CIMAT

Centro de Investigación en Matemáticas, A.C.

PROBLEMS IN STATISTICAL GENETICS: CLASSIFICATION AND TESTING FOR NETWORK CHANGES

T E S I S

Que para obtener el grado de

Doctor en Ciencias

con Orientación en

Probabilidad y Estadística

Presenta

Adolphus Wagala

Directores de Tesis:

Dra. Graciela González Farías

y

Prof. Terence Paul Speed

Guanajuato, Gto., 07 de marzo de 2018

CIMAT **Centro de Investigación en Matemáticas, A.C.**

PROBLEMS IN STATISTICAL GENETICS: CLASSIFICATION AND TESTING FOR NETWORK CHANGES

T E S I S

Que para obtener el grado de

Doctor en Ciencias

con Orientación en

Probabilidad y Estadística

Presenta

Adolphus Wagala

Directores de Tesis:

Dra. Graciela González Farías

y

Prof. Terence Paul Speed

Autorización de la versión final

Guanajuato, Gto., 07 de marzo de 2018

CIMAT

Centro de Investigación en Matemáticas, A.C.

Problems in Statistical Genetics: Classification and Testing for Network Changes

Tesis

Que para obtener el Grado de:

Doctorado en Ciencias con Orientación en Probabilidad y Estadística

P R E S E N T A:

Adolphus Wagala

Directores de Tesis:

Dra. Graciela González Farías

y

Prof. Terence P. Speed

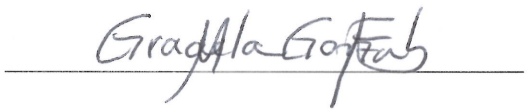
Guanajuato, Guanajuato, México

07 Marzo de 2018

Integrantes del Jurado

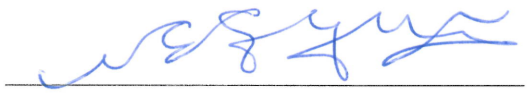
Presidente: Dr. Johan Jozef Lode Van Horebeek
CIMAT
Secretario: Dra. Lilia Leticia Ramírez Ramírez
CIMAT
Vocal: Dr. José Ulises Márquez Urbina
CIMAT-CONACYT
Vocal: Dr. Gabriel Arcángel Rodríguez Yam
Universidad Autónoma de Chapingo
Director de tesis: Dra. Graciela María de los Dolores González Farías
CIMAT
Lector especial: Prof. Terence Paul Speed
Walter and Eliza Hall Institute, Melbourne, VIC, Australia

Asesor:



Dra. Graciela María González Farías

Sustentante:



M.C. Adolphus Wagala

To my dad William,

to my late mum Abigail (R.I.P),

to Abbie,

to William.

Acknowledgements

First and foremost I would like to thank the Almighty God for having given me good health and sane mind as I pursued and completed my doctorate studies.

I am very grateful to my advisor Dr. Graciela Gonzalez Farías for having accepted to mentor me, for the patience and guidance throughout my studies. It was an honour to study under your guidance.

To Prof. Terence Speed, thanks for accepting to be my co-advisor, mentoring and guidance throughout the period I was working with you. Thanks for hosting me at WEHI/University of Melbourne during my academic visit to the Speed Lab. My appreciation to Prof. Ian Wicks of the Inflammation Division, WEHI for having incorporated me in the ARF project from where I learnt a lot and formed a motivation of my second problem for the thesis. I extend my appreciation to the members of the WEHI, Bioinformatics Division from whom I learnt a lot and also made my stay at WEHI to very enjoyable. Special thanks to Jo Keeble who had a lot of patience to take me through lots of biology and who was always ready to assist whenever I needed help. Thank you to Willy-Johns and Jo for the discussions and numerous lunches during my stay at WEHI.

Lots of gratitude to my PhD committee members (Drs. Johan, Gabriel, Ulises and Leticia) who provided lots of suggestions that tremendously improved the thesis.

I also appreciate my CIMAT professors with a special mention of Dr. Rogelio Ramos and Dr. Oscar Dalmau for their assistance with various aspects of the thesis.

To my family back in Kenya especially my dad William, I am grateful for your support and encouragement when I was in these turbulent waters of PhD.

Thanks to the Secretaría de Relaciones Exteriores (SRE) for the first year scholarship and Consejo Nacional de Ciencias y Tecnología (CONACyT) for providing scholarship (Numero de beca 384101) for the remainder of the period of my studies. Further financial support was received from the proyecto de CONACyT CB 252996. Muchas gracias.

Finally, to CIMAT for providing a conducive atmosphere, facilities and partial financial support during my studies. I appreciate.

Abstract

This thesis addresses the problems of classification of microarray data and the statistical integration of molecular data to test for network changes. For the classification problem, we consider the unpreprocessed and preprocessed microarray data sets. We implement an extension of the partial least squares generalized linear regression (PLSGLR) Bastien et al. (2005) achieved by combining it with the logistic regression to get partial least squares generalized linear regression-logistic regression model (PLSGLR-log) and also with the linear discriminant analysis to get the partial least squares generalized linear regression-linear discriminant analysis denoted by (PLSGLRDA). These two classification methodologies are then compared with the classical methodologies namely the k-nearest neighbours (KNN), linear discriminant analysis (LDA), partial least squares discriminant analysis (PLSDA), ridge partial least squares (RPLS), the support vector machine (SVM). Furthermore, we implement a recent algorithm by Dalmau et al. (2015) known as kernel multilogit algorithm (KMA). The results indicate that for the noisy unpreprocessed data, the KMA emerged as the clear "winner" based on based on their low misclassification error rates. For the preprocessed normalized data, there was no clear "winner" since there was no single method that performed outstandingly better than the rest. The KNN emerged as a clear "loser" since it consistently had a relatively higher rate of misclassification both when applied to the un-preprocessed and preprocessed data sets.

The statistical integration of molecular data to test for network changes considers an experiment involving two main groups namely the healthy (H) and acute rheumatic fever (ARF) subjects. For each group, each specimen is divided in two portions so that one portion is *group A streptococcus* (GAS) stimulated while the other is unstimulated so that we end up with four sub groups: Healthy GAS stimulated, Healthy unstimulated, ARF-GAS stimulated and ARF unstimulated. As a result, we have dependence within the groups and independence between the groups. For all the groups, p genes are measured for expression. We identify a prior network from the curated literature and online sources. The genes considered in the experiment are then matched with the ones in the prior network so that we reduce the prior network to only the genes that are found in the experimental data. We then construct two networks, one for the healthy and the other one for the ARF. The nodes are coloured based the log fold changes to indicate the genes

that remain unchanged, up or down regulated. A group or cluster of genes that constitute a certain important functional group or that have some known interactions are identified and their sub networks extracted from both groups. A likelihood ratio test statistic for testing for network changes assumption of variance-covariance matrix is unknown is therefore developed. A simulation study is done to demonstrate the applications of the developed statistics. The experiments confirm that the test statistics follow a chi-square distribution. This research contributes a theoretical analysis motivated by a practical problem for which no formal statistical method is in use.

Table of contents

1 Introduction **1**

1.1 Introduction to the classification problem 1

1.2 Statistical integration of molecular data 3

1.3 Preliminaries of genetics 5

1.3.1 Microarray technology 7

1.3.2 Next generation sequencing (NGS) technology 10

1.4 Multiple hypothesis testing in the high-dimensional data 11

1.5 Outline of this thesis 13

2 Methods for classification **15**

2.1 Introduction 15

2.2 Basics of classification 16

2.3 Partial least squares (PLS) and some of its applications in genomics . 26

2.4 PLS regression (PLSR) algorithm 27

2.5 PLS generalized linear regression algorithm 31

3 Applications to real data sets **33**

3.1 Introduction 33

3.2 Analysis of the unpreprocessed data sets 33

3.2.1 Some exploratory analysis 34

3.2.2 Results from the analysis of the unpreprocessed data 38

3.2.3 Feature selection 42

3.2.4 Analysis of the preprocessed data sets 42

3.2.5 Results and discussions for analysis of preprocessed data . . . 43

3.3 Summary on classification methods 51

4 Statistical integration of molecular data to test for network changes 53
4.1 Introduction 53
4.2 Example of a pilot laboratory experiment data 55
4.3 RNA-seq data 56
4.4 Statistical integration of molecular data 58
4.5 The likelihood ratio testing 63
4.5.1 Likelihood ratio test for network changes 65
4.6 Simulation study for the multivariate problem 70
4.7 Some special cases 82
4.7.1 One gene problem 82
4.7.2 Simulation study of the one gene problem (p=1) 85
4.8 Summary for the testing for network changes 91

5 Summary, conclusions and future work 93
5.1 Summary and conclusions 93
5.2 Future work 95

List of Figures

1.1 **Structure of the DNA** (**source:** Wikipedia). The DNA consists of two strands, each with a linear backbone of alternating sugar (deoxyribose) and phosphate residues. Four bases namely, adenine (A), guanine (G), cytosine (C) and thymine (T) are covalently attached to the backbone. The two strands of the DNA are connected by hydrogen bonds between two complementary opposing bases, that is thymine (T) connects with adenine (A) while cytosine (C) connects withs guanine (G) so that the resulting DNA resembles a ladder commonly described as a double helix. 6

3.1 **Box plot for the non-preprocessed colon data**. The box plot for unpreprocessed data clearly shows that the data is noisy and has a lot of variations. The data has some unwanted variations that are expected to affect its analysis. It also lacks symmetry. 34

3.2 **Box plot for the preprocessed colon data**. This plot looks reasonable with less variations. The data seem to have a symmetric distribution and does not show the presence of unwanted variation. From the two figures, it is expected that the preprocessed data would be easy to analyze. 35

3.3 **RLE plots for the non-preprocessed and pre-processed colon data.** The RLE plot for the unprocessed data shows the presence of a lot of heterogeneity which reveals that the data has variations that do not necessarily come from the biological factors. However, the RLE plot for the processed data shows homogeneity and lack of unwanted noise and should give relatively good results when analyzed statistically. 36

3.4 **PCA plot for the nonpreprocessed Colon data.** The PCA plots show that the it is harder to separate/classify the unpreprocessed data. 37
3.5 **PCA plots for the preprocessed Colon data.** It is relatively easier to separate/classify compared preprocessed data. 38
3.6 **Box plots for the error rates for the unpreprocessed Colon data**. The errors for all the classifiers are not symmetric except the SVM. The boxplots confirm that the top best classifiers are the KMA, PLSDA and RPLS. The KNN is outstandingly performing poor. . . 40
3.7 **Box plots for the error rates for the unpreprocessed Leukemia data**. This set of data had a relatively lower rate of missclassification. The best classifiers are the KMA, PLSDA, RPLS and PLSGLRDA. The KNN remains consistent in it's poor performance. 41
3.8 **Box plots for the error rates for the unpreprocessed Prostate data**. The best methodologies remain the KMA, PLSDA, RPLS and SVM. The KNN remains the worst. 41
3.9 **Box plots for the Test Errors for the Colon Data.** The box plot shows that the distribution of errors for PLSDA and RPLS are the same for all the number of genes selected. The distributions are symmetric with the mean below 0.2. The LDA has a higher error rate for $p = 50$ while PLSGLRDA has higher error rates for $p = \{500, 1000\}$. 48
3.10 **Box plots for the Test Errors for the Leukemia Data.** For the leukemia data, most of the classifiers have a low miss-classification rate. The RPLS and SVM have very low error rates for $p = \{300, 500, 1000\}$. 49
3.11 B**ox plots for the Test Errors for the Prostate Data.** The outstanding bars are the ones for KNN which seem have a relatively higher rate of mis-classification and this rate increases with the increase in the number of genes p selected. In general, the test error rate for this data set for all the algorithm have many outliers. 50

4.1 **Sub network 1 consisting mainly of the functional group Th1/Th17**. The choice of the genes used in this subnetwork 1 is based on other experiments carried out in the Wicks Lab that led to the hypothesis that pathogenic Th1/Th17 T cells are key mediators of the heart inflammation and damage in ARF. The edge information is obtained from the PINA network. The nodes (genes) are coloured in such a way that the yellow nodes represent upregulated genes, light blue nodes are downregulated genes (or groups of genes) while the white ones are neither up nor down regulated. We aim to develop a statistical framework to test for the changes in the similar genes for each of the subnetwork for healthy and ARF subjects. For example, in the healthy subjects network, the gene LCN2 is neither up or down regulated while in the ARF subjects the same gene is upregulated. Furthermore, in the healthy PBMCs, the gene RELA is upregulated while in the ARF patients PBMCs the same gene's average expression level is unchanged. It is these kind of changes we refer to as changes and we want to develop a statistical framework to test them. 62

4.2 **Sub network 2 for functional group Th2**. Bhatnagara et al. (1999) also found out that chronic rheumatic heart disease (CRHD) patients secreted IL-4 and IL-10 in large amounts, i.e. Th2 type of cytokine profile. The genes in this subnetwork would help determine the changes of the Th2 group for the ARF patients and healthy subjects. The edge list is obtained from the prior network (PINA). The nodes (genes) are coloured in such a way that the yellow nodes are upregulated, light blue nodes are downregulated while the white ones are neither up nor down regulated. The aims and objectives for this figure are similar to those discussed for Figure 4.1. 63

4.3 **Histograms for the simulated data for p=2 and p=5 when Σ is known.** 73

4.4 **Histograms for the simulated data for p=8 and p=15 when Σ is known.** 74

4.5 **Histograms for the simulated data for p=2 and p=5 when Σ is unknown.** 75

4.6 **Histograms for the simulated data for p=8 and p=15 when Σ is unknown.** 76
4.7 **Histograms for the simulated data for p=20, m=200 & k=190 when Σ is known.** 78
4.8 **Histograms for the simulated data for p=20, m=300 & k=350 when Σ is known.** The histograms in Figures 4.7 and 4.8 exhibit the same properties as the ones discussed in Figures 4.5 and 4.6. . . . 79
4.9 **Histograms for the test statistic computed from resampling the simulated data for p=20 and Σ is unknown.** This set of histograms exhibit the same properties as the ones already discussed in 4.5 and 4.6. 80
4.10 **Histograms for the test statistic computed from resampling the simulated data for p=20 and Σ is unknown.** This set of histograms exhibit the same properties as the ones discussed in the previous figures. 81
4.11 **Histograms for the test statistics computed by resampling the simulated data for p=1, different values of m and k when σ is known.** 87
4.12 **Histograms for the test statistics computed by resampling the simulated data for p=1, different values of m and k when σ is known.** 88
4.13 **Histograms for the test statistics computed by resampling the simulated for p=1, different values of m and k when σ is unknown.** 89
4.14 **Histograms for the test statistics computed by resampling the simulated for p=1, different values of m and k when σ is unknown.** 90

List of Tables

1.1 Proposed strategy 3
1.2 Errors committed when testing n hypotheses 11

2.1 PLS Regression Algorithm 30

3.1 Percentage missclassification for the different methods when applied to the unpreprocessed data sets 39
3.2 Percentage Misclassification for the Colon Data Set 44
3.3 The proportions of False Positives and False Negatives for the Colon Data 44
3.4 Percentage Misclassification for the Leukemia Data Set 45
3.5 Proportions for types of misclassification for the Leukemia Data . . . 46
3.6 Percentage Misclassification for the Prostate Data Set 46
3.7 The proportions False Positives and False Negatives for the Prostate Data 47

4.1 Major and minor Jones criteria for the diagnosis of acute rheumatic fever 54
4.2 Illustration of a table of read counts for the two categories of the samples and groups. 56

4.3 Calculated test statistics when Σ is known 71
4.4 Calculated test statistics when Σ is unknown 71
4.5 Calculated test statistics and p-values when Σ is known for different values of m and k. 77

4.6 Calculated test statistics and p-values when Σ is unknown for different values of m and k. 77
4.7 Calculated test statistics and p-values when Σ is known for different values of m and k for the one gene problem. 85
4.8 Calculated test statistics and p-values when Σ is unknown for different values of m and k for the one gene problem. 86

Chapter 1

Introduction

This dissertation looks into two major themes in the biological data analysis namely the classification and statistical integration of molecular data to test for network changes. Furthermore, we give a brief review of the developments in the field of genomics as regards to data generation and analysis.

1.1 Introduction to the classification problem

The field of genomics has witnessed a tremendous increase in the data generation due to biotechnological advances like the microarrays and the next-generation sequencing platforms. These biotechnological advances have made it possible to simultaneously monitor expression levels in cells for thousands of genes and thus help in solving particular problems related to the identification of molecular variations in genes, classification, diagnosis, prognosis and treatment. The high dimensional data generated from microarray technology involve many thousands of genes measured simultaneously using several microarrays, that is, different microarray for each individual. This definitely introduces some noise and unwanted variations that might be from technical or unknown sources.

In a microarray experiment let n and p be the numbers of the samples and genes respectively so that the generated data is a $n \times p$ matrix. The main challenge with these technologies is that the resultant data generated is noisy due to biological and technological variations and at the same time usually have more variables (high dimension) but low sample size (few samples), that is, $n << p$. This condition $n << p$ makes the direct application of most classical statistical methodology

implausible and so there have been attempts to find a solution to this problem by different researchers.

Normally, before the down stream analysis of the data generated from DNA microarray, preprocessing and normalization is done to it so as to remove the noise, filtering out the genes with low expression values, missing values are addressed and the data is standardized via log-transformation. One of the most used preprocessing procedure for the microarray data is the one one proposed by Dudoit et al. (2002) which entails three basic steps namely thresholding, filtering out of genes with a given minimum/maximum intensities and finally, standardization of the expression values by taking log transformation Alshamlan et al. (2013); Dudoit et al. (2002).

In this disseration, the classification problems for microarray data sets are considered under two conditions, namely the un-preprocessed and the preprocessed one. In the un-preprocessed data, we use all the genes in the study while in the preprocessed one, only the subset of genes believed to play important role towards the biological problem of interest are used. We extend the Partial Least Squares Generalized Linear Regression (PLSGLR) algorithm Bastien et al. (2005) by combining it with the logistic regression (PLSGLR-log) and also with the Linear Discriminant Analysis to come up with (PLSGLRDA). Furthermore we compare their performance with those of kernel multilogit algorithm (KMA) proposed by (Dalmau et al., 2015) and the classical methods namely, the k-Nearest Neighbour (KNN), Ridge Partial Least Squares (RPLS), Partial Least Squares-Linear Discriminant Analysis (PLSDA), the usual Linear Discriminant Analysis (LDA) and the Support Vector Machines (SVM) when applied to three sets of microarray data, namely the Colon (Alon et al., 1999), Leukemia (Golub et al., 1999) and the Prostate (Singh et al., 2002) data sets. We evaluate the classifiers not only on the misclassification rates but also on the proportion of false negatives percentages attributed to each of them for the data sets considered.

In many studies involving classification problems in microarrays with higher dimensional data ad lower number of samples (or subjects), the two stage strategy has been used for example by (Nguyen and Rocke, 2002a,b). It is worth noting that,most studies including those by Nguyen and Rocke (2002a,b) involve the use the original PLS to build the components even though the response variables are discrete. This is intuitively not correct since the original PLS is an algorithm best suited for the continuous response variables. Secondly, in most of the procedures,

a variable (gene) selection step is involved while in our case, we study the models with and without the gene selection step in order to evaluate the performance of each classifier. Furthermore the original PLS used can not handle the missing values unlike the PLSGLR.

Therefore, we propose to use the two stage strategy in solving the classification problem as follows.

Table 1.1: Proposed strategy

Steps
Step 1: Dimension reduction In this stage, we propose to use the PLSGLR to project the high dimensional data to a low dimension space thus resulting in new components (latent variables) which have information about the intrinsic structure of the data. Use the algorithm presented in Section 2.5.
Step 2: Use of latent variables for classification to use the obtained latent variables with a lower dimension with the classical statistical classifiers: (i) PLSGLR components with logistics regression to get the PLSGLR-logistic model denoted as (PLSGLR-log) (ii) PLSGLR components with linear discriminant analysis model to get PLSGLR-Linear Discriminant Analysis model denoted as (PLS-GLRDA)

To the best of our knowledge, the proposed combination of PLS generalized linear regression algorithm with logistic and discriminant analysis have not been so far used in the cases where $n << p$ to evaluate its effectiveness in the classification problems. The PLS generalized linear regression algorithm is simple and a good performance compared to the classical methods would make it an attractive alternative.

1.2 Statistical integration of molecular data

The second major problem addressed in this thesis is the statistical integration of molecular in order to test the network changes. This study is motivated by the on-

going research on acute rheumatic fever (ARF) at the Speed and Wicks Labs at the Walter & Eliza Hall Institute of Medical Research, Melbourne, VIC. The research in these labs are geared towards understanding the type of inflammation occurring in ARF patients with the ultimate goal being to find new diagnostic markers to diagnose ARF and new drugs. The ARF still remains a major challenge to the developing countries and to the poor people of the developed countries living in poor unhygenical conditions e.g the Aboriginal and Torres Islanders of New Zealand and Australia.

The key idea in this part of the thesis is to integrate a known network (also referred to as the prior network), usually the protein-protein interaction (PPI) network is used with the experimental molecular data. Specifically, we consider an experiment in two groups of subjects namely the healthy control (HC) and the ARF subjects are involved. Each sample is further divided into two sub-samples whereby one sub-sample is Group A *Streptococcus* (GAS) stimulated while the other sub-sample is unstimulated. As a result, we now have paired samples for the HC and also another paired samples for the ARF, resulting into four different subgroups. The samples from all these groups are then sequenced to measure the expression levels for the p genes under consideration. The genes whose expression levels are measured are the same for all the subgroups. It is expected that the GAS stimulation of HC and ARF subjects will help in understanding how the GAS affects the HC and ARF subjects thereby possibly able to identify the biomarkers associated with the ARF. Assume that the sample sizes for HC and ARF are m and k respectively and that p genes are considered in the experiment. The paired measurements are correlated within subjects, but independent between subjects, as well as being independent between HC and ARF group. Furthermore, since the genes usually act in a group, the p genes are expected to be correlated. The fact that genes interact with each other can be captured through a prior network like the PPI network. The network is obtained from the online and curated literature sources. Now, we integrate this prior network with the experimental molecular data so that we get two networks, one for the HC and another one for the ARF. To integrate the experimental data, we get the edge list from the prior network and use the obtained list of edges for constructing networks for HC and ARF. The nodes for the constructed networks are coloured to reflect the changes in the genes with regards to up or down regulations or no change. For the two networks, we expect that with the GAS treatment, some

genes will be upregulated, others downregulated while some will remain the same for different network. These changes are easily measured using the log fold change (logFC).

The aim is to develop a statistical framework for comparing the changes in expression levels in the different sets of genes between the two networks. These genes are selected from the different sub-networks that are believed to be important functional groups in the ARF disease. We use the well known likelihood ratio theory to formally derive a new test for measuring the difference in the differences of the mean expression levels for the healthy and ARF subjects in the context of network changes.

Even though the motivation for the test statistic developed in this part of the thesis is from the research on ARF, this kind of study is very useful in other general situations involving hypothesis testing for the differences of means across groups where the measurements within the subjects are dependent, while the subjects and groups are independent.

1.3 Preliminaries of genetics

All living organisms consist of cells which in turn consist of molecules. Every human cell except the red blood cells contain a nucleus which has chromosomes that carry the individual's genetic information. These chromosomes contain the deoxyribonucleic acid (DNA) and proteins. The DNA is the main information carrier molecule required for the development and functioning of an organism. The DNA consists of two strands, each with a linear backbone of alternating sugar (deoxyribose) and phosphate residues. Four bases namely, adenine (A), guanine (G), cytosine (C) and thymine (T) are covalently attached to the backbone. The two strands of the DNA are connected by hydrogen bonds between two complementary opposing bases, that is thymine (T) connects with adenine (A) while cytosine (C) connects with guanine (G) so that the resulting DNA resembles a ladder commonly described as a double helix. Therefore the two DNA fragments only differ with respect to the arrangement of the bases Ziegler and König (2008). The basic building blocks in the nucleic acids are known as nucleotides consisting of a phosphates, pentose sugar and a heterocyclic amine. The order in which the bases occur determines the information for protein synthesis which is basically a two-step process including, first the transcription in

which information is read from the sequence of bases to make amino acids and ribonucleic acid (RNA). The second step is the translation by which the RNA form proteins (Mitra et al., 2014).

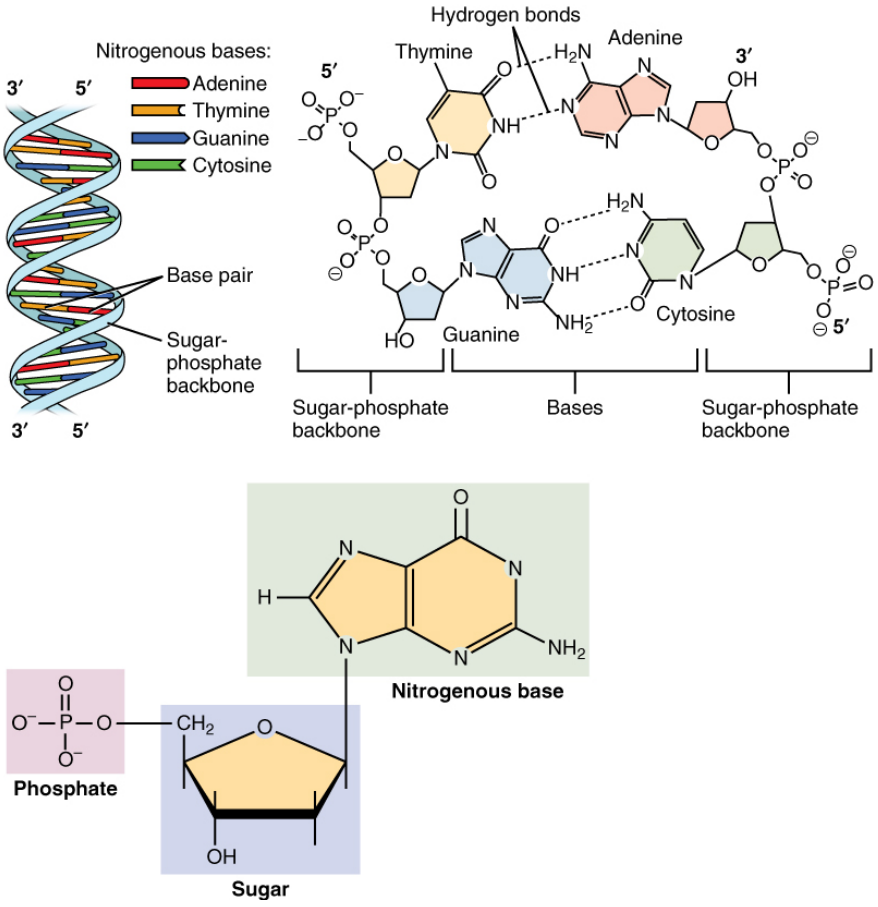


Figure 1.1: Structure of the DNA (**source:** Wikipedia). The DNA consists of two strands, each with a linear backbone of alternating sugar (deoxyribose) and phosphate residues. Four bases namely, adenine (A), guanine (G), cytosine (C) and thymine (T) are covalently attached to the backbone. The two strands of the DNA are connected by hydrogen bonds between two complementary opposing bases, that is thymine (T) connects with adenine (A) while cytosine (C) connects withs guanine (G) so that the resulting DNA resembles a ladder commonly described as a double helix.

The RNA is smaller and much shorter than the DNA. It is constructed like the DNA but has the following major differences: the RNA is a single strand, its sugar component is composed of ribose instead of deoxyribose and also it has uracil (U) instead of thymine (T) that exist in the DNA.

The basic physical and functional unit of heredity in humans are the genes, which

are made up of DNA. More importantly, the DNA encodes the genes that in turn produce compounds of amino acids known as proteins which are essential elements in most of the cellular functions like biochemical reactions, cell signalling, metabolism, cell cycle and immune responses. This transfer of biological information from the DNA to proteins is usually referred to as the central dogma of molecular biology and it involves two major steps, the first being *transcription*; a process in which the information encoded in DNA is transcribed by a polymerase into ribonucleic acid (RNA) and the second step is *translation*, a step in which the RNA is synthesized into proteins by ribosomes (Babu, 2004; Ziegler and König, 2008).

Despite the fact that all the cells in the human body contain identical genetic material, the same genes are not active in every cell therefore studying the active and inactive genes in different cell types helps scientists to understand both how these cells function normally and how they are affected when various genes do not perform properly. For instance, all the cells of the human body contain the same DNA, yet there are hundreds of different types of cells, each expressing a unique configuration of genes from the DNA. Previously, it was only possible to conduct these genetic analyses on a few genes at once. However, with the invention of microarrays and various sequencing technologies, it is now possible to examine thousands of genes simultaneously. The DNA encodes genes and regulatory elements control whether genes are on or off (Ziegler and König, 2008).

The gene expression measurement involves the measuring of the abundance of the RNA transcripts in order to find out some aspects of the cell function. This approach of measuring the levels of RNA transcripts is cheaper in the high-throughput technology than measuring the protein levels in the translation stage Brazma and Vilo (2000). Two main technologies that have been extensively used for gene expression measurement among others include the DNA microarray and the next generation sequencing technologies.

1.3.1 Microarray technology

A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). It consists of ordered probes which include nucleic acids, proteins, carbohydrates, tissues, cells and polymers which are to be investigated in general to detect a biological target (Ventimiglia and Petralia, 2013; Babu, 2004).

Two types of microarrays that are popular among the others include;

- **Spotted or cDNA microarray**: Uses complementary copy of the original DNA and each probe represents one gene. In this technology the probes are synthesized apart and printed mechanically on the slide. This is also referred to as the two colour array.
- **Oligonucleotide chips (Affymetrix)**: In this case, the probes are directly synthesized on the surface. The sythesis process allows to create only small fragments so that a gene is not represented by one probe but a set of probes. This technology uses only one sample per chip, it simplifies the experiment and is much more sensitive.

To carry out microarray sequencing, a gene is activated thereby igniting cellular machinery to copy certain segments of that gene resulting in a product is known as messenger RNA (mRNA), a body's template responsible for protein creation. The mRNA binds to the original portion of the DNA strand from which it was copied due to the fact that it is complementary. The genes which are turned on and the ones turned off are identified by first collecting the mRNA molecules in a particular cell and then each mRNA is labelled by using a reverse transcriptase enzyme (RT) to generate a complementary cDNA from the mRNA. The cDNA refers to an mRNA transcript's sequence, expressed as DNA bases (GCAT) rather than RNA bases (GCAU). The fluorescent nucleotides are attached to the cDNA during this process. Next is to label the samples for example if they are tumor or normal with different fluorescent dyes and then onto a DNA microarray slide. The labeled cDNAs will hybridize to their synthetic complementary DNAs attached on the microarray slide, leaving its fluorescent tag. **Hybridization** is the process of combining two complementary single-stranded DNA or RNA molecules and allowing them to form a single double-stranded molecule. A special scanner is then used to measure the fluorescent intensity for each spot/areas on the microarray slide. A very active gene will produce more messenger RNA, thus, more labeled cDNAs, which hybridize to the DNA on the microarray slide to generate a very bright fluorescent area. The less active ones will produce fewer mRNAs, thereby less labeled cDNAs, resulting to fluorescent spots with low intensity. If a gene is inactive then there will be no fluorescence, implying that none of the messenger molecules have hybridized to the DNA. As an example, when co-hybridizing tumor samples (Red Dye) and

normal sample (green dye) together compete for the synthetic complementary DNAs on the microarray slide. Consequently, if the spot is red, then that specific gene is more expressed in tumor than in normal (upregulated in cancer). A green spot means that the gene under consideration is more expressed in the normal tissue (downregulated in cancer). On the other hand, a yellow spot shows that a gene is expressed in both normal and tumor (see, National Human Genome Institute on `https://www.genome.gov/10000533/dna-microarray-technology/`).

A microarray experiment results in an image of the microarray, in which each spot that corresponds to a gene has an associated fluorescence value representing the relative expression level of that gene across two conditions.The gene expressions data obtained from the experiment are then processed using three main stages, namely; image processing, transformation and normalization.

Image processing involves the identification of the spots and distinguishing them from spurious signals. Thereafter, the determination of the local region to estimate background hybridization follows. Summary statistics are then reported followed by assigning spot intensity after subtracting for background intensity. A very active gene produces many molecules of mRNA, thus, more labeled cDNAs, which hybridize to the DNA on the microarray slide and generate a very bright fluorescent area. On the other hand, genes that are less active produce fewer mRNAs, thus, less labeled cDNAs, which results in dimmer fluorescent spots. If there is no fluorescence, no mRNA molecules have hybridized to the DNA, indicating that the gene is inactive. In this manner, the activity of various genes at different times are examined, see (Babu, 2004).

The relative expression level for the cDNA microarray a gene is measured as the amount of red or green light emitted after excitation. The most common metric used in microarray data analysis is called expression ratio. The expression ratio is a relevant way of representing expression differences in a very intuitive manner. For example, genes that do not differ in their expression level will have an expression ratio of 1. Thus up-regulation is blown up and mapped between 1 and infinity, whereas down-regulation is compressed and mapped between 0 and 1 (Babu, 2004). A logarithmic to base 2 is often used for data transformation and is considered by some researchers to be a better alternative to the ratio. On the other hand, the Affymetrix chips represent each gene as a set of probes corresponding each one to one short oligonucleitotide chain. Each probe is a probe pair consisting of perfect

match (PM) probe that corresponds to the original DNA and a miss-match (MM) probe whose central nucleotide has been changed. The key idea here is that anything that hybridizes with the miss-match probe does not represent the real expression but anything else that is background see (Babu, 2004).

Data normalization is a term that is used to describe the process of eliminating various variations to allow appropriate comparison of data obtained from the two samples. Once the data has been preprocessed it can then be represented in the form of a matrix, (called gene expression matrix) that contains rows representing genes and columns representing particular sample. Each entry is a value, given in arbitrary units, that reflects the expression level of a gene under a corresponding sample (Babu, 2004).

1.3.2 Next generation sequencing (NGS) technology

The NGS technology has revolutionized research in the fields of computational biology, pharmacology, medicine and many other fields of research involving the molecular biology and computations. The NGS technology can be classified into two main categories namely the high-end and bench-top platforms. The former being more expensive, bulky instrument, higher cost of setup and offers long reads (for example, Illumina-HiSeq) and therefore more appropriate large sequencing centers and core facilities while the latter (for example Ion PGM, MiSeq) are less costly and more suitable for microbial applications (Mitra et al., 2014; van Dijk et al., 2014).

NGS platforms have three steps module namely *library preparation* in which the genomic DNA is extracted and purified, *polymerase chain reaction (PCR) amplification* which involves the cloning of the DNA molecules in the library, preparing them for the final step which is *sequencing* where the base pairs are read. The “reads" are the final products of all the next generation sequencing platforms.

Once the reads are obtained, it is usually necessary to align and merge fragments from a longer DNA sequence in order to reconstruct the original genome, a process referred to as genome assembly. This process heavily depends on the computer algorithms. For more details see Metzker, M. L. (2010); Mitra et al. (2014); van Dijk et al. (2014).

1.4 Multiple hypothesis testing in the high-dimensional data

In many classical hypothesis testing problems at α level, the probability of observing at least one significant hypothesis by chance when $\alpha = 0.05$ for one hypothesis is just 0.05 which is low and reasonable. However, as the number of hypotheses to be tested increase, the probability of committing type 1 error or of finding a significant hypothesis by chance also increases. As an illustration, if we have 100 hypotheses tested simultaneously at $\alpha = 0.05$ then by assuming independence the probability of finding at least one significant hypothesis by chance is given by

$$\begin{aligned}\mathbb{P}(\text{At least 1 significant hypothesis}) &= 1 - \mathbb{P}(\text{no significant results}) \\ &= 1 - (1 - 0.05)^{100} \\ &= 0.9940795.\end{aligned}$$

The illustration in Table 1.2 seeks to highlight the problem of multiple hypothesis testing for high dimensional data that are nowadays the order of the day due to advances in technology in the fields of biology, chemometrics and many other fields. The probability of getting a significant result simply due to chance tends to 1 as the number of hypotheses increases and so there has been many suggestions for controlling this probability of false discovery. Some of the suggested methods to address the above problem are presented next.

In a multiple comparison setting for n hypotheses, four outcomes are possible, as presented in Table 1.2 where U, V, S, T are unobserved random variables while R is the observed one with the quantities of interest being the sizes of V and R. Two

Table 1.2: Errors committed when testing n hypotheses

	Accepted	Rejected	Total
True null hypothesis	U	V	g_o
False null hypothesis	T	S	$g - g_o$
Total	$g - R$	R	g

general ways of addressing the multiple hypothesis problem and include the control of

the family wise error (FWER) and the False Discovery Rate (FDR). FWER control can be defined as the control of probability of any error, that is, FWER=$\mathbb{P}(V \geq 1)$. Some procedures that control FWER include;

- **Bonferroni's method**: controls FWER at level α in strong sense such that $\mathbb{E}(V) \leq \frac{g_o}{g}\alpha$. The adjusted p-values for the Bonferroni's method is given by $\tilde{p}_i = \min\{gp_i, 1\}$. This method is often criticized as being too conservative.

- **Sidak's procedure**: considers each test under independence improves on the Bonferroni's bound by rejecting the null hypothesis when $p_i \leq 1 - (1-\alpha)^{1/N}$ so that the corresponding adjusted p-value is given by $\tilde{p}_i = 1 - (1-p_i)^N$.

- **Holm's procedure**: is a more elaborate procedure for controlling familywise error. Let the ordered p-values be;

  $$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(g)}$$

  a null hypothesis is rejected when its corresponding p-value corresponding to $p_{(i)}$ is

  $$p_{(j)} \leq \frac{\alpha}{n - j + 1}$$

  for $j = 1, 2, ..., i$. It is worth noting that the Bonferroni correction is usually very conservative.

- Others include the Westfall and Young procedures.

Procedures for controlling FDR include:

- **Benjamini-Hochberg (BH) procedure**: define FDR as the expectation of **Q** where $\mathbf{Q} = \mathbf{V}/(\mathbf{V}+\mathbf{S}) = \mathbf{V}/\mathbf{R}$. The procedure is as follows; for g hypotheses $H_1, ..., H_g$ based on the p-values $p_1, p_2, ..., p_g$. Let the ordered p-values be $p_{(1)} \leq p_{(2)} \leq, ..., \leq p_{(g)}$. Now, denoting the i^{th} hypothesis as $H_{(i)}$ corresponding to $p_{(i)}$ then BH procedure is given by: let k be the largest i for which $p_{(i)} \leq \frac{i}{g}q^*$ then reject all $H_{(i)}, i = 1, 2, ..., k$. This controls the FDR at q^* for any independent test statistics (Benjamini and Hochberg, 1995), in other words FDR is controlled by controlling the proportion of false discovery. In practice, each $p_{(i)}$ is compared against $\frac{i}{g}q^*$.

The decision on whether to choose the FWER or FDR procedures depends on whether researcher is afraid of getting stuff in the significant list that should not be there and so in that case the FWER methods should be used. However, if the researcher is afraid of missing out on some interesting genes and does not mind having more significant stuff in the list then the FDR methods would be a better choice.

1.5 Outline of this thesis

The remainder of this thesis is organized as follows; Chapter 2 reviews some of the statistical methods used in classification problems in genomics. A comparative study of the different classification methods to real data sets are presented in Chapter 3. The statistical integration of molecular data to test for network changes and a framework for testing for network changes is presented in Chapter 4. Conclusions, summary and future works are presented in Chapter 5.

Chapter 2

Methods for classification

2.1 Introduction

In this chapter, we review the some of the existing classification methods which have been used independently or in combination with other methods. In particular, we review the logistic regression, linear discriminant analysis (LDA), the k-nearest neighbours (KNN), ridge partial least squares (RPLS), support vector machines (SVM) and the kernel multilogit algorithm (KMA). Statistical decision theory which form the framework for the above mentioned models is also reviewed.

Classification involves predicting a certain response variable based on a given set of explanatory variables. An algorithm is usually developed from the training set is then used to discover relationships between the attributes thus making it possible to predict the response.

The problem of classification is not new but then the modern day challenges with regards to classification stems from the complex, high dimensional-low sample data that are generated by different technologies in various disciplines for instance in genomics and signal processing. The structure of the data does not allow for the direct application of the classical multivariate classification techniques. Consequently, there is an obvious need to develop new methods and or adjust the old methods in tandem with the current data structure (Alshamlan et al., 2013).

2.2 Basics of classification

Following Dudoit and Fridlyand (2003), classification is a learning problem consisting of a variable to be predicted which consists of K predefined, unordered set $\{c_1, c_2, ..., c_K\}$ which are arbitrarily assigned labels say $0, \ldots, K-1$ or any other convenient labelling scheme. The K values are predefined according to a given class for example “infected" vs “not infected" depending on the context of the problem at hand. Each object is this case is associated with a corresponding response variable or class label , $Y \in \{1, 2, ..., K\}$ and a set of G predictor variables $X = (X_1, ..., X_G)$. The feature vector X belongs to the feature space $\mathcal{X}$ i.e. $(\mathbb{R}^G)$. Thus the main task is to classify an object into one of the possible K classes based on the observed data X. This in other words implies predicting Y based on X. A classifier is a rule $\mathcal{C}$ that reveals the connection between the response and predictor variables. The classifier $\mathcal{C}$ maps the feature space $\mathcal{X}$ into $\{1, 2, ..., K\}$, $\mathcal{C} : \mathcal{X} \longmapsto \{1, 2, ..., K\}$. In this way, the feature space $\mathcal{X}$ is partitioned by the classifier $\mathcal{C}$ into disjoint K subsets that are exhaustive.

In gene expression data, the approaches for deriving classifiers can broadly be categorized into two: (a) *Simple “manual” methods* usually univariate. (b) *Statistical learning methods* which are often multivariate, complicated but give better performance. The recipe for building a classifier using statistical learning involves first, choosing a classification method. Then a feature selection/dimension reduction is implemented. The classifier is thus trained and finally the performance of the trained classifier is assessed. Typically, to build a classifier, the data is first partioned into the learning set $\mathcal{L}$ and the test data $\mathcal{T}$. This partioning should be done arbitrarily in a random manner so as to minimize the risk of biaseness. Dudoit et al. (2002), recommend a division of a ratio of $2:1$ in favor of the learning set.

Decision theory

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ denote a real valued random explanatory and response variables respectively with a joint distribution $P(X, Y)$. To predict the values of Y given X we need a function $f(X)$ and a loss function to penalize the prediction errors. Denote the squared loss function as $L(Y, f(X)) = (Y - f(X))^2$ so that the expected prediction error (EPE) is given by

$$EPE(f) = E(Y - f(X))^2 \tag{2.1}$$

$$= \int [y - f(x)]^2 P(dx, dy). \tag{2.2}$$

Noting that $P(X, Y) = P(Y/X)P(X)$, that is, conditioning on X then 2.2 can be written as

$$EPE(f) = E_{Y|X}([Y - f(X)]^2|X) \tag{2.3}$$

which minimizes EPE pointwise

$$f(x) = \arg\min_c E_{Y|X}([Y - c]^2|X = x) \tag{2.4}$$

whose solution is basically a regression function $f(x) = E(Y|X = x)$. Thus the best prediction is given by the conditional mean with reference to the mean squared error, see (Hastie et al., 2009). Now, considering a categorical variable Y with K elements then the estimate $\hat{Y}$ is expected to assume the values in the space $\mathcal{Y}$ of all possible classes same as G. The loss function will be a $K \times K$ matrix $\mathbf{L}$ with zeros on the diagonals and non-negatives on off the diagonal and $K = card(\mathcal{Y})$. $L(k, l)$ is the price paid for classifying a sample as belonging to class $\mathcal{Y}_k$ as $\mathcal{Y}_l$. The expected prediction error is therefore given by

$$EPE = E\{L(Y, \hat{Y}(X))\}. \tag{2.5}$$

Taking expectation with respect to to the joint distribution and conditioning, we get;

$$EPE = E_X \sum_{k=1}^{K} L[\mathcal{Y}_k, \hat{Y}(X)]P(\mathcal{Y}_k|X) \tag{2.6}$$

to minimize the EPE pointwise

$$\hat{Y}(x) = \arg\min_{y \in \mathcal{Y}} \sum_{k=1}^{K} L(\mathcal{Y}_k, y)P(\mathcal{Y}_k|X = x) \tag{2.7}$$

which simplifies to

$$\hat{Y}(x) = \arg\min_{y \in \mathcal{Y}}[1 - P(y|X = x)] \tag{2.8}$$

$$\hat{Y}(x) = \mathcal{Y}_k \text{ if } P(\mathcal{Y}_k|X = x) = \max_{y\in\mathcal{Y}} P(Y|X = x) \tag{2.9}$$

also known as the Bayes classifier classifies using the most probable class based on the conditional distribution $P(Y|X)$ (which is just the class posterior distribution). For a K-class problem, the Bayes Classifier can be presented as $E(Y_k|X) = P(Y = \mathcal{Y}_k|X)$ via estimation of the squared loss function (Dudoit and Fridlyand, 2003).

Linear discriminant analysis (LDA)

Let the prior probability of class k be given by $P(Y = k) = \pi_k$, where $\sum_{k=1}^{K} \pi_k = 1$ and the conditional density $P(X = x|Y = k) = f_k(x)$, it follows that the posterior distribution is given by

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_k(x)\pi_l}. \tag{2.10}$$

Now given a value, say x, in order to assign it to a given class, a good strategy would be to consider the class with the highest posterior probability. In that sense, x is assigned to k if $\frac{\mathrm{P}(k|x)}{\mathrm{P}(l|x)} > 1$ or when $\frac{f_k(x)}{f_l(x)} > \frac{\pi_k}{\pi_l}$ (for all l not equal to k) while on the boundary $\{x \in \mathbb{R} : \frac{f_k(x)}{f_l(x)} = \frac{\pi_k}{\pi_l}\}$ then the assignment to a particular class can be resolved by tossing a fair coin.

The LDA utilizes the multivariate Gaussian density with the assumption that the classes have a common variance such that $\Sigma_k = \Sigma\ \forall k$. To compare classes k and l, one strategy would be to compare the log odds ratio so that

$$\log \frac{P(Y = k|X = x)}{P(Y = l|X = x)} \quad = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \tag{2.11}$$

Equation 2.11 is linear and implies that the decision boundary between the classes k and l is the set where $P(Y = k|X = x) = P(Y = l|X = x)$ is linear in x; in p dimensional hyperplane. From equation 2.11 linear discriminant functions are given by $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k$ and are equivalent in terms of the decision rule with $Y(x) = \arg\max_k \delta_k(x)$ with unknown parameters estimated from the learning set. When the covariances of k classes are not assumed equal then we

end up with the quadratic discriminant function presented as

$$\delta_k(x) = \log \pi_k - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) \tag{2.12}$$

with the decision boundary for classes k and l described by $\{x : \delta_k(x) = \delta_l(x)\}$. For a comprehensive discussion see Hastie et al. (2009); Mitchell (1994). One problem with the quadratic discriminant analysis (QDA) is caused when some of the attributes have zero variance in one class resulting in to non-invertible covariance matrix. This problem is usually avoided by adding a small positive constant term to the diagonal terms of the covariance matrix or solved by using a combination of the class covariance and the pooled covariance (Mitchell, 1994).

The k-nearest neighbour (KNN)

This approach focuses on the distance (usually the euclidean distance) between elements of a data set especially the closest elements without taking into consideration the distributional assumptions. In this method k is a number to be determined by the researcher for instance when $k = 1$ then only the nearest neighbour is taken into consideration and any new object will assigned to the class of its nearest neighbour. The classification should be quite straightforward in situations where the $k > 1$ nearest neighbours are all of the same class otherwise a majority vote is considered for decision making. There is no universal rule of thumb for the choice of k, optimal value of k can be picked by trying various values on the data under consideration and checking the respective performance of various values of k. Another alternative is to use the leave one out cross validation to pick the optimal value of k.

Logistic regression

Consider a data set of sample size n and p covariates with the data points being (x_i, y_i) with $i = 1, \ldots, n$, where y_i is the response variable in 0/1 and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Rewriting in matrix form, then $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{X} = (x_1^T, \ldots, x_p^T)^T$. The logistic regression is thus expressed as

$$\text{logit}\pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{Z}_i^T\boldsymbol{\beta} \tag{2.13}$$

where $\pi_i = E(Y_i), \mathbf{Z} = \{(1,\ldots,1)^T, \mathbf{X}\}, \boldsymbol{\beta} = \{\beta_0,\ldots,\beta_p\}$. The log-likelihood function for this model is then written as

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i\zeta_i - \log(1+\exp(\zeta_i))] \tag{2.14}$$

where $\zeta_i = \mathbf{z}_i^T\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is estimated by the usual maximum likelihood estimation via iterative algorithms. In some cases when $n << p$ then the MLE of $\boldsymbol{\beta}$ may not even exist and so a penalized logistic regression like the popular lasso L_1 and $L_{1/2}$ penalization which was recently proposed by Liang et al. (2013) may be utilized.

The ridge partial least squares (RPLS)

The RPLS was developed by Fort and Lambert-Lacroix (2005) for situations in which the response variable was binary unlike in the original PLS where the response variable is continuous. RPLS works by substituting the categorical response variable of the PLS by a continuous-valued pseudo-response variable whose expected value has a linear relationship with the covariates. When $n << p$ the usual Iterative reweighted least squares (IRLS) algorithm (to be introduced next) no longer works since the limiting pseudo-response variable is infinite in norm. Therefore, the likelihood criterion is penalized to constrain the pseudo-response variable to be finite. The RPLS algorithm broadly contains two procedures: first is to combine the ridge penalty step with the PLS step then dimension reduction is incorporated in the classification step.

Iterative reweighted least squares (IRLS) and ridge IRLS (RIRLS)

From 2.14, the estimate of $\boldsymbol{\beta}$ can be computed as the limit of a converging Newton-Raphson sequence also known as the IRLS algorithm (Fort and Lambert-Lacroix, 2005) where each iteration is divided into two steps:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{m+1} &= (\mathbf{Z}^T\mathbf{V}^{(m)}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{V}^{(m)}\boldsymbol{\vartheta}^{(m)}\\ \boldsymbol{\vartheta}^{(m+1)} &= \mathbf{Z}\hat{\boldsymbol{\beta}}^m + (\mathbf{V}^{(m)})^{-1}\left[\mathbf{y}-\boldsymbol{\pi}^{(m)}\right].\end{aligned}$$

Here, $\boldsymbol{\vartheta}^m$ is the pseudo-variable while $\hat{\boldsymbol{\pi}}^{(m)} = \left[\hat{\pi}_1^{(m)}, \ldots, \hat{\pi}_n^{(m)}\right]$ is the vector of the estimated probabilities of success for each observation, $\hat{\pi}_i^{(m)} = \text{logit}^{-1}\left(\mathbf{z}_i^T \hat{\boldsymbol{\beta}}^{(m)}\right)$, $\mathbf{V}^{(m)} = \text{diag}\left(v_i^m, \ldots, v_n^m\right)$ is the diagonal empirical variance matrix of observations y_i at step m where, $v_i^m = \hat{\pi}_i^m\left[1 - \hat{\pi}_i^m\right]$. The IRLS algorithm achieves the successive resolution of a weighted least squares regression since each step can be viewed as a regression of the pseudo-variable $\boldsymbol{\vartheta}^m$ against $\mathbf{Z}$ with the weight matrix being $\mathbf{V}^{(m)}$. Accordingly, the pseudo-variable $\boldsymbol{\vartheta}^\infty$ is produced as the limit of the sequence $\left(\vartheta^m\right)_{m\geq 1}$ computed during each iteration. That is, $\boldsymbol{\vartheta}^\infty = \mathbf{Z}\hat{\boldsymbol{\beta}}^\infty + \boldsymbol{\epsilon}$, where $\hat{\boldsymbol{\beta}}^\infty$ is the solution of the likelihood optimization while $\boldsymbol{\epsilon}$ is a vector of noise of the covariance matrix $(\mathbf{V}^\infty)^{-1}$, where $\mathbf{V}^\infty$ is the limit of the matrix sequence $\left(\mathbf{V}^m\right)_{m\geq 1}$.

When $\mathbf{Z}$ is not of full rank the MLE parameter is not unique if it exist and may not even exist due to the infinite norm especially so when $n << p, n = \text{rank}(\mathbf{Z})$ and so regularization by applying the l_2 norm penalty constraint on the co-efficients is done to get

$$\log L(\boldsymbol{\beta}) - 0.5\lambda\boldsymbol{\beta}^T\boldsymbol{\Sigma}^2\boldsymbol{\beta} \tag{2.15}$$

where $\boldsymbol{\Sigma}^2$ is diagonal and is the empirical variance of $\mathbf{Z}$ and $\lambda > 0$. The optimization of the regularized log-likelihood function leads to the Ridge IRLS (RIRLS) so that the weighted regression of each IRLS iteration is replaced by the ridge reweighted regression, hence

$$\hat{\boldsymbol{\beta}}^{(m+1)} = (\mathbf{Z}^T\mathbf{V}^{(m)}\mathbf{Z} + \lambda\boldsymbol{\Sigma}^2)^{-1}\mathbf{Z}^T\mathbf{V}^{(m)}\vartheta^{(m)} \tag{2.16}$$

which guarantee a unique solution which is computed as the limit of $(\hat{\boldsymbol{\beta}}^{(m)})_{m\geq 1}$ see (Fort and Lambert-Lacroix, 2005).

Weighted PLS (WPLS)

Consider the response vector $\mathbf{y} \in \mathbb{R}^n$ and $\boldsymbol{X}$ a $n \times p$ dimensional data matrix and $\boldsymbol{V}$ being a symmetric positive definite $n \times n$ matrix. According to Fort and Lambert-Lacroix (2005), the PLS defines κ, V-orthogonal scores $t_k, k = 1, \ldots, \kappa$ which are linear combinations of $\mathbb{Z}$ for all $k, \mathbb{1}_n' V t_k = 0$ and also performs a V-weighted least

squares regression of the response $\mathbf{y}$ on $(\mathbb{1}_n, t_1, \ldots, t_\kappa)$ to yield

$$\mathbf{y} = b_0\mathbb{1}_n + b_1 t_1 + \cdots + b_\kappa t_\kappa + e_{\kappa+1} = Z\hat{\boldsymbol{\beta}}^{PLS,\kappa} + e_{k+1}$$

where $e_{\kappa+1}$ is the error term which is V-orthogonal to the vector $(\mathbb{1}_n, t_1, \ldots, t_\kappa)$. The usual PLS algorithm is derived using $V = \mathbb{I}$ an identity matrix Fort and Lambert-Lacroix (2005) for the weighted PLS algorithm.

Therefore, the RPLS basically involves two major steps with the first one being to build the the pseudo-response variable which is continuous $\boldsymbol{\vartheta}^{(\infty)}$ with a 'dispersion matrix' $[\mathbf{V}^\infty]^{-1}$ then the second step is to implement a weighted PLS. RPLS depends mainly on the two parameters λ and κ where λ is determined in the first step by Bayesian Information Criterion (BIC) criterion while the determination of κ that is, the number of PLS components to be used is still an open problem even though there are several proposals for determining it such as cross validation and hard thresholding depending on the context, see (Fort and Lambert-Lacroix, 2005).

Support vector machines (SVM)

The review in this section comes mainly from and Cortes and Vapnik (1995) Hastie et al. (2009) among other references. Given a training set consisting of $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$; SVM finds an hyperplane $\{x : f(x) = x^T\beta + \beta_0 = 0\}$ that creates the biggest margin $m = 2/\|\beta\|$ between the training points for the two classes. If the classes are assumed separable then function $f(x) = x^T\beta + \beta_0$ with $y_i f(x_i) \geq 1, \forall i$ can be found. This optimization problem can be solved by the Lagrange multiplier equation L given as

$$La = \frac{1}{2}\|\beta\|^2 - \sum_i \alpha_i(y_i(x^T\beta + \beta_0) - 1). \tag{2.17}$$

where $\alpha_i' s$ are Lagrangian multipliers, one for each data point. The parameters β, β_0 and α_i determine the unique maximal margin (m) boundary line solution and are determined by taking partial derivatives with respect to each parameter respectively. For the maximum margin, the positive and negative data points on the edges of the margin with non-zero alpha values are known as the support vectors and are associated with the weights α_i which determine the amount of influence on the surrounding region. The vectors with zero weights α_i are known as the non-support

vectors and usually dont influence the decision boundary.

When the data points are not linearly separable then one strategy is to introduce variables $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ which allow some individual observations to be on the wrong side of the hyperplane leading to a convex optimization problem of

$$y_i(x_i^T\beta + \beta_0) \geq m(1 - \xi_i), \tag{2.18}$$

$\forall i, \xi_i \geq 0, \sum_{i=1}^{N} \xi \leq k$, where k is a constant. Thus equation 2.18 leads to the standard support vector classifier (Hastie et al., 2009) which can conveniently be re-expressed as

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i, \text{ subject to } \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \, \forall i, \tag{2.19}$$

where C is the cost and equals infinity when the classes are separable. The Lagrange function for 2.19 is given by

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{N} \mu_i\xi_i, \tag{2.20}$$

where α_i and μ_i are positivity constraints.

Equation 2.20 is minimized with respect to the unknown parameters to obtain the following $\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \sum_{i=1}^{N} \alpha_i y_i$ and $\alpha = C - \mu_i, \forall i$ and also $\xi_i \geq 0 \, \forall i$. These results are substituted into equation 2.20 to obtain the Lagrange dual objective function

$$L_D = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j x_i^T x_j. \tag{2.21}$$

Equation 2.21 is maximized subject to $0 \leq \alpha_i \leq C$ and $0 = \sum_{i=1}^{N} \alpha_i y_i = 0$.

Noting that equation 2.17 basically requires us to compute the dot products of β and x amounts to requiring the computation of the "distance" between β and x. Consider a function ϕ such that $\phi : \mathbb{R}^N \to H$, i.e., ϕ maps from the original space to a higher dimensional space H. The original data $\mathcal{X}$ input space with the classification rule $G(x) = \text{sign}(x^T\beta + \beta_0)$ subject to $\beta = \sum_{i=1}^{N} \alpha_i y_i x_i$ can be mapped

into a higher dimenional space feature space $\phi(x)$ such that $G(x) = \text{sign}(\phi x^T \beta + \beta_0)$ subject to $\beta = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)$. The transformed feature vectors $\phi(x_i)$ need not be known explicitly but any function defined by $K(.,.) : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ and is positive definite (satisfying the Mercer condition) guarantees existence of ϕ, so that $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ (Cortes and Vapnik, 1995). Thus the Lagrange dual function for the transformed feature space is given by

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_i) \rangle. \tag{2.22}$$

The solution function $f(x)$ can therefore be written as

$$f(x) = \phi(x)^T \beta + \beta_0 = \sum_{i=1}^{N} \alpha_i y_i \langle \phi(x), \phi(x_i) \rangle + \beta_0. \tag{2.23}$$

The parameters α_i, β_0 can be obtained by solving $y_i f(x_i) = 1$ in equation 2.22 for all x_i for which $0 < \alpha_i < C$. The solution for equation 2.23 is consequently written as $\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0$. Some of the kernel functions that can be used for $K(.,.)$ include, the linear kernel $K(x_i, x_j) = \langle x_i, x_j \rangle$, gaussian radial basis function (RBF) kernel $K(x_i, x_j) = \exp\left[-\sigma \|x_i - x_j\|^2\right]$, the Laplace radial basis function (LRBF) $K(x_i, x_j) = \sigma \|x_i - x_j\|$, the polynomial kernel and the linear splines kernel in one dimension among many others.

The choice of an appropriate kernel is usually a non-trivial task and regardless of the kernel chosen, the kernel parameter needs to be tuned in order to get a good performance. One of the most popular tuning method that has frequently been employed is the K-fold cross validation.

Kernel multilogit algorithm (KMA)

The KMA was recently proposed by Dalmau et al. (2015). This algorithm works by first transforming a categorical response variable to a continuous one via multilogit transformation. The transformed variable is then used with the explanatory variables in a regression model for classification and prediction. Finally, the new predicted variables are transformed back using the inverse multilogit function to the original space to enable classification.

Let the response variable vector $\boldsymbol{y}$ be a categorical with class labels $\{1, 2, \ldots, C\}$,

to classify a discrete variable from predictor variables $\boldsymbol{x}$, the first step is to transform the response variable $\boldsymbol{y}$ into a new space using the multilogit function. The multinomial logit model with C as the reference category can be given as

$$\begin{aligned}\Pr(\boldsymbol{y}=j|\boldsymbol{x}) &= \frac{\exp\{f(\boldsymbol{x};\boldsymbol{\theta}_j)\}}{1+\sum_{i=1}^{C-1}\exp\{f(\boldsymbol{x};\boldsymbol{\theta}_i)\}}, \quad j=\{1,2,\ldots,C-1\} \\ \Pr(\boldsymbol{y}=C|\boldsymbol{x}) &= \frac{1}{1+\sum_{i=1}^{C-1}\exp\{f(\boldsymbol{x};\boldsymbol{\theta}_i)\}},\end{aligned} \tag{2.24}$$

where $f(\boldsymbol{x};\boldsymbol{\theta}_i)=\boldsymbol{x}^T\boldsymbol{\theta}_i$. The expected value of $\boldsymbol{y}$ being multinomial random variable is given by $E(\boldsymbol{y}|\boldsymbol{x})=[\Pr(\boldsymbol{y}=1|\boldsymbol{x}),\Pr(\boldsymbol{y}=2|\boldsymbol{x}),\ldots,\Pr(\boldsymbol{y}=C|\boldsymbol{x})]^T$. Now, denoting $\boldsymbol{t}=E(\boldsymbol{y}|\boldsymbol{x})$, the original response variable $\boldsymbol{y}$ is not used but instead a transformed version $\boldsymbol{\vartheta}=\text{logit}(\boldsymbol{t})$ is used. The logit transformation is done with C as the reference category as follows

$$\vartheta_j=\text{logit}(t_j)=\log\frac{t_j}{t_C}, \quad j=\{1,2,\ldots,C-1\} \tag{2.25}$$

where $\vartheta_j\in\boldsymbol{\vartheta}, t_j\in\boldsymbol{t}$.

In the second step a parametric linear model is proposed and its parameter estimates can be obtained via the standard Bayesian formula $\Pr(\boldsymbol{\vartheta}|\boldsymbol{x}) = \Pr(\boldsymbol{x}|\boldsymbol{\vartheta})\Pr(\boldsymbol{\vartheta})/\Pr(\boldsymbol{x})$ where $\Pr(\boldsymbol{\vartheta}|\boldsymbol{x})$ is the posterior probability distribution, $\Pr(\boldsymbol{x}|\boldsymbol{\vartheta})$ is the likelihood function and $\Pr(\boldsymbol{x})$ is the normalization constant. Assuming that $\boldsymbol{\vartheta}\in\mathbb{R}^{C-1}$ for a given $\boldsymbol{x}\in\mathbb{R}^m$ follows a multivariate normal distribution $\boldsymbol{\vartheta}|\boldsymbol{x}\sim\mathcal{N}(\Theta^T\mathbf{x},\alpha^2\mathbf{I})$, $\Theta\in\mathbb{R}^{m\times C-1}$ and $\Pr(\boldsymbol{\vartheta}|\boldsymbol{x})$ is also multivariate normally distributed. Furthermore, the prior parameters are assumed to follow a normal distribution, i.e. $\boldsymbol{\theta}\sim\mathcal{N}(\mathbf{0},\beta^2\mathbf{I})$ where β is known. The parameter matrix $\boldsymbol{\Theta}$ is thus estimated by optimizing an equivalent of regularized least squares function

$$\begin{aligned}\hat{\boldsymbol{\Theta}} &= \arg\min_{\Theta}\mathbf{U}(\boldsymbol{\Theta}) \\ U(\boldsymbol{\Theta}) &= \|\boldsymbol{\vartheta}-\mathbf{X}\boldsymbol{\Theta}\|_F^2+\lambda\|\boldsymbol{\Theta}\|_F^2,\end{aligned} \tag{2.26}$$

where $\boldsymbol{\vartheta}=[\boldsymbol{\vartheta}^{(i)}]_{i=1,2,\ldots,n}^T$, $\mathbf{X}=[\boldsymbol{x}^{(i)}]_{i=1,2,\ldots,n}^T$, $\|.\|_F$ is the Frobenius norm of a matrix and λ is the regularization parameter. The result is a closed form estimate given by

$\hat{\mathbf{\Theta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{\vartheta}$. To capture non-linearities which may be present, a dual representation $\mathbf{\Theta} = \mathbf{X}^T\mathbf{\Gamma}$ is taken so that

$$U(\mathbf{\Gamma}) = \|\boldsymbol{\vartheta} - \mathbf{X}\mathbf{X}^T\Gamma\|_F^2 + \lambda\|\mathbf{X}^T\mathbf{\Gamma}\|_F^2$$

then $U(\mathbf{\Gamma})$ is optimized to get $\hat{\mathbf{\Gamma}} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\boldsymbol{\vartheta}$, where $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ is the Gram matrix, $K_{ij} = \langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \rangle + 1$. However a more general kernel $K_{ij} = ((\phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(j)}))$ where $\phi(\cdot)$is a nonlinear mapping, is preferred in practice.

The final step of the algorithm involves prediction/classification given a new set of response variables $\boldsymbol{x}^{new}$. This entails estimation of $\boldsymbol{\vartheta}^{new}$ by $\boldsymbol{\vartheta}^{new} = \hat{\mathbf{\Gamma}}^T\hat{\boldsymbol{x}}^{new}$, but $\hat{\boldsymbol{x}}^{new} = K((\phi(\boldsymbol{x}^{(i)}), \phi(\boldsymbol{x}^{(new)}))$. The computed $\boldsymbol{\vartheta}^{new}$ is used to estimate $\mathbf{t}^{new}$ by using $\boldsymbol{t}^{new} = \text{logit}^{-1}(\boldsymbol{\vartheta}^{new})$. The inverse of a logit function is given by

$$\begin{aligned} t_j^{new} &= \frac{\exp\{\vartheta_j^{new}\}}{1 + \sum_{i=1}^{C-1} \exp\{\vartheta_j^{new}\}}, \quad j = \{1, 2, \ldots, C-1\} \\ t_J^{new} &= \frac{1}{1 + \sum_{i=1}^{C-1} \exp\{\vartheta_j^{new}\}}. \end{aligned} \tag{2.27}$$

The class labels associated with $\boldsymbol{x}^{new}$ are then computed using the estimated conditional distribution by finding the components that maximises the components of $\boldsymbol{t}^{new}$ i.e. using the Bayes rule. The computed $\mathbf{t}^{new}$ is then used to get the class label ($\hat{\boldsymbol{y}}^{new}$) of the new data , for details see Dalmau et al. (2015).

2.3 Partial least squares (PLS) and some of its applications in genomics

The PLS is a very useful approach because it is able to analyze data with many, noisy, collinear as well as incomplete variables. PLS is usually utilized in data reduction when there is multicollinearity or when the data has more variables than the number of samples. Essentially, the PLS aims at maximizing the covariance between the response variables $\boldsymbol{Y}$ and the predictors $\boldsymbol{X}$, i.e., $cov(\boldsymbol{X}^T\boldsymbol{Y})$ of highly multidimensional data by finding a linear subspace of the explanatory variables (Wold et al., 2001; Höskuldsson, 1988). Some literature on PLS can be found on Wold et al.

(2001, 1984); Höskuldsson (1988) among others.

The research on PLS is still very active due to its ability to address problems associated with the high dimensional data such as multicolinearity, high dimensionality, among others. In recent past, the PLS has been utilized predominantly in the high dimensional data in different fields like chemometrics and the "omics" like genomics, proteomics and many other fields that generate large amounts of data like spectroscopy (Gromski et al., 2015). Some recent applications of PLS in microarray studies include, Huang et al. (2013) who applied PLS regression (PLSR) in breast cancer intrinsic taxonomy, for classification of distinct molecular sub-types by using PAM50 signature genes as predictive variables in PLS analysis and the latent gene component in binary the logistic regression for each molecular sub-type. Telaar et al. (2013) extended the notion of PLS-discriminant analysis (PLS-DA) to Powered PLS-DA (PPLS-DA), introducing a so called 'power parameter', which is maximized towards the correlation between the components and the group-membership thereby achieving a minimal classification error. Furthermore, Xi et al. (2014) discussed the PLS-DA with applications to metabolites data. Other articles involving the usage of PLS include, Dong et al. (2014) who used PLS to investigate the underlying mechanism of the post-traumatic stress disorder (PTSD) using microarray data. Gusnanto et al. (2013) made gene selection based on partial least squares and logistic regression random-effects (RE) estimates for evaluation in classification models. Gene selection involving PLS was also done by Wang et al. (2015). The sparse PLS has also been utilized by many researchers for instance Chun and Keles (2009); Lee et al. (2011) provided an efficient algorithm for the implementation of the sparse PLS for variable selections in high dimensional data. Furthermore, Lê Cao et al. (2008) used a sparse PLS for variable selection when integrating omics data. They implemented sparsity via lasso penalization of the PLS loading vectors when computing the singular value decomposition.

2.4 PLS regression (PLSR) algorithm

Starting with the response variables' matrix Y and the predictors' matrix X with dimensions $N \times K$ and $N \times M$ respectively. The matrices are scaled and centered previously to make their distributions fairly asymmetric. Höskuldsson (1988) and Wold et al. (2001) give a good summary of the PLS with application in Chemometrics.

The PLSR uses the estimates of the latent variables or the PLS components as the new variables denoted by $(\mathbf{t}_h, h = 1, ..., H)$. These new variables are estimated as a linear combination of the of the original variable $\mathbf{x}_k$ with the co-efficients, "weights" w^*_{kh}.

$$t_{ih} = \sum_k W^*_{kh} X_{ik}; \quad \mathbf{T} = \mathbf{X}\mathbf{W}^* \tag{2.28}$$

where $\mathbf{T}$ is a matrix of PLS components/X-scores, $\mathbf{W}$ is a matrix of weights. The $\mathbf{t}_h$'s have the property that when multiplied by the loadings p_{hk} they give good summaries of $\mathbf{X}$ so that the residuals e_{ik} are small.

$$X_{ik} = \sum_h t_{ih} p_{hm} + e_{ik}; \quad \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}. \tag{2.29}$$

In a similar manner for $\mathbf{Y}$ when $(M > 1)$, the corresponding "Y-score" $(\mathbf{u}_h)$, when multiplied by weights c_{hm} gives good summaries of $\mathbf{Y}$ so that the residuals g_{im} are small

$$y_{im} = \sum_h u_{ih} c_{hm} + g_{im}; \quad \mathbf{Y} = \mathbf{U}\mathbf{C}^T + \mathbf{G}. \tag{2.30}$$

The X-scores (t_h)'s are good predictors of $\mathbf{Y}$ so that

$$y_{im} = \sum_h c_{mh} t_{ih} + f_{im}; \quad \mathbf{Y} = \mathbf{T}\mathbf{C}^T + \mathbf{F} \tag{2.31}$$

where f_{im}, are residuals for the observed responses and the modelled ones. Rewriting equations 2.28 and 2.29 as a multiple regression we get;

$$y_{im} = \sum_h c_{mh} \sum_k w^*_{kh} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im}; \quad \mathbf{Y} = \mathbf{X}\mathbf{W}^* + \mathbf{F} = \mathbf{X}\mathbf{B} + \mathbf{F}. \tag{2.32}$$

Also, the regression parameters can be written as

$$b_{mk} = \sum_h c_{mh} w^*_{kh}; \quad \mathbf{B} = \mathbf{W}^*\mathbf{C}^T \tag{2.33}$$

The estimated coefficients are not independent unless the number of PLS components are equal to the number of X-variables and so their confidence-intervals are infinite according to the classification statistics (Wold et al., 2001). The matrix $\mathbf{X}$ is usually deflated after each component h by subracting $\mathbf{X} - \mathbf{t}_h \mathbf{p}_h^T$ making it

possible to express the PLS model in terms of weights $\mathbf{w}_h$ referring to the residuals after previous dimension $\mathbf{E}_{h-1}$ instead of relating to X-variables themselves (Wold et al., 2001). Thus we write 2.28 as

$$\begin{aligned} t_{ih} = \sum_k W^*_{kh} e_{ik,h-1}; \quad \mathbf{t}_h = \mathbf{E}_{h-1}\mathbf{W}_h \\ e_{ik,h-1} = e_{ik,h-2} - t_{i,h-1}p'_{h-1}; \quad \mathbf{E}_{h-1} = \mathbf{E}_{h-1} - \mathbf{t}_{h-1}\mathbf{p}'_{h-1} \\ e_{ik,0} = X_{ik}; \quad \mathbf{E}_0 = \mathbf{X}. \end{aligned} \tag{2.34}$$

The weights $\mathbf{W}$ can be transformed to $\mathbf{W}^*$ which is directly related to the original variables as $\mathbf{W}^* = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}$. The matrix $\mathbf{Y}$ can also be deflated by subtracting $\mathbf{t}_h\mathbf{c}'_h$ but the deflation or non deflation of $\mathbf{Y}$ does not affect the results. According to Höskuldsson (1988); Wold et al. (2001), the first weight vector $\mathbf{w}_1$ is the first eigenvector of the combined variance-covariance matrix, $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$ and the weight vector component $\mathbf{h}$ are the eigenvectors to the deflated matrix $\mathbf{Z}'_h\mathbf{Y}\mathbf{Y}'\mathbf{Z}'_h$ where $\mathbf{Z}_h = \mathbf{Z}_{h-1} - \mathbf{T}_{h-1}\mathbf{P}'_{h-1}$. The first component $\mathbf{t}_1$ is an eigenvector to $\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{Y}'$ and $\mathbf{t}_h = \mathbf{Z}_h\mathbf{Z}'_h\mathbf{Y}\mathbf{Y}'$. The vectors $\mathbf{w}_h$ form orthogonal set, furthermore the vectors $\mathbf{t}_h$ are orthogonal to each other. However, the loadings $\mathbf{p}_h$ are not orthogonal to each other and neither are $\mathbf{u}_h$'s. On the other hand, the $\mathbf{u}$'s and the $\mathbf{p}$'s are orthogonal to $\mathbf{t}$'s and $\mathbf{w}$'s respectively. It implies that $\mathbf{u}'_i\mathbf{t}_j = 0$ and $\mathbf{p}'_i\mathbf{w}_j = 0$ if $i > j$ and also, $\mathbf{w}'_i\mathbf{p}'_h = 1$ (Wold et al., 2001). There are several variants of the PLSR algorithms that have been developed by different researchers. Some vital properties to consider when designing an algorithm include orthogonality between model components, good summarizing properties of the components $\mathbf{t}_h$ and interpretability of the model. Now we present the NIPALS algorithm (Wold et al., 2001, 1984). The algorithm starts with optionally standardized or mean centred data $\mathbf{X}$ and $\mathbf{Y}$. The steps in Table 2.1 are then implemented.

Table 2.1: PLS Regression Algorithm

Algorithm:

1. Set $\mathbf{u}$ to be the first column of $\mathbf{Y}$
2. Calculate the weights $\mathbf{w}$:$\mathbf{w} = \mathbf{X}^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$
3. Normalize the $\mathbf{w}$ as : $\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|$
4. Obtain X-scores: $\mathbf{t} = \mathbf{X}\mathbf{w}$
5. Calculate Y-scores: $\mathbf{c} = \mathbf{Y}'/(\mathbf{t}'\mathbf{t})$
6. Normalize $\mathbf{c}$ to be of length one
7. Update the Y-scores: $\mathbf{u} = \mathbf{Y}'\mathbf{c}/(\mathbf{c}'\mathbf{c})$
8. If there is convergence then go to step 9 otherwise 2
9. X-loadings: $\mathbf{p} = \mathbf{X}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$
10. Y-loadings: $\mathbf{q} = \mathbf{Y}'\mathbf{u}/(\mathbf{u}'\mathbf{u})$
11. Regress $\mathbf{u}$ on $\mathbf{t}$: $b = \mathbf{u}'\mathbf{t}/\mathbf{u}'\mathbf{u}$
12. Obtain the residual matrices: $\mathbf{X} \longrightarrow \mathbf{X} - \mathbf{t}\mathbf{p}'$ and $\mathbf{Y} \longrightarrow \mathbf{Y} - \mathbf{t}\mathbf{q}'$

The convergence is tested by determining the change in $\mathbf{t}$ of the norms "old" and "new" values divided by the norm of the "old" values. The next set of iterations begin with the residual matrices obtained in step 12 and continue until $\mathbf{X}$ contains zeros or a stopping rule can be used.

Other important issues in PLS include determination of the number of components to be included. Care must be taken so as to avoid over-fitting. A practical way to determine the number of components would be to use cross validation (CV). Depending on the sample size, we can do k-fold cross validation or leave one out CV. Another issue is the model validation to determine whether the model consistently predicts Y with a new set of predictors. In this case, the CV still is a formidable tool.

2.5 PLS generalized linear regression algorithm

In this section, we present an algorithm that can be applied to any Generalized Linear Regression which was developed by Bastien et al. (2005). Consider the response data $\mathbf{y}$ with the explanatory variables $\mathbf{x_1}, \ldots, \mathbf{x_p}$ then a PLS-General Linear Regression (PLSGLR) can be written as

$$g(\theta) = \sum_{h=1}^{m} c_h \Big(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \Big), \tag{2.35}$$

where θ a conditional expectation of the variable $\mathbf{y}$ if its distribution is continuous or a vector of probabilities if the variable $\mathbf{y}$ follows a discrete distribution with a finite support, while $g(.)$ is the link function chosen according to the probability distribution of $\mathbf{y}$. The PLS components are given by $t_h = \sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j, j = 1, \ldots, p, h = 1, \ldots, m$. To compute the PLS components let $\mathbf{X} = \mathbf{x_1}, \ldots, \mathbf{x_p}$ be a matrix of p centred explanatory variables $\mathbf{x}_j$'s. The key objective is to determine m orthogonal PLS components defined as a linear combination of the $\mathbf{x}_j$'s. The algorithm is presented as follows:

1. Computation of the first PLS component t_1 : First, the regression coefficients a_{1j} of $\mathbf{x}_j$ are computed using the usual GLM procedure of $\mathbf{y}$ on $\mathbf{x}_j, j = 1, ..., p$. The column vector $\mathbf{a}_1$ which contains a_{1j} is then normalized : $\mathbf{w}_1 = \mathbf{a}_1 / \|\mathbf{a}_1\|$. Finally, the component t_1 is computed as $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 / \mathbf{w}_1' \mathbf{w}_1$.

2. Computation of the first PLS component t_2 : Involves the computation of the linear model coefficients a_{2j} of $\mathbf{x}_j$ in the GLM setting of $\mathbf{y}$ on t_1 and $\mathbf{x}_j, j = 1, ..., p$. Since the main idea of PLS is to create the orthogonal components t_2, the component t_1 is added as a variable in estimating $\mathbf{y}$ on t_1 and $\mathbf{x}_j, j = 1, ..., p$. This is because the structure of PLSGLR does not allow the residuals of y to be obtained in each iteration that would aid in construction of orthogonal components. The column vector $\mathbf{a}_2$ which contains a_{2j} is normalized: $\mathbf{w}_2 = \mathbf{a}_2 / \|\mathbf{a}_2\|$ and thereafter, the residual matrix $\mathbf{X}_1$ is obtained via the regression of $\mathbf{X}$ on t_1. The use of residual matrix in the obtaining the next component ensures orthogonality between the different components. The component t_2 is calculated by $\mathbf{t}_2 = \mathbf{X}_1 \mathbf{w}_2 / \mathbf{w}_2' \mathbf{w}_2$. Finally, $\mathbf{t}_2$ is expressed in terms of $\mathbf{X}$: $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2^{*}$.

3. Computation of the h^{th} PLS Component t_h: Consider the already been computed components $t_1, ..., t_{h-1}$, the final component t_h is computed by calculating the GLM coefficients a_{hj} of $\mathbf{x}_j$ by fitting $\mathbf{y}$ on $t_1, ...t_{h-1}$ and $\mathbf{x}_j, j = 1, ..., p$. Next, the column vector $\mathbf{a}_h$ which contains a_{hj} are normalized as: $\mathbf{w}_h = \mathbf{a}_h/\|\mathbf{a}_h\|$. The residual matrix $\mathbf{X}_{h-1}$ of the regression of $\mathbf{X}$ on $t_1, ..., t_{h-1}$ is then computed. The use of the residual matrix and the previously obtained $t_1, ...t_{h-1}$ as covariables in calculating the GLM coefficients helps with creating orthogonal components as previously explained. The final component t_h is thus computed as $\mathbf{t}_h = \mathbf{X}_{h-1}\mathbf{w}_h/\mathbf{w}'_h\mathbf{w}_h$. Finally, $\mathbf{t}_h$ is expressed in terms of $\mathbf{X}$: $\mathbf{t}_h = \mathbf{X}\mathbf{w}^*_h$.

Bastien et al. (2005) note that while computing the components t_h, the nonsignificant elements in a_h can be set to zero in order to simplify calculations since only the significant response variables are needed to build the PLS components. The number of m components to be used can be determined through cross-validation or by hard thresholding. The iteration can be stopped once there are no more significant coefficients in a_h.

Consider $x_{h-1,i}$ a column vector of the transpose of the ith row of X_{h-1}, then $t_{hi} = x'_{h-1,i}w_h/w'_hw_h$ of the ith case on the component t_h. This is basically the slope of the fitted line of the univariate OLS linear regression without intercept for $x_{h-1,i}$ on w_h which can be estimated even with some data missing. Consequently, the component is computed based on the available data. Therefore the PLSGLR algorithm by Bastien et al. (2005) effectively copes up with missing data.

Chapter 3

Applications to real data sets

3.1 Introduction

This section presents the data, methodology and explanations of the various aspects of the application of the classification methods to three sets of real microarray data. Each of these data sets is analysed under two different conditions, preprocessed and un-preprocessed. For the preprocessed sets, the PLS Generalized Regression combined with Logistic Regression and also Linear Discriminant Analysis are implemented and their performance compared with the classical methodologies like KNN, LDA, RPLS, PLSDA and SVM. The same is done for the un-preprocessed data sets but in addition to the previous methodologies, the KMA algorithm is implemented.

3.2 Analysis of the unpreprocessed data sets

Colon data is due to Alon et al. (1999) obtained from the `R` package `plsgenomics` is a (62×2000) matrix giving the expression levels of 2000 genes for the 62 Colon tissue samples. Out of the 62 tissues, 22 are healthy while 40 had Colon cancer. The Leukemia data was a matrix of dimension (38×7129) where 11 patients suffered from acute myeloid Leukemia (AML) while 27 were acute lymphoblastic Leukemia (ALL) patients. The third data set was the Prostate cancer dataset is due to Singh et al. (2002) was a (102×12600) in dimension. Out of it, 50 were normal and 52 were tumor. The unpreprocessed Leukemia and Prostate cancer data were downloaded freely from `www.stats.uwo.ca/faculty/aim/2015/9850/microarrays/FitMArray/data/`.

3.2.1 Some exploratory analysis

We use the Colon data to visualize the differences in the unpreprocessed and preprocessed microarray data sets. The preprocessing is done using the `R` package `plsgenomics` see `https://rdrr.io/cran/plsgenomics/` that implements the recommendations of (Dudoit et al., 2002). To visualize the differences between the preprocessed and non-preprocessed data sets, we consider the pairs of box plots, relative log expression (RLE) and principal components analysis (PCA) plots presented in Figures 3.1,3.2, 3.3, 3.4 and 3.5 respectively.

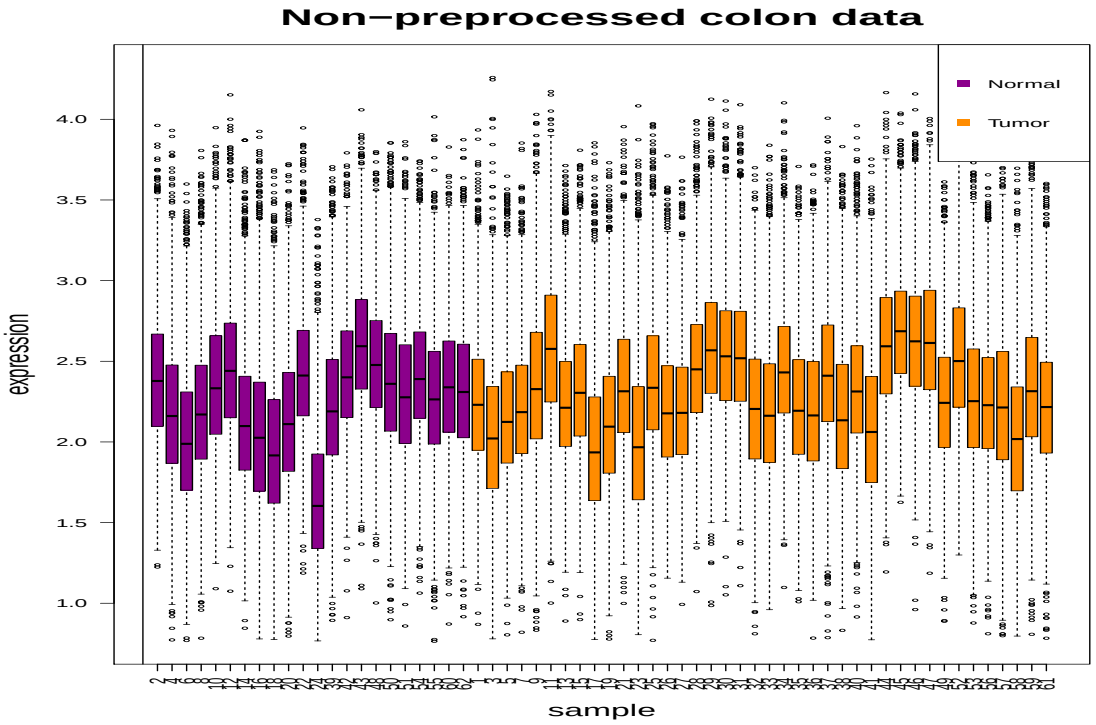


Figure 3.1: Box plot for the non-preprocessed colon data. The box plot for unpreprocessed data clearly shows that the data is noisy and has a lot of variations. The data has some unwanted variations that are expected to affect its analysis. It also lacks symmetry.

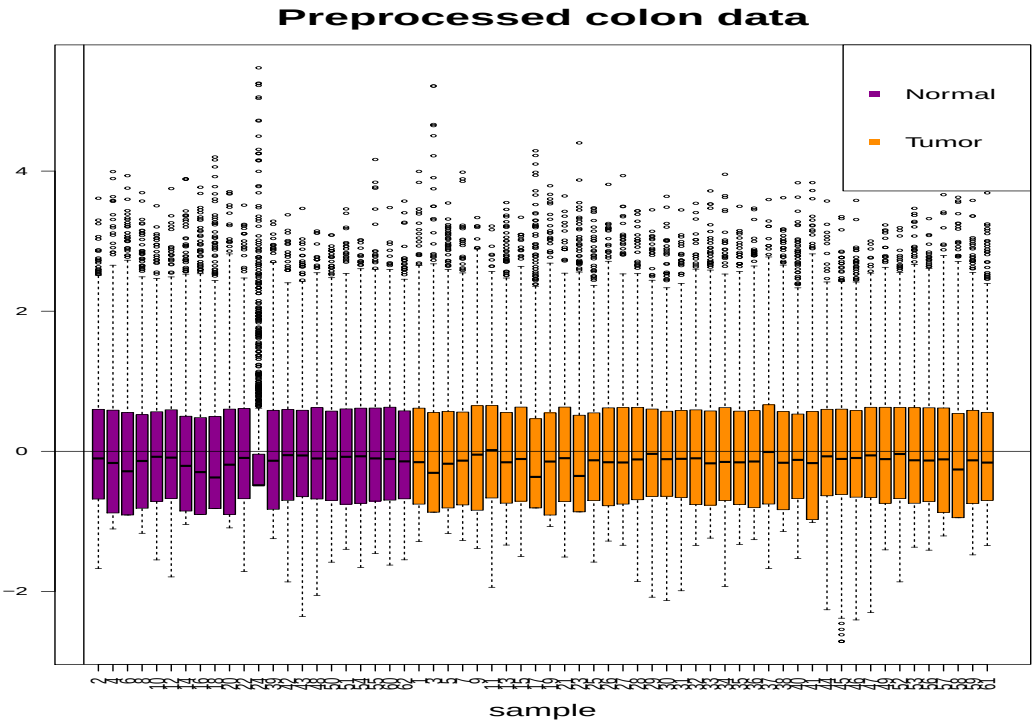


Figure 3.2: Box plot for the preprocessed colon data. This plot looks reasonable with less variations. The data seem to have a symmetric distribution and does not show the presence of unwanted variation. From the two figures, it is expected that the preprocessed data would be easy to analyze.

Next the same pair of data sets is examined using RLE plots to show how the preprocessed data compares with the un-preprocessed one with regards to the batch effect or any other abnormality. The RLE plots have been extensively used in the studies of the microarray data to reveal the effectiveness of data normalization for example see Gagnon-Bartsch and Speed (2011). The RLE plots are simple yet very powerful in the visualization of data to detect unwanted variations. To understand how an RLE plot is constructed, consider a data matrix $\boldsymbol{X}_{p\times n}$ where p is the number of genes while n the number of microarray sample and so the element of the data matrix x_{ij} represents the i^{th} gene in the j^{th}sample. To construct the RLE plot, we calculate the median across each of the p rows and then subtract the respective median across each row of the data matrix $\boldsymbol{X}$, i.e $(x_{ij} - \text{median } x_{i*})$. The median is used because it is robust and not affected by the outliers. A box plot is then generated for each of the n samples and a good one will be centered around zero and its width (interquartile range) should be equal to or less than 0.2 see (Gagnon-Bartsch and Speed, 2011).

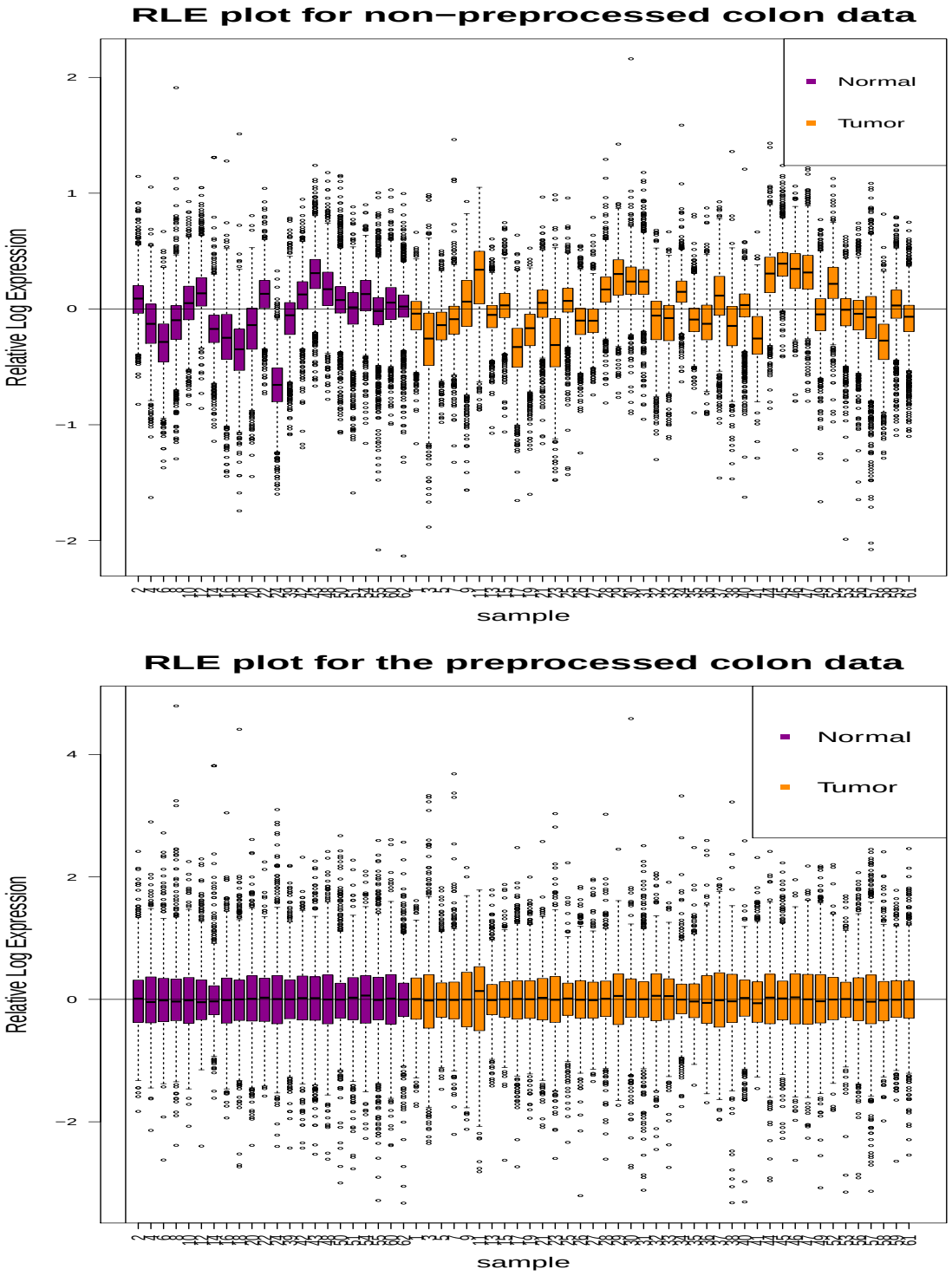


Figure 3.3: RLE plots for the non-preprocessed and pre-processed colon data. The RLE plot for the unprocessed data shows the presence of a lot of heterogeneity which reveals that the data has variations that do not necessarily come from the biological factors. However, the RLE plot for the processed data shows homogeneity and lack of unwanted noise and should give relatively good results when analyzed statistically.

Finally we compare the ease of classification between the un-preprocessed and preprocessed data. The simplest way to visualize the separability of categories in a given data set is the use of principal components analysis (PCA) plots.

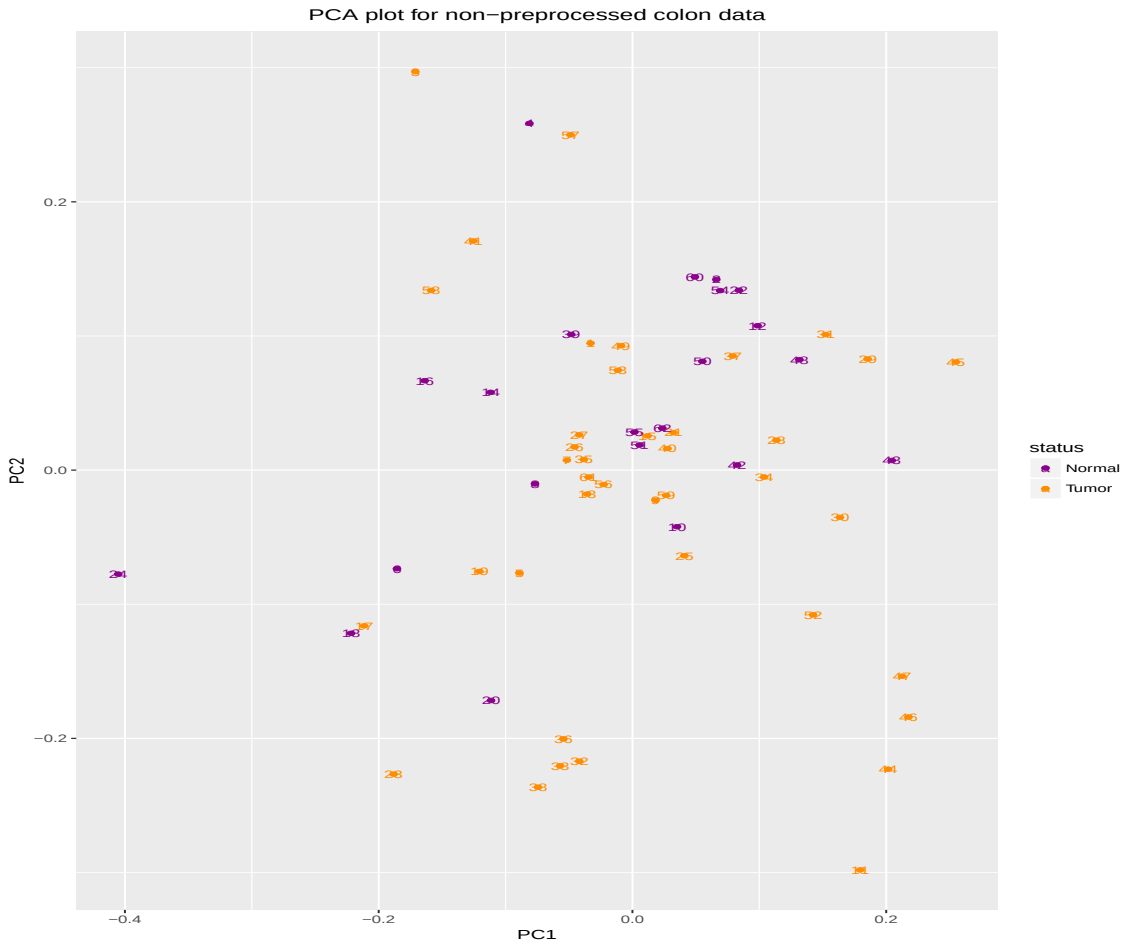


Figure 3.4: PCA plot for the nonpreprocessed Colon data. The PCA plots show that the it is harder to separate/classify the unpreprocessed data.

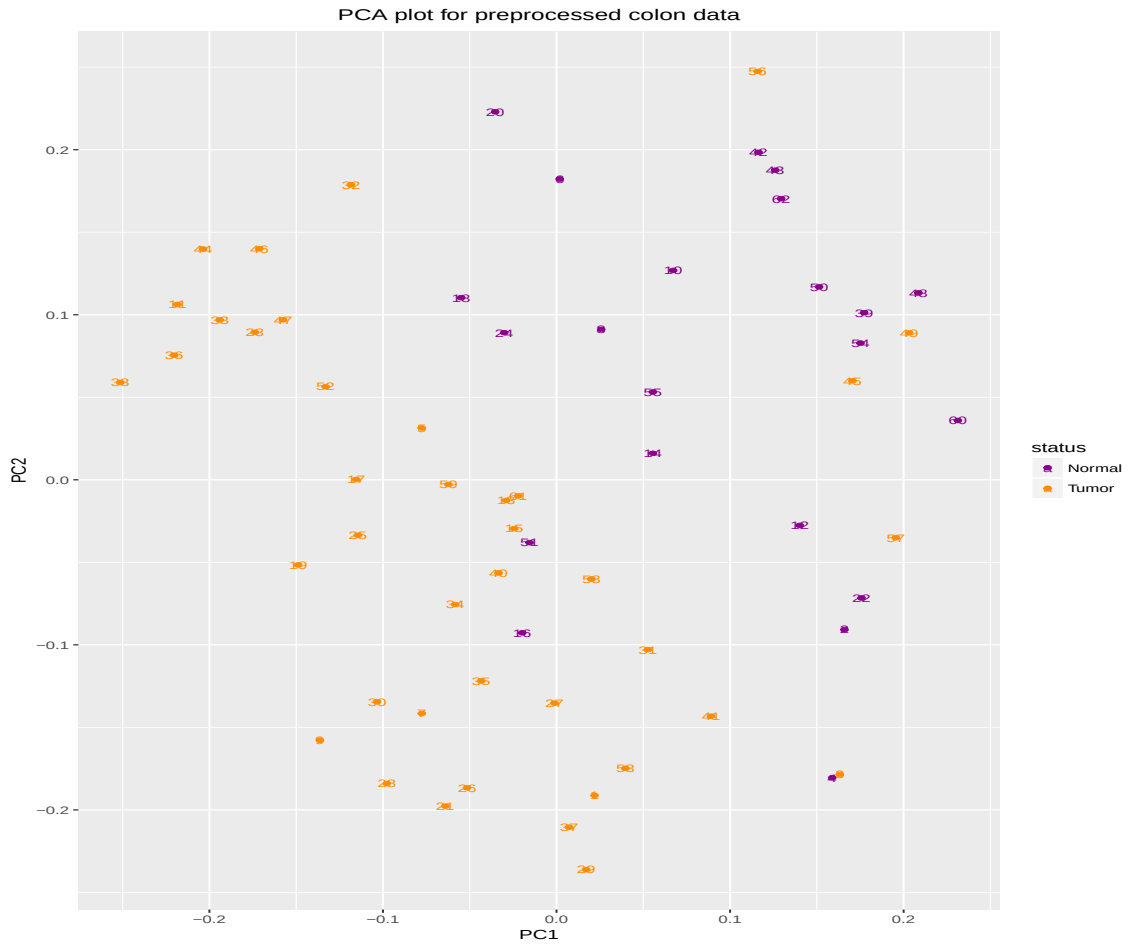


Figure 3.5: PCA plots for the preprocessed Colon data. It is relatively easier to separate/classify compared preprocessed data.

Gagnon-Bartsch and Speed (2011) note that one of the key challenges of the removal of unwanted variation is the difficulty in distinguishing the unwanted variations from the biological variation of interest. Further more they note that the most appropriate way to deal with unwanted variation depends so much on the final objective of the analysis for instance, differential expression (DE), classification, or clustering.

3.2.2 Results from the analysis of the unpreprocessed data

The key objective of this set of analyses was to compare the performance of our proposed model extensions PLSGLR-log, PLSGLRDA and the KMA Dalmau et al. (2015) with the classical methods when the data had not been preprocessed nor

variables selected. In other words, to test the performance of the classification algorithms in the presence of noise. In that regard, none of the data sets was preprocessed and neither was the feature selection implemented for any of them. The performance of the methodologies were then compared using the cross validation. In this case a 10 fold cross validation (10-CV) was carried out and corresponding missclassification percentages computed. The results are presented in Table 3.1.

Table 3.1: Percentage missclassification for the different methods when applied to the unpreprocessed data sets

DATA	PLSGLR-log	PLSGLRDA	KNN	LDA	PLSDA	RPLS	SVM	KMA
Colon	38.3	31.7	60.0	25.0	11.7	15.0	18.3	1.7
Leukemia	5.6	1.4	34.7	5.6	1.4	1.4	100.0	1.3
Prostate	11.6	8.0	100.0	9.8	7.1	6.2	7.1	0.8

A particular method is judged to be the "best" if it has a misclassification percentage relative to the other methods, otherwise its a poor classifier. The results based on minimal cross validation misclassification percentages indicate that for the colon data, the KMA emerged the best followed by PLSDA, RPLS while the worst was the KNN and PLSGLR-log. For the Leukemia data, KMA was the best, while the second best had a tie between, PLSGLRDA, PLSDA and RPLS while KNN and the SVM were the worst classifiers in this case. Finally, for the Prostate data, KMA still emerged as the best followed by PLSGRDA, PLSDA and SVM while KNN retained the worst performance while the other methodologies performed fair. The results suggest that KMA, PLSGRDA, RPLS and PLSDA seem to perform well in the absence of preprocessing and gene selection. In overall, for the un-preprocessed noisy data sets, it is clear that KMA is the best classifier, KNN is the worst while for the rest there is no clear "winner".

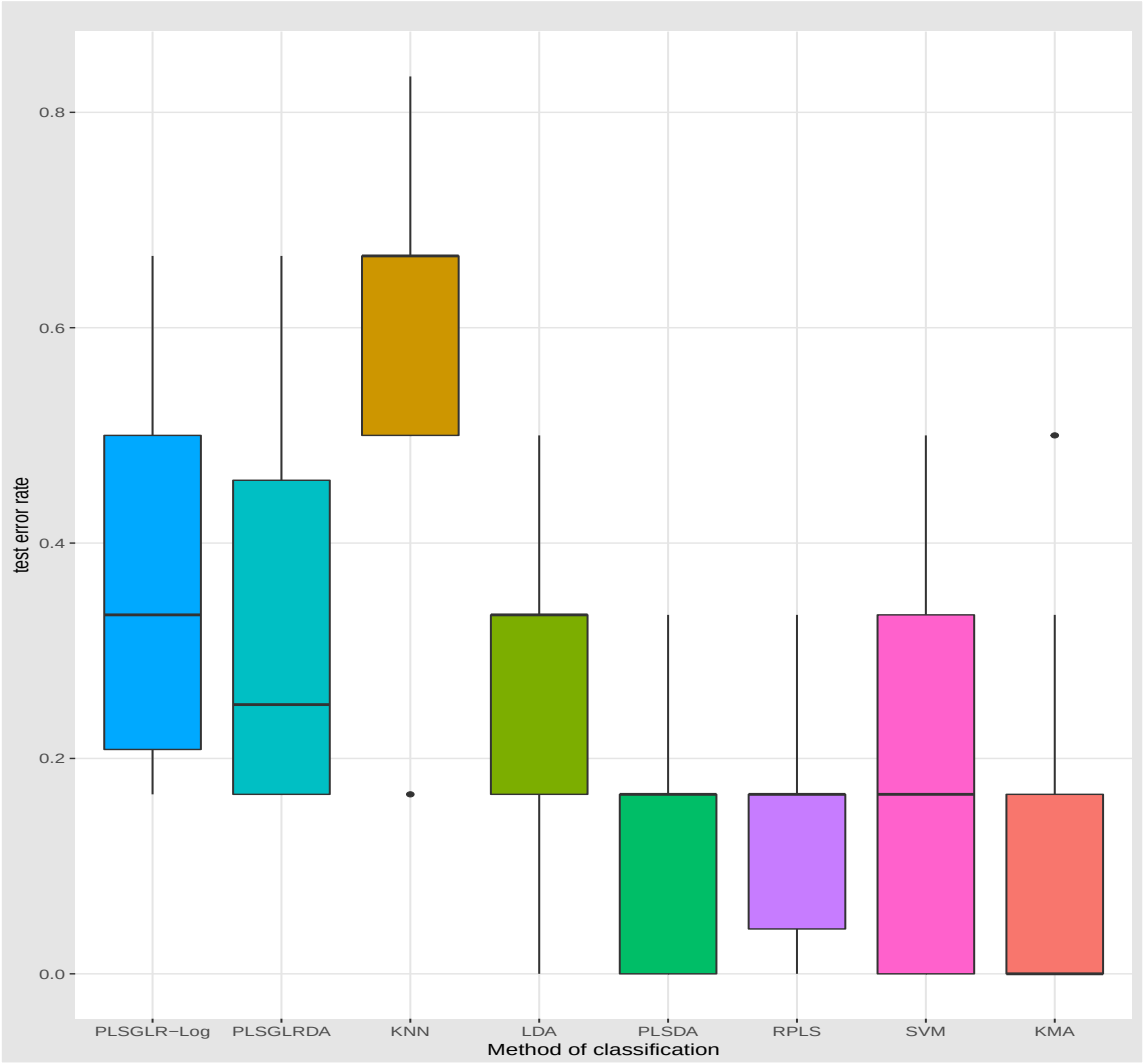


Figure 3.6: Box plots for the error rates for the unpreprocessed Colon data. The errors for all the classifiers are not symmetric except the SVM. The boxplots confirm that the top best classifiers are the KMA, PLSDA and RPLS. The KNN is outstandingly performing poor.

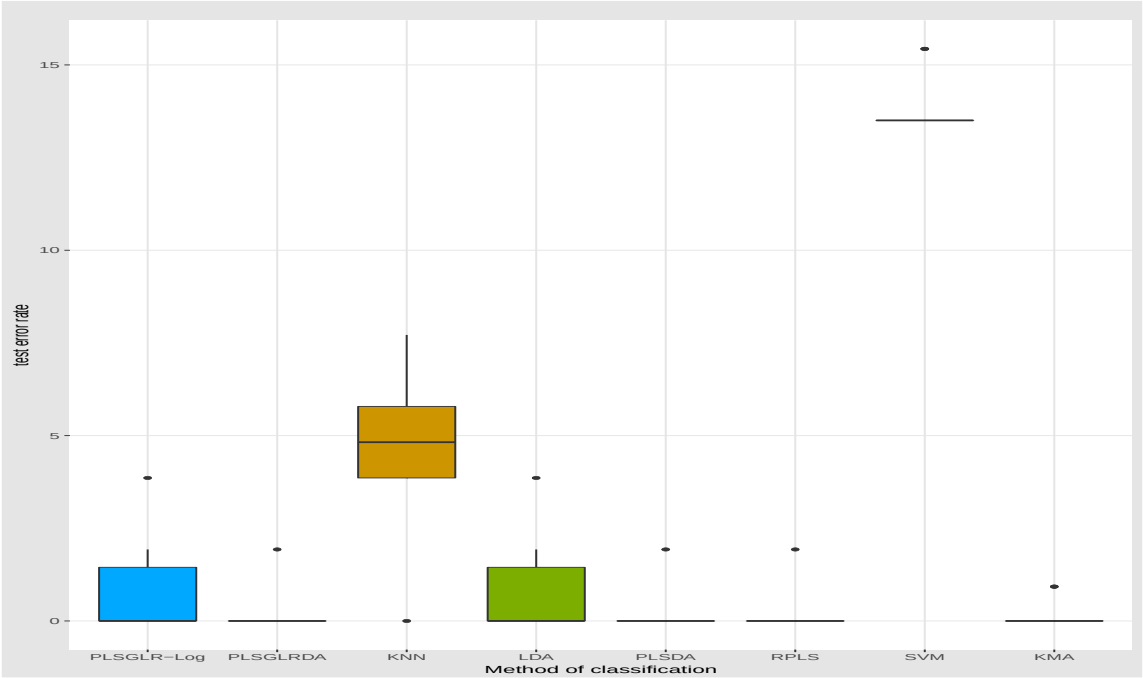


Figure 3.7: Box plots for the error rates for the unpreprocessed Leukemia data. This set of data had a relatively lower rate of missclassification. The best classifiers are the KMA, PLSDA, RPLS and PLSGLRDA. The KNN remains consistent in it's poor performance.

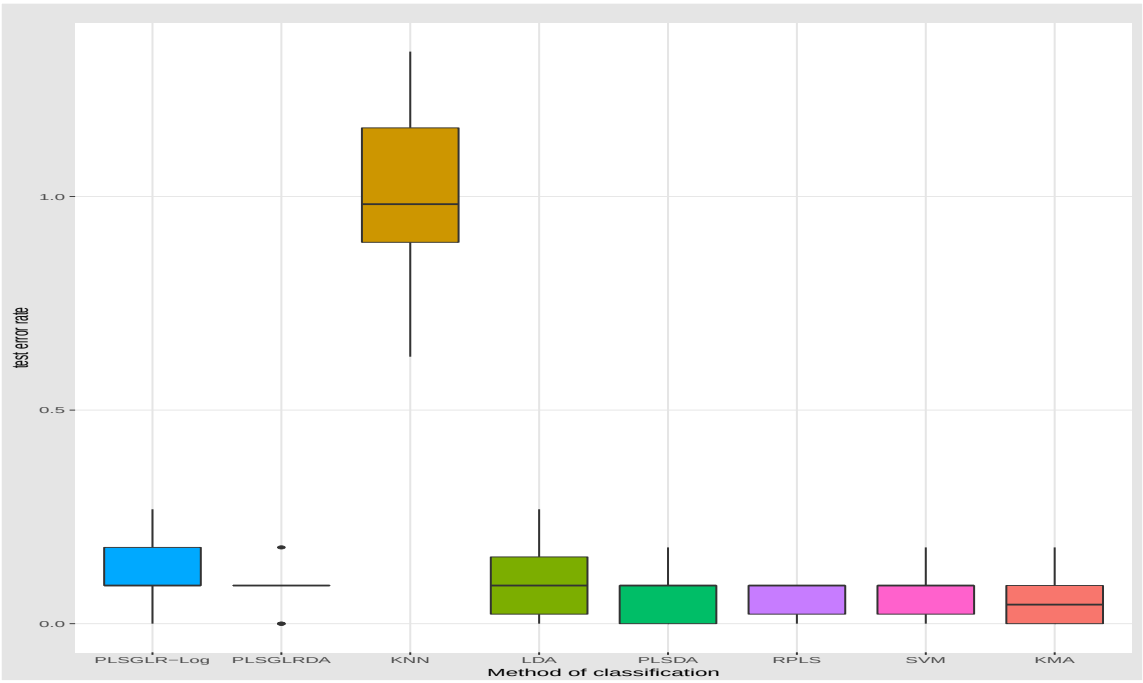


Figure 3.8: Box plots for the error rates for the unpreprocessed Prostate data. The best methodologies remain the KMA, PLSDA, RPLS and SVM. The KNN remains the worst.

3.2.3 Feature selection

Feature selection is a very important step in microarray data analysis because out of the thousands of variables (genes) generated, only a handful play an important role towards the biological problem of interest. The thousands of data points, are likely to be noisy due to biological or technical reasons. Thus the feature selection extracts a subset of the genes that are most informative thereby reducing the noise in the data and at the same time improving the efficiency of the classifiers. It seeks to reduce the number of features by targeting an optimum subset of features and removing the irrelevant or redundant features (Awada et al., 2012; Dudoit et al., 2002). Most commonly used feature selection methods involve ranking the genes based on some value of a univariate statistic like t-statistic, F-statistic, Wilcoxon, and Kruskal-Wallis statistics. A cut-off point based on either the number of genes or p-value is imposed so as to determine the number of variables to be used. Dudoit et al. (2002) suggest a gene selection method based on ranking. This is achieved by finding the ratio of between-group to within-group sum of squares (BSS/WSS) so that for a gene j,

$$BSS_j/WSS_j = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2} \tag{3.1}$$

where $\bar{x}_{.j}$ and $\bar{x}_{kj}$ are the average expression level of gene j and across all samples in class k respectively. The p genes with the biggest ratio are selected. In this study, we adopted the Dudoit et al. (2002) method of feature selection.

3.2.4 Analysis of the preprocessed data sets

Each data set was divided at random into a training and test sets of approximately 2:1 samples respectively. For instance, if a data set had n observations, then it was split into the training n_1 and test n_2 sets respectively where $n = n_1 + n_2$. The classifiers were then built using the n_1 training set and then prediction/classification done using the n_2 test set. The proportion of the misclassified labels are determined and thus the method with the lowest misclassification proportion was judged to be the best. A re-randomization study was implemented by repeating splitting of the data 100 times by re-sampling the samples and thereafter the proportion of misclassification was measured for each subdivision.

The preprocessing was done on the training set using the recommendations of Dudoit et al. (2002). Furthermore, the gene selection was done following the steps recommended by Dudoit et al. (2002). Both the preprocessing and gene selection were implemented in the `R` package `plsgenomics`. The top p genes were thus selected using Equation 3.1 for the implementation of the classification methods. The number of the top p genes to be used was arbitrarily chosen. The same procedure was repeated for all the other methods under consideration.

The colon and the leukemia data sets were preprocessed using the recommendations of Dudoit et al. (2002) while the Prostate data was obtained from the website `http://stat.ethz.ch/~dettling` in a preprocessed data by (Dettling and Bühlmann, 2002). Colon data had $p = \{50, 100, 500, 1000\}$, leukemia data had $p = \{100, 300, 500, 1000\}$ and the Prostate data had $p = \{100, 300, 500, 1500\}$.

3.2.5 Results and discussions for analysis of preprocessed data

As described in the introduction of Subsection 3.2.4, each data set was subdivided into two sets namely, the training set and the test set at a ratio of 2:1 for training and testing respectively. A resampling study of 100 random subdivisions was done and the test error for each subdivision summarized in tables and the box plots.

For the determination of PLSGLR components, the p genes were selected from the preprocessed data set, the first stage thus was to determine the genes that significantly contribute to the response variable. This was done by running separate logistic regressions for each of the selected p genes. The logistic regressions that were significant at 0.05 level were retained for building the first component t_1. The hard thresholding was implemented in which the non significant coefficients of the logistic regressions were recoded to zero in order to eliminate their effects in the construction of the PLS components. The PLSGLR algorithm was then implemented for the remaining components and stopped when there were no more significant GLM coefficients. The computed PLS components were then used in the PLSGLR-log and the PSGLRDA.

The classical classification methods namely, the k-nearest neighbour (KNN), the linear discriminant analysis (LDA), the ridge partial least squares (RPLS) , the partial least squares linear discriminant analysis (PLS-LDA) and SVM were also

implemented in order to compare them with the proposed methods. To choose k in the KNN, a 10-fold cross-validation was used and k that leads to the smallest number of misclassifications was chosen. For the SVMs, the cost parameter C and parameter gamma were determined through the 10-fold cross validation. Furthermore, the choice of kernel was influenced by the misclassification rates for different kernels so that the kernels that produced the lowest misclassification were preferred. The results for the analysis of Colon data are presented in Table 3.2.

Table 3.2: Percentage Misclassification for the Colon Data Set

p	PLSGLR-log	PLSGLRDA	KNN	LDA	PLSDA	RPLS	SVM	KMA
50	15.1	14.1	21.0	27.0	13.4	15.3	17.7	14.1
100	17.0	14.8	20.5	17.8	13.9	15.6	17.2	14.6
500	20.8	16.7	20.2	14.5	13.7	15.7	16.4	15.2
1000	20.8	16.8	21.2	14.8	13.9	15.1	16.8	14.6
Average	18.4	15.6	20.7	18.5	13.7	15.4	17.0	14.6

Table 3.2 shows that for classification of the colon data set, the PLSDA had the lowest percentage misclassification followed by the KMA, RPLS and PLSGLRDA. The difference in the percentage misclassification for the four top (best) classifiers is however marginal and so we can say that the three perform in almost the same manner. The highest misclassification percentage was observed with the KNN method. In this data set, the LDA performed poorly just like the KNN. Next we examine how each classifier performed in terms of False Positive and False Negative proportions.

Table 3.3: The proportions of False Positives and False Negatives for the Colon Data

p	50		100		500		1000		Average	
Method	C.C.N	N.C.C	C.C.N	N.C.C	C.C.N	N.C.C	C.C.N	N.C.C	C.C.N	N.C.C
PLSGLR-Log	6.9	8.0	7.8	9.2	9.7	11.10	9.6	11.1	8.5	9.9
PLSGLRDA	8.5	5.6	9.1	5.7	10.0	6.7	9.62	7.1	9.3	6.3
KNN	9.1	11.8	9.2	11.3	8.7	11.5	9.2	12.0	9.1	11.7
LDA	15.7	11.3	9.7	8.1	8.3	6.2	8.6	6.3	10.6	8.0
PLSDA	6.7	6.7	6.7	7.2	6.5	7.1	6.6	7.2	6.6	7.1
RPLS	8.4	6.9	9.1	6.6	9.0	6.76	8.5	6.6	8.8	6.7
SVM	10.6	7.1	10.1	7.1	9.7	6.7	10.5	6.2	10.2	6.8
KMA	6.7	7.4	7.6	7.1	8.5	6.7	7.8	6.9	7.7	7.0

In Table 3.3 for the Colon data, we look at the proportion of Cancer tissues

that were classified as normal abbreviated as C.C.N and the proportion of normal tissues classified as cancer (N.C.C). In our opinion, it is much worse to classify a cancer tissue as normal than to classify a normal tissue as cancerous. As a result, a classifier with a higher proportion of C.C.N performs relatively poorly compared to the one with higher proportion of N.C.C. In comparing the misclassification rates and proportions of C.C.N, then PLSDA and KMA emerge as the better options among the top four classifiers with lower misclassification rates. However, the RPLS and the PLSGLRDA equally performed well with regards to proportion of C.C.N. For the two methodologies, the difference in the proportions of C.C.N were small and thus a clear 'winner' was not established.

Table 3.4: Percentage Misclassification for the Leukemia Data Set

p	PLSGLR-log	PLSGLRDA	KNN	LDA	PLSDA	RPLS	SVM	KMA
100	4.1	4.0	5.7	3.5	4.3	2.7	2.6	4.2
300	4.4	3.1	5.6	1.9	2.7	1.9	1.9	2.4
500	3.4	3.2	4.8	2.6	2.4	1.5	1.2	1.9
1000	3.0	3.2	2.9	4.2	1.5	1.2	1.1	0.0
Average	3.7	3.4	4.8	3.1	2.7	1.8	1.7	2.1

Table 3.4 presents the misclassification percentages for the Leukemia data. This data set is perceived to be "easy" or "less problematic" to classify so that it is possible to achieve excellent classification accuracy in this data set even with trivial methods, see (Fort and Lambert-Lacroix, 2005; Boulesteix, 2004). In our case, SVM had the lowest misclassification percentage followed by KMA, RPLS, PLSDA, LDA and the PLSGLR-log while KNN had the highest percentage. Once again, the difference between the top best methods are not big. It is worth noting that unlike in the colon data see Table 3.3, the SVM performed very well in terms of the misclassification error rates. The leukemia data was used to classify two types of cancers namely; acute lymphoblastic Leukemia (ALL) and acute myeloid Leukemia (AML). The proportions of ALL classified as AML or otherwise are next presented in Table 3.5.

Table 3.5: Proportions for types of misclassification for the Leukemia Data

p	100		300		500		1000		Average	
Method	LCM	MCL	LCM	MCL	LCM	MCL	LCM	MCL	LCM	MCL
PLSGLR-Log	2.4	1.7	1.9	2.5	0.9	2.5	0.77	2.2	1.5	2.2
PLSGLRDA	3.2	0.8	2.8	0.3	0.3	2.9	3.08	0.2	2.3	1.0
KNN	5.7	0.0	5.6	0.0	4.8	0.0	2.92	0.0	4.8	0.0
LDA	2.1	1.5	1.4	0.5	2.2	0.4	3.31	0.9	2.3	0.8
PLSDA	2.8	1.5	1.7	1.0	1.7	0.7	0.85	0.7	1.8	1.0
RPLS	1.2	1.5	0.8	1.1	0.8	0.7	0.69	0.5	0.9	0.9
SVM	1.2	.5	0.9	0.9	0.5	0.7	0.6	0.5	0.8	0.65
KMA	3.5	0.6	1.9	0.5	1.5	0.4	0.0	0.0	1.7	0.38

LCM: lymphoblastic leukemia (ALL) classified as myeloid leukemia (AML); MLC: myeloid leukemia (AML) classified as lymphoblastic leukemia (ALL)

Table 3.6 presents the results for the Prostate data set. The task here was to classify two tissues as either tumors or non tumor. The results are similar to the one for Colon data whereby the best methods were SVM, RPLS, PLSDA, PLSGLRDA and KMA. Once again, the difference in misclassification percentages is minimal between the top four methods. The KNN once again emerged as the worst option due to its high misclassification rate, a result consistent with Fort and Lambert-Lacroix (2005).

Table 3.6: Percentage Misclassification for the Prostate Data Set

p	PLSGLR-log	PLSGLRDA	KNN	LDA	PLSDA	RPLS	SVM	KMA
100	10.6	8.5	11.9	18.5	7.3	9.2	10.0	10.9
300	12.1	8.7	15.8	8.9	7.0	8.4	8.8	13.7
500	12.1	9.3	18.6	8.5	7.1	8.4	8.2	13.5
1500	12.4	10.3	22.1	8.4	7.6	8.2	8.2	0.0
Average	11.8	9.2	17.1	11.1	7.3	8.5	8.8	9.5

Table 3.7 presents the types of misclassifications with the non tumor classified as tumor abbreviated as N.C.T while the tumor classified as non tumor is abbreviated as T.C.N. We once again suggest that a good classifier should have a lower proportion of T.C.N. From the table, the SVM, RPLS and PLSDA methods had a relatively lower proportion of T.C.N followed PLSGLRDA and KMA while KNN had a relatively high proportion of the same.

Table 3.7: The proportions False Positives and False Negatives for the Prostate Data

p	100		300		500		1500		Average	
Method	TCN	NCT	TCN	NCT	TCN	NCT	TCN	NCT	TCN	NCT
PLSGLR-Log	5.8	4.8	6.5	5.5	6.7	5.5	7.15	5.3	6.5	5.3
PLSGLRDA	5.1	3.4	5.0	3.7	4.9	4.4	5.32	4.9	5.1	4.1
KNN	9.5	2.4	13.6	2.2	16.2	2.4	18.71	3.4	14.5	2.6
LDA	9.6	8.9	5.2	3.6	4.9	3.7	5.38	3.1	6.3	4.8
PLSDA	4.5	2.8	4.0	3.0	4.2	3.0	4.71	2.9	4.4	2.9
RPLS	4.6	4.5	4.2	4.1	4.2	4.2	4.21	4.0	4.3	4.2
SVM	4.3	5.7	3.9	4.91	3.7	4.5	3.4	4.7	3.8	5.0
KMA	5.3	5.7	7.6	6.1	7.6	6.0	0.0	0.0	5.1	4.5

NCT: Non Tumor Classified as Tumor; TCN: Tumor classified as Non Tumor

Now focusing on the box plots Figures 3.9, 3.10, 3.11 for the three sets of data with the corresponding classification method. For the Colon data, the error rates for PLSGLRDA, PLSDA and RPLS are generally lower than the rest. In the case of Leukemia data, all the classification methods seem to have low test errors and at the same time have more or less similar distribution of the errors. Furthermore, the RPLS and the PLSDA exhibit the lowest error rates. For the Prostate data set, the PLSGLRDA, PLSDA and RPLS have an almost similar distribution of classification errors.

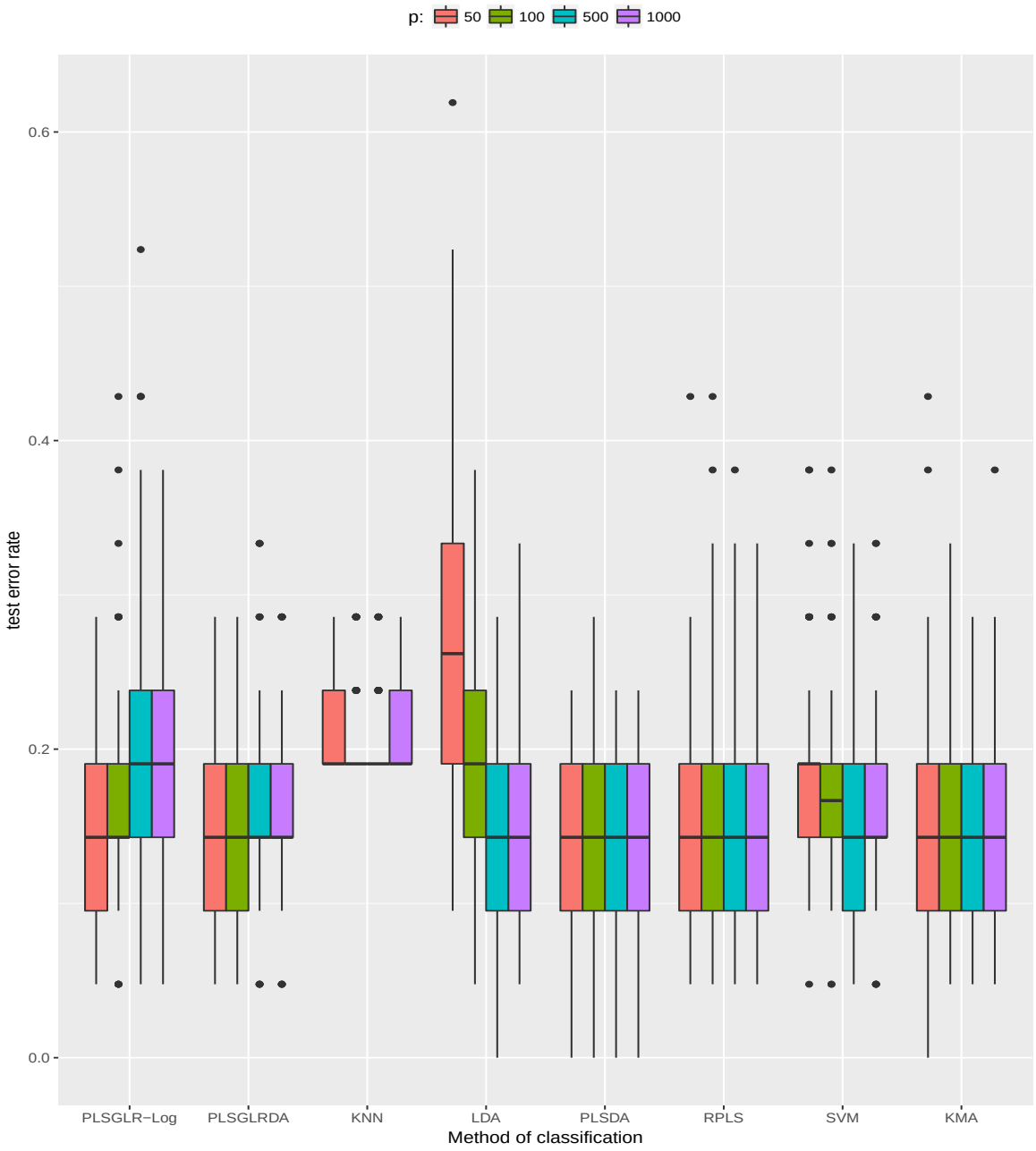


Figure 3.9: Box plots for the Test Errors for the Colon Data. The box plot shows that the distribution of errors for PLSDA and RPLS are the same for all the number of genes selected. The distributions are symmetric with the mean below 0.2. The LDA has a higher error rate for $p = 50$ while PLSGLRDA has higher error rates for $p = \{500, 1000\}$.

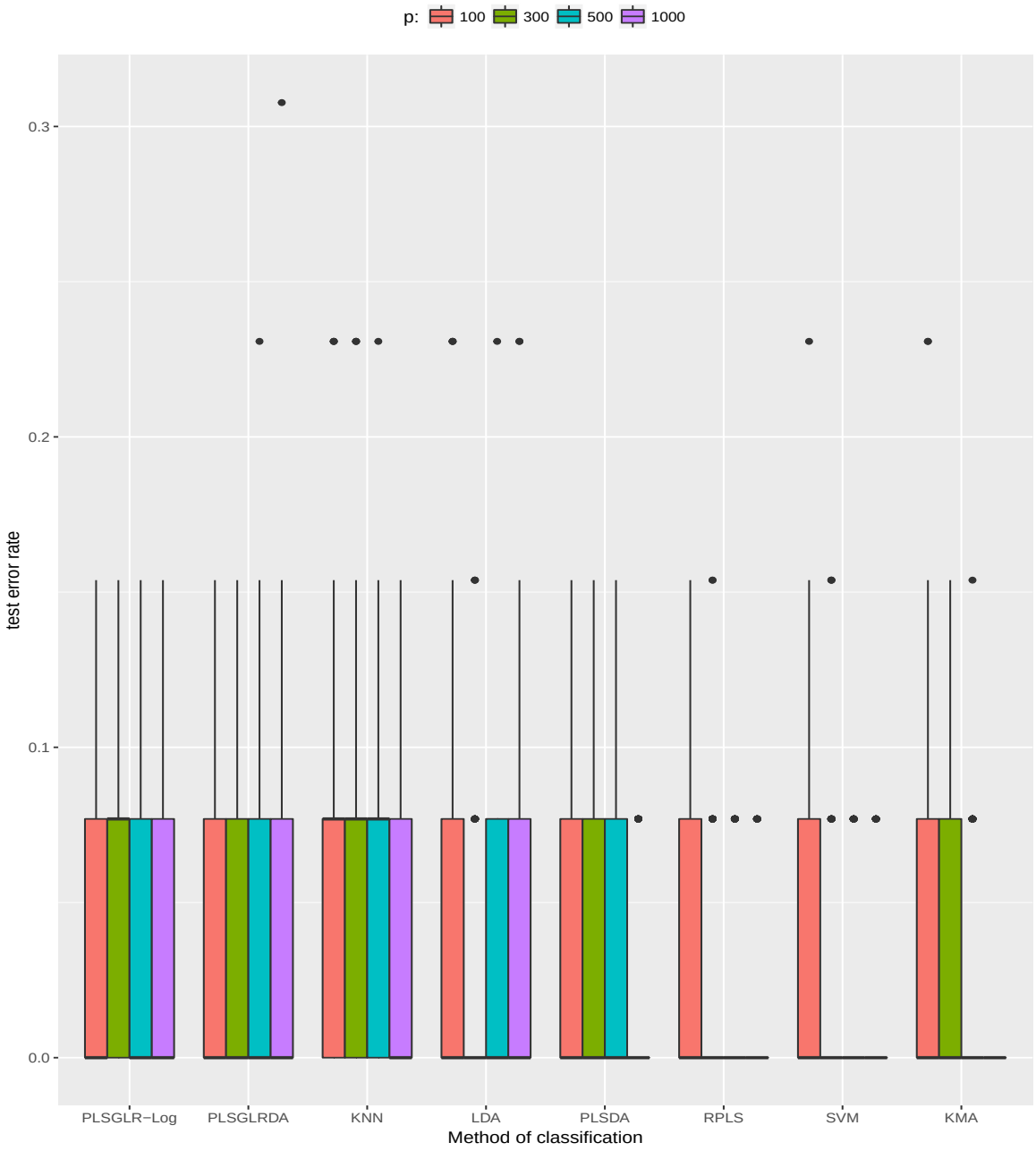


Figure 3.10: Box plots for the Test Errors for the Leukemia Data. For the leukemia data, most of the classifiers have a low miss-classification rate. The RPLS and SVM have very low error rates for $p = \{300, 500, 1000\}$.

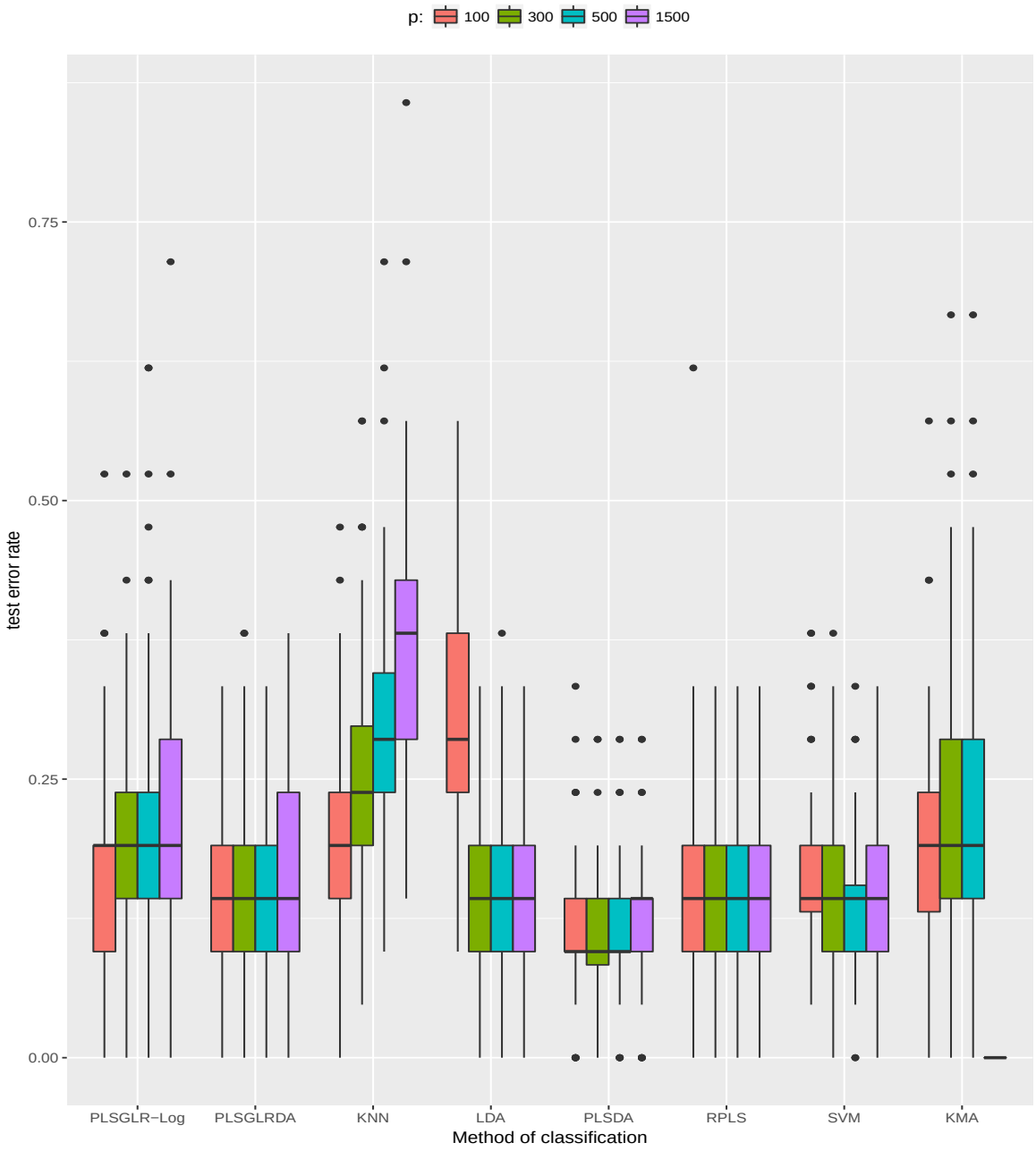


Figure 3.11: Box plots for the Test Errors for the Prostate Data. The outstanding bars are the ones for KNN which seem have a relatively higher rate of mis-classification and this rate increases with the increase in the number of genes p selected. In general, the test error rate for this data set for all the algorithm have many outliers.

3.3 Summary on classification methods

For the un filtered/un-preprocessed Colon data without gene selection, KMA emerged as the best with lowest misclassification error followed by PLSDA and RPLS while KNN, PLSGLR-log and PLSGLRDA were the worst. For leukemia data, the KMA emerged the best followed PLSGLRDA, PLSDA, RPLS while KNN remained the worst. For Prostate data, KMA was the best followed by PLSGLRDA, PLSDA, SVM and PLSGLRDA-log while KNN performed the worst. The KMA perform very well for the large sample sizes compared to when the data has a smaller sample size.

The results suggest that KMA and the suggested PLSGLR methodologies perform well in the presence of noise and with many features. The KMA and the suggested PLSGLR extensions have proven to perform well in the presence of noise without variable selection. However, the KMA emerges as a clear “winner" in the un-preprocessed data. Furthermore The KMA and the suggested PLSGLR extensions are relatively simple to implement.

Furthermore, we have extended the Bastien et al. (2005) PLS Generalized to PLS Generalized Linear Regression-Linear Discriminant Analysis (PLSGLRDA) using the two step procedure consisting of dimension reduction followed by application of standard statistical procedures like logistic regression and the LDA to classify three microarray data sets namely the Colon, Leukemia and the Prostate data sets. Our proposed combination (PLSGLRDA) has proved to be competitive with the RPLS, SVM, PLSDA and so can be used as an alternative classification method for the classification problems in microarray data since it equally easier to implement and perform as well as the existing classical methodologies. The PLSGLR therefore can be considered as an important addition to the existing classes of data reduction methodologies in the microarray data analysis and other data types with similar structure as the microarray data. It is important to note that KNN, LDA and PLSGLR-Log consistently performed poorly in terms of misclassification errors and as such are the worst option for these kinds of data considered in this study.

Chapter 4

Statistical integration of molecular data to test for network changes

4.1 Introduction

The work in this section is motivated by ongoing research on the acute rheumatic fever (ARF), an autoimmune disease which is consequence of infection with *group A streptococcus* (GAS) at the Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia. Among others symptoms, ARF causes an acute generalised inflammatory response and an illness that selectively affects the heart, joints, brain and skin. It is important to note that despite the acute nature of the infections, the ARF does not leave lasting damage to the brain or joints but the leaves long term damage to heart valves leading to a condition known as known as rheumatic heart disease (RHD) which can become a chronic condition leading to congestive heart failure, strokes, endocarditis which is inflammation of the inside lining of the heart chambers and heart valves (endocardium). ARF may also cause death (Carapetis et al., 2007; Seckeler and Hoke, 2011).

GAS, also known as *Streptococcus pyogenes* is a human-restricted pathogen that can spread through direct contact with the mucus or sore skin and is responsible for a wide range of both invasive and noninvasive infections such as commonly mild superficial infections of the pharynx, skin and life-threatening ones such as necrotizing fasciitis is a serious skin infection, spreads quickly and kills the body's soft tissue. GAS infection leads to acute rheumatic fever (ARF), sudden appearance of red blood cells in urine and probably, paediatric autoimmune neuropsychiatric

disorders associated with streptococcal infections (Martin et al., 2015).

The ARF is usually diagnosed using the Jones criteria Jones (1944) which involves the major and minor signs. Diagnosis is made by the presence of either two major or one major and two minor criteria in addition to evidence of recent streptococcal infection.

Table 4.1: Major and minor Jones criteria for the diagnosis of acute rheumatic fever

Major	Minor
Migratory polyarthritis	Arthralgia
Carditis	Fever
Erythema marginatum	First-degree heart block
Sydenham chorea	Elevated markers (ESR, CRP)
Subcutaneous nodules	
CRP: C-reactive protein, ESR	Erythrocyte sedimentation rate

Martin et al. (2015) explain that the evidence of previous streptococcal infection is usually confirmed by elevated or rising serum antibodies to streptococcal antigens, such as streptolysin O and deoxyribonuclease B. This is because the throat culture tends to be negative for GAS for the ARF patients. ARF is usually treated with a goal to eradicate streptococcal organisms and bacterial antigens from the pharyngeal region using penicillin for persons who are not at risk of allergic reaction.

The ARF and RHD are not common in the developed countries where there is proper hygiene, improved nutrition, less crowding and access to medical facilities. However, the disease is still prevalent in the developing countries and amongst the poor, mainly indigenous populations in the developed countries including Australia. Despite an obvious clinical need, there is no definitive method for diagnosing ARF, in fact the current diagnosis lacks specificity and sensitivity. Current blood tests to assist in the diagnosis of ARF involve the measurement of antibody titres to streptococcal antigens, streptolysin O and DNase B. These markers increase in numerous group A streptococcal infections that do not lead to ARF resulting into uniformly high background titres of these antibodies in the remote Australian Aboriginal communities. Misdiagnosis of ARF is a major contributor to the high rates of RHD seen in Aboriginal communities in the Northern Territory, Australia, with 29% of patients with supposed “primary" episodes of ARF already having established RHD. Timely diagnosis of an initial ARF episode and subsequent use of antibiotic prophylaxis is currently the best method of preventing RHD. There is

currently no treatment for the acute episode of ARF that alters the development of RHD. In clinical trials, non-steroidal anti-inflammatory medications and corticosteroids do not appear to have long-term benefit (Carapetis et al., 2005, 2007; Martin et al., 2015).

Research on ARF is very active and several experiments are being carried out to achieve different objectives. However, despite decades of research, there is still no diagnostic test or vaccine for ARF. Of interest is the research that has been going on in the Prof. Ian Wicks lab at the Walter and Eliza Hall Institute of Medical Research. The researchers in this lab have been working towards understanding the type of inflammation occurring in ARF patients with the ultimate goal being to find new diagnostic markers to diagnose ARF and new drugs.

4.2 Example of a pilot laboratory experiment data

The RNA-seq data considered in this section of the thesis is from pilot experiments done at the Wicks Laboratory, Inflammation Division at the Walter and Eliza Hall Institute of Medical Research (WEHI), Melbourne, Australia and was partially analyzed at Smyth Lab, Bioinformatics Division in the same institute. The experiment involved two groups of subjects namely the healthy donors and the ARF patients.

The experiment involved two groups of subjects namely the healthy and the ARF. The healthy subjects were 3 while the ARF ones were 25 Aboriginal people. Out of the ARF group, 18 had were ARF patients, 5 were healthy controls (no ARF infection) and 2 had alternate diagnoses - RHD and a cardiac disease respectively. For each subject, a specimen of peripheral blood mononuclear cell (PBMC) was taken and divided in two portions. One portion is GAS stimulated while the other portion remain unstimulated. The samples were then sequenced at the Australian Genome Research Facility (AGRF) on the Illumina HiSeq platform under five different conditions: unstimulated 0 hour, unstimulated 24 hours, GAS stimulated 24 hours, hydrochloroquine (HCQ) stimulated 24 hours and GAS & HCQ stimulated 24 hours. However, in this thesis our focus is on studying the PBMCs GAS stimulated 24 hours vs unstimulated 24 hours.

4.3 RNA-seq data

Recently, in many experiments gene expressions levels are usually measured through technologies like the next generation sequencing technology that produces the RNA-seq data. The RNA-seq data is usually summarized by a data matrix consisting of counts. The count data matrix for this can conveniently presented in a format such as in Table 4.2.

Table 4.2: Illustration of a table of read counts for the two categories of the samples and groups.

	Healthy subjects					
	GAS stimulated			Unstimulated		
	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
gene 1	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
gene 2	h_{21}	h_{22}	h_{23}	h_{24}	h_{25}	h_{26}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
gene p	h_{p1}	h_{p2}	h_{p3}	h_{p4}	h_{p5}	h_{p6}
Library size	N_1	N_2	N_3	N_4	N_5	N_6

	ARF subjects					
	GAS stimulated			Unstimulated		
	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6
gene 1	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}
gene 2	a_{21}	y_{22}	a_{23}	a_{24}	a_{25}	a_{26}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
gene p	a_{p1}	a_{p2}	a_{p3}	a_{p4}	a_{p5}	a_{p6}
Library size	K_1	K_2	K_3	K_4	K_5	K_6

Table 4.2 shows the read counts per sample often referred to as the library. The total number of reads on the other hand is known as the library size which in this case are $N_i, i = 1, 2 \ldots, 6$ and $K_i, i = 1, 2 \ldots, 6$ healthy and ARF subjects respectively. The aim of such an experiment is to identify the genes that are differentially expressed between the treated samples and the untreated ones. The samples in these matrix are paired even though the labelling is continuous. The pairing is usually captured

by the design matrix during the data analysis. Same genes are considered in both the the two categories i.e. for healthy and the ARF subjects. Typically in such experiments a total number of $p = 20,000$ genes are considered.

Considering the healthy subjects then using the approach of Chen et al. (2014), let h_{gi} denote the read count for a particular gene g in the ith sample given the experimental conditions. The sequence depth is then given by

$$E(h_{gi}) = \mu_{gi} = \lambda_{gi} \cdot N_i, \tag{4.1}$$

where N_i is the library size and λ_{gi} is the expected proportion of reads mapped to gene g for the treated and unstimulated groups respectively. Denote the parameters associated with the treated samples as $\lambda_{g1} = \lambda_{g2} = \lambda_{g3} = \lambda^t$ while the ones associated with the untreated samples as $\lambda_{g4} = \lambda_{g5} = \lambda_{g6} = \lambda^u$, where g=(gene 1,..., gene p), then the hypothesis to be tested for a differential analysis can be formulated as

$$H_o : \lambda^t = \lambda^u \text{ versus } H_1 : \lambda^t \neq \lambda^u. \tag{4.2}$$

The read count h_{gi} is usually assumed to follow a generalized linear model of the negative binomial family with a logarithmic link $h_{gi} \sim \text{NB}(\mu_{gi}, \alpha_g)$ where μ_{gi} is the mean and α_g the dispersion parameter. The parameter λ_{gi} is assumed to be presented by a log-linear model

$$\log\lambda_{gi} = x_i^T \beta_g, \tag{4.3}$$

where x_i is a vector indicating different treatment conditions applied to sample i while β_g is a vector of regression coefficients for covariate effects for gene g. Consequently, we have

$$\log\mu_{gi} = x_i^T \beta_g + \log N_i. \tag{4.4}$$

Now, putting the covariates x_i into a design matrix X, the vector of linear predictors for gene g is given by $X\beta_g$ and therefore in our example of treated/unstimulated conditions, the regression coefficients will be $\beta_g = [\beta_{g1}, \beta_{g2}]'$ where β_{g1} and β_{g2} represent the log-expression for the GAS stimulated and unstimulated samples respectively. The parameters $\beta_{g1} = \log\lambda_g^t$ and $\beta_{g2} = \log(\lambda_g^t/\lambda_g^u)$ leading to the

hypothesis as

$$H_0 : \beta_{g2} = 0 \text{ versus } H_0 : \beta_{g2} \neq 0 \tag{4.5}$$

for all the genes. In other words we are testing the null hypothesis that the logarithmic fold change between the treatment and control is for a gene g is zero. This set of hypotheses can be tested using the asymptotic chi-square approximation to the likelihood ratio statistic and can be implemented using a `Biocundoctor` software package like `edgeR` by (Chen et al., 2014).

4.4 Statistical integration of molecular data

Numerous studies have shown that genes tend to act in groups and that the ones which have correlated expression changes over different conditions are likely to be involved in similar functional or cellular processes. This is because they most likely share DNA sequence elements and are therefore regulated by common transcription factors. These relationships between genes from a given functional or cellular groups can be represented as a network. There exist many biological networks in the literature and online data bases that provide important information about the protein-protein interactions, gene networks, functional pathways among others. The molecular networks and interactions provide a convenient way to study the changes in gene expression and integration of a number of measurements (Ideker et al., 2001).

Several studies of gene expressions have revealed that the genes tend to interact with and respond to an organism's environment. Some genes are always expressed regardless of the stimuli in the organism's environment, while some tend to be turned on or off depending on the stimuli. Using the prior network from literature, we want to develop a statistical framework to test for some changes in the genes expression with regards to up regulation, down regulation or no changes.

The protein-protein interactions (PPI) or protein-DNA interactions have been widely used to integrate the network information (that is, interactions between the proteins/genes) with the 'omics' data to generate statistical hypotheses that reveal some underlying mechanisms observed in the gene expressions. This is usually done by identifying the most active hubs or subnetworks. The hubs in this case implies the nodes that are highly connected with other nodes or in other words are the nodes that have higher degree within a network or subnetwork.

Some specific examples include, Ideker et al. (2001) build, test and analyze changes to critical pathway components using DNA microarrays and quantitative proteomics. Their approach consist of four major steps with the first one being to define all of the genes and the subset of genes, proteins or any other molecules contained in the pathway of interest. An initial model of the interactions governing pathway functions is defined from existing literature. Next, each pathway component is perturbed by changing different experimental conditions using technologies and then the expression values are measured. In this step, a generalized likelihood ratio test (GLRT) is calculated for each gene to determine the ones whose mRNA levels differed significantly from the reference under some changes. The identified genes are then clustered using self organizing maps so that each cluster contain genes with similar expression responses over all changes see Ideker et al. (2001). To check whether the changes in mRNA expression are also reflected at the level of protein abundance, the protein abundance is examined under different experimental conditions then a ratio is calculated that is compared to that of the mRNA-expressions via correlation coefficient. The third step is to integrate the observed gene or protein data with the known network of protein-protein or protein-DNA interactions networks. Here, the authors assemble a network curated from existing literatures. The interactions between the genes from the mRNA expressions are then compared based on the information from the interactions in the catalogued network. Finally, new hypotheses are formulated to explain changes that are not predicted by the model.

The integrative approach has also been utilized by Ideker et al. (2002) who introduced an algorithm that uses a statistical scoring method which captures the changes in gene expression within a given set of genes to find subnetworks in which the connected sets of genes seem to have high levels of differential expressions. This methodology starts by calculating the p-values for the expression changes for each gene and then a $z-$score corresponding to each p-value is calculated. An aggregate $z-$score is then computed for a given subnetwork by finding the mean of the $z-$scores for all the genes within that subnetwork. A higher agregate $z-$score indicates a higher biological activity within a particular subnetwork. An algorithm using the Monte Carlo approach and simulated annealing is then devised to find the highest scoring subnetwork and at the same time capture the connection between expression and network topology.

Other integrative "omics" studies have been carried out by Taylor et al. (2009); Han et al. (2004); Schramm et al. (2013) and Jayaswal et al. (2013). Their techniques in general involve differential network mapping which combines the gene expression data with some predetermined interaction networks (for instance, protein-protein) from curated literature and or online high-throughput database sources to identify the differentially expressed interactions under different conditions (Schramm et al., 2013; Taylor et al., 2009). Once the interaction networks have been selected from the curated sources, the identification of the nodes that interact with many partners follows. These nodes that interact with more partners are referred to as the 'hubs' and the number of interacting partners that qualify a particular node to be a 'hub' is selected by the researcher. The next step is to quantify the interactions between the hubs and their interacting partners via some association measure depending on the number of conditions under consideration. For two conditions, a measure like the correlation coefficient may be used while for more than two conditions an F-statistic may be appropriate. Once a measure of association has been calculated, the estimate of p-value for the test statistic is calculated using a permutation test.

To develop ideas and motivation for this work, we use the pilot laboratory experimental expression data in Section 4.2 in conjunction with a prior PPI network information to illustrate the problem at hand. Specifically, we aim to integrate prior PPI network curated from literature and existing database with the experimental gene expression data in order to properly formulate a statistical framework for testing changes in a network.

We start by identifying a prior known network, which in this case we chose a comprehensive web resource, which includes a database of unified protein-protein interaction known as Protein Interaction Network Analysis (PINA) `http://omics.bjcancer.org/pina/` as the source of the prior network. The identification of the genes or proteins of interest that match for the experimental data and the known curated PINA network follows. In this research, the Th-specific genes were chosen based on previous studies for instance Zhu et al. (2010) as well as the recent study in the Wicks laboratory, WEHI showed that Th1 and Th17-like T cells played an important role in the pathogenesis of acute rheumatic fever and rheumatic heart disease.

The experimental data is thereafter mapped on to the known prior network using the gene list from the experimental data and the edge list is retrieved from the curated

prior network. The nodes are coloured according to the $\log_2$Fold Change ($\log_2$FC) calculated based on the experimental RNA-seq data to reflect whether a gene expression is up or down regulated or unchanged. An arbitrary threshold of $-0.5 < \log_2\text{FC} < 0.5$ is chosen but guided by looking at the corresponding number of genes that are either up or down regulated in order to make an informed decision. Subnetworks are then chosen based on functional groups of interest. Examples are given in Figures 4.1 and 4.2.

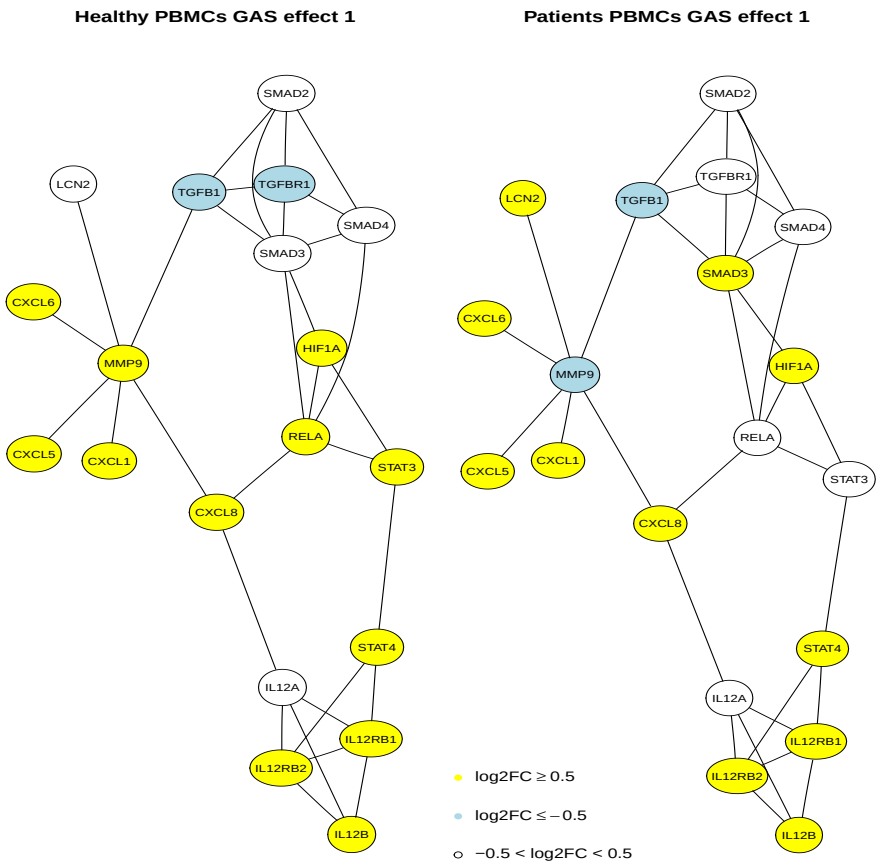


Figure 4.1: Sub network 1 consisting mainly of the functional group Th1/Th17. The choice of the genes used in this subnetwork 1 is based on other experiments carried out in the Wicks Lab that led to the hypothesis that pathogenic Th1/Th17 T cells are key mediators of the heart inflammation and damage in ARF. The edge information is obtained from the PINA network. The nodes (genes) are coloured in such a way that the yellow nodes represent upregulated genes, light blue nodes are downregulated genes (or groups of genes) while the white ones are neither up nor down regulated. We aim to develop a statistical framework to test for the changes in the similar genes for each of the subnetwork for healthy and ARF subjects. For example, in the healthy subjects network, the gene LCN2 is neither up or down regulated while in the ARF subjects the same gene is upregulated. Furthermore, in the healthy PBMCs, the gene RELA is upregulated while in the ARF patients PBMCs the same gene's average expression level is unchanged. It is these kind of changes we refer to as changes and we want to develop a statistical framework to test them.

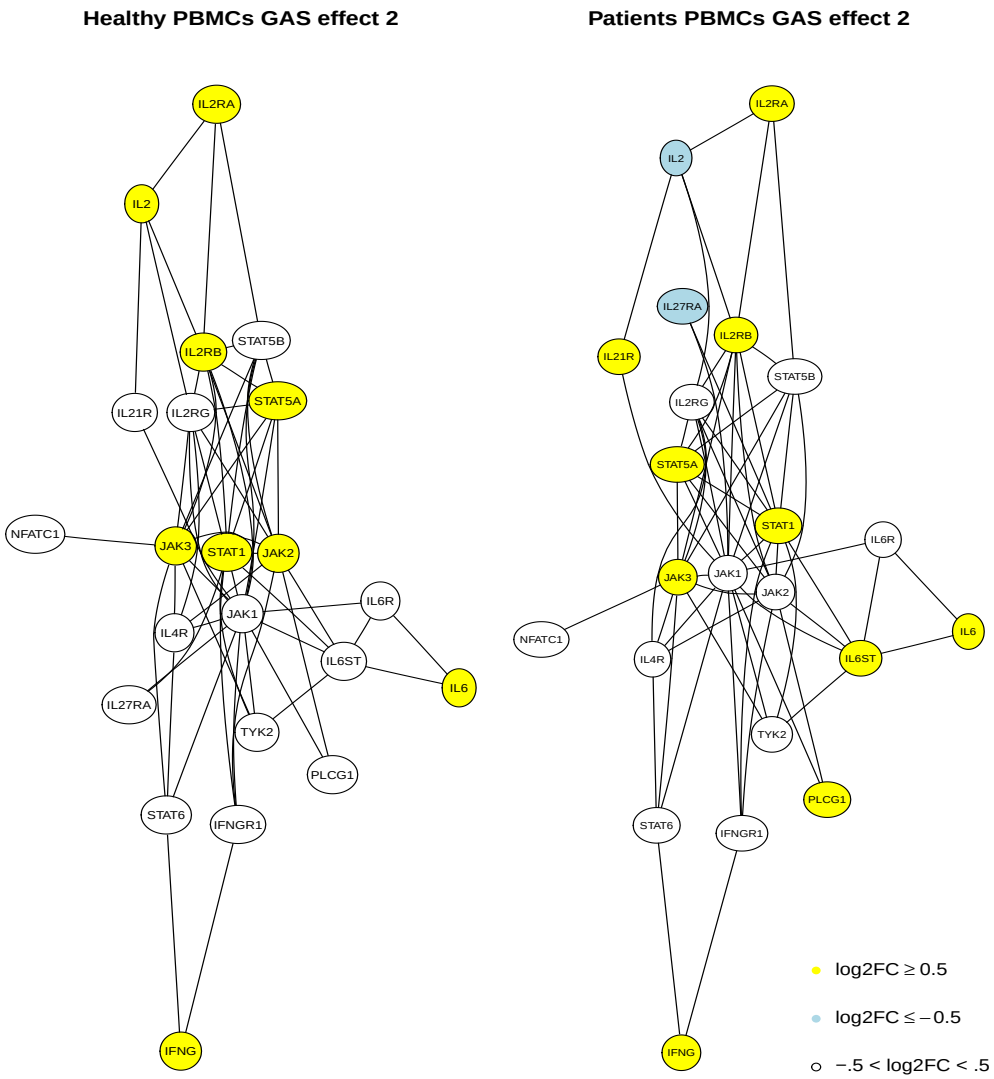


Figure 4.2: Sub network 2 for functional group Th2. Bhatnagara et al. (1999) also found out that chronic rheumatic heart disease (CRHD) patients secreted IL-4 and IL-10 in large amounts, i.e. Th2 type of cytokine profile. The genes in this subnetwork would help determine the changes of the Th2 group for the ARF patients and healthy subjects. The edge list is obtained from the prior network (PINA). The nodes (genes) are coloured in such a way that the yellow nodes are upregulated, light blue nodes are downregulated while the white ones are neither up nor down regulated. The aims and objectives for this figure are similar to those discussed for Figure 4.1.

4.5 The likelihood ratio testing

The theory of the likelihood ratio test is well understood and has been utilized extensively in the field of statistical inference. Most standard multivariate statistics

books like for example Anderson (2003); Seber (2004); Mardia et al. (1980); Johnson and Wichern (2007), to mention but a few, contain comprehensive treatment of this subject matter. In general, let $\boldsymbol{\theta}$ be the parameter vector for the likelihood function $L(\boldsymbol{\theta})$ with observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with a density function given by $f(\boldsymbol{x}; \boldsymbol{\theta})$. Now, if the parameter space is given by $\boldsymbol{\Theta}$ and suppose we want to test the null hypothesis $H_o : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ where $\boldsymbol{\Theta}_0$ is a subset of $\boldsymbol{\Theta}$. The parameter space $\boldsymbol{\theta}$ is referred to as unconstrained while $\boldsymbol{\theta}_0$ is constrained. The likelihood ratio statistic is given by

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_o} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta})}. \tag{4.6}$$

The null hypothesis H_o is rejected when $\Lambda < k$ where k is a critical value depending on the type-I error. The likelihood ratio test has good power properties asymptotically and usually is as good or better than many other test statistics Seber (2004). The log likelihood ratio (LRT) statistic under general conditions and with large samples are approximately $\chi^2_{(d)}$ distributed where d is the degree of freedom which in general is given by the total number of variables under consideration. The LRT is given by

$$-2\text{Log}\Lambda = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \{-2\text{Log L}(\boldsymbol{\theta})\} - \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \{-2\text{Log L}(\boldsymbol{\theta})\}. \tag{4.7}$$

Some outstanding common problems that have been tackled in the said standard multivariate statistics analysis books with regards to the likelihood ratio test include the following.

- Suppose we have N observations on $\boldsymbol{X}$ that is multivariate normally distributed according to $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, a test statistic is derived to test for the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ when Σ is unknown. The obvious MLE for Σ in this case is the sample covariance. The resultant test statistics is the T^2 statistics which follows the T^2 Hotelling distribution. This test can be used for testing the hypothesis about the mean vector $\boldsymbol{\mu}$ of the population and obtaining the confidence region for the unknown vector $\boldsymbol{\mu}$ see Anderson (2003); Seber (2004); Mardia et al. (1980); Johnson and Wichern (2007).

- The two sample problem with unequal covariance matrices is addressed. In this case, let $\left\{\boldsymbol{y}_j^{(i)}\right\}, j = 1, \ldots, N$ be samples from $\mathcal{N}(\boldsymbol{\mu}^{(i)}, \Sigma_i), i = 1, 2$ a test statistic for testing $H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$ is developed. The distribution for the respective sample mean vectors is given by $\text{E}(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)}) = \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}$ while

the covariance for the difference $\text{Cov}(\bar{\boldsymbol{y}}^{(1)} - \bar{\boldsymbol{y}}^{(2)}) = \frac{1}{N_1}\Sigma_1 + \frac{1}{N_2}\Sigma_2$. It is shown that when $N_1 \neq N_2$ and assuming that $N_1 < N_2$ then a suitable test would still be a T^2 test with $(N_1 - 1)$ degrees of freedom can be used, see Anderson (2003).

- When Σ_1 and Σ_2 are assumed to be equal and unknown, then a pooled sample covariance is used as an estimate. The test statistic is found to be the usual T^2 which follows the T^2 distribution see Anderson (2003); Seber (2004).

- The topic of paired comparisons is also treated especially in Johnson and Wichern (2007) in which for the paired samples, the difference between them is calculated. The T^2 test is then applied to the differences.

- Most of the likelihood problems tackled in these standard multivariate books only compare two mean vectors and the resultant statistic is the T^2 with a certain degree of freedom depending on the problem set-up.

4.5.1 Likelihood ratio test for network changes

To set-up a statistical framework for testing network changes, consider an experiment consisting of two groups namely Healthy (H) and ARF (A). The healthy (H) group has m subjects while the ARF (A) group has k subjects. For each group, two measurements are done for the same subject so that we have paired measurements. The measurements are administered in a similar manner between the groups. For instance, each subject has a measurement when it is unstimulated and when GAS stimulated for p different genes. As an illustration, for healthy subjects each gene has m paired measurements $[(h_{u1}, h_{s1}), (h_{u2}, h_{s2}), \ldots, (h_{um}, h_{sm})]$ where the first measurement is for unstimulated while the second one is for GAS stimulated specimen for the same subject. In a similar fashion, the ARF subjects have k paired measurements for each gene $[(a_{u1}, a_{s1}), (a_{u2}, a_{s2}), \ldots, (a_{uk}, a_{sk})]$ for the unstimulated and GAS stimulated specimens in each pair respectively. We wish to test the hypothesis that there is no difference in the difference of the means for the GAS stimulated and unstimulated subjects for healthy subjects and ARF patients. This kind of hypothesis would give an important insight in to how differently or similarly the healthy people and the ARF patients react to GAS treatment. The overall objective is to derive a likelihood ratio test for the p genes that are assumed to be

correlated based on some protein-protein interaction (PPI) network.

The m healthy subjects assumed to come from a multivariate normal distributions $\binom{\boldsymbol{h}_u}{\boldsymbol{h}_s} \sim \mathcal{N}_{2p}\left[\binom{\boldsymbol{\mu}_u}{\boldsymbol{\mu}_s}, \boldsymbol{\Sigma}\right]$ and the k ARF subjects are also assumed to come from a multivariate normal distribution $\binom{\boldsymbol{a}_u}{\boldsymbol{a}_s} \sim \mathcal{N}_{2p}\left[\binom{\boldsymbol{\nu}_u}{\boldsymbol{\nu}_s}, \boldsymbol{\Sigma}\right]$.

The matrix $\boldsymbol{\Sigma}$ can be partitioned as $\left[\begin{array}{c|c} \Sigma_{11} & C_{12} \\ \hline C_{21} & \Sigma_{22} \end{array}\right]$. Note that $C_{21} = C_{12}'$ and we assume that these matrices of covariance within each of the groups are equal. In the block matrix, Σ_{11} and Σ_{22} represent the variance-covariance matrices for the unstimulated and GAS stimulated subjects respectively.

The hypothesis to be tested is

$$H_0 : (\boldsymbol{\mu}_u - \boldsymbol{\mu}_s) = (\boldsymbol{\nu}_u - \boldsymbol{\nu}_s) \text{ versus } H_a : (\boldsymbol{\mu}_u - \boldsymbol{\mu}_s) \neq (\boldsymbol{\nu}_u - \boldsymbol{\nu}_s).$$

Case 1: Assuming the covariance matrix Σ is known

For m healthy subjects denote a $2p \times 1$ vector of parameters $\boldsymbol{\mu} = \binom{\boldsymbol{\mu}_u}{\boldsymbol{\mu}_s}$ for the random vector $\boldsymbol{h} = \binom{\boldsymbol{h}_u}{\boldsymbol{h}_s}$ where the first p elements represent the elements of $\boldsymbol{h}_u$ while the remaining p represents the $\boldsymbol{h}_s$. Similarly for the k ARF subjects we have the vector of parameters being $\boldsymbol{\nu} = \binom{\boldsymbol{\nu}_u}{\boldsymbol{\nu}_s}$ and is associated with random variables $\boldsymbol{a} = \binom{\boldsymbol{a}_u}{\boldsymbol{a}_s}$ and $\boldsymbol{\nu}$ is of $2p \times 1$ dimension.

Therefore, the joint probability density function is given as

$$f(\boldsymbol{h}, \boldsymbol{a}) = (2\pi)^{-p}|\boldsymbol{\Sigma}|^{-1}\exp\Big\{ -\frac{1}{2}\big[(\boldsymbol{h} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{h} - \boldsymbol{\mu}) + (\boldsymbol{a} - \boldsymbol{\nu})'\Sigma^{-1}(\boldsymbol{a} - \boldsymbol{\nu})\big]\Big\}. \tag{4.8}$$

A reduced -2 Log of the likelihood function (dropping the terms that do not include the parameters) in terms of sufficient statistics is given by

$$-2\text{LogL}(\boldsymbol{\mu}, \boldsymbol{\nu}) = B + m(\bar{\boldsymbol{h}} - \boldsymbol{\mu})'\Sigma^{-1}(\bar{\boldsymbol{h}} - \boldsymbol{\mu}) + k(\bar{\boldsymbol{a}} - \boldsymbol{\nu})'\Sigma^{-1}(\bar{\boldsymbol{a}} - \boldsymbol{\nu}). \tag{4.9}$$

where B is a constant that does not contain the parameters under consideration and will vanish during the optimization.

Maximum Likelihood Estimates (MLEs)

To get the MLEs under H_o, we consider the parameter space given by $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\nu} : -\infty < -\infty < \boldsymbol{\mu}, \boldsymbol{\nu} < \infty\}$ and then we optimize the constrained Log-likelihood function using the Lagrangian $S(\boldsymbol{\Theta}, \boldsymbol{\lambda}) = -2\text{LogL}(\boldsymbol{\mu}, \boldsymbol{\nu}) + \boldsymbol{\lambda}'(\boldsymbol{\mu}_u - \boldsymbol{\mu}_s - \boldsymbol{\nu}_u + \boldsymbol{\nu}_s)$. The constraint on $2p\times 1$ can be conveniently expressed in matrix form as $A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0$ where $A = (\boldsymbol{I}, -\boldsymbol{I})$ and $\boldsymbol{I}$ is a $p \times p$ identity matrix. The constraint to be added to $-2\text{LogL}(\boldsymbol{\mu}, \boldsymbol{\nu})$ is of the form $2(\boldsymbol{\mu}-\boldsymbol{\nu})'A'\boldsymbol{\lambda} = 2[\boldsymbol{\lambda}'A(\boldsymbol{\mu}-\boldsymbol{\nu})]'$. We then find the MLEs as follows

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\mu}} = -2m\Sigma^{-1}(\bar{\boldsymbol{h}} - \boldsymbol{\mu}) + 2A'\boldsymbol{\lambda} \tag{4.10}$$

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\nu}} = -2k\Sigma^{-1}(\bar{\boldsymbol{a}} - \boldsymbol{\nu}) - 2A'\boldsymbol{\lambda} \tag{4.11}$$

$$\frac{\partial S(\boldsymbol{\Theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 2A(\boldsymbol{\mu} - \boldsymbol{\nu}). \tag{4.12}$$

Equating 4.10, 4.11 and 4.12 to zero and simplify get

$$\Sigma^{-1}(\bar{\boldsymbol{h}} - \boldsymbol{\mu}) - \frac{1}{m}A'\boldsymbol{\lambda} = 0 \tag{4.13}$$

$$\Sigma^{-1}(\bar{\boldsymbol{a}} - \boldsymbol{\nu}) + \frac{1}{k}A'\boldsymbol{\lambda} = 0 \tag{4.14}$$

$$A(\boldsymbol{\mu} - \boldsymbol{\nu}) = 0 \tag{4.15}$$

Now subtracting equation 4.13 from 4.14 and with some algebraic manipulations results in

$$
\begin{aligned}
\Sigma^{-1}(\bar{\boldsymbol{a}} - \boldsymbol{\nu} - \bar{\boldsymbol{h}} + \boldsymbol{\mu}) + \Big(\frac{1}{m} + \frac{1}{k}\Big)A'\boldsymbol{\lambda} &= 0 \\
(\bar{\boldsymbol{a}} - \boldsymbol{\nu} - \bar{\boldsymbol{h}} + \boldsymbol{\mu}) &= -\Big(\frac{1}{m} + \frac{1}{k}\Big)\Sigma A'\boldsymbol{\lambda} \\
A(\boldsymbol{\mu} - \boldsymbol{\nu}) + A(\bar{\boldsymbol{a}} - \bar{\boldsymbol{h}}) &= -\Big(\frac{m+k}{mk}\Big)A\Sigma A'\boldsymbol{\lambda} \\
A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) &= \Big(\frac{m+k}{mk}\Big)A\Sigma A'\boldsymbol{\lambda} \\
\boldsymbol{\lambda} &= \Big(\frac{mk}{m+k}\Big)(A\Sigma A')^{-1}A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \\
\boldsymbol{\lambda} &= \Big(\frac{mk}{m+k}\Big)(A\Sigma A')^{-1}\Delta \qquad (4.16)
\end{aligned}
$$

where $\Delta = A\bar{\boldsymbol{h}} - A\bar{\boldsymbol{a}}$. From equations 4.13 and 4.14, we get

$$\hat{\boldsymbol{\mu}}_0 = \bar{\boldsymbol{h}} - \frac{1}{m}\Sigma A'\boldsymbol{\lambda} \qquad (4.17)$$

$$\hat{\boldsymbol{\nu}}_0 = \bar{\boldsymbol{a}} + \frac{1}{k}\Sigma A'\boldsymbol{\lambda} \qquad (4.18)$$

The next set of MLEs under the alternative hypothesis H_a obtained by maximizing the unconstrained likelihood function are given by; $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{h}}$ and $\hat{\boldsymbol{\nu}} = \bar{\boldsymbol{a}}$.

Let $\boldsymbol{\theta}$ be the parameter vector for the likelihood function $L(\boldsymbol{\theta})$ with observations from the paired samples of healthy and ARF subjects as previously explained. Consider the parameter space given by $\boldsymbol{\Theta}$; we wish to test the null hypothesis $H_o : \boldsymbol{\theta} \in \boldsymbol{\Theta}$ versus the alternative $H_a : \boldsymbol{\theta} \notin \boldsymbol{\Theta}$. Recall that -2log likelihood ratio statistic is given by 4.7.

Substituting the MLEs under H_0 4.17 and 4.18 into the log likelihood function 4.9 we get

$$
\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} \\
&= B + m\Big(\frac{1}{m}\Sigma A'\boldsymbol{\lambda}\Big)'\Sigma^{-1}\Big(\frac{1}{m}\Sigma A'\boldsymbol{\lambda}\Big) + k\Big(\frac{1}{k}\Sigma A'\boldsymbol{\lambda}\Big)'\Sigma^{-1}\Big(\frac{1}{k}\Sigma A'\boldsymbol{\lambda}\Big) \\
&= B + \frac{1}{m}\Big(\boldsymbol{\lambda}'A\Sigma\Big)\Sigma^{-1}\Big(\Sigma A'\boldsymbol{\lambda}\Big) + \frac{1}{k}\Big(\boldsymbol{\lambda}'A\Sigma\Big)\Sigma^{-1}\Big(\Sigma A'\boldsymbol{\lambda}\Big) \\
&= B + \frac{1}{m}\Big(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\Big) + \frac{1}{k}\Big(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\Big) \\
&= B + \frac{(k+m)}{mk}\Big(\boldsymbol{\lambda}'A\Sigma A'\boldsymbol{\lambda}\Big) \qquad (4.19)
\end{aligned}
$$

Substituting for the values of $\boldsymbol{\lambda}$ from 4.16 then equation 4.19 can be written as

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_o}\{-2\text{Log L}(\boldsymbol{\theta})\} = \\
&B + \frac{(k+m)}{mk}\Big(\frac{mk}{(m+k)}(A\Sigma A')^{-1}\Delta\Big)'(A\Sigma A')^{-1}\Big(\frac{mk}{(m+k)}(A\Sigma A')^{-1}\Delta\Big) \\
&= B + \frac{mk}{(m+k)}\Delta'(A\Sigma A')^{-1}\Delta.
\end{aligned} \tag{4.20}$$

Under the unconstrained hypothesis $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{-2\text{Log L}(\boldsymbol{\theta})\} = B$ and therefore from 4.20

$$-2\text{Log}\Lambda = \frac{mk}{(m+k)}\Delta'(A\Sigma A')^{-1}\Delta \tag{4.21}$$

The distribution of $\Delta = A\bar{\boldsymbol{h}} - A\bar{\boldsymbol{a}}$ is $\Delta \sim N\left((A(\boldsymbol{\mu}-\boldsymbol{\nu}), \frac{(k+m)}{mk}(A\Sigma A')^{-1}\right)$. Now, if H_0 is true then $A(\boldsymbol{\mu}-\boldsymbol{\nu}) = 0$ so that $\Delta \sim N\left(0, \frac{(k+m)}{mk}(A\Sigma A')^{-1}\right)$. It is well known that given that $X \sim N_p(0, V)$ then $V^{-\frac{1}{2}} \sim N(0, I)$ implying that $(V^{-\frac{1}{2}}X)^T(V^{-\frac{1}{2}}X) \sim \chi^2_{(p)}$ and so $X^TV^{-1}X \sim \chi^2_{(p)}$, thus

$$-2\text{Log}\Lambda = \frac{mk}{(m+k)}\Delta'(A\Sigma A')^{-1}\Delta \sim \chi^2_{(p)}. \qquad \blacksquare$$

Case 2: Assuming the covariance matrix Σ is unknown

To estimate the covariance matrix, the -2log likelihood is re-written as follows

$$\begin{aligned}
l = {} & mp\log(2\pi) + m\log|\Sigma| + \text{tr}\Sigma^{-1}\boldsymbol{S}_h + \text{tr}\Sigma^{-1}(\bar{\boldsymbol{h}}-\boldsymbol{\mu})(\bar{\boldsymbol{h}}-\boldsymbol{\mu})' \\
& + kp\log(2\pi) + k\log|\Sigma| + \text{tr}\Sigma^{-1}\boldsymbol{S}_a + \text{tr}\Sigma^{-1}(\bar{\boldsymbol{a}}-\boldsymbol{\nu})(\bar{\boldsymbol{a}}-\boldsymbol{\nu})'
\end{aligned} \tag{4.22}$$

where $S_h = \sum_{i=1}^m(\boldsymbol{h}_i - \bar{\boldsymbol{h}})(\boldsymbol{h}_i - \bar{\boldsymbol{h}})'$ and $S_a = \sum_{j=1}^k(\boldsymbol{a}_j - \bar{\boldsymbol{a}})(\boldsymbol{a}_j - \bar{\boldsymbol{a}})'$. The partial derivative with respect to Σ^{-1} is obtained as

$$\begin{aligned}
\frac{\partial l}{\partial\Sigma^{-1}} &= -m\left((\boldsymbol{\Sigma}^{-1})^{-1}\right)' - k\left((\boldsymbol{\Sigma}^{-1})^{-1}\right)' + S_h + S_a + \left[m(\bar{\boldsymbol{h}}-\boldsymbol{\mu})(\bar{\boldsymbol{h}}-\boldsymbol{\mu})'\right]' \\
&\quad + \left[k(\bar{\boldsymbol{a}}-\boldsymbol{\nu})(\bar{\boldsymbol{a}}-\boldsymbol{\nu})'\right]' \\
&= -(m+k)\Sigma + S_h + S_a + m(\bar{\boldsymbol{h}}-\boldsymbol{\mu})(\bar{\boldsymbol{h}}-\boldsymbol{\mu})' + k(\bar{\boldsymbol{a}}-\boldsymbol{\nu})(\bar{\boldsymbol{a}}-\boldsymbol{\nu})'
\end{aligned} \tag{4.23}$$

The estimator for the variance-covariance matrix is given by

$$\hat{\Sigma} = \frac{1}{(m+k)}\left[S_h + S_a + m(\bar{\boldsymbol{h}} - \hat{\boldsymbol{\mu}})(\bar{\boldsymbol{h}} - \hat{\boldsymbol{\mu}})' + k(\bar{\boldsymbol{a}} - \hat{\boldsymbol{\nu}})(\bar{\boldsymbol{a}} - \hat{\boldsymbol{\nu}})'\right]. \tag{4.24}$$

Now, substituting the plug-in estimator for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ which are $\bar{\boldsymbol{h}}$ and $\bar{\boldsymbol{a}}$ respectively, we get the plug-in estimator for the covariance matrix as

$$\hat{\Sigma} = \frac{1}{(m+k)}\left[S_h + S_a\right]. \tag{4.25}$$

The estimator $\hat{\Sigma}$ is then plugged-in into the test statistic given by 4.21 which has $\chi^2_{(p)}$ distribution to get

$$-2\text{Log}\Lambda = \frac{mk}{(m+k)}\Delta'(A\hat{\Sigma}A')^{-1}\Delta. \tag{4.26}$$

Proposition

Denote 4.21 by $\Lambda_1 = \frac{mk}{(m+k)}\Delta'(A\Sigma A')^{-1}\Delta$ and 4.26 by $\Lambda_2 = \frac{mk}{(m+k)}\Delta'(A\hat{\Sigma}A')^{-1}\Delta$ and noting that $\hat{\Sigma}$ is a consistent estimator of Σ, since $\Lambda_1 \overset{d}{\sim} \chi^2_{(p)}$ then $\Lambda_2 \overset{a}{\sim} \chi^2_{(p)}$, where $\overset{d}{\sim}$ means exactly distributed while $\overset{a}{\sim}$ stands for asymptotically distributed.

Proof

Since $\hat{\Sigma} \xrightarrow{p} \Sigma$ as $n \to \infty$ where $n = m+k$ and the fact that $(A\Sigma A')$ is positive definite, we had shown in case 1 that $A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \overset{d}{\sim} N\left(0, \frac{m+k}{mk}A\Sigma A'\right)$ under H_0 then it follows that in a similar manner $A(\bar{\boldsymbol{h}} - \bar{\boldsymbol{a}}) \overset{a}{\sim} N\left(0, \frac{m+k}{mk}A\hat{\Sigma}A'\right)$ under H_0. Consequently the test statistic $\frac{mk}{(m+k)}\Delta'(A\hat{\Sigma}A')^{-1}\Delta \overset{a}{\sim} \chi^2_{(p)}$. ■

4.6 Simulation study for the multivariate problem

In this section, we carry out a simulation experiment and use it with the likelihood ratio statistics that has been developed in this thesis. The summary of the simulation set-up is given below.

- The mean vector for the "healthy unstimulated" is obtained by simulating p uniform random variables in the range of $(0, 0.5)$ to be the vector $\boldsymbol{\mu}_u$.

- Similarly, we generate p uniform random variables in the interval $(0.6, 0.75)$ to create $\boldsymbol{\mu}_s$ which is the "healthy stimulated".
- For the "ARF unstimulated", the values for simulating $\boldsymbol{\nu}_u$ were the range $(0, 0.55)$ to generate uniform random variables of dimension p.
- The $\boldsymbol{\nu}_s$ are obtained by generating a p uniform random variables of the interval $(0.001, 0.2)$ to obtain the mean vector for the "ARF stimulated".

The covariance matrices are generated using the `R` package `clusterGeneration` in order to get a $p \times p$ positive definite matrix. The number of subjects for the healthy group is arbitrarily set at 100 while the ARF group is set at 105.

Results for the analysis of the multivariate data

Simulation experiment I

The data was simulated for four different values of $p = \{2, 5, 8, 15\}$ while the sample sizes were fixed at $m = 100$ and $k = 105$. A test statistic and corresponding p-value calculated when Σ is assumed to be unknown and when it is known. A resampling distribution was then obtained from which an approximate p-value is then computed. The results are shown in Tables 4.4 and 4.3 in addition to Figures 4.3 and 4.6.

Table 4.3: Calculated test statistics when Σ is known

	p=2	p=5	p=8	p=15
Log RT	8.46	18.84	22.43	56.09
calculated p-value[e]	0.016	0.002	0.004	0.000
p-value from resampling	0.01	0.000	0.006	0.000

[e] p-values calculated from the exact $\chi^2_{(p)}$ distribution

Table 4.4: Calculated test statistics when Σ is unknown

	p=2	p=5	p=8	p=15
Log RT	10.65	21.32	25.81	84.07
calculated p-value[a]	0.005	0.001	0.001	0.000
p-value from resampling	0.005	0.001	0.002	0.000

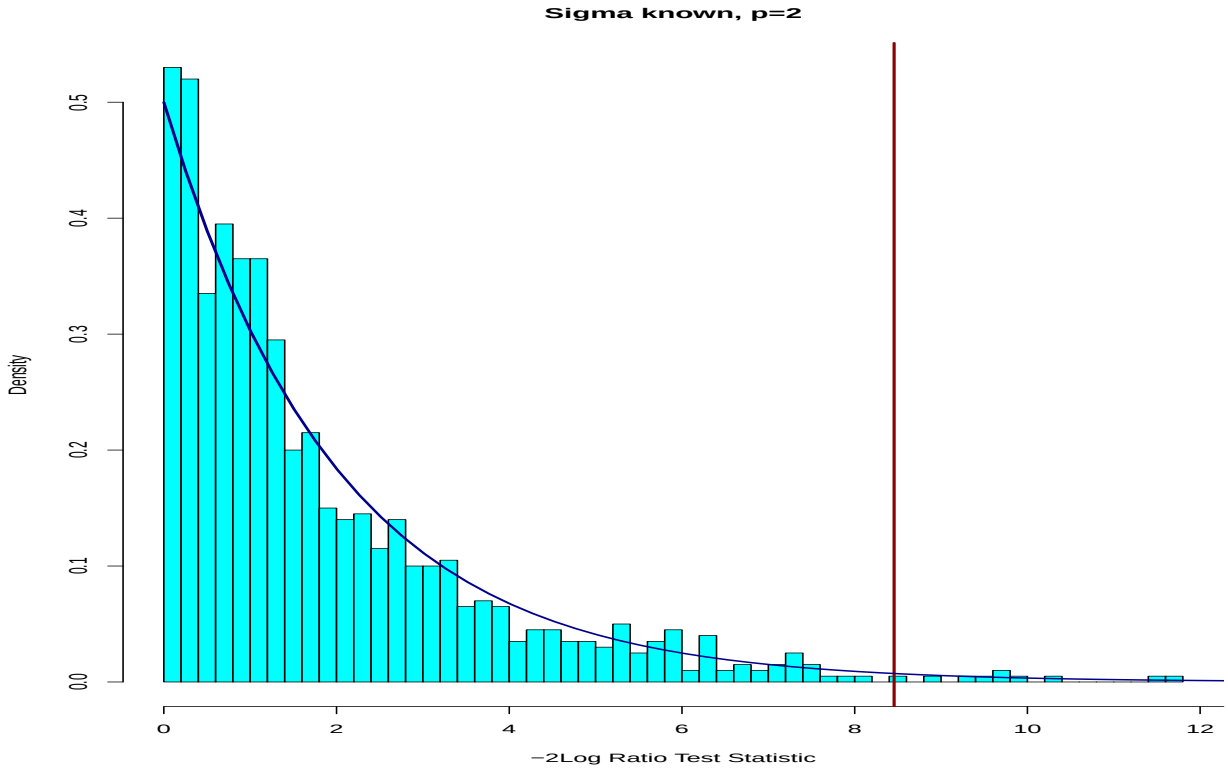
[a] p-values calculated from the asymptotic $\chi^2_{(p)}$ distribution

The results reveal that both the calculated p-value and the one obtained from resampling lead to the same conclusions regarding the hypothesis testing. In this

case, for all the cases, the difference in the means was statistically significant at 5% level. Next we look at the sampling distribution of the test statistics for different numbers of p.

The Figures 4.3 and 4.4 show the histograms are based on the test statistics calculated from resampling the simulated data and the curves are the chi-squared densities for the corresponding degrees of freedom p. The plots show that the distributions for the -2 log likelihood test statistic follow a chi-square distribution and are also positive skewed. However, as the number of p increases, the distributions look like normal distribution and the skewness is less when the degree of freedom is higher. The normal looking distribution are still a chi-squared, for they approach $N(p, 2p)$ as the degree of freedom gets large. The red vertical lines shows the position of the computed stastistic for the unresampled data.

Figures 4.5 and 4.6 show the histograms are obtained from the simulated data while the curves are chi-squared densities with p degrees of freedom. This set of histograms exhibit the same properties as the ones discussed in Figures 4.3 and 4.4.



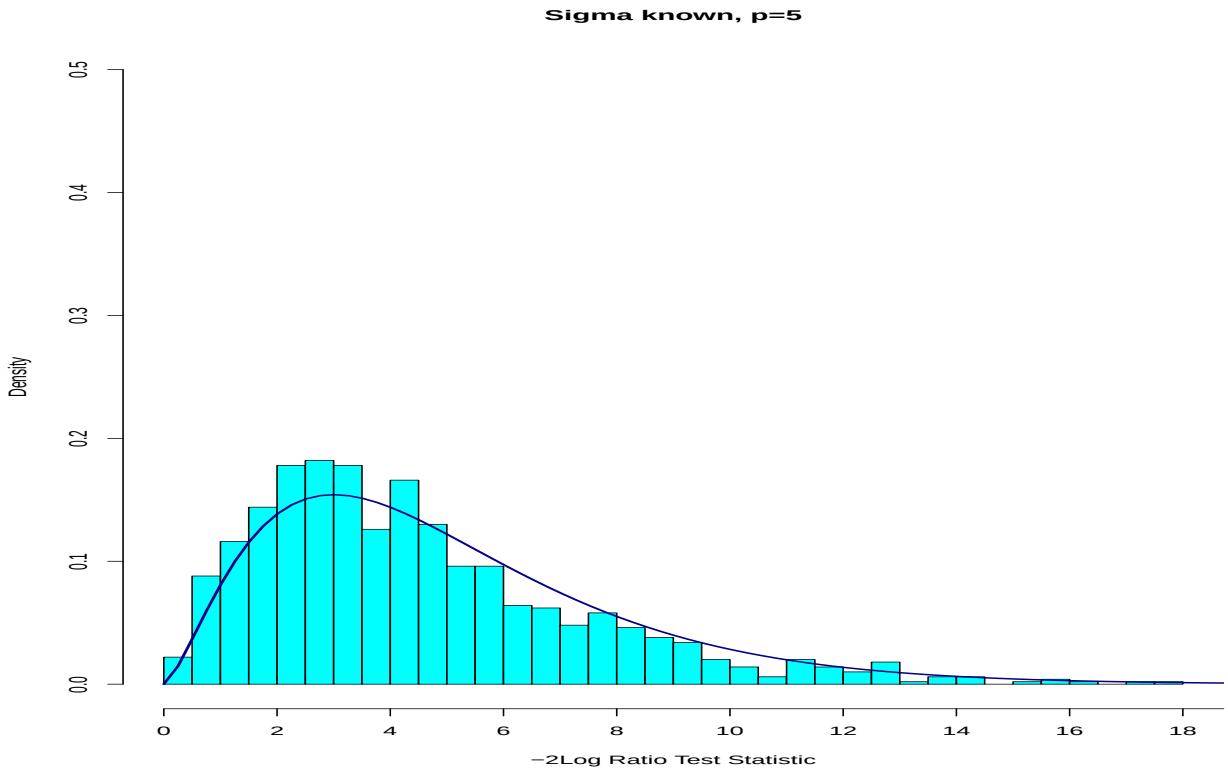


Figure 4.3: Histograms for the simulated data for p=2 and p=5 when Σ is known.

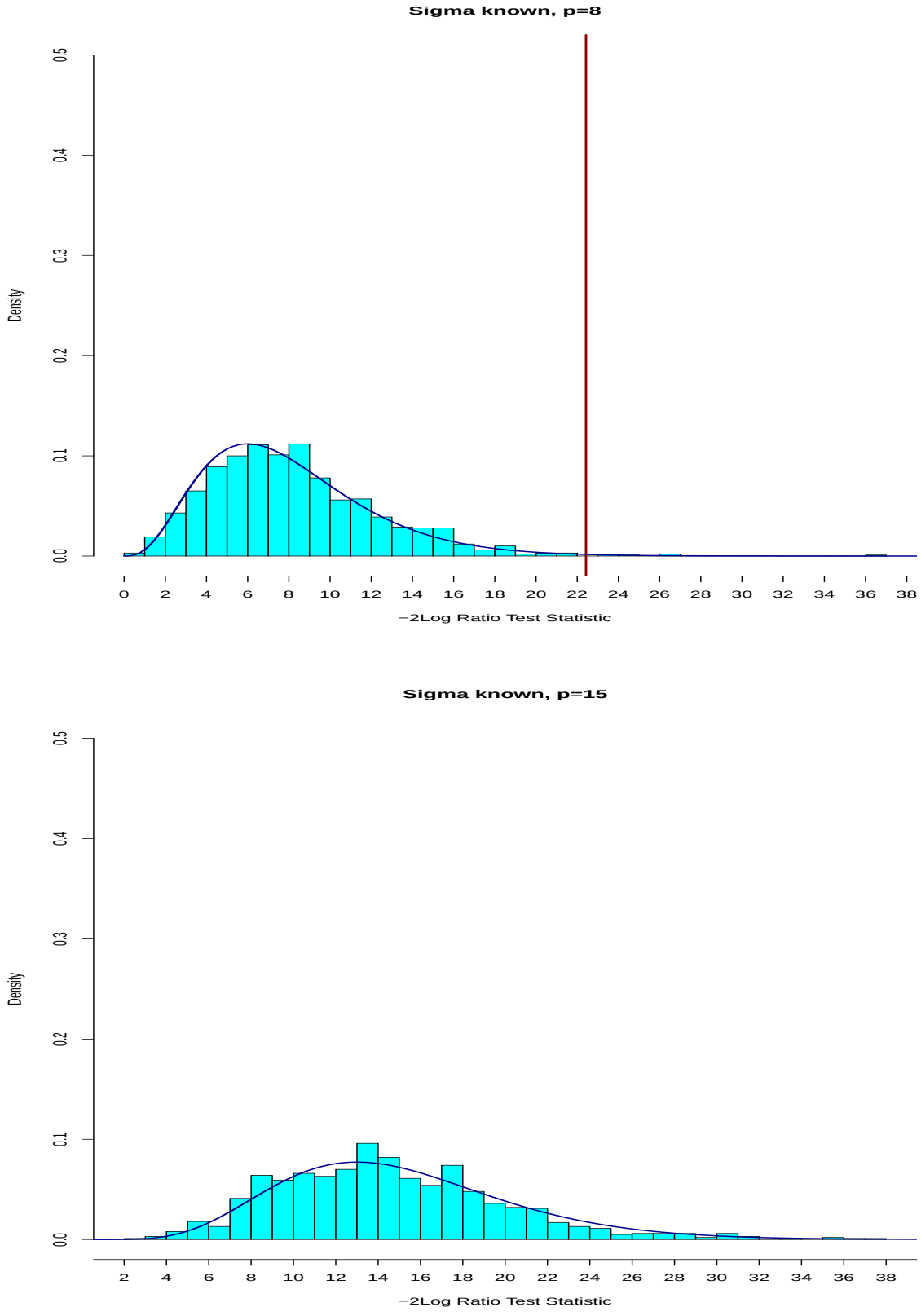


Figure 4.4: Histograms for the simulated data for p=8 and p=15 when Σ is known.

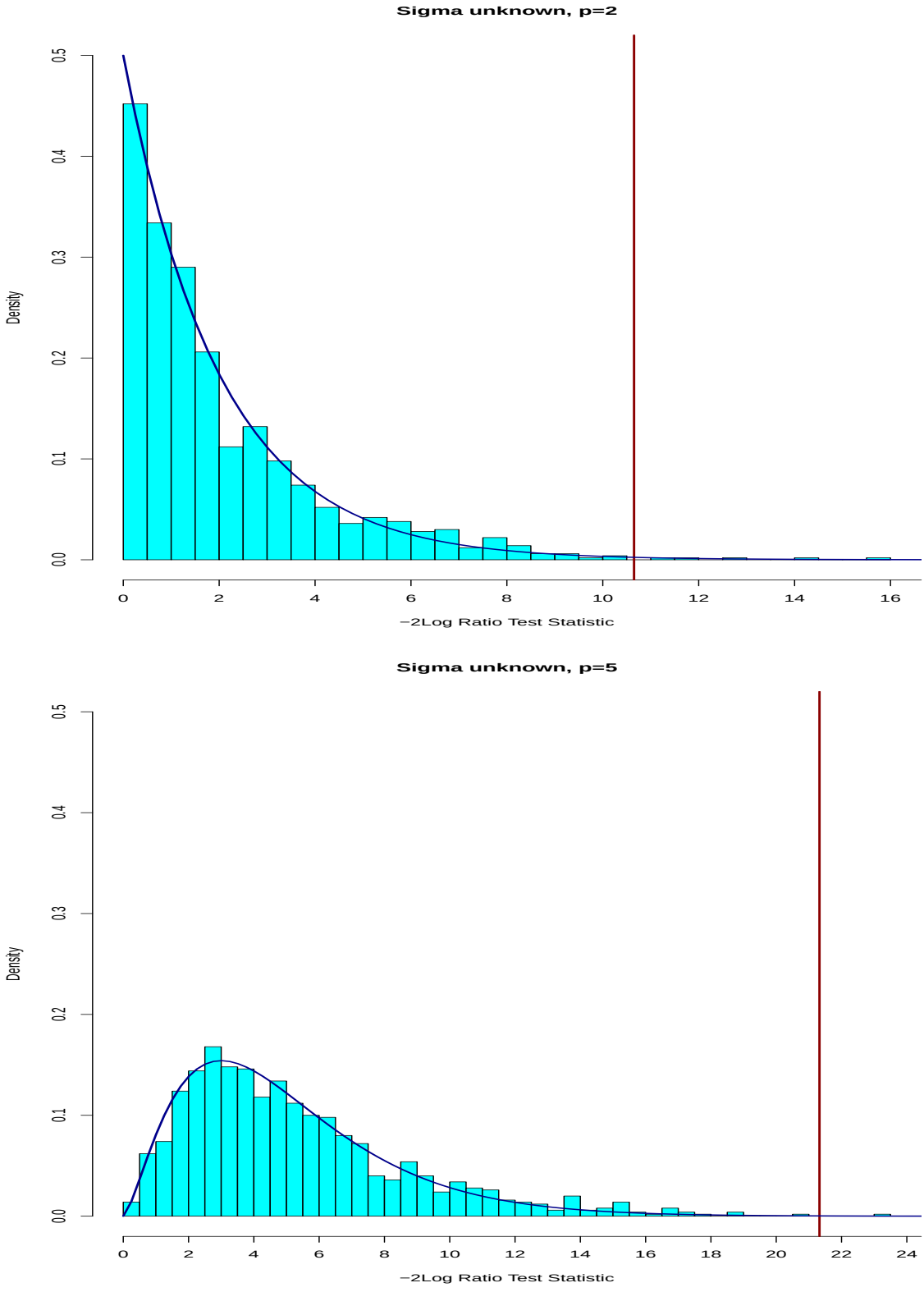


Figure 4.5: Histograms for the simulated data for p=2 and p=5 when Σ is unknown.

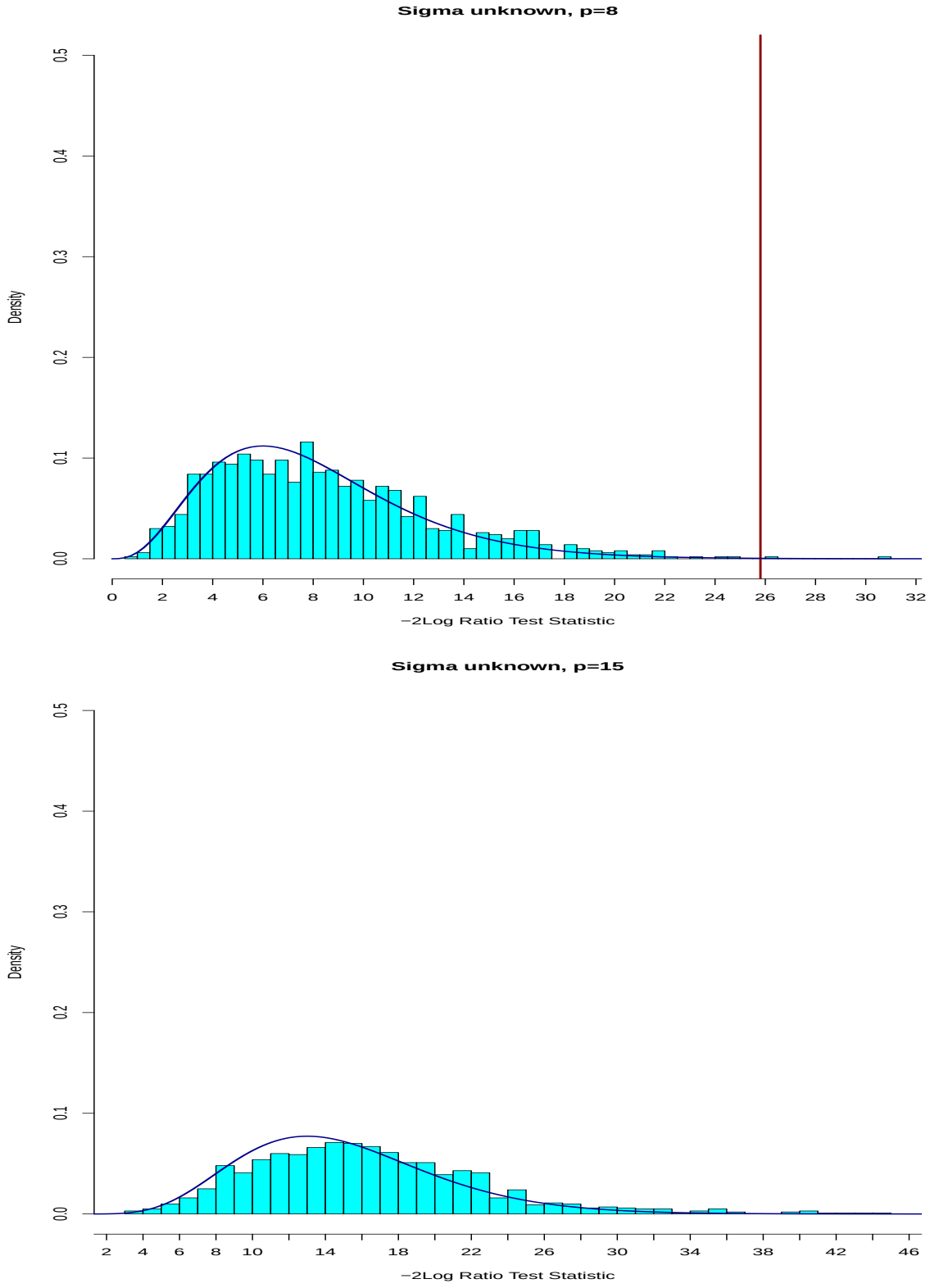


Figure 4.6: Histograms for the simulated data for p=8 and p=15 when Σ is unknown.

.

Simulation experiment II

In this experiment, we fix $p = 20$ but vary the sample sizes m and k during the simulation. On the same data sets, we apply our test statistics when Σ is assumed known and then when it is assumed to be unknown. The test statistic is computed and then a resampling procedure is carried out to get a sampling distribution for both cases.

Table 4.5: Calculated test statistics and p-values when Σ is known for different values of m and k.

	m=150, k=180	m=200, k=190	m=k=250	m=300, k=350
Log RT	60.98	56.62	91.42	81.32
calculated p-value[e]	0.001	0.000	0.000	0.000
p-value from resampling	0.006	0.000	0.000	0.000

[e] p-values calculated from the exact $\chi^2_{(p)}$ distribution

Table 4.6: Calculated test statistics and p-values when Σ is unknown for different values of m and k.

	m=150, k=180	m=200, k=190	m=k=250	m=300, k=350
Log RT	44.61	69.45	105.20	103.76
calculated p-value[a]	0.001	0.000	0.000	0.000
p-value from resampling	0.006	0.000	0.000	0.000

[a] p-values calculated from the asymptotic $\chi^2_{(p)}$ distribution

Both Tables 4.5 and 4.6 show that the difference in the mean differences is significant at 5%. It is worth noting that for both the cases when Σ is assumed known and unknown, the respective test statistics lead to the same conclusion of rejecting a null hypothesis in this case.

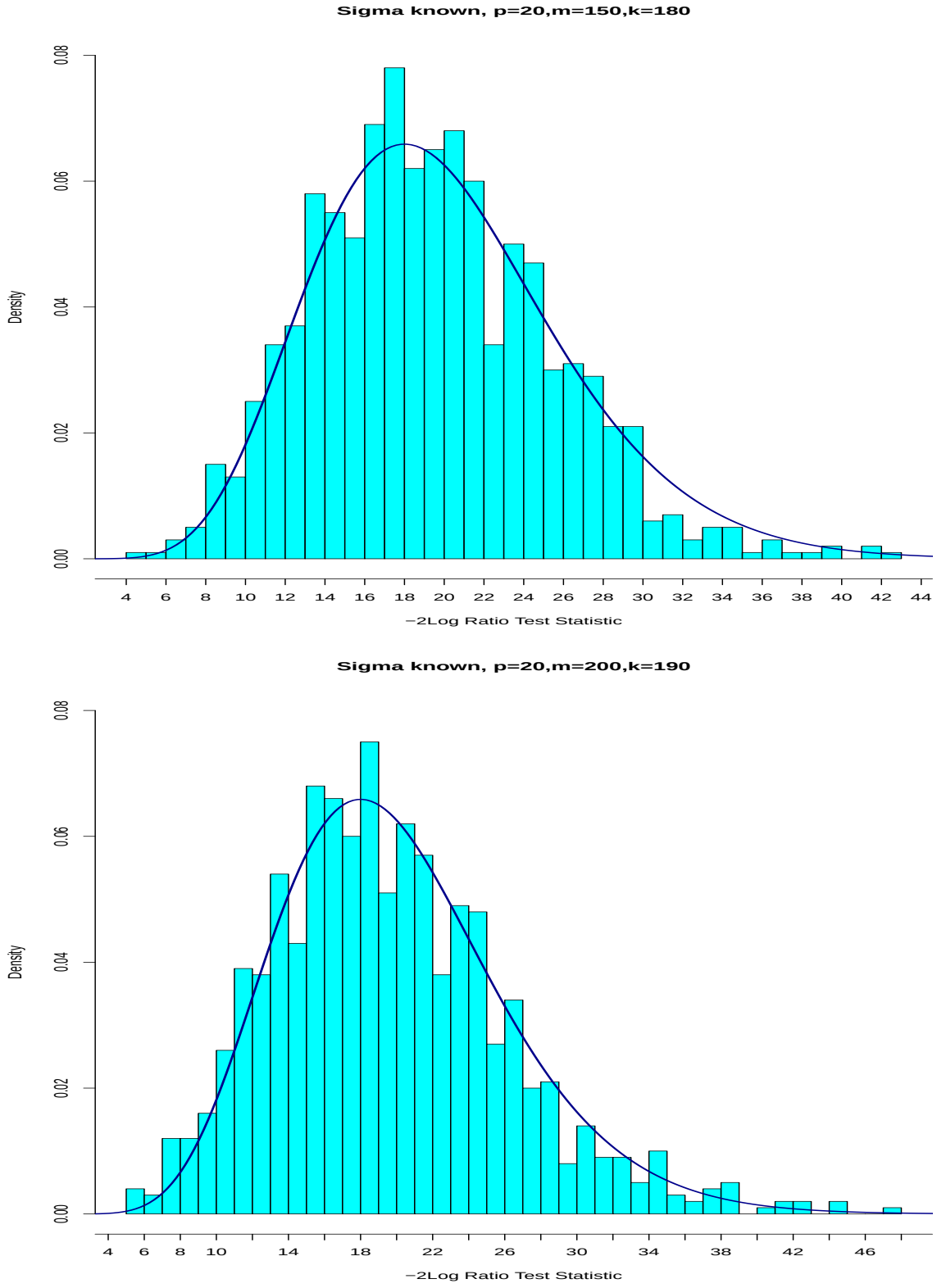


Figure 4.7: Histograms for the simulated data for p=20, m=200 & k=190 when Σ is known.

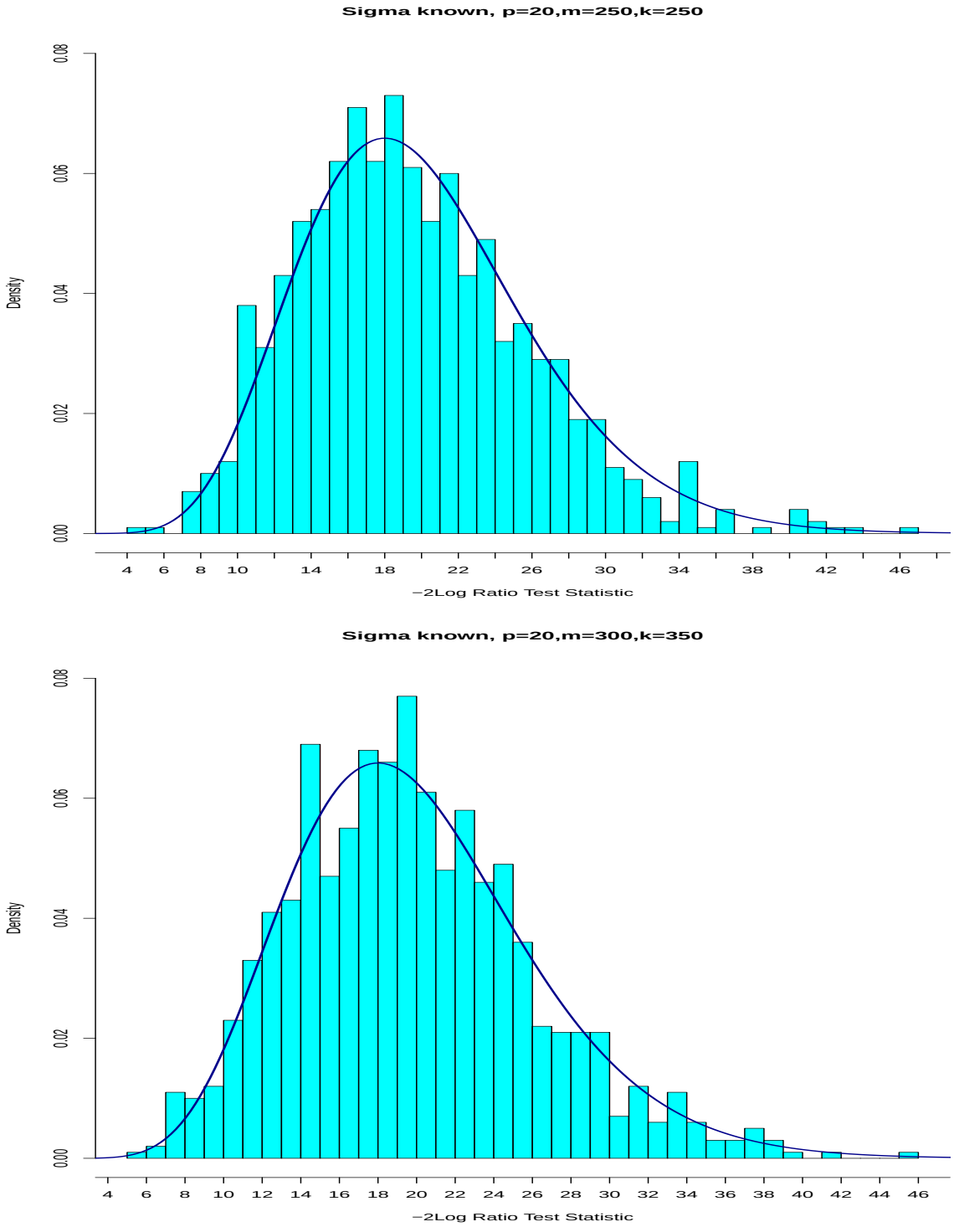


Figure 4.8: Histograms for the simulated data for p=20, m=300 & k=350 when Σ is known.

The histograms in Figures 4.7 and 4.8 exhibit the same properties as the ones discussed in Figures 4.5 and 4.6.

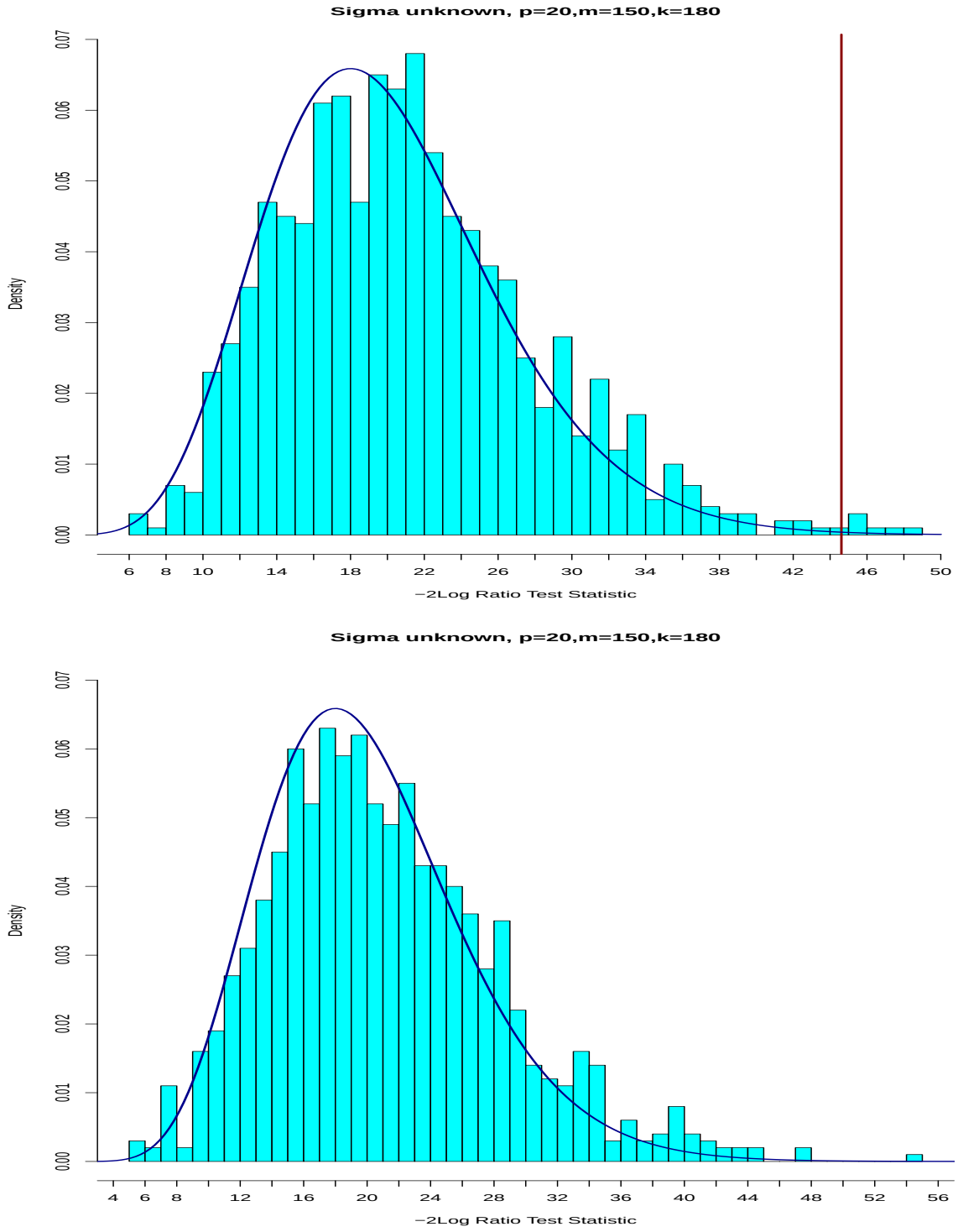


Figure 4.9: Histograms for the test statistic computed from resampling the simulated data for p=20 and Σ is unknown. This set of histograms exhibit the same properties as the ones already discussed in 4.5 and 4.6.

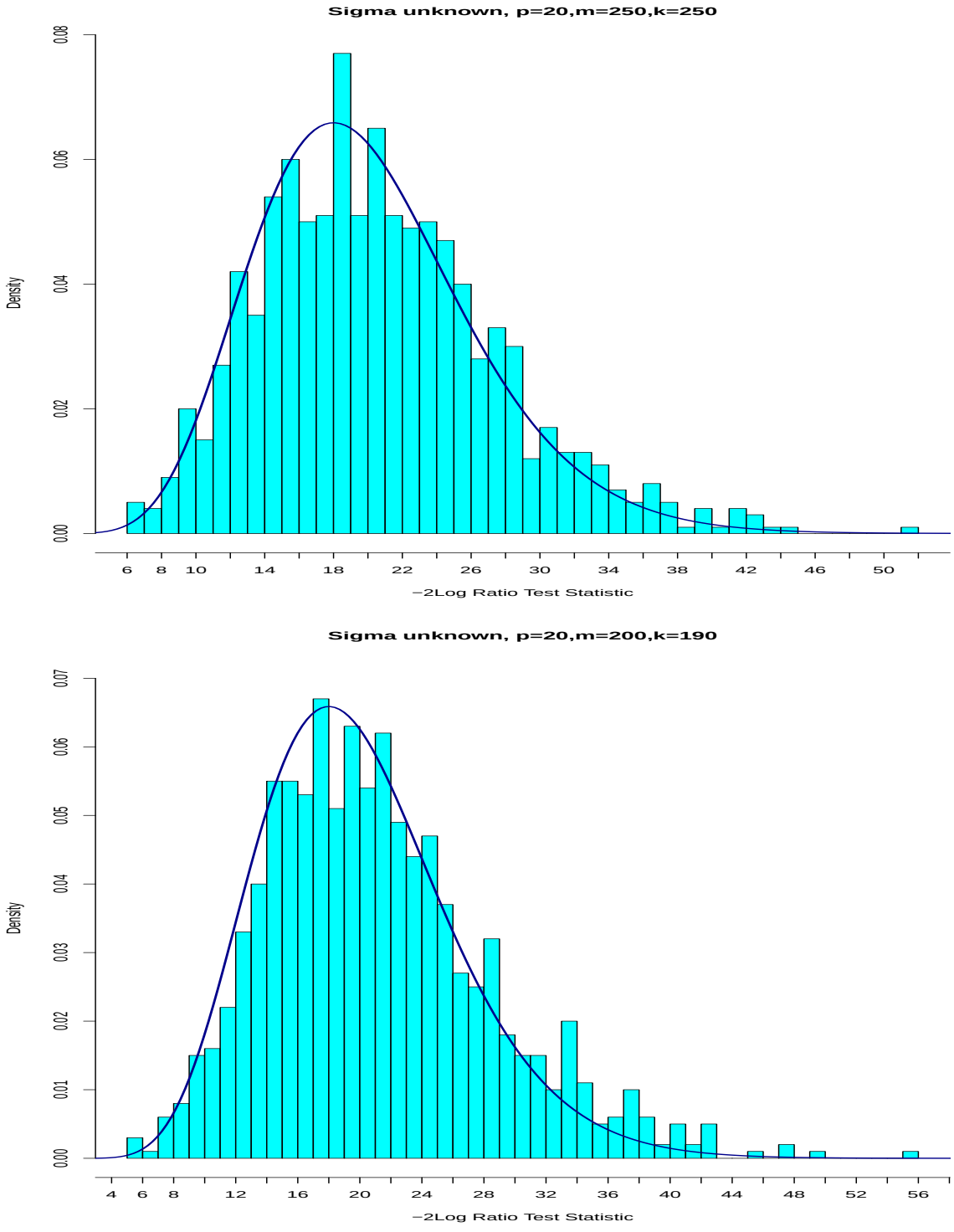


Figure 4.10: Histograms for the test statistic computed from resampling the simulated data for p=20 and Σ is unknown. This set of histograms exhibit the same properties as the ones discussed in the previous figures.

4.7 Some special cases

In this subsection, we present special cases for comparing the healthy and ARF groups while integrating the graph network information when $p = 1$ with Σ being known and when it is unknown.

4.7.1 One gene problem

Consider the paired measurements framework described in Subsection 4.5.1 and let $p = 1$. In this case, the healthy and ARF subjects come from a bivariate normal distributions $\binom{h_u}{h_s} \sim \mathcal{N}\left[\binom{\mu_u}{\mu_s}, \Sigma\right]$ and $\binom{a_u}{a_s} \sim \mathcal{N}\left[\binom{\nu_u}{\nu_s}, \Sigma\right]$ respectively. The variance-covariance matrix given by

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{us} \\ \sigma_{su} & \sigma_s^2 \end{bmatrix}, \qquad \sigma_{us} = \rho\sigma_u\sigma_s.$$

Assuming that $\sigma_u = \sigma_s = \sigma$ and is known, it can be shown that the parameter estimates under the null hypothesis H_0 are obtained by optimizing the constrained likelihood function with respect to $\mu_u, \mu_s, \nu_u,$ and ν_u to get

$$\begin{aligned} \hat{\mu}_u &= \bar{h}_u + \frac{m}{2(k+m)}\Delta \\ \hat{\mu}_s &= \bar{h}_s - \frac{m}{2(k+m)}\Delta \\ \hat{\nu}_u &= \bar{a}_u - \frac{k}{2(k+m)}\Delta \\ \hat{\nu}_s &= \bar{a}_s + \frac{k}{2(k+m)}\Delta. \end{aligned} \tag{4.27}$$

where $\Delta = \Delta h - \Delta a$, $\Delta h = \bar{h}_u - \bar{h}_s$ and $\Delta a = \bar{a}_u - \bar{a}_s$. The MLEs under H_a are obtained by optimizing the unrestricted log likelihood function and are given by

$$\hat{\mu}_u = \bar{h}_u, \hat{\mu}_s = \bar{h}_s, \hat{\nu}_u = \bar{a}_u, \hat{\nu}_s = \bar{a}_s. \tag{4.28}$$

The -2 Log likelihood ratio test is given as

$$-2\text{Log}\lambda = \frac{mk}{8(1-\rho)(k+m)\sigma^2}\Delta^2, \tag{4.29}$$

The variance $\text{var}(\Delta) = \frac{2(m+k)\sigma^2(1-\rho)}{mk}$ and therefore

$$\frac{mk}{8(1-\rho)(k+m)\sigma^2}\Delta^2 \sim \chi^2_{(1)}. \qquad \blacksquare$$

On the other hand, when we assume that $\sigma_u = \sigma_s = \sigma$ and unknown, we can get the MLE expression for σ^2 by optimized the constrained likelihood function with respect to σ^2 to get

$$\begin{aligned}\hat{\sigma}^2 = \frac{1}{(m+k)(1-\rho^2)}\Bigg\{ &\Big[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 + \\ &\sum_{i=1}^{k}(a_{s_i}-\bar{a}_s)^2 + \sum_{i=1}^{k}(a_{ui}-\bar{a}_u)^2\Big] \\ &+ \left[m(\bar{h}_u-\mu_u)^2 + m(\bar{h}_s-\mu_s)^2 + k(\bar{a}_s-\nu_s)^2 + k(\bar{a}_u-\nu_u)^2\right] \\ &- 2\rho\left[m(\bar{h}_u-\mu_u)(\bar{h}_s-\mu_s) + k(\bar{a}_u-\nu_u)(\bar{a}_s-\nu_s)\right]\Bigg\}.\end{aligned} \tag{4.30}$$

To get the expression for the estimate of ρ let, $B = \Big[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + m(\bar{h}_u-\mu_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 + m(\bar{h}_s-\mu_s)^2 + \sum_{i=1}^{k}(a_{ui}-\bar{a}_u)^2 + k(\bar{a}_u-\nu_u)^2 + \sum_{i=1}^{k}(a_{si}-\bar{a}_s)^2 + k(\bar{a}_s-\nu_s)^2\Big]$ and C=$\Big[m(\bar{h}_u-\mu_u)(\bar{h}_s-\mu_s) + k(\bar{a}_u-\nu_u)(\bar{a}_s-\nu_s)\Big]$ then the expression for ρ can be found as

$$(m+k)(\rho-\rho^3) - \frac{\rho}{\sigma^2}B + (1+\rho^2)C = 0. \tag{4.31}$$

The MLE expressions for the variance and correlation coefficient under H_0 are obtained by substituting 4.27 into 4.30 to get respectively as

$$\begin{aligned}\hat{\sigma}_0^2 = \frac{1}{(m+k)(1-\hat{\rho}_0^2)}\Bigg\{ &\Big[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 + \sum_{i=1}^{k}(a_{si}-\bar{a}_s)^2 \\ &+ \sum_{i=1}^{k}(a_{ui}-\bar{a}_u)^2\Big] + \frac{(1-\hat{\rho}_0)km}{2(k+m)}\Delta^2\Bigg\},\end{aligned} \tag{4.32}$$

and

$$(m+k)(\hat{\rho}_0 - \hat{\rho}_0^3) - \frac{\hat{\rho}_0}{\hat{\sigma}_0^2}B_0 + (1+\hat{\rho}_0^2)C_0 = 0. \tag{4.33}$$

where $B_0 = \Big[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2+\sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2+\sum_{i=1}^{k}(a_{u_i}-\bar{a}_u)^2+\sum_{i=1}^{k}(a_{si}-\bar{a}_s)^2+ \frac{km}{2(k+m)}\Delta^2\Big]$ and $C_0 = \frac{km}{4(k+m)}\Delta^2$ where $\Delta = \Delta h - \Delta a$, $\Delta h = \bar{h}_u - \bar{h}_s$ and $\Delta a = \bar{a}_u - \bar{a}_s$.

Substituting the MLEs 4.28 into 4.30 we get that the expressions for the variance and correlation coefficients are given respectively under H_a as

$$\begin{aligned}\hat{\sigma}_1^2 = \frac{1}{(m+k)(1-\rho^2)} \Bigg\{ &\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 \\ &+ \sum_{i=1}^{k}(a_{si}-\bar{a}_s)^2 + \sum_{i=1}^{k}(a_{ui}-\bar{a}_u)^2 \Bigg\},\end{aligned} \tag{4.34}$$

and

$$(m+k)(\hat{\rho}_1 - \hat{\rho}_1^3) - \frac{\hat{\rho}_1}{\hat{\sigma}_1^2} B_1 = 0. \tag{4.35}$$

where

$$B_1 = \left[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 + \sum_{i=1}^{k}(a_{u_i}-\bar{a}_u)^2 + \sum_{i=1}^{k}(a_{si}-\bar{a}_s)^2 \right].$$

The estimates $\hat{\sigma}_0, \hat{\rho}_0, \hat{\sigma}_1$ and $\hat{\rho}_1$ can be obtained using numerically. The log likelihood ratio function under H_0 is given by

$$\begin{aligned}&\mathrm{Log}L(\boldsymbol{\Theta}_0) = \\ &-\frac{(m+k)}{2}\left[2\log\hat{\sigma}_0^2 + \log(1-\hat{\rho}_0)\right] - \frac{1}{2(1-\hat{\rho}_0^2)\hat{\sigma}_0^2}\Bigg\{ \Big[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 \\ &\qquad + \sum_{i=1}^{k}(a_{si}-\bar{a}_s)^2 + \sum_{i=1}^{k}(a_{ui}-\bar{a}_u)^2 \Big] + \frac{(1+\hat{\rho}_0)km}{2(k+m)}\Delta^2 \Bigg\}.\end{aligned} \tag{4.36}$$

The unrestricted log likelihood function is given by

$$\begin{aligned}
&\mathrm{Log}L(\boldsymbol{\Theta}_1) = \\
&= -(m+k)\log(2\pi) - \frac{(m+k)}{2}\left[2\log\hat{\sigma}_1^2 + \log(1-\hat{\rho}_1)\right] \\
&- \frac{1}{2(1-\hat{\rho}_1^2)\hat{\sigma}_1^2}\Big[\sum_{i=1}^{m}(h_{ui}-\bar{h}_u)^2 + \sum_{i=1}^{m}(h_{si}-\bar{h}_s)^2 + \sum_{i=1}^{k}(a_{s_i}-\bar{a}_s)^2 + \sum_{i=1}^{k}(a_{u_i}-\bar{a}_u)^2\Big].
\end{aligned} \tag{4.37}$$

From 4.36 and 4.37 we get the -2 Log Likelihood ratio test by

$$-2\mathrm{Log}\lambda = -2\mathrm{Log}L(\boldsymbol{\Theta}_0) + 2\mathrm{Log}L(\boldsymbol{\Theta}_1) \tag{4.38}$$

The distribution of 4.38 is not explicitly known, we can use the resampling methods like the permutation test in order to find the its sampling distribution for making inferences.

4.7.2 Simulation study of the one gene problem (p=1)

Simulation Experiment III

In this simulation experiment, we set $p = 1$ and simulate the data with the following means and variance-covariance matrix

$$\Sigma = \begin{bmatrix} 6.6 & 0.06 \\ 0.06 & 6.6 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}.$$

The simulation was done for different values of m and k. When $\sigma_u = \sigma_s = \sigma$ and assumed known, we use the test statistic given in 4.29. The statistics and p-values shown in Table 4.7.

Table 4.7: Calculated test statistics and p-values when Σ is known for different values of m and k for the one gene problem.

	m=150, k=180	m=200, k=190	m=k=250	m=300, k=350
Log RT	7.52	6.58	6.84	9.34
calculated p-value[e]	0.006	0.010	0.009	0.002
p-value from resampling	0.006	0.000	0.000	0.000

[e] p-values calculated from the asymptotic $\chi^2_{(1)}$ distribution

When $\sigma_u = \sigma_s = \sigma$ and assumed unknown, we still use the data that was previously simulated but in this case we now estimate the correlation coefficient and the variance using numerical methods because ρ does not have a closed form solution in this case. The `R` function `nlminb` is used to implement the numerical estimation and then the $-2\text{Log}\lambda$ is computed for each of the 1000 permutations of the labels of the healthy and the ARF subjects. Since the distribution of the test statistic is not known, permutations are used to compute the sampling distribution of the statistic given in 4.38 and the results given in Table 4.8.

Table 4.8: Calculated test statistics and p-values when Σ is unknown for different values of m and k for the one gene problem.

	m=150, k=180	m=200, k=190	m=k=250	m=300, k=350
$-2\log$ RT	59.32	50.01	56.10	76.97
p-value from resampling	0.006	0.000	0.000	0.000

The results for when sigma is known and when unknown, both lead to the same conclusions of significant difference in the difference of the means for the two categories under consideration. The sampling distributions for various cases are presented in Figures 4.11, 4.12, 4.13 and 4.14 which exhibit the same properties as the ones discussed in Figures 4.3 and 4.4.

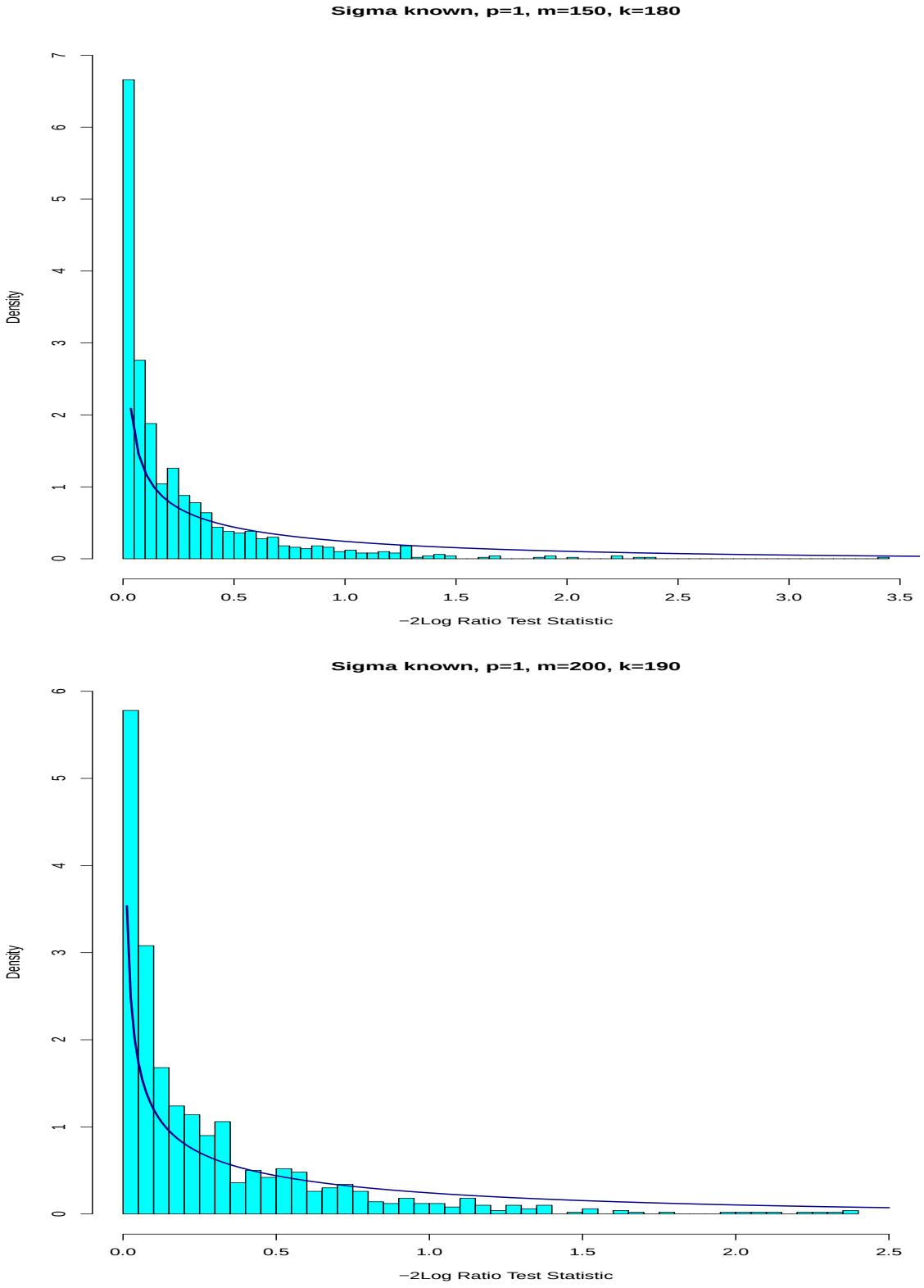


Figure 4.11: Histograms for the test statistics computed by resampling the simulated data for p=1, different values of m and k when σ is known.

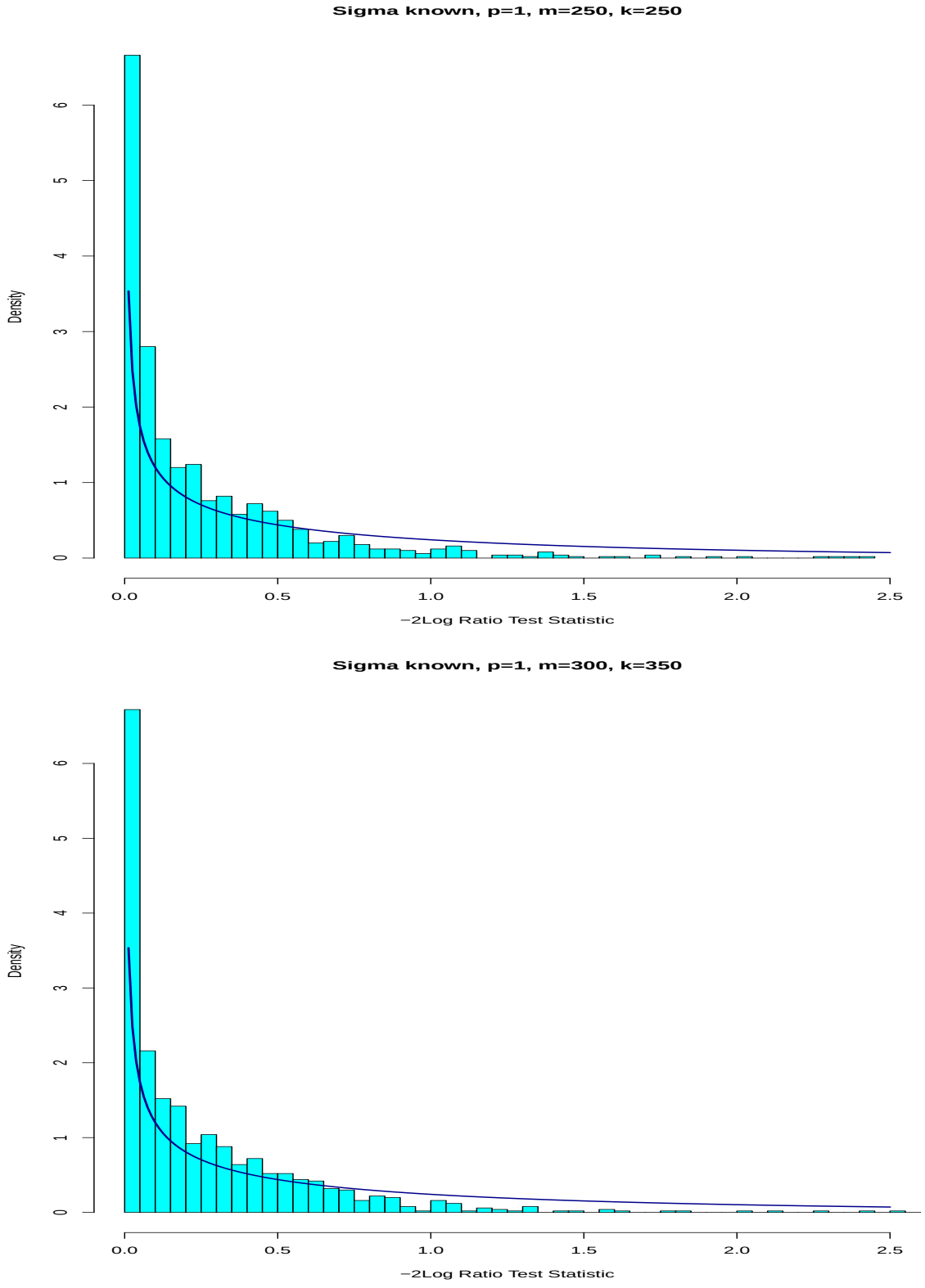


Figure 4.12: Histograms for the test statistics computed by resampling the simulated data for p=1, different values of m and k when σ is known.

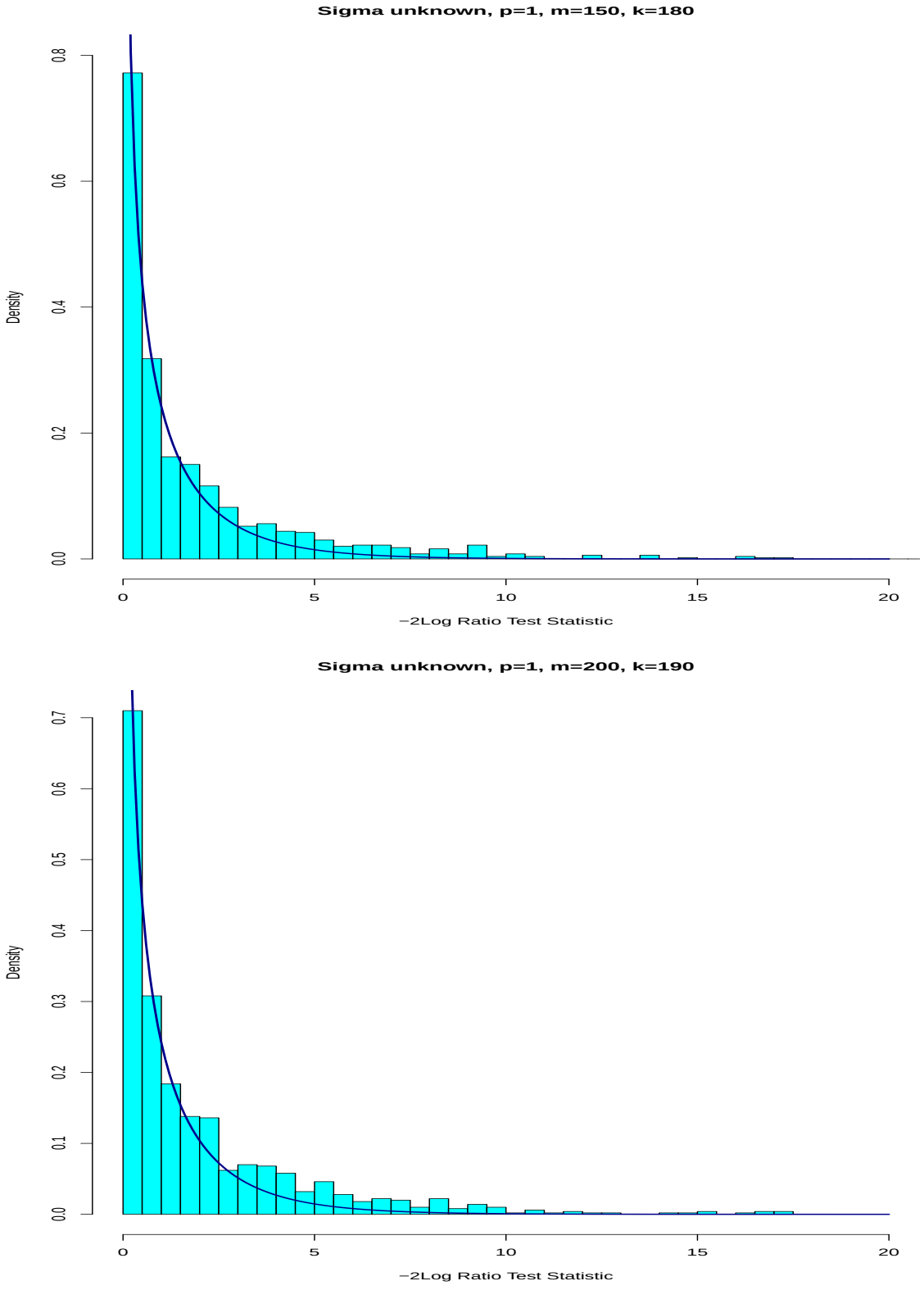


Figure 4.13: Histograms for the test statistics computed by resampling the simulated for p=1, different values of m and k when σ is unknown.

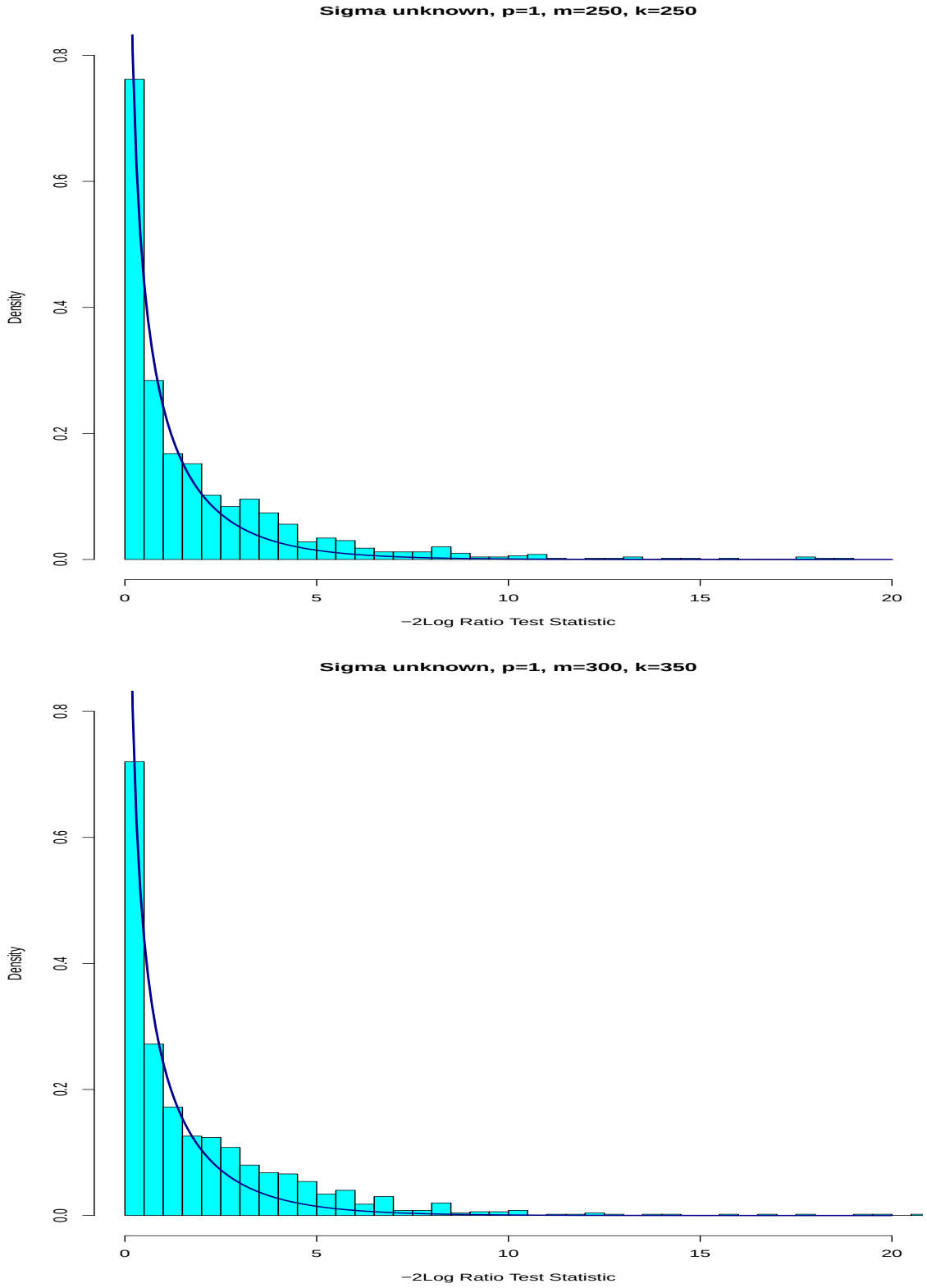


Figure 4.14: Histograms for the test statistics computed by resampling the simulated for p=1, different values of m and k when σ is unknown.

4.8 Summary for the testing for network changes

To integrate network information, the first step is to identify the appropriate prior network that contains the same genes as the experimental data. From the prior network, a list of edges is extracted. This list is used with the experimental data to build a graph network where the genes are represented by the nodes while the edges are the connectivity information obtained from the list of edges. The nodes are coloured to reflect up/down regulation or no change in the gene expression values. The colouring of the nodes is aided by use of the log fold change calculated from the experimental data. From the network, subnetworks that represent a certain functional group are chosen. The genes in the chosen subnetwork(s) are utilized to select the subset of the original experimental data used in testing for network changes.

We have used the known likelihood ratio theory to develop a statistic that can be used to formally test for network changes. The simulation study has shown that the developed likelihood ratio tests are capable of testing for the changes from a practical point of view. For a small number of variable p, the sampling distributions are chi-square. As the degree of freedom increases, the distributions tend to look like a normal distribution. That is due to the fact that the skewness decreases as the degrees of freedom increases.

Chapter 5

Summary, conclusions and future work

5.1 Summary and conclusions

In thesis, two major themes have been looked in to: the classification and statistical integration of the molecular data to test for network changes.

The classification problem has been addressed using two different approaches; first using the un-preprocessed data and secondly using the preprocessed data. The methodology of Bastien et al. (2005) has been combined with the logistic regression (PLSGLR-log) and also with the linear discriminant analysis (PLSGLRDA), then applied to the microarray data sets. The performance of these two extensions are then compared with with the classical methodologies like KNN, LDA, RPLS, PLSDA, SVM in addition to the KMA, an algorithm that was recently proposed by (Dalmau et al., 2015). For the un-preprocessed data, using the 10-fold cross validation, the KMA emerges as clear winner, the new extensions PLSGLR-log and PLSGLRDA perform competitively well relative to the classical methodologies. The worst classifier is the KNN due to its consistent high error rates compared to the other methodologies. On the other hand, for the preprocessed data, the methods are assessed based on the error rates and the type of misclassification (i.e whether a normal tissue is classified as tumor and vice versa, especially for the Colon and Prostate data sets). The distribution of the classification error rates are also studied through box plots. For the preprocessed Colon data, PLSDA emerged as the best, followed by RPLS, PLSGLRDA while KNN emerged as the worst methodology.

Looking at the type of misclassification, PLSDA had the lowest proportion of cancer tissues classified as normal. Furthermore, the RPLS and PLSGLRDA also had a relatively lower proportion of cancer tissue classified as normal. For the preprocessed Leukemia data, SVM emerged as the best followed by RPLS, PLSDA and LDA, with KNN being the worst classifier. For the preprocessed Prostate data, PLSDA was the best followed by RPLS, SVM and PLSGLRDA in that order. The SVM and the RPLS have a relatively lower proportion of tumor tissues classified as cancer and so in addition to their low classification errors, they can be considered as good classifiers. In general, it is important to note that the difference in the classification error rates between the methodologies is very small and that no particular methodology has been declared the “winner" in all the cases for the preprocessed data. As such it is fair to conclude that there is no clear winner for the classifications of preprocessed data sets. However, there is a clear “loser” which is the KNN since it has consistently performed poor.

For the statistical integration of molecular data to test for network changes, we start by first identifying the prior network to be used. The network is curated from the literature and (or) online databases. The genes considered in the experimental data are then identified so that only the nodes of the prior network containing the said genes are retained. From the prior network, an edge list containing the connections between different nodes (gene) are identified. We take advantage of the fact that genes usually act in groups according to some pathological functional groups and this kind of relation is reflected the adjacency matrix from the prior network. Recall that the experimental data has two main groups healthy (H) and ARF. These groups have two subgroups each namely, the GAS stimulated and unstimulated. We use the prior network edge list to construct the networks for each of the two groups (H and ARF). Furthermore, we colour the nodes for each network to reflect the up or down regulation or no change for each of the gene expression. We identify a sub network from a given group of genes guided by biology (e.g some functional group or genes associated with some disease). From this information, we extract subnetworks from network for the healthy and the one for the ARF. We then test for the changes in the nodes (genes) with regards to log fold change. These changes in the nodes is what we refer to as changes. We have shown that this is not a trivial problem and so derived a likelihood ratio test for these changes. The derived test do follow a chi-square distribution with p degrees of freedom when the variance-covariance matrix

is known. We have assumed that the variables (genes) follow a multivariate normal distribution with a known variance-covariance matrix which can be deduced from the prior network that has been chosen. Finally, the likelihood ratio test statistic for changes has been derived when the variance-covariance matrix is unknown. A simulation study has been done and demonstrated that the developed tests can be useful for testing the network changes and can be applied to other cases which have similar problem set-ups. The test statistics have a chi-square distribution when the number of variables are few but tends to a normal distribution as the number of variables increase.

5.2 Future work

With respect to the classification problem, it would be interesting to study the mathematics behind the good performance of the kernel multilogit algorithm's superior performance in the classification problems when applied to the noisy unpreprocessed data sets.

For the part on testing for network changes the following are possible open problems for future work.

- In this thesis we developed methods for analyzing only two groups (Healthy and ARF) but these methods can be extended to situations involving more than two groups.

- It would be to derive a test statistic for testing the changes by constraining the covariance matrix to reflect the structure of the prior network and considering the whole network of p genes and not the subnetworks. That is, by using the structure of the adjacency matrix of the prior network used in the integration of network information with the molecular data.

- In situations for testing networks where the number of samples is less than the number of genes then a regularized estimate for Σ would be needed.

- Since the multivariate tests are usually very sensitive to outliers, it would be important to develop a robust variant of the methods developed in this part of the thesis.

- The developed test can be extended from the normal distribution to a more general family of distributions for instance the elliptical family.

Bibliography

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96(12):6745–6750.

Alshamlan, H., Badr, G., and Alohali, Y. (2013). A study of cancer microarray gene expression profile: Objectives and approaches. In *Proceedings of the World Congress on Engineering, London, U.K.*, volume II.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience.

Awada, W., Khoshgoftaar, T., Dittman, D., Wald, R., and Napolitano, A. (2012). Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on. In *A Review of the Stability of Feature Selection Techniques for Bioinformatics Data*, pages 356–363.

Babu, M. (2004). Introduction to microarray data analysis. In Grant, R., editor, *Computational Genomics: Theory and Application*. Horizon Bioscience, U.K.

Bastien, P., Vinzi, E. V., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics and Data Analysis*, 48:17–46.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach for multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bhatnagara, A., Grover, A., and Ganguly, N. (1999). Superantigen-induced t cell responses in acute rheumatic fever and chronic rheumatic heart disease patients. *Clinical and Experimental Immunology*, 116(1):100–106.

Boulesteix, A. (2004). PLS dimension reduction for classification with microrray data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–30.

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS Letters*, 480(1):17–24.

Carapetis, J., Brown, A., Wilson, N., and Edwards, K. (2007). An australian guideline for rheumatic fever and rheumatic heart disease: an abridged outline. *Medical Journal of Australia*, 186(11):581–586.

Carapetis, J., McDonald, M., and Wilson, N. (2005). Acute rheumatic fever. *The Lancet*, 366(9480):155 – 168.

Chen, Y., Lun, A. T. L., and Smyth, G. (2014). *Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR*, pages 51–74. Springer International Publishing, New York.

Chun, H. and Keles, S. (2009). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 72(1):3–25.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Dalmau, O., Alarcón, T. E., and González, G. (2015). Kernel multilogit algorithm for multiclass classification. *Computational Statistics and Data Analysis*, 82:199–206.

Dettling, M. and Bühlmann, P. (2002). Supervised clustering of genes. *Genome Biology*, 3(12):1–15.

Dong, K., Zhang, F., Zhu, Z., Wang, Z., and Wang, G. (2014). Partial least squares based gene expression analysis in posttraumatic stress disorder. *European Review for Medical and Pharmacological Sciences*, 18:2306–2310.

Dudoit, S. and Fridlyand, J. (2003). Statistical analysis of gene expression microarray data. In Speed, T. P., editor, *Statistical Analysis of Gene Expression Microarray Data*, chapter 3, pages 100–165. Chapman & Hall/CRC, London.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–86.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 7:1104–1111.

Gagnon-Bartsch, J. A. and Speed, T. (2011). Using control genes to correct for unwanted variation in microarray data. *Biostatistics (Oxford, England)*, 13(3):539–552.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(531-537).

Gromski, S., Muhamadali, H., Ellis, D., Xu, Y., Correa, E., Turner, M., and Goodcare, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, 879:10–23.

Gusnanto, A., Ploner, A., Shuweihdi, F., and Pawitan, Y. (2013). Partial least squares and logistic regression random-effects estimates for gene selection in supervised classification of gene expression data. *Journal of Biomedical Informatics*, pages 697–709.

Han, J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer-Verlag, 2nd edition.

Höskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2:211–228.

Huang, C., Tu, S., Huang, C., Lien, H., Lai, L., and Chuang, E. (2013). Multiclass prediction with partial least square regression for gene expression data:

Applications in breast cancer intrinsic taxonomy. *BioMed Research International*, pages 1–9.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl.1):S233–S240.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934.

Jayaswal, V., Schramm, S., Mann, G., Wilkins, M., and Yang, Y. (2013). Van: an R package for identifying biologically perturbed networks via differential variability analysis. *BMC Research Notes*, 6:430–430.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Analysis*. Prentice Hall, Upper Saddle River, New Jersey.

Jones, T. (1944). The diagnosis of rheumatic fever. *Journal of the American Medical Association*, 126(8):481 – 484.

Lê Cao, K., Rossouw, D., Robert-Granieé, C., and Besse, P. (2008). A Sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), Article 35.

Lee, D., Lee, W., Lee, Y., and Pawitan, Y. (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemometrics and Intelligent Laboratory Systems*, 109(1):1 – 8.

Liang, Y., Liu, C., Luan, X., Leung, K., Chan, T., Xu, Z., and Zhang, H. (2013). Sparse logistic regression with a l1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics*, 14(1):198.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980). *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press, 1 edition.

Martin, W., Steer, A., Smeesters, P., Keeble, J., Inouye, M., Carapetis, J., and Wicks, I. P. (2015). Post-infectious group A streptococcal autoimmune syndromes and the heart. *Autoimmunity Reviews*, 14(8):710 – 725.

Metzker, M. L. (2010). Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, 11(1):31–46.

Mitchell, J. (1994). Classical statistical methods. In Michie, D., Spiegelhalter, D., and Taylor, C., editors, *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.

Mitra, R., Gill, R., Datta, S., and Datta, S. (2014). *Statistical Analyses of Next Generation Sequencing Data: An Overview*, chapter 1, pages 1–24. Springer International Publishing, Cham.

Nguyen, D. and Rocke, D. M. (2002a). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18(9):1216–1226.

Nguyen, D. and Rocke, D. M. (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50.

Schramm, S., Li, S., Jayaswal, V., Fung, D. C. Y., Campain, A., Pang, C. N. I., Scolyer, R. A., Yang, Y. H., Mann, G. J., and Wilkins, M. R. (2013). Disturbed protein-protein interaction networks in metastatic melanoma are associated with worse prognosis and increased functional mutation burden. *Pigment Cell & Melanoma Research*, 26(5):708–722.

Seber, G. (2004). *Multivariate Observations.* Wiley & Sons, New York.

Seckeler, M. and Hoke, T. (2011). The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. *Clinical Epidemiology*, 3:67–84.

Singh, D., Febbo, F., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., and Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., and Wrana, J. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotech*, 27(2):199–204.

Telaar, A., Liland, K., Repsilber, D., and Nürnberg, G. (2013). An extension of PPLS-DA for classification and comparison to ordinary PLS-DA. *PLoS ONE 8*, 2:e55267.

van Dijk, E., Jaszczyszyn, Y., and Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research*, 322(1):12 – 20.

Ventimiglia, G. and Petralia, G. (2013). Recent advances in DNA microarray technology: an overview on production strategies and detection methods. *BioNanoSci*, 3:428 – 450.

Wang, A., An, N., Chen, G., Li, L., and Alterovitz, G. (2015). Improving pls–rfe based gene selection for microarray data classification. *Computers in Biology and Medicine*, 62:14–24.

Wold, S., Ruhe, A., Wold, W., and Dunn III, W. J. (1984). The collinearity problem in linear regression, the partial least squares approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Wold, S., Sjöström, M., and Erikson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130.

Xi, B., Gu, H., Baniasadi, H., and Raftery, D. (2014). Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Methods Mol Biol.*, 1198:333–353.

Zhu, J., Yamane, H., and Paul, W. (2010). Differentiation of effector cd4 t cell populations. *Annual Review of Immunology*, 28(1):445–489. PMID: 20192806.

Ziegler, A. and König, I. (2008). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley-VCH Verlag GmbH & Co. KGaA.