Arab American University

Faculty of Graduate Studies

Using Principal Component Analysis and Linear Discriminant Analysis as Dimensional Reduction Techniques

Prepared By

Sajeda Rasim Frehat

Supervised By

Prof. Hasan Hamdan Hanashi

This thesis was submitted in partial fulfillment of the requirements for the Master‘s degree in Applied Mathematics

November, 2021

Thesis Approval

Using Principal Component Analysis and Linear Discriminant Analysis as a Dimensional Reduction Technique

By

Sajeda Rasim Frehat

This thesis was defended successfully on 25th/7/2021 and approved by:

Committee members	Signature
1. Supervisor ~~Name~~: Prof. Hasan Hamdan Hanashi	.....
2. Internal Examiner ~~Name~~: Dr. Mahmoud AlManassra	.....
3. External Examiner ~~Name~~: Dr. Inad Nawajah	.....

Declaration

My name is Sajeda Frehat, and I am the student owner of the AAUP university number 21520138. I acknowledge that I have complied with all the regulations, instructions, Arab American University standards of Academic codes of conduct. I bear all the consequences and results and responsibility if it shows otherwise. I also adhere to the Deans Council regulations in withdrawing their confirmation of this degree in case of any violations.

Name: Sajeda Rasim Frehat

Signature: ساجدة فريحات

Date: 15/12/2021

Dedication

With my great love, I dedicate this thesis

To my parents,

To my life partner (Mohammad),

To my brothers and sisters,

To my lovely son (Amr),

To my relatives and friends, and, To everyone who encouraged me to continue this work.

Acknowledgments

First of all, my great continuous thanks to Allah who gave me the ability to finish this work.

I want to express my deepest gratitude and sincerest appreciation to my supervisor Prof. Hasan Hanayshi, who guided and encouraged me in every level of preparing this work.

My sincere thanks to the examination committee for their careful reading of my thesis and their valuable feedback.

Special Thanks also goes to my university and particularly to the department of Mathematics at Arab American University for a great education and research experience.

Abstract

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) will be used as dimensional reduction techniques.

In particular, PCA will be compared with other dimensional reduction technique which is Linear Discriminant Analysis(LDA), these two methods and others are used to reduce the number of random variables and obtaining a set of principal variables that retains a large percentage of the total variation.

These two techniques will be applied on a dataset and explored and compared.

The comparison will be done between the two mentioned methods. We have relied on the number of components after reduction to give the best proportion of variance retained, so the total variance after reduction with the same number of components will determine the best method.

Recommendation will be made and the results will be presented.

Contents

List of Tables

List of Figures

Chapter 1

Introductive and Literature Review

1.1 General Introduction

Data revolution stormed all aspects of our lives as we can see everywhere around us. Real-world data exists in all fields of science, and are easily obtained and generated. Therefore, we get a tremendous size of high dimensional data sets that must be summarized to be accessible which is a challenging task. That entails questions that comes to mind:

First, resources needed such as storage, computational time, human resources, programmers,. . . ,etc.

second, complexity of the data, in addition, we will ask ourselves are all of the collected data useful and important? Can we take part of them to make things easier? If we want to analyze this data, does it need days, months, or more? How much this cost money? Can we perform any meaningful analysis? What is the amount of computational power this will take? Let us agree that in general a huge size of data is desired. It is a mighty thing to be accurate in your result due to size, it will take a lot of time, effort and money, also the process considerably becomes slow. It is axiomatic, this takes long time and hard effort so a radical solution is needed to facilitate the handling of the available data.

Therefore, the need of improved methods in data summaries, data mining and extractions is growing every day, i.e. the need for developing samples representative or techniques for the generated data.Suppose we have a data set with five or more dimensions.

So the question now is there any way to somehow we can reduce the extra dimensions to visible one in some sense?

The mathematical procedures and successful research investigations make the idea of dimension reduction possible. They help data scientists evaluate new algorithms that aim eliminating the curse of multidimensionality. They find of domains and find the most relevant features that plays a dominant role in determining and influencing the output results. This leads to a reduced training time and cost. Further, it shows effective visuals and leads to have a better idea.

This issue, dimensionality reduction methods, has become one of the most popular methods in data science. It is widely used in different fields in our daily life for example in information theory, in machine learning, in computational biology, in computer science and so on. It also plays a major part in image, video and audio analysis particularly in the field of medical research by diagnosing and curing diseases, it also reaches to politics and many other applications in several sectors. Therefore, and as mentioned above, this area of research will be our focus in this thesis.

Data Analysis and dimension reduction:

Nowadays, the expression 'huge data analytics' is buzzword and everyone who is specialized in data sciences cares about learning how to deal with such data and applying the suitable analyses.

For any collected dataset, we tend to add more and more features to be more accurate and extract more helpful indicators, without a clear answer when to stop. It is known in statistics that after a threshold point, the performance of the system will be affected. It is a well-known phenomenon, which is called 'The Curse of Dimensionality'. This curse happened because the relation between the classifier's performance and the number of dimensions is exponentially distributed i.e. when the number of the dimension increases, the classifier's performance increases until reach to the optimal number of features, after that whatever we add more features without enough samples i.e.(further increasing the dimensionality without increasing the number of training samples) this leads to a decrease in classifier performance and the dimensionality for features will grow and be sparser and sparser [1]. By this sparsity the production of an analysis is closely coincides or exactly

to a particular set of data, and therefore it may fail to fit additional data and makes the generalization prediction for future unreliable [2]. More details about this phenomenon will be discussed in the next chapter. The question now: how can we beat this curse and get rid of overfitting when we have several features with a limited number of samples? One way out is 'dimensionality reduction techniques' [1]. More details about this phenomenon will be discussed in the next chapter.

Dimensionality reduction in simple word, the set of algorithms used to reduce the number of variables and features for any high dimensional dataset problem. It compresses a large set of variables onto a smaller group of important variables without much loss of helpful and necessary information (i.e. We will lose information but not much).

As we know, when we get a higher number of features, it becomes harder to visualize the dataset. It is probably that the majority of these features are mostly correlated and some of them are superfluous. In these conditions, dimensionality reduction is a good choice to be applied to extract the useful information with fewer dimensions[3].

Therefore, it is critical for every statistical scientist, researcher, data users and machine learning experts to be professional in understanding the concept of this topic and be clever when and why to use dimensionality reduction[3].

To apply dimensionality reduction on a given dataset we can use any of the following ways:

(i) Feature selection: briefly, in this method we keep only the most pertinent variables from our dataset i.e. subset of the data, using one of three common strategies: filter strategy, wrapper strategy and embedded strategy [3].

(ii) Feature projection: here we transfer the data from the high- dimensional space to a fewer one. This transformation not require the model to be linear as in PCA, there exist several non-linear techniques such as: kernel PCA, locally linear embedding (LLE) and many other methods [3].

In this study, we will focus only on the linear techniques for dimensionality reduction, in particular the following main linear methods:

1. Principal Component Analysis (PCA).

2. Linear Discriminant Analysis (LDA).

We desire to find a combination of the input variables to be our new variables that contain most of the information. This is the intrinsic meaning dimensionality reduction: find the best linear combination of the original variables that retain largest possible percentage of the original variations.

1.2 Motivation

The suitable dimensionality reduction techniques to be used surely depends on the characteristics of the data to be reduced, its size and which distribution does it belong to.

Of course, there is no certain method that does the best in all given scenarios. For example, a simple model may lose some important information, so does a more complex model. Even if the later fits well with the data, but it may not give a good predictive accuracy. Various data were selected to examine the suitability of the chosen method for each example and compare its performance with the other method if it used instead of the other one. There are many reductions methods can be applied such as principal component analysis (PCA), linear discriminant analysis (LDA), generalized discriminant analysis (GDA), low variance filter, non-negative matrix factorization (NMF), graph-based kernel PCA and many many others. The use of each depends on the nature according to the nature of the data and its properties [4].

Our focus will be mainly on comparing the performance of PCA with that of the LDA. This is because PCA and LDA are the most commonly used techniques in Machine Learning and Data Science. The popularity in their use stems from their solid theoretical basis and performance.

1.3 Research Questions

In this study, we will focus on two areas of dimensionality reduction techniques using a real-life data sets. Then the performance of the two methods will be compared. The comparison will be made to decide which of them is more efficient, more practical and under what conditions.

Suggested questions that will be answered by this thesis:
In particular, we address two main questions:

1. Which of the two techniques is more robust to the model assumption?
2. Which of the two techniques is more sensitive to outliers and extreme values?
   If time allows we may explore two more questions further, otherwise it will be part of future work:
3. Can the mixture method be used as a dimensional reduction technique?
4. Is there a one dimensionality reduction method which is suitable for all data?

1.4 The Objectives of the Study

The goal of this thesis is to apply and compare the PCA and LDA as dimensional reduction techniques to data sets with large number of variables and observations. We will show four real-life examples that come from different sources and with different research questions. We will apply the PCA and LDA on each example and compare the accuracy of the two methods using graphs and visualization method. Graphical techniques make calculation easier and faster without big change of the information. Certain practical guideline will be developed on how and when to use PCA or LDA according to the given data. More related information will be provided.

1.5 Methodology

As we mentioned before, the scope of the research is to apply PCA and LDA on the data set, which are chosen diligently to serve the goals of our case study. Then, the comparison between the mentioned techniques will be done. Also evaluating the performance will be done using cross-validation techniques such as the bootstrap and the Jackknife for (PCA). Outliers are investigated using 'Mahlanobis distance' which is measure how many standard deviations that any point such "p" is away from the mean of the distribution "D".

All statistical analysis will be done using the open source R Software.

1.6 Organization of the Thesis

This thesis consists of 5 chapters.

In chapter 1, a brief background is presented to serve the purpose of the thesis and it's methodology. The use of dimensionality reduction techniques- PCA and LDA- in this thesis were described briefly along with a modest review of the literature.

In chapter 2, we will talk about the key features, properties and assumptions of PCA. Also, the performance of PCA will be discussed step by step.

In chapter 3, LDA is introduced with its properties and assumptions. Also, a real life example will be presented.

In chapter 4, the theoretical framework done using examples of PCA and LDA and a comparison of results are made.

In chapter 5, which is the final one, the performance of the two techniques is compared, the results are discussed and analyzed and the research recommendations will be made.

1.7 Literature Review

Dimensionality reduction is a common technique that is used in statistics and machine learning to simplify the complexity of the data. If the number of features is very large relative to the number of observations in the data set, we need a certain algorithm that attempt to reduce the number of random variables while retaining the important information to get an effective model.

Therefore, we can catch useful indicators and be accurate in our results by obtaining a set of principal features without losing too much information.

This method (dimensional reduction) is an old one in statistics. In fact, principal component analysis or what is known as (PCA) is the oldest and most popular dimensionality reduction technique [5].

The PCA used mostly in all of science disciplines, it's origin may be related to more than one scientist, it can be traced to Karl Pearson, Cauchy, Jordan, Caylay, Silverster or even Hamilton. But its modern instantiation was formulated by Hotelling in 1933 [44].

Linear Discriminant Analysis (LDA) is also another dimensionality reduction technique, which implies to reduce the number of dimensions (i.e. variables) in a given dataset and retain as much information as possible by identifying the linear combination of feature variables that can presents a best grouped data into separate classes [6].

The LDA, which is classified as a feature selection model, is an extremely popular dimensionality reduction technique, which was developed as early as 1936 by Ronald A. Fisher. It is suitable for recognition problems such as classifications. An important additional information regarding the LDA is that it assumes normal distributed data, independent features, and identical covariance for every class for the optimality criterion. Therefore, other naming of this method is 'normal discriminant analysis (NDA)' because of normality assumption. Whereas if you are using PCA for modelling purposes, then normality would be ideal. But if it is for data reduction purposes, then normality is not a strict requirement [7].

Brief History of PCA

Sometimes it's difficult to trace the beginning and the origins of the PCA. However, several studies showed that the earliest dates of PCA goes back to Karl Pearson. It was just an earliest description of the PCA technique (1901) that can be found in [8].

Pearson (1901) tried to fit an N-dimension space in a line through projections. This attempt led to the discovery of the PCA as a geometric optimization technique. He applied his method to 4-dimensional space but the difficulty of calculations made this technique not feasible [8].

The development of the technique has been rather uneven in the ensuing years, but there was a great deal of activity in the 1930s.
In (1931), Thurstone started working on the ideas of PCA, but his paper was more about factor analysis rather than PCA.

Then in (1933), Hotelling goes along with the same lines as Thurston. He wrote his own paper which consists of two parts, and from the collection of Pearson's paper together with the important part of Hotelling, Bryant and Atchley made an edited paper (1975).

"Hotelling's motivation is that there may be a smaller 'fundamental set of independent variables which determine the values' of the original variables. He notes that such variables have been called 'factors' in the psychological literature, but introduces the alternative term 'components' to avoid confusion with other uses of the word 'factor' in mathematics. Hotelling chooses his 'components' so as to maximize their successive contributions to the total of the variances of the original variables, and calls the components that are derived in this way the 'principal components'. The analysis that finds such components is then christened the 'method of principal components.'

Hotelling's derivation of PCs is similar to that given above, using Lagrange multipliers and ending up with an eigenvalue-eigenvector problem, but it differs in three respects. First, he works with a correlation, rather than covariance matrix; second, he looks at the original variables expressed as linear functions of the components rather than components expressed in terms of the original variables; and third, he does not use matrix notation." [9].

Girshick (1936) introduced an alternate derivations of the PCs. Then in (1939) he

investigated the asymptotic sampling distribution of coefficients and variances of PCs [10].

The following four scientists published four papers, which are now substantial references for the topic PCA:

- Anderson (1963) discussed the asymptotic sampling distributions of the coefficients and variances of the sample PCs (this paper described as the most theoretical of the four)[11].
- Rao's (1964) gave new ideas about using and interpretation and extensions of PCA[12].
- Gower (1966) discussed the relation between the other statistical techniques and PCA and gave geometric prudence[13].
- Jeffers (1967) discussed two case studies and he used the simple dimension reduction tool, so that he addressed the practical side of the subject[14].

These papers were interpreted in a book by Preiesendorfer and Mobley in (1988)[15].

In (1989), George H. Dunteman published a book "principal component analysis", and this book was a must for anyone in need of a concise, introductory guide to principal components analysis. This book describes PCA through an effective use of simple mathematical-geometrical and multiple real-life examples (such as crime statistics, indicators of drug abuse, and educational expenditure). By minimizing the use of matrix algebra, the reader can quickly master and put this technique to immediate use. His book view main examples to illustrate the myriad applications of PCA.

In (1994), Krzanowski, W.J. and Marriott, F.H.C authored a statistical book with title: "Multivariate Analysis: part 1: Distributions, Ordination and Inference" (Kendll's library of statistics) which is published by Hodder Educations publishers in that year.

In the period (1999-2000), the website of science identified that over than 2000 articles published that include the phrases "Principal Component Analysis" Which made a quick leap in this topic.

In (2000), Krzanowski, published an article in Oxford university press with title "Practical Guidelines for Learning Bayesian Networks from small Datasets which is about Principals of Multivariate analysis (A user's perspective)".

In 2003, J. EDWARD wrote a very helpful book “A User’s Guide to Principal Component”. The purpose of this book to show why this technique is useful in statistics and how it will be carried out also it explains why PCA works according to the given examples [17].

In 2009, Mark Richardson wrote an essay about PCA. It aimed to explain the theoretical side of PCA, and to provide examples and applications. Also the writer viewed a method for implementing PCA for real-world data, and he concluded that there is a close connection between PCA and the singular value decomposition (SVD) from numerical linear algebra [18].

In Jan, 2012 Hui Zou, Trevor Ha stie & Robert Tibshirani introduced a new method which is known as “Sparse principal component analysis (SPCA)”, it is a modified principal component with sparse loading, SPCA is applied to the real life and simulated data and give an encouraging result. A new formula is given to calculate the total variance of the modified PCA.

In 2015 a web tool “ClustVis” for visualizing clustering of multivariate data using Principal Component Analysis and heat map was constructed by Tauno Metsalu, Jaak Vilo to help scientists with little programming skills to make plots of their own data so that they can visualizing two of the components on the scatterplot and do more advanced requirements to get the actual PCA. his web server can be found freely at http://biit.cs.ut.ee/clustvis/ [19].

In 2019, Abbas F.M. Alkarkhi, Wasin A.A. Alqaraghuli publish their book “Easy Statistics for Food Science with R” this book present several applications about food science to relief researchers and students who are specialized in food science and food engineering to choose the suitable multivariate statistical technique using R. Many examples are analyzed by PCA with R commands such as “addresses the physicochemical properties of bananas” and “considers the antioxidants’ activity and antioxidative compounds” [21].

In brief, there is so much literature in all fields of sciences and psychology related to the development of the PCA because ot its importance and its wide range of applications and potential use.

Brief History of LDA

Linear discriminant analysis (LDA) or what is known as Fisher Discriminant Analysis (FDA) is one of the most prevalent statistical method for dimensional reduction. It is particularly popular because it is both a classifier and a dimensionality reduction technique.Moreover; it can be considered as a visualization technique. It can be used in statistics, pattern recognition, and machine learning to get a linear combination of variables that separating two or more classes of objects or events [22].

LDA is also closely related to (PCA) and to factor analysis. All aim at finding the linear combinations of variables that produce best explanation of the data. LDA tries to model differences between classes while factor analysis contrast combinations of features based on differences [23]. The general LDA approach is extremely similar to PCA, but it has an additional property "to find the component axes that maximize the variance of our data and focusing on the axes that maximize the separation between multiple classes" [24].

LDA has a powerful ability to solve the univariate and multivariate problems and convert it to an easy solved problem by projecting the dataset onto a lower-dimensional space with good class-separation to avoid overfitting and get a smaller computation costs. Also , it is useful to check whether the variables are effective for predicting category membership [25].

Discriminant analysis was first introduced by Sir Ronald Fisher in 1936, his goal was to find a linear projection of the data to maximize the variance between classes compared to the variance for data in the same class. This approach is called as Fisher's linear discriminant analysis. In that time, Fisher express a criterion for a two-class problems very gracefully and supply us with its derivation. Later, C. R. Rao made a generalization to be a "multi-class Linear Discriminant Analysis" or "Multiple Discriminant Analysis" in 1948 [26].

In (2001), R. Duda et al. develop a new technique in dimensionality reduction and classification in which LDA provides class separability by drawing a decision region between the different classes. So that LDA can maximize the ratio of the between-class variance and the within-class variance, and by assuming uniform prior class probability as in the

case of Naïve Bayes. We thus obtain the within class scatter matrix [27].

In 2002, Davis found a function which can find a relationship between the groups that have the maximum difference. It is closely related to multiple regression and trend-surface analysis (Press and Wilson, 1978). The idea behind this statistical method is to reduce the multi-dimensional space that has many parameters into just one dimension, so this will maximize the separation that could be noticed in this dimension and the problem solving would be easier [28].

In 2012, a new manifold learning method called improved semi-supervised local fisher discriminant analysis (iSELF) for gene expression data classification is proposed by HongHuang, JianweiLi and JiaminLiu. The purpose of this method is to preserves the global structure of unlabeled samples in addition to separating labeled samples in different classes from each other [29].

In 2019, Jason Brownlee discovered an algorithm for the Linear Discriminant Analysis (LDA) to make sensitive classification predictive modeling problems which contains several classes (more than two classes) instead of “Logistic regression” which is a traditional classification algorithm can only be applied for two-class classification problems [30].

Chapter 2

Overview of Dimensionality Reduction Methods (PCA and LDA)

In this chapter, the theoretical details of the two methods namely principal component analysis (PCA) and Linear Discriminant Analysis (LDA) are presented and discussed.

2.1 Introduction

The advancements made in collecting data and storage capacity made it easy to produce datasets with large size. This is a new challenge in data in the all fields of science. The traditional statistical methods break down with this large number of variables associated of each observation, these kind of data set are called “High Dimensional Data”. One way of managing the analysis of this is to ignore the redundant and dependent features by only extracting the useful information. The process of mapping the high dimensional data to somehow a lower reduced (one or two even three) humanly dimensional space(This is done by dismissing the uninformative variance from the origin data to get a subspace, in which we can easily deal with data with preserving the relationships among them) is called “Dimensionality Reduction”[31].

In order to understand the concept of “Dimensional Reduction” from our daily life, imagine that you want to upload and send a very huge number of data files to your work partner using e-mail, without any doubt this process needs too long time to be done!

Technological scientists discover a new method to speed up this process by compressing "zipping" these files and send the compressed files instead of sending the original ones via e-mail. This process let us to deal with a smaller sets of files which are equivalent to the large quantity of data. This makes things easier. We can say that dimensional reduction is the same idea of zipping the data[32].

Dimensional reduction considered to be 'the lens' for dealing with classification problems. It shrinks an enormous and redundant numbers of features into a new subset features while maintaining the important information.

Before we start focusing on the two main techniques, let us quickly talk about dimensionality reduction's applications and its benefits. Also we will look at the problem "cures of dimensionality". Obviously this method is popular in many real applications such as: machine learning, data mining[33], face recognition, image segmentation, bioinformatics[34], biometrics[35], cross-media retrieval, pattern recognition, text categorization and data classification.It also famous on databases including: artificial intelligence, visualization and optimization.

Now, suppose you have a large dataset of texts, sounds or even images and we aim to present these data by its variables (features or parameters). Surely dealing with this large data that have a lot of variables will increase the space. Furthermore it causes a slowdown of the needed time for training the model which introduces more noise than expected.

Dimensional reduction considered to be key criteria in many cases; since it helps improving the efficiency of the computational intensive duties. The idea is to transfer every object to another one and hence the model can easily be described throw less information, after ensuring that most of the sensitive information is maintained. Of Course, less dimensions lead to less computations and less training time. And as the number of dimensions comes down, the space required to store the data will be reduced. Also, it is very useful for the indexing purposes because it raises the speed of filtering steps. Finally, it helps in visualizing data, since it is very difficult to visualize data in higher dimensions. So after reducing the given high space to 2D or 3D may allow us to plot and observe patterns more clearly.

Sometimes there are algorithms that don't work efficiently when the dimensions are

huge. Therefore, applying dimension reduction is required for that algorithm to be useful. Dimension reduction cares about “multicollinearity”; it removes redundant features. “For example, suppose that you have two variables: ‘time spent on treadmill in minutes’ and ‘calories burnt’. These variables are highly correlated (as the more time you spend running on a treadmill, the more calories you will burn). Hence, there is no point in storing both as just one of them does what you require”[36].

2.2 Curse of Dimensionality:

The phenomena “the cures of dimensionality” first introduced by Bellman[37].

It is an obstacle for solving problems in the very high dimensionality space. It makes machine learning problems more complicated. That happened when we add more and more features to be more accurate in the results. The question now, can we be perfectly precise even more many new features are added? Initially that will be great until a specific point where you can perceive that increasing the dimensions causing a degrade the performance of the model. The following Fig.2.1 illustrates the meaning of Curse of Dimensionality:

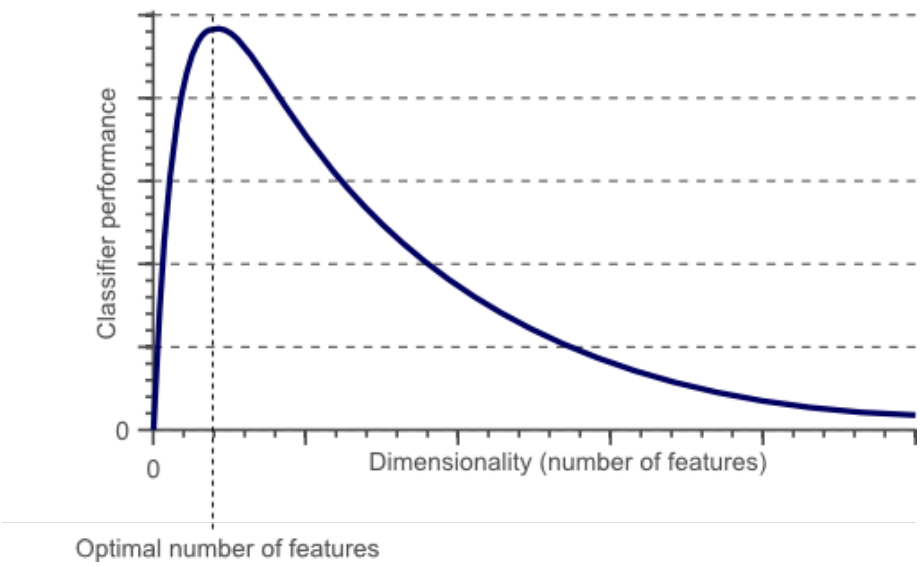


Figure 2.1: The Curse of Dimensionality

In terms of classification, which uses several input variables to produce output variable,

the more input variables we have the less efficient the model is.

We can see that the performance of a classifier decreases when the dimensionality of the problem becomes too large! But how to avoid this curse of dimensionality?

Unfortunately, there is no specific criteria which tells how many number of features is enough for dimensionality reduction purposes. In general, there are three main factors: how many training data available, the technique of reductions used, and the complexity of the decision boundaries.

"The smaller the size of the training data, the less features should be used. If N training samples suffice to cover a 1D feature space of unit interval size, then N^2 samples are needed to cover a 2D feature space with the same density, and N^3 samples are needed in a 3D feature space. In other words, the number of training instances needed grows exponentially with the number of dimensions used"[38].

Suppose there is a dataset with L features, how can we extract the "optimal" subset of H features where H must be less than L?

Many approaches can be used to replace the set of L features by a set of H features where (H<L). Feature selection, feature extraction, and cross-validation are worked in order to avoid overfitting due to the curse of dimensionality[38].

In general, dimensionality reduction can be divided into feature selection and feature extraction.

Feature selection attempts to choose a subset of the original features to be used in the machine learning model, so we can get rid of redundant and irrelevant features without losing important information.

Feature extraction or which is called 'feature projection' brings new features by projecting the data from a high dimensional space to a fewer dimension one. In this way, we can also derive informative and non-redundant features. Whereas feature selection returns a subset of the original features.

Sometimes we can use a mixture of feature selection and feature extraction together. For example, feature extraction can be performed on selected variables which contains

relevant information instead of using the original features and by this approach we will have better human interpretations through a simplification model[39].

As an important relevant addition to this topic that is worth mentioning is that there are two main types of tasks within the field of machine learning that are supervised and unsupervised learning. In the first case the supervised method, the training set is labeled and the results of training set are known. However, in the second case, in the unsupervised approach, labeling classes of the data is not required and the final outcome is unknown. Principal Component Analysis (PCA) is the most popular applicative method of the unsupervised approach[40].

In the case where the labeled data are limited in large data, labeling the rest can be costly. But when the data are unlabeled, things will be easier. And to get an effective utilize semi-supervised learning, was propose gathering both labeled and unlabeled data at the same time. This idea now used in many real life applications [41].

Dimensionality reduction techniques can be classified to be a supervised method like (linear discriminant analysis, LDA, and support vector machines, SVM) or unsupervised as (principal component analysis PCA, singular value decomposition, SVD, and independent component analysis, ICA) and there is some semi-supervised algorithm for dimensionality reduction. In this regards, we need to know that LDA is a supervised method and PCA is unsupervised since these techniques are substantial in this research[31].

The main objective for applying dimension reduction remains to find the most relevant features that plays a dominant role in interpretation the results by representing the data in a lower-dimensional space and preserving the helpful properties of the original data. More specific definition compared to the one presented in page (12), is done by viewing dataset with large number of observations as a large matrix.We can abstract this matrix by finding “narrower” matrices that have only a small number of rows or a small number of columns and close to the original as possible. Therefore, we can use it more efficiently than the large matrix. The process of finding these narrow matrices is called “dimensionality reduction”[43].

Equation (2.1) below shows the reduction process from a high-dimensional data space

(data [N]) in to lower-dimensional space [K] [31].

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \longrightarrow reduce \quad dimensionality \longrightarrow y = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \tag{2.1}$$

where (k << n)

Now, we can take a deep dive in the main key techniques that are the subject of this research, which are principal component analysis (PCA) and linear discriminant analysis (LDA).

2.3 Principal Component Analysis (PCA):

Principal component analysis is very versatile, it considered to be the oldest and very well-known technique of multivariate analysis and dimensionality reduction methods. Its usage was not famed until the invention of electronic computers. Nowadays, it can be found in most of statistical computer packages[44].

PCA is one of the family of an unsupervised dimensionality reduction techniques. It is basically taking the high-dimensional data and converting it to a lower dimension and more tractable data using the dependencies between variables. It emphasizes variation and bring out strong patterns in the dataset. It is done by finding a few orthogonal linear combinations which are called the principal components “PCs” of the original variables that has the largest variance so that the data become more easy explored and visualized. PCA is also one of the simplest dimension reduction techniques. It is a robust way for doing such dimensionality reduction. PCA has been rediscovered many times by various scholars in several fields, so it acquired many names. It is known as the Karhunen-Loève transformation, the Hotelling transformation, the method of empirical orthogonal functions, and singular value decomposition, also it sometimes called Hotelling transform singular value decomposition. For simplicity we will only call it PCA [45].

2.3.1 Key features of PCA:

- PCA attempts to represents a data frame containing correlated variables in terms of uncorrelated components.
- The principal components identified account for successively smaller amounts of the variability in the data frame.
- By selecting those components that account for relatively large amounts of variability, PCA can be used to reduce a large number of correlated variables to a smaller number of uncorrelated components.
- PCA can help identifying the under-lying structure in the data and provide clues about causal connections.

But very simply, principal component analysis converts correlated variables into uncorrelated components. It accomplishes this by identifying directions in the data (called components) where the variation is at a maximum and uses linear combinations of the observed variables that is used to describe the component. Below is the general form for the formula to compute scores on the first component extracted in a principal component analysis:

Principal Component 1 = β_{11} Variable 1 + β_{12} Variable 2 +... + β_{1k} Variable k where β_{1k} is the coefficient of Variable k in component 1.

Principal component 1 accounts for the largest amount of variation in the data that can be accounted for by a single linear model. Additional principal components can be derived by applying other linear models which identify sources of variance uncorrelated with the first principal component. Principal component 2 can therefore be computed using the linear model:

Principal Component 2 = β_{21} Variable 1 + β_{22} Variable 2 +... + β_{2k} Variable k where β_{2k} is the coefficient of Variable k in component 2.

Principal component 2 accounts for the largest amount of variation in the data that can be accounted for by a single linear model after principal component 1 has been accounted for. This process continues with additional components being computed which account for successively smaller amounts of variance in the data until all variance in the data has

been accounted for. This happens when the number of components equals the number of variables"[46].

Next, after this brief description, let us move to the theoretical meaning of the PCA and its definition.

2.3.2 Definition of PCA:

"The most common definition of PCA, due to Hotelling is that, for a given set of data vectors x_i, i $\in$ 1 ...t, the d principal axes are those orthonormal axes onto which the variance retained under projection is maximal"[47].

2.3.3 The PCA Space:

Suppose we have a dataset with number of dimension n and we want to apply PCA as a dimensional reduction technique.

Given a set of data on n dimensions, PCA aims to find a linear subspace of dimension lower than n such that the data points lie mainly on this linear subspace. Such a reduced subspace attempts to maintain most of the variability of the data.

The linear subspace can be specified by d orthogonal vectors which form a new coordinate system and are called the 'principal components'. The principal components are orthogonal, linear transformations of the original data points, so there can be no more than n of them. However, the hope is that only d<n principal components are needed to approximate the space spanned by the n original axes[48].

2.3.4 Assumptions for the PCA:

There are key properties and assumptions that govern PCA, and these are:[49]
Orthogonality: this technique supposes the principal components to be orthogonal to each other.

Here in Fig. 2.2 are some example datasets where PCA fails because the above as-

sumptions do not hold:

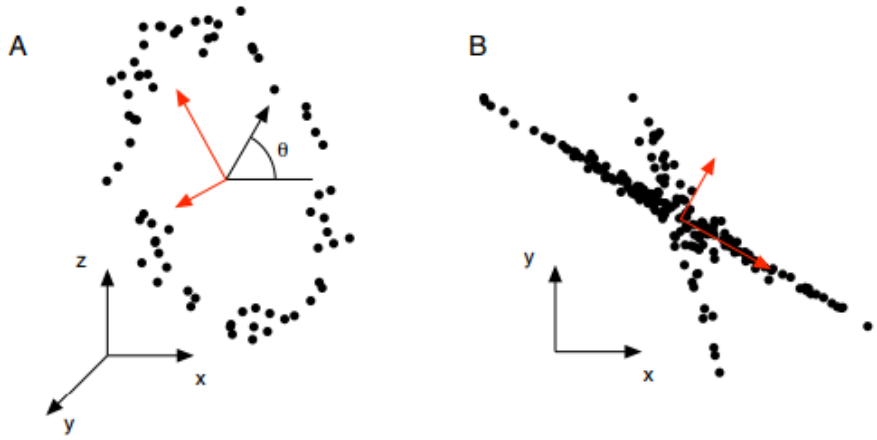


Figure 2.2: OrthogonalityAssumption

Where the red arrows represent the principal axes, which do not capture the complete signal[49].

Linearity: the principal components are assumed to be a linear combination of the original variables. If this condition is not satisfied, then the PCA will not work well and produce insensible results.

Large variance implies more structure: in this method the importance of a particular dimension is measured using the variance. Where the axes with high variance considered to be the principal components. Conversely, the directions with very little variation express the linearity among the origin features and these are noisy.

2.3.5 Performing Principal Components Analysis

The following example shows the general step for performing PCA.

Step1: Get some data

The data in the next table 2.1 will be used to apply PCA:

Table 2.1: Get some data

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Step 2: Subtract the mean

The next step is to adjust the data by Finding the mean of x and y and then subtracting the mean from each value in the data.

To find the mean, we need to calculate the sum of values and divide it on the number of values which is 10.

$$\bar{x} = 1.81$$

$$\bar{y} = 1.91$$

Then, we'll subtract the mean from each value in each column to get the adjusted data as shown in the following table 2.2.

Table 2.2: Adjusted data

Adjusted Data =

x	y
0.69	0.49
-1.31	-1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	-0.31
-0.81	-0.81
-0.31	-0.31
-0.71	-1.01

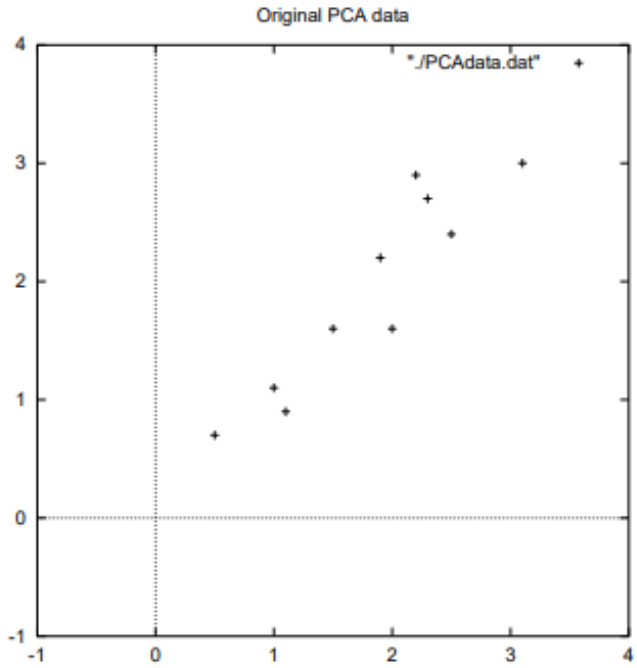


Figure 2.3: PCA example data, original data on the left, data with the means subtracted on the right, and a plot of the data

Step 3: Calculate the covariance matrix

The next step is to calculate the covariance matrix(C). Since the data is 2 dimensional, the covariance matrix will be 2×2 matrix.

$$\begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix}$$

Some points to note: Down the main diagonal, you see that the covariance value is between one of the dimensions and itself. These are the variances for that dimension i.e. cov(x,x) = var(x).

The other point is that since cov(x,y) = cov(y,x), the matrix is symmetrical about the main diagonal.

So that, we need to find var(x), var(y) and cov(x,y) using the following equations.

$$var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{2.2}$$

$$var(y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{2.3}$$

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{2.4}$$

Let's start finding the var(x):

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.69	0.4761
-1.31	1.7161
0.39	0.1521
0.09	0.0081
1.29	1.6641
0.49	0.2401
0.19	0.0361
-0.81	0.6561
-0.31	0.0961
-0.71	0.5041
	Sum=5.549

Now, using the sum from the above table and dividing it by (n-1), which is equal 9, then we get var(x) = 0.616555556, which is the first value in covariance matrix.

In the same way we calculate var(y).

$y_i - \bar{y}$	$(y_i - \bar{y})^2$
0.49	0.2401
-1.21	1.4641
0.99	0.9801
0.29	0.0841
1.09	1.1881
0.79	0.6241
-0.31	0.0961
-0.81	0.6561
-0.31	0.0961
-1.01	1.0201
	Sum=6.449

using the var(y) equation, we get var(y) = 0.7165555, and this value is the forth value in covariance matrix.

Now, to find the second and third value in covariance matrix we will calculate cov(x,y) and cov(y,x) which are equal.

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
0.69	0.49	0.3381
-1.31	-1.21	1.5851
0.39	0.99	0.3861
0.09	0.29	0.0261
1.29	1.09	1.4061
0.49	0.79	0.3871
0.19	-0.31	-0.0589
-0.81	-0.81	0.6561
-0.31	-0.31	0.0961
-0.71	-1.01	0.7171
		Sum=5.539

using the table above, we find cov(x,y)=cov(y,x) = 5.539/9 = 0.6154444.

So, now the covariance matrix will the following 2×2 matrix.

$$\begin{bmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{bmatrix}$$

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

The next step will be to calculate the eigenvalues and eigenvectors of the covariance matrix.

Firstly, let's start finding λ using the following formula:

$$\left|A - \lambda I\right| = 0 \tag{2.5}$$

Where A is the covariance matrix and I is the identity matrix.

$$\begin{vmatrix} 0.6165 - \lambda & 0.61544 \\ 0.61544 & 0.71655 - \lambda \end{vmatrix} = 0 \tag{2.6}$$

Solving the above determinant:

$$\lambda^2 - 1.33305\lambda + 0.4417963 = 0 \tag{2.7}$$

After solving the above quadartic equation, we find:

$$\lambda = 0.04908$$

Or

$$\lambda = 1.28402771$$

Now, substitute the value of $\lambda = 0.04908$ in $A - \lambda I$ to find the B matrix, where B is a square matrix used to solve the system $B\vec{x} = 0$.

$$\text{B} = \begin{bmatrix} 0.5675 & 0.61544 \\ 0.61544 & 0.66755 \end{bmatrix}$$

Now, find the value of x1 and x2 by solving the system:

$$\begin{bmatrix} 0.5675 & 0.61544 \\ 0.61544 & 0.66755 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$0.5675x_1 + 0.61544x_2 = 0 \tag{2.8}$$

$$0.61544x_1 + 0.66755x_2 = 0 \tag{2.9}$$

The first eigenvector which generated by solving Eq.2.8 is:

$$\begin{bmatrix} -0.735178656 \\ -0.677873399 \end{bmatrix}$$

And the second eigenvector which generated by solving 2.9 is:

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

Look at the Fig. 2.4, it shows that the data has a strong pattern.

As expected from the covariance matrix, the x and y increase together.

In Fig. 2.4, the two eigenvectors appear as a diagonal dotted lines, and they are perpendicular to each other.
One of the eigenvectors goes through the middle of the points. This eigenvector is the most important than the other.

Now, the remaining steps involves transforming the data in terms of the two lines.

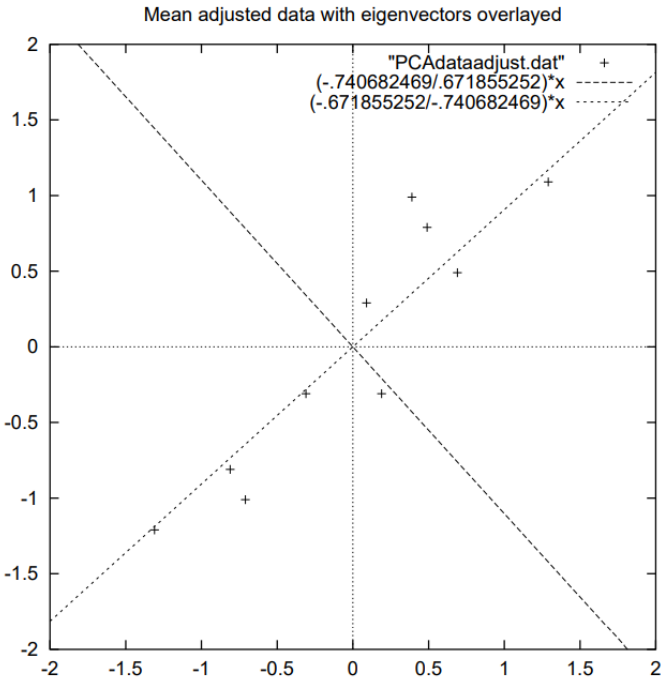


Figure 2.4: A plot of the normalised data (mean subtracted) with the eigenvectors of the covariance matrix overlayed on top

Step 5: Choosing components and forming a feature vector:

Look at the eigenvectors and eigenvalues, the eigenvector with the highest eigenvalue is the principal component of the data set.

In this example, the eigenvector with the highest eigenvalue is the one that pointed down the middle of the data:

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

Now, after eigenvectors are found from covariance matrix, we order them based on the value of eigenvalue from highest to lowest. This gives you the components in order of significant.

Now, we need to form a feature vector by taking eigenvectors that you need to keep.

$$FeatureVector = (eigenvector_1 \qquad eigenvector_2 \qquad eigenvector_n)$$

In our example, we have two choices, we can form a feature vector from the 2 eigenvectors which are:

$$\begin{bmatrix} -0.677873399 & -0.735178656 \\ -0.735178656 & 0.677873399 \end{bmatrix}$$

$FeatureVector = (eigenvector_1 \qquad eigenvector_2)$

Or we can ignore the smaller (less significant one), and only take the following eigenvector:

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

$FeatureVector = (eigenvector_1)$

Step 6: Deriving the new dataset:

This is the final step, here we simply take the transpose of the vector and multiply it on the left of the original dataset transposed.

Final Data = Row Feature Vector(transposed of eigenvector) $\times$ Row data adjust(mean-adjust data transposed)

The transformed data using two eigenvectors:

Table 2.3: Transformed data using two eigenvectors

Transformed Data =

x	y
-0.827970186	-0.175115307
1.77758033	0.142857227
-0.992197494	0.384374989
-0.274210416	0.130417207
-1.67580142	-0.209498461
-0.912949103	0.175282444
0.0991094375	-0.349824698
1.14457216	0.0464172582
0.438046137	0.0177646297
1.22382056	-0.162675287

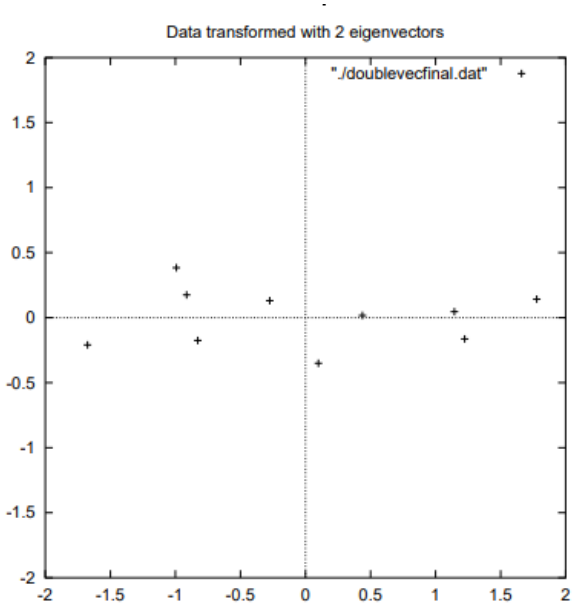


Figure 2.5: The table of data by applying the PCA analysis using both eigenvectors, and a plot of the new data points.

The transformed data using single eigenvector:

Table 2.4: The data after transforming using only the most significant eigenvector

Transformed Data =

x
-0.827970186
1.77758033
-0.992197494
-0.274210416
-1.67580142
-0.912949103
0.0991094375
1.14457216
0.438046137
1.22382056

2.3.6 Problems with PCA:

Although PCA is primarily the best known and a commonly used technique for applying dimensionality reduction, there are some limitations and problems facing the users:

Firstly, PCA assumes the space of the input data to be approximately normally distributed. But sometimes it can deal with data which is not normal and give a new lower space. However, this method probably failed when the data is on a complicated manifold.[51].

PCA also depend on the "linear assumption" that is when you have variables that are linearly correlated in order to find an orthogonal projections with the highest variance. But when the data is not linear this method will fail. Look at the following example to make sense:

Imagine two variables that represents the size of something in cm and inch respectively (the values of those variables are correlated by the formula 2.54 cm = 1 inch), if you add noise and plot the data you will get something similar to this picture:

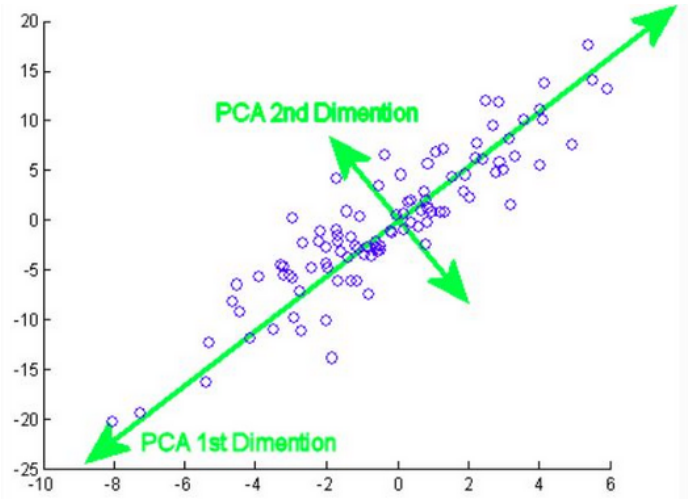


Figure 2.6: Orthogonality Assumption

But if the data is not linearly correlated (for example, in spiral, where $x = t * cos(t)$ and $y = t * sin(t)$), PCA is not enough: [52]

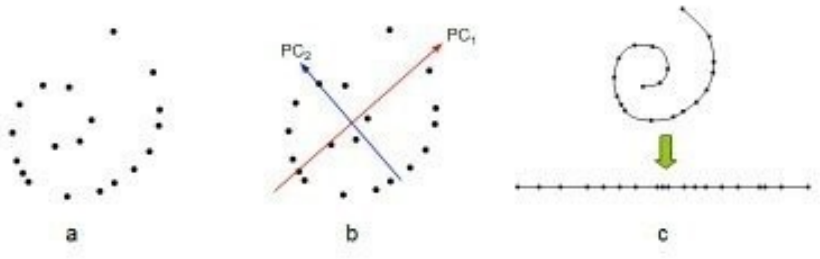


Figure 2.7: When the data is not linearly correlated

In addition, the principal components rely on the "orthogonal transformations" where the components are orthogonal to the others. This restriction is required in order to find the projection with maximal variance as the next graph explains:

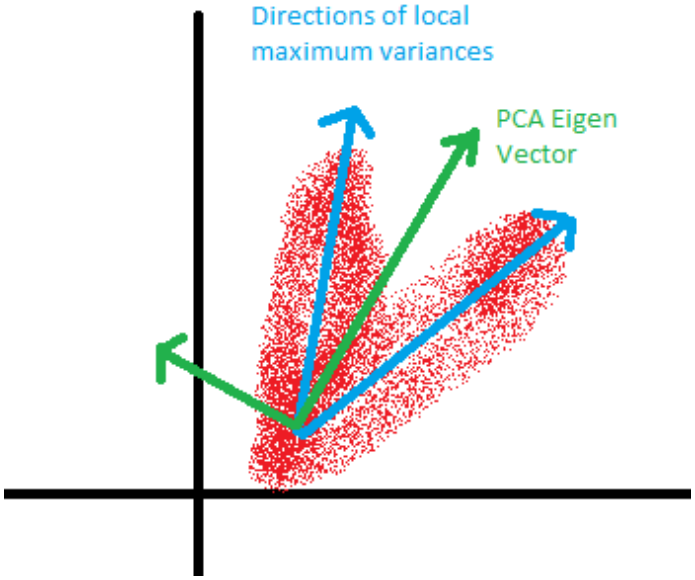


Figure 2.8: Orthogonal Transformation

PCA deal with known assumption that is "large variance = low covariance = high importance" but this depends on the nature of the data and what the problem of interest required. In particular,

- If you want to compress or remove noise from your dataset this assumption is an advantage
- For mostly any other problem (like Blind Source Separation) it is not useful. Based on Independent Component Analysis theory: uncorrelated is only partly independent (unless it is totally normal).

As we see before, the PCA always concerned by the mean and the covariance of the data set. But these not necessarily can investigate and describe many distributions.

Finally, suppose that you have to change the scale of some of the variables, the results will differ when applying PCA before and after change the scale.

Although the PCA facing some limitations, it still a preferred technique and has various advantages that are branch around the main idea of the PCA which is “elimination of the correlated variables”. Since this process does not make big changes in the final results due to removing the correlated variables. per contra, this reduction of number of features leads to:

1. Lessening overfitting in the given model.
2. Reducing the training time.
3. Improving the performance of the algorithms.
4. Improving visualization by converting the graph of features from N dimensions into only two visible dimensions[53].

Chapter 3

Linear Discriminant Analysis (LDA)

3.1 Introduction:

Linear discriminant analysis, or simply (LDA) is a type of Linear combinations. It is a successful and well-known mathematical process of dimensionality reduction and classification technique. It is sometimes named as “Fisher Discriminant Analysis” (FDA) which related to the statistics professor at University College London “Sir Ronald Fisher” who developed this method in 1936.

The basic idea of the linear combinations goes back to the sixth decade of the 20th century. It was appearing with the ‘Altman Z-scores ‘for bankruptcy. Nowadays, LDA can keep preventative data which consists of two classes in the cases when the ‘Logistics Regression’ is failed.[55]

LDA technique is considered to be one of the most well-used data reduction techniques which helps to represent data for two or more classes, when logistic regression is not sufficient. [56] It appears extensively in many statistical areas such as: pattern recognition problems, speech recognition, face recognition, predictive analysis in marketing, image recognition, multimedia information retrieval, noise removal and geology fields.[57]

The main goal of this method is to separate samples of distinct groups through transforming the data to another space that is ”optimal” to discriminate the classes, [58] i.e. LDA tries to find the vectors that make the best discrimination between classes instead of

focusing on the data that gives the best description of the origin datasets (by projecting data points into a line). More precisely, projecting data into a surface of dimension D-1 (as when your hand's shadow falls in a wall) This is what happened with the data points when it falls in the new line. See the figure below Fig. 3.1:[59]

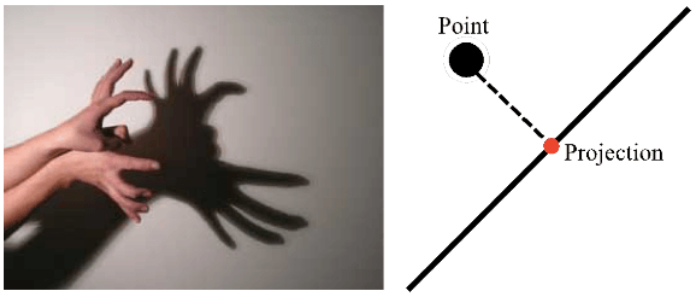


Figure 3.1: Projecting data onto a surface

This method produces a linear combination of independent variables that gives the highest mean differences between such classes. LDA maximizes the ratio of 'between-class variance' to 'within-class variance' in the used dataset in a way that guarantee having the maximum separation.[60]

This classical multivariate technique (LDA) converts the high dimensional space into a subspace with lower dimension. In this subspace LDA works as a simple prototype classifier with linear decision boundaries.[62]

LDA seeks a linear combination, the discriminant function which acts in a way that maximizes the ratio of the between- group sum of squares compared with the within-group sum of squares and find a relation between group that has maximal difference on it to achieve the required separation analyze which consists of multiple classes of items.[61]

The next photo Fig. 3.2 was taken by 'Jamie McCaffrey' [63]. We can see that the bride stands in the middle in a way which separates men and women. This explains the actual meaning of the LDA where the bride represents the axes that separates the two classes men and women.

Figure 3.2: Photo by Jamie McCaffrey, ‘Representation of LDA Models’

LDA is the preferred linear classification technique, due it’s is simplicity in preparation and application, it provides statisticians with algebraic and graphically description of the features of observations from the collected dataset. It aims to find the good discriminants which can make the most possible separation in order to sort the observations into two or more labeled classes [64].

LDA has a powerful ability for solving multivariate problems to get a small sample problem with easy steps to be solved and separated into predefined groups. So, this technique is a like between the univariate and multivariate statistical problems.[65]

Dimensionality reduction techniques in general have two mainly different approaches, which are supervised and unsupervised approaches. In the supervised approach, the dimension reduction techniques require labeling the classes of the data. However, the unsupervised approach doesn’t take into consideration the labels of the classes.[66]

Non-negative Matrix Factorization (NMF) and Independent Component Analysis (ICA) [67] are famous examples of unsupervised dimensionality reduction techniques, whereas PCA is the most popular technique as discussed in the previous chapter.

On other hand, the Linear Discriminant Analysis (LDA) is the most familiar supervised technique in addition to other technique such as Neural Networks (NN) [68] and Mixture Discriminant Analysis (MDA).[69]

LDA technique is divided into two types depending on the data used. The first type is a class-dependent LDA (CDLDA) focuses on the own features of a certain class that discriminates it from other classes. Also, the separation score to a lower dimension must be calculated for each class and the projection here done on the same class. While in the second type: the class independent LDA) CILDA (the focus is on the general features of the whole classes and every class will be treated as a separate class versus. In this type there is only a unique lower space that all classes can project its data on it. [70][71]

3.2 Preparing the Data for Applying the LDA:

Here are some suggested points that should be considered in preparing the data for the best effective use of LDA:

- **Removing outliers from data:**
  LDA separating of data is highly influenced by outliers, and this causes a skewness in the basic statistics concepts. For example, the mean and the standard deviation that help in separating the classes in LDA algorithm.
- **Same Variance:**
  It standardizes the data i.e. the mean of the data must equal to zero and it have one standard deviation. This is an important assumption of the LDA that states that all variables in the input data must have the same variance.
- **Normality assumption:** LDA model assumes the input variables follow "the normal distribution with its popular bell-shaped curve.

Sometimes certain transformations are used to convert the data to be more normal distribution looking. Box-Cox is used in the skewed distributions cases and logarithm or root for exponential-like distributions.

the below example shows how to convert an exponential distribution to a normal one. The next histogram fig.3.3 shows the log normal distribution.

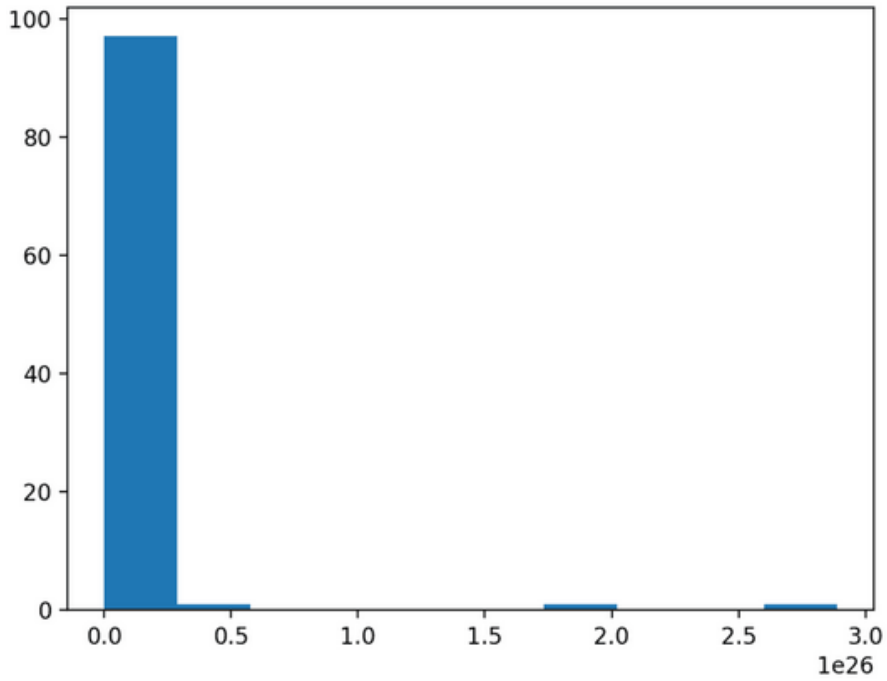


Figure 3.3: Histogram of a Log Normal Distribution

Taking the square root and the logarithm of the observation in order to make the distribution normal belongs to a class of transforms called power transforms. The Box-Cox method is a data transform method that is able to perform a range of power transforms, including the log and the square root. The method is named for George Box and David Cox.

More than that, it can be configured to evaluate a suite of transforms automatically and select a best fit. It can be thought of as a power tool to iron out power-based change in your data sample. The resulting data sample may be more linear and will better represent the underlying non-power distribution, including Gaussian.

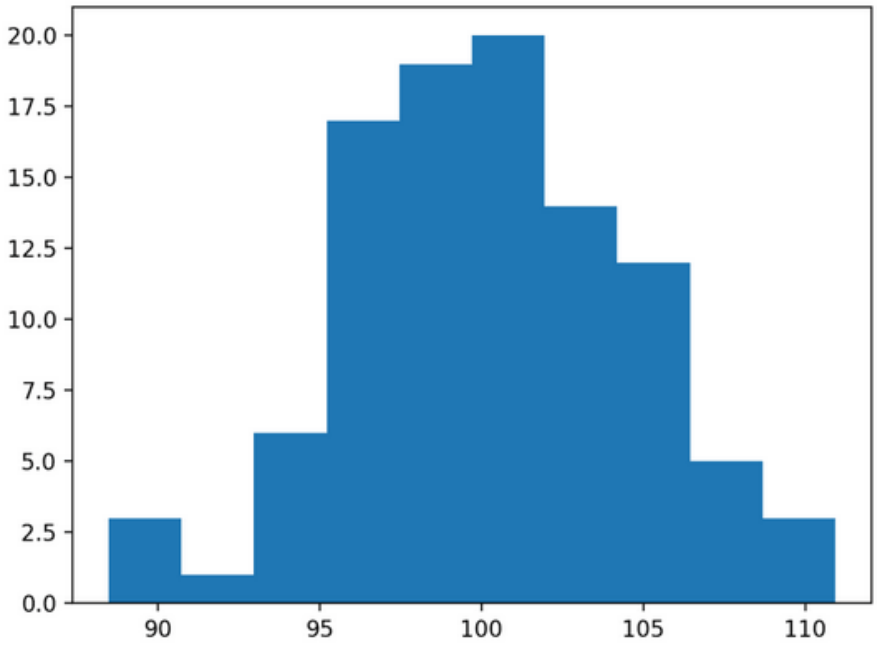


Figure 3.4: Histogram Plot of Box Cox Transformed Exponential Data Sample

Note that, the graph Fig.3.4 shows the bell-shaped for the transformed data set.

3.3 Assumptions for the LDA:

The LDA is sensitive for the following assumptions:

1. **Sample size:** There is no objection to existing samples of different sizes, but the size of the set should exceed the number of variables. For example, the maximum number of variables = N -2, where N is the sample size.

2. **Multivariate normality:** variables that are independent must be normally distributed in any stage of grouping the variables (for each level of the grouping variable).[72]

3. **Non-multicollinearity:** There is an inverse relation between the correlation between variables and the power of prediction, that is increasing levels of the correlation will cause a decrease in the predictive power. This assumption confirms that LDA requires the independency of variables. [18]

4. **Independency:** observations for a certain variable assumed to be independent.[72]

5. **Homogeneity of variance-covariance:** It also called “homoscedasticity”: here also the variance among groups must be the same in any level of predictions (i.e. $\sum 1 = \sum 2 = \sum 3 \ldots = \sum K$), where the $\sum K$ represent the variance among the group K, and K is number of classes in the dataset. ‘Box’s M statistic’ is a statistical test which used to examine equality of multiple variance-covariance matrices. [73] In the cases where the covariances among groups are not equal, another type of discrimination can be used which is that quadratic discriminant analysis.

   In general, the mentioned assumptions of the LDA are similar for that in MANOVA (Multivariate analysis of variance, which is an extension of ANOVA the univariate analysis of variance). Also, this method is susceptible to outliers and the following condition must be achieved: the smallest group must have a size that is greater than the number of predictor variables. [73]

   In general, the above assumptions are always not easy to be achieved. which make the LDA not flexible with many kinds of datasets especially in behavioral sciences. [74]

   Finally, when the LDA acts as a dimensionality reduction tool, these assumptions can be relaxed. That is LDA in this case is a robust technique and can work reasonably well. [75]

3.4 Key Features of LDA:

1. Linear discriminant analysis is used as a classification method and as a robust - dimension reduction technique. It also appears in data visualization. [76]

2. The LDA algorithm can be used when either the data are normal or not i.e. normality assumptions- when LDA works as a dimensional reduction tool- are often violated. This illustrate the strength of LDA. [77]

3. The LDA still works as a dimensionality reduction technique even though the ‘common covariance matrix among groups’ assumption is not satisfied. [77]

4. In spite of the simplicity of the LDA, it is considered to be the first and the ‘benchmarking’ technique before thinking of another complex and modern flexible methods. [76]

In general, the LDA performs well. Here are two distinct examples of applications of LDA:

(i) Bankruptcy prediction: Bankruptcy of a Company can be predicted using Edward Altman's model which back to 1968. The probability bankruptcy for any company small or large, private or public can be expected using 'trained LDA coefficients' with high accuracy sometimes reach to (80% - 90%).

(ii) Facial recognition: as mentioned before, the features in Principal Component Analysis in general (PCA) are known as 'Eigenvectors'. While the features in the LDA called 'Eigen Faces' especially in the face recognition applications. This naming refers to the statistician Sir Ronald Fisher. [76]

3.5 Performing Linear Discriminant Analysis LDA:

There are five main steps for applying LDA:

1. Calculate the "d-dimensional mean vectors" for each class in the dataset.
2. Compute the scatter matrices: the within-class scatter matrix and the between-class scatter matrix.
3. For the scatter matrices, compute the eigenvectors (e_1,e_2,...,e_d) and its corresponding eigenvalues (λ_1,λ_2,...,λ_d).
4. Order the eigenvectors with decreasing eigenvalues and choose "k "eigenvectors that have the largest eigenvalues in order to form "d$\times$k dimensional matrix" W (note that: every column represents an eigenvector).
5. The yield d$\times$k dimensional matrix can be used to transform the samples to a new subspace. This can be presented by the matrix multiplication Y=X$\times$W s.t(X is a n$\times$d-dimensional matrix with n samples, y represents the transformed n$\times$k-dimensional samples into the new subspace. [78]

3.6 Real Life Applications of LDA:

Here are some areas and examples related to the PCA:

1. **Face recognition:** pictures of faces consist of huge numbers of pixels. LDA plays main role in computerized facial image classification. It extracts the important features and reduces the high dimension to a lower dimensional space where the dimensions are a linear combinations of the origin pixels. These linear combinations are called 'Fisher faces' because Fisher's linear discriminant is used to extract them.

2. **Agriculture applications:** In this kind of applications, certain computational models are used to classify the unknown samples into 'pre-defined species'. Together with various variants of LDA which shrinks the big numbers of dimension in the collected data as in Plant identification approach. See [83] where LDA is used in Classifying the olive oil by using high throughput flow finger printing. [84]

3. **Biometrics applications:** LDA technique and its variant properties nowadays applied in many fields such as 'Biometrics Systems'. It is divided into two essential steps and these are: feature extraction step and recognition step. Feature extraction step contains extract and pre-processing the features from the origin data. However, in the recognition step the attention goes toward identifying the unknown samples such as unknown face image.

   In general, LDA applications can be found in face recognition [85], Ear [86], gait [87], speech [88], fingerprint applications [89]. Moreover, LDA sometimes used in animal biometrics. For more details, see the references.

4. **Medical applications:** In this field, the data in general has a high number of features which represents the dimensions such as the DNA and RNA microarray gene expression data and cancer classification [90]. These kind of data require expensive and extensive efforts to train their models. Also, the high dimensionality sometimes affects classification performance.

   LDA technique plays a major role in solving such these problems. It is done by reducing the original features and constructing a new and smaller set of important features.[91]

5. **Marketing:** Here LDA can determine reasons for identifying various kinds of products and customers according to the collected data. Moreover, LDA can be applied in positioning and product management fields.

6. **Detecting credit card fraud:** here linear discriminant functions act a good role in detecting credit card fraud using "Modified Fisher Discriminant Analysis".[92]

7. **Bankruptcy prediction:** LDA was the first statistical technique that explains which companies entered bankruptcy verse survived. Edward Altman's model 1968 is still the leader model in this kind of applications.[93]

8. **Earth science:** this sort of applications focuses on separation the 'alteration region'. LDA tries to find the pattern for available data sets comes from different zones and produces an efficiently classification method. For example: discriminant analysis used for alteration separation; sungun copper deposit, East Azerbaijan, Iran. Australian as mentioned in this reference [94]

3.7 How Does the LDA Work?

In order to understand how the LDA works, let us discuss the following real life example which is about cancer drug, and this drug works great for some people, however it makes other people feel worse. The important question is: how do we decide who to give the drug to or who should not receive it?

Gene expression can help in deciding.

First, let use one gene, let's say Gene-X, to decide who to give drug and who should not.

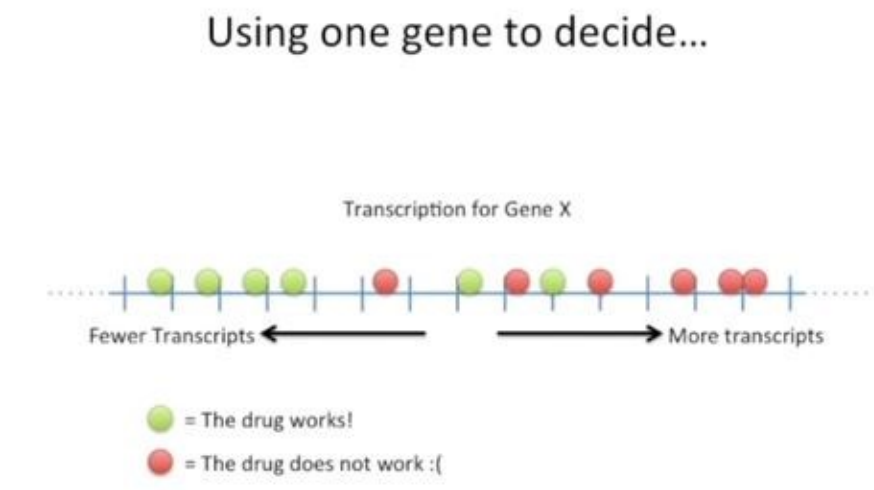
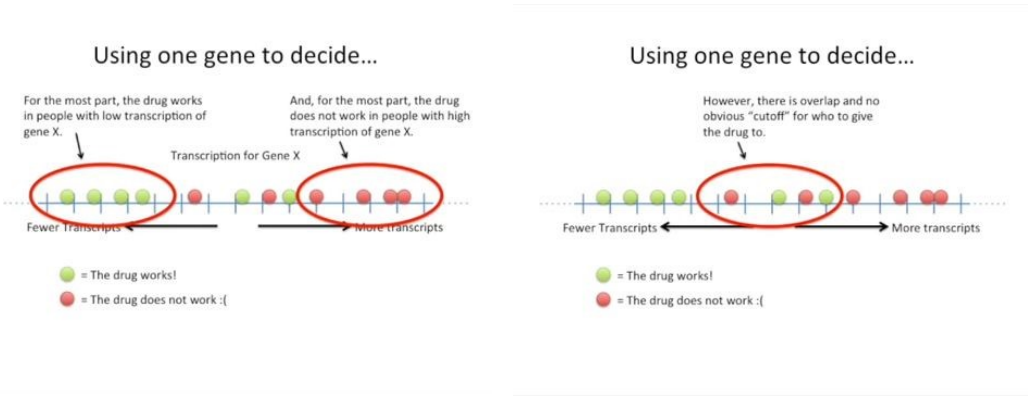


Figure 3.5: Using one gene to decide

In the above Fig. 3.5, the dots represent individual people: the green dots for people the drug works for them while the red dots represent people the drug does not work for them. In Fig. 3.6a, the people have fewer transcripts of Gene-X than people on the right on Fig. 3.6b who have more transcripts.



(a) (b)

Figure 3.6: (a) People have fewer transcripts of Gene-X (b) people have more transcripts of Gene-X

So as seen in the above figure, mostly, the drug fit with people who are less transcripts of Gene-X. While most of people that the drug doesn't work who have highly transcript of Gene-X. However, things are not obvious i.e. there is an overlap and "cutoff" for whom the drug can be given to in the middle area.

In short, Gene-X gives good classifier for telling us who should take the drug and who shouldn't, but can we do better if we use more than one gene?

Let's use two genes to decide (Gene-X and Gene-Y).

As shown in the following figure Fig. 3.7, using two genes do a better job separating the two categories, however it is not perfect!

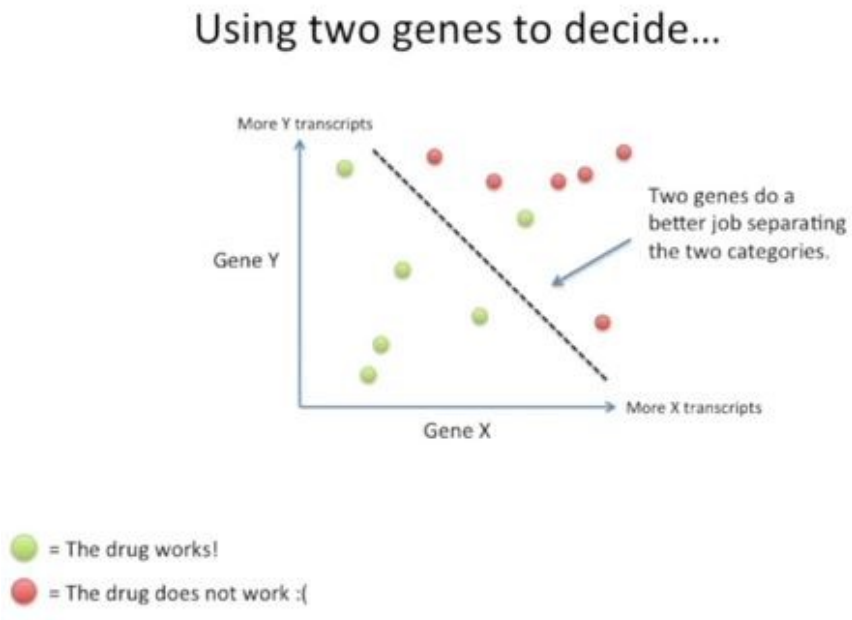


Figure 3.7: Using two genes to decide

Now, would three genes be even better? Let's see the figure below Fig. 3.8 which represents using Gene-X, Gene-Y and Gene-Z.

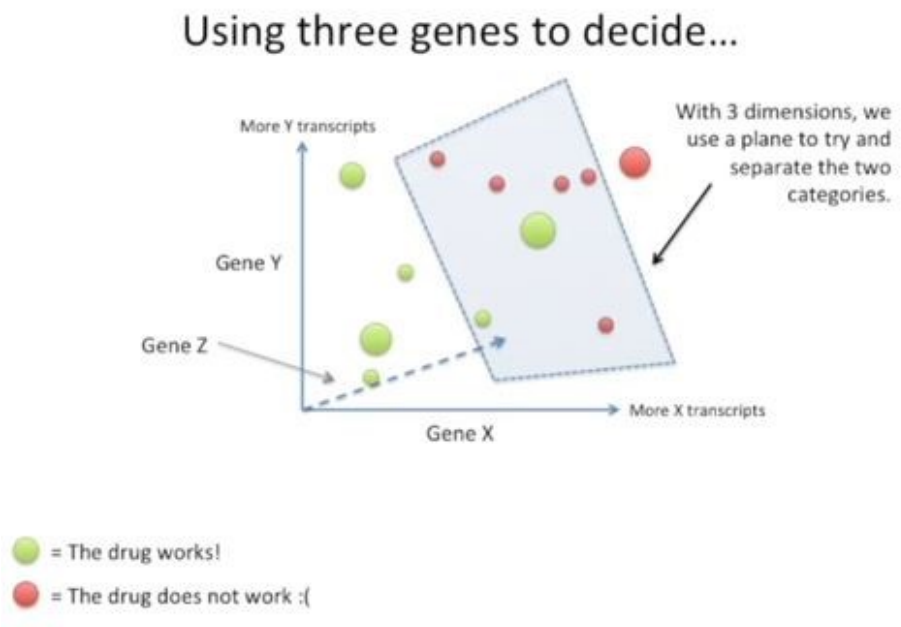


Figure 3.8: Using three genes to decide

And as we can see, the Gene-Z is on the Z-axis, so imagine a line going through this page into the wall behind it and the big circles are the ones that are closer to you and the smaller circles are further away. When we have 3 dimensions, we use a plane to try to separate the two categories and as you can see it is very difficult to decide something about it.

What if we need four, five or more genes to separate the two categories?!

The first problem we can't draw 4-D graph or 1000-D graph, so here the LDA can be really vital.

Linear Discriminant analysis focuses on maximizing the seperability among known categories.

Let us explain this with a simple example of reducing a 2-D graph to a 1-D graph as shown in Fig. 3.9 and Fig. 3.10.

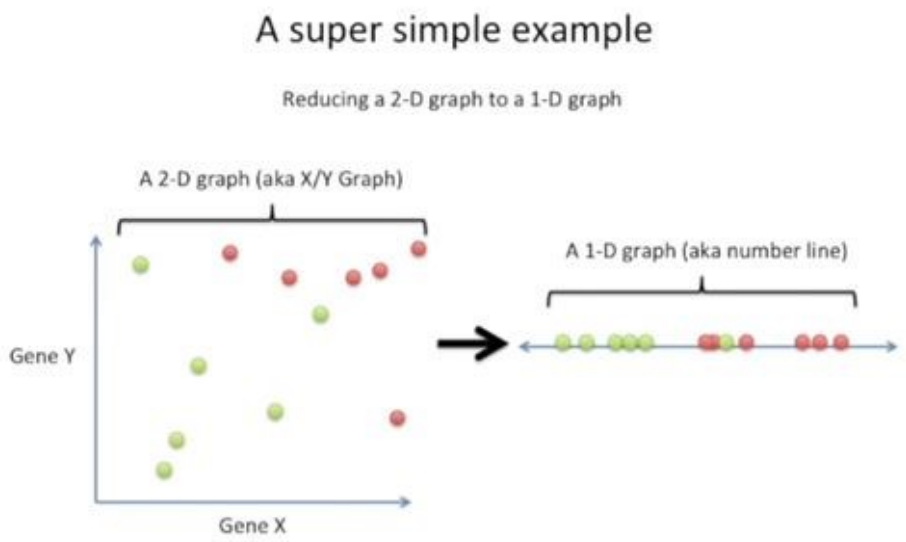


Figure 3.9: A super simple example of reducing 2-D graph to 1-D graph

Here, we want to reduce the two dimensional graph to only one dimensional graph in such a way that maximizes the separability between two categories, but how do we do this?

We can do this by ignoring Gene-Y (here only we need to project the data down to the x-axis. This is a bad option! Since it disregards the useful information that Gene-Y provides.

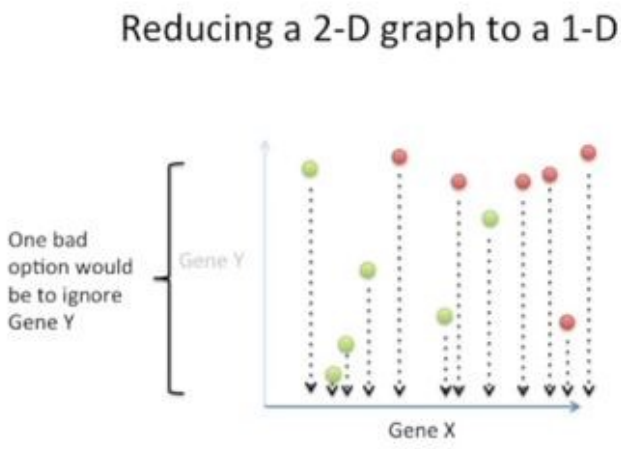
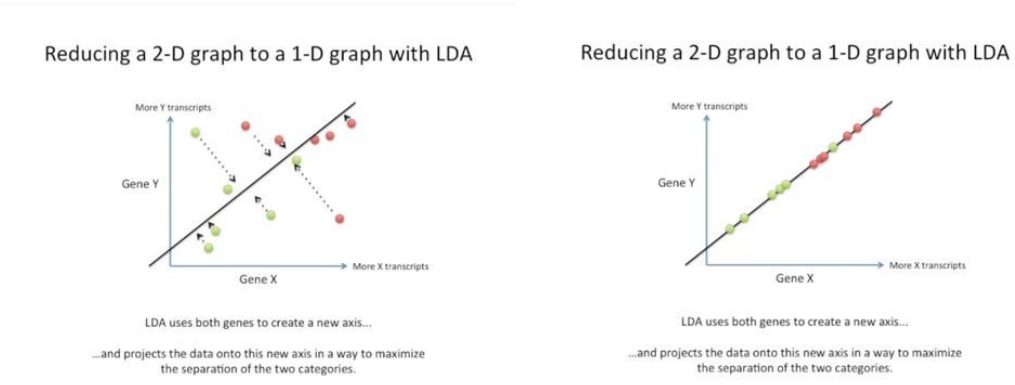


Figure 3.10: Reducing a 2-D graph to 1-D graph

LDA can solve the problem in a better way, it maintains using information from both Genes to create a new axis, and it projects the data to this new axis in a way that maximizes the separation between the two categories as shown below in Fig. 3.11a.



(a) (b)

Figure 3.11: (a) Reducing a 2-D graph to 1-D with LDA to create new axis (b) Reducing a 2-D graph to 1-D with LDA to create new axis

Now, let us get in more details in how LDA creates the new axis. The new axis is created according to two criteria:

1. Maximize the distance between means as in the figure below. We'll have a new μ character which represents the mean for the green category and a red μ character which represents the mean for the red category.

2. Minimize the variation "scatter" and is represented by s2 within each category.

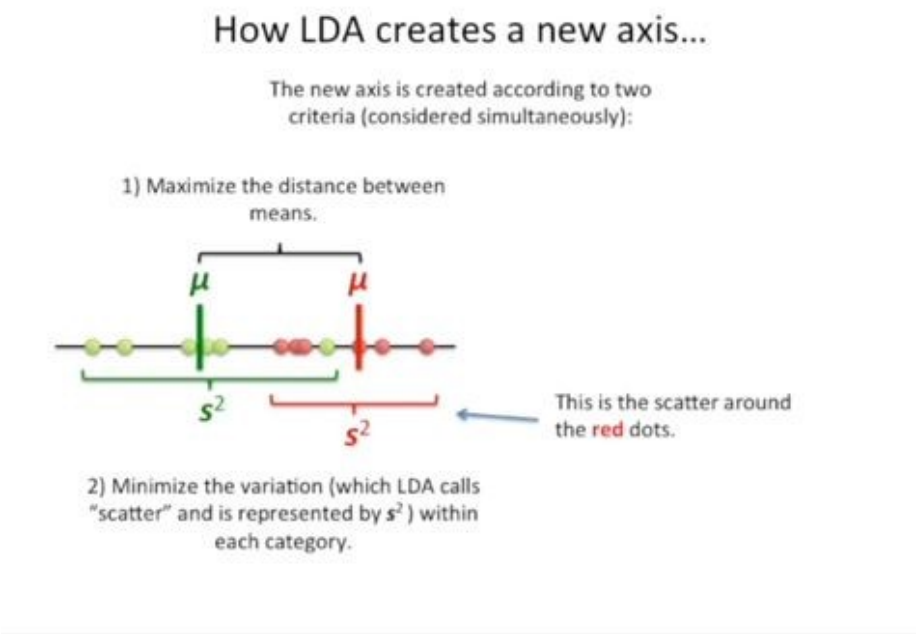


Figure 3.12: How to create a new axis with LDA

And we can use the relation $(\mu_{green} - \mu_{red})^2 \ /(S^2_{green} + S^2_{red})$ to consider the two criteria simultaneously, and to make this equation simpler and we can represent $(\mu_{green} - \mu_{red})$ with d for distance.

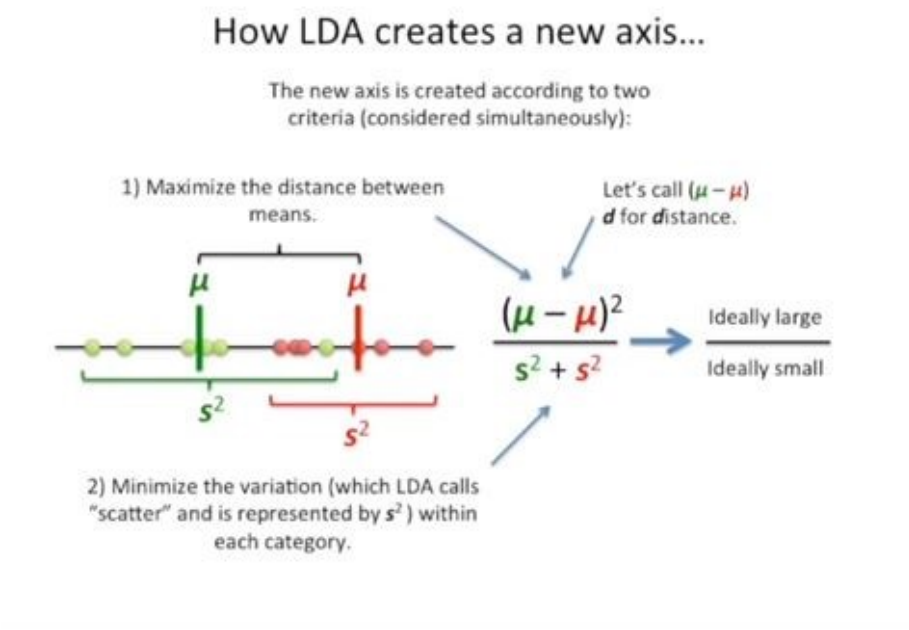


Figure 3.13: How to create a new axis with LDA

Let's see an example of why both distance and scatter are important. In next figure Fig. 3.14 which represents 2-D data, if we only maximize the distance between means, then we'll get a result as shown in the next figure, and the separation isn't great. However, if we optimize the distance between means and scatter, then we get a nice separation as show in the following figure.

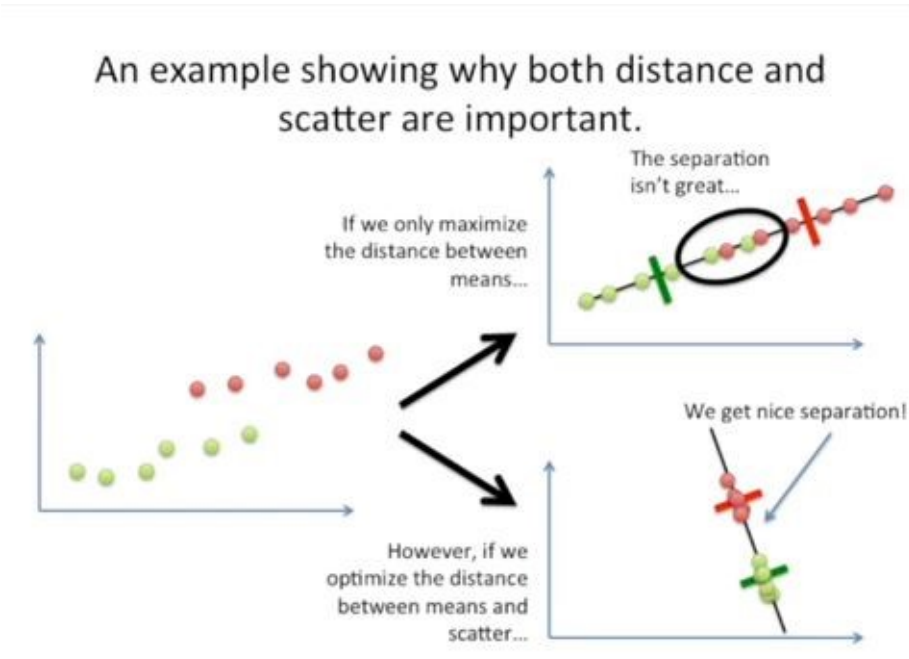


Figure 3.14: An example showing why both distance and scatter are important

So, now what if we have more dimensions, the process is the same which is creating an axis that maximizes the distance between the means for the two categories and minimizes the scatter as explained in next Fig. 3.15.

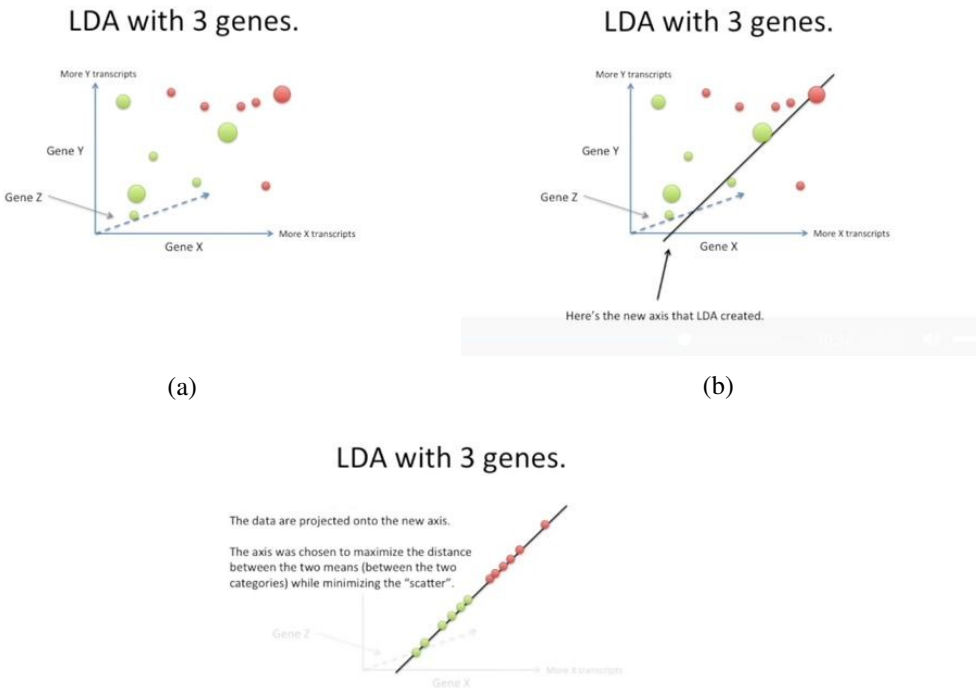


Figure 3.15: (a) Creating a new axis for three dimensions step 1 (b) Creating a new axis for three dimensions step 2 (c) Creating a new axis for three dimensions step 3

Now, if we have 3 categories, two things change, but hardly.

The first difference is how you measure the distance among the means, we first find a point that is central to all of the data, then we measure the distance d^2 between a point that is central in each category and the main central point.

Then we maximize the distance between each category and the central point while minimizing the scatter for each category, and for this we use the relation $(d^2_{green} + d^2_{red} + d^2_{blue})/(s^2_{green} + s^2_{red} + s^2_{blue})$.

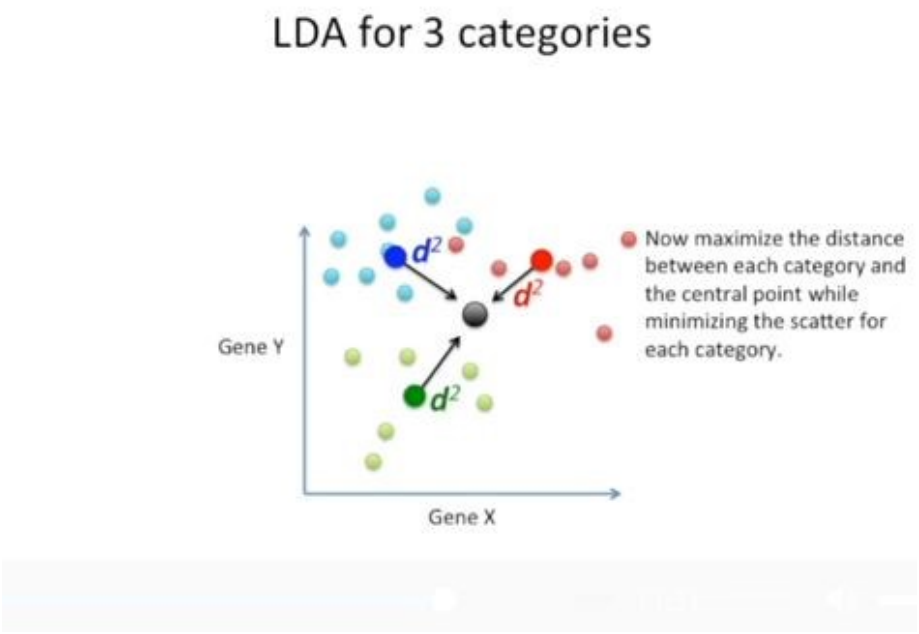


Figure 3.16: LDA for 3 categories

The second difference is LDA creates 2 axes to separate the data. This is because the 3 central points for each category define a plane and this is shown in Fig. 3.17.

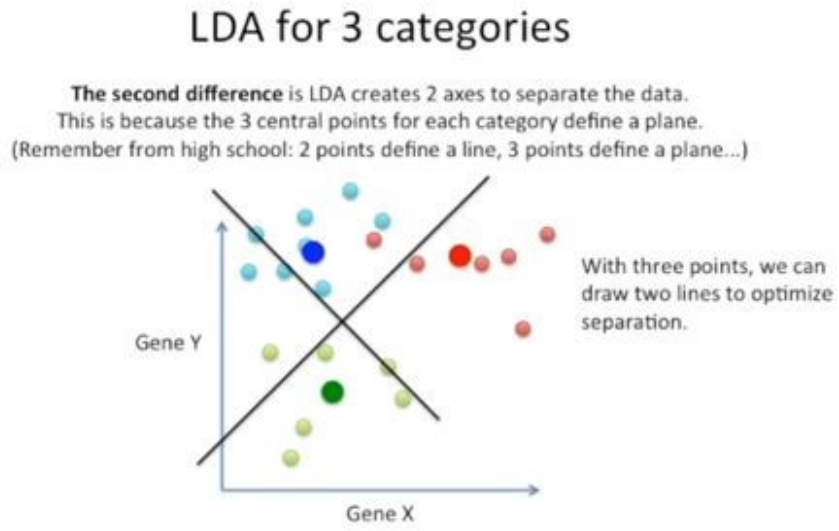


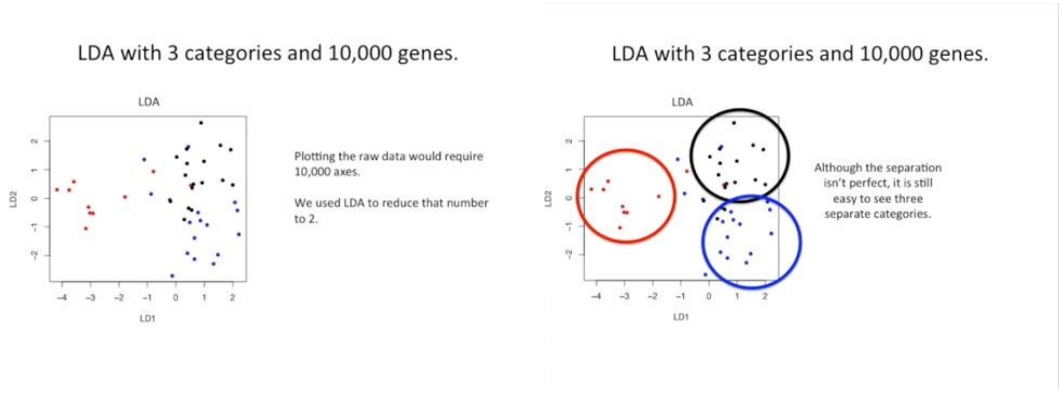
Figure 3.17: LDA for 3 categories

When we only use 2 genes, this is no big deal. The data started out on a X-Y plot and

plotting them on a new X-Y plot doesn't change much.

But if we used data from 10,000 genes, that would mean we'd need 10,000 dimensions to draw the data.

In the next figure 3.18 is a real data, and we try to separate 3 categories of 10,000 genes. Plotting the raw data would require 10,000 axes. So, we use LDA to reduce that number to 2.



(a) (b)

Figure 3.18: (a) LDA with 3 categories and 1000 genes (b) LDA with 3 categories and 1000 genes

3.8 Performing Linear Discriminant Analysis numerical example

Let's take an example and workout step by step.

Let's take a 2-D $(x_1 and x_2)$ dataset with two classes C_1 and C_2.

$C_1 \rightarrow x_1 = (x_1, x_2) = \{(4,1),(2,4),(2,3),(3,6),(4,4)\}$

$C_2 \rightarrow x_2 = (x_1, x_2) = \{(9,10),(6,8),(9,5),(8,7),(10,8)\}$

Step 1: Compute within-class scatter matrix (S_W)

$$S_W = S_1 + S_2 \tag{3.1}$$

Where S^-1 is the covariance matrix for the class C_1 and S_2 is the covariance matrix for class C_2.

So, let's now find the covariance matrix of each class:-

$$S_1 = \sum_{x \in C_1} (x - \mu_1)(x - \mu_1)^T \tag{3.2}$$

Where μ_1 is the mean of C_1, which is computed by:-

$$\mu_1 = \begin{bmatrix} \frac{4+2+2+3+4}{5} & \frac{1+4+3+6+4}{5} \end{bmatrix}$$

$$\mu_1 = \begin{bmatrix} 3 & 3.6 \end{bmatrix}$$

Similarly, The mean of class 2 C_2 is:

$$\mu_2 = \begin{bmatrix} 8.4 & 7.6 \end{bmatrix}$$

Our goal is to find the covariance of class 1 C_1. We can do this by substituting μ_1 in eq. 3.2.

$$S_1 = \sum_{x \in C_1} (x - \mu_1)(x - \mu_1)^T$$

$$x_1 - \mu_1 = \begin{bmatrix} 1 & -1 & -1 & 0 & 1 \\ -2.6 & 0.4 & -0.6 & 2.4 & 0.4 \end{bmatrix}$$

Now for each x, we are going to calculate $(x - \mu_1)(x - \mu_1)^T$. So, we will have 5 such matrices.

$$\begin{bmatrix} 1 \\ -2.6 \end{bmatrix} \begin{bmatrix} 1 & -2.6 \end{bmatrix} = \begin{bmatrix} 1 & -2.6 \\ -2.6 & 6.76 \end{bmatrix}$$

$$\begin{bmatrix} -1 \\ 0.4 \end{bmatrix} \begin{bmatrix} -1 & 0.4 \end{bmatrix} = \begin{bmatrix} 1 & -0.4 \\ -0.4 & 0.16 \end{bmatrix}$$

$$\begin{bmatrix} -1 \\ -0.6 \end{bmatrix} \begin{bmatrix} -1 & -0.6 \end{bmatrix} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 0.36 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 2.4 \end{bmatrix} \begin{bmatrix} 0 & 2.4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 5.76 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 0.4 \end{bmatrix} \begin{bmatrix} 1 & 0.4 \end{bmatrix} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 0.16 \end{bmatrix}$$

Adding the above five matrices, and taking average, we get covariance matrix S_1.

$$S_1 = \begin{bmatrix} 0.8 & -0.4 \\ -0.4 & 2.6 \end{bmatrix}$$

Similarly, for class 2, the covariance matrix is given by:

$$S_2 = \begin{bmatrix} 1.84 & -0.04 \\ -0.04 & 2.64 \end{bmatrix} \text{ where } \mu_2 = \begin{bmatrix} 8.4 & 7.6 \end{bmatrix}$$

So, the within class scatter matrix is:-

$$S_W = S_1 + S_2$$
$$S_W = \begin{bmatrix} 2.64 & -0.44 \\ -0.44 & 5.24 \end{bmatrix}$$

Step 2: Compute between-class scatter matrix (S_B) :

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{3.3}$$

$S_B =$

$$\begin{bmatrix} -5.4 \\ -0.4 \end{bmatrix} \begin{bmatrix} -5.4 & -4 \end{bmatrix} = \begin{bmatrix} 29.16 & 21.6 \\ 21.6 & 16.00 \end{bmatrix}$$

Step 3: Find the best LDA eigenvector

Similar to Principal Component Analysis, we find this using eigenvalues that having the largest eigenvalue.

$$S_W^{-1} S_B V = \lambda V \tag{3.4}$$

$$\left| S_W^{-1} S_B - \lambda I \right| = 0$$
$$\begin{vmatrix} 11.89 - \lambda & 8.81 \\ 5.08 & 3.76 - \lambda \end{vmatrix} = 0$$

Solving λ, we get:

$$\lambda = 0 \text{ or } \lambda = 15.65$$

From above we take the largest eigenvalue which is $\lambda = 15.65$ and substitute it in Eq.3.4, we get:

$$\begin{bmatrix} 11.89 & 8.81 \\ 5.08 & 3.76 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix} = 15.65 \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

we get

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0.39 \end{bmatrix}$$

Step 4: Dimension Reduction

$$y = w^T x \tag{3.5}$$

where w is the eigenvector, and x is the input data sample.

$$\underbrace{\begin{bmatrix} 1 & -2.6 \\ -1 & 0.4 \\ -1 & -0.6 \\ 0 & 2.4 \\ 1 & 0.4 \\ 0.6 & 2.4 \\ -2.4 & 1.6 \\ 1.4 & -2.4 \\ \ldots & \ldots \\ \ldots & \ldots \\ \ldots & \ldots \end{bmatrix}}_{2-D} \begin{bmatrix} 0.91 & 0.39 \end{bmatrix} = \underbrace{\begin{bmatrix} \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \\ \ldots\ldots \end{bmatrix}}_{1-D}$$

3.9 Problems with LDA:

Despite the robustness of LDA and its common usage in data reduction techniques, there are two main difficulties:

1. **Small Sample Size (SSS):** or which called 'under sampled problem' is can be a serious problem with LDA. The problem appears as a result of the high-dimensional pattern classification tasks. Moreover, the problem happens when there is a low number of samples for each class when it compared with the dimensionality of the sample space. [25] [26]

   The problem comes when SW matrix is singular i.e. (it is a square matrix with no inverse, with zero determinant) which means that all columns and rows are sometimes not independent.

   So, when the dimensions are much higher than the number of samples in the data matrix, LDA fails to find the lower dimensional space.[27]

   Note: the SSS problem always appear in face recognition applications, since each image consists of 100X100 pixels. This represents high-dimensional features which causes the singularity problem. I.e. it requires a very large sample.

2. **Linearity Problem:** it is also known that LDA tries to find 'linear transformation' that discriminates between different classes. However, LDA fails when the classes are non-linearly separable. In this case, LDA can't switch to a lower dimensional space because the discriminatory information can't be found in the means of the classes, this happened when the means of the classes are equal or approximately equal. [28]

Chapter 4

Comparing Principal Component Analysis (PCA) with Linear Discriminant Analysis (LDA) as Dimensionality Reduction Techniques.

In this chapter, we will focus on the practical implementation of the two popular dimensionality reduction techniques PCA and LDA and evaluate empirically both techniques to decide which is more efficient in reducing dimensions and under what conditions.

4.1 Introduction

In our study, the criterias that we'll use to evaluate both methods and to decide which method is better are:

1. The number of components used.
2. The variance retained and the cumulative variance for the components used.

Another method can be used to decide which is better is the use of a classification table. However, we will not use it because it is outside the scope of our thesis.

4.2 Real datasets criteria:

In order to achieve the main goal of this research which is "comparing the two dimensionality reduction techniques PCA and LDA":

1. Four datasets will be used to compare the two methods.
2. These datasets are real life data: The Frog Dataset, The Female Empowerment, The Corruption Dataset and Pizza Dataset.
3. The datasets are interesting in their own rights, so we will describe each one of them in details:

4.2.1 Frog Dataset:

This dataset is collected by the Department of Biology at James Madison University. The data classifies different types of frogs according to biological measurements of: eye, arm, head, the distance between nostrils and mandible, and the decomposition of the lower teeth where measurements in the unit centimeter. There are 251 frogs' observations on 17 different properties or variables of this dataset. This is retrieved by the str() function in R which shows the structure of the dataset where the variables are numeric/quantitative variables.
The classification is decided based on the frog Clade designation which is based on DNA sequence data.

Character abbreviations for frog data variables:
ED - Eye diameter
EN - Eye-nostril distance
ES - Eye-Snout distance
FEL - Femur Length
FOL - Foot Length
HL - Head Length
HW - Head Width
IN - Internarial distance

IO - Interorbital distance

LAL – Lower arm length

MN - Mandible-nostril distance

PAL - Palmer Length

SVL - Snout-Vent Length

TBL - Tibia Length

UEW - Upper Eyelid Width

OL - Odontoid Length (measured from the bottom of the mandible to the top of the odontoid)

MD - Mandible depth (the thickness of the mandible at the base of the odontoid. This is used to calculate the actual length of the odontoid by subtracting MD from OL).

4.2.2 Female Empowerment:

This Data is from UNDP and CIA sites, which published in 2004. Its aim is to rank 177 of countries around the world if they empower their women or not based on 18 variables.

The classification is decided based on Gender Empowerment Measure (GEM). The countries are classified to 7 levels from A to G, where A is a country the is extremely female empowered whereas the G is a country that females are rarely empowered.

Character abbreviations for female empowrement variables:

SR - sex ration at birth male to female births

EYS - expected years of schooling for females in years.

PSE - population with at least some secondary education (female ages 25 and older).

ABR - adolescent birth rate (births per 1000 women - ages between 15 and 19)

EGNI - Estimated gross national income per capita

FSE - Female share of employment in senior and middle management.

CHI - coefficient of human inequality

HDI - human development index

IE - inequality in education.

II - inequality in income

IAE - inequality adjusted in education.

LFPR - labor force participation rate (ages 15 and older)

LE - life expectancy at birth for female in years.

SENA - share of employment in non-agriculture.

SSP - share of seats in parliament (held by women)

SRF - suicide rate for females.

TUR - total unemployment rate (female to male ratio).

YER - youth unemployment rate (female to male ratio).

4.2.3 Corruption:

The data are from the Enterprise Surveys organization (www.enterprisesurveys.org).

Corruption by public officials may present a major administrative and financial burden on firms. Corruption creates an unfavorable business environment by undermining the operational efficiency of firms and raising the costs and risks associated with doing business.

The classification is based on the Corruption Perceptions Index (CPI) which is an index published annually by Berlin-based Transparency International since 1995 which ranks countries ”by their perceived levels of public sector corruption, as determined by expert assessments and opinion surveys and the data we use is from the 2017 report.

The countries are classified to 10 levels according the CPI and as the CPI increases the country considered to has less corruption.

Character abbreviations for corruption dataset variables:

CI - Corruption index

BI - Bribery incidence (percent of firms experiencing at least one bribe payment request)

BD - Bribery depth (of public transactions where a gift or informal payment was re-

quested)

PoF.Tax - Percent of firms expected to give gifts in meetings with tax officials

PoF.Gov - Percent of firms expected to give gifts to secure government contract

VoG.Gov - Value of gift expected to secure a government contract (of contract value)

PoF.Lic - Percent of firms expected to give gifts to get an operating license

PoF.Import - Percent of firms expected to give gifts to get an import license

PoF.const - Percent of firms expected to give gifts to get a construction permit

PoF.Elec - Percent of firms expected to give gifts to get an electrical connection

PoF.Water - Percent of firms expected to give gifts to get a water connection

PoF.GPO - Percent of firms expected to give gifts to public officials "to get things done"

PoF.Constraint - Percent of firms identifying corruption as a major constraint

PoF.Court - Percent of firms identifying the courts system as a major constraint

4.2.4 Pizza:

Who likes pizza? I mean, there are so many things to like. Let's take a closer look! The data set pizza contains measurements that capture the kind of things that make a pizza tasty. Can you determine which pizza brand works best for you and explain why?

This data is a subset data of the USDA National Nutrient Database. It has 300 observations with 9 variables.

The variables in the data set are:

id - Sample analysed

mois - Amount of water per 100 grams in the sample

prot - Amount of protein per 100 grams in the sample

fat - Amount of fat per 100 grams in the sample

ash - Amount of ash per 100 grams in the sample

sodium - Amount of sodium per 100 grams in the sample

carb - Amount of carbohydrates per 100 grams in the sample

cal - Amount of calories per 100 grams in the sample

4.3 Testing Assumptions:

Let us remember those assumptions that were discussed in the previous chapters.

4.3.1 Normality assumption:

The PCA and LDA require that the data variables have a normal distribution or almost normal distribution. Recall that the normality can be checked graphically by density plot, Q-Q plot or histogram, however a formal test is needed which are significant methods such as (Kolmogorov-Smirnov normality test K-S, Shaprio-Wilk's test and Royston's Multivariate test).

Let us use for simplicity Shapiro-Wilk test.

Let us talk in short about these tests:

(I) **Graphical Checks of Normality:**

(a) **Density plot:**

which represents the data in away similar to the histogram. It allows to analyze the spread and the shape of the distribution so a visual judgment about the distribution to be normal or not (i.e. it is bell shaped) can be done.

Fig. 4.1 represents the density plot of first variable in the Frog Dataset, which is the eye diameter. We can see that the curve is too close to the normal distribution.

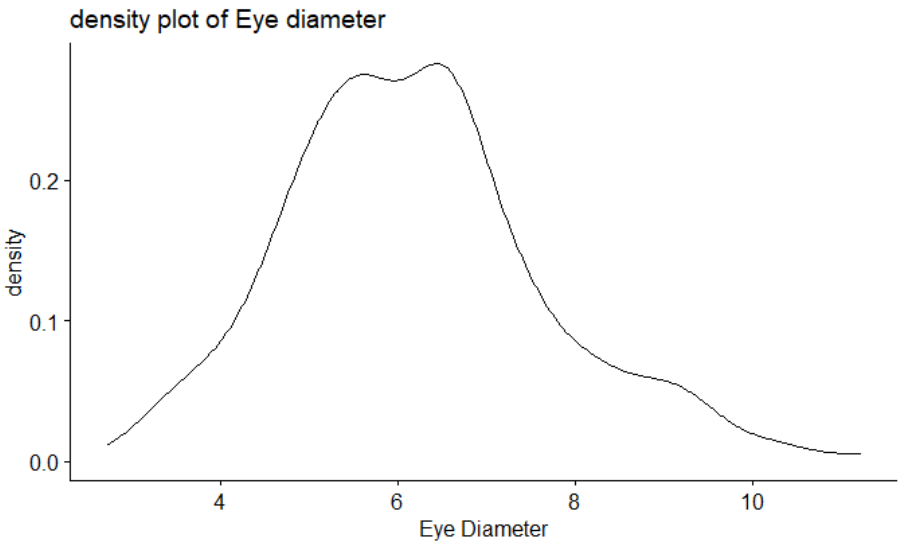


Figure 4.1: Density plot for Eye-Diameter variable in Frog dataset

The following Fig. 4.2 represents the density plot for the variable "Expected years of schooling for females" in the female empowerment dataset, and as we can see it is very close to be normal.

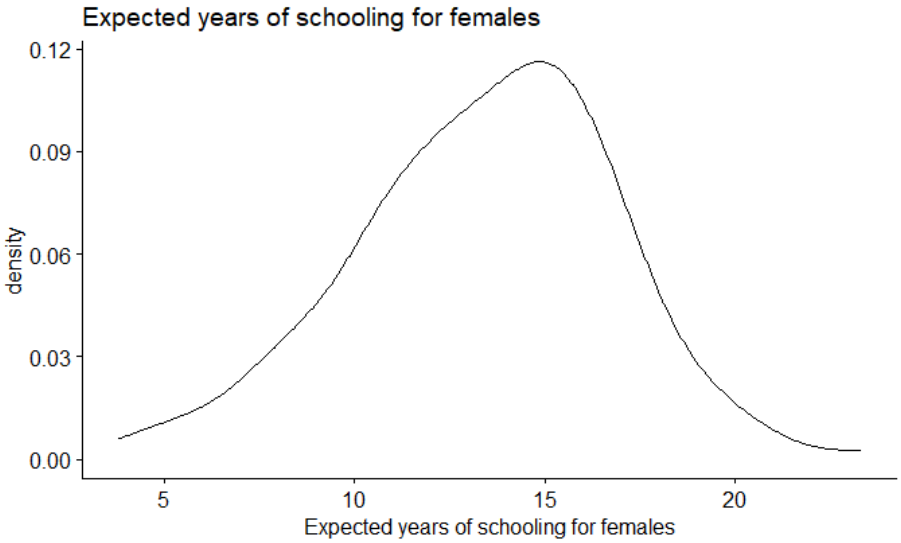


Figure 4.2: Density plot for Expected years of schooling for Females variable

The below Fig. 4.3 shows the density plot for the variable "Bribery Incidence" in the corruption dataset, and as we can notice it is somewhat close to normal distribution.

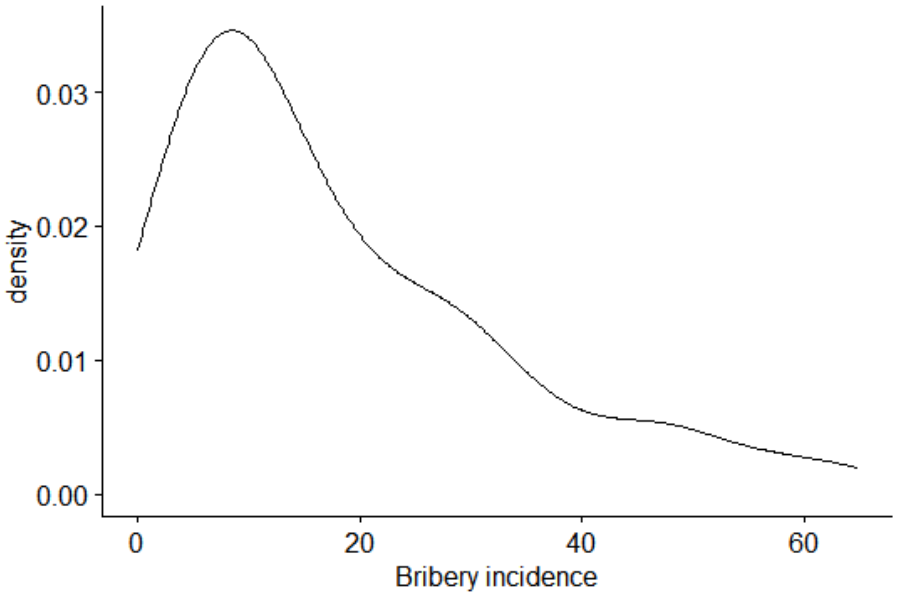


Figure 4.3: Density plot for bribery incidence variable in corruption dataset

The final dataset is the Pizza, and the next Fig. 4.4 shows the density plot for one of its variables which is (cal);the amount of calories per 100 grams in the sample.

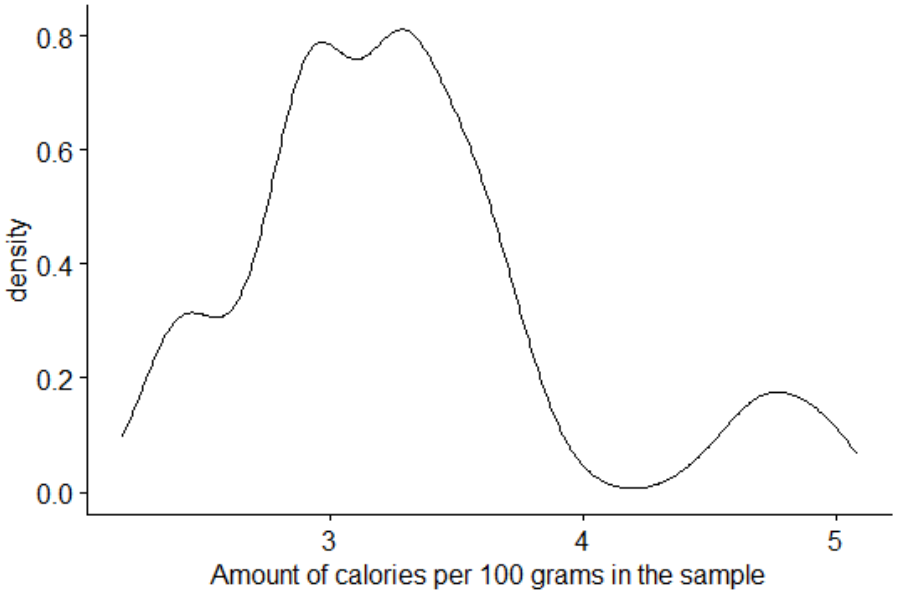


Figure 4.4: Density plot for amount of calories variable in Pizza dataset

(b) **QQ plot (Quantile-Quantile plot):**

which provides the graph of correlation between a given data set and the normal distribution where each observation is represented as a dot. When these dots form together a straight line we can say the data are normally distributed.

Fig. 4.5 shows the QQ-plot for the frog dataset variables where most of the dots

falls on the normality line which implies that the variables are normal.

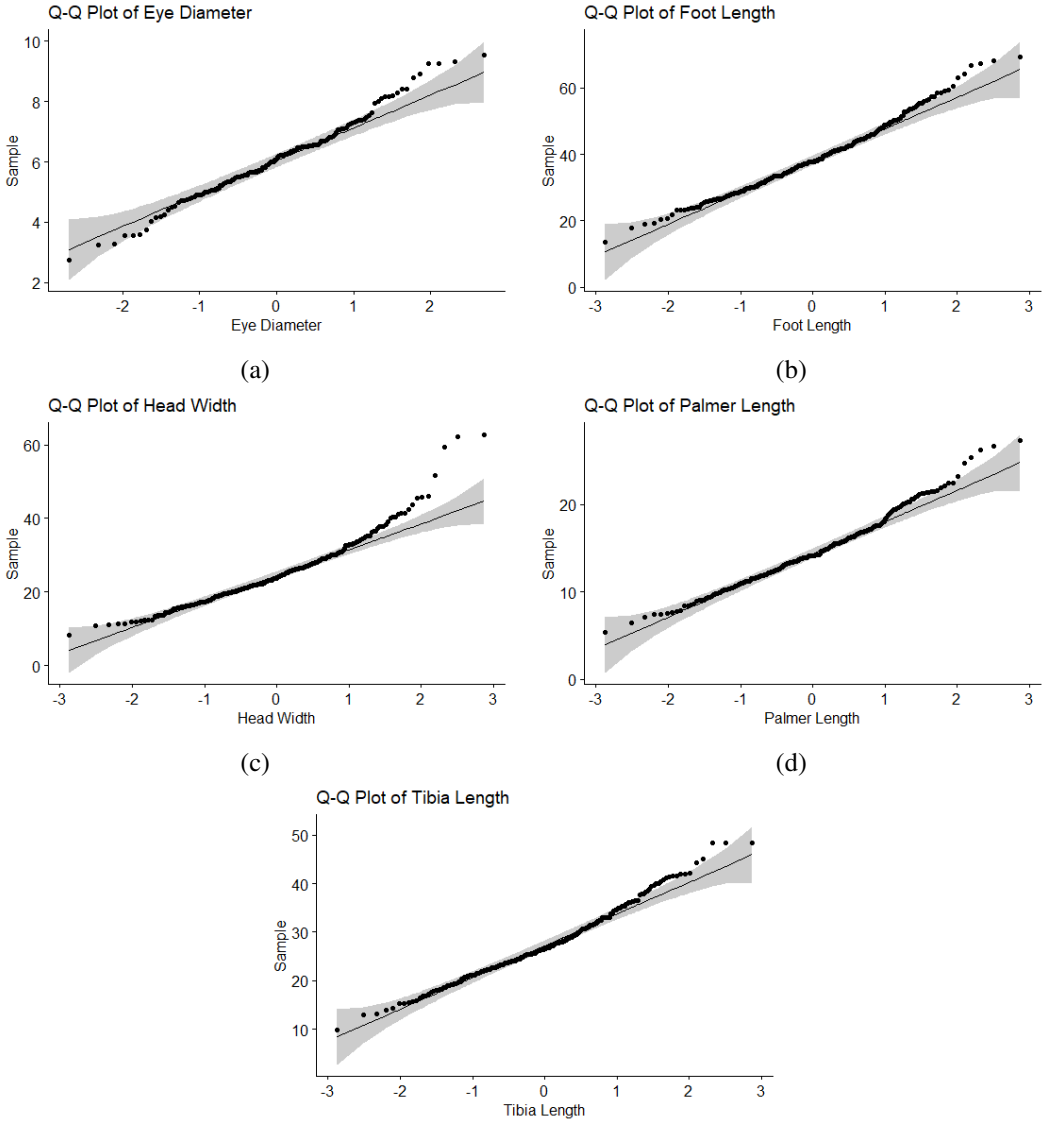


Figure 4.5: (a) QQ-plot of eye diameter variable of frog data set (b) QQ-plot of foot length variable of frog data set (c) QQ-plot of head width variable(d) QQ-plot of Palmer length variable (e) QQ-plot of Tibia length variable

The following Fig.4.6 shows the QQ-plot for the Female Empowerment dataset variables where most of the dots falls on the normality line which implies that the variables are normal.

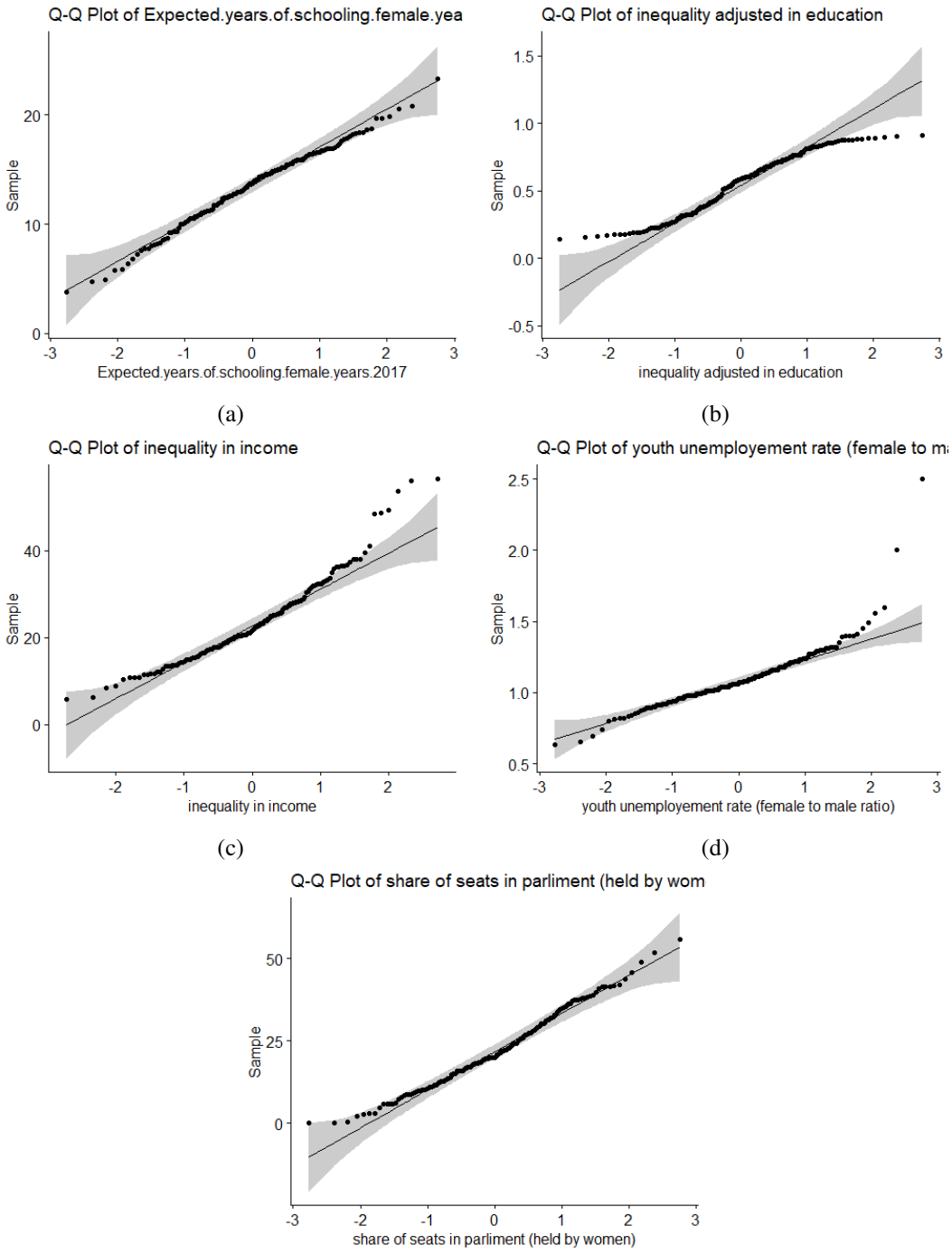


Figure 4.6: (a) QQ-plot of expected years of schooling variable of female empowerment dataset (b) QQ-plot of inequality adjusted in education variable of female empowerment dataset (c) QQ-plot of Inequality in Income variable of female empowerment data set (d) QQ-plot of youth employment rate variable of female empowerment data set (e) QQ-plot of share of seats in parliament (held by women) variable of female empowerment data set

The below Fig. 4.7 shows the QQ-plot for the corruption dataset variables where most of the dots falls on the normality line which implies that the variables are approximately normal.

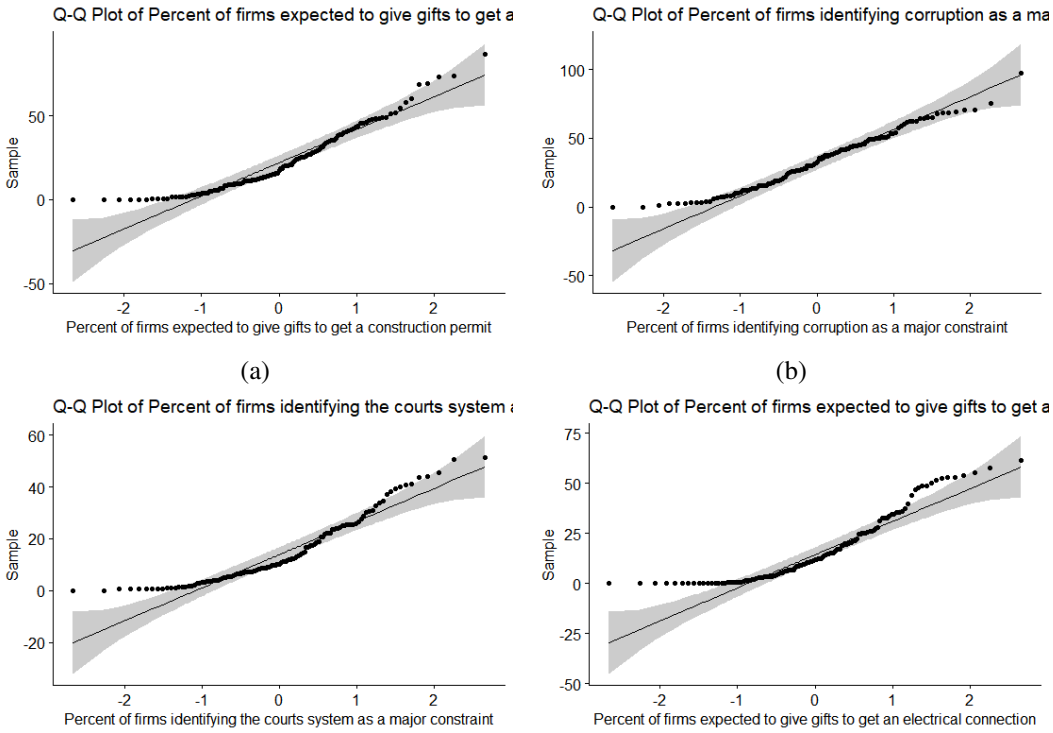


Figure 4.7: (a) QQ-plot of Percent of firms expected to give gifts to get a construction permit variable of Corruption data set (b) QQ-plot of Percent of firms identifying corruption as a major constraint variable of Corruption data set (c) QQ-plot of Percent of firms identifying the courts system as a major constraint variable of corruption data set (d) QQ-plot of Percent of firms expected to give gifts to get an electrical connection variable of Corruption data set

The following Fig. 4.8 shows the QQ-plot for the Pizza dataset variables where a lot of the dots falls on the normality line which implies that the variables are approximately normal. There are a lot of outliers, so that as we did for all datasets we'll combine the visual Q-Q test with the shapiro-wilk test to check the normality.

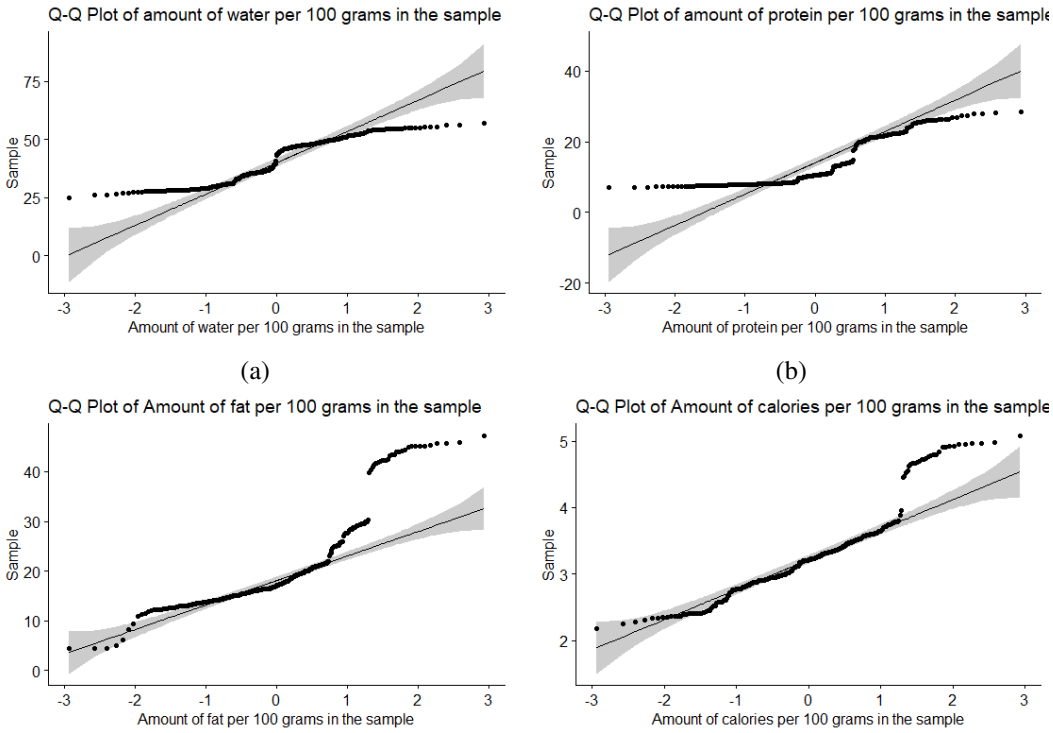


Figure 4.8: (a) QQ-plot of amount of water per 100 grams in the sample variable of Pizza dataset (b) QQ-plot of amount of protein per 100 grams in the sample variable of Pizza dataset (c) QQ-plot of amount of fat per 100 grams in the sample variable of Pizza dataset (d) QQ-plot of Percent of amount of calories per 100 grams in the sample variable of Pizza dataset

As we can see from the QQ-plots in the above figures, there are some outliers in the datasets, and they are summarized in Table 4.1:

Table 4.1: Summary of outliers check in all datasets

Dataset	**Variables with outliers**	**overall present of outliers or not(Low, Med, Hig)**
Frog	HW PAL	Low
Female Empowerment	II IAE YER	Med
Corruption	Most variables	High
Pizza	FAT CAL Prot	Med

(II) **Formal Test of Normality:**

which is more reliable than the mentioned visual inspection, here a significance test comparing the sample distribution to a normal one is made. Which help in deciding whether the data are normal or not.

Royston's MVN test

It is a multivariate normality test, used to test normality for multivariate data. We used the MVN function from MVN package in R to test multivariate normality by setting the argument "multivariatetest="royston".

The difference between Royston's MVN test and the Shapiro-Wilk test that the Royston is used for the multivariable dataset whereas the Shapiro-Wilk's test always used for the univariate datasets.

Shapiro-Wilk's test:

Which is a widely common test of normality. It was discovered in 1965 by Samuel Sanford Shapiro and Martin Wilk. It is focus on the correlation between the data and the corresponding normal scores.[54]

Note: Shapiro-Wilk test is the one which will be used to evaluate normality in our datasets.

The effect of the normality or lack of normality will be studied and investigated. So we have two hypotheses:

(a) **Null Hypothesis:** The variable follow normal distribution (i.e. the distribution is normal).

(b) **Alternative Hypothesis:** The distribution is not normal.

In general, we hope to achieve the null hypothesis that is when the P-value is > 0.05, we reject the alternative hypothesis and accept the null hypothesis, which implies that the distribution is normal. However, if the P-value < 0.05, the distribution is not normal.

We use the mvn function from MVN package in R to use the Shaprio-Wilk's test by setting the argument unnivariatetest="SW" in mvn function.

The Tables below Table 4.2 and Table 4.3 and Table 4.4 and Table 4.5 shows the results of applying Shapiro-Wilk test for the entire datasets which have all observations.

Table 4.2: Applying Shapiro-Wilk's test for Frog Dataset with all observations

Dataset	Variables	P-Value of Shapiro-Wilk test
Frog	ED	0.002500000000
	EN	0.000000002945
	ES	0.000000054720
	FEL	0.000072870000
	FOL	0.004545000000
	HL	0.000000000069
	HW	0.000000001217
	IN	0.000357000000
	IO	0.000000002128
	LAL	0.000005359000
	MN	0.000000000058
	PAL	0.001806000000
	SVL	0.000027240000
	TBL	0.002430000000
	UEW	0.000003585000
	OL	0.000000000026
	MD	0.000000000009

Table 4.3: Applying Shapiro-Wilk's test for Female Empowerment Dataset with all observations

Dataset	Variables	P-Value of Shapiro-Wilk test
Female Empowerment	SR	0.000000000001
	EYS	0.342500000000
	PSE	0.000000319700
	ABR	0.000000000486
	EGNI	0.000000000000
	FSE	0.000817400000
	CHI	0.000020980000
	HDI	0.000028300000
	IE	0.000000002813
	II	0.000005108000
	IAE	0.000008671000
	LFPR	0.001119000000
	LE	0.000000792300
	SENA	0.000000000140
	SSP	0.036390000000
	SRF	0.000000075460
	TUR	0.000000000000
	YER	0.000000000002

Table 4.4: Applying Shapiro-Wilk's test for Corruption Dataset with all observations

Dataset	Variables	P-Value of Shapiro-Wilk test
Corruption	CI	0.001805000000
	BI	0.000000015900
	BD	0.000000000621
	PoF.Tax	0.000000000380
	PoF.Gov	0.000012430000
	VoG.Gov	0.000000000000
	PoF.Lic	0.000000000607
	PoF.Import	0.000000000006
	PoF.const	0.000000588700
	PoF.Elec	0.000000004123
	PoF.Water	0.000000000122
	PoF.GPO	0.000000000313
	PoF.Constraint	0.007449000000
	PoF.Court	0.000000041040

Table 4.5: Applying Shapiro-Wilk's test for Pizza Dataset with all observations

Dataset	Variables	P-Value of Shapiro-Wilk test
Pizza	mois	0.00000000000035830
	prot	0.00000000000000022
	fat	0.00000000000000022
	ash	0.00000000000000221
	sodium	0.00000000000000022
	carb	0.00000000000000022
	cal	0.0.00000000000064340

And if we use the mvn function to test normality (univariate) with Shapiro-Wilk's test we get the following results:

According to the table above, the p-value of Shapiro-Wilk test for the frog dataset variables is less than 0.05, so it rejects the null hypothesis that the data is normal, however because our sample size is large > 50, the Shapiro-Wilk test is very sensitive even to a minor deviation from normality, so we will depend on the QQ-plot approach to test normality.

According to QQ-plot as illustrated in the above figures for frog dataset, we see that they are normal.

And according the MVN test, which uses the univariate normality test "Shaprio-Wilk" and a multivariate normality test "Royston" with a small subset of data (approximately 50 observation) we get the following results shown in Table 4.6 for multivariate Royston test and as shown in Table 4.7

Table 4.6: MVN Royston of multivariate normality test for Frog Dataset.

Test	**H**	**p-value**	**MVN**
Royston	6.483034	0.08916893	YES

Table 4.7: MVN Shaprio-Wilk univariate normality test for Frog Dataset.

Test	**Variable**	**Statistic**	**p-value**	**Normality**
Shapiro-Wilk	ED	0.9493	0.0321	NO
Shapiro-Wilk	EN	0.9677	0.1871	YES
Shapiro-Wilk	ES	0.9833	0.6952	YES
Shapiro-Wilk	FEL	0.9888	0.9130	YES
Shapiro-Wilk	FOL	0.9655	0.1506	YES
Shapiro-Wilk	HL	0.9789	0.5046	YES
Shapiro-Wilk	HW	0.9842	0.7363	YES
Shapiro-Wilk	IN	0.9861	0.8181	YES
Shapiro-Wilk	IO	0.9793	0.5242	YES
Shapiro-Wilk	LAL	0.9709	0.2507	YES
Shapiro-Wilk	MN	0.9754	0.3770	YES
Shapiro-Wilk	PAL	0.9766	0.4201	YES
Shapiro-Wilk	SVL	0.9526	0.0438	NO
Shapiro-Wilk	TBL	0.9449	0.0212	NO
Shapiro-Wilk	UEW	0.9844	0.7444	YES
Shapiro-Wilk	OL	0.9219	0.0028	NO
Shapiro-Wilk	MD	0.9734	0.3170	YES

Now, let we focus on the Table 4.3 above for the p-value of Shapiro-Wilk test for the second data set “Female Empowerment”,We can see that the p-value of Shapiro-Wilk test for all Female Empowerment variables) except Eye variable) is less than 0.05 which implies to reject the null hypothesis that the data is normal. Also here we can depend on the QQ-plot approach to test normality since the sample size in this data set is large (177 obs.) and the Shapiro-Wilk test is very sensitive even to a minor deviation from normality.

According to QQ-plot as illustrated in the above figures Fig. 4.6 for the Female Empowerment Dataset, we see that they are almost normal.

Note that for the EYS(Expected years of schooling for females) the p-value in the table was 0.342500000000 which is > 0.05 here the null hypotheses is not rejected which means the variable is normal. Now let’s apply the same MVN univariate and multivariate normality tests on the second dataset which is the female empowerment. The results of the MVN Royston test is shown in Table 4.8 and the results of Shapiro Wilk is shown in Table 4.9.

Table 4.8: MVN Royston of multivariate normality test for Female Empowerment dataset.

Test	**H**	**p-value**	**MVN**
Royston	5.333725	0.08630606	YES

Table 4.9: MVN Shaprio-Wilk univariate normality test for Female Empowerment Dataset.

Test	**Variable**	**Statistic**	**p-value**	**Normality**
Shapiro-Wilk	SR	0.8293	0.0155	NO
Shapiro-Wilk	EYS	0.9133	0.2034	YES
Shapiro-Wilk	PSE	0.9052	0.1578	YES
Shapiro-Wilk	ABR	0.8247	0.0136	NO
Shapiro-Wilk	EGNI	0.8979	0.1251	YES
Shapiro-Wilk	FSE	0.9654	0.8337	YES
Shapiro-Wilk	CHI	0.8469	0.0260	NO
Shapiro-Wilk	HDI	0.9232	0.2766	YES
Shapiro-Wilk	IE	0.9093	0.1795	YES
Shapiro-Wilk	II	0.9212	0.2603	YES
Shapiro-Wilk	IAE	0.9295	0.3361	YES
Shapiro-Wilk	LFPR	0.9260	0.3022	YES
Shapiro-Wilk	LE	0.9539	0.6582	YES
Shapiro-Wilk	SENA	0.8502	$<$0.001	NO
Shapiro-Wilk	SSP	0.9489	0.0308	NO
Shapiro-Wilk	SRF	0.9315	0.0063	NO
Shapiro-Wilk	TUR	0.4857	$<$0.001	NO
Shapiro-Wilk	YER	0.8674	$<$0.001	NO

The Shapiro-Wilk' test for the third dataset (Corruption) is also is less than 0.05, so like other datasets we will depend on the QQ-plot to check normality, and according the QQ-plot figure of the corruption dataset, it shows that the variables are nearly normal.

And also we checked the univariate normality with a small size dataset of female by the mvn function with Shapiro-Wilk's test and get the following results shown in Table 4.11

Table 4.10: MVN Royston of multivariate normality test for Corruption Dataset.

Test	**H**	**p-value**	**MVN**
Royston	12.49348	0.01313076	NO

Table 4.11: MVN Shaprio-Wilk univariate normality test for Corruption Dataset.

Test	Variable	Statistic	p-value	Normality
Shapiro-Wilk	CI	0.9632	0.8313	YES
Shapiro-Wilk	BI	0.8207	0.0351	NO
Shapiro-Wilk	BD	0.8427	0.0619	YES
Shapiro-Wilk	PoF.Tax	0.8675	0.1157	YES
Shapiro-Wilk	PoF.Gov	0.9317	0.4976	YES
Shapiro-Wilk	VoG.Gov	0.7047	0.0016	NO
Shapiro-Wilk	PoF.LIC	0.9302	0.4829	YES
Shapiro-Wilk	PoF.Import	0.8368	0.0532	YES
Shapiro-Wilk	PoF.const	0.9525	0.7172	YES
Shapiro-Wilk	PoF.Elec	0.7920	0.0165	NO
Shapiro-Wilk	PoF.Water	0.8214	0.0357	NO
Shapiro-Wilk	PoF.GPO	0.8409	0.0591	YES
Shapiro-Wilk	PoF.Constraint	0.8624	0.1019	YES
Shapiro-Wilk	PoF.Court	0.7765	0.0110	NO

Now, let us apply the MVN functions: Royston multivariate test and shapiro-Wilk univariate test on the last dataset which is the Pizza. Table 4.12 below shows the results of applying Royston multivariate test on Pizaa dataset.

Table 4.12: MVN Royston of multivariate normality test for Pizza Dataset.

Test	H	p-value	MVN
Royston	21.08965	0.000118434	NO

And the next Table 4.13 shows the results of applying MVN Shapiro- Wilk's test on Pizaa dataset with a subset of the observations.

Table 4.13: MVN Shaprio-Wilk univariate normality test for Pizza Dataset.

Test	Variable	Statistic	p-value	Normality
Shapiro-Wilk	mois	0.959	0.081	YES
Shapiro-Wilk	prot	0.8229	0.001	NO
Shapiro-Wilk	fat	0.9293	0.0047	NO
Shapiro-Wilk	ash	0.9766	0.4061	YES
Shapiro-Wilk	sodium	0.9347	0.0076	NO
Shapiro-Wilk	carb	0.9464	0.0224	NO
Shapiro-Wilk	cal	0.9761	0.3896	YES

Finally, it is an important point to be mentioned that the normality test is sensitive to sample size. That is when the samples are very small, the normality tests mostly go well.

So it is necessary to gathering both the significance test with visual inspection which help to decide whether the distribution is normal or not.

Note: the normality assumption is must be checked for both techniques (LDA &PCA)

4.3.2 Equality/ Homogeneity of Covariance Matrix(Only needed for LDA):

Which is an important assumption in multivariate statistics. It is a multivariate version of the univariate assumption of Homogeneity of variance. This assumption can be tested using the Box M test which should be non-significant with its two hypotheses:
Null Hypothesis: The covariance matrices for groups are equal, i.e. $\sum 1 = \sum 2$.
Alternative Hypothesis: The covariance matrices for groups are not equal, i.e. $\sum 1 \neq \sum 2$.
Also here we hope to have the p-value > 0.05 at the 0.05 level of significance to reject the alternate hypothesis and accept the null hypothesis.
Note: this assumption (equality of covariance) is only required for the LDA.

After applying BoxM test on the Frog Data set, the result of p is 0.00000000000000022 which is less than 0.05 and so we reject the null hypothesis that assumes equality.

When we apply the BoxM test on the Female Empowerment Dataset, we found the p-value to be 0.00000000000000022, which is less than 0.05, and so we reject the null hypothesis that assumes equality.
When we apply the BoxM test on the Corruption and Pizza datasets, we find the p-value to be 0.00000000000000022 which is less than 0.05, so all the datasets reject the null hypothesis that assumes equality between groups.

4.4 Exploratory Data Analysis:

In statistics; 'exploring data, data screening and describing the main characteristics with important features' is an approach to analyzing and summarizing data sets. These steps are usually done as initial procedures before starting analysis process. Exploratory data analysis aims to explore the data to find out if there are any errors, such as missing in-

formation, correct alignment of the variables, outliers..., etc. After this is done, we can treat with the missing values and making transformations of variables as needed in order to checking assumptions and testing hypothesis for a certain model. Moreover, it gives a roughly an idea about how the correlation behaves between the variables. This can be done easily using "scatter plots" and correlations.

1. **Scatter plots:**

   Which is also known as 'scatter chart' or 'scatter graph' it is show us the trends between the explanatory numerical variables when the data are taken as a whole. It revels any patterns. Dots are using to represent the values for two different variables. So, we can easily observe the relationships between variables and identifying the correlation among them.

2. **Correlation matrix:**

   Now, let we apply the scatter plot on each data set to find the correlation matrix.

The following Fig. 4.9 represents the scatter plot for the frog dataset.

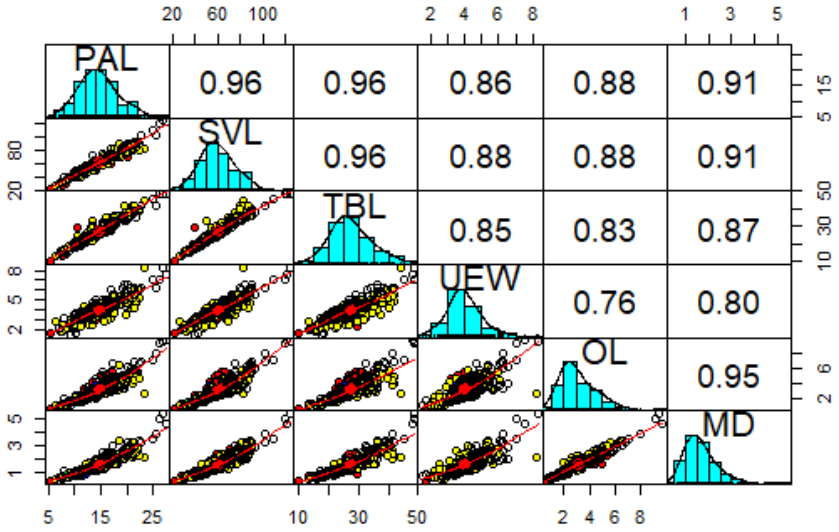


Figure 4.9: Scatter and correlation plot for the Frog Data set

Where the graph represents a matrix of scatter plots between the frog variables. In

general, the first impression when we look at the whole plot shows that there are strong positive linear relationships between the 17 variables.

The SVL variable has very strong positive linear relationships with all other variables. Also, PAL has strong positive linear relationships with all remaining variables. Moreover, the relationship between SVL and TBL is a very strong positive relationship. Also, the MD variable has strong linear relationships with other remaining variables, but the relationship between the MD and OL variables is the best. As we can see, The relationship between UEW and OL is the weakest but it still has good correlation. And the relationships between the UEW variable with all remaining are moderate linear relationships.

The next Fig. 4.10 represents the scatter plot for the female empowerment dataset.

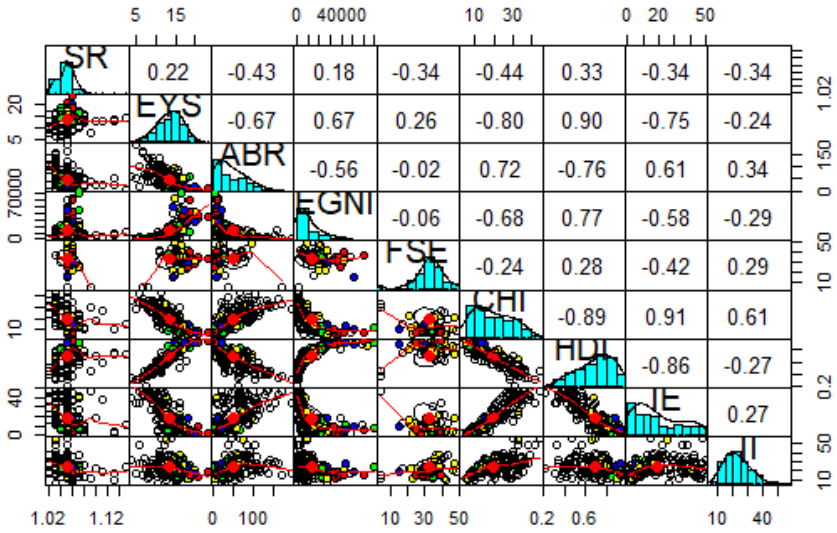


Figure 4.10: Scatter and correlation plot for the Female Empowerment Data set

Fig. 4.10 represents a matrix of scatter plots between the Female Empowerment variables. In general, the first impression when we look at the whole plot shows that there are moderate linear relationships between the most of the variables. The HDI variable has very strong positive linear relationships with EYS variable. There is a moderate strength linearity between the variables IE and ABR. However there is a weak linear relationship between the variables CHI and FSE. However, there is a negative strong linear relationship between HDI and CHI variables. And for the variables FSE and EGNI we can notice that

there is no linear relationship can be seen.

The following Fig. 4.11 shows the scatter plot and correlation for the corruption dataset.

and we can see there are a good linear relationship between all variables. There is a very positive correlation between BD and BI, and there is moderate negative correlation between PoF Gov and CI. There is a weak relationship between PoF LIC and VoG GOV.

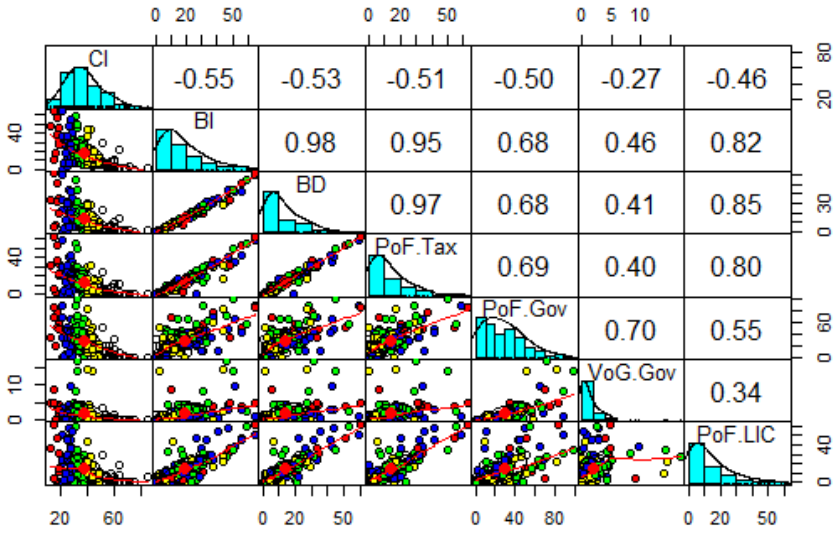


Figure 4.11: Scatter and correlation plot for the Corruption Data set

The following Fig. 4.12 show the scatter plot and correlation for the Pizza dataset.

We can see a very strong positive linear relationship between sodium and fat variables, and a very strong negative linear relationship between carb and ash variables.Some variable have a moderate linear relationship between each other like cal and ash and some other variable have no relationship between them like cal and carb variables.

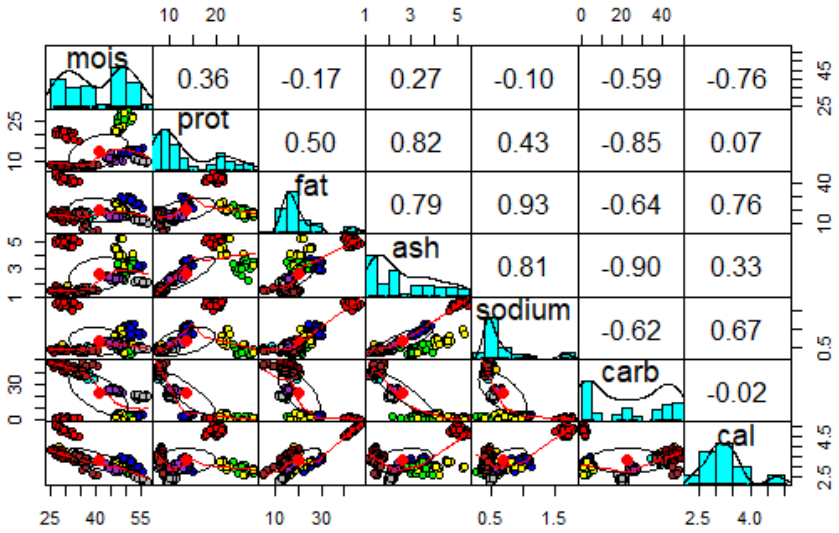


Figure 4.12: Scatter and correlation plot for the Pizza Data set

Overall, the correlations are strong in Frog and Pizza examples suggesting a PCA analysis.

4.5 Applying Principal Component Analysis:

We use the prcomp function in R to perform the PCA analysis on our data.

The PCA results of applying prcomp function on the frog data set is shown below in Table 4.14

Table 4.14: Principal components for Frog Dataset

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	3.9198	**0.9038**	**0.90380**
PC2	0.65303	**0.02509**	**0.92888**
PC3	0.4825	0.0137	0.94260
PC4	0.45557	0.01221	0.95479
PC5	0.4040	0.0096	0.96440
PC6	0.35659	0.00748	0.97187
PC7	0.32289	0.00613	0.97800
PC8	0.29822	0.00523	0.98324
PC9	0.24508	0.00353	0.98677
PC10	0.22924	0.00309	0.98986
PC11	0.20283	0.00242	0.99228
PC12	0.18128	0.00193	0.99421
PC13	0.16583	0.00162	0.99583
PC14	0.1596	0.0015	0.99730
PC15	0.15277	0.00137	0.99870
PC16	0.12431	0.00091	0.99961
PC17	0.08124	0.00039	1.00000

To specify the number of principal components that we must take to interpret the data, we can use different methods such as (Skree-plot, Kaiser's rule or Horn method).

1. Skree plot Method (Graphical tool): In the skree plot, we stop where the elbow is (that is the point after which the remaining eigenvalues decline in approximately linear fashion drop).

   The skree plot is a graphical tool that calls for a relative judgment of the amount of variance accounted for by the retained components.

2. Kaiser's rule: Retain only the PC with eigenvalues exceeding unity (assuming that we are dealing with standarized variables in the analysis). This rule reflects the common sense that any principal component should account for at least as much variations as any of the original variables X (because it is a measure of the common variance).

3. Horn's procedure: Use as a cutoff the eigenvalue from principal complements of random data with the same number of observations.

   (Generate random data with same p and same n then apply the PCA to the random data and plot the eigenvalues into skree plot along with the eigenvalues from the actual data, then we only keep the PCA that have eigenvalues that exceed the Horn's

cutoff), the rational behind this method is that PCA tends to capitalize of random error.

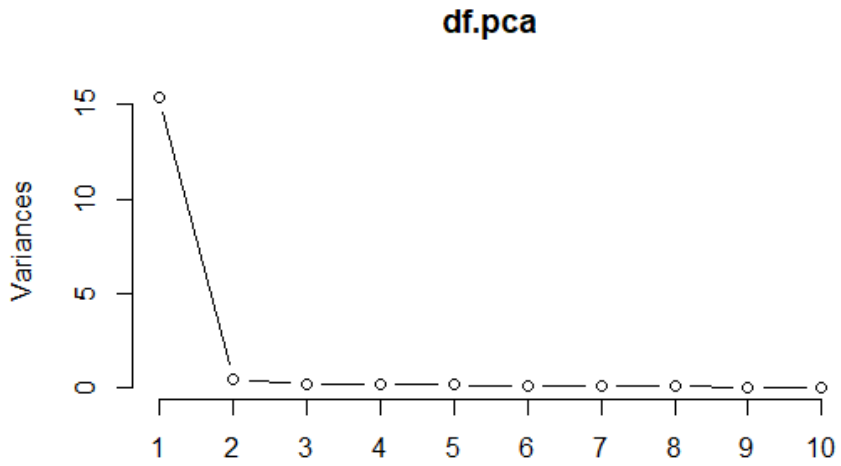


Figure 4.13: Skree plot of the Frog Data set

The above Fig. 4.13 shows the skree plot for the PCA of Frog Dataset, and as we can see there a very large decrease in variance from PC1 to PC2 and then to PC3, So we can keep the first two PCs.

Now, let's see the results of applying PCA on the Female Empowerment Dataset as shown in Table 4.15.

Table 4.15: Principal components for Female Empowerment data set

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	3.0092	**0.5031**	**0.50310**
PC2	1.55620	**0.1346**	**0.63760**
PC3	1.25482	0.08748	0.72510
PC4	1.08685	0.06562	0.79072
PC5	1.00512	0.05613	0.84685
PC6	0.88774	0.04378	0.89063
PC7	0.68620	0.02616	0.91679
PC8	0.5832	0.0189	0.93570
PC9	0.54878	0.01673	0.95242
PC10	0.48449	0.01304	0.96546
PC11	0.45451	0.01148	0.97693
PC12	0.39258	0.00856	0.98550
PC13	0.34801	0.00673	0.99222
PC14	0.25296	0.00355	0.99578
PC15	0.23065	0.00296	0.99873
PC16	0.12189	0.00083	0.99956
PC17	0.07845	0.00034	0.99990
PC18	0.04199	0.00010	1.00000

Fig. 4.14 shows the scree-plot for the PCA of the female empowerment dataset, and as we can see there a very large decrease in variance from PC1 to PC2 and then a minimal decrease, so we can keep the first two PCs.

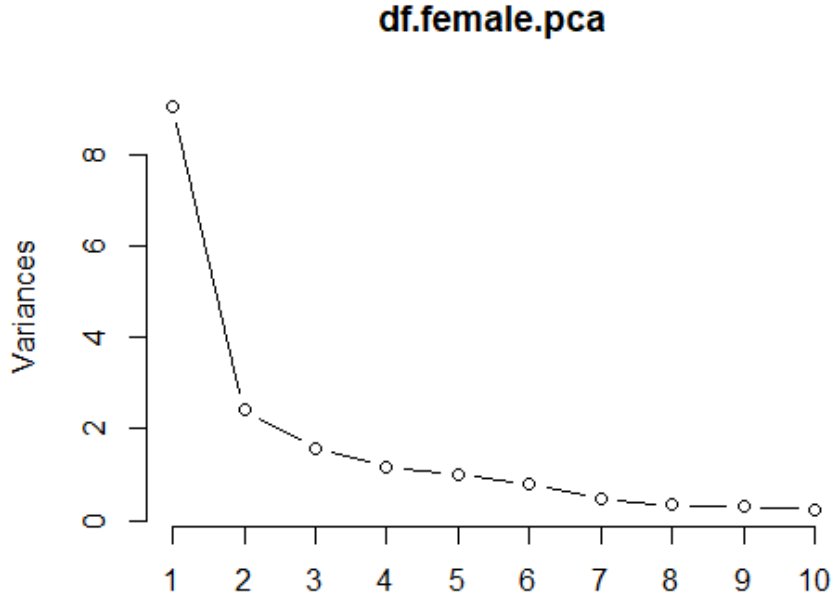


Figure 4.14: Skree plot of the Female Empowerment Data set

Next, the results of applying PCA on the Corruption Dataset is shown below in Table 4.16.

Table 4.16: Principal components for Corruption Data set

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	2.8293	**0.5718**	**0.5718**
PC2	1.3111	**0.1228**	**0.6945**
PC3	1.05941	**0.08017**	**0.77471**
PC4	0.77181	0.04255	0.81726
PC5	0.74675	0.03983	0.85709
PC6	0.64847	0.03004	0.88713
PC7	0.62057	0.02751	0.91464
PC8	0.56422	0.02274	0.93737
PC9	0.50932	0.01853	0.95590
PC10	0.47596	0.01618	0.97208
PC11	0.43374	0.01344	0.98552
PC12	0.40344	0.01163	0.99715
PC13	0.17346	0.00215	0.99930
PC14	0.09916	0.00070	1.00000

Fig. 4.15 represents the Skree plot for the PCA of Corruption Dataset, and as we can see there is a sharp decrease from PC1 to PC2 to PC3, so we can keep the first three PCs for this dataset.

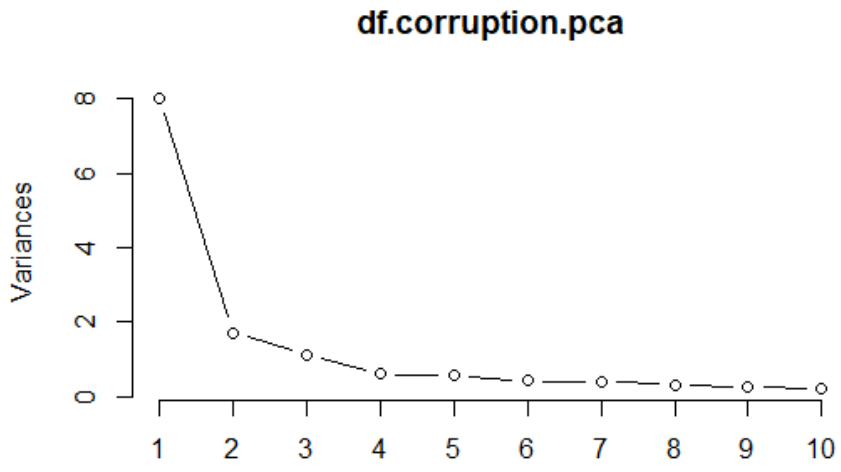


Figure 4.15: Skree plot of the Corruption Data set

Finally, the results of applying PCA on the Pizza Dataset is shown below in Table 4.17.

Table 4.17: Principal components for Pizza Data set

Principal Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	2.042	**0.596**	**0.596**
PC2	1.5134	**0.3272**	**0.9232**
PC3	0.64387	**0.05922**	**0.9824**
PC4	0.3085	0.0136	0.9960
PC5	0.16636	0.00395	0.99995
PC6	0.01837	0.00005	1.0000
PC7	0.003085	0.00000	1.0000

Fig. 4.16 represents the Skree plot for the PCA of Pizza dataset, and as we can see there is a sharp decrease from PC1 to PC2 and to PC3, so we can keep the rst three PCs for this dataset.

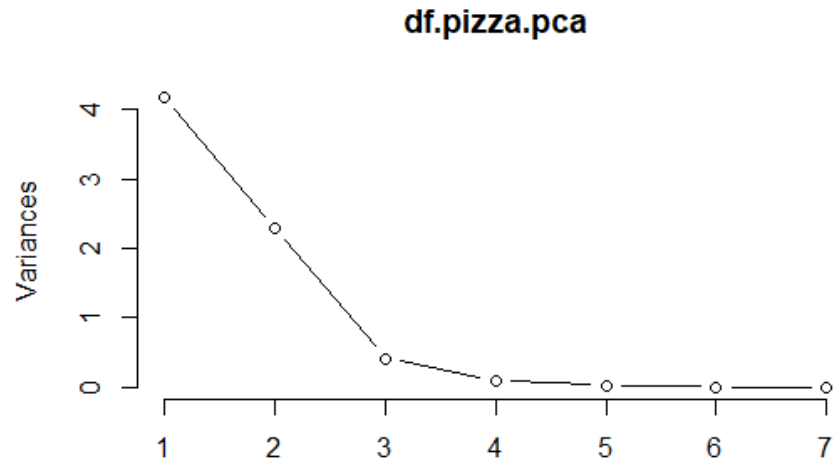


Figure 4.16: Skree plot of the Pizza Data set

4.5.1 Bi-plot for PCA

The most used method to visualize the PCA is the biplot which is done using the ggbiplot package in R.

PCA plot shows clusters of samples based on their similarity.

A loading plot shows how strongly each characteristic influences a principal component. Their project values on each PC show how much weight they have on that PC.

Another nice thing about loading plots are the angles between the vectors which tell us how characteristics correlate with one another.

When two vectors are close, forming a small angle, the two variables they represent are positively correlated.

If they meet each other at 90°, they are not likely to be correlated.

When they diverge and form a large angle (close to 180°), they are negative correlated. Fig. 4.17 shows bi-plot for the frog dataset PCA.

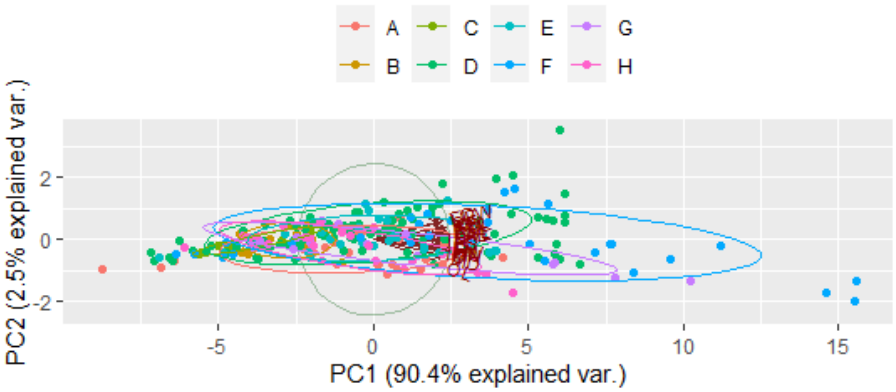


Figure 4.17: Bi-plot of Frog Dataset PCA

As we can see in Fig. 4.17, all of the variables are very correlated, and all of them are on the same direction, and they have very influence on PC1 which shows that PC1 can represents most of the data.

The below following Fig. 4.18 shows the bi-plot for the female empowerment dataset PCA.

We can see that there are a very good correlation between some of variables like IE and CHI or ABR and YER, and also there are a good negative correlation between EGNI and ABR. However, some of variables have a 90 degree angle between them which implies weak or no correlation between them such as SR and SENA.

The variables have a good projection values on both PC1 and PC2, so PC1 and PC2 share most of the data.

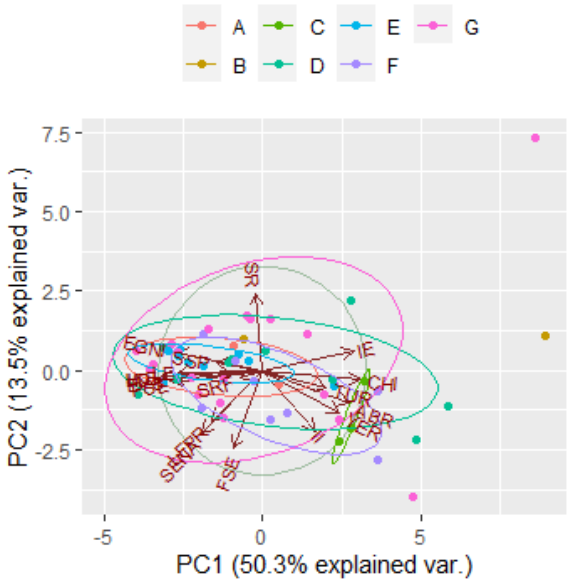


Figure 4.18: Bi-plot of Female Empowerment Dataset PCA

Fig. 4.19 shows the bi-plot for the Corruption Dataset. We can see that there are a very good correlation between most of variable (PoF.Gov, PoF.Elec, PoF.Import,...etc).However,some variables have no correlation with the variables as we can see there is a 90 degree angle between them and the other variables such as (PoF.Court, PoF.Constraint and CI), and this of course will have a bad effect on PCA results.

Most of variables projection are PC1, so PC1 will have most the data.

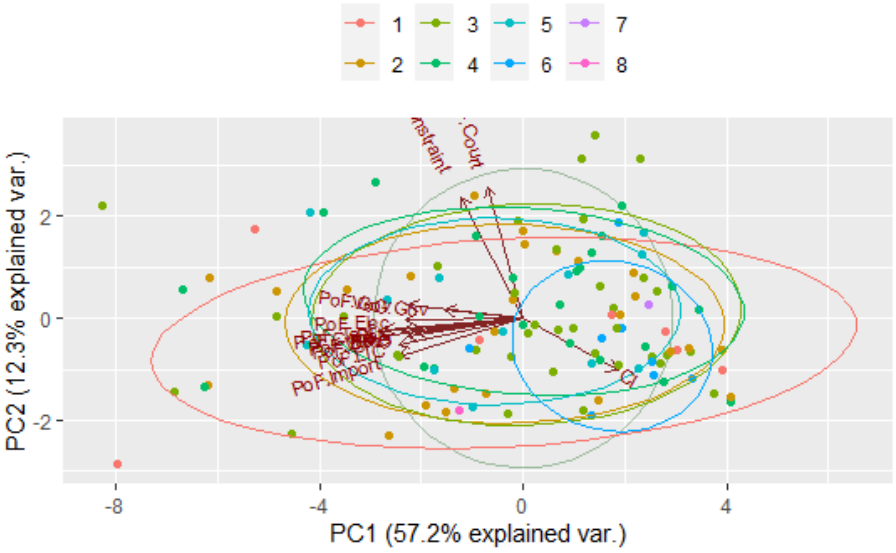


Figure 4.19: Bi-plot of Corruption Dataset PCA

The following Fig. 4.20 shows the bi-plot for the Pizza dataset. It appears most of variables have a good correlation with each other like Sodium and fat or ash and prot as there is a very small angle between them, and also there is a very good negative correlation between the variables carb and prot.However, some variables have an approxmately 90 degree between them such as cal and prot or sodium and mois which implies a weak correlation between them.

We can see that most variables projection are lies on PC1 and PC2, so this will make PC1 and PC2 will have the most data after reduction.

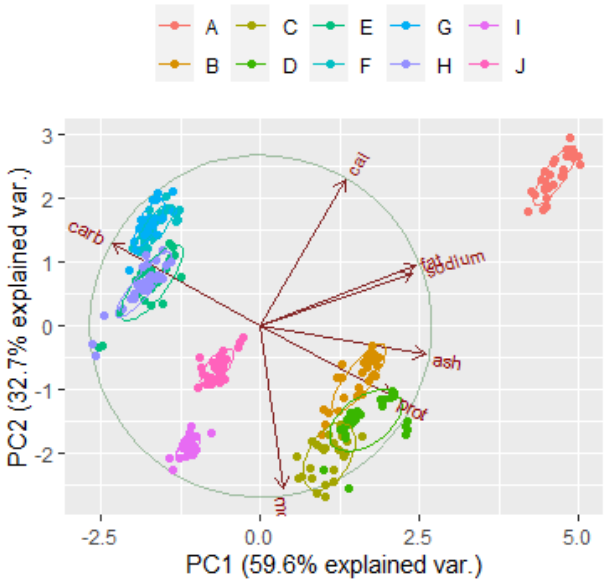


Figure 4.20: Bi-plot of Pizza Dataset PCA

4.6 Applying Linear Discriminant Analysis:

Tables below show linear discrimination coefficients for the dataset, which in turn measures the relative weight and importance of all variables in each data, such that the variable that possesses the absolute larger coefficient means that it has the heaviest of the variables and vice versa. This means that the groups are most affected by this variable.

The following Tables 4.18, 4.19, 4.20 and 4.21 show the results of applying LDA to Frog, Female Empowerment, Corruption, and Pizza data sets respectively.

Table 4.18: LDA results of Frog Dataset

Linear Discriminant	**Proportion of trace**
LD1	**0.4045**
LD2	**0.3098**
LD3	0.1114
LD4	0.0928
LD5	0.0338
LD6	0.0329
LD7	0.0147

Table 4.19: LDA results of Female Empowerment Dataset

Linear Discriminant	Proportion of trace
LD1	**0.5349**
LD2	**0.1858**
LD3	0.1011
LD4	0.0866
LD5	0.0674
LD6	0.0242

Table 4.20: LDA results of Corruption Dataset

Linear Discriminant	Proportion of trace
LD1	**0.9770**
LD2	0.0122
LD3	0.0051
LD4	0.0040
LD5	0.0013
LD6	0.0003

Table 4.21: LDA results of Pizza Dataset

Linear Discriminant	Proportion of trace
LD1	**0.7693**
LD2	**0.1391**
LD3	**0.0854**
LD4	0.0056
LD5	0.0006
LD6	0.0001
LD7	0.0000
LD8	0.0000

For the Frog Dataset LDA as shown in table 4.18, we can keep the First 2 LDAs to have approximately 0.7143 proportion of variance.

For the Female empowerment LDA as shown in table 4.19, we can also keep the first two LDAs to have approximately 0.7207 of the total variance.

For the Corruption LDA as shown in Table 4.20, the first LD represent 0.9770 of total variance, so we can just take this variable.

For the Pizza dataset, we can keep the first three LDs to have approximately 0.9938 of the total variance.

4.6.1 Bi-plot for LDA

An LDA (Linear Discriminant Analysis) biplot is designed to show how individuals and groups are different.

LDA biplots are read just like PCA biplots – with the projections of the points, the angles between the arrows, and the length of the arrows all used to interpret the map.

The "arrows" (vectors) represent the variables. Longer arrows represent more discrimination.

The angles between arrows show the relationships between measures.

Respondent values (usually ratings) are read as projections on the arrows.

LDA biplots generally show separation between groups or segments better than PCA biplots, so they may be better suited for identifying the most discriminating attributes.

The following Fig. 4.21 shows the bi-plot for the Frog Dataset LDA. We can see that arrows goes in all directions, and this make both LD1 and LD2 share the information. The length of arrows are short which make them less discriminate the values. This biplot gives an indication that LDA is not that good in reducing dimensions.

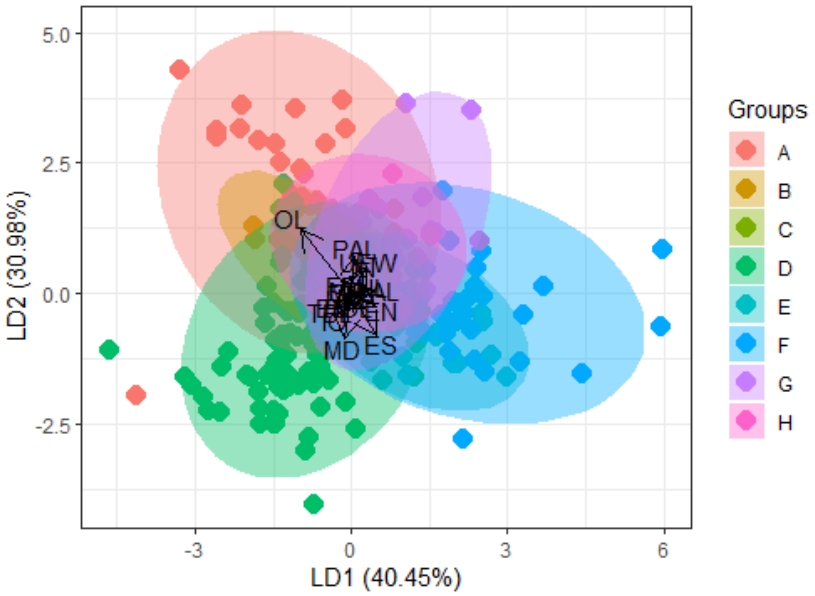


Figure 4.21: Bi-plot of Frog Dataset LDA

The next Fig. 4.22 shows the bi-plot for the female empowerment dataset LDA. We can notice most arrows are short and their projection are mostly on LD1, and this make LD1 discriminate most of the data. However, the variables IAE, HDI and SR discriminate values mostly by LD2, so LD1 and LD2 share in reducing dimensions for most data. The biplot indicates that LD1 and LD2 are very good in reducing dimensions for this dataset.

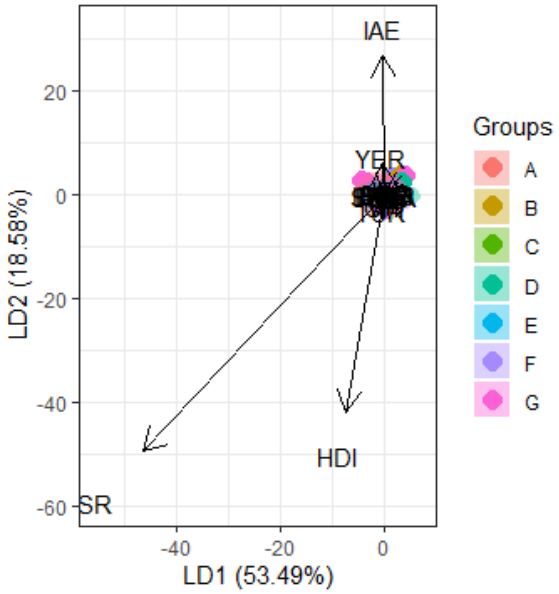


Figure 4.22: Bi-plot of Female Empowerment Dataset LDA

The below Fig. 4.23 shows the bi-plot for the corruption dataset LDA. The arrows are long in the direction of LD1 and their projection on LD1 have a high ratings, so LD1 discriminate most the data. LD2 has a very low discriminating as we can see from the arrow projections on LD2. The biplot shows that LDA will reduce the dimensions very well by LD1 only.

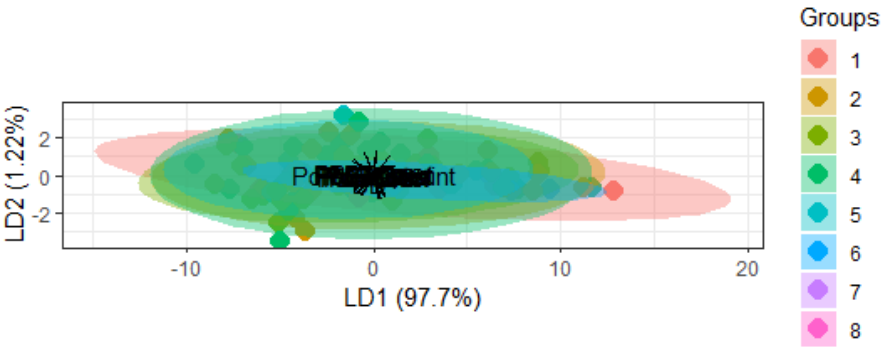


Figure 4.23: Bi-plot of Corruption Dataset LDA

The below Fig. 4.24 shows the bi-plot for the Pizza dataset LDA. The arrow are short, but most of them projects on LD1 and this makes LD1 discriminate most of the data. However, the variable sodium has a very good influence on LD2, and this will make LD2 also share the dimension reduction for this dataset.

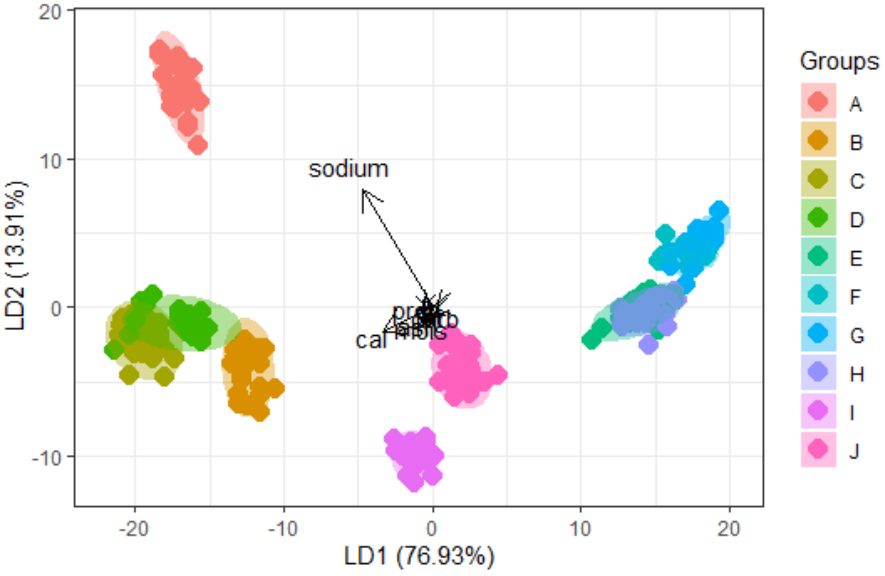


Figure 4.24: Bi-plot of Pizza Dataset LDA

Chapter 5

Conclusion and Recommendation

5.1 Introduction:

In this study, the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were presented, applied and compared in which method is better in dimensionality reduction.

We have relied on the number of components after reduction to give the best proportion of variance retained, so the total variance after reduction with the same number of components will determine the best method.

We used four real datasets to evaluate the performance of each method. These datasets are Frog dataset, Female empowerment dataset, corruption dataset and pizza dataset. We applied PCA on each dataset and specify the number of PCs after reduction and how much variance they have retained. Then, we applied LDA on the same dataset and specified the number of LDAs and how much variance they have retained as well.

The first dataset is the Frog. The dimensionality reduction of this dataset by PCA leads to two principal components which has 92.888% of variance. The dimensionality reduction by LDA on the Frog Dataset gives us two LDAs with 71.43% total variance. So, the result of PCA is better than LDA for the Frog Dataset as a dimensionality reduction.

The next dataset is the Female Empowerment dataset. The first two PCs after PCA is applied is 63.760% of the total variance, while LDA gives two LDAs with 72.07% of

variance. So LDA is better than PCA as dimensionality reduction on this dataset.

The third dataset is the Corruption Dataset. The first two PCs of PCA is 69.45% of the total variance, while LDA gives one LDA with 97.70% of total variance. So this time LDA is better than PCA as a dimensionality reduction on this dataset based on number of components and the total variance.

The final dataset was the Pizza Dataset, PCA has Two PCs with 92.32% of total variance, and LDA has two LDAs with 90.84% of total variance. As we can see, PCA is slightly better than LDA. They are very close to each other on this dataset.

The overall observation based on these examples is that, neither PCA or LDA is better all times in dimensionality reduction, sometimes PCA is better and in other LDA is better and this depends on the dataset characteristics.

The following table gives a summary of the above results:

Table 5.1: Summary of PCA and LDA results for the four datasets

Example	**n (No of observations)**	**p (No. of variables)**	**Minimum Correlation**	**Average correlation**	**No. of PC taken after reduction**	**% of total variance retained after PCA**	**No. of LDA components taken**	**% of total variance retained after LDA**	**which is better?**
Frog	251	17	0.757	0.909	2	92.888%	2	71.43%	PCA
Female Empowerment	177	19	0.007	0.460	2	63.760%	2	72.07%	LDA
Corruption	128	14	0.018	0.538	2	69.45%	1	97.70%	LDA
Pizza	300	7	0.023	0.608	2	92.32%	2	90.84%	PCA

5.2 Recommendations:

Based on inferences, we recommend that:

1. Before using one of the classification methods, it's important to check the data and explore the data graphically and numerically for assumptions.
2. PCA is better than LDA when there are a strong correlation between the variables, so when we check the correlation between the dataset varaibles and see a good correlation between variables, then it is recommended to use PCA not LDA.
3. PCA performs better in case where number of samples per class is less. Whereas LDA works better with large dataset having multiple classes.
4. The effect of sample size: PCA and LDA normality assumption is sensitive to sample size. Available tests work well with moderate sample sizes.As sample increases any small departure from normal leads to the rejection of normality.
5. Large datasets with many class labels gives LDA better chance in reducing dimensions than PCA.
6. LDA is better when there are many outliers as PCA is very sensitive to outliers.
7. This study is limited in scope, so we need to increase the variety of examples to make more concrete recommendations.

5.3 Future work:

We used the number of PCs and LDs with how much variance they have to compare between the two methods.

We can extend our evaluation method and use a classification method and decide which classification is better when we use PCA or when we use LDA.

Notice that these two tools are great in machine learning and data science.

Bibliography

[1] Vincent Spruyt. *The Curse of Dimensionality in classification*.

[2] Definition of "overfitting" at OxfordDictionaries.com: this definition is specifically for statistics.

[3] Deniz Erdogmus, Umut Ozertem and Tian Lan. *Information Theoretic Feature Selection and Projection*. (Studies in Computational Intelligence book series (SCI, volume 83), 2008.

[4] Lan Huong Nguyen & Susan Holmes.(2019) *"Ten quick tips for effective dimensionality reduction "*.

[5] Ali Ghodsi. *"Dimensionality Reduction- A Short Tutorial"*. Department of Statistics and Actuarial Science. University of Waterloo, Ontario, Canada, 2006.

[6] Cory Maklin.(2019). *Linear Discriminant Analysis in Python. Stories in Towards Data Science.*

[7] *Separability-Oriented Subclass Discriminant Analysis.* Article in IEEE Transactions on Software Engineering PP 99, (February 2017).

[8] Pearson, K. *"On Lines and Planes of Closest Fit to Systems of Points in Space"*. Philosophical Magazine. 2 (11): 559–572, 1901.

[9] Hotelling, H. *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, 24, 417–441, and 498–520, 1933.

[10] Hotelling, H. *"Relations between two sets of variates"*. Biometrika: 321–377, 1936.

[11] Anderson, T.W. (1963). Asymptotic theory for principal component analysis. Ann. Math. Statist., 34, 122–148.

[12] Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. Sankhya A, 26, 329–358.

[13] Gower, J.C. and Hand, D.J. (1996). Biplots. London: Chapman and Hall.

[14] Jeffers, J.N.R. (1967). Two case studies in the application of principal component analysis. Appl. Statist., 16, 225–236.

[15] Preisendorfer, R.W. and Mobley, C.D. (1988). Principal Component Analysis in Meteorology and Oceanography. Amsterdam: Elsevier.

[16] M. S. Wagner and David G. Castner. (2001). Characterization of Adsorbed Protein Films by Time-of-Flight Secondary Ion Mass Spectrometry with Principal Component Analysis. Langmuir 2001, 17, 15, 4649–4660.

[17] Jackson, J. E. *A User's Guide to Principal Components.* WileySeries in Probability and Statistics. John Wiley and Sons, 2003.

[18] Soren Hojsgaard. *Examples of multivariate analysis Principal component analysis (PCA).* Statistics and Decision Theory Research Unit, Danish Institute of Agricultural Sciences, 2009.

[19] Tauno Metsalu, Jaak Vilo. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Research, Volume 43, Issue W1, 1 July 2015, Pages W566–W570.

[20] Fang Yu & J. Meza. *in Proteomic Profiling and Analytical Chemistry, 2nd edition*. 2016.

[21] Abbas Alkarkhi & Wasin Alqaraghuli. *"Easy Statistics for Food Science with R", 1st Edition*. (Sep.2018).

[22] Anukrati Mehta. *Everything You Need to Know About Linear Discriminant Analysis.* Min Read, (Feb 2019).

[23] Martinez, A. M. Kak, A. C. (2001). "PCA versus LDA". IEEE Transactions on Pattern Analysis and Machine Intelligence. pp.228–233.

[24] Martinez, A. M.; Kak, A. C. *"PCA versus LDA" (PDF).* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001.

[25] Sebastian Raschka. *Implementing a Principal Component Analysis (PCA) in Python, step by step.* (Apr 13, 2014).

[26] Green, S.B. Salkind, N. J. & Akey, T. M. *Using SPSS for Windows and Macintosh: Analyzing and understanding data.* New Jersey: Prentice Hall, 2008.

[27] R. Duda et al. *Pattern Classification. John Wiley.* (pp. 117-124), 2001.

[28] John C. Davis. *Statistics and Data Analysis in Geology. Third edition.* (May 2002).

[29] *Expert Systems with Applications.* Pages 2314-2320. 15 February 2012.

[30] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2017). An Introduction to Statistical Learning with Applications in R. Springer.

[31] Telgaonkar Archana H.& Deshmukh Sachin. July 2015. Dimensionality Reduction and Classification through PCA and LDA.Intrnational Journal of Computer Applications.Volume122-No.17.

[32] Hussein Abdullatif.18october, 2018.Dimensionality Reduction For Dummies-Part1: Intuition.

[33] M. Bramer, Principles of Data Mining, 2nd edn, Springer.2013

[34] Y. Saeys, I. Inza and P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23(19) (2007), 2507–2517. doi:10.1093/bioinformatics/btm344.

[35] M.M. Tantawi, K. Revett, A. Salem and M.F. Tolba, Fiducial feature reduction analysis for electrocardiogram (ecg) based biometric recognition, Journal of Intelligent Information Systems 40(1) (2013), 17–39. doi:10.1007/s10844-012-0214-7.

[36] Pulkit Sharma. 27 August, 2018. The Ultimate Guide to 12 Dimensionality Reduction Techniques (with Python codes).

[37] Bellman R.E. Adaptive Control Processes. Princeton University Press, Princeton, NJ, 1961.zbMATHGoogle Scholar.

[38] Vincent Spruyt.2014." The Curse of Dimensionality in classification".

[39] Chuan Xu. 14 Jun, 2018 ·"Why is Dimensionality Reduction So Important?". University of San Francisco.

[40] R.O. Duda, P.E. Hart and D.G. Stork. 2012. 'Pattern Classification'. 2nd edn. Wiley.

[41] Gan, H., Sang, N., Huang, R., Tong, X. & Dan, Z. Using clustering analysis to improve semi-supervised classification. 2013. Neurocomputing 101, 290–298.

[42] Feiping Nie & Dong Xu & Xuelong Li. 2011. "Semi supervised Dimensionality Reduction and Classification through Virtual Label Regression".

[43] Jure Leskovec & Anand Rajaraman & Jeffrey David Ullman. 2014. Mining of Massive Datasets. 2nd edition. Cambridge University Press.

[44] Jolliffe I.T. (2002). “Principal Component Analysis”, Series: Springer Series in Statistics,2nd edition.

[45] Y.pang, Y. Yuan, and X. Li. 2008. ”Effective feature extraction in high dimensional space”, IEEE Trans. Syst.

[46] Graeme Hutcheson. Manchester University. Dictionary of Quantitative Methods in Management. Sage Publications.

[47] H. Hotelling. Analysis of a complex of statistical variables into components. J. ofEducational Psychology, 24:417441, 1933.

[48] Mia Hubert, Peter J Rousseeuw & Karlien Vanden Branden. (2012). A New Approach to Robust Principal Component Analysis. Technometrics Volume 47 pp.64-79.

[49] Jonathon Shlens. April 7, 2014.A Tutorial on Principal Components Analysis. Google Research. Mountain View, CA 94043 (Dated: April 7, 2014; Version 3.02)

[50] Smith, L. I. (2002). A tutorial on Principal Components Analysis (Computer Science Technical Report No. OUCS-2002-12).

[51] I. T. Jolliffe. (2002). Principal Component Analysis. Springer Science & Business Media.

[52] Roberto Reif. May 8, 2018. Limitations of Applying Dimensionality Reduction using PCA.

[53] Jonathon Shlens. April 7, 2014.A Tutorial on Principal Components Analysis. Google Research. Mountain View, CA 94043 (Dated: April 7, 2014; Version 3.02)

[54] Ghasemi, Asghar, and Saleh Zahediasl. 2012. “Normality Tests for Statistical Analysis: A Guide for Non-Statisticians.” Int J Endocrinol Metab 10 (2): 486–89. doi:10.5812/ijem.3505

[55] Amir Ali. Mar 10, 2019. Dimensionality Reduction (PCA and LDA) with Practical Implementation.

[56] Quarizmi AdTech. Nov 26, 2015. ‘A full introduction to The Linear Fisher Discriminant Analysis’.

[57] Hua Yu, Jie Yang. 23 October 2000. A direct LDA algorithm for high-dimensional data with application to face recognition. Carnegie Mellon University.

[58] F.R. Lopez, H. Jimenez-Salazar, D. Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science Vol 4394, pp. 468-475. Springer, 2007.

[59] Quarizmi AdTech. Nov 26, 2015. 'A full introduction to The Linear Fisher Discriminant Analysis'.

[60] R.A. Fisher. 1938. 'The Statistical Utilization of Multiple Measurements'. Annals of Eugenics.

[61] McLachlan, G.J. 1992. 'Discriminant Analysis and statistical Pattern Recognition'. John Wiley. New York.

[62] Volker Roth & Volker Steinhage. Nonlinear Discriminant Analysis using Kernel Functions. University of Bonn, Germany.

[63] Pritha Saha. May 25 2020. 'Linear Discriminant Analysis - An Introduction.

[64] RICHARD A. JOHNSON, DEAN W. WICHERN. 'Applied Multivariate Statistical Analysis'. Sixth edition.

[65] Davis, J.C., 2002. 'Statistics and data analysis in geology'. third edition. john Wiley & son's publication.

[66] M. Kirby. 2000. 'Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns'. Wiley.

[67] A. Hyvarinen, J. Karhunen and E. Oja. 2004. 'Independent Component Analysis'. Wiley.

[68] G.E. Hinton and R.R. Salakhutdinov .2006. 'Reducing the dimensionality of data with neural networks'.

[69] T. Hastie and R. Tibshirani. 1996. 'Discriminant analysis by Gaussian mixtures'. Journal of the Royal Statistical Society.

[70] P. Viszlay, M. Lojka and J. Juhár. 2014. 'Class-dependent two dimensional linear discriminant analysis using two-pass recognition strategy. in: Proceedings of the 22nd European Signal Processing Conference (EUSIPCO).

[71] S. Balakrishnama and A. Ganapathiraju. 1998. ‘Linear discriminant analysis-a brief tutorial’. Institute for Signal and Information Processing.

[72] BOKEOGLU ÇOKLUK, O, & BUYUKOZTURK, Ş. (2008). Discriminant function analysis: Concept and application. Eğitim araştırmaları dergisi .

[73] Box, G.E.P. December 1949. ”A General Distribution Theory for a Class of Likelihood Criteria”. Biometrika. 36 (3–4): 317–346.

[74] Buyukozturk, Şener; Çokluk-Bokeoglu, Omay. Oct2008.” Discriminant Function Analysis: Concept and Application”. Eurasian Journal of Educational Research (EJER).

[75] Tao Li, Shenghuo Zhu, and Mitsunori Ogihara. “Using Discriminant Analysis for Multi-Class Classification: An Experimental Investigation.” Knowledge and Information Systems 10, no. 4 (2006): 453–72.)

[76] YANG Xiaozhou. May 9 2020. ‘Linear Discriminant Analysis Explained’. Stories in Towards Data Science. The National University of Singapore.

[77] Duda, Richard O, Peter E Hart, and David G Stork. 2001. Pattern Classification. New York: Wiley.

[78] Sebastian Raschka. Aug 3, 2014. “Linear Discriminant Analysis– Bit by Bit”.

[79] R. Huang, Q. Liu, H. Lu and S. Ma, Solving the small sample size problem of lda, in: Proceedings of 16th InternationalConference on Pattern Recognition, 2002, Vol. 3, IEEE, 2002,pp. 29–32.

[80] J. Lu, K.N. Plataniotis and A.N. Venetsanopoulos, Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, Pattern Recognition Letters 26(2) (2005).

[81] T.-t. Feng and G. Wu, A theoretical contribution to the fast im-plementation of null linear discriminant analysis method us-ing random matrix multiplication with scatter matrices, 2014,arXiv preprint arXiv:1409.2579

[82] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim and Aboul Ella Hassanien. May, 2017. ”Linear discriminant analysis: A detailed tutorial”.Ai Communications (AI COMMUN) Journal. IOS Press.

[83] A. Tharwat, T. Gaber and A.E. Hassanien. 2016. ‘One-dimensional vs. two-dimensional based features: Plant identification approach’. Journal of Applied Logic.

[84] S. Rezzi, D.E. Axelson, K. Héberger, F. Reniero, C. Mariani and C. Guillou.2005. 'Classification of olive oils using high throughput flow 1 h nmr fingerprinting with principal component analysis, linear discriminant analysis and probabilistic neural networks'. Analytica Chimica Acta.

[85] S. Noushath, G.H. Kumar and P. Shivakumara. 2006. 'An efficient approach for face recognition, Pattern Recognition'.

[86] L. Yuan and Z.-c. Mu. 2007. 'Ear recognition based on 2d images, in: Proceedings of the First IEEE International Conference on Biometrics: Theory, Applications, and Systems.

[87] N.V. Boulgouris and Z.X. Chi.2007. 'Gait recognition using Radon transform and linear discriminant analysis, IEEE Transactions on Image Processing .

[88] R. Haeb-Umbach and H. Ney. 1992.'Linear discriminant analysis for improved large vocabulary continuous speech recognition'. International Conference on Acoustics, Speech, and Signal Processing.

[89] C.H. Park and H. Park. 2005. 'Fingerprint classification using fast Fourier transform and nonlinear discriminant analysis'.

[90] A. Sharma and K.K. Paliwal, Cancer classification by gradient lda technique using microarray gene expression data, Data & Knowledge Engineering 66(2) (2008).

[91] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim and Aboul Ella Hassanien. May, 2017. "Linear discriminant analysis: A detailed tutorial".Ai Communications (AI COMMUN) Journal. IOS Press.

[92] NaderMahmoudi, EkremDuman . 1 April 2015. "Detecting credit card fraud by Modified Fisher Discriminant Analysis" .Expert Systems with Applications. Volume 42, Issue 5. https://doi.org/10.1016/j.eswa.2014.10.037

[93] Mihalovic, M. (2016), Performance Comparison of Multiple Discriminant Analysis and Logit Models in Bankruptcy Prediction, Economics and Sociology, Vol. 9, No 4, pp. 101-118.

[94] Tahmasebi, P., Hezarkhani, A., Mortazavi, M. 2010. "Application of discriminant analysis for alteration separation; sungun copper deposit, East Azerbaijan, Iran. Australian" (PDF). Journal of Basic and Applied Sciences.

[95] Quarizmi AdTech. Nov 26, 2015. 'A full introduction to The Linear Fisher Discriminant Analysis'.

[96] Ricardo Gutierrez-Osuna. Introduction to Pattern Recognition. Lecture 6: Dimensionality reduction (LDA).Right State University.

الملخص

إعداد: ساجدة فريحات

إشراف: أ.د. حسن حنايشة

اقتحمت ثورة البيانات جميع جوانب حياتنا، حيث يمكننا مشاهدة ذلك في جميع مناحي الحياة حولنا والسبب في ذلك أنه أصبح يتم الحصول على المعلومات وتوليدها بسهولة وبالتالي انتاج كمية هائلة من البيانات.

هذا الكم الهائل من البيانات أصبح متعدد الأبعاد بشكل كبير، لذلك أصبح واجبا علينا إيجاد طريقة لتلخيص وعرض هذه البيانات بحيث يسهل فهما والتعامل مها.

إن عملية تحويل هذا البيانات من بيانات متعددة الأبعاد الى بيانات ذات أبعاد أقل تحتاج الى قدرة تخزينية ومعالجة كبيرة وبالتالي تصبح العملية صعبة جدا. ولحل هذه المشكلة ظهر عدة تقنيات لتقليل الأبعاد من اهمها تقنية تحليل المكونات الرئيسية وتقنية التحليل التمييزي اللتان هما موضوع هذه الرسالة.

تم دراسة تقنيات تقليل الأبعاد المختلفة وتم التعمق وعمل مقارنة بين طريقة تحليل المكونات الرئيسية وطريقة التحليل التمييزي وتم استخدام بيانات عالمية وتطبيق الطريقتين على البيانات بشكل منفصل للتعرف على ايهما أفضل في تقليل الأبعاد وبالتالي تقديم توصيات حول متى يكون استخدام أحدهما أفضل من الأخرى وما هي العوامل الذي تساعد في تفضيل أحدهما على الأخرى.

بعد الدراسة المستفيضة للطريقتين والمقارنة بينهما، وجدنا ان تحليل المكونات الرئيسية تكون أفضل من التحليل التمييزي عندما تكون العلاقة بين المتغيرات قوية أو عندما تكون عدد العينات لكل صنف قليل. من ناحية أخرى تكون طريقة التحليل التمييزي أفضل من تحليل المكونات الرئيسية عندما تكون البيانات كبيرة جدا وعدد التصنيفات فيها كبير جدا او عند وجود قيم متطرفة بشكل كبير في البيانات.