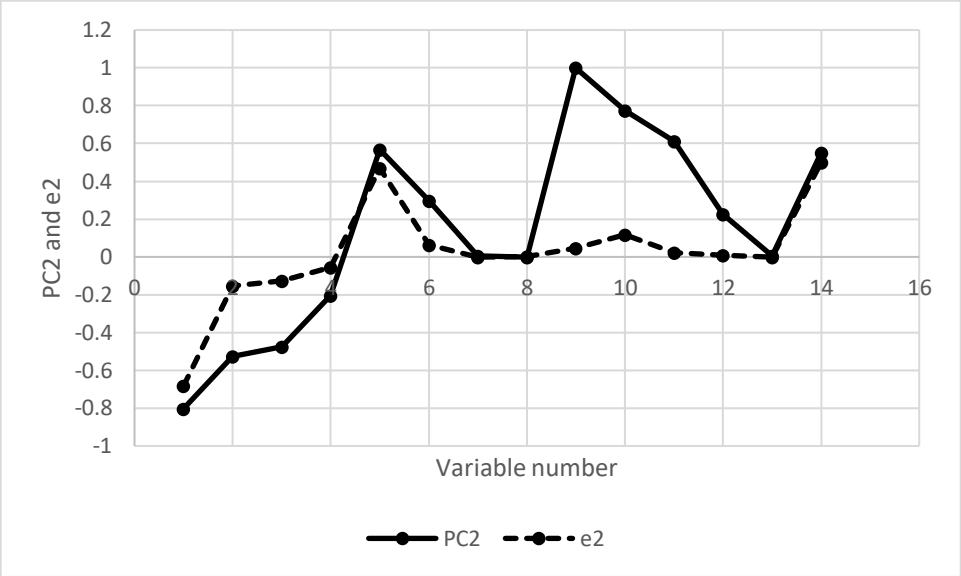

Figure 5.6: Relationship between the random variables X_i and PC2

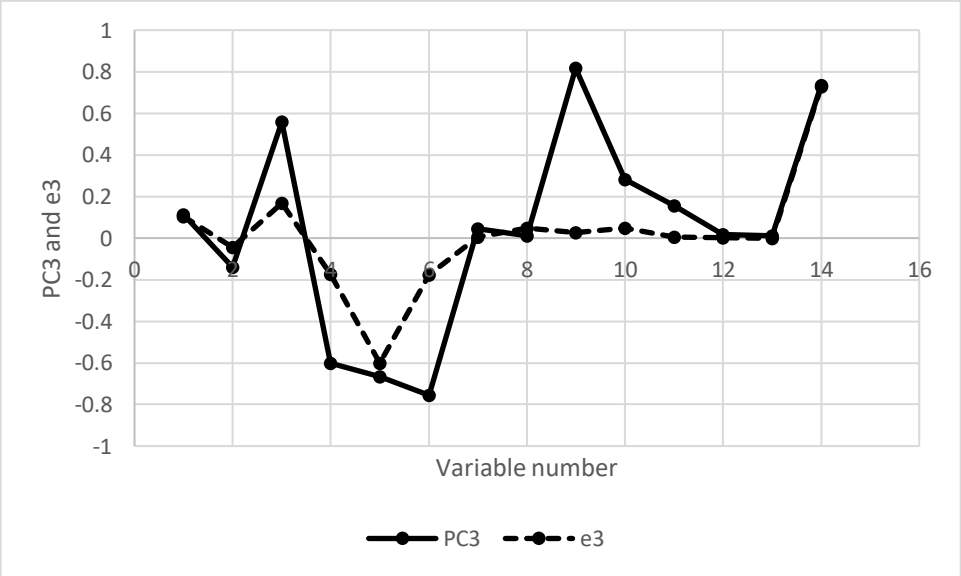

Figure 5.7: Relationship between the random variables X_i and the PC3

The same interpretations given for PC1 in Figure 5.5 are valid for PC2 and PC3 shown in Figures 5.6 and 5.7 respectively.

One other important point in estimation problems is the error committed. For the data used taking into account the level of linear correlation on one hand and the amount of contribution of each variable to each PC on the other, obtained bandwidth, MSE and the ratio of MSE to the average variance of estimated values (AVE), are summarized in Tables B1, B2, B3, B4, B5, B6 and B7 in Appendix B. These tables are the summary of output obtained from kernel regression, using different band widths and different *dx* increments to adjust the number of iterations of computing the kernel values. An easy tool for the assessment of error levels is the ratio of MSE to the average variance of estimated values, MSE/AVE. In general the following points are observed,

i. An increase in the linear correlation between the PC and the variable tends to reduce the error level.

ii. A decrease in the contribution of a variable in determining the value of a PC, results in larger error levels.

5.2 A Summary of PCs Using the Correlation Matrix

Use of the correlation matrix is more appropriate when the variables are inhomogeneous or have very high range of variation. The correlation coefficient **R** matrix obtained from data given in Table 5.1 is given below

$$\mathbf{R} = \begin{pmatrix}
1.0000 & 0.2663 & 0.4214 & 0.2657 & -0.2787 & -0.3458 & -0.1909 & -0.2022 & -0.3987 & -0.4333 & -0.4583 & -0.4804 & -0.1879 & -0.1430 \\
0.2663 & 1.0000 & 0.1527 & 0.0152 & -0.1685 & 0.1054 & 0.6041 & 0.5995 & -0.5920 & -0.5735 & -0.5477 & -0.5019 & -0.2890 & -0.5261 \\
0.4214 & 0.1527 & 1.0000 & -0.6253 & -0.5752 & -0.7394 & -0.0258 & -0.0561 & -0.1201 & -0.2142 & -0.2510 & -0.3182 & 0.3247 & 0.1442 \\
0.2657 & 0.0152 & -0.6253 & 1.0000 & 0.3377 & 0.6438 & -0.1576 & -0.1333 & -0.2755 & -0.1648 & -0.1259 & -0.0240 & -0.5100 & -0.4324 \\
-0.2787 & -0.1685 & -0.5752 & 0.3377 & 1.0000 & 0.6661 & -0.0009 & 0.0237 & 0.2098 & 0.2814 & 0.3236 & 0.3749 & -0.2320 & -0.0464 \\
-0.3458 & 0.1054 & -0.7394 & 0.6438 & 0.6661 & 1.0000 & 0.2481 & 0.2703 & -0.0249 & 0.1362 & 0.2039 & 0.3315 & -0.5714 & -0.4334 \\
-0.1909 & 0.6041 & -0.0258 & -0.1576 & -0.0009 & 0.2481 & 1.0000 & 0.9983 & -0.2201 & -0.1838 & -0.1999 & -0.1842 & -0.1310 & -0.2221 \\
-0.2022 & 0.5995 & -0.0561 & -0.1333 & 0.0237 & 0.2703 & 0.9983 & 1.0000 & -0.2380 & -0.1995 & -0.2151 & -0.1981 & -0.1459 & -0.2438 \\
-0.3987 & -0.5920 & -0.1201 & -0.2755 & 0.2098 & -0.0249 & -0.2201 & -0.2380 & 1.0000 & 0.9734 & 0.9551 & 0.8872 & 0.4041 & 0.8291 \\
-0.4333 & -0.5735 & -0.2142 & -0.1648 & 0.2814 & 0.1362 & -0.1838 & -0.1995 & 0.9734 & 1.0000 & 0.9924 & 0.9583 & 0.2054 & 0.6932 \\
-0.4583 & -0.5447 & -0.2510 & -0.1259 & 0.3236 & 0.2039 & -0.1999 & -0.2151 & 0.9551 & 0.9924 & 1.0000 & 0.9829 & 0.1621 & 0.6360 \\
-0.4804 & -0.5019 & -0.3182 & -0.0240 & 0.3749 & 0.3315 & -0.1842 & -0.1981 & 0.8872 & 0.9583 & 0.9829 & 1.0000 & 0.0343 & 0.5043 \\
-0.1879 & -0.2890 & 0.3247 & -0.5100 & -0.2320 & -0.5714 & -0.1310 & -0.1459 & 0.4041 & 0.2054 & 0.1621 & 0.0343 & 1.0000 & 0.7447 \\
-0.1430 & -0.5261 & 0.1442 & -0.4324 & -0.0464 & -0.4334 & -0.2221 & -0.2438 & 0.8291 & 0.6932 & 0.6360 & 0.5043 & 0.7447 & 1.0000
\end{pmatrix}$$

The eigenvalues and eigenvectors obtained from **R**

$$\mathbf{E}_R = \begin{pmatrix}
-0.2027 & -0.1676 & -0.3000 & -0.5668 & -0.4948 & -0.1068 & -0.1580 & -0.0844 & 0.0650 & -0.4470 & 0.1180 & -0.0690 & -0.0909 & 0.0145 \\
-0.3099 & 0.0616 & 0.2672 & -0.2809 & -0.1064 & -0.0635 & 0.8209 & -0.1785 & -0.0263 & 0.1466 & -0.0805 & -0.0091 & 0.0021 & -0.0014 \\
-0.0926 & -0.4318 & 0.0815 & -0.3313 & 0.1850 & -0.3637 & -0.0744 & 0.6315 & 0.0275 & 0.3130 & -0.1090 & 0.0662 & 0.0509 & -0.0087 \\
-0.0814 & 0.3533 & -0.3633 & -0.0704 & -0.3975 & 0.3806 & -0.0057 & 0.3462 & -0.3269 & 0.4007 & -0.1494 & 0.0941 & 0.0999 & -0.0181 \\
0.1262 & 0.3639 & -0.0440 & 0.1804 & -0.2768 & -0.8261 & -0.0481 & -0.0821 & -0.1926 & 0.1021 & -0.0253 & 0.0035 & 0.0287 & -0.0077 \\
0.0172 & 0.5114 & 0.0140 & 0.0024 & 0.0004 & -0.0128 & 0.1214 & 0.4475 & 0.6861 & -0.2098 & 0.0186 & -0.0640 & -0.0581 & 0.0166 \\
-0.1562 & 0.1387 & 0.5701 & -0.0907 & -0.1664 & 0.0844 & -0.2749 & 0.0513 & -0.0887 & -0.0578 & 0.1288 & -0.3039 & 0.6155 & -0.1055 \\
-0.1610 & 0.1528 & 0.5616 & -0.0601 & -0.1686 & 0.0781 & -0.2928 & 0.0332 & -0.1035 & 0.0163 & -0.0802 & 0.3459 & -0.6025 & 0.1017 \\
0.4135 & -0.0355 & 0.1010 & -0.1540 & -0.1028 & 0.0343 & 0.0416 & -0.0657 & 0.0801 & -0.1750 & -0.4471 & 0.4350 & 0.1959 & -0.5581 \\
0.4075 & 0.0558 & 0.0891 & -0.2463 & 0.0200 & 0.0541 & -0.0270 & -0.0288 & -0.0865 & 0.0617 & -0.4426 & -0.6834 & -0.2782 & 0.0761 \\
0.4048 & 0.0880 & 0.0741 & -0.2449 & 0.0662 & 0.0259 & 0.0869 & 0.0335 & -0.1304 & -0.1678 & 0.0105 & 0.3114 & 0.2707 & 0.7305 \\
0.3840 & 0.1551 & 0.0533 & -0.2705 & 0.1471 & 0.0326 & 0.1344 & 0.1472 & -0.2732 & 0.0002 & 0.6784 & -0.0221 & -0.1945 & -0.3417 \\
0.1598 & -0.3478 & 0.1438 & 0.4719 & -0.4195 & 0.0463 & 0.2965 & 0.4045 & -0.1906 & -0.3517 & 0.0257 & -0.1095 & -0.0787 & 0.0403 \\
0.3274 & -0.2512 & 0.1013 & 0.0203 & -0.4539 & 0.0597 & -0.0349 & -0.2002 & 0.4682 & 0.5252 & 0.2541 & 0.0136 & 0.0049 & 0.0907
\end{pmatrix}$$

and

$$\Lambda_S = \{5.4893, 3.5707, 2.2866, 0.8872, 0.6728, 0.5147, 0.3110, 0.2006, 0.0391, 0.0155, 0.0114, 0.0011, 0.0003, 0.00001\}$$

From the $\mathbf{E}_R$ matrix similar to $\mathbf{E}_S$, it is possible to write *p* different PCs. However, for dimension reduction it is desired to reduce the number of PCs such that they will still represent a high percentage of variation in the data (preferably more that 80% of the variation). In this data it is observed that the first 4 PCs represents 87.38% of total variation in the data.

Once the representative number of PCs are determined, all steps followed for the PCs obtained from the covariance matrix, can be repeated. Results to be obtained can similarly be interpreted.

The PCs computed from the correlation coefficient matrix. See in appendix C, table C2.

The correlation coefficients between the first PC and X variables are

$r_{Y_1Z_1} = -0.4749$, $r_{Y_1Z_2} = -0.7261$, $r_{Y_1Z_3} = -0.2169$, $r_{Y_1Z_4} = -0.1907$, $r_{Y_1Z_5} = 0.2957$,

$r_{Y_1Z_6} = 0.0403$, $r_{Y_1Z_7} = -0.3659$, $r_{Y_1Z_8} = -0.3772$, $r_{Y_1Z_9} = 0.9688$, $r_{Y_1Z_{10}} = 0.9547$,

$r_{Y_1Z_{11}} = 0.9484$, $r_{Y_1Z_{12}} = 0.8997$, $r_{Y_1Z_{13}} = 0.3744$, $r_{Y_1Z_{14}} = 0.7671$.

Based on all the computation done so far, the PCs computed from the covariance matrix are different from the PCs computed from the correlation matrix. The relation or the significance of variables in PCs computed using covariance or correlation matrix is different from one case to another. Furthermore, there is no linear relation between the PCs computed using one or another matrix. This leads to the conclusion that the standardizing of variables is usually required for inhomogeneous variables or for variables which have very high range of variation.

The PCs computed from this correlation matrix are given in Table 5.3.

Table 5.3: PCs computed using eigenvectors obtained from correlation matrix

$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	$\hat{Y}_5$	$\hat{Y}_6$	$\hat{Y}_7$	$\hat{Y}_8$	$\hat{Y}_9$	$\hat{Y}_{10}$	$\hat{Y}_{11}$	$\hat{Y}_{12}$	$\hat{Y}_{13}$	$\hat{Y}_{14}$
-2.2608	-0.0063	0.2177	1.131	0.1917	0.8974	0.5936	0.7238	-0.3314	0.0452	-0.0126	-0.0415	0.0033	0.001
-3.09	0.0125	0.3103	1.2264	0.1682	0.8746	0.5838	0.6995	-0.3396	0.0465	-0.013	-0.0419	0.0033	0.001
-3.4228	0.0795	0.3917	1.1378	0.2058	0.8512	0.5951	0.7216	-0.3294	0.0456	-0.0133	-0.0416	0.0033	0.001
-2.7166	0.2732	0.5006	1.1988	0.1705	0.8958	0.566	0.7063	-0.3247	0.0459	-0.013	-0.0415	0.0033	0.001
-4.281	0.2499	0.1588	1.1492	0.203	0.8728	0.5729	0.7096	-0.3338	0.0456	-0.0124	-0.0415	0.0033	0.001
-2.8356	0.3928	0.2205	1.1638	0.1729	0.8708	0.6174	0.7159	-0.3317	0.0465	-0.0131	-0.0414	0.0033	0.001
-3.5799	0.2435	0.4324	1.2426	0.2216	0.8613	0.5898	0.7272	-0.3308	0.0455	-0.0131	-0.0417	0.0033	0.001
-3.06	0.2086	0.327	1.2124	0.2124	0.884	0.6035	0.7135	-0.3258	0.0459	-0.0124	-0.0422	0.0033	0.001
-3.0605	0.0828	0.1864	1.2615	0.2619	0.8832	0.5804	0.7151	-0.3256	0.0465	-0.0136	-0.0415	0.0033	0.001
-3.0431	0.0433	0.4238	1.1304	0.1954	0.8342	0.5973	0.709	-0.3314	0.0459	-0.013	-0.0416	0.0033	0.001
-3.3691	0.1346	0.4492	1.3125	0.1838	0.8846	0.5911	0.7279	-0.3354	0.0464	-0.0126	-0.0413	0.0033	0.001
-2.7843	0.2889	0.3976	1.2107	0.2489	0.889	0.5981	0.7027	-0.3404	0.0452	-0.0133	-0.0416	0.0033	0.001
-3.1389	0.1014	0.3847	1.2767	0.2556	0.8817	0.6002	0.7148	-0.335	0.0461	-0.0123	-0.0416	0.0032	0.001
-2.531	0.217	0.3694	1.1503	0.2569	0.846	0.583	0.7059	-0.3298	0.0463	-0.0122	-0.0414	0.0033	0.001
-2.9705	0.2128	0.3991	1.0165	0.2346	0.8889	0.5762	0.7289	-0.3382	0.0466	-0.0129	-0.0418	0.0033	0.001
-2.2205	0.2664	0.2349	1.2962	0.1982	0.8394	0.5695	0.7254	-0.3365	0.0458	-0.0129	-0.0418	0.0033	0.001

The first 2 PCs that represents 65% of total variation, can be used to diagnose any extreme values that may exist in the data set. For this purpose, the scatter diagram of estimated PCs $\hat{Y}_1$ and $\hat{Y}_2$ is drawn. Then, the axes of ellipsoid are computed using the following formula $\frac{\hat{Y}_1}{\lambda_1}+\frac{\hat{Y}_2}{\lambda_2} \le \chi_2^2(0.05)$ where $\lambda_1 = 0.2566$, $\lambda_2 = 0.0130$ and $\chi_2^2(0.005) = 5.99$. The major and minor semi-axes of the ellipsoid are therefore $\mathrm{M} = \sqrt{\chi_2^2(0.05)\lambda_1} = 1.2398$ and $\mathrm{m} = \sqrt{\chi_2^2(0.05)\lambda_2} = 0.2791$. Then the control ellipsoid is drawn on the scatter diagram of the first two estimated PCs $\hat{Y}_1$ and $\hat{Y}_2$.

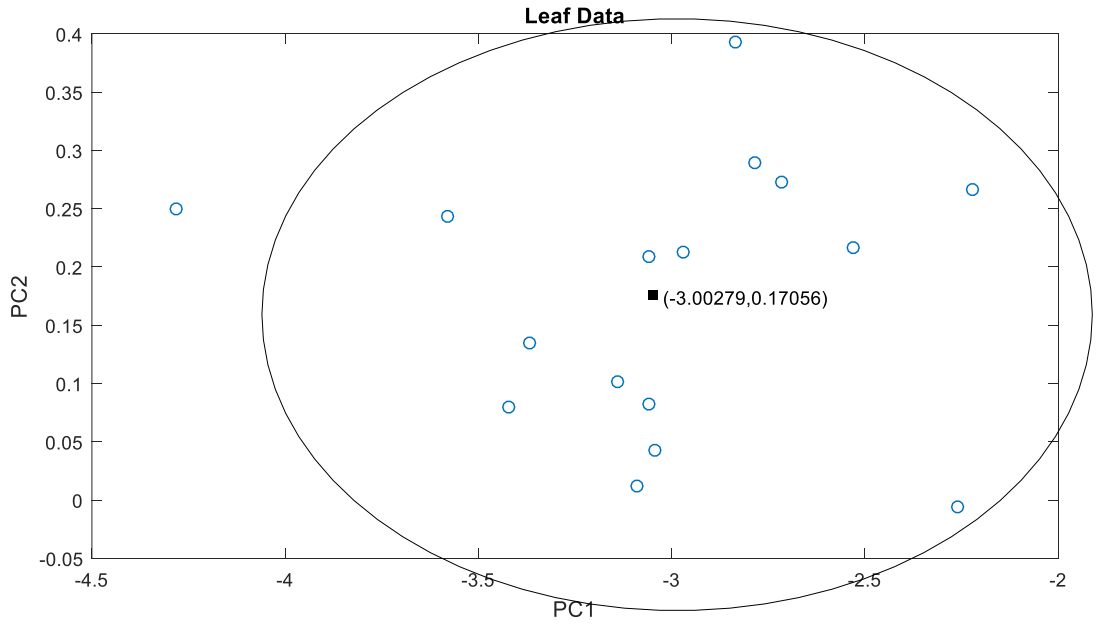

Figure 5.8: Control Ellipsoid

Any point from the scatter diagram that falls outside the ellipsoid is considered as an outlier or extreme value. From the Figure 5.8 one outlier is detected. It is on the left hand side has a PC1 value $\hat{y}_1 = -4.281$ PC2 value around $\hat{y}_2 = 0.2499$. An inspection of $\hat{y}_1$ and $\hat{y}_2$ values of this point identifies it as the 5^{th} value in Table 5.1. These PCs were computed using the 5^{th} raw data values from Table 5.1. A quick inspection of the data in raw 5 indicates that $x_{5,8} = 4.2904$ and $x_{5,13} = 0.0000725$ are standing out. Looking at the percentiles of these, we obtain the percentile for $x_{5,8} = 4.2904$ as 0.012, and the percentile for $x_{5,8} = 4.2904$ as 0.08. Clearly $x_{5,8}$ can be considered as an outlier due to its very low percentile value. $x_{5,13}$ may or may not be considered as an outlier, since its percentile value is above 5%.

Chapter 6

CONCLUSION

Dimension reduction in large data sets where the number of variables are expressed with tens or hundreds is an essential issue. PCA is the technique that does this very efficiently. Estimation of the value of a variable in the absence of population parameters is best done by kernel regression. In this thesis both methods are initially explained in detail. Subsequently an attempt is made towards the integration of the results of two methods to obtain better estimates via kernel regression by the use of a pilot data set.

In Chapter 3 where the kernel regression is explained, special emphasis is put on highlighting important points to be observed while applying this technique. Bandwidth is the most important parameter in kernel regression as it determines the amount of smoothing, as well as influencing the variance - bias balance. In the application example parallel to the increase of the bandwidth, the following became evident.

i. MSE increase.

ii. Variance of estimates decrease.

iii. Bias increase.

Decision whether to use the covariance or correlation matrix in the computation of PCs is a very important issue. This is highlighted in Chapter 4 and also in Chapter 5 while applying the PCA theory.

The proposed idea of estimating independent variables used in kernel regression by assuming them as dependent variables on the PCs has produced satisfactory results. Points taken into account in this application are

i. Linear correlation between a PC $\hat{Y}_j$ and each variable X_i.

ii. Contribution of each variable X_i to the computation of each $\hat{Y}_j$.

Level of correlation and amount of contribution are found not to be correlated with each other. But their influence on the MSE values while estimating X_i using $\hat{Y}_j$ has been recorded. It is generally observed that an increase in correlation and/or contribution values parallels a decrease in MSE levels.

REFERENCES

[1] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis,* Wiley.

[2] Arrow, K. J. & Lehmann, E. L. (2005). Harold Hotelling. *National Academy of Sciences.* 8-9.

[3] Banerjee, A. (2012). Impact of Principal Component Analysis in the Application of Image Processing. *International Journal of Advanced Research in Computer Science and Software Engineering,* 2, 1-8.

[4] Bezergianni, S. & Kalogianni, A. (2008). Application of Principal Component Analysis for Monitoring and Disturbance Detection of a Hydrotreating Process. *Ind. Eng. Chem. Rhs.* 47, 6972-6982.

[5] Cline, D. B. H. (1988). Admissible Kernel Estimators of a Multivariate Density. *Ann. Statist.* 16**,** 1421-1427.

[6] Dunteman, G.H. (1989). *Principal Components Analysis*. USA: SAGE Publication. 10-60.

[7] Farrell, R. H. (1972). On the Best Obtainable Asymptotic Rates of Convergence in Estimation of a Density Function at a Point. *Ann. Math. Statist.* 43, 170-80.

[8] Fix, E. & Hodges, J. L. (1951). *Discriminatory Analysis Nonparametric Discrimination: Consistency Properties.* University Of California, Texas. 50-200.

[9] Hardle, W. & Simar, L. (2007). *Applied Multivariate Statistical Analysis*. New Jersey.

[10] Hardle, W. (1991). *Smoothing Techniques: With Implementation in S*. NY: Springer-Verlag. 100-252.

[11] Mathematics Stack Exchange. (2013, October). Retrieved from https://math.stackexchange.com/questions/564751/how-can-i-simply-prove-that-the-pearson-correlation-coefficient-is-between-1-an

[12] Solutions on the Kernel Density Estimation. Retrieved from http://dutiosb.twi.tudelft.nl/~cai/AS2015/solutions-kernel-density.pdf

[13] Irizarry, R.A. (2001). Applied Nonparametric and Modern Statistics. *In, Resampling Methods: Bias, Variance, and Trade off*. 140. 43-62.

[14] Jeffers, J. N. R. (1967). Two Case Studies in the Application of Principal Component Analysis. *Journal of the Royal Statistical Society*. 16 (3), 225-236.

[15] Johnson, R. A. (1982). *Applied Multivariate Statistical Analysis,* Prentice Hall.

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

[17] Jolliffe, I. (2002). *Principal Component Analysis.* New York: Springer. 30-59

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

[19] Music, A. Palalic, S.G. (2016). *Classification of Leaf Type Using Multilayer Perceptron, Naïve Bayes and Support Vector Machine Classifiers.* 5 (2), 16-20.

[20] Ramsay, J.O. & Silverman, B.W. (2005). *Functional Data Analysis.* USA: Springer Science + Business Media. 70-80.

[21] Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *The Indian Journal of Statistics*. 26 (4), 329-358.

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

[23] Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*. 27 (3), 832-837.

[24] Ruppert, D. & Cline, D. B. H. (1994). Bias Reduction in Kernel Density Estimation by Smoothed Empirical Transformations. *Institute of Mathematical Statistics*. 22 (1), 185-210.

[25] Scott, D. W. (1979). On Optimal and Data-Based Histograms. *Biometrika.* 66**,** 605-610.

[26] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. USA: Springer Science + Business Media. 75-93.

[27] Wand, M.P. & Jones, M.C. (1995). *Kernel smoothing*, NY Chapman & Hall. 10-85.

[28] Zhang, W. & Wei, Y. (2012). Regression Based Principal Component Analysis for Sparse Functional Data with Applications to Screening Pubertal Growth Paths. *The Annals of Applied Statistics.* 9 (2), 597-620.

APPENDICES

Appendix A: Matlab Code for Computing PCs from Leaf Data

```
FD='c:\PcaAnalysis\LeafData\leaf11.txt'
H=importdata(FD)
S=cov(H)
R=corr(H)
%E1 vector of eigenvector and L1 is matrix of eigenvalues
[E1,L1]=eig(S)
[E2,L2]=eig(R)
TotEval=sum(diag(L1));
TotVar=sum(diag(S));
Y1=H*E1(:,14); %PC1
Y2=H*E1(:,13); %PC2
Y3=H*E1(:,12); %PC3
Y4=H*E1(:,11); %PC4
Y5=H*E1(:,10); %PC5
Y6=H*E1(:,9); %PC6
Y7=H*E1(:,8); %PC7
Y8=H*E1(:,7); %PC8
Y9=H*E1(:,6); %PC9
Y10=H*E1(:,5); %PC10
Y11=H*E1(:,4); %PC11
Y12=H*E1(:,3); %PC12
Y13=H*E1(:,2); %PC13
Y14=H*E1(:,1); %PC14
Y=[Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9 Y10 Y11 Y12 Y13 Y14]
plot(Y1,Y2,'o');
xlabel('PC1')
ylabel('PC2')
title('Leaf Data')
text(-3.02279,0.175056,'(-3.00279,0.17056)')
evec1=E1(:,14)
coY1X1=[0.2069  -0.6109  0.0609  0.1283  -0.0268  -0.2771  -0.9999  -0.9998
0.4826 0.2135 0.1773 0.0663 0.0481 0.2532]
coY1X1trans=coY1X1'
xaxes=[1 2 3 4 5 6 7 8 9 10 11 12 13 14]
plot(xaxes,evec1,xaxes,coY1X1trans)
evec2=E1(:,13)
coY2X2=[-0.8024  -0.5253  -0.4575  -0.2052  0.5689  0.2951  0.0068  0.000004
0.9999 0.7755 0.6128 0.2264 0.0072 0.5512]
coY2X2trans=coY2X2'
xaxes1=[1 2 3 4 5 6 7 8 9 10 11 12 13 14]
plot(xaxes1,evec2,xaxes1,coY2X2trans)
evec3=E1(:,12)
coY3X3=[0.1126  -0.1395  0.5586  -0.6024  -0.6656  -0.7557  0.0459  0.0101
0.8168 0.2815 0.1566 0.0182 0.0099 0.7348]
coY3X3trans=coY3X3'
xaxes2=[1 2 3 4 5 6 7 8 9 10 11 12 13 14]
plot(xaxes2,evec3,xaxes2,coY3X3trans)
```

Appendix B: Contribution of each variable to each PC

Table B1: Contribution of each variable to PC1

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y1,X2)= -0.602	-0.611	0.033576	0.05	0.1	0.000478	0.620486
Cov(Y1,X7)= -0.997	-0.999	0.017543	0.05	0.1	3.14E-07	0.001636
Cov(Y1,X8)= -0.999	-0.999	0.00277	0.05	0.1	0.00071	0.001578

Table B2: Contribution of each variable to PC2 when dx=0.05

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y2,X1)= -0.801	-0.802	0.303798	0.05	0.02	0.001874	0.215931
Cov(Y2,X9)= 0.7438	0.999	0.019858	0.05	0.02	1.21505E-05	0.322477
Cov(Y2,X10)= 0.734	0.775	0.052451	0.05	0.02	7.73E-05	0.239633
Cov(Y2,X14)=0.5511	0.551	0.22672	0.05	0.01	0.004359	0.439803

Table B3: Contribution of each variable to PC2 when dx=0.005

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y2,X1)= -0.801	-0.802	0.303798	0.005	0.02	0.001874	0.193821
Cov(Y2,X9)= 0.7438	0.999	0.019858	0.005	0.02	1.21505E-05	0.285928
Cov(Y2,X10)= 0.734	0.775	0.052451	0.005	0.02	7.73E-05	0.201439
Cov(Y2,X14)= 0.551	0.551	0.22672	0.005	0.008	0.004863	0.519498

Table B4: Contribution of each variable to PC2 when dx=0.002

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y2,X1)= -0.801	-0.802	0.303798	0.002	0.02	0.001937	0.197733
Cov(Y2,X9)= 0.7438	0.999	0.019858	0.002	0.02	1.21505E-05	0.282999
Cov(Y2,X10)= 0.734	0.775	0.052451	0.002	0.02	7.73E-05	0.198103
Cov(Y2,X14)= 0.551	0.551	0.22672	0.002	0.008	0.003746	0.448267

Table B5: Contribution of each variable to PC3 when dx=0.05

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y3,X5)=-0.66	-0.665	0.282083	0.05	0.02	0.003755	0.769589
Cov(Y3,X6)=-0.75	-0.755	0.083161	0.05	0.02	0.000165	0.360047
Cov(Y3,X9)= 0.415	0.816	0.012711	0.05	0.008	2.84E-05	1.081868
Cov(Y3,X14)= 0.73	0.734	0.34146	0.05	0.02	0.003023	0.311415

Table B6: Contribution of each variable to PC3 when dx=0.005

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y3,X5)= -0.66	-0.665	0.282083	0.005	0.015	0.003437	0.633569
Cov(Y3,X6)= -0.75	-0.755	0.083161	0.005	0.015	0.000139	0.282799
Cov(Y3,X9)= 0.415	0.816	0.012711	0.005	0.008	2.72E-05	0.890691
Cov(Y3,X14)= 0.73	0.734	0.34146	0.005	0.015	0.002783	0.21955

Table B7: Contribution of each variable to PC3 when dx=0.002

	r	e	dx	h	MSE	MSE/VarEstX
Cov(Y3,X5)= -0.66	-0.665	0.282083	0.002	0.02	0.003755	0.750868
Cov(Y3,X6)= -0.75	-0.755	0.083161	0.002	0.02	0.000165	0.350661
Cov(Y3,X9)= 0.415	0.816	0.012711	0.002	0.008	2.84E-05	0.868739
Cov(Y3,X14)= 0.73	0.734	0.34146	0.002	0.02	0.003023	0.246745

Appendix C: PC Tables from Covariance and Correlation Matrix

Table C1: First three PCs from covariance matrix

$$Y_1 = 0.0396X_1 - 0.0400X_2 + 0.0038X_3 + 0.0076X_4 - 0.0050X_5 - 0.0134X_6 - 0.0209X_7 - 0.9967X_8 + 0.0033X_9 + 0.0073X_{10} + 0.0014X_{11} + 0.0006X_{12} + 0.00003X_{13} + 0.0517X_{14}$$

$$Y_2 = 0.6823X_1 - 0.1528X_2 - 0.1269X_3 - 0.0540X_4 + 0.4707X_5 + 0.0634X_6 - 0.0006X_7 + 0.0019X_8 + 0.0446X_9 + 0.1178X_{10} + 0.0215X_{11} + 0.0091X_{12} + 0.0002X_{13} + 0.5001X_{14}$$

$$Y_3 = 0.1046X_1 - 0.0443X_2 + 0.1692X_3 - 0.1731X_4 - 0.6014X_5 - 0.1773X_6 + 0.0044X_7 + 0.0488X_8 + 0.0271X_9 + 0.0467X_{10} + 0.0060X_{11} + 0.0008X_{12} + 0.0003X_{13} + 0.5001X_{14}$$

Table C2: First four standardized PCs from correlation matrix

$$Y_1 = -0.2027X_1 - 0.3099X_2 - 0.0926X_3 - 0.0814X_4 + 0.1262X_5 + 0.0172X_6 - 0.1562X_7 - 0.1610X_8 + 0.4135X_9 + 0.4075X_{10} + 0.4048X_{11} + 0.3840X_{12} + 0.1598X_{13} + 0.3274X_{14}$$

$$Y_2 = -0.1676X_1 + 0.0616X_2 - 0.4318X_3 + 0.3533X_4 + 0.3639X_5 + 0.5114X_6 + 0.1387X_7 + 0.1528X_8 - 0.0355X_9 + 0.0558X_{10} + 0.0880X_{11} + 0.1551X_{12} - 0.3478X_{13} - 0.2512X_{14}$$

$$Y_3 = -0.3000X_1 + 0.2672X_2 + 0.0815X_3 - 0.3633X_4 - 0.0440X_5 + 0.0140X_6 + 0.5701X_7 + 0.5616X_8 + 0.1010X_9 + 0.0891X_{10} + 0.0741X_{11} + 0.0533X_{12} + 0.1438X_{13} + 0.1013X_{14}$$

$$Y_4 = -0.5668X_1 - 0.2809X_2 - 0.3313X_3 - 0.0704X_4 + 0.1804X_5 + 0.0024X_6 - 0.0907X_7 - 0.0601X_8 - 0.1540X_9 - 0.2463X_{10} - 0.2449X_{11} - 0.2705X_{12} + 0.4719X_{13} + 0.0203X_{14}$$

Appendix D: Temperature and Relative Humidity Data

	Temperature	Relative Humidity
1	9.900000095	65.19999981
2	10.57500005	63.25
3	10.7750001	62.54999924
4	10.94999981	62.07499981
5	10.97500014	66.5
6	11.00000024	61.44999981
7	11.07500005	62.22499943
8	11.125	60.20000076
9	11.17499995	68.47500038
10	11.32499981	70.20000076
11	11.375	70.52499962
12	11.4749999	58.17499924
13	11.47500014	60.02500057
14	11.5999999	61.92499924
15	11.67499995	57.09999943
16	11.67499995	63.65000057
17	11.82499981	58.02500057
18	11.89999986	67.44999886
19	11.9000001	57.39999962
20	11.96666686	54.90000025
21	11.9749999	61.12500095
22	12	58.85000038
23	12	62.92499924
24	12.32499981	66.15000057
25	12.3499999	65.35000038
26	12.4000001	63.87500095
27	12.5999999	63.47499943
28	12.60000014	54.20000076
29	12.65000033	57.90000057
30	12.75	63.22500038
31	12.77500033	55.82499981

Kernel data continued

	Temperature	Relative Humidity
32	13.05000019	57.90000057
33	13.125	54.75
34	13.35000014	61.34999943
35	13.5999999	54.59999943
36	13.62499976	62.05000019
37	13.625	56.29999924
38	13.6500001	49.47500134
39	13.75	53.35000134
40	13.875	53.875
41	14.04999995	57
42	14.0666666	65.69999949
43	14.0999999	49.80000019
44	14.12499976	50.02499866
45	14.125	55.57500076
46	14.27499986	55.44999981
47	14.27500033	50.57500076
48	14.72500014	55.92500019
49	14.75	52.07499981
50	14.75	59.22499943
51	14.77499986	54.32499981
52	14.79999995	53.92499924
53	14.82500005	54.67500114
54	14.82500005	60.55000114
55	15.25000024	61.44999886
56	15.47500014	51.875
57	15.52499986	52.92500019
58	15.69999981	60.15000057
59	16.0250001	50.92499924
60	16.25	50.95000076
61	16.6500001	57.375
62	16.70000029	48.92500019
63	18.2750001	52.62500095