UNIVERSITY OF NAIROBI

PRINCIPAL COMPONENT ANALYSIS AND LINEAR DISCRIMINANT ANALYSIS IN GENE EXPRESSION DATA

by

EDWIN MUNENE KAGEREKI

A thesis submitted in partial fulfillment for the
degree of MASTERS OF SCIENCE IN MEDICAL STATISTICS

in the

COLLEGE OF HEALTH SCIENCES
UNIVERSITY OF NAIROBI INSTITUTE OF TROPICAL AND INFECTIOUS DISEASES

November 2013

Declaration of Authorship

I, EDWIN MUNENE KAGEREKI, declare that this thesis titled, 'PRINCIPAL COMPONENT ANALYSIS AND LINEAR DISCRIMINANT ANALYSIS IN GENE EXPRESSION DATA' and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University of Nairobi.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

The thesis of **Edwin Munene Kagereki** is approved:

Supervisor:

DR. ANNE WANGOMBE

Senior lecturer, University of Nairobi, School of Mathematics.

Signed: _____

Date: _____

UNIVERSITY OF NAIROBI, 2013

'I believe in evolution in the sense that a short-tempered man is the successor of a crybaby'

Criss Jami

Abstract

The datasets from microarray experiments enables the measurement of gene expression profile of in cells. Statistical models maybe used for classify the samples into various physiological categories based on the gene expression profile. However gene classification as a domain of research is not straight-forwad due to some inherent properties of the data; mainly multidimensionality and the noise.

The thesis studied three aspects of gene expression analysis. That is dimension reduction, classification of the expression profiles and described the variability of the gene expression data due to the covariates like age and gender. The dataset used in the thesis is the GEO dataset *GSE34105* . Principle Component Analysis and Eigen-R^2 methods were applied to dissect the overall variation. Subsequently a linear discriminant classifier was built and the effect of the number of principal components retained on the accuracy of the linear discriminant classifier was assessed using the leave-one-out cross-validation approach. All the data analysis was done in ***R 3.0.1 and R 2.6.2*** and the relevant packages.

The first three components accounted for a cumulative 33.34 % of the total variance (23.26 % , 6.02 % and 4.06 % respectively). The error rate of the linear discriminant classifier systematically increased at the number of retained principal components increased from three to seventy (6 % to 33 %). In our study the age explained 0.8 % of the variance, the disease condition 26.5 % and gender only 1.59 %. The accuracy of the linear discriminant classifier was highly dependent on the number of principal components retained. The error rate increased systematically from 6 % to 33% when the components retained were increased from 3 to 70.

The fact that the first few principal components explained a large proportion of the variance suggests that there were only a few genes that accounted for the signiffcant amount of the variance.This aligns with the knowledge that only a few number of genes present relevant attributes and that the gene expressed data comes with presence of noise which can be termed as technical and biological distortions of the data.

In conclusion the proper understanding of the variability of gene expression data is key to making proper biological conclusions. The appreciation of the contribution of the variability contributed to other biological factors is important in the study design.

Acknowledgements

I would like to gratefully and sincerely thank Anne Wangombe for her guidance, understanding and patience during my Msc studies at the University of Nairobi. Her mentorship was paramount in providing a well rounded experience consistent with my long-term career goals.

My sincere thanks also goes to my friends Siker Kimbung and Ben Weaver for your continous assistance.

My classmates - Valeria, Elizabeth,Chepkutto, and Augustine - I highly appreciate the team effort. I am glad we made it.

Last but not the least, I would like to thank my my parents Luke Kagereki and Anne Wambere for the contribution they have made in my life.

Contents

List of Figures

List of Tables

This work is dedicated to Bobbie for when the world said, "Give up" you always whispered, "Try it one more time!"

Chapter 1

INTRODUCTION

1.1 Motivation

A reliable, precise and timely diagnosis and classification of neoplasms is essential for successful treatment. Conventional methods for diagnosis and classifying human malignancies rely on a variety of morphological, clinical, and molecular variables. However due to the fundamental role played by genes in the neoplastic processes, gene expression profiling may provide more efficient solution to this complex problem. Golub et al [1], proposed the classification of cancers from gene-expression data. Currently there is no single classification methodology that is universally accepted, and the accuracy has been proposed to be dependent on classification method, gene selection method, and the dataset itself [2, 3].

The area of genetics has grown by leaps and bounds in the last decade. One of the huge developments is the micro-array gene-expression technology which has spread across the research community with immense speed. These experiments give a detailed snap-shot of genetic mechanisms in the cell at various physiological states. Thus micro-array technology can automate the diagnostic task and improve the accuracy of the traditional diagnostic techniques based on the molecular activities in the cell. With micro-arrays, it is possible to examine the expression of thousands of genes at once. Testing for differentially expressed genes can assist in prediction,diagnosis and classification of cancer cases. Micro-array experiments are thus creating a unique opportunity to improve our knowledge of the cellular machinery; best done by comparing activity in various states, eg diseased and normal. These experiments provide quantitative information about the

whole transcription profile of cells.. The data is subsequently subjected to various statistical and analytical techniques to provide biological knowledge. However gene classification poses multiple unique challenges with no universally accepted method known to achieve biology relevance and have classification accuracy. Gene classification as a domain of research poses unique challenges due to the nature of the data. First, most of these datasets have a small sample size (usually below 200), while having thousands to hundred thousands of genes presented in each tuples. Second, only a few numbers of these (genes) presents relevant attributes to the investigated disease. Third, comes from the presence of noise (biological and technical distortions) inherent in the dataset. Fourth is the challenge of achieving biological relevancy as well as acceptable classification accuracy.

Principal component analysis is a standard dimension reduction tool for multivariate data. In micro-array datasets the standard practice is to have the phenotype which has a large number of dimensions, represented by a small number of principal components sometimes referred to as super-genes [4]. In statistical parlance, the measured features are considered a set of related predictor variables used to predict the physiological condition. The same principal has been extended to include more functional analysis by including gene annotation. However in common practice dissecting the variation of transcriptional levels of thousands of genes in terms of relevant biological variables like age and sex is commonly ignored. This would be relevant to account for some variation in the data in form of the noise (biological and technical distortions). Recently extensions of the principal component analysis have however been proposed that can be used here [5]. From a biological point of view this would help biologists understand the level of variation in the gene expression data due to the inherent nature of the tissues as well as epigenetic factors.

Oral squamous cell carcinoma (OSCC) is a frequent neoplasm, which is usually aggressive and has unpredictable biological behavior and unfavorable prognosis. Early diagnosis of the disease, would be of immense help to the patient because prognosis depends on the progress of the disease. Classification of the disease depending on the unique characteristics with a bearing to the progress speed, metastasis, prognosis and treatment options would be of utmost benefit to the patient. Currently there exists no known molecular sub-types of OSCC, thus a proper understanding of the gene expression profiles in this condition would help

to identify any unknown molecular sub-types that may be relevant in the clinical management of the patients.

Thus it is necessary to develop proper profiling methods for oral squamous cell carcinoma. Furthermore, the gene expression of the oral and circum-oral tissues affected by other physiological conditions needs to be described and where need-be controlled in the study design.

1.2 Research question

Does the number of Principal components retained affect the accuracy of the linear discriminant classifier in gene expression data?

1.3 Study objectives

- To reduce the dimensionality of the gene expression data into a small
- To describe the source of variability in the gene expression micro-array data
- Build a linear discriminant function for the data
- Assess the accuracy of the linear discriminant function

1.4 Biological background

The aim of this section is to provide a basic understanding of human genetics. The contents of this introductory material can be found in basic biology textbooks and is thus summarised from one main sources [6]

The material in this section is divided into various conceptual blocks, building up from the structure of DNA to the process of protein transcription.

The discovery of the DNA structure by James Watson and Francis Crick paved way for a race for the human genome, characterised by controversies and awe-inspiring medical implications. This coupled with the ever increasing rate at which genomes are being sequenced has opened a new area of genome research, functional

genomics, which is concerned with assigning biological function to DNA sequences and extending this to other relevant aspects of molecular biology. With the ability to sequence the human genome, an essential and formidable task is to define the role of each gene and understand how the genome functions as a whole. Other potential areas of interest are profiling the phenotype based on the sequence. Innovative approaches, such as the c DNA and oligonucleotide microarray technologies, have been developed to exploit DNA sequence data and yield information about gene expression levels for entire genomes.

A gene consists of a segment of DNA which codes for a particular protein, the ultimate expression of the genetic information. A deoxyribonucleic acid or DNA molecule is a double stranded polymer composed of four basic molecular units called nucleotides. Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases. Each of the nucleotide is formed of a phosphate group, a sugar and a nitrogenous base. The nucleotides are connected by a phosphor-diester bond, in which a phosphate group links the 3-hydroxyl group of one nucleotide to the 5-hydroxyl group of the next, giving rise to directionality in the polynucleotide chain. Figure 1.1 is an illustration of the famous double helix structure of DNA that consists of two long strands of polymers formed by nucleotides, and the backbone of each strand is composed of sugars and phosphate groups joined by ester bonds. The figure also shows the chemical structure of DNA double helix, from which we can also see that attached on each sugar in the phosphate-deoxyribose backbone is one of the four different nucleobases, (cytosine, guanine, adenine, and thymine). It is the sequence of these four different nucleobases that encodes the hereditary information of biological organism. In the double helix of DNA, adenine pairs and bonds with thymine, and guanine pairs and bonds with cytosine.

While a DNA molecule is built from a four-letter alphabet, proteins are sequences of twenty different types of amino acids. The expression of the genetic information stored in the DNA molecule occurs in two stages: (i) transcription, during which DNA is transcribed into messenger ribonucleic acid or mRNA, a single-stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine; (ii) translation, during which mRNA is translated to produce a protein. The correspondence between DNA's four-letter alphabet and a protein's twenty-letter alphabet is specified by the genetic code, which relates

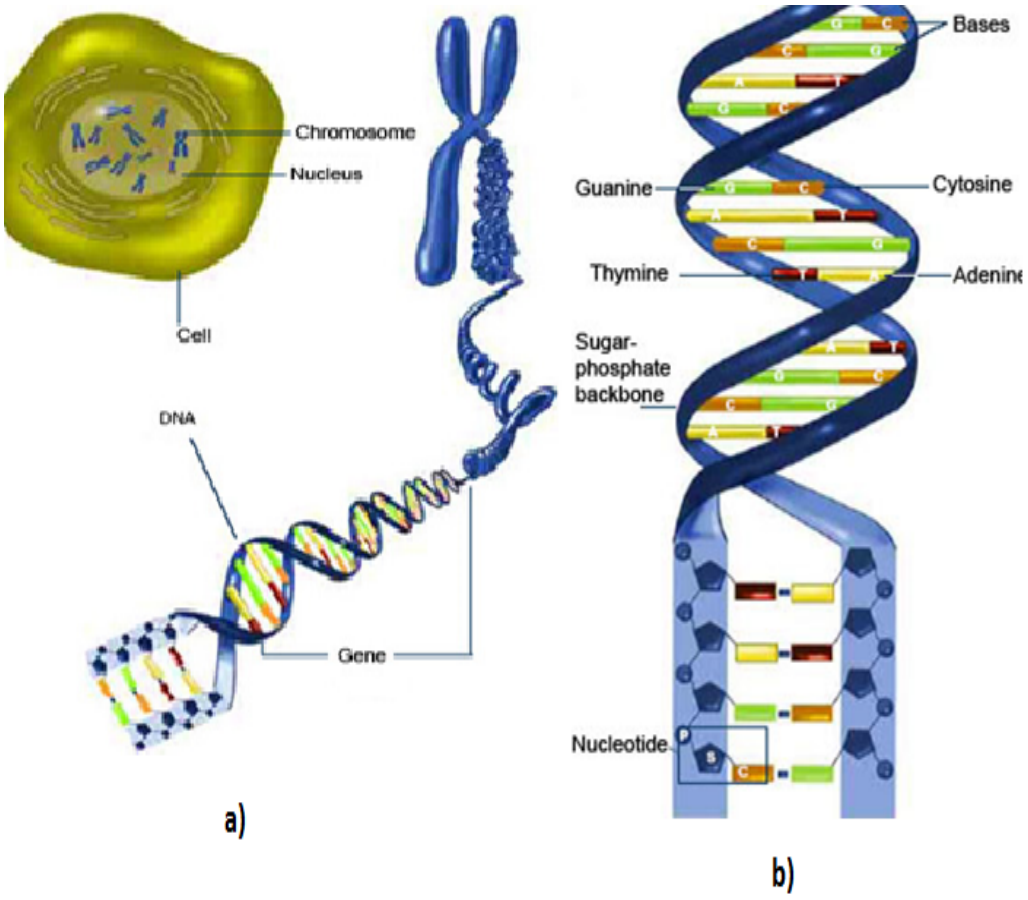

FIGURE 1.1: Structure of DNA

nucleotide triplets to amino acids. We refer the reader to the NIH educational website [6] for an introduction to the relevant biology.

The central dogma of molecular biology deals with the flow of information from the nucleic acid to protein within the biological system. It was first stated by Francis Crick in 1958 and re-stated in a Nature paper [7] This has simply been described as "DNA makes RNA makes protein". Two fundamental processes are involved in the processing of the genes - translation and transcription. The most fundamental pathway in biological system is the transcription of DNA to messenger RiboNucleic Acid (mRNA) and the translation of mRNA to protein that is the building block of biological system. Figure 1.2 illustrates the processes of transcription and translation. Within the cell nucleus, when transcription starts, the double helix of DNA opens up. The sequence information of a DNA strand is complimentary transcribed to a single-strand mRNA, which is also a long chain of nucleotides. Transcription basically transfers the coding information of DNA to mRNA.

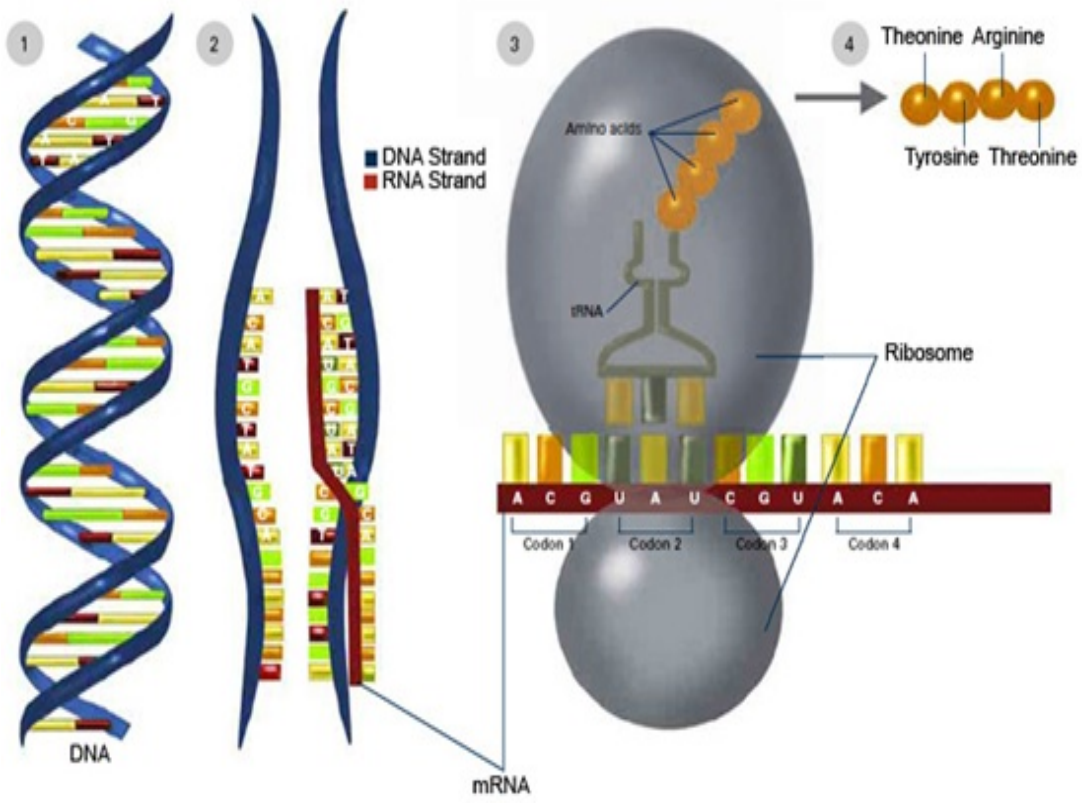

FIGURE 1.2: Central dogma

After transcription and several further processes on mRNA, e.g. capping, polyadenylation, and splicing, the mRNA matures and is transported to the cytoplasm, where amino acids are assembled according to the sequence information encoded in mRNA to form a chain, with help from other molecules and complexes, like transfer RNA (tRNA) and ribosome. This is the process of translation. The assembled amino acid chain after further processes, such as folding, becomes protein.

1.5 Micro-array experiments and bioinformatics

We now have the ability to attach a piece of every gene in a genome (all of an organism's genes) to a postage stamp-sized glass microscope slide. This ordered series of DNA spots is called a DNA micro-array, a gene chip or a DNA chip. Different properties of gene expression can be studied using microarrays, such as expression at the transcription or translation level, and sub-cellular localization of

gene products. However, currently much focus has been given on the expression at the transcription stage, i.e., on mRNA levels. Although the regulation of protein synthesis in a cell is by no means controlled solely by mRNA levels, mRNA levels sensitively reflect the type and state of the cell. Micro-arrays derive their power and universality from a key property of DNA molecules described above: complementary base-pairing. The term hybridization refers to the annealing of nucleic acid strands from different sources according to the base-pairing rules. To utilize the hybridization property of DNA, complementary DNA or cDNA is obtained from mRNA by reverse transcription. Multiple micro-array systems have been described and used depending on the manufacturer.

cDNA micro-arrays consist of thousands of individual DNA sequences printed in a high density array on a glass microscope slide using a robotic arrayer. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples or targets are reverse transcribed into cDNA, labeled using different fluorescent dyes (e.g. a red-fluorescent dye Cy5 and a green-fluorescent dye Cy3), then mixed and hybridized with the arrayed DNA sequences or probes. After this competitive hybridization, the slides are imaged using a scanner and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA sequence in the two nucleic acid samples.

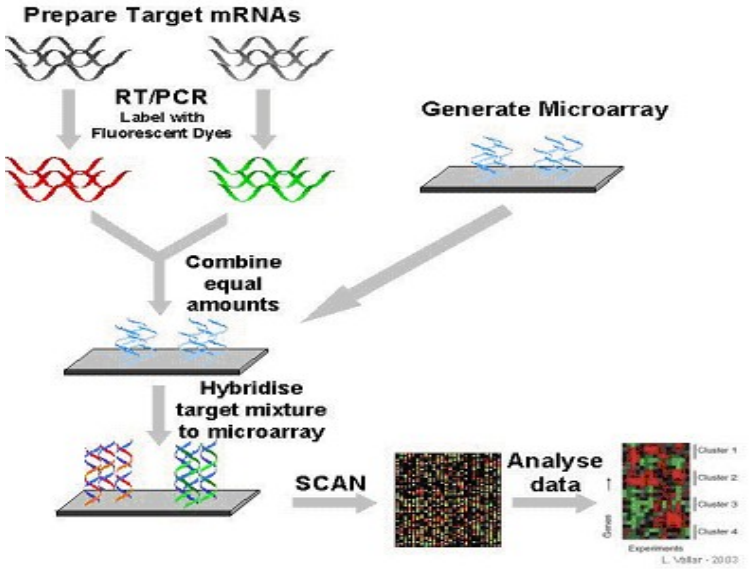

FIGURE 1.3: Steps in micro-array experiment(http://csmbio.csm.jmu.edu)

The data generated thereafter poses some inherent challenges in performing gene classification. Firstly, the curse of dimensionality whereby there are much more variables as compared to the sample size; where most of these datasets has sample size below 200, vs. thousands of genes presented in each tuples. Secondly, most of the genes in the dataset are not relevant to the disease, thus some methods of classification applies gene selection before the classification. Thirdly, is the presence of noise (biological and technical distortion).

The gene expression data set from the micro-array experiment can follow the following general representation [7]

$$\{G = (i, j) \,|1 \leq i \leq j \leq m\}$$

Where the columns

$$G = \{\overrightarrow{g_1}, \overrightarrow{g_2}, ...\overrightarrow{g_m}\}$$

form the expression pattern of genes and the rows

$$S = \{\overrightarrow{s_1}, \overrightarrow{s_2}, ...\overrightarrow{s_m}\}$$

To illustrate the micro-array dataset, an hypothetical table 1.1 is used. It shows the organisation of the data into m columns to represent the number of genes and n rows to illustrate the number of samples.

TABLE 1.1: Illustration of the gene expression dataset

	Genes				
Sample	$Gene1$	Gene 2	.....	$Genem$	$Class$
1	$G(1,1)$	$G(1,2)$	.....	$G(1,m)$	$OSCC$
2	$G(2,1)$	$G(2,2)$	.....	$G(2,m)$	$NORMAL$
....	....	.....	.....	.....	....
n	$G(n,1)$	$G(n,2)$	....	$G(n,m)$	$OSCC$

According to a review of the micro-array datasets done [8] , the number of gene (m) ranges from thousand to hundreds of thousands while the number of samples (n) is almost always less than 200.

The typical objectives of micro-array experiment are clustering (grouping the groups of genes differentially expressed), classification (profiling into phenotypes based on the gene expression), modeling (study the joint probability of the gene groups).

One necessary part of multivariate statistical analysis in such applications is dimension reduction.This is done by either selecting a subset of interesting genes (gene selection), or producing gene components or super genes combinations of genes (dimension reduction), or using combination of the strategies.

1.6 Organisation of the thesis

The main research topics discussed in this dissertation include application of PCA in dimension reduction of the gene expression data, expression profiling by LDA

and assessment of the classification. This thesis is divided into six chapters. Following this introductory chapter is Chapter 2 which discusses some common dimension reduction and classification methods used for cancer expression data as found in the literature.

Chapter 3 mainly discussed the methodology. The first part discussed the research question. This discusses the dataset, statistical methodology and the dataset used on the thesis, including the set-up of the experiment, the sample size and the data access. The results of evaluations are presented in Chapter 4. Then Chapter 5 discusses my interpretations and opinions and seeks to explain the implications of my findings, and suggest future research. Chapter 6 is the conclusion section. Thereafter appendices A and B present the code that we used for the analysis. Finally is the section on the references to the literature used in the thesis.

Chapter 2

LITERATURE REVIEW

As the bulk of publicly available expression data has grown, a variety of successful techniques have been proposed for its analysis. Broadly these techniques aim at assessing the classification or the clustering of the genes. However data collected by DNA micro-arrays are not suitable for direct statistical analysis, and thus several approaches have been suggested [1].

There is no one-size-fits-all solution for the analysis and interpretation of genome-wide expression data [9] , due to the complexity of biological systems, various goals of disease study and different experiment designs. Many bioinformatics tasks have been proposed for the analysis of gene expression data, for example, detection of differentially expressed genes between different biological conditions (or time points) [10, 11], detection of co-expressed genes/samples [1, 12], classification of new samples into known disease/phenotype categories [1, 12], and inference of gene regulatory networks and pathways [13].

The nature of the gene expression dataset poses a few unique challenges. Snousy et al [11] summarized these challenges.The first challenge is the curse of dimensionality, whereby the sample size is almost always les than 200, while the variables(genes) run into hundreds or into thousands. Secondly is the issue that only a few genes have relevance to the condition of interest. Thirdly is the issue of noise which is the presence of biological and technical distortion of the data. Forth is the challenge of interpretation and application of the results to bring biological relevance.

Expression classifiers are important because they can be used for diagnosis purposes in medicine and because they can help to understand the dependencies between classes (diseases) and features (gene expression values) [14].However, the problem of building classifiers from micro-array experiments is dimensionality sparseness: in general, there are a large number of features (gene expression measurements) against few examples (patients monitored). In high dimensional domains like this it is well known that many induction algorithms degrade in performance accuracy and run time. In fact many machine learning algorithms were developed to deal with high dimensionality. Therefore it is not always straightforward using such algorithms directly in these datasets. One possible solution to this problem consists in reducing high dimensional datasets through feature selection [15, 16] This approach has found use in this area with weighted voting of informative genes is used by Golub et al [1] whereas Chow et al [17]] employ support vector machines (SVM); Such classifiers provide high predictive accuracy but since they include many features they are not useful for human expert interpretation.

An alternative way to see this problem consists in preserving the logical connections among features enabling the induction comprehensible classifiers by human experts, where classifiers are expressed as rules for the labels of gene expression data [14, 18].

Besides the high dimensionality (large number of features), as mentioned before, the gene expression domain suffers from dimensionality sparseness: the high number of features contrasts with a very small number of examples. It is known from the literature that in such cases classifiers are prone to over fitting [19] because actually weak/irrelevant features can appear to be relevant simply by chance to machine learning algorithms due the available data sample [20] Over-fitted classifiers are characterized by low specialization error but high generalization error, in other words, there is a significantly increased error on unseen examples when compared to the training set error [21]

The classification may thus be divided into three broad topics: gene selection, classification, and performance testing.

2.1 Dimensionality Reduction

Due to the large dimension of the variable (genes) in the gene expression datasets dimension reduction has been adopted as the first step in solving the classification problem.These methods have been shown highly useful for classification with gene expression data. However, there is lack of comparison studies on those methods with the relative performance of their performance largely unknown.

Many approaches can be used to meet this goal. Although no standard method of categorizing then has been adopted, we discuss them depending on the technical approach used. One approach is to selects a subset of genes based on certain criteria such that this subset of genes is believed to best predict the outcome.This selection method uses univariate statistical methods such as t-test and rank test to to relate the individual gene expression level and the outcome of interest.[22, 23]. Another strategy is to construct gene component which are weighted combination of genes of lower dimension to represent the total variation of the data These combinations have been termed as super-genes. Representative approaches are principle component analysis (PCA) and partial least squares (SLR) [1].

Many methods have been proposed and applied in the data reduction methods,Li et al [2] compared the result of multi-class classification using many feature selections and classification methods. Eight methods were applied in this study information gain, twoing rule, sum minority, max minority, Gini index, and sum of variances, one-dimensional SVM, and t-statistics. The benefit of dimension reduction in for classification was clearly exemplified. Pei et al clearly showed that accuracy of the classification method increased upon reduction of the high number of the genes. There are different methods for subset selection and each has its own limitations ([19]).

2.1.1 Principal Component analysis

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set [24]. It accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal. PCA is a widely used method of dimension reduction in many areas of research .

These linear combinations are known as the principal components (PCs). Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability and are ordered in terms of the amount of variation of the original data they account for. The first PC accounts for as much of the variability as possible, and each succeeding component accounts for as much of the remaining variability as possible. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem. The projection vectors (or called the weighting vectors) u can be obtained by eigenvalue decomposition on the covariance matrix XS ,

Another interesting issue about the PCA is the choice of the number of the PCAs to retain. Two approaches may be used to select the number of components to retain. One can use components that correlate with a phenotype of interest [25] or use enough components to include most of the variation in the data [26]. However it has been suggested that that no matter what feature selection method is employed, at least 50 (and frequently more) features would need be chosen and used for classification in general [27].

Apart from PCA there are a diversity of methods that have been proposed and used in the literature fo rdimension reduction of the gene expression datasets. T-test has been applied in this aspect [23] Using this method, t-scores are computed for all genes and the top p* genes with the best scores are retained. We use both random subset selection and the t-score based gene selection in the assessment studies. The challenge with this method is that an arbitrary method for selecting the level of significance has to be introduced.

Partial least squares regression (PLS) is a statistical method that bears some relation to principal components regression; it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. The objective of constructing components in PLS is to maximize the covariance between the response variable y and the original predictor variables X. It seeks for a linear combination of attributes whose correlation with the class attribute is maximized. In PLS regression the task is to build a linear model, Y = BX + E, where B is the matrix of regression coefficients and E is the matrix of error coefficients. In PLS, this is done via the factor score matrix Y = WX with an appropriate weight matrix W. Then it considers the linear model, Y = QY + E, where Q is the matrix of regression coefficients for Y . Computation of Q will

yield Y = BX + E, where B = WQ. PLS has been adopted from the field of chemometrics [28] In this example the performance of PLS is compared to that of PCA. PLS proves to perform slightly better on some of the selected datasets.

Sliced inverse regression (SIR) is one of the sufficient dimension reduction methods. It is a supervised approach, which utilizes response information in achieving dimension reduction. This method has also been successfully used in some micro-array data dimension reduction [10]. This method, however, has limitations with problems where the number of predictors, p, exceeds the sample size, n, and can suffer when there is high collinearity among the predictors. Unfortunately this is the usual situation in the gene expression datasets.

Chi-square (v2) attributes evaluate. The chi-square (v2) method evaluates features individually by measuring their chi-squared statistic with respect to the classes. The v2 value where V is the set of possible values for a, n the number of classes, Ai(a =V) the number of samples in the i th class with a= v, and Ei(a = v) the expected value of Ai(a = v); Ei(a =v) = P(a = v)P(ci)N, where P(a= v) is the probability of a= v, P(ci) the probability of one sample labeled with the i th class, and N the total number of samples [13].

This being an active area of research, multiple methods have been proposed and used in the area of genomic. For example [10]compared eight methods which included the lesser used methods like information gain, twoing rule, sum minority, max minority, Gini index, and sum of variances, one-dimensional SVM, and t-statistics. Symmetry uncertainty- the principle behind this is the mutual dependence of variable. Machine learning algorithms such as LASSO and Random Forest have embedded capacity to select variables while simultaneously making predictions, and can be used to accommodate high dimensional micro-array data.

2.2 Classification Algorithm

On the other hand, recent comparative studies([29, 30]) suggest that, as far as gene selection is applied reasonably, simple and classical classifier such as k-nearest neighbor(k-NN) perform as well as or even better than more complex methods including SVMs. Hybrid methods have been suggested and and tested ([31])

The machine learning methods that have been shown to perform best are the SVM and Knn [3]

Commonly used classifiers that have been used in this field of study include neighborhood analysis [1], support vector machine (SVM) [8], k-nearest neighbor (KNN) [29], and linear discriminant analysis (LDA) The literature has multiple algorithms produced for classification models. These methods have developed from the earlier methods like nearest neighbor analysis and decision tree to the newer methods like support vector machines (SVM) [32]. Many studies have showed comparison of the various methods using a wide variety of datasets. First we introduce the documentation of the various methods and then we review the studies that compare the performances of the various methods. compared the performances of the NB and the DT, SVM and k-NN while applying various gene selection methods. In their experiment the accuracy ranged between 69.33 % and 90.01 %. In this study the experimenters used four attribute selection methods Chi-square, Information gain, Relief- F and symmetric uncertainty. Peter et al compared several methods and the accuracy attained ranged between 61.2 % to 99.4 %, however the gene selection method he used was partial least squares. Hong Hu et al brings on board a few new methods and compared them with the traditional methods. The accuracy ranged between 60 % to 98.9 %. Aik Choon et al [17] compared the various decision tree algorithms (single decision tree and ensemble based decision trees Bagging and AdaBoost). The accuracy of their experiment ranged between 52.38 % and 93.29 %. Pie et al compared various methods using two datasets . For binary classification (cancer vs. normal) the highest accuracy (close to 95 % for GSE3 and more than 99 % for (SE8054) was achieved with AdaBoost and a linear kernel SVM. For multi-class classification (SRBCT tumor subtypes) we achieve an accuracy of 100 % with a linear kernel SVM without feature selection and 98 % after reducing the feature dimension by 4 using the correlation coefficient feature selection technique. Classification of gene expression data may be categorised into two broad approaches: class prediction and class discovery. In the first part we concentrate on the supervised approach, learning from data with class labels from a dataset consisting of several gene expression values. Such classifiers are important because they can be used for diagnosis purposes in medicine and because they can help to understand the dependencies between classes (diseases) and features (gene expression values) [14]. In fact many machine learning algorithms were not developed to deal with high dimensionality. Therefore it is not always straightforward using such algorithms directly in these datasets. One possible solution to this

problem consists in reducing high dimensional datasets through feature selection [10, 15], which can provide the building of more accurate classifiers.

For classifier strategies, Dudoit et al. [1] carried out a comparison of current methods. The authors concluded that the diagonal linear discriminant analysis (DLDA) and nearest neighbors (NN) methods were among the few most accurate and stable classifiers.

Li et al [2] did a comparison for multi-class classifiers: SVM, KNN, and Decision Tree. They discussed that the SVM was the best classifiers for tissue classification based on gene expression. However, the best decomposition method for SVM appears to be problem-dependent. The KNN classifier gave good performance on most of the datasets which means it is not problem-dependent.

After dimension reduction, standard statistical models can be used for class prediction based on the smaller number of new predictors (e.g. [1]). The class prediction model we use for this study is the logistic discrimination (LD). This model has been widely used for binary class prediction problems and has been shown to perform well in previous studies [15].

All in all the NNs has previously demonstrated very good performance in several bioinformatics tasks ([23]);

The weighted voting method has been proposed and applied in gene expression data [1, 33] for classifying binary class data. The assignment of classes is based on the weighted voting of the expression values of a group of informative genes in the test tuple. The informative genes are genes that have high correlation with the class labels. Let the expression values of gene g in n training samples be represented by an expression vector g = (e1, e2, . . . , en), where ei denotes the expression value of g in tuple i. Let vector c = (c1, c2, . . . , cn) be the class vector denoting the classes of tuple i. Let (1(g), 1(g)) and (2(g), 2(g)) denote the mean and the standard deviation of the log10 of the expression values of g in class 1 and class 2 respectively. Then, the level of correlation, P(g, c), between the expression values of gene g and the class vector c is measured using signal-to-noise ratio(SNR). P(g, c) = (1(g) 2(g))/(1(g) + 2(g)) Intuitively, this metric favors genes with expression values that span a big range, has small variation within the same class and big variation between different classes. The value of —P(g, c)— is proportional to the correlation between the gene expression vector and the class vector. The sign of P(g,c) denotes which of the two classes the

gene is more correlated with and the magnitude denotes the degree of correlation. Positive P-values denotes higher correlation with class 1 and negative P-values denotes higher correlation with class 2. The larger the magnitude, the stronger the correlation. The informative genes, IG, are selected as follows: let L be the user input parameter for the number of informative genes to be selected. Then the GS method selects L/2 genes having the highest positive P values and L/2 genes having the highest negative values. For each g IG, define parameters (ag, bg), where ag = P(g, c), bg = (1(g)+2(g))/2. ag reflects the correlation of the expression values of g in the training data with the classes. bg denotes the average of the mean log10 expression values of g of training tuples in the two classes. Let and denote the mean and the standard deviation of the expression values of gene g in the training tuples. Given a test tuple s, where s = (s1, s2, . . . , sm). The class label of s is determined as follows: For each gene g IG with expression value in s denoted by sg, the normalized log10 expression value of g is defined as Norg = log10((sg)/). Define the vote of gene g as vg = ag(Norg bg), where the sign of the vote indicates the class(positive for class 1 and negative for class 2). Intuitively, each informative gene casts a weighted vote for one class, where the magnitude depends on the expression level of the gene in the test tuple and the degree of correlation of that gene has over the training set. The total vote for class 1, V1, by IG is the sum of all the positive votes, and the total vote for class 2, V2, is the sum of all the absolute values of the negative votes. Let Vwin be the total vote of the class that has the higher total votes, and Vlose be the total vote of the class with lower total votes. Then the prediction strength, PS, of the vote cast by IG is defined as PS = (V_win Vlose)/(Vwin + Vlose). PS denotes the relative margin on victory over the vote. A prediction strength threshold, pst, is used to determine if the prediction of the weighted voting is strong enough to assign the majority class to the test tuple. If PS pst, then the winning class is assigned to be the class label of s, otherwise, the weighted voting is considered to be too weak to assign the test sample to the voted class, thus assigning Uncertain as the class label to the test tuple.

Decision Tree is a flow-chart like structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes). A path from root to leaf represents classification rules. In decision analysis a decision tree and the closely related influence diagram is used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are

calculated. A decision tree consists of 3 types of nodes: 1. Decision nodes - commonly represented by squares 2. Chance nodes - represented by circles 3. End nodes - represented by triangles Various decision tree methods exist, Mohmad et al divides them into two single decision tree methods (C4.5, CART, REPTree, Random trees and Decision Stump) and the other gropu being the ensample decision trees ADTree, Random Forests, Bagging, and AdaBooost.

Khan et al [26] used neural networks for cancer type prediction. The method consists of three major steps: principle component analysis, relevant gene selection and artificial neural network prediction.

2.2.1 Discriminant analysis

First applied in 1935 by M. Barnard discriminant analysis is based on finding linear combinations xa of the gene expression levels x = (x1; : : : ; xp) Discriminant analysis is a classification problem, where two or more groups are known a priori and one or more new observations are classified into one of the known populations based on the measured characteristics. Here, we shall make the following standard assumptions:

1. The data from group i has common mean vector i

2. The data from group i has common variance-covariance matrix .

3. Independence: The subjects are independently sampled.

4. Normality: The data are multivariate normally distributed.

Linear discriminant analysis is used when the variance-covariance matrix does not depend on the population from which the data are obtained. In this case, our decision rule is based on the so-called Linear Score Function which is a function of the population means for each of our g populations i, as well as the pooled variance-covariance matrix.

2.3 Performance testing

There are many methods for estimating classification error. The performance of the proposed algorithm was evaluated by computing the percentages of Sensitivity (SE), Specificity (SP) and Accuracy (AC). As given in [5],

Sensitivity: is the fraction of real events that are correctly detected among all real events.

Specificity: is the fraction of nonevents that has been correctly rejected.

Sensitivity, specificity and accuracy of prediction have been calculated according to the following formulas:

Mohmad et al applied this method successfully in comparing the various classification approaches in decision trees. Dudoit et al [23] also applied a re-sampling and permutation method in testing the performance of the various methods in classifying various Leukemias.

In this method we use k-fold cross-validation with k=n, the number of samples in the training set. In each fold we use n-1 samples as training set and test the classifier on the remaining sample. This procedure is repeated for all samples. The estimated error is simply the fraction of wrongly classified samples. This method is computationally expensive as it requires the construction of n different classifiers. However, it uses almost all the samples in each training subset, thus it is more suitable for smaller datasets. The focus of this study will be on this cross-validation method (LOOC). We will perform external 10-fold cross validation as proposed by Ambroise et al [7] Although the LOOCV is not the best error rate estimator for the small size sample, an advantage of using LOOCV is it give almost unbiased estimation and the most important thing is its computational time is faster than the bootstrap method. The LOOCV also has been used in many studies in microarray classification. It is acceptable to perform the LOOCV for parameter analysis.

Chapter 3

METHODOLOGY

3.1 Dataset

The study dataset was obtained from the public database GEO.

The GEO accession GSE34105 dataset is a gene expression profiling of archival tongue carcinoma and normal tongue tissue.RNA extracted from 78 tongue samples, 62 tongue carcinomas and 16 non-malignant controls, were successfully analysed using the whole genome array to obtain gene expression profiles. We accessed the normalized dataset. The study population was from Umea City, Sweden. Dataset made public on May 11, 2012.

3.2 Statistical analysis

The analysis was done in a multi-stage approach to achieve the objectives. This is illustrated graphically below:

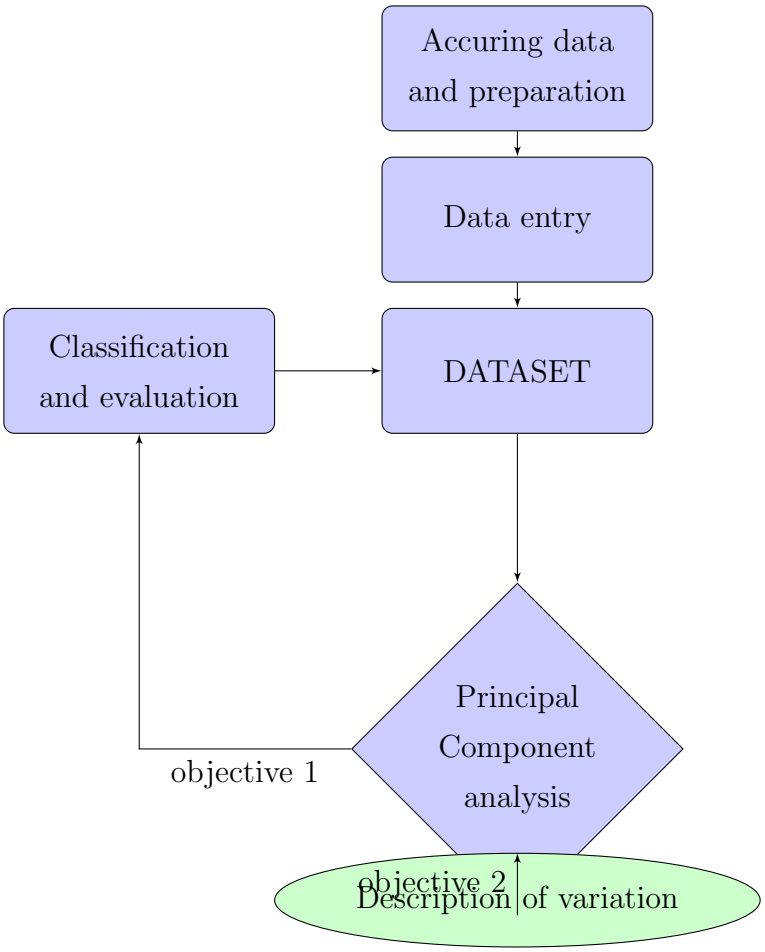

FIGURE 3.1: Work-flow chart

Data analysis was done using R-2.6.2, R-3.0.1 and the relevant packages - MASS,ggbiplot,ggplot2,

3.2.1 Dimension reduction

Principal component analysis was done to reduce the dimension of the dataset from the initial 29377. A scree plot was subsequently drawn to guide on the choice of the number of components to include for the next stage of the analysis. An extended version of the PCA - eigen R^2 was used to estimate the amount of variance that each of the covariates contributed.

3.2.2 Linear discriminant analysis

This will be done in a step-wise method as outlined below:

Step 1: Collection of the ground truth or training data.

Ground truth or training data are data with known group memberships. The datasets from the GEO databases with the classification were used.

Step 2: Computation of the discriminant functions/The classification rule.

Step 3: Use cross validation to estimate misclassification probabilities (Detailed section 3.3 below)

This is achieved by transforming the (p) original variables X = [x_1,x_2 x_p] to a new set of K predictor variables, T = [t_1, t_2, , t_K], . which are linear combinations of the original variables. Where (k) is always less than (p).

To illustrate the PCA algorithm the following equations are used.

Suppose that we have a random vector (X).

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

with population variance-covariance matrix

$$\text{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}$$

Consider the linear combinations

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

Each of these can be thought of as a linear regression, predictingY_i from X_1,X_2, ... , X_p. There is no intercept, but e_i1, e_i2, ..., e_ip can be viewed as regression coefficients.

Note that Y_i is a function of our random data, and so is also random. Therefore it has a population variance

$$\text{var}(Y_i) = \sum_{k=1}^{p} \sum_{l=i}^{p} e_{ik} e_{il} \sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_i$$

Moreover, Y_i and Y_j will have a population covariance

$$\text{cov}(Y_i, Y_j) = \sum_{k=1}^{p} \sum_{l=i}^{p} e_{ik} e_{jl} \sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_j$$

Here the coefficients $e_i j$ are collected into the vector

$\mathbf{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix}$ First Principal Component (PCA1): Y1

The first principal component is the linear combination of x-variables that has maximum variance (among all linear combinations), so it accounts for as much variation in the data as possible.

Specifically we will define coefficients $e_1 1$, $e_1 2$, ... , $Y e_1 p$ for that component in such a way that its variance is maximized, subject to the constraint that the sum of the squared coefficients is equal to one. This constraint is required so that a unique answer may be obtained.

More formally, select $e_1 1$, $e_1 2$, ... , $Y e_1 p$ that maximizes

$$\text{var}(Y_1) = \sum_{k=1}^{p} \sum_{l=i}^{p} e_{1k} e_{1l} \sigma_{kl} = \mathbf{e}_1' \Sigma \mathbf{e}_1$$

subject to the constraint that

$$\mathbf{e}_1' \mathbf{e}_1 = \sum_{j=1}^{p} e_{1j}^2 = 1$$

Second Principal Component (PCA2): Y2

The second principal component is the linear combination of x-variables that accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0

Select e_21, e_22, ... , e_2p that maximizes the variance of this new component...

$$\text{var}(Y_2) = \sum_{k=1}^{p}\sum_{l=i}^{p} e_{2k}e_{2l}\sigma_{kl} = \mathbf{e}_2'\Sigma\mathbf{e}_2$$

subject to the constraint that the sums of squared coefficients add up to one,

$$\mathbf{e}_2'\mathbf{e}_2 = \sum_{j=1}^{p} e_{2j}^2 = 1$$

along with the additional constraint that these two components will be uncorrelated with one another.

$$\text{cov}(Y_1, Y_2) = \sum_{k=1}^{p}\sum_{l=i}^{p} e_{1k}e_{2l}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_2 = 0$$

All subsequent principal components have this same property they are linear combinations that account for as much of the remaining variation as possible and they are not correlated with the other principal components

We will do this in the same way with each additional component. For instance: ith Principal Component ($PCA - i$): Y_i

We select e_1i, e_2i, ... , e_pi that maximizes

$$\text{var}(Y_i) = \sum_{k=1}^{p}\sum_{l=i}^{p} e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_i$$

subject to the constraint that the sums of squared coefficients add up to one...along with the additional constraint that this new component will be uncorrelated with all the previously defined components.

$\mathbf{e}_1'\mathbf{e}_1 = \sum_{j=1}^{p} e_{1j}^2 = 1$

$textcov(Y_1, Y_i) = \sum_{k=1}^{p}\sum_{l=i}^{p} e_{1k}e_{il}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_i = 0,$

$\text{cov}(Y_2, Y_i) = \sum_{k=1}^{p}\sum_{l=i}^{p} e_{2k}e_{il}\sigma_{kl} = \mathbf{e}_2'\Sigma\mathbf{e}_i = 0,$

$\vdots$

$\text{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^{p} \sum_{l=i}^{p} e_{i-1,k} e_{il} \sigma_{kl} = \mathbf{e}'_{i-1} \Sigma \mathbf{e}_i = 0$

Therefore all principal components are uncorrelated with one another.

3.2.3 Cross-validation

To assess ho well the model classification performs to an independent dataset.

In this type of validation, one case in our data set is used as the test set, while the remaining cases are used as the training set. We iterate through the data set, until all cases have served as the test set. In order to implement the iteration in R, we introduce an extra column, that is used as an index to identify the leave-out-case. Here is the code, for the gala data set in the faraway package.

Cross validation was done using the leave-one-out-cross-validation(loocv) method was used. For a dataset with N examples, perform N experiments. For each experiment N-1 samples were used for training and the remaining example for testing.

This was evaluated by computing the Sensitivity (SE), Specificity (SP) and Accuracy (AC). These terms are defines as [5]:,

Sensitivity the proportion of cases with disease who are correctly predicted

Specificity proportion of cases without disease who are correctly predicted

Sensitivity, specificity and accuracy of the particular method was calculated as follows:

$$SENSITIVITY = \frac{TP \times 100}{TP + FN}$$

$$SPECIFICITY = \frac{TN \times 100}{TN + FP}$$

$$ACCURACY = \frac{TP + TN \times 100}{TP + FN + FP + TN}$$

Where

- TP Number of predicted positive cases that are actually positive
- TNNumber of predicted positive cases that are actually negative
- FP Number of predicted cases that are actually negative
- FNNumber of predicted negatives cases that are negative

Chapter 4

RESULTS

A total of 78 people were involved in the study. The average age of the participants was 56.35 ± 17.5. The female were 36 and the male were 42. The number of samples from the patients who had Oral Squamous Cell carcinoma was 62 as compared to 16 which were collected from patients with no OSCC. A total of 29377 markers were sequenced per sample. There was a no significant difference in the age for cancer patients (M= 58.19 ±17.44) and normal (M= 49.19 ± 16.32); t= -1.94, p = 0.064. There was also no gender difference in the prevalence of cancer amongst the participants;X^2, df=1, p=0.95.

Principal components were calculated which represented the aggregated trends in the gene expression profiles. The first PC was the linear combination of the gene expression profiles that explained the most variation in the data. The second PC was the linear combination of the gene expression profiles that explained the most variation in the data once the first PC had been removed, and so on. The top-3 PCs which explained 23.26 %, 6.02 % and 4.06 % of the total variation,respectively. It was noteworthy that the first 10 PCs explained 51.94 % variation, suggesting that the gene expression profiles might be affected by only few but significant factors. Using eigen-R^2 the variability contributed by age was estimated at 0.8 % , the disease condition 26.5% and gender 1.59 %. The fraction of total variance in the data as explained or represented by each PC was represented in a scree plot.

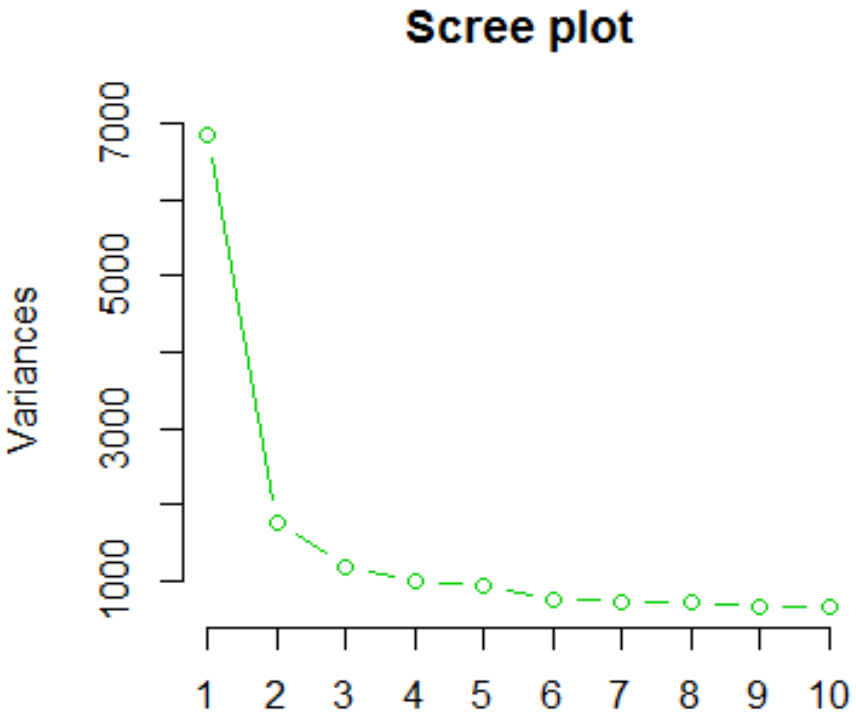

FIGURE 4.1: Scree plot

Guided by the scree-plot the first three Pcs were retained. There was a strong negative correlation between PC1 and PC2 (-0.86) but very weak correlations between the PC2 and PC3 (0.05)m and between PC1 and PC3 (0.005)

Differences between the two groups were examined. T -tests were done to check the difference between the PCs amongst the tumour patients and the normal patients. The results were tabulated below:

TABLE 4.1: Summary of the t-test(disease PCs)

	mean control	mean tumour	Pvalues
PC1	38083.37	-15788.85	<0.05
PC2	-7090.13	60650.46	<0.05
PC3	58820.40	50106.70	0.11

Further the difference between the two groups was graphically shown.:

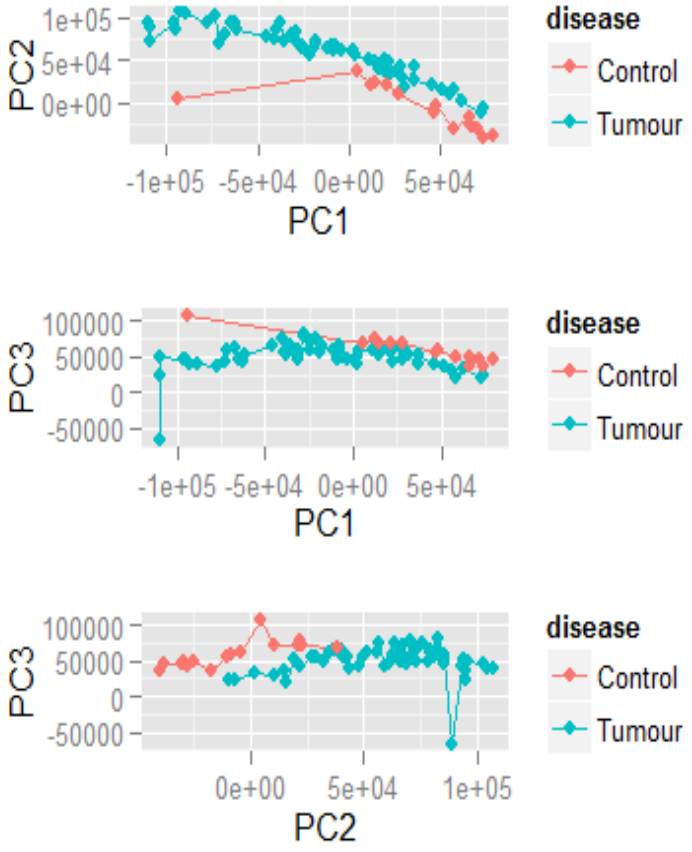

FIGURE 4.2: Graphs of first 3 PCs

The differences in the other variables was further plotted as shown the the tables 4.4,4.5 and 4.3 below.

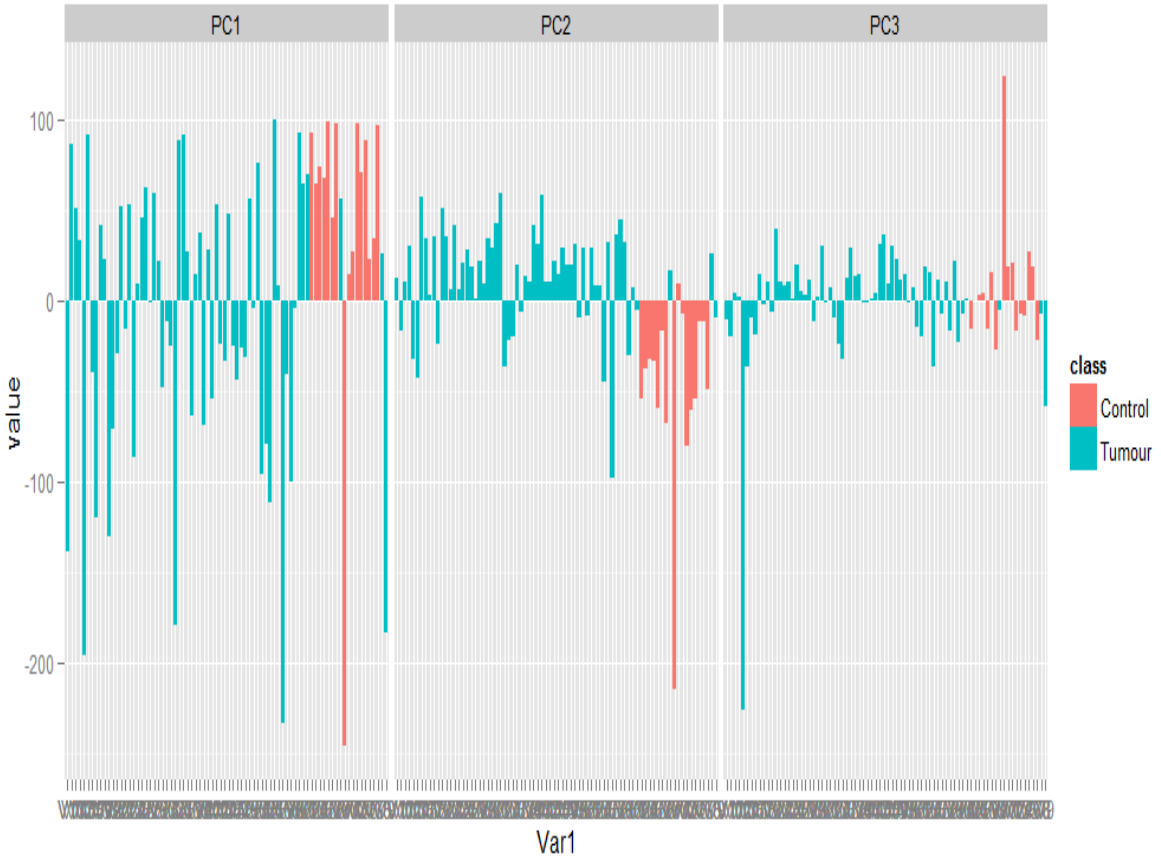

FIGURE 4.3: The explained variance by disease status

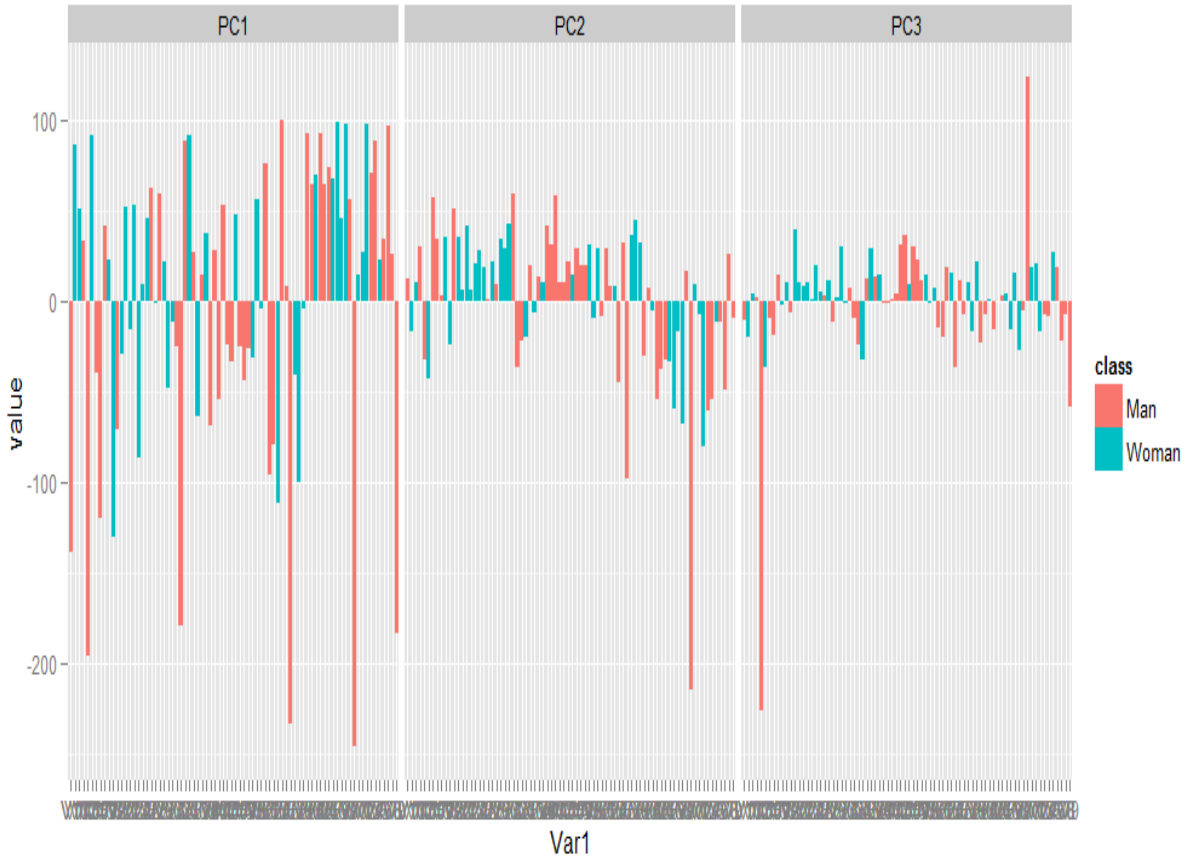

FIGURE 4.4: The explained variance by gender

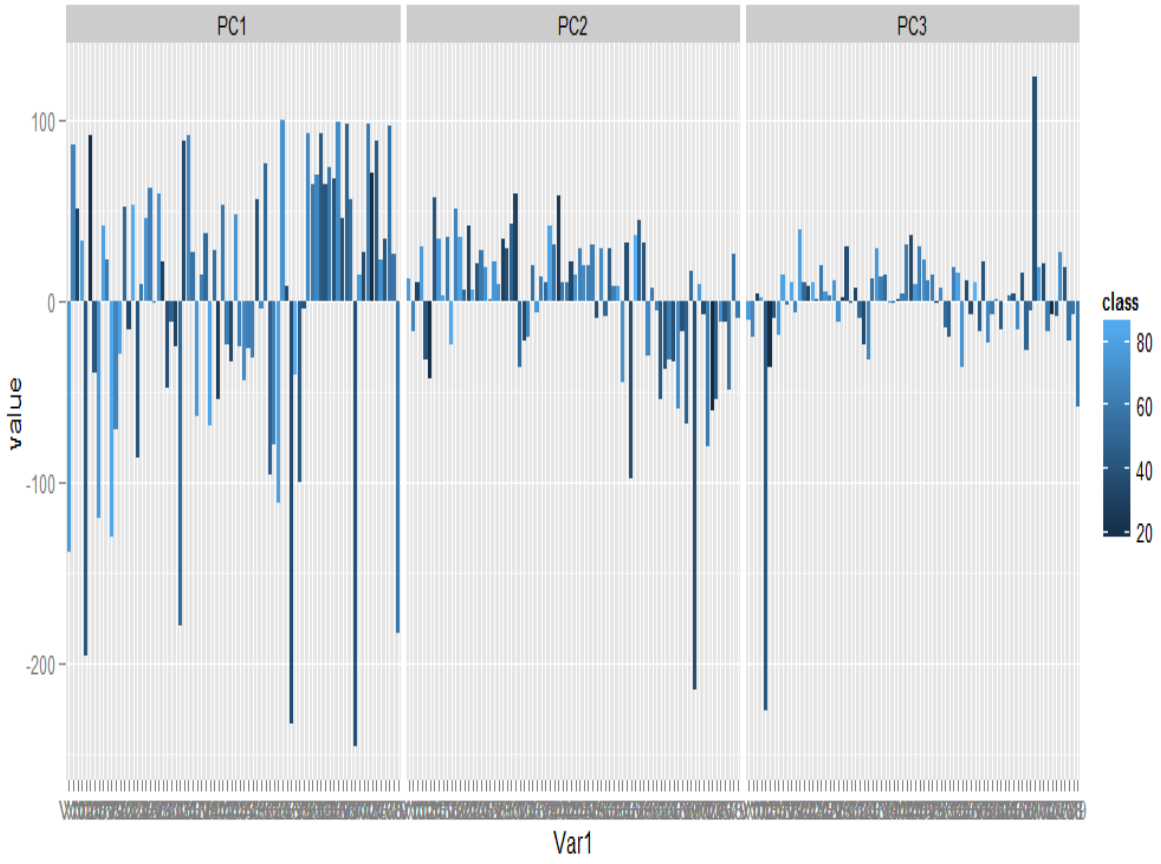

FIGURE 4.5: The explained variance by disease age

The dataset was trained with the 78 samples and testing was done using two approaches. Initially the whole dataset was used as the test database. Subsequently the LOOCV approach was used and the results tabulated below. To illustrate this the table below shows the confusion matrices for the testing process when three components were retained.

1

TABLE 4.2: Confusion matrices when PCs retained=3

TABLE 4.3: A

	Control	Tumour
Control	16	0
Tumour	0	62

TABLE 4.4: B

	Control	Tumour
Control	9	20
Tumour	7	42

Table A: Actual values and Table B: LOOCV predicted values.

The error rate was subsequently analysed. The range of our analysis was from 2 PCs to 70 PCs. The general trend was that the more we increased the number of the PCs the less accurate our classification became. The error rate increased

from 6% to 33 %. There was a strong correlation (0.77) between the the number of PCs and accuracy. A scatter plot describes this phenomenon:

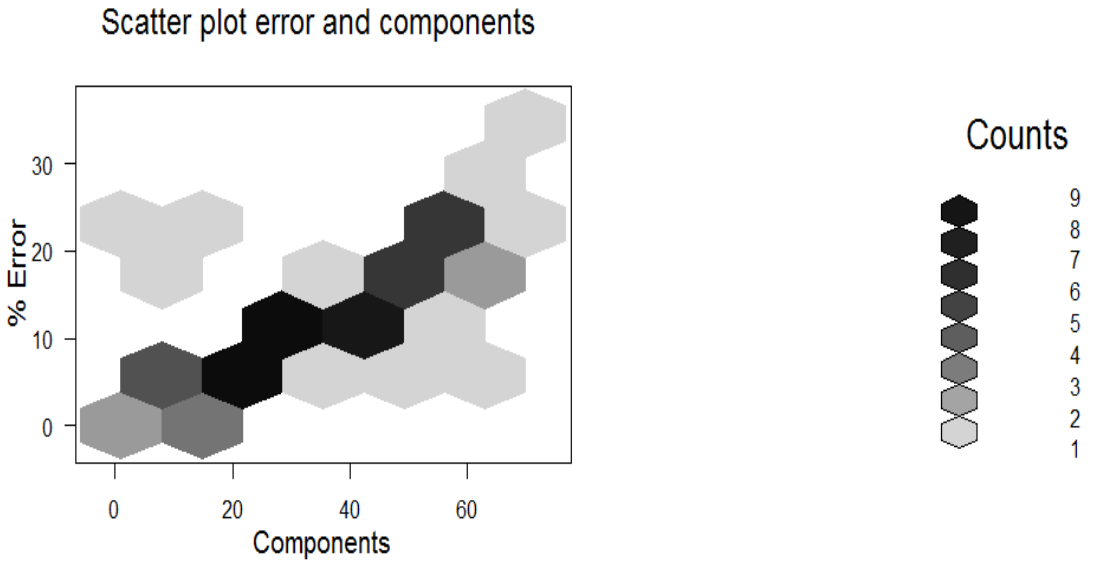

FIGURE 4.6: Scatter plot of error rate and number of Components

Chapter 5

DISCUSSIONS

Oral squamous cell carcinoma is a highly prevalence cancer of the oral and circumoral tissues. In the current study we have examined three aspects of gene expression analysis. That is dimension reduction, classification of the expression profiles and factors that may contribute to variance in the data.

PCA is a classical technique to reduce the dimensionality of the data set by transforming the large dataset to a new set of variables (the principal components) to summarize the features of the data. Principal components (PCs) are uncorrelated and ordered depending on the variance. It is very popular in micro-array data analysis, where the principal components are interpreted as the (few) physiological processes driving the variability in the dataset. The k-th PC can be interpreted as the direction that maximizes the variation of the projections of the data points such that it is orthogonal to the first k-1 PCs.

The set of principal components is often reduced to a set of size k, where $1 < k < pc$. The objective of dimension reduction is to make analysis and interpretation easier, while at the same time retaining most of the information (variation) contained in the data. Clearly, the closer the value of k is to pc the better the PCA model will fit the data since more information has been retained, while the closer k is to 1, the simpler the model.

Many methods have been proposed to determine the number k, that is, the number of components to retain. The complexity of the methods is varied; some using simple graphs while others are computationally intensive.These methods include [24, 34] (among others): the broken stick model, the Kaiser-Guttman test,

Log-Eigenvalue (LEV) diagram, Velicer's Partial Correlation Procedure, Cattell's SCREE test, cross-validation, bootstrapping techniques, cumulative percentage of total of variance, and Bartlett's test for equality of eigenvalues. .

Thus the various techniques are prone to suffer from subjectivity or have a tendency to under estimate or over estimate the true dimension of the data [35]. Thus there is no ideal solution to the problem of dimensionality in a PCA as noted by others among Jolliffe et al that [24] notes ' it remains true that attempts to construct rules having more sound statistical foundations seem, at present, to offer little advantage over simpler rules in most circumstances.'

The traditional approach is to use the first few PCs in data analysis since they capture most of the variation in the original data set. In contrast, the last few PCs are often assumed to capture only the residual noise in the data. However, deciding how many and which components to use in the subsequent analysis is a major challenge. It has been suggested that[25, 26], one can use components that correlate with a phenotype of interest or use enough components to include most of the variation in the data.

It has been suggested that for the purpose of classification the number of components to be retained has to be at-least 50 (and sometimes more) [27]. In our study the first 70 components that cumulatively accounted for 98.39% of the variance in the dataset for purposes of classification. However the scree plot suggests that only the first two PCs should be retained. For the classification purpose we chose to compare the accuracy on retaining the various number of components. However for the graphical illustrations we chose the first 3 components as guided by the scree plot . These first three components accounted for a cumulative 33.34 % of the total variance. This suggested that there were only a few genes that accounted for the significant amount of the variance. This aligns with the knowledge that only a few number of genes present relevant attributes and that the gene expressed data comes with presence of noise which can be termed as technical and biological distortions of the data. It would be worthwhile note that the the first 3 components accounted for 23.26 % , 6.02 % and 4.06 % of the variance respectively.

In the analysis of the T-tests we found that the first two components were different amongst the two categories of the participants. This would be explained by the fact that these two PCs contributed alot of variance in the initial analysis. Ignoring potential sources of experimental bias, such as assuming the variance in the

gene expression dataset is due to the disease condition alone, may yield misleading results. This practice overlooks important variables that may have small, but reproducible, changes in expression. We cannot ignore the known biological phenomenon of epigenetic that the gene expression level are affected by other factors that have been established. Biological data is inherently variable, and statistical inference is required in order to draw conclusions from data and add to the body of knowledge. Collecting data and acquiring knowledge are not the same thing. Good design and sound statistical inference will be a crucial factor in determining whether micro-arrays fulfill their potential.

Jin et al [36] estimated the contributions of sex, genotype and age to transcriptional variation in *Drosophila melanogaster*. They found that expression variation is mostly explained by sex, genotype and their interactions, and less explained by age. Brem [28] dissected the variation of expression in yeast according to genotype, based on recombinant lines derived from two distinct isogenic strains. The genome-wide transcription variation explained by population structure has been estimated in the teleost fish and humans [11]. In all of these studies, one can think of each feature as being a response variable, where a key summary statistic is the proportion of variation among the thousands of response variables explained by the independent variables of interest.An even more relevant study has been to estimate the various factors that affect the degree of variation in blood gene expression profiles.

Thus from these studies there are some factors that affect the gene expression profiles in specific tissues. Statistically failure to factors these aspects in the study design or in the data analysis would introduce some bias to the inference made.

For a single response variable, the proportion of variation explained by independent variables is usually accomplished by calculating R^2. This is computed as the ratio of the variance of the fitted model to the variance of the response variable. Eigen-R-square is a high-dimensional version of the classic R-square statistic. It can be applied when one wants to determine the aggregate R-square value for many related response variables according to a common set of independent variables. In our study the age explained 0.8 % of the variance, the disease condition 26.5% and gender 1.59 %. Clearly most of the variance would be attributed to the disease condition. These figures are relatively less compared to a previous study which showed that gender and age contributed 8.3 % and 9.2 % respectively in the blood transcription profile [37]. However it would be biologically wrong to compare these

figures since difference tissues have difference transcription profiles. Thus it would be nice to have estimates of the different factors that influence the gene expression in different tissues. This would form a basis for proper study design to avoid confounding effect.

We therefore described the variation of the other variables in the study.

One of the main goals of analysing gene expression data is for purposes of classification. In this aspect the identification of an ideal classification method is crucial so that the research may be translated to the clinical use, and thus better patient care.

Multiple methods have been suggested and subsequently compared. The classification is almost always preceded by dimension reduction. So in the literature, a classification method is assumed to be the combination of a dimension reduction method and the subsequent classification algorithm. In the same approach, our method could be termed as PCA-LDA method. The success of the method described as in the methodology section. However,factors other than accuracy contribute to the merits of a given classifier [23]. These include simplicity and insight gained into the predictive structure of the data.

In our study it was notable that the accuracy of the classification method reduced as we increased the number of components. Although previously it has been noted that the ideal number of components that we should retain for purposes of classification should be at-least 50 [27], the larger number proved to be less accurate. Our results however resonate with previous findings where the accuracy of the number of PCs retained was checked in a classification experiment [38]. In this experiment where the k-nearest neighbor was used as the classification method, the results of increasing the number of PCs was similar to our findings. Actually, simply setting $k = 1$ gave the best result.

Despite the fact the there is no gold standard in this area,linear discriminant analysis remains a relatively common method of classification. In this study we came up with an classification function and then attempted to estimate the error rate of the results. In our study we studied the accuracy of the discriminant function on altering the number of the principal components. generally the trend was that as we increased the number of the components to retain our accuracy reduced.

Chapter 6

CONCLUSION

In the study described here, we intended to asses the accuracy of linear discriminant analysis in gene expression of oral squamous cell carcinoma. However we also wished to describe the characterize the major sources of variation in the same dataset.

We used the PCA and Eigen-R^2 method to dissect the overall variability of gene expression data and associate the major sources of variation with the predefined biological variables.

Only a few PCs explain a large proportion of the variance, with the first three components accounting for a cumulative 33.34 % of the total variance. This resonates well with the fact that there is a lot of noise in the gene expression data. It has been suggested that only a few attributes are relevant to the disease. Thus analysis of gene expression data should only incorporate these few PCs for accurate results. This is also confirmed in the classification, where the more the number of PCs retained the less accurate the classification. This confirms that an accurate classification model should only incorporate the PCs explaining maximum variance.

The results indicated that the variation in the gene expression dataset could also be attributed to these other factors. The physiological factors (age and gender) were found to be associated with some proportion of variation in the gene expression profiles. In our study the age explained 0.8 % of the variance, the disease condition 26.5% and gender only 1.59 %. Remarkably the disease status represents the most significant portion of the overall variation explained. Recently, molecular biology

techniques, especially epigenetic and genetic techniques, have been developed that have enabled us to gain a greater insight into the molecular pathways underlying the cancers. In translating the research into a format that will facilitate effective molecular classification, support personalized treatment and determine prognosis remains a challenge. In this thesis, the possibility of gene expression dependency on physiological factors is highlighted. Thus the same idea maybe extended to a clinical setting for patients with epigenetics and potential confounder. Thus the need to factor them in when doing the classification of the datasets cannot be ignored. In the study design, physiological factors need to be well controlled and have them equally distributed between the comparison groups.

Based on the results of the study, it was identified that carrying out the classification process requires consideration of the other factors that would affect gene expression. Biologically these processes called epigenetics may not have found very a lot of use in the statistics. Thus the various physiological states that have been attributed to gene expression but may not be well related to the disease must be controlled. Although the dataset considered a few covariates, we were able to demonstrate the role they play. Notably some pioneer work [37] has been done in the to identify the factors that may alter the variation in of gene expression profiles.

Thus identifying it is important to identify the sources of variation in gene expression profiles in various tissues. This will help in proper study designs to limit the confounding effect to the minimum., such variability thus improving the accuracy and reproducibility of the gene expression studies.

Appendix A

The R Code (EDA and PCA)

```
# Analysis scheme:

# 1. Import the data into R - 2 datasets have been imported - One contains the actual micro-arra
#status
#2. Exploratory of the dataset - all the variables
#3. Associations between the various variables with keen interest on the variable of interest -
#4. Principal component analysis

library(stats)
library(MASS)


setwd("C:/Users/EDWIN/Desktop/gene express t")
data<-read.table("data2.csv", sep=",", h=F)
data2<-t(data)

label<-read.table("labels.csv", sep=",", h=FALSE)
sample.lables<-t(label)
desc<-read.table("desc.csv", sep=",", h=TRUE)

ngenes<-nrow(data)
nsample<-nrow(desc)
meanage<-mean(desc$age)
sdage<-sd(desc$age)
male <- desc[ which(desc$gender=='Man'),]
female <- desc[ which(desc$gender=='Woman'),]
nmale<-nrow(male)
nfemale<-nrow(female)
tumour <- desc[ which(desc$disease=='Tumour'),]
control <- desc[ which(desc$disease=='Control'),]
ntumour<-nrow(tumour)
ncontrol<-nrow(control)


tbl<-table(desc$gender,desc$disease)
test1<-chisq.test(tbl)
```

```
tdisease<-t.test(desc$age~desc$disease)
tsex<-t.test(desc$age~desc$gender)
pt2<-tsex$p.value
pt2<-round(pt2,3)
tstatdis<-tdisease$statistic
tstatsex<-tsex$statistic
pt<-tdisease$p.value
pt<-round(pt,3)
pchi<-round(test1$p.value,2)
chistatistic<-round(test1$statistic,2)
tstatistic<-round(t$statistic,2)
control <- desc[ which(desc[1]=='Control'), ]
disease <- desc[ which(desc$disease=='Tumour'), ]
meancancer<-mean(disease$age)
sdcancer<-sd(disease$age)
meannormal<-mean(normal$age)
sdnormal<-sd(normal$age)

pca.m <- prcomp(data2, scale=TRUE)
chosen.components <- 1:70
feature.vector <- pca.m$rotation[,chosen.components]
compact.data <- t(feature.vector) %*% t(data2) ## This thing has given me a whole week of headac
# Thus our new dataset is a 70*78 matrix - 70 loadings and 78 samples- If we chose 70 PCs
compact.data2<-t(compact.data)

# Then the plots
par(mfrow=c(3,2))
plot(summary(pca.m)$importance[3,], type="l", ylab="%variance", xlab="nth component (decreasing
abline(h=0.99,col="red")
abline(v=50,col="red",lty=10)#
mtext("A")

plot(summary(pca.m)$importance[3,], type="l", ylab="%variance ", xlab="nth component")
abline(h=0.99,col="red")
abline(v=20,col="red",lty=10)#
mtext("B")

plot(summary(pca.m)$importance[3,], type="l", ylab="%variance ", xlab="nth component")
abline(h=0.99,col="red")
abline(v=10,col="red",lty=10)#
mtext("C")

plot(summary(pca.m)$importance[3,], type="l", ylab="%variance", xlab="nth component")
abline(h=0.99,col="red")
abline(v=70,col="red",lty=10)# NB - by changing the data here we are able to visualize the effec
mtext("D")
# Chosen.........QUite a nice thing to know.....by KAgereki at 1:07 AM on 29/10/2013
title("Comparisons of variance contribution", outer=TRUE)

#Variance contributed by the PCs cumulatives
var1<-pca.m$sdev[1]^2/sum(pca.m$sdev^2) #First PC
var2<-pca.m$sdev[1:2]^2/sum(pca.m$sdev^2) # First and second PC
var3<-pca.m$sdev[1:3]^2/sum(pca.m$sdev^2) # Third PC

var10<-sum(pca.m$sdev[1:10]^2/sum(pca.m$sdev^2)) # First 10 PCS
```

```
var20<-sum(pca.m$sdev[1:20]^2/sum(pca.m$sdev^2)) # etc
var30<-sum(pca.m$sdev[1:30]^2/sum(pca.m$sdev^2))
var40<-sum(pca.m$sdev[1:40]^2/sum(pca.m$sdev^2))
var50<-sum(pca.m$sdev[1:50]^2/sum(pca.m$sdev^2))
var60<-sum(pca.m$sdev[1:60]^2/sum(pca.m$sdev^2))
var70<-sum(pca.m$sdev[1:70]^2/sum(pca.m$sdev^2))

# Subsequenly to get the percentage variance contributed we multiply by 100, given below is an e
percvar1<-round((var1*100),2) # Percentage of variance by 1st component
percvar2<-round((var2*100),2) # Percentage of variance by the  first and second PCs
percvar3<-round((var3*100),2) # Percentage of variance by the first three PCs

# To get the variance distribution in the other covariates:
classage<-desc$age
scores2<-pca.m$x
melted <- cbind(classage, melt(scores2[,1:3]))

barplot <- ggplot(data=melted) +
  geom_bar(aes(x=Var1, y=value,fill=classage),stat="identity") +
  facet_wrap(~Var2)
#title("Comparisons of variance contribution", outer=TRUE)

classgender<-desc$gender
scores2<-pca.m$x
melted <- cbind(classgender, melt(scores2[,1:3]))
barplot <- ggplot(data=melted) +
  geom_bar(aes(x=Var1, y=value,fill=classgender),stat="identity") +
  facet_wrap(~Var2)
#title("Comparisons of variance contribution", outer=TRUE)

classdisease<-desc$disease
scores2<-pca.m$x
melted <- cbind(classdisease, melt(scores2[,1:3]))

barplot <- ggplot(data=melted) +
  geom_bar(aes(x=Var1, y=value,fill=classdisease),stat="identity") +
  facet_wrap(~Var2)
#title("Comparisons of variance contribution", outer=TRUE)


# Scree plot
screeplot(pca.m, type="lines",col=3,main="Scree plot")




# To calculate the EigenR-squared
#Here we used an earlier version of R (R version 2.6.2  - Released on 2008-02-08)because the pac
#


library(eigenR2)


mod1 <- model.matrix(pca.m$x~1+desc$age)
```

```
eigenR2.age <- eigenR2(dat = compact.data, model = mod1)
eigenR2.age$eigenR2*100

mod2 <- model.matrix(pca.m$x~1+desc$disease)
eigenR2.disease <- eigenR2(dat = compact.data, model = mod2)
eigenR2.disease$eigenR2*100

mod3 <- model.matrix(pca.m$x~1+desc$gender)
eigenR2.gender <- eigenR2(dat = compact.data, model = mod3)
eigenR2.gender$eigenR2*100
```

Appendix B

R Code (LDA and cross validation)

```
library(MASS)

dathelp=data.frame(compact.data2)
lda1=lda(desc$disease~ . , data=dathelp ,CV=FALSE, method="moment")

#Confusion matrix - Training data used as testing data
summary(lda1)
pred = predict(lda1,dathelp)
names(pred)
table1<-table(desc$disease,pred$class)
table1

# Error rate for the the confusion table above
error1 = sum(table1[row(table1) != col(table1)]) / sum(table1)
error1

# A function to run the LOOCV for the dataset
vlda = function(v,formula,data,cl){
   require(MASS)
   grps = cut(1:nrow(data),v,labels=FALSE)[sample(1:nrow(data))]
   pred = lapply(1:v,function(i,formula,data){
            omit = which(grps == i)
            z = lda(formula,data=data[-omit,])
            predict(z,data[omit,])
            },formula,data)

   wh = unlist(lapply(pred,function(pp)pp$class))
   table(wh,cl[order(grps)])
}

# The sample below shows  the generic code to cross-validate the dataset, with the choice of n #
 #iterated for the 70 PCs
```

```
chosen.components3 <- 1:3
feature.vector <- pca.m$rotation[,chosen.components]
compact.data <- t(feature.vector3) %*% t(data2)
compact.data<-t(compact.data)
dathelp=data.frame(compact.data)
lda=lda(desc$disease~ . , data=dathelp,CV=FALSE, method="moment")
pred= predict(lda,dathelp)
names(pred)
table<-table(desc$disease,pred$class)
disease<-desc$disease
data<-cbind(disease,dathelp)
data<-data.frame(data)
tt = vlda(3,disease~.,data,data$disease)
error = sum(tt[row(tt) != col(tt)]) / sum(tt)

# The errors were tabulated and the hex correlation plot done with the code below:

library(hexbin)
bin<-hexbin(Pc, data, xbins=5) # Where PCs is the Principal components and data is the errors
plot(bin, main="Scatter plot error and components ",xlab="Components",ylab="% Error")
```

Bibliography

[1] P. Tamayo M. Gaasenbeek C. Huard J.P. Mesirov H. Coller M. Loh J.R. Downing M.A. Caligiuri C.D. Bloomfield T.R. Golub, D.K. Slonim and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science,*, (7):531537, October 1999.

[2] Wang H Daggard G Shi M. Hu H, Li J. A maximally diversified multiple decision tree algorithm for micro-array data classification. *Conferences in Research and practice in information technology(CRPIT)*, 73(37), 2006.

[3] Coleman M. Peterson L. Machine learning-based receiver operating characteristics (roc) curves for crisp and fuzzy classification of dna micro-arrays in cancer research. *Int J Approx Reason*, 47(38):17:36, 2008.

[4] Shuangge Ma1 and Michael R. Kosorok2. Gene expression identification of differential gene pathways with principal component analysis. *BIOINFORMATICS*, 25(24):882889, 2009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732304/pdf/btp085.pdf.

[5] Storey JD Chen LS. Eigen-r2 for dissecting variation in highdimensional studies. *Bioinformatics*, 24(25):2260–2262, 2008.

[6] Nih educational website. (31). URL ://publications.nigms.nih.gov/thenewgenetics/chapter1.html.

[7] Ambroise C. and McLachlan G. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, (1):6562–6566., 2002.

[8] Ying Lu Jiawei Han. Cancer classification using gene expression data;. *Department of Computer Science University of Illinois at Urbana-Champaign Urbana, IL 61801, USA*, (9).

[9] D. K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32(18):502–8, 2002.

[10] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology,pp. research0032.1-research0032*, (19):11, 2001.

[11] J. T. Leek R. G. Tompkins J. D. Storey, W. Xiao and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20):12837–42, 2005.

[12] T. J. Giordano R. Kuick D. E. Misek G. Rennert D. R. Schwartz S. B. Gruber C. Logsdon D. Simeone S. L. Kardia J. K. Greenson K. R. Cho D. G. Beer E. R. Fearon K. A. Shedden, J. M. Taylor and S. Hanash. Accurate molecular classification of human cancers based on gene expression using a 156 simple classifier with a pathological tree-based framework. *American Journal of Pathology*, 163(21), 2003.

[13] A. Regev D. Pe'er D. Botstein D. Koller E. Segal, M. Shapira and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(22):166–177, 2003.

[14] Aharonov R. Meiri E. Rosewalt S. Rosenfeld, N. and Y. Spector. Micrornas accurately identify cancer tissue origin. *Nature Biotechnology*, 26(4)(2): 462469, 2008.

[15] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence,*, 97(12)(12), 1997.

[16] H. Liu and editors Motoda, H. Feature extraction, construction and selection: A data mining perspective. *Kluwer Academic Publishers.*, (4), 1998.

[17] Moler E. Chow, M. and I. Mian. Identifying marker genes in transcription profile data using a mixture of feature relevance experts. *Physiol. Genomics,*, 5(5):99–111, 2001.

[18] Lavrac N. Zelezny F. Gamberger, D. and J. Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37(17), 2004.

[19] Tibshirani R. Hastie, T. and J. Friedman. The elements of statistical learning, data mining, inference and prediction. *Berlin: Springer.*, (14), 2001.

[20] Kela I. Getz G. Givol D. Ein-Dor, L. and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics,*, 21(2)(15), 2005.

[21] P. Domingos. The role of occams razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(16):409425., 1999.

[22] Wang and van der Laan. Machine learning-based receiver operating characteristics (roc) curves for crisp and fuzzy classification of dna micro-arrays in cancer research. *BMC Bioinformatics*, 12(39):312, 2011. URL http://www.biomedcentral.com/1471-2105/12/312.

[23] J. Fridlyand S. Dudoit and T. P. Speed. Comparison of discrimination methods for the classification of tumours using gene expression data. *Journal of the American Statistical Association,*, 97(457):(10):77:87, 2002.

[24] Jolliffe I. Principal component analysis. *Springer, New York*, (51), 2002.

[25] W. Landgrebe, J. Wurst and G. Welzl. 3. *Genome Biol.*, RESEARCH0019 (41), 2002.

[26] J. et al. Khan. 7. *Nat. Med.*, (42):673679, 2001.

[27] B. Dolenko R. L. Somorjai and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, (43):19(12):14841491,, 2003.

[28] R.B. et al. Brem. Genetic dissection of transcriptional regulation in budding yeast. Science(32):752–755, 2002.

[29] Zhang C. Li, T. and M. Ogihara. M., a comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression,. *Bioinformatics,*, 20(15):24292437,(33), 2004.

[30] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification,. *IEEE Trans. Pattern Anal. Mach. Intell.,*, 18(6):607615,(34), 1996.

[31] Satoru Kuhara Satoshi Niijim. Eective nearest neighbor methods for multiclass cancer classiffication using microarray data. (36), http://www.jsbi.org/pdfs/journal1/GIW05/GIW05P051.pdf.

[32] Abdel Moniem NK. Sweilam NH, Tharwat AA. Support vector machine for diagnosis cancer diseases: a comparative study. *Egyptian Inform*, (11).

[33] J. Mesirov T. Golub D. Slonim, P. Tamayo and E. Lander. Class prediction and discovery using gene expression data. *In Proc. 4th Int. Conf. on Computational Molecular Biology(RECOMB),*, (8):263272, 2000.

[34] Jackson J. A user's guide to principal components. *New York, John Wiley and Sons*, (50), 1991.

[35] Jackson D. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, 74(8)(52):2204–2214, 1993.

[36] W. et al. Jin. The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nat. Gen.*, 29(66):389395., 2001.

[37] Wu F Liu F Ye X et al Xu Q, Ni S. Investigation of variation in gene expression profiling of human blood by extended principle component analysis. *PLoS ONE*, (23):6(10): e26905. doi:10.1371/journal.pone.0026905, 2011. URL http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0026905&representation=PDF.

[38] Sampath Deegalla and Henrik Boström. Classification of microarrays with knn: Comparison of dimensionality reduction methods. (60). URL http://su.diva-portal.org/smash/get/diva2:305374/FULLTEXT01.