3-29-2010

Statistical Learning and Behrens Fisher Distribution Methods for Heteroscedastic Data in Microarray Analysis

Nabin K. Manandhr-Shrestha

University of South Florida

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the American Studies Commons, Mathematics Commons, and the Statistics and Probability Commons

Scholar Commons Citation

Manandhr-Shrestha, Nabin K., "Statistical Learning and Behrens Fisher Distribution Methods for Heteroscedastic Data in Microarray Analysis" (2010). Graduate Theses and Dissertations.
http://scholarcommons.usf.edu/etd/3513

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Statistical Learning and Behrens-Fisher Distribution Methods
for Heteroscedastic Data in Microarray Analysis

by

Nabin K. Manandhar Shrestha

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Kandethody M. Ramachandran, Ph.D.
G. S. Ladde, Ph.D.
Marcus M. McWaters, Ph.D.
Tapas K. Das, Ph.D.

Date of Approval:
March 29, 2010

Keywords: Genes, False Discovery Rate, Multiple Testing, Correlation, Classification

© Copyright 2010, Nabin K. Manandhar Shrestha
Dedication

To My Parents
Acknowledgements

I express my sincere and deepest gratitude to my research supervisor and dissertation advisor Professor K. M. Ramachandran. Also, I would like to pay my gratitude to Professor G. Ladde for his guidance throughout my graduate studies and providing valuable suggestions in this study. This work could never have been completed without their constant guidance and supports. I would also like to thank Professor Marcus M. McWaters and Professor Tapas K. Das for their advice, support and serving in my dissertation committee.

A special thank goes to Professor Gordon Fox for his kind willingness to chair the defense of my dissertation. Last but not the least, I am thankful to my parents, my uncle Badri K. Shrestha, my wife Sharmila and my son Sangam Shrestha for their constant supports and encouragement.
Table of Contents

List of Tables iv
List of Figures vi
Abstract vii

1 Microarrays and Selection of Differentially Expressed Genes 1
  1.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
  1.2 Gene Expression and Main Question of Interest . . . . . . . . . . . . 3
  1.3 What is microarray and how does it work ? . . . . . . . . . . . . . . 3
  1.4 Oligonucleotide microarray . . . . . . . . . . . . . . . . . . . . . . . 4
  1.5 cDNA Microarray . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  1.6 Measuring the Expression Levels of a Gene . . . . . . . . . . . . . 7
  1.7 Statistical Methods for Differentially Expressed Genes . . . . . . . . 8
  1.8 Methods for cDNA Data . . . . . . . . . . . . . . . . . . . . . . . . 9
  1.9 Methods for Oligonucleotide Data . . . . . . . . . . . . . . . . . . . . 11
  1.10 Assessing the Reliability of Tests . . . . . . . . . . . . . . . . . . . 14
  1.11 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

2 Behrens-Fisher distribution for selecting Differentially expressed Genes 17
  2.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
  2.2 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
  2.3 Multiple Testing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
  2.4 Measures of Erroneous Rejection of Null Hypotheses . . . . . . . . 22
  2.5 Two-Sample Permutation Test . . . . . . . . . . . . . . . . . . . . . 25
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6 Significance Analysis of Microarrays (SAM)</td>
<td>26</td>
</tr>
<tr>
<td>2.7 Sampling Distribution for Non-homogeneous Variance</td>
<td>27</td>
</tr>
<tr>
<td>2.8 Bayesian Approach</td>
<td>27</td>
</tr>
<tr>
<td>2.9 Test Statistic</td>
<td>31</td>
</tr>
<tr>
<td>2.10 Calculation of Prior d.f. and Prior Variance</td>
<td>34</td>
</tr>
<tr>
<td>2.11 Estimation of Hyperparameters</td>
<td>34</td>
</tr>
<tr>
<td>2.12 Simulation and Result</td>
<td>39</td>
</tr>
<tr>
<td>2.13 Power Comparisons</td>
<td>41</td>
</tr>
<tr>
<td>2.14 Comparison of DE Genes using the Window Method with Other Methods</td>
<td>46</td>
</tr>
<tr>
<td>2.15 Comparison of Proposed Method with other Methods</td>
<td>50</td>
</tr>
<tr>
<td>2.16 Selection of Differentially Expressed Genes in the Golub Data</td>
<td>51</td>
</tr>
<tr>
<td>2.17 Transformation and Test of Normality Assumptions</td>
<td>55</td>
</tr>
<tr>
<td>2.18 Conclusion</td>
<td>58</td>
</tr>
<tr>
<td>3 Support Vector Machines and Other classification Methods</td>
<td>63</td>
</tr>
<tr>
<td>3.1 Summary</td>
<td>63</td>
</tr>
<tr>
<td>3.2 Introduction</td>
<td>63</td>
</tr>
<tr>
<td>3.3 Assumptions on Classification Methods</td>
<td>64</td>
</tr>
<tr>
<td>3.4 Support Vector Machines (SVM) Methods</td>
<td>64</td>
</tr>
<tr>
<td>3.5 Kernel Matrix and Kernel Tricks</td>
<td>67</td>
</tr>
<tr>
<td>3.6 High Dimensional Feature Space for Large $p$ Small $n$</td>
<td>67</td>
</tr>
<tr>
<td>3.7 Nearest Shrunken Centroids Method</td>
<td>69</td>
</tr>
<tr>
<td>3.8 Weighted Voting Method</td>
<td>72</td>
</tr>
<tr>
<td>3.9 Dudoit’s Multi-class Classification Method</td>
<td>74</td>
</tr>
<tr>
<td>3.10 Gene Selection by Behrens-Fisher Statistic</td>
<td>74</td>
</tr>
<tr>
<td>3.11 Choosing the number of genes required for classification</td>
<td>77</td>
</tr>
<tr>
<td>3.12 Simulation and Results</td>
<td>78</td>
</tr>
<tr>
<td>3.13 Datasets Pre-processing and Filtering</td>
<td>79</td>
</tr>
<tr>
<td>3.14 MLL Leukemia Data</td>
<td>79</td>
</tr>
<tr>
<td>3.15 Golub Leukemia Data</td>
<td>82</td>
</tr>
<tr>
<td>3.16 Subtypes of Pediatric Acute Lymphoblastic Leukemia</td>
<td>85</td>
</tr>
</tbody>
</table>
3.17 Conclusion ................................................. 88

4 Selection of Differentially Expressed Genes in Correlated Statistics 89
4.1 Summary .................................................. 89
4.2 Introduction .............................................. 89
4.3 Effects of Correlations ................................. 90
4.4 Application ................................................ 92
4.5 Computation of Effect of Correlations .............. 98
4.6 Conditional and Unconditional \( p \) - values ........ 99
4.7 Dispersion Parameter .................................. 100
4.8 Conclusion ................................................ 102

5 Performance of a Classifier taking into account of Correlated Genes 103
5.1 Performance of Classifiers ............................ 103
5.2 McNemar’s Test and Confusion Matrix .......... 104
5.3 Filtering the Highly Correlated Genes ............. 105
5.4 Classification using SVM ............................. 106
5.5 Application ............................................... 107
5.6 Confusion Matrices ................................... 111
5.7 Binary Accuracy Measures ......................... 113
5.8 Classification using the genes selected by the Correlated test Statistics 114
5.9 Conclusion ............................................... 114

6 Future Research ........................................... 115

References .................................................... 116

About the Author ............................................. End Page
List of Tables

2.1 Multiple testing Procedure in Simultaneous Hypothesis Testing . . . . 20
2.2 Table of Power Comparison . . . . . . . . . . . . . . . . . . . . . . 47
2.3 Comparison of BF test with other tests based on the proportion of actually DE genes selected in Window Method. . . . . . . . . . . . . 48
2.4 Comparision of BF test with other tests based on the proportion of actually DE genes selected in Similar Variance Method. . . . . . . . 48
2.5 Comparision of BF test with other tests based on the proportion of actually DE genes selected in Resampling Method. . . . . . . . . . . 49
2.6 Comparison of three methods ( I = Window Method, II = Similar Variance Method, III = Resampling Method ) according to Proportion of truly DE genes selected, FDR and FNR. . . . . . . . . . . . . . . 51
2.7 Dependence of Variance on Expression level . . . . . . . . . . . . . . 57
2.8 Comparision of Differentially expressed genes in Golub data controlling the FDR at the level $\alpha = 0.05$ . . . . . . . . . . . . . . . . . . . . . . 58
3.1 Accuracy Rate for the Simulated Test data . . . . . . . . . . . . . . . 79
3.2 Comparison of Classification Performance on MLL Leukemia Data. . 82
3.3 Comparison Classification of Golub Data using all the genes . . . . . . 84
3.4 Comparison Classification of Golub Data using BF selected genes . . 84
3.5 Comparison of Classification Performance on Golub Data. . . . . . . 85
3.6 Comparison of Classification Performance on ALL-7 Data. . . . . . . 87
3.7 Confusion matrix for the MLL-3 training data by Dudoit method, and Golub Test Data by NSC Method. . . . . . . . . . . . . . . . . . . . . . 87
3.8 Confusion matrix for the ALL-7 test data by SCRDA method. . . . . 88
4.1 Goodness-of-Fit test of the test-statistics obtained by the BF method for the Golub Data ........................................ 93
4.2 Effect of Correlation on the standard deviations of the middle and tail count of the transformed test statistics ......................... 99
4.3 Effect of correlation on p-value on test statistics in Golub data ... 100

5.1 Contingency table for misclassification .......................... 105
5.2 Confusion Matrix for misclassification ............................. 105
5.3 Classification of Golub Data taking into Correlation among the genes 108
5.4 Comparison of Classification Performance on Golub Data after taking into correlation between the genes. Different percentage points were used to determine the optimal cut-point. .................. 109
5.5 Comparison of Classification Performance on ALL-7 Data using the genes Selected by the BF method and taking into correlation between the genes. The classes were compared using one versus another. ... 110
5.6 Confusion Matrices of Test samples of Golub Data by Weighted Vote, SVM and Dudoit methods ................................. 111
5.7 Confusion Matrix of MLL data applying the Dudoit Classification method after taking correlation ............................... 111
5.8 Confusion Matrix of MLL data applying the Weighted Vote Classification method after taking correlation ......................... 112
5.9 Confusion Matrix of MLL data applying the SVM Classification method after taking correlation ................................. 112
5.10 Contingency table for misclassification of Golub Test Data ... 112
5.11 Contingency table for misclassification of MLL Test Data .......... 113
5.12 Confusion matrix derived accuracy measures, \( N = a + b + c + d \) ... 113
5.13 Accuracy measures for the Wt. Vote, SVM, Logistic Discrimination and Dudoit Multi-class Classifiers ................................. 113
5.14 Classification of Golub Data taking into Correlated test statistics .. 114
List of Figures

2.1 Power comparison for different sample size and priors using Method 1 42
2.2 Power of the Proposed Method at m=n=3 43
2.3 Power of the Proposed Method at m=n=5 44
2.4 Graph of DE genes 45
2.5 Histograms of Unprocessed Data Golub Data 52
2.6 Histogram of log-2 transformed Golub Data 53
2.7 QQ-Plot of Golub Data 54
2.8 Q-Q plot of some of the genes before Yeo-Johnson Transformation 60
2.9 Histogram of the T-values obtained by the BF Method 61
2.10 Q-Q plot of some of the genes after Yeo-Johnson Transformation 62

3.1 Histogram of Pre-processed AML Data 80
3.2 Histogram of ALL,MLL and AML Training and Testing Data after Pre-processing 81
3.3 Cross-Validation Error of Nearest Shrunken Centroid Method for MLL data 83
3.4 Error Comparison for Golub data 86

4.1 Histogram and the fitted Cauchy density curve for the BF test statistics 94
4.2 Histogram of transformed test-statistics 95
4.3 Density, $g(\rho)$, of correlations between the genes 96
The aim of the present study is to identify the differentially expressed genes between two different conditions and apply it in predicting the class of new samples using the microarray data. Microarray data analysis poses many challenges to the statisticians because of its high dimensionality and small sample size, dubbed as "small n large p problem". Microarray data has been extensively studied by many statisticians and geneticists. Generally, it is said to follow a normal distribution with equal variances in two conditions, but it is not true in general. Since the number of replications is very small, the sample estimates of variances are not appropriate for the testing. Therefore, we have to consider the Bayesian approach to approximate the variances in two conditions. Because the number of genes to be tested is usually large and the test is to be repeated thousands of times, there is a multiplicity problem. To remove the defect arising from multiple comparison, we use the False Discovery Rate (FDR) correction. Applying the hypothesis test repeatedly gene by gene for several thousands of genes, there is a great chance of selecting false genes as differentially expressed, even though the significance level is set very small. For the test to be reliable, the probability of selecting true positive should be high. To control the false positive rate, we have applied the FDR correction, in which the $p$-values for each of the gene is compared with its corresponding threshold. A gene is, then, said to be differentially expressed if the $p$-value is less than the threshold.

We have developed a new method of selecting informative genes based on the Bayesian Version of Behrens-Fisher distribution which assumes the unequal variances in two conditions. Since the assumption of equal variances fail in most of the sit-
uation and the equal variance is a special case of unequal variance, we have tried to solve the problem of finding differentially expressed genes in the unequal variance cases. We have found that the developed method selects the actual expressed genes in the simulated data and compared this method with the recent methods such as Fox and Dimmic’s $t$-test method, Tusher and Tibshirani’s SAM method among others.

The next step of this research is to check whether the genes selected by the proposed Behrens -Fisher method is useful for the classification of samples. Using the genes selected by the proposed method that combines the Behrens Fisher gene selection method with some other statistical learning methods, we have found better classification result. The reason behind it is the capability of selecting the genes based on the knowledge of prior and data. In the case of microarray data due to the small sample size and the large number of variables, the variances obtained by the sample is not reliable in the sense that it is not positive definite and not invertible. So, we have derived the Bayesian version of the Behrens Fisher distribution to remove that insufficiency. The efficiency of this established method has been demonstrated by applying them in three real microarray data and calculating the misclassification error rates on the corresponding test sets. Moreover, we have compared our result with some of the other popular methods, such as Nearest Shrunken Centroid and Support Vector Machines method, found in the literature.

We have studied the classification performance of different classifiers before and after taking the correlation between the genes. The classification performance of the classifier has been significantly improved once the correlation was accounted. The classification performance of different classifiers have been measured by the misclassification rates and the confusion matrix.

The another problem in the multiple testing of large number of hypothesis is the correlation among the test statistics. we have taken the correlation between the test statistics into account. If there were no correlation, then it will not affect the shape
of the normalized histogram of the test statistics. As shown by Efron, the degree of the correlation among the test statistics either widens or shrinks the tail of the histogram of the test statistics. Thus the usual rejection region as obtained by the significance level is not sufficient. The rejection region should be redefined accordingly and depends on the degree of correlation. The effect of the correlation in selecting the appropriate rejection region have also been studied.
Chapter 1

Microarrays and Selection of Differentially Expressed Genes

1.1 Introduction

Deoxyribonucleic Acid (DNA) microarray technology was first mentioned in an article by Schena et. al. [38] published in the ’Genome Issue’ of Science in 1995. After its publication, this technology attracted the attention of genome researchers. Nowadays, it is one of the most advanced technologies to know the gene expression. In the course of understanding and deciphering the genomes of many organisms, there was a need for functional studies of thousands of genes across tissue samples. There was a need for identification of expression patterns of genes under normal and pathological conditions. The access of genome sequencing was due to the development of high throughput DNA sequencing technology which created the system to approach biology. The most important throughput technology is the DNA microarrays technology which allows the researchers to make snapshots of genes in an organism in a single experiment. This technology allows the researchers to identify genes that are expressed in different cell types and conditions, to learn how their expression level changes in different developmental stages or disease states; and to identify the cellular process in which they participate. This technology produces a huge amount of information that can provide clues about how genes and gene product interacts and their interaction networks. But, transforming this data into knowledge is not a trivial task. Analysis using multiple techniques is needed to provide the comprehensive view of the underlying biology.
In 1953, James Watson and Francis Crick [50] established the structure of DNA. The structure of DNA is a double helix, which is like a twisted ladder. Genes are made up of DNA and RNA (Ribonucleic Acid). Both DNA and RNA are polymers, that is, molecules that are constructed by sequentially binding the members of a small subunits called nucleotides into a linear strand or sequence. Each nucleotide consists of a base, attached to a sugar, which is attached to a phosphate group. The linear strand consists of alternate sugars and phosphates, with the bases protruding from the sugars. In DNA, the sugar is deoxyribose and the bases are Cytosine(C), Guanine (G), Thymine(T), Adenine(A). In RNA, the sugar is ribose and the bases are Cytosine(C), Guanine (G), Uracil(U) and Adenine(A). The genetic information of cellular organism is stored in long sequence of these four different bases that bond in a certain way - A bonds with T (or U), and C bonds with G via the hydrogen bonds. The sugar-phosphate backbone can, for the purpose of informatics, be considered as straight; though actually it has all sorts of twists, kinks and loops. The bases that protrude from the backbone are far more informative. The DNA is, thus, can be regarded as double stranded polymer. These DNA strands are complementary to each other meaning that every guanine (G) in one strand corresponds to a cytosine (C) in other complementary strand, and every adenine (A) in one strand corresponds to thymine (T) in the other complementary strand. The strings of nucleotides (bases) are the DNA molecules that compose the genome of an organism. These genome contain segments of DNA that encode genes. Genes are the functional and physical units of heredity that are passed from parent to offspring. Genes can be thought as a segment of DNA sequence that corresponds to a particular protein. Genome is the set of DNA molecules. DNA has two strands:

1. Sense strand

2. Anti-sense strand

Genes are transcribed into RNA called messenger RNA (mRNA) and are translated to form proteins. The RNA in the intermediate stage is called mRNA because it is used as a platform to form proteins from DNA and to pass the information to proteins.
that is encoded by gene or DNA. These proteins are building blocks and functional units of a living cell. The process converting gene into proteins is called the gene expression. It occurs in two steps.

1. Transcription : The process of converting DNA into the mRNA
2. Translation : The process of converting mRNA into protein.

The information transfer from DNA to mRNA and mRNA to protein is called the ”Central Dogma of Biology.”

1.2 Gene Expression and Main Question of Interest

Each of the genes encoded in the DNA molecule either transforms into proteins through the transcription and translation or remains unchanged under that condition. If the gene is transformed, then we say the gene is expressed in that condition. DNA is a stable molecule and the same genomic DNA is present, with a few exception, in all cells of living organism. Despite this all the cells are not the same, like-hair, muscle, skin cell etc. Why? The difference between the cells is due to the different subset of genes that are expressed in each of the different cell types. Different subset of genes are expressed in response to stimuli so that the pattern of gene expression level reflects both- type of cell and its condition. The amount of each mRNA detected in the cell can provide information on the corresponding protein and the relationship between abundance of mRNA and formation of protein.

1.3 What is microarray and how does it work ?

A microarray is typically a glass/ polymer slide onto which DNA molecules are attached at fixed locations, called spots or features. There may be tens of thousands of spots on an array- each spot containing tens of millions of identical DNA molecules. For gene expression, each of these DNA molecules should identify (transcribe) to a single mRNA molecule in a genome. The features are either printed on the microarrays by a robot or a jet; or synthesized in situ by photolithography, or by ink-jet
printing. The most popular microarray application is to compare the expression levels of genes in two different samples; e.g. the same cells under two different conditions. This is done by labeling the mRNA obtained by the reverse transcription method, extracted from each sample of two different ways by two colors-green from the normal and red from the experimental condition. The hybridized microarrays is excited by a laser and scanned at wavelengths suitable for the detection of red and green dye intensities. The amount of fluorescence emitted upon laser excitation corresponds to the amount of nucleic acid bound to each spot. If the nucleic acid from sample condition 1 is abundance, the spot appears green, while if from condition 2 is abundance, it appears red; if both are equal, it appears yellow; and if neither present, appears black. Thus from the fluorescence intensities and colors of each spot, the relative expression levels of the genes in both samples are estimated. Hence, thousands of data points and information of expression levels of a particular transcript can be obtained from a single experiment. Two types of microarrays are currently available.

1. cDNA microarray

2. Oligonucleotide microarray

1.4 Oligonucleotide microarray

These are the list of terms that are used in this type of microarrays. **Probe:** oligonucleotides of 25 base pair length used to probe RNA targets. **Perfect Match:** probes intended to match perfectly the target sequence. **PM:** intensity value read from the perfect matches. **Mismatch:** the probes having one base mismatch with the target sequence intended to account for non-specific binding. **MM:** intensity value read from the mis-matches. **Probe Pair:** a unit composed of a perfect match and its mismatch.

Oligonucleotide arrays are used to measure the abundance of mRNA transcripts for many genes simultaneously. 11 to 20 perfect match (PM) and mismatch (MM) probe pairs are used to measure the expression level of each gene. Here is an example
of probe and PM and MM: Probe Pair: This consists of the 4 kinds of bases: A, C, G, T. Generally, 25 of these base sequence is selected from the gene of interest.

\[
\begin{align*}
5' & \ldots \ldots A \ G G G \ T \ G \ C C C C T T T G \ A A A \ldots . . . . . . 3' \text{ (sense strand)} \nonumber \\
3' & \ldots \ldots T \ C C C \ A \ C \ G G G G A A C \ T T T \ldots . . . . . . 5' \text{ (antisense strand)}
\end{align*}
\]

to

\[
\begin{align*}
5' & \ldots \ldots A \ G G G \ U \ G \ C C C C U U U U G \ A A A \ldots . . . . . . 3' \nonumber \\
3' & \ldots \ldots U \ C C C \ A \ C \ G G G G A A C \ U U U \ldots . . . . . . 5'
\end{align*}
\]

translation into mRNA

The only difference between DNA and RNA is that the base thymine (T) is replaced by the base uracil (U).

Examples of PM and MM probes:

PM probe: ATGATCTCGAATAGCGTGCGCGAAT

MM probe: ATGATCTCGAATGCCTGCGCGAAT

This is an example of probe pair. 11-20 such probe pairs of 25-bases is chosen in each spot of microarray to get the expression of a gene. The polymer such formed is also called a probe for a gene. The two probes are called complementary to each other. The only difference between PM probe and MM probe is that the middle base (A in the above example) is flipped to its complementary base T. Then simple or weighted/robust average of the difference PM-MM of all probe sets is taken. This is called the average difference which measures the abundance of mRNA in the oligonucleotide affymetrix data. The human genome has about \(2.8 \times 10^9\) base pairs and it encodes at least 40,000 genes.
1.5 cDNA Microarray

The mRNA from two tissues are extracted, separately reverse transcribed, and labeled with different colors- green (Cy3), red (Cy5). The mixture labeled cDNA are hybridized onto the different spots of glass slides (called microarray). These spots already contains the abundance identical probe sequence of complementary DNAs. Thus the colored cDNAs compete to bind (hybridize) to their complementary cDNA in each spot. After hybridization, the microarray slide is washed and they are scanned at different wavelengths by a laser or by a charged-coupled device (CCD) camera to obtain numerical intensity of each dye. Since the values thus obtained are so large, the statistical analysis are done transforming into $\log_{2}(\text{cy5/cy3})$. Generally base 2 is taken because, it equivalently transforms up-regulated and down-regulated genes. The intensities ranges from 0 to $2^{16}$.

The advantage of cDNA microarray is that they can be prepared directly from the isolated clones. Once the set of corresponding PCR products has been generated, microarrays can be created in multiple versions containing the entire set of cDNA sequences, resulting in the large-scale arrays for identification of differentially expressed genes of interest. It is less expensive to prepare. The cross-hybridization between homologous sequence is problematic for cDNA microarrays. The advantage of oligonucleotide microarrays is that they can be synthesized either in plates or directly on solid surfaces; making it easier to prepare. They are expensive and the probes in such array can be designed to represent unique gene gene sequences so that the cross hybridization between related gene sequences is minimized to a degree dependent upon the completeness of available sequence information.

Microarrays are applied to different situations. It helps to find the genes that are differentially expressed in different conditions. So, it has enormous importance in the field of Bio-medicine and Genomic. For example, as mentioned in the paper of Efron et al. [12], some cancer patients have severe life-threatening reactions to radiation treatment. So, it is important to recognize the basis of this sensitivity, so that such
patients can be identified before treatment is given. The treatment is life threatening, so if given to such patients who are sensitive to radiation it can threaten life and hence decision should be taken before use.

1.6 Measuring the Expression Levels of a Gene

The starting point of any statistical analysis begins from the estimation of expression level of each gene in the microarray. To measure the real gene expression, we have to measure the abundance of proteins. However, DNA microarray experiments measure the abundance of mRNA, but not the protein abundance. According to the simple traditional view of gene expression, there is a direct one-to-one mapping from DNA to mRNA to protein. To put in another way, a specific gene (i.e. genomic DNA sequence) always produce one and the same amino acid sequence of corresponding protein, which then fold to assume its native state. Given this simplified scheme, measuring the mRNA abundance would provide us with highly accurate information on protein abundance, as protein and mRNA abundance are proportional to direct mapping. There are thousands of spots (features) in a microarray slide. Each spot is associated with just one gene. For each gene/spot/feature, the amount of mRNA present is measured by the principle:

\[ \text{Amount of fluorescence emitted} \propto \alpha m, \text{ where, } m \text{ is the amount of mRNA present in the spot.} \]

To measure the expression level, the hybridized microarray is excited by the laser and scanned at the wavelengths suitable for the detection of both red and green intensities. From the fluorescence intensities and color of each spot/gene, we measure the expression level of the gene in the spot. Suppose that for gene \( g \),

\[ R_g = \text{expression level (florescent intensity) of gene } g \text{ in query sample} \]
\[ G_g = \text{expression level(florescent intensity) of gene } g \text{ in reference sample, and} \]

\[ T_g = \frac{R_g}{G_g} \]
is the ratio of intensities between query and reference sample.

- If the spot appears red, then $T > 1$.
- If the spot looks yellow, then $T = 1$.
- If the spot appears green, then $T < 1$.
- If the spot appears black, $T$ is undefined.

Similarly, for the two sample with replication, we measure the expression levels of green and red intensities of the same gene at more than one spot and we get the values of expression for two different condition.

The quality filtering is implemented to produce good quality of intensity measurement. For this, one uses the multiple spot replication slides. Multiple spotting of target DNA on a slide provides a means to assess the quality of data for a gene on that slide. Suppose each gene is spotted $p$ times on the slide. For each spot, a ratio of Cy3 and Cy5 intensity is calculated as $m = Cy_3/Cy_5$. Let $CV = \sigma/\bar{m}$ be the coefficient of variation of the set of ratios $m_1, m_2, ..., m_p$ on the multiple spots. The quality of the data on the expression level of each gene is inversely related to its CV. For each gene, a window subset containing 50 genes whose mean intensities are closest to the gene of interest is constructed. The CV of each gene is calculated and ordered in the increasing order. If the CV of the gene of interest falls within the top 10% of the CV’s then we discard this gene saying it has poor quality.

1.7 Statistical Methods for Differentially Expressed Genes

An exciting development in genomic is the use of microarray technology to simultaneously monitor the expression levels of thousands of genes. A common task is to compare the expression levels of genes in samples drawn from two different conditions. Specially, it is of interest to detect genes with differential expression under two different conditions. In early days, the simple method of fold-change was used and now it is known to be unreliable (Chen et al. [7]) because statistical variability
was not taken into account. Since then many more sophisticated statistical methods have been proposed (e.g. Chen et al.[7]; Efron et al. [12]; Newton et al.[29]; Tusher et al.[48]; Pan et al. [30]). It has also been noted that the data based on a single array may not be reliable and may contain high noises. As the technology advances, microarray experiments are becoming less expensive, which makes the use of multiple arrays (or multiple spots on each array) feasible. Hence it is possible to use the test that requires replicated measurements of expression levels of each gene under two different conditions. A straightforward method is to use the traditional two sample t-test. Thomas et al. [46] proposed a regression modeling approach. Pan et al. ([30]) suggested the a mixture model approach, which follows the basic idea of Efron et al. [12] and Tusher et al. [48]. On the other hand there are different methods using Bayesian and empirical Bayes method. Other methods are the linear models and empirical Bayes method of Smyth (2003) and ANOVA approach of Kerr et al.[22]. The other efficient methods are given by Florence et al. 2007 based on the Johnson’s distribution.

1.8 Methods for cDNA Data

A number of methods have been suggested for the identification of differentially expressed genes in single-slide two-color microarray experiments. In such experiments, the data for each gene consist of two fluorescence intensity measurements. Let \( R \) and \( G \) be the expression level of the gene in the red (Cy5) and green (Cy3) labeled mRNA samples respectively. Generally, the data consists of the value of the logarithmic base two of \( R/G \). The advantage of this transformation is that it produces a continuous spectrum of values for differentially expressed genes while treating up-and down-regulated genes equivalently. For example, if the expression ratio \( R/G = 4, 2, 1, 1/2, 1/4 \), the logarithmic base 2 has values \( 2, 1, 0, -1, -2 \). So, if the data is transformed by logarithmic base 2, then an up-regulated gene by 4 has value 2, and down-regulated gene by 4 (ratio = 1/4) has value –2.

Early analysis of microarray data (Schena et al., [38]) relied on the fold change cut-offs to identify differentially expressed genes. Typically a fold change equal to
2 or 3 is taken as the cut-off. If a gene has logarithmic base 2 ratio greater than 2 then the gene is said to have differentially expressed. Schena et al.[38] used a spiked control in mRNA samples to normalize the signals for the two fluorescent dyes and declared a gene as differentially expressed if the difference of the expression levels is more than 5 in two mRNA samples.

Let \( m_g = \log_2 \frac{R_g}{G_g} \) be the log expression of gene \( g \). If \( m'_g = \log_2 \frac{R'_g}{G'_g} \) be the log ratio of \( N \) ”housekeeping genes”, that is the genes believed not to be differentially expressed between two conditions of interest. Let \( m' \) be the mean and \( s' \) be the standard deviation of these housekeeping genes. De Risi et al.(1996)[8] declared the genes as DE if

\[
| \frac{m_g - m'}{s'} | > 3 \tag{1.8.1}
\]

A slightly more sophisticated approach involves calculating the mean and the standard deviation of the distribution of the ratio \( m_g \) and defining the global fold change difference and confidence. The ratio-intensity plot reveals that the data has more variability at lower intensities and less variability at higher intensities. So, using a sliding window at each gene we can access the local structure of the data to determine its differentiability. Let \( \bar{x}_g \) and \( s_g \) be the mean and standard deviation of log-2 ratio of gene \( g \) calculated by taking all the genes within the window, then

\[
z_g = \frac{\bar{x}_g}{s_g} \tag{1.8.2}
\]

is normally distributed with mean zero and standard deviation 1. Declare the genes as differentially expressed if \( |z_g| > 1.96 \). At higher intensities, \( s_g \) will be bigger and allows changes to be identified where the data is more variable.

Kerr et al. [22] introduced the use of ANOVA models that accounted for array, dye, and treatment effects for cDNA arrays. In this fashion, normalization was accomplished intrinsically without preliminary data manipulation. The model they proposed can be written as

\[
y_{ijk} = \mu + A_i + T_j + D_k + G_g + AG_{ig} + TG_{jg} + \epsilon_{ijk} \tag{1.8.3}
\]
where, \( \mu \) is the mean expression, \( A_i \) is the effect of the \( i \)th array, \( T_j \) is the effect of the \( j \)th treatment, \( D_k \) is the effect of the \( k \)th dye, \( G_g \) is the effect of the \( g \)th gene, and \( AG_{ig} \) and \( TG_{jg} \) are the interaction effects. Of interest for testing the differential expression are the interaction effects, \( TG_{jg} \), for which appropriate contrast can be estimated for each gene. In this model all effects were considered as fixed effects and other terms could be incorporated in the model.

### 1.9 Methods for Oligonucleotide Data

To detect the DE genes between two different conditions, the two sample \( t \)-test and its variants have been frequently used. Specifically, if we have samples from two conditions, the \( t \)-statistic is given by

\[
t = \frac{\bar{x}_2 - \bar{x}_1}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

where, \( \bar{x} \) are the means and \( s \) is the pooled sample standard deviation.

It is hard to verify the underlying assumption of normality because of small sample sizes perfectly, but with the common technologies this assumption is reasonable for the logarithm of the expression levels [1], [13], [11]. Baldi and Long [1] proposed a Bayesian version \( t \)-statistic using the priors. More specifically, if the \( x_{c1}^c, x_{c2}^c, ..., x_{cn_c}^c \) and \( y_{t1}^t, y_{t2}^t, ..., y_{tn_t}^t \) be the log transformed expression levels in the control and treatment conditions respectively, they have assumed the data has been transformed into the form such that normality assumption holds. So \( x \sim N(\mu, \sigma^2) \). The priors for means and variance has been chosen as

\[
p(\mu|\sigma^2) = N(x; \mu_0, \frac{\sigma_0^2}{\lambda_0}) \quad p(\sigma^2) = IG(x; \nu_0, \sigma_0^2)
\]

where, \( IG \) is the scaled inverse gamma pdf with degree of freedom \( \nu_0 > 0 \) and scale \( \sigma_0 > 0 \)

\[
IG(x; \nu_0, \sigma_0^2) = \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \sigma_0^{\nu_0} x^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\sigma_0^2}{2x}\right)
\]
The prior is obtained as

\[ p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2). \]

With this prior, the posterior was of the same form as prior, hence it was the \textit{conjugate} prior. Taking \( \mu_0 = \bar{x} \) and using the mean of the posterior distribution, the maximum a posteriori (MAP) estimate for mean \( \mu \) and variance \( \sigma^2 \) was shown to be

\[ \hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{\nu_0 \sigma_0^2 + (n - 1)s^2}{\nu_0 + n - 2} \]  

(1.9.6)

Other researchers used the log-normal and gamma-gamma model to detect the DE genes (Newton \textit{et al.}, [29]). Since the small variance gives rise to large \( t \)-statistic, empirical Bayes (EB) method of analyzing microarray data was used by Efron \textit{et al.}[12] without assuming any distributional assumption of the data. They slightly tuned the \( t \)-statistic by adding a suitable constant, \( a_0 \), which is generally taken as 90th percentile of standard deviation, on the \( Z \)-score obtained like in \( t \)-test,

\[ Z = \frac{\bar{D}_i}{a_0 + S_i}, \]

where, \( S_i \) is the standard deviation of the \( i \)th gene. If \( p_1 (p_0) \) are the prior probability that the gene is expressed (not expressed), \( p_1(z) (p_0(z)) \) are the posterior probabilities of a gene is expressed (not expressed), \( f_1(z) (f_0(z)) \) be the density of the expressed genes, then the mixture density for a gene is given by

\[ f(z) = p_0 f_0(z) + p_1 f_1(z). \]  

(1.9.7)

The densities \( f_0(z) \) and \( f_1(z) \) were estimated from the data without assuming any distribution, so it was called the \textit{Empirical Bayes} approach. The posterior probabilities can be written as

\[ p_1(z) = 1 - p_0 \frac{f_0(z)}{f_1(z)}, \quad p_0(z) = p_0 \frac{f_0(z)}{f_1(z)} \]  

(1.9.8)
The ratio \( \frac{f_0(z)}{f_1(z)} \) was obtained by fitting the logistic regression model

\[
\text{logit}(p_1(z)) = \beta_0 + \beta_1 z
\]  

(1.9.9)

and the priors \( p_0 \) and \( p_1 \) were estimated by

\[
p_1 \geq 1 - \min_z \frac{f(z)}{f_0(z)}, \quad p_0 \leq \min_z \frac{f(z)}{f_0(z)}
\]  

(1.9.10)

The estimate of the sample variance is not reliable for small sample data, specially in the case of microarray data, Tusher et al. (2001) [48] proposed the statistical analysis of microarray (SAM) method to stabilize the effect of large variances arising from low expressed genes. In this method, the SAM statistic is the same as in (1.9.4) but a small fudge factor \( s_0 \) is added in the denominator. The value of the fudge factor is chosen so that the coefficient of variation of the \( t \)-statistics is constant as a function of standard deviations. Generally, 90th percentile of \( s_i \) is used as the fudge factor.

After estimating the fudge factor, the \( t \)-scores is ranked in the decreasing order of \( t \)-statistics, i.e. \( t(1) \geq t(2) \geq \ldots \geq t(p) \), where, \( p \) is the number of genes. To select the threshold \( \Delta \) to determine the genes with \( t \)-scores greater and/or smaller than the threshold are differentially expressed, SAM uses the permutation of columns(arrays). The first \( n_1 \) of the permuted columns are taken to be as from condition 1 and the remaining columns are taken to be as from condition 2. We permute the arrays, say- \( B \) times. In each permutation \( b = 1, 2, \ldots, B \), the \( t \)-score thus obtained is ranked, say \( t_b(1) \geq t_b(2) \geq \ldots \geq t_b(p) \). The expected order statistics of gene \( g \) is calculated by \( t_E(g) = \frac{1}{B} \sum_{b=1}^{B} t_b(g) \). This means that look the \( g \)th descending ordered row \( g = 1, 2, \ldots, p \) of each column and take the average. This gives the expected order of \( g \)th ordered gene \( t(g) \) in the original ordering. The false discovery rate is calculated by

\[
FDR = \frac{1}{B} \sum_{b=1}^{B} \frac{N_b}{N}
\]

where, \( N_b \) is the number of genes in the \( b \)th permutation that has expected score smaller or greater than lower cut point, \( t(g_0) \), and upper cut point, \( t(g^*) \), respec-
tively, and \( N \) is the number of genes that has original \( t \)-score smaller or greater than \( t(g_0) \) and \( t(g^*) \) respectively.

1.10 Assessing the Reliability of Tests

The \( p \)-value was invented for testing the single hypothesis. But, in microarray data, there are thousands of genes, so, we need to test thousands of hypothesis simultaneously. In the case of repeated testing of hypotheses, the \( p \)-value is conceptually associated with the specificity of the test, \( i.e. \) it is used to control the false positive rate of a test. Declaring a test to be significant when \( p \)-value \(< 0.05 \) means that we are setting specificity 0.95. False discovery rate (FDR) of a test is defined as the expected proportion of false positives among the declared positives/significant results. If we declare 100 genes are significant and FDR is 0.1, then this means that we expect maximum of 10 out of 100 genes are false positives. No such interpretation is available from the \( p \)-value. The false negative rate (FNR) is defined as the expected proportion of true negatives out of actual positives. When controlling the FDR, one need to control the sensitivity or FNR. Setting FDR too low means that FNR is high. This means that the chance of including truly differentially expressed genes as not differentially expressed is high. So, FDR must be accompanied by sensitivity. Factors determining FDR:

1. Proportion of truly differentially expressed (DE) genes
2. Distribution of true differences
3. Measurement variability
4. Sample size.

What is the relation between sample size and FDR? Any statistical testing procedure applied on the gene by gene basis is characterized as follows:

- Compute the relevant test statistic for each gene
• Sort the statistic in order

• Determine a cut-off point beyond which all the genes are significant

Pawitan et. al. [32] has shown that, if we declare top \((1-p_0) \times 100\%\) as DE genes, then FDR=FNR. For a microarray test, generally one expects \(FDR = FNR\). Now to answer the question: To answer how the sample size affects the different rates (i.e. FDR, FNR, sensitivity), they proved that, as the sample size increases then the FDR decreases and the sensitivity (power) increases. FDR and sensitivity totals 1, so FNR decreases as the sample size increases. Given a true proportion of not-DE genes, the paper discusses the different rates as a function of the critical values. It then discusses the effect of sample size on the different rates. Again it discusses the different rates as the function of percentage of significant genes (obtained according to the top ranking genes). When one declares a small proportion of top genes as differentially expressed, it can lead to low sensitivity / large FNR. Sensitivity of a test increases with the number of samples per group. One can get high sensitivity declaring more proportion of top genes, but the small FNR has to pay a price for high FDR. So, declaring few top genes as DE is not only the solution of the problem of gene selection. What should one expect if one declares genes in the basis of critical values, as in \(t\)-statistic? When one declares genes on the basis of two sided test and critical values like \(t\)-test, then again the problem is due to the high FDR. In this case one must have big enough sample size to get high sensitivity and low FDR. It does not work for the small sample sizes like 5 samples per group but works well for 30 samples per group. Furthermore, it depends on the true proportion \(p_0\) of non-DE genes. If \(p_0\) is 0.99, this method also does not produce the satisfactory result even for sample size is large. But, on the other hand, if \(p_0\) is small, say 0.9, then it works well. If one chooses the method of fold change, one must choose the fold change of at least 3 times standard deviation to get better result.
1.11 Conclusion

In this chapter we have reviewed the literature about the microarray data analysis. The basic concepts of microarray data, how the expression of genes are measured and how these data are useful for identifying the differentially expressed genes. We have revisited some of the tools to analyze these data, why the microarray data is difficult to analyze and simple statistical tool is not effective for analyzing such a data. Furthermore, we have introduced the factors that should be taken care into account while identifying the differentially expressed genes.
2.1 Summary

In this chapter, we will derive an expression for the test statistics in the case of heterogeneous variances in the two samples using the Behrens Fisher distribution. Since the sample variances obtained from the microarray data are not a better estimates of the population variances, we use the priors for the estimation of variances. Then we compare the powers in the case of equal and unequal variances and study the effects of the variance in calculating the test statistics for the thousands of genes. We have compared the proposed method with other existing methods such as the equal variance method of Fox and Dimmic [13], the data dependent method of SAM [48], simple two sample $t$-test [4] and LIMMA [41] for the small sample size settings.

2.2 Introduction

The two sample $t$-test is one of the most popular methods for testing the difference between two samples [1], [13], [48]. There are different versions and variants of these tests: non-parametric and Bayesian version. In such tests, one is primarily interested on the identification of differentially expressed genes under two different conditions, so that those particular genes of interest are further studied. Most of the existing methods used in the literature for identification of differentially expressed genes are two-sample $t$-tests [35], [17], [1], [48] and its variants, SAM [48] and regularized $t$-test.
In the $t$-test, the variances of each of the genes in account together with the means. In this test, a gene is said to be expressed if $|t|$ exceeds a certain threshold depending on the confidence level selected. Since the distance between the sample means are standardized by the variances, this approach is better than the fold-change method. The gene expression data shows that there is an inherent limitation behind using the simple two-sample $t$-test. The variances of the genes depends on the expression level [35], [1], [48]. To identify the genes that are actually expressed, one should consider this fact. This fact was considered in the earlier works [17], [48].

There are two inherent problems in microarray experiments: First, the number of replications is very small, and second the number of genes to be tested is usually large and the test is to be repeated thousands of times. Since the small population size is very common in microarray studies, the sample estimates of variances are not appropriate for the testing. The variance of genes depends on the expression level [48], [35], [1]. Therefore, we have to consider the Bayesian approach to approximate the variances in two conditions. To remove the second defect arising from multiple comparison, we use the False Discovery Rate (FDR) correction [54]. Applying the hypothesis test repeatedly gene by gene for several thousands of genes, there is a great chance of selecting false genes as differentially expressed, even though the significance level is set very small. For the test to be reliable, the probability of selecting true positive should be high. To control the false positive rate, we have applied the FDR correction, in which the $p$-values for each of the gene is compared with its corresponding threshold. A gene is, then, said to be differentially expressed if the $p$-value is less than the threshold.

The Behrens - Fisher problem arises when one seeks to make inferences about the means of two normal populations without assuming the variances are equal. But, in many practical situations the populations do not have same variances. Although the Satterthwaites $t$-test deals with the unequal variances case, the degrees of freedom is small and hence it does not produce better estimate of the variances [1]. Here, we
briefly review the two sample $t$-test and introduce the proposed Behrens Fisher (BF) distribution.

Let $x = (x_1, x_2, ..., x_m)$ and $y = (y_1, y_2, ..., y_n)$ be two independent samples from two normal populations with means $\mu_x$ and $\mu_y$ and equal variances $\sigma^2_x = \sigma^2_y = \sigma^2$ respectively. If $\bar{x}$ and $\bar{y}$ are means; $s^2_x$ and $s^2_y$ are variances of $x$ and $y$ respectively; and $s^2 = \frac{(m-1)s^2_x + (n-1)s^2_y}{m+n-2}$ is the pooled sample variance, the two sample test statistic is given by

$$t = \frac{\delta - (\bar{y} - \bar{x})}{\sqrt{\frac{1}{m} + \frac{1}{n}} s^2}$$

is distributed as student’s $t$-statistic with $(m + n - 2)$ degrees of freedom.

Now, let the samples $x$ and $y$ are from normal distributions with unequal variances $\sigma^2_x$ and $\sigma^2_y$ respectively. In this case neither a pivotal statistic nor an exact confidence interval procedure exist [23]. We can take a statistic

$$t^* = \frac{\delta - (\bar{y} - \bar{x})}{\sqrt{s^2_x/m + s^2_y/n}} \sim t_{[\min(\nu_1, \nu_2)]}$$

where, $\nu_1 = m - 1$, $\nu_2 = n - 1$. If the sample sizes $m$ and $n$ are large, then the both $t$ and $t^*$ statistics give almost the same result. In the microarray experiments the sample sizes are relatively small, thus motivating us to look for an alternative.

2.3 Multiple Testing

To select the differentially expressed genes from thousands of genes, there are thousands of hypotheses each belonging to each gene. So the hypotheses must be tested simultaneously. For this, the hypothesis tests should be run thousands of times repeatedly. A problem with doing so many tests is that the number of false positives may be increased. This phenomenon is called multiple testing. The simultaneous hypotheses is:
<table>
<thead>
<tr>
<th>Number of Genes</th>
<th>Declared non-DE</th>
<th>Declared DE</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True non DE</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>True DE</td>
<td>$T$</td>
<td>$S$</td>
<td>$p - m_0$</td>
</tr>
<tr>
<td>$p - R$</td>
<td>$R$</td>
<td>$p$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Multiple testing Procedure in Simultaneous Hypothesis Testing

\[
H_0 : \begin{cases} 
\text{Gene 1 is } \text{not} \text{ differentially expressed} \\
\text{Gene 2 is } \text{not} \text{ differentially expressed} \\
\ldots \ldots \ldots \ldots \ldots \ldots \ldots \\
\text{Gene } p \text{ is } \text{not} \text{ differentially expressed.} 
\end{cases} \quad (2.3.1)
\]

\[H_1: \text{At least one Gene is differentially expressed.}\]

While testing the simultaneous hypotheses, one gets the number of hypotheses as in the table. Similar to testing a single hypothesis, the idea here is to control the number of false positives, $V$. This number is a random variable whose value differs from one test to another test. Let $\alpha$ be the type-I error that one makes when testing for a gene $i$. Then, $\alpha = \text{Probability of rejecting } H_0 \text{ in fact } H_0 \text{ is true}$. In terms of above hypothesis, $\alpha = \text{Probability of selecting a false gene as differentially expressed}$. Such a gene is called a \textit{false positive}. So, expected number of selecting false positives from a set of $p$ genes is $\alpha p$. In other words, since $p$ is very big integer, the number of false positives in the experiment is very big, even though we choose $\alpha$ very small.

This means that, probability of \textit{not} selecting a false gene as differentially expressed \hspace{1cm} = 1 − $\alpha$. In other words, probability of making the right decision for a gene $= 1 − \alpha$. Hence, the probability of making correct decision for all $p$ genes $= (\text{probability of making correct decision for gene 1}).(\text{probability of making correct decision for gene 2})\ldots\ldots(\text{probability of making correct decision for gene } p) \hspace{1cm} = (1 − \alpha)^p$. From this we see that, as $p$ increases, the probability of making correct decision decreases. So, the probability of at least one false positive somewhere is:

\[
\text{Type I Error} \quad = \quad 1 - (1 - \alpha)^p \quad \quad (2.3.2)
\]
This *Type I Error* is also called the *family-wise error rate* (FWER). There are few approaches to minimize the family-wise error rates. If the FP Rate is the error measure used, then a simple p-value threshold of $\alpha$ guarantees that the expected number of false positives, $V$, when testing all $p$ hypothesis/genes is $E(V) \leq \alpha p$.

### 1. Sidak and Bonferroni Correction

This multiple correction method is one of the earliest method introduced by Bonferroni [54]. Let $p$ be the number of tests performed for each gene. Let us consider a problem of achieving a global significance level $\alpha$. Now the question is what value of gene-wise significance level $\alpha_g$ should be specified to achieve this goal? From (2.3.2), this means that,

$$
\text{rection} \alpha = 1 - (1 - \alpha_g)^p \\
\text{or,} \quad \alpha_g = 1 - (1 - \alpha)^{\frac{1}{p}}
$$

(2.3.3)

The above equation (2.3.3) is called the Sidak Correction for multiple testing. This means that if we want to achieve the global significance level $\alpha_g$, we have to set the significance level for each gene as $1 - (1 - \alpha_g)^{\frac{1}{p}}$. Expanding the (2.3.3) by Binomial theorem and taking the first two terms we get the *Bonferroni Correction*:

$$
\alpha = 1 - (1 - \alpha)^p = 1 - (1 - p\alpha_g + ...) = p\alpha_g
$$

(2.3.4)

or,

$$
\alpha_g = \frac{\alpha}{p}
$$

(2.3.5)

This means that instead, we have to set the significance level as $\alpha$ divided by the number of genes (tests to be performed). In other words, if the error measure is FWER, then the probability that at least one false positive gene will be selected by the rule when we set $\alpha_g = \frac{\alpha}{p}$ does not exceed $\alpha$. From the above, the significance levels are the same for each gene, no matter of their $p$-values. To avoid the situation
the following method was proposed.

2. Holm’s Step-wise Correction:

While the Sidak and Bonferroni approaches [54] are effective to avoid too many false positives, but the worst that can happen is: we get none of the genes significantly expressed because we take the same significance level for all tests (genes) and the significance level $\frac{\alpha}{p}$ is too small for large $p$. So, this method is too conservative in the sense that it selects the only strong truly DE genes. Because of large number of hypotheses, $p$, there is many false positive by chance, it is more appropriate to choose the significance level for a gene according to its $p$-value. This method adjusts more to the genes that have smaller $p$-values than on larger $p$-values.

**Algorithm:**

1. Choose the global significance level $\alpha$.

2. Order the genes according to their $p$-values in ascending order.

3. Compare the $p$-value ($p_i$) of the $i$-th gene in the ordered list with the threshold $\tau_i = \frac{\alpha}{p-i+1}$.

4. Report the gene $i$ in the order as significantly expressed if $p_i < \tau_i$.

Here, what we see is that the threshold for the genes is chosen according to their $p$-values. If the $p$-value for a gene is smaller then the order $i$ is smaller. Hence $\tau_i$ is also smaller. This makes more sense than the uniform significance level.

2.4 Measures of Erroneous Rejection of Null Hypotheses

There are two measures of false decision in the multiple testing context as described in [54]:

1. Familywise Error Rate (FWER)
2. False Discovery Rate (FDR)

The FWER is defined as the probability of at least one false positive:

\[ FWER = \text{Prob}(V \geq 1). \]

The false discovery rate is defined as given by

\[ FDR = E\left[ \frac{V}{R} | R > 0 \right] \times \text{Prob}[R > 0]. \]

where, \( V = \) number of false positives; \( R = V + S, \) number of positive findings. Generally, \( FWER \geq FDR. \) Equality holds if all null hypotheses are true. But, in practice, some null hypotheses are actually false, hence FDR control is less strict than the FWER control, hence it has more power [52]. So, instead of controlling the FWER, Benjamini and Hochberg [3] proposed a method to control the FDR. The FDR is estimated by the permutation scheme. Depending on the chosen cut-off value(s) \( \alpha \) for the test statistic \( T_g \) one can estimate the FDR as follows:

- Estimate the number of non-differentially expressed genes, \( m_0. \) This can be done in the following way:
  - Calculate the \( p \)-values for each of the genes. A gene \( g \) with \( p_g > 0.5 \) is usually not differentially expressed.
  - Since the \( p \)-values of non-differentially expressed genes should be distributed uniformly on \([0, 1]\), the estimate of \( m_0, \hat{m}_0 = 2 \times \# \{ g : p_g > 0.5 \}. \)

- Compute the number of significant genes under permutations of the sample labels. The average of these numbers, multiplied with \( \hat{m}_0/N \) gives an estimate of the expected number of false positives, \( \hat{E}(V). \)

- Estimate FDR of that test, \( E(V/R) \) by \( \hat{E}(V)/R. \)
In other words, if $\alpha$ be the significant level of the test, $m_0$ be the truly non-differentially expressed genes and $R$ be the declared differentially expressed genes, then

$$FDR = \frac{\alpha \times m_0}{R}.$$ 

If all the null hypotheses were true, then the FDR is equal to the familywise error rate. However, this rarely happens in reality. In general, the more the number of hypotheses that are truly false, the smaller is the FDR. Therefore, the control of FDR tends to be more relaxed than control of FWER at the same level of significance. To determine the differentially expressed genes using the FDR procedure, one uses the following steps:

- Choose the global significance level $\alpha$
- Order the genes according to their $p$-values in the increasing order.
- Compare the $p$-values ($p_i$) of the $i$-th gene in the ordered list with threshold $\tau_i = \frac{i}{p} \alpha$.
- Report those genes $i$ as differentially expressed which satisfies $p_i < \tau_i$.

Storey and Tibshirani [43] noted that an adjustment is only necessary when there are positive findings, i.e. there are cases when the null hypotheses are rejected. They proposed the modified version of the FDR, called the positive false discovery rate (pFDR):

$$pFDR = E[\frac{V}{R} | R > 0]$$

Since the number of false positives are unknown, we have to estimate $V$ in order to estimate pFDR. Suppose we have a dataset of $N$ genes in two conditions replicated $n_1$ and $n_2$ times. Then change the class labels by permuting each gene $B$ times. Suppose that the average number $R*$ of genes have the $p$-values smaller than threshold $\alpha$ (where $\alpha$ is the genewise significant level) over the $B$ permuted data sets. Then the estimate of the pFDR is:
\[ pFDR = \frac{R*}{R} \]

Since the calculation of FDR is based on the gene scores from permutations of the data, the correlation in the genes are accounted for. Use of the permutation distribution avoids the parametric assumptions about the distribution of individual genes. Hence, the FDR criteria of selecting genes is usually better than the FWER criteria of selecting genes, because FDR takes account of correlation among the genes and it is more powerful.

### 2.5 Two-Sample Permutation Test

In the two sample \( t \)-test, one requires the following three assumptions:

- the samples are randomly selected from two populations
- the populations have normal distributions; and
- the variances of two populations are same.

If anyone of the above assumption fails, then it is likely that \( t \)-test can not give more accurate result. Furthermore, the sample size in the microarray experiment is usually too small. So, the random assignment of the observed intensities of each gene \( i \) in two different conditions provides the basis for drawing statistical inference about the effect of the treatment over the normal condition. The argument goes as follows: Consider a single gene \( i \). If there is no difference between normal and treatment conditions (\( i.e. \), it expresses similarly in both conditions), then assigning randomly all the values obtained between two conditions would have an equal chance of being observed in the study.

For example, let there are 3 values under normal condition and 2 values under treatment condition of gene \( i \). Then there are \( C(5, 2) = 10 \) combinations of different genes altogether, in each conditions. These are the types of datasets one would expect to observe if the gene \( i \) under two conditions are equally expressed. Since we have permuted the label of each gene in two conditions, we expect that the mean of two
samples are equal. This means we can use the difference of means as the test statistic. Let $D_{\text{obs}}$ be the difference between two observed sample means. Then we have 10 different means of differences $D$ of permuted samples. Then for the two-tailed test, the $p$-value for the gene $i$ is given by:

$$p_i = \frac{\text{number of } |D|'s \geq |D_{\text{obs}}|}{\text{number of permutations}}$$

With this $p$-value, reject the null hypothesis at the level of 5% significance level, i.e. if $p$-value < 0.05.

2.6 Significance Analysis of Microarrays (SAM)

Although the permutation based approach proposed by Westfall and Young (1991) [51] defines weak control of the error rate and considers the correlation among genes, it is still too stringent to their data. In their experiment, they identified either zero or 300 significant genes depending on how the $p$-value is corrected. To address the above challenges, Tusher et al.[48] proposed the Significance Analysis of Microarray (SAM). Basically, SAM assigns a score for each gene according to the change in gene expression. Genes with greater than the threshold is considered to be potentially significant. To control false positives, SAM uses the permutation of measurements to estimate the pFDR [48]. The score threshold for genes is then adjusted iteratively according to the pFDR until a set of significant genes have been identified. To each gene $i$, SAM assigns a score

$$d(i) = \frac{\bar{x}_1(i) - \bar{x}_2(i)}{s(i) + s_0},$$

(2.6.6)

where $\bar{x}_1(i)$ and $\bar{x}_2(i)$ are the average level of expressions for gene $i$ in the classes 1 and 2 respectively. If there are $n_1$ and $n_2$ replications of gene $i$ in two classes,

$$s(i) = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Compared to $t$-statistic, the SAM’s score adds a "fudge" term, $s_0$, in the denom-
The reason behind it is, the variance \( s(i) \) tends to be smaller at the lower expression levels. This makes the score \( d(i) \) dependent in \( s(i) \). To compare the values of \( d(i) \) across all genes, the distribution of \( d(i) \) should be independent of expression levels and hence independent of \( s(i) \). To address this problem, SAM seeks to find the fudge factor \( s_0 \) such that the dependency of \( d(i) \) with \( s(i) \) is as small as possible. Such an \( s_0 \) is obtained by certain percentile of \( s_i \) values depending on the data. The FDR in the SAM procedure is obtained by the permutation method, which is different than the Benjamini and Hochberg procedure. It is given by

\[
pFDR = \frac{1}{B} \sum_{b=1}^{B} \frac{C_b}{C}
\]

where \( B \) is the number of permutations of samples, \( C_b \) is the number of potentially significant genes in the \( b \)-th permutated data set, and \( C \) is the # of potentially significant genes in the original data set.

### 2.7 Sampling Distribution for Non-homogeneous Variance

### 2.8 Bayesian Approach

In the following, we are going to derive the marginal posterior distribution of the difference of two population means from two samples with different variances. We have taken the priors as in [1], and we have compared the result with the existing method [13]. Since there are few samples in the microarray data, all the information required to estimate the parameters are not sufficient. So, as in Baldi and Long [1] and [13], the Bayesian approach has been proposed to estimate the distribution and estimate the parameters.

For the samples \( \mathbf{x} = (x_i) \overset{iid}{\sim} N(\mu, \sigma^2) \) and \( \mathbf{y} = (y_j) \overset{iid}{\sim} N(\mu + \Delta \mu, \tau^2) \), where, \( i = 1, 2, ..., m; \) and \( j = 1, 2, ..., n. \)
The density for $x = (x_1, x_2, \ldots, x_m)$ can be written as,
\[
f(x) = \frac{1}{\sigma^m(2\pi)^{m/2}} \exp \left[ -\frac{1}{2\sigma^2} \left\{ (m - 1)s_x^2 + m(x - \mu)^2 \right\} \right] \tag{2.8.7}
\]

Similarly, the density of $y$ is,
\[
f(y) = \frac{1}{\tau^n(2\pi)^{n/2}} \exp \left[ -\frac{1}{2\tau^2} \left\{ (n - 1)s_y^2 + n(y - \mu - \Delta \mu)^2 \right\} \right] \tag{2.8.8}
\]

The frequentist approach requires a large number of observations before making the probabilistic judgements about the outcomes. In the microarray data, the sample size is small thus it underestimates the population variance and gives unreliable estimate of the population variance. Classical statistics are directed towards the use of sample information in making inferences. But in Bayesian analysis, one combines the information obtained from sample and other relevant aspects of the problem in order to make the best decision. So, we use the bayesian analysis for estimating the variances. For microarray data the conjugate prior seems to be more suitable and flexible because of its convenient form. Assuming the independency of the location parameter $\mu$ and scale parameter $\sigma^2$ as in [13], the joint prior for $\mu$ and $\sigma^2$ is the product
\[
p(\mu, \sigma^2) = p(\mu)p(\sigma^2) \tag{2.8.9}
\]

Similarly, the joint prior for $\mu + \Delta \mu$ and $\tau^2$ is
\[
p(\mu + \Delta \mu, \tau^2) = p(\mu + \Delta \mu)p(\tau^2) \tag{2.8.10}
\]

Finally, the joint prior for $\mu, \sigma^2, \mu + \Delta \mu, \tau^2$ is
\[
\text{prior} = p(\mu, \sigma^2, \mu + \Delta \mu, \tau^2) = p(\mu)p(\sigma^2)p(\mu + \Delta \mu)p(\tau^2) \tag{2.8.11}
\]

Since $(\bar{x}, s_x^2)$ and $(\bar{y}, s_y^2)$ are sufficient statistics for $(\mu, \sigma^2)$ and $(\mu+\Delta \mu, \tau^2)$ respectively,
we have the joint posterior distribution is given by the Bayes’ Rule [15],

\[ p(\mu, \Delta \mu, \sigma^2, \tau^2 | x, y) \propto (\text{prior}) \frac{1}{\sigma^{m+\nu_0}} \exp\left(-\frac{C_\mu}{2\sigma^2}\right) \exp\left(-\frac{D_{\mu+\Delta \mu}}{2\tau^2}\right) \]  

(2.8.12)

where,

\[ C_\mu = (m - 1)s^2_x + m(\bar{x} - \mu)^2; \]

\[ D_{\mu+\Delta \mu} = (n - 1)s^2_y + n(\bar{y} - (\mu + \Delta \mu))^2 \]

We obtain the marginal posterior density of \((\Delta \mu, \sigma^2, \tau^2)\) by integrating (2.8.12) with respect to \(\mu\). The marginal posterior is

\[ p(\Delta \mu, \sigma^2, \tau^2 | x, y) \propto \int_{-\infty}^{\infty} (\text{prior}) \left[ \frac{1}{\sigma^{\nu_0+\tau_0}} \exp\left(-\frac{C_\mu}{2\sigma^2} - \frac{D_{\mu+\Delta \mu}}{2\tau^2}\right) \right] d\mu \]  

(2.8.13)

Let us assume that the priors for \(\mu\) and \(\mu + \Delta \mu\) are flat priors, i.e. \(p(\mu) = 1\), \(p(\mu + \Delta \mu) = 1\) and the priors for \(\sigma^2\) and \(\tau^2\) are scaled inverse-\(\chi^2\) distributions as in [13]. The pdf of inverse chi-square distribution with \(\nu = \alpha\) and scale = \(\beta^2\) is given by

\[ f(x; \alpha, \beta^2) = \left(\frac{\beta^2}{\Gamma\left(\frac{\alpha}{2}\right)}\right) x^{-(\frac{\alpha}{2}+1)} \exp\left(-\frac{\alpha \beta^2}{2x}\right) \]

With these priors the posterior is of the same form [1]. So, they are the conjugate priors for the normal likelihoods. i.e. \(p(\sigma^2) = I(\sigma^2; \nu_0, \sigma_0^2)\) and \(p(\tau^2) = I(\tau^2; \eta_0, \tau_0^2)\)

where,

\[ p(\sigma^2) \propto \sigma^{-(\nu_0+2)} \exp\left(-\frac{1}{2\sigma^2} \nu_0 \sigma_0^2\right) \]  

(2.8.14)

\[ p(\tau^2) \propto \tau^{-(\eta_0+2)} \exp\left(-\frac{1}{2\tau^2} \eta_0 \tau_0^2\right) \]  

(2.8.15)

where, \(\alpha = (\nu_0, \eta_0, \sigma_0^2, \tau_0^2)\) is the hyper-parameters that should be estimated from the data.

Hence from equations (2.8.13), (2.8.14) and (2.8.15), the marginal posterior density of \((\Delta \mu, \sigma^2, \tau^2)\) is

\[ p(\Delta \mu, \sigma^2, \tau^2 | x, y) \propto \frac{1}{\sigma^{m+\nu_0+2}} \cdot \frac{1}{\tau^{n+\eta_0+2}} \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{C_\mu + \nu_0 \sigma_0^2}{2\sigma^2}\right) \exp\left(-\frac{D_{\mu+\Delta \mu} + \eta_0 \tau_0^2}{2\tau^2}\right) d\mu \]
\[ \begin{align*}
&= \int_{-\infty}^{\infty} \left[ \frac{1}{\sigma_{m+\nu_0+2}} \exp\left( - \frac{C_\mu + \nu_0 \sigma_0^2}{2\sigma^2} \right) \right] \left[ \frac{1}{\tau_{n+\eta_0+2}} \exp\left( - \frac{D_{\mu+\Delta\mu} + \eta_0 \tau_0^2}{2\tau^2} \right) \right] d\mu \\
\text{The marginal posterior of } \Delta\mu \text{ is obtained by}
&\quad \text{p}(\Delta\mu|x, y) \propto \int_{\mu=-\infty}^{\infty} \left[ \int_{0}^{\infty} \frac{1}{\sigma_{m+\nu_0+2}} \exp\left( - \frac{C_\mu + \nu_0 \sigma_0^2}{2\sigma^2} \right) d\sigma \right] d\mu \\
&\quad \left[ \int_{0}^{\infty} \frac{1}{\tau_{n+\eta_0+2}} \exp\left( - \frac{D_{\mu+\Delta\mu} + \eta_0 \tau_0^2}{2\tau^2} \right) d\tau \right] d\mu \\
&\quad = \int_{\mu=-\infty}^{\infty} I_1. I_2 \ d\mu \quad (2.8.16)
\end{align*} \]

Now,
\[ I_1 = \int_{0}^{\infty} \frac{1}{\sigma_{m+\nu_0+2}} \exp\left( - \frac{A}{2\sigma^2} \right) d\sigma^2, \]

where,
\[ A = C_\mu + \nu_0 \sigma_0^2 \]

Changing the variable \( u = \frac{A}{\sigma^2} \) and after some computation, we get
\[ I_1 = \int_{0}^{\infty} (\frac{A}{2u})^{-(m+\nu_0+2)} e^{-u(\frac{A}{2\nu_0})} du \]
\[ = A \left( \frac{m+\nu_0}{2} \right) \int_{0}^{\infty} e^{-u} u^{-(m+\nu_0+2)\frac{1}{2}} du \]
\[ = A \left( \frac{m+\nu_0}{2} \right) \int_{0}^{\infty} e^{-u} u^{-(m+\nu_0+1)} du \]

This being non-normalized gamma integral, the above integral is,
\[ I_1 \propto A^{-\frac{(m+\nu_0)}{2}} \]
\[ = (C_\mu + \nu_0 \sigma_0^2)^{-\frac{(m+\nu_0)}{2}} \]
\[ = [(m-1)s_x^2 + m(\bar{x} - \mu)^2 + \nu_0 \sigma_0^2]^{-\frac{(m+\nu_0)}{2}} \]
\[ I_1 \propto \left[ 1 + \frac{m(\bar{x} - \mu)^2}{v_m\sigma_m^2} \right]^{-\frac{(m+\nu_0)}{2}} \]

i.e.

\[ I_1 \propto \left[ 1 + \frac{m(\bar{x} - \mu)^2}{v_m\sigma_m^2} \right]^{-\frac{(m+\nu_0)}{2}} \]

(2.8.17)

where, \( v_m = m + \nu_0 - 1 \), \( v_m\sigma_m^2 = (m - 1)s^2_x + \nu_0\sigma_0^2 \).

Similarly, we get the another factor in (2.8.16) as

\[ I_2 \propto \left[ 1 + \frac{n(\bar{y} - \mu - \Delta\mu)^2}{w_n\tau_n^2} \right]^{-\frac{(n+\eta_0)}{2}} \]

(2.8.18)

where, \( w_n = n + \eta_0 - 1 \), \( w_n\tau_n^2 = (n - 1)s^2_y + \eta_0\tau_0^2 \).

Substituting (2.8.17) and (2.8.18) in (2.8.16), we get

\[
p(\Delta\mu|\mathbf{x}, \mathbf{y}) = k \int_{\mu=-\infty}^{\infty} \left[ 1 + \frac{m(\bar{x} - \mu)^2}{v_m\sigma_m^2} \right]^{-\frac{1}{2}(m+\nu_0)} \left[ 1 + \frac{n(\bar{y} - (\Delta\mu + \mu))^2}{w_n\tau_n^2} \right]^{-\frac{1}{2}(n+\eta_0)} d\mu
\]

(2.8.19)

This is the pdf of the Behrens-Fisher distribution, where \( k \) is given by,

\[
k = \left[ \text{Beta} \left( \frac{v_m}{2}, \frac{1}{2} \right) \text{Beta} \left( \frac{w_n}{2}, \frac{1}{2} \right) \sqrt{v_m w_n} \right]^{-1}
\]

Now, we can apply the Behrens-Fisher distribution for testing the hypothesis regarding the two population means, using two samples drawn from the population with different means and different variances.

2.9 Test Statistic

Let us define the statistic, called the BF-statistic as in [5]:
\[ B = \frac{\Delta \mu - (\bar{y} - \bar{x})}{(\frac{\sigma_m^2}{m} + \frac{\tau^2}{n})^{\frac{1}{2}}} \]
\[ = \frac{(\mu + \Delta \mu) - \bar{y}}{\tau_n/\sqrt{n}} \cos \theta - \frac{(\mu - \bar{x})}{\sigma_m/\sqrt{m}} \sin \theta \]
\[ = B_y \cos \theta - B_x \sin \theta \]

where,
\[ \tan \theta = \frac{\sigma_m/\sqrt{m}}{\tau_n/\sqrt{n}}, \quad 0 \leq \theta \leq \frac{\pi}{2} \]

and, \( B_x \) and \( B_y \) are independently distributed as \( t \)-statistics, \( t(v_m) \) and \( t(w_n) \) respectively. Hence, under the sampling distribution, \( p(x, y|\mu, \sigma^2, \tau^2) \), the statistic \( B \) is distributed as the Behrens-Fisher distribution with \( v_m \) and \( w_n \) degrees of freedom. That is,
\[ B \sim BF(v_m, w_n, \theta) \]

with pdf
\[ f(\beta|\mu, \sigma^2, \tau^2) = k \int_{-\infty}^{\infty} \left[ 1 + \frac{(\alpha \cos \theta - \beta \sin \theta)^2}{v_m} \right]^{-\frac{v_m+1}{2}} \left[ 1 + \frac{(\alpha \sin \theta + \beta \cos \theta)^2}{w_n} \right]^{-\frac{w_n+1}{2}} d\alpha, \]
\[ (2.9.20) \]

where,
\[ \alpha = B_y \sin \theta + B_x \cos \theta, \quad \beta = B_y \cos \theta - B_x \sin \theta \]

which is same as (2.8.19). Hence, we have proved the following:

**Theorem:** Let \( x \) and \( y \) be two independent samples with sample sizes \( m \) and \( n \) respectively from normal distributions with different means \( \mu \) and \( (\mu + \Delta \mu) \) and variances \( \sigma^2 \) and \( \tau^2 \). If the priors \( \nu_0 \) and \( \eta_0 \) for means are flat priors and priors for
variances $\sigma_0^2$ and $\tau_0^2$ are scaled inverse $\chi^2$-distributions, the posterior distribution of $\Delta_\mu$ is the Behrens-Fisher distribution with $v_m = m + \nu_0 - 1$, $w_n = n + \eta_0 - 1$ degrees of freedom.

Due to the complexity of the pdf of the BF-distribution as given in (2.9.20), it is very hard to compute the corresponding probabilities, especially due to the possibility of the fractional degrees of freedom. In addition, there are no uniformly most powerful unbiased tests for all sample sizes for the BF-problem [4]. Because of this, there are various types of approximations available in the literature [31],[4]. Due to the simplicity of application as well as availability of R-code (www.r-project.org) for computing $t$-values even for fractional degrees of freedom, we use Patil’s approximation [31] in this work as follows.

Let

$$f_1 = \left(\frac{w_n}{w_n - 2}\right) \cos^2 \theta + \left(\frac{v_m}{v_m - 2}\right) \sin^2 \theta$$

$$f_2 = \frac{w_n^2}{(w_n - 2)^2(w_n - 4)} \cos^4 \theta + \frac{v_m^2}{(v_m - 2)^2(v_m - 4)} \sin^4 \theta$$

$$a^2 = \frac{(b-2)}{b} f_1$$

$$b = 4 + \frac{f_2}{f_1}$$

$$\cos^2 \theta = \frac{\frac{b^2}{a^2} + \frac{\tau_0^2}{\sigma_0^2}}{\left(\frac{b^2}{a^2} + \frac{\tau_0^2}{\sigma_0^2}\right)}$$

$$\sin^2 \theta = 1 - \cos^2 \theta.$$

Then, the statistic

$$\frac{B}{a} \sim t(b).$$

That is, $B$ has approximately $t$-distribution with $b$ degrees of freedom ($b \geq 1$, $b$ may not be an integer [5] ) and scale parameter $a$. This statistic $B$ can also be denoted as $B \sim t(0, a^2, b)$. It was noted [31] that the formula (2.9.21) is valid only for $v_m, w_n \geq 5$ and works quite well for $v_m, w_n \geq 7$.
The corresponding Bayesian test statistic [13] for the equal variance case is

\[ E = \frac{\left( \Delta \mu - (\bar{y} - \bar{x}) \right)}{\nu \sqrt{\frac{1}{m} + \frac{1}{n}}} \]  

(2.9.22)

which is distributed as \( t \) distribution with \( \delta \) degrees of freedom, where,

\[
\begin{aligned}
\delta &= m + n + \nu_0 - 2 \\
\delta \nu^2 &= \nu_0 \sigma_0^2 + (m - 1)s_x^2 + (n - 1)s_y^2
\end{aligned}
\]  

(2.9.23)

2.10 Calculation of Prior d.f. and Prior Variance

The observed dependence between population mean (\( \mu \)) and population variance (\( \sigma^2 \)) is established by calculating the prior variance (\( \sigma_0^2 \)) based on the variance of similarly expressed genes. However, in this formulation, instead of taking the average standard deviation of similarly expressed genes we choose to estimate the prior variance by totaling the sum of squared difference for each similar genes and dividing by the total prior degrees of freedom. This is more statistically rigorous way of incorporating prior information and leads to more consistent test. There are many possible ways one can choose \( p \) and \( q \) (and so prior variances and prior means) in the above expressions. In each of the methods, we use the sample variances of only particular genes in both control and treatment conditions, and apply equations (2.11.30),(2.11.31),(2.11.32) and (3.4). In this work we choose \( p \) and \( q \) in the following manners and compare the effect of each of these in terms of FDR.

2.11 Estimation of Hyperparameters

There are two different methods to estimate the hyperparameters: Empirical Bayes (Robbins) method and Evidence Method (MacKay 1992). In the evidence method the evidence, which is the integration of the posterior distribution of data, is maximized with respect to the hyperparameter of interest. The reason for this is similar to
maximising the posterior distribution under the squared error loss function \( L(\theta, a) = (\theta - a)^2 \) while estimating the parameter \( \theta \) by \( a \). The evidence method has been used to estimate the parameter. We take the estimate of degrees of freedom from \( m \) control samples as \( \hat{\nu}_0 = m - 1 \), and from \( n \) treatment samples as \( \hat{\eta}_0 = n - 1 \). The evidence is given by the integral

\[
P(x|\nu_0, \sigma_0^2) = \int_{\mu=-\infty}^{\infty} \int_{\sigma^2=0}^{\infty} \frac{1}{\sigma^{m+\nu_0-2}} \exp\left\{-\frac{1}{2\sigma^2} \left[(m-1)s_x^2 + m(\bar{x} - \mu)^2 + \nu_0\sigma_0^2\right]\right\} d\sigma^2 d\mu
\]

Writing

\[
\sigma^2 = u, \quad \frac{m + \nu_0 - 2}{2} = a, \quad (m-1)s_x^2 + \nu_0\sigma_0^2 = b \quad \text{and} \quad \mu = v,
\]

the integral (2.11.24) reduces to

\[
P(x|\nu_0, \sigma_0^2) = \int_{v=-\infty}^{\infty} \int_{u=0}^{\infty} \frac{1}{u^a} \exp\left\{-\frac{1}{2u} \left[b + m(\bar{x} - v)^2\right]\right\} du dv
\]

\[
= \int_{u=0}^{\infty} \frac{1}{u^a} \left[ \int_{v=-\infty}^{\infty} \exp\left\{-\frac{1}{2u} \left[b + m(\bar{x} - v)^2\right]\right\} dv \right] du
\]

\[
= \int_{u=0}^{\infty} \frac{1}{u^a} \exp\left(-\frac{b}{2u}\right) \left[ \int_{v=-\infty}^{\infty} \exp\left\{-\frac{1}{2u} \left[m(\bar{x} - v)^2\right]\right\} dv \right] du
\]

Integrating the interior integral and simplifying, we get the above integral as

\[
P(x|\nu_0, \sigma_0^2) = \int_{u=0}^{\infty} \frac{1}{u^a} \exp\left(-\frac{b}{2u}\right) \left[ \sqrt{\frac{4\pi u}{m}} \right] du
\]

\[
= \sqrt{\frac{4\pi}{m}} \int_{u=0}^{\infty} \frac{1}{u^{a-1/2}} \exp\left(-\frac{b}{2u}\right)
\]

(2.11.25)

Substituting \( \frac{b}{2u} = z \) in the above integral and changing the limit, we get

\[
P(x|\nu_0, \sigma_0^2) = \sqrt{\frac{4\pi}{m}} \left(\frac{b}{2}\right)^{\frac{1-a}{2}} \int_{z=0}^{\infty} z^{a-\frac{3}{2}} \exp(-z) dz
\]

\[
= \sqrt{\frac{4\pi}{m}} \left(\frac{b}{2}\right)^{\frac{1-a}{2}} K(a)
\]

(2.11.26)
where, the integral above is unnormalized gamma, hence a function of $a$, $K(a) > 0$.

After replacing the value of $a$ and $b$, we get

$$P(x|\nu_0, \sigma^2_0) = \sqrt{\frac{4\pi}{m}} \left( \frac{(m - 1)s^2_x + \nu_0\sigma^2_0}{2} \right)^{\frac{1}{2}} K\left(\frac{m + \nu_0 - 2}{2}\right)$$  \hspace{1cm} (2.11.27)

Differentiating with respect to $\sigma^2_0$ and equating it to 0, we get

$$\frac{\partial P}{\partial \sigma^2_0} = 0$$

$$\sigma^2_0 = \frac{(m - 1)s^2_x}{\nu_0}$$

$$\hat{\sigma}^2_0 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{\nu_0}$$  \hspace{1cm} (2.11.28)

Similarly, the estimate of $\tau^2_0$ is obtained as

$$\hat{\tau}^2_0 = \frac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{\eta_0}$$  \hspace{1cm} (2.11.29)

For estimating the number of similarly expressed genes, we have the following three methods:

**Method 1: Window Method**

The window method was introduced by Baldi et al.[1]. In this method, we calculate the prior degrees of freedom and prior variances by taking those genes that are similar in variances to that of the gene of interest both in control and treatment condition within pre-chosen window size. To calculate the prior variance of a gene $g$ in control condition, we calculate its variance in control condition. Then we calculate the variance of all other genes and take only those $p$ genes whose variances are close to that of gene $g$. Now the prior variance for gene $g$ is the mean of variances of these $p$ genes in control condition. Similarly the prior variance for gene $g$ in treatment condition is calculated by taking $q$ genes with similar variances to that of gene $g$.

For each gene $g$, the prior variances and prior means will be different. We have
dropped the subscript $g$ from them. For $p$ genes with similar variances and each having $m$ replicates in control condition, the prior degree of freedom for the variance can be calculated as,

$$\hat{\nu}_0 = p(m - 1)$$

(2.11.30)

Similarly, for $q$ genes with similar variances and each having $n$ replicates in treatment condition, the prior degrees of freedom for the variance is given by

$$\hat{\eta}_0 = q(n - 1)$$

(2.11.31)

The prior variances for control and treatment conditions are calculated as sample variance of similar genes, which are the means of variances of similar genes in these two conditions respectively.

$$\hat{\sigma}_0^2 = \frac{1}{\nu_0} \sum_{k=1}^{p} \sum_{i=1}^{m} (x_{k,i} - \bar{x}_k)^2$$

(2.11.32)

$$\hat{\tau}_0^2 = \frac{1}{\eta_0} \sum_{k=1}^{q} \sum_{j=1}^{n} (y_{k,j} - \bar{y}_k)^2$$

(2.11.33)

where,

- $\bar{x}_k$ is the mean response of gene $k$ in the control condition
- $\bar{y}_k$ is the mean response of gene $k$ in the treatment condition
- $x_{k,i}$ is the response $i$ of gene $k$ in the control condition
- $y_{k,i}$ is the response $i$ of gene $k$ in the treatment condition

**Method 2: Similar Variance Method**

In the following, we present the a method of estimating hyperparameter.

Another method is to choose the genes that have the absolute difference of variances in both control and treatment group using some cut-off values, so that we will have certain level of confidence choosing them. For a gene $g$, let $s_{cg}^2$ and $s_{tg}^2$ be the sample variances in control and treatment conditions respectively. Then choose $p_g=$
number of genes $j$ such that $|s_{cg}^2 - s_{cj}^2| \leq k$. We choose $k$ by the 95% CI for the population variance, which uses the $\chi^2$ distribution to find the confidence interval. So, $k = |\min\{s_{cg}^2 - 95\% \text{ Lower } \chi^2 \text{ CL of } s_{cg}^2, s_{cg}^2 - 95\% \text{ Upper } \chi^2 \text{ CL of } s_{cg}^2\}|$ in the control condition. Similarly, choose $q_g = \text{number of genes } j \text{ such that } |s_{tg}^2 - s_{tj}^2| \leq k$, where $k = |\min\{s_{ct}^2 - 95\% \text{ Lower } \chi^2 \text{ CL of } s_{ct}^2, s_{ct}^2 - 95\% \text{ Upper } \chi^2 \text{ CL of } s_{ct}^2\}|$ in the treatment condition. So, in this method the values of $p$ and $q$ varies according to each gene. By choosing the number $k$ in such a way we are 95% confident that the genes used for the calculation of prior variance and the degrees of freedom have similar variance to the gene $g$. For simplicity, we omit the subscript $g$ in the notation.

This method was proposed by us.

**Method 3: Resampling Method**

To estimate the precision of sample statistics by using the subsets of available data by drawing randomly with replacement resampling method is used Good [?]. In the case of estimating variances, instead of taking those genes which have similar variances within a window or as in similar variance method, we have used this method to calculate prior variances for the actual data. Because, in the window method the researcher should consider all possible window size and the determination of it may be time consuming. On the other hand, the similar variance method selects huge number of genes and so the computation time is exponentially increased. We take the $p = q = 1\%$ bootstrap samples from the pool of all genes together with the gene of interest from the control and treatment conditions respectively to determine the prior d.f. and variance for a gene. Then for $p$ genes each having $m$ replicates in the control condition, the prior degrees of freedom is given by $\nu_0 = p(m - 1)$ and the prior variance $\sigma_0^2$ is estimated by the variance of the $p$ bootstrap samples in the control condition. The prior degrees of freedom in the treatment condition $\eta_0 = q(n - 1)$ and variance $\tau_0^2$ are calculated similarly from the treatment condition.

The algorithm for selecting the differentially expressed genes consist of the follow-
ing steps:

- Use the method of Dudoit et al [10] to pre-process the data, filter the genes, and normalize the data.

- Take the control group of the training sample.

- Calculate the variances of each gene.

- Set the bounds for error: For each gene $g$

  \[
  \text{lower bound} = \text{Variance of } g - \text{Error} \\
  \text{upper bound} = \text{Variance of } g + \text{Error} \\
  \text{Error} = \frac{1}{2}(V_{\text{max}} - V_{\text{min}})
  \]

  where, $V_{\text{max}} = \max_g \text{Variances}$ and $V_{\text{min}} = \min_g \text{Variances}$

- Choose those variances that are in between lower and upper bound and take a mean of those variances to estimate the prior variances. This gives the variances of each gene as the mean of those variances which satisfy the similar variance criteria in the training samples.

- Repeat the above 5 procedures in the control group of test sample.

2.12 Simulation and Result

To implement the theory, we have simulated 10,000 genes from normal distributions having mean $\mu_1$ variance $\sigma_1^2$ in the normal conditions and mean $\mu_2$ and variance $\sigma_2^2$ in the treatment conditions that was found as in Golub data. Each gene was replicated 10 times in each of the control and treatment conditions. Without loss of generality, we have set the first 2 percent of the genes as differentially expressed. The assumptions that we made to simulate the data were they are independent and normally distributed.
1. Out of 10,000 genes, simulate 98,00 genes from a normal distribution with mean \( \mu \sim U[0, 1] \) and variance \( \sigma^2 \sim Inv - \chi^2(df = 20, scale = 2) \).

2. Without loss of generality, simulate first 100 genes from normal distribution with mean \( \mu \sim U[5, 10] + \frac{1}{2} \) and variance \( \sigma^2 \sim Inv - \chi^2(df = 20, scale = 4) \) in the control condition and mean \( \mu \sim U[5, 10] - \frac{1}{2} \) and variance \( \sigma^2 \sim Inv - \chi^2(df = 18, scale = 6) \) in the control condition.

3. Simulate another 100 genes from normal distribution with mean \( \mu \sim U[5, 10] - \frac{1}{2} \) and variance \( \sigma^2 \sim Inv - \chi^2(df = 18, scale = 6) \) in the control condition and mean \( \mu \sim U[5, 10] + \frac{1}{2} \) and variance \( \sigma^2 \sim Inv - \chi^2(df = 18, scale = 4) \) in the control condition.

In the simplest case, when we choose \( \nu_0 = \eta_0 = 0 \), then the BF statistic reduces to the two sample t-test. The parameters \( \nu_0 \) and \( \eta_0 \) represents the degree of confidence in the background variances \( \sigma_0^2 \) and \( \tau_0^2 \) versus the empirical variances of control and treatment respectively. We have chosen the values of these two variances in the wide range beginning from 0. The values of \( \nu_0 \) and \( \eta_0 \) are increased. In each of the calculation of variances of similar genes in the two conditions, we have taken the window size equal to the the replicates in each condition, i.e., taking \( p/2 \) genes in control group immediately above and below the gene under considerations. The mean is taken of those ordered variances corresponding to the gene of interest. Since the Behrens-Fisher distribution does not give the good result unless \( v_m \) and \( w_n \) is greater than 5 and gives very good result when it exceeds 7 [31], we have chosen the values of the prior degrees of freedom such that \( v_m \) and \( w_n \) exceeds 7. We have chosen those genes as differentially expressed whose p-values are smaller than ranked values of the statistic given by FDR criterion. Furthermore we have chosen the level of significance \( \alpha = 0.05 \). The number of genes that are found differentially expressed in our method and in the equal variance cases are compared in Table 2.2. We have run the data 5 times and averaged the number of genes selected as differentially expressed in each time. In analyzing the fold change method with the BF method, we have found that all the genes associated with the large - fold change are not necessarily statistically
significant in the Bayesian BF method.

2.13 Power Comparisons

Here, we have compared the power obtained by our method with that of equal variance test of Fox-Dimmic [13]. The application of treatment not only affects the averages but it affects the expression measurements, so it affects variances as well. It is a common experience in statistical analysis that as the number of samples increases, the power of the test also increases. But in our case, the power depends on the priors as well. Even for the small values of priors, the power of the proposed method seems optimum than the equal variance method.

The power of a test is defined as the probability of rejecting the false null hypothesis. If $\beta$ is the probability that a true alternative hypothesis is falsely rejected, then the power of that test is given by $1 - \beta$. The power depends on the sample size, the sample standard deviation and the difference of the null and alternative hypothesized mean. We have chosen the standardized effect, $\frac{\mu_1 - \mu_2}{\text{sd.error}}$ as 2 for power calculation.

We see from the graph of power in Fig. 2.1 and power comparison in Table 2.2 that, at small values of $\nu_0$ the power obtained by the equal variance test is preferred than our method. But, eventually, our proposed method seems to have better result because we have considered the unequal variances in the two samples. In small values of $m$ and $\nu_0$, the corresponding values for $v_m = m + \nu_0 - 1$ are smaller and it does not give appropriate values unless it exceeds 7. Similarly, for the small values of $w_n$. So, we have to choose the values of $\nu_0$ or $\eta_0$ and $m$ or $n$ such that the values of $v_m$ and $w_n$ is at least 5. Holding $\nu_0 = \eta_0 = 4$ we see the difference of power in EV method and the new method increases significantly as the sample size increase from 2 to 5. The difference in the power between this proposed method and the EV method is less pronounced for the small sample sizes and small values of the prior degrees of freedom. But it is more distinct as the prior degree of freedom increases. We have chosen the prior variances $\sigma_0^2=0.7$ and $\tau_0^2=0.3$. 

41
Figure 2.1: Power comparison for different sample size and priors using Method 1
No. of samples per condition = 3

Figure 2.2: Power of the Proposed Method at m=n=3
No. of samples per condition = 5

![Graph showing the power of the Proposed Method at m=n=5](image)

Figure 2.3: Power of the Proposed Method at m=n=5
Figure 2.4: Graph of DE genes
The Figures 2.2 and 2.3 show the power comparison of the proposed BF method and the method used by the Fox and Dimmic in the case of samples of sizes 3 and 5. In this small sample case, the power of the proposed BF method is found to be larger than the Fox-Dimmic method. The Fig. 2.4 is the graph of the proposed Behrens-Fisher test. The line in this graph is the cut-off for controlling the false discovery rate of 0.05. The points below the line is the genes that are differentially expressed.

In this simulated study, we have used the proportion of genes that are actually differentially expressed and detected by our method as the criterion for comparing the proposed method. As this method is not suitable if there is small values of \( m \) or \( n \) and \( p \) or \( q \) that makes \( v_m \) and \( w_n \) small, which is seen clearly from the table (Table 2) that for \( m = n = 2 \) and choosing any values of the priors \( \nu_0 \) and \( \tau_0 \) only selects 10% of the genes that are actually expressed, although it reports large amount of genes as differentially expressed. We have found that the genes marked as differentially expressed by the fold-change (fold change = 2) method is almost 32% of the total number of genes included in the simulated study. This result is highly unacceptable as we have assumed only 2% of the genes as differentially expressed. On the another extreme, we ran two-sample \( t \)-test and found that it selected few genes as differentially expressed as the sample sizes was increased to 4 and the window size was 20. After that it selected many genes and the result was getting better when we took sample size of 10. In this case, it selected almost the same genes as our method. So, this method failed for the small sample size, because it could not select the genes that are actually differentially expressed.

2.14 Comparison of DE Genes using the Window Method with Other Methods

Table 2.3 compares the different methods with the proposed method. All three tests, except fold-change method, gave almost the same conclusion as our proposed method. Initially, when the sample size and window size both are small, \( m = 2 \) and \( p = 2 \),
then the method of equal variance test seemed better in terms of proportion of actual genes selected. This is natural because our method does not give the better result for small values of $v_m$ or $w_n$. However, as we have greater values of sample size and window size, our method excels the equal variance Bayesian test counterpart. We have seen from the result that almost all of the actually DE genes were selected by this proposed method when $m = n = 10$ and $p = q = 8$. But, the proportion of genes selected compared to the actual set of DE genes is just 75.7%.

The maximum proportion of genes selected is 88.7% when $m = n = 10$ and $p = q = 2$. In this case all of the genes actually DE are selected. Our method and equal variance test selected almost the same number of genes on average. Unless $p = q = 8$ and for $m = n = 2$, both the equal variance and unequal variance method selected enormous number of genes, most of them being bogus (false positive). The proportion of true genes is very low, about 11%. When $m = n = 3$ and $p = q = 2$, our method selected 45 genes of them 20 are actually DE, but equal variance method selected about 134 genes, of them only 58 are actually DE. Hence the proportion of true genes are 0.44 and 0.42 respectively. Taking the sample sizes constant, we found

<table>
<thead>
<tr>
<th>$m = n$</th>
<th>$v_0 = \eta_0$</th>
<th>Power BF</th>
<th>Power EV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>16</td>
<td>0.5617153</td>
<td>0.4532222</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>0.7179875</td>
<td>0.6035196</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>0.820609</td>
<td>0.717347</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>0.8865956</td>
<td>0.8020212</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0.4969753</td>
<td>0.4067113</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0.6431226</td>
<td>0.5496432</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.7496756</td>
<td>0.6657278</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0.8262618</td>
<td>0.7566512</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.3985896</td>
<td>0.3444951</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.5350779</td>
<td>0.4867308</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0.6495701</td>
<td>0.6098946</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.8395813</td>
<td>0.7105832</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.3057716</td>
<td>0.2803869</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.4221825</td>
<td>0.429857</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.545814</td>
<td>0.5631983</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.6547838</td>
<td>0.6739527</td>
</tr>
</tbody>
</table>

Table 2.2: Table of Power Comparison
Table 2.3: Comparison of BF test with other tests based on the proportion of actually DE genes selected in Window Method.

<table>
<thead>
<tr>
<th>Replications</th>
<th>DE genes</th>
<th>Common in actual</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>m = n, p = q</td>
<td>FC</td>
<td>t-test</td>
<td>EV</td>
</tr>
<tr>
<td>2, 2</td>
<td>3421</td>
<td>1</td>
<td>157.5</td>
</tr>
<tr>
<td>2, 4</td>
<td>3430.5</td>
<td>0.5</td>
<td>544.5</td>
</tr>
<tr>
<td>2, 8</td>
<td>3439.67</td>
<td>0</td>
<td>1145.67</td>
</tr>
<tr>
<td>2, 20</td>
<td>3459</td>
<td>0</td>
<td>1093</td>
</tr>
<tr>
<td>2, 100</td>
<td>3374.5</td>
<td>0</td>
<td>2126.5</td>
</tr>
<tr>
<td>3, 2</td>
<td>3347</td>
<td>0.33</td>
<td>134.33</td>
</tr>
<tr>
<td>3, 4</td>
<td>3392</td>
<td>0.33</td>
<td>344</td>
</tr>
<tr>
<td>3, 8</td>
<td>3314.33</td>
<td>0</td>
<td>569.33</td>
</tr>
<tr>
<td>3, 20</td>
<td>3371</td>
<td>0</td>
<td>832</td>
</tr>
<tr>
<td>3, 100</td>
<td>3320.5</td>
<td>0.5</td>
<td>1032</td>
</tr>
<tr>
<td>4, 2</td>
<td>3293.33</td>
<td>0.5</td>
<td>165</td>
</tr>
<tr>
<td>4, 4</td>
<td>3317.4</td>
<td>0.33</td>
<td>279.4</td>
</tr>
<tr>
<td>4, 8</td>
<td>3353.33</td>
<td>1.2</td>
<td>409</td>
</tr>
<tr>
<td>4, 20</td>
<td>3378.5</td>
<td>0.67</td>
<td>568</td>
</tr>
<tr>
<td>4, 100</td>
<td>3311</td>
<td>1</td>
<td>648.5</td>
</tr>
<tr>
<td>5, 2</td>
<td>3332.33</td>
<td>1</td>
<td>208</td>
</tr>
<tr>
<td>5, 4</td>
<td>3273</td>
<td>36.33</td>
<td>260</td>
</tr>
<tr>
<td>5, 8</td>
<td>3267</td>
<td>23</td>
<td>320.5</td>
</tr>
<tr>
<td>5, 20</td>
<td>3295</td>
<td>36.5</td>
<td>424.33</td>
</tr>
<tr>
<td>5, 100</td>
<td>3251.67</td>
<td>51.67</td>
<td>453.33</td>
</tr>
<tr>
<td>10, 2</td>
<td>3169</td>
<td>28.67</td>
<td>226.5</td>
</tr>
<tr>
<td>10, 4</td>
<td>3133.67</td>
<td>215</td>
<td>245.33</td>
</tr>
<tr>
<td>10, 8</td>
<td>3145.33</td>
<td>212.33</td>
<td>264.33</td>
</tr>
<tr>
<td>10, 20</td>
<td>3168.33</td>
<td>217</td>
<td>269.67</td>
</tr>
<tr>
<td>10, 100</td>
<td>3094</td>
<td>211.5</td>
<td>266.5</td>
</tr>
</tbody>
</table>

Table 2.4: Comparison of BF test with other tests based on the proportion of actually DE genes selected in Similar Variance Method.
that, the number of genes selected by all three tests by Method 1 is proportional to the window size. But, most of them were bogus, i.e., genes seem to be differentially expressed but are not really. Even though the window size is small, the proportion of genes selected by these methods increased as window size decreased. This means that for a fixed sample size, small values of the priors $\nu_0$ and $\tau_0$ are preferred.

To see the performance of Method II, we have simulated 1000 genes as in the other two methods. Because it selects huge number of genes with similar variance according to our cut-of criterion, the computation is time consuming (which may not be good for large number of genes and low memory computer). This method compares with the equal variance and SAM as seen from the proportion and the number of truly differentially expressed genes selected. In this case we have introduced only 20 genes as differentially expressed. All three tests—BF, EV and the SAM were quite competitive in this method. But, EV and BF are relatively more competitive. We see that SAM selects very few genes when the sample sizes are small. But as the sample size was increased to 15, it selected more genes. On the other hand, both BF and the EV method selected almost the similar number of genes even though the sample size was increased. The result is shown in Table 2.4.

Table 2.5 shows the number of gene selected by the Method III. We have compared
the actual number of genes selected by different methods including SAM. We see that
the genes selected by our method is comparable to EV test and SAM. We notice that
SAM selects small number of genes relatively in the small sample sizes. Although
our BF method chooses more genes, it selects genes that are actually expressed. The
proportion of actually DE genes selected by BF method is comparable with the EV
test, and performs better than the SAM method. The number of genes selected by
different methods increases as the sample size increases. But the EV test and BF
test select more actually expressed genes than the SAM and t-test. This means the
the false positive rate of these two tests are smaller than that of SAM and t-test.
The false positive rate is defined as probability of false positives among the positive
findings. False Positive Rates for the samples having $m = n = 5$ are: (BF, 0.025),
(EV, 0.020), (SAM, 0.072) and (t-test, 0.42). This means that 2.5% genes selected by
BF method are bogus whereas 7.2% genes selected by the SAM method are bogus.
The false discovery rate is defined as the expected proportion of false positives among
the positive findings. So, our proposed method (Method II) seems better than SAM
and t-test while selecting the truly differentially expressed genes.

2.15 Comparison of Proposed Method with other Methods

Table 2.6 shows the comparison of aforementioned three methods. Here we have
calculated the false discovery rate (FDR), which is the expected proportion of false
positives among the positive findings; Proportion of DE genes; and False Negative
rate (FNR), which is the expected proportion of true negatives among the truly DE
genes for each sample and compared them. It has been found that the proportion of
truly DE genes in the Method III is higher than in the other two methods, taking the
same sample sizes in all three methods. This means that bootstrap method selects
more truly DE genes than other two method. Similarly, the FDR is also small in
bootstrap method than other two methods. And the FNR is smaller in method III
than in method I and comparable to that of Method II. Hence it seems that the
bootstrap method is the best method in selecting the truly DE genes than window
and similar variance method in our simulated data.
Table 2.6: Comparison of three methods ( I = Window Method, II = Similar Variance Method, III = Resampling Method ) according to Proportion of truly DE genes selected, FDR and FNR.

<table>
<thead>
<tr>
<th>Samples m = n</th>
<th>Proportion of DE Genes</th>
<th>FDR</th>
<th>FNR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>3</td>
<td>0.35</td>
<td>0.68</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>0.56</td>
<td>0.79</td>
<td>0.98</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
<td>0.91</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>0.74</td>
<td>0.87</td>
<td>0.97</td>
</tr>
<tr>
<td>10</td>
<td>0.82</td>
<td>0.8</td>
<td>0.96</td>
</tr>
</tbody>
</table>

2.16 Selection of Differentially Expressed Genes in the Golub Data

The selection of differentially expressed genes in the micro-array data has been one of the best studied area in the microarray literature. In this chapter, I am going to analyze the real microarray Affymetrix data to see whether the proposed Behrens Fisher method competes with the other methods like Significance Analysis of Microarray (SAM) introduced by Tusher et al. [48]), Linear Methods for Microarray Analysis (LIMMA) introduced by Gordon Smyth [41], Bayesian t-test introduced by Fox and Dimmic [13] and Empirical Bayes method introduced by Newton and Kendziorski [29]. The Golub data is an Affymetrix data that consists of two types of cancers of 72 patients. This data was first analyzed by Golub et al. [16] to classify Acute Lymphoid Leukemia (ALL) with Acute Myeloid Leukemia (AML). There are 38 ALL samples and 34 AML samples. There are 7129 genes and 72 samples. The first step in analyzing an affymetrix data is to pre-process the data to remove the bias and outliers because of background effect measured when collecting the data. It distorts the outcome of the analysis if not performed correctly. Here, I have used the preprocessing step as follows:

1. Set floor of 100 units and ceiling of 16,000 units.

2. Filter the low quality genes.

3. Transform the data by using base-10 logarithm.

The threshold was set with floor of 100 units and ceiling of 16,000 units. A ceiling of 16,000 units was chosen because it is at this level that we observe the fluorescence
Figure 2.5: Histograms of Unprocessed Data Golub Data
Figure 2.6: Histogram of log-2 transformed Golub Data
Figure 2.7: QQ-Plot of Golub Data
saturation of the scanner, below or above this values above this can’t be reliable measure, and called them as outliers. Similarly a floor of 100 units was chosen to minimize the noise and maximise the interpretation of marker genes due to the correlation of genes. We have filtered out (excluded) those low quality genes that have ratio \((\max/\min) < 5\) and \((\max – \min) < 500\) across all of the samples. To make the data symmetrical (normal), base-10 logarithm has been used for the transformation. Because of this logarithmic transformation, the skewed data will be almost normal and symmetrical.

The histograms in the Figure 2.5 are the raw expression data and after logarithmic transformation. The histogram in Fig. 2.6 are the histograms of the overall expression level of Golub data in the training and the test samples after the logarithmic transformation. So, from the histograms we see that the overall expression patterns in each of the samples from the two conditions are symmetrical and normal after transformation.

### 2.17 Transformation and Test of Normality Assumptions

In the golub data, after the standard pre-processing step as mentioned in the above section, the normality assumption was checked using the Goodness-of-Fit tests such as Shapiro-Wilks test for normality [40]. The null hypothesis is the data \(x_1, x_2, \ldots, x_n\) come from the normally distributed population. The test statistic is

\[
SW = \frac{\left(\sum_{i=1}^{n} a_i x^{(i)}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

where, \(\bar{x}\) is the mean of samples, \(a_i\)'s are constants given by

\[
(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V m)^{1/2}}
\]

and \(m = (m_1, m_2, \ldots, m_n)\) and \(m_i\)'s are expected values of order statistics, and \(V\) is the covariance matrix of these order statistics. If the p-value of the test statistic \(SW\) is
less than confidence level $\alpha$, then reject the null hypothesis, data come from normal.

This test can be used instead of the Kolmogorov-Smirnov test (Lilliefors test), because some studies has shown that this test has good power in many situations than other goodness of fit test of the composite hypothesis of normality (Shapiro, Wilk and Chen 1960). This test is not affected by the ties in the values. In this test, we were 95% confident that the genes satisfy the normality assumption. Furthermore, those genes which did not satisfy the normality assumption were transformed to normality using the Yeoh-Johnson transformation [53]. It is given by

$$
\psi(\lambda, x) = \begin{cases} 
\frac{(x + 1)^\lambda - 1}{\lambda}, & (x \geq 0, \lambda \neq 0); \\
\log(x + 1), & (x \geq 0, \lambda = 0); \\
-\frac{((-x + 1)^{2-\lambda} - 1)}{(2 - \lambda)}, & (x < 0, \lambda \neq 2); \\
-\log(-x + 1), & (x < 0, \lambda = 2).
\end{cases}
$$

(2.17.35)

where, $\lambda$ is a parameter, that is estimated from the data assuming the data come from normal population using the maximum likelihood method.

The Fig. 2.8 shows the Q-Q plot of some of the genes in the Golub data before Yeoh-Johnson transformation. We can see from the graph that the distribution of the some of the genes are not normal. The Yeoh-Johnson transformation is similar to the Box-Cox transformation except the former is also useful for transforming positive or negative values. We can see from the Fig. 2.10 that the genes satisfies normality after transformation.

Here in the following, we have computed the mean and variance of first few genes of the Golub data [16]. We see that the variance depends on the mean expression level. But after the estimation of prior variances, the posterior variance is stabilized and this has the effect on selecting the actual expressed genes.

Differentially expressed genes were selected using the Behrens-Fisher (BF) distribution. To overcome the false genes which arises due to the multiple testing, we have controlled the familywise error rate (FWER) at the level $\alpha = 0.05$ using the Bonferroni
### Table 2.7: Dependence of Variance on Expression level

<table>
<thead>
<tr>
<th>Genes</th>
<th>Mean</th>
<th>Var Before</th>
<th>Var After</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX - BioDn - 3_at</td>
<td>-1.01454519</td>
<td>0.22596073</td>
<td>0.281732</td>
</tr>
<tr>
<td>AFFX - BioB - 5_st</td>
<td>-0.89731151</td>
<td>0.24048058</td>
<td>0.281736</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - MA_at</td>
<td>-1.18813202</td>
<td>0.10450397</td>
<td>0.281698</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - MB_at</td>
<td>-0.89940262</td>
<td>0.15503013</td>
<td>0.281712</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 3_at</td>
<td>0.21463536</td>
<td>0.25394972</td>
<td>0.28174</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 1_at</td>
<td>0.34598585</td>
<td>0.39634255</td>
<td>0.282684</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 5_st</td>
<td>0.0673037</td>
<td>0.5154905</td>
<td>0.282492</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 3_at</td>
<td>0.13062448</td>
<td>0.25132857</td>
<td>0.282435</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 1_at</td>
<td>3.0651817</td>
<td>3.35897426</td>
<td>0.28178</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 5_at</td>
<td>3.0229095</td>
<td>2.93944803</td>
<td>0.28182</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 3_at</td>
<td>3.13561769</td>
<td>2.73528247</td>
<td>0.281739</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 1_at</td>
<td>3.2402988</td>
<td>2.71219933</td>
<td>0.281805</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 5_at</td>
<td>-0.52114347</td>
<td>0.48562603</td>
<td>0.281714</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 3_at</td>
<td>-1.16505316</td>
<td>0.16032256</td>
<td>0.281698</td>
</tr>
<tr>
<td>AFFX - HUMISGF3A/M97935 - 1_at</td>
<td>-0.56158528</td>
<td>0.10416494</td>
<td>0.281733</td>
</tr>
<tr>
<td>AFFX - M27830 - 5_at</td>
<td>0.91157724</td>
<td>0.23132494</td>
<td>0.282428</td>
</tr>
<tr>
<td>AFFX - M27830 - 1_at</td>
<td>1.74595905</td>
<td>0.94761702</td>
<td>0.281934</td>
</tr>
<tr>
<td>AFFX - M27830 - 3_at</td>
<td>0.2323194</td>
<td>0.3217269</td>
<td>0.281762</td>
</tr>
<tr>
<td>AFFX - HSAC07/X00351 - 3_at</td>
<td>1.13089924</td>
<td>0.39854292</td>
<td>0.28178</td>
</tr>
<tr>
<td>AFFX - HSAC07/X00351 - 5_at</td>
<td>-0.7434625</td>
<td>0.27957232</td>
<td>0.281747</td>
</tr>
<tr>
<td>AFFX - HSAC07/X00351 - 3_at</td>
<td>-0.4391125</td>
<td>0.3516393</td>
<td>0.281767</td>
</tr>
<tr>
<td>AFFX - HSAC07/X00351 - 1_at</td>
<td>-0.33299838</td>
<td>0.3902128</td>
<td>0.281778</td>
</tr>
<tr>
<td>A28102_at</td>
<td>-0.56174915</td>
<td>0.35655727</td>
<td>0.281769</td>
</tr>
<tr>
<td>AB000115_at</td>
<td>-0.5748046</td>
<td>0.4019783</td>
<td>0.281781</td>
</tr>
<tr>
<td>AB000409_at</td>
<td>-1.15104296</td>
<td>0.05802834</td>
<td>0.281685</td>
</tr>
</tbody>
</table>
correction method and false discovery rate (FDR) using the Benjamini and Hochberg method. The genes selected by the BF method is compared with other current gene selection methods such as: SAM, LIMMA and FD methods. This method selects most of the genes that are also selected by the other methods. SAM and LIMMA do not require the data to be normally distributed. As seen from the table 2.8, the proposed method is competitive to these method. It has selected the genes in the midway between SAM and LIMMA.

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of Genes</th>
<th>Common Genes</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF</td>
<td>713</td>
<td>BF, SAM</td>
<td>713</td>
</tr>
<tr>
<td>SAM</td>
<td>841</td>
<td>BF, LIMMA</td>
<td>670</td>
</tr>
<tr>
<td>LIMMA</td>
<td>678</td>
<td>SAM, LIMMA</td>
<td>678</td>
</tr>
<tr>
<td>FD</td>
<td>703</td>
<td>FD, BF</td>
<td>658</td>
</tr>
<tr>
<td>All 4</td>
<td>655</td>
<td>BF, LIMMA, SAM</td>
<td>670</td>
</tr>
</tbody>
</table>

Table 2.8: Comparision of Differentially expressed genes in Golub data controlling the FDR at the level $\alpha = 0.05$

2.18 Conclusion

We have proposed a new method to get the differentially expressed genes if the variances are different in samples. In all of those, our proposed method performed better than the equal variance (EV) test (see Table 2.3) and SAM (see Table 2.4 and Table 2.5). Furthermore, we have compared the three methods based on the proportion of truly differentially expressed genes selected, FDR, and FNR criterion (see Table 2.6) and found that the bootstrap method gave the best result among three proposed methods. We have implemented our theory in R statistical software, http://www.r-project.org. In the above the FDR criterion to control the false positives is preferred than the familywise error rate (FWER) control criteria, because $\text{FDR} \leq \text{FWER}$. When some of the genes are actually not differentially expressed, then the FDR is strict than the FWER therefore FDR controlling multiple testing procedure is more powerful than the FWER controlling procedure [?].

In this chapter, we derived an expression for the posterior Behrens-Fisher distri-
bution for the heterogeneous variance using the uniform and scaled inverse chi-square priors. The theory thus derived was applied to select the differentially expressed genes between two samples. We applied the result for the simulated data and real data. The multiplicity problem that arises while testing thousands of hypotheses simultaneously was controlled using the false discovery rate. We have compared our method of selecting the genes with other popular methods such as LIMMA, SAM and FD in the sections 2.11 and 2.12. We have found that the proposed gene selection methods performs better than the other mentioned methods.
Figure 2.8: Q-Q plot of some of the genes before Yeo-Johnson Transformation
Figure 2.9: Histogram of the T-values obtained by the BF Method
Figure 2.10: Q-Q plot of some of the genes after Yeo-Johnson Transformation
Chapter 3

Support Vector Machines and Other classification Methods

3.1 Summary

In this chapter we will discuss different classification methods, such as Support Vector Machines (SVM), Weighted Vote (WV) Method, Nearest Shrunken Centroid (NSC) Method, and Multiclass Classification Method. We have compared the classification performance of these methods on the real and simulated data using the genes selected by the Behrens Fisher Distribution in Section 3.10. The performance of these classification methods have been compared. We have found that the classification performance improves as we take the genes obtained by the Beherns-Fisher method. In the simulated data, the ranking of the methods were: Weighted-Vote (WV), Support Vector Machines (SVMs), Multiclass method, Nearest Shrunken Centroid (NSC) and Shrunken Centroid Regularised Discriminant Analysis (SCRDA) respectively. In the real data also the Weighted Vote and Support Vector methods were the best for classifying the tumor data that we have analyzed.

3.2 Introduction

Support vector machines (SVMs) introduced by Vapnik in 1995 [49] are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized classifiers and is considered as a special case of Tikhonov regularization. A special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin. Hence it is also
known as maximum margin classifier. A learning method (classification method) is called unsupervised if it learns (classifies) in the absence of a teacher signal. Unsupervised learning method begins with a definition of similarity (measure of distance) between expression patterns, but with no prior knowledge of true functional class of genes. Genes are then grouped by using a clustering algorithm.

3.3 Assumptions on Classification Methods

The assumptions made by these five different methods are as follows:

- **Weighted Vote:** This method assumes that the data are independent and identically distributed. It does not assume the normality.

- **Support Vector Machines:** This method also assumes the data are independent but does not assume any normality.

- **Multiclass Method:** This method assumes that the data are \( i.i.d. \).

- **Nearest Shrunken Centroid:** This method assumes that the data are \( i.i.d \) from normal distribution.

- **SCRDA Method:** This method also assumes that the data are \( i.i.d \) from normal distribution.

The above five methods are the brief summary of the classification methods we will use for the classification of samples. More about these methods can be found on [16], [14], [47], and [49]. In these methods, the non-homogeneity of the samples have not been used. So, one would expect that if the non-homogeneity was taken into account, the classification method separates the data more accurately into their true classes.

3.4 Support Vector Machines (SVM) Methods

The support vector machine (SVM) is one of the methods that has been successfully applied to the cancer diagnosis problem in previous studies Mukherjee et al.[28], 1999;
Ramaswamy et al., 2001)\cite{34}. In this method, one finds the separating hyperplane

\[ f(x) = b + w'x \]  \hspace{1cm} (3.4.1)

from the training samples \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \) where \( y_i = \pm 1 \) are binary class labels, that correctly classifies the training samples and maximizes the margin. The class of a new sample \( x \) is determined by the sign\([f(x)]\). All the training samples are classified correctly if

\[ y_i(b + w'x_i) \geq 1 \quad \text{for all} \quad i \]

The hyperplanes

\[ b + w'x = \pm 1 \]

are called the canonical hyperplanes and the distance \( 1/||w|| \) between one of the canonical hyperplanes and separating hyperplane (3.4.1) is the margin. So, the optimization problem can be rephrased as

\[ \min_{w} ||w|| \]

subject to

\[ y_i(b + w'x_i) \geq 1 \quad \text{for all} \quad i \]

For the non-separable case, we still maximize the margin but we allow some points on the wrong side of the hyperplane defining the slack variables \( \xi = (\xi_1, \xi_2, \ldots, \xi_n) \). The optimization problem is

\[ \min ||w|| + \gamma \sum_{i=1}^{N} \xi_i \]

subject to

\[ y_i(b + w'x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \text{for all} \quad i \]

where, \( \gamma \) is the cost parameter that is determined by cross-validation. Using the Karush-Kuhn-Tucker condition, the optimal values of the parameters of the hyper-
plane are obtained as
\[
\hat{w} = \sum_{i \in SV} \alpha_i y_i x_i
\]
where, \(0 < \alpha_i \leq \gamma\). The sum \(\sum_i \alpha_i y_i = 0\) corresponds to the support vectors \(x_i\). The bias parameter
\[
\hat{b} = \frac{1}{n_0} \left\{ \sum_{i \in SV} y_i - \sum_{i,j \in SV} \alpha_i y_i x_i' x_j' \right\}
\]
where \(n_0\) is the number of support vectors. Since \(\alpha_i = 0\) for the non-support vectors \(x_i\), the summation indices \(i\) and \(j\) are only for the support vectors. The separating hyperplane is thus given by
\[
f(x) = \hat{b} + \hat{w}'x
\]
and the corresponding decision rule for a new sample \(x^*\) is:
\[
f(x^*) = \hat{w}'x^* + \hat{b}\begin{cases} > 0, & \text{Classify to class 1;} \\ < 0, & \text{Classify to class 2.} \end{cases}
\]

Using the non linear basis functions \(\phi(x_i)\), one can map the input space into a high dimensional feature space. Then the samples are classified by the linear boundaries in the feature space using the kernel \(K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)\) which corresponds to the non linear boundaries in the input space. The separating hyperplane in the feature space is
\[
f(x) = \hat{b} + \sum_{i=1}^{N} \alpha_i y_i K(x_i, x)
\]
Different kernels are obtained by
\[
K(y, x) = \phi'(y) \phi(x)
\]
where \(\phi(y)\) and \(\phi(x)\) can be any linear or non linear transformations of \(y\) and \(x\) and must satisfy the Mercer’s Conditions [21]. But for this work, we use simple linear
kernel,

\[ K(y, x) = \phi'(y)\phi(x) = y'x + 1 \]

### 3.5 Kernel Matrix and Kernel Tricks

The matrix \( K = K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle \) is called the kernel matrix [21]. This will be of particularly important in the extensions of the algorithm. In the kernel based learning methods, the input features are mapped into a high dimensional feature space by a projection map \( \Phi : X \rightarrow Y \), such that the dot product in the feature space \( Y \) is represented by a kernel \( k(x, y) = \langle \Phi(x), \Phi(y) \rangle \). The map \( \Phi \) need not to be determined explicitly and the computation in feature space can be done simply by taking the dot product. Hence once the suitable kernel (which is symmetric and positive definite) function is chosen, the user even do not need to worry about the features being used. After this transformation, the classifier is used to classify the samples in the feature space. This trick of computation is called the kernel trick.

### 3.6 High Dimensional Feature Space for Large \( p \) Small \( n \)

In the DNA microarray technology to classify the positive samples from the negative samples, the main difficulty arises from the fact that the number of samples, \( n \), is small, generally less than one hundred, relative to the number of genes, \( p \), which is in several thousands. Golub et al. (1999) used the gene expression data to classify between acute myeloid leukemia (ALL) and acute lymphocytic leukemia (ALL). Since then, disease classification using microarray data has been focus of intensive research with the aim of providing more accurate diagnostic tools than what traditional pathological method alone can provide.

The basic idea of high dimensional comes from the fact that, the training pairs \((x_i, y_i), \quad i = 1, 2, ..., n\) are separated by maximizing the margin between the support
vectors by the hyperplane

\[ \{ \mathbf{x} : D(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0 \} \]

where \( \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \). In the cases when it is impossible to get a hyperplane that classify the points in low dimension, we map those points into high dimensional space thus making it possible to classify easily.

Let \( \phi \) be a feature map that maps each point \( \mathbf{x} \) into \( \phi(\mathbf{x}) \). Then the decision function in the feature space is given by

\[ D(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b \]

where \( \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathbf{x}_i) \). So, the decision surface is given by

\[ D(\mathbf{x}) = \sum_{i \in S} \alpha_i y_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b \]

According to the Hilbert-Schmidt theory, the inner product in the feature (Hilbert) space can be represented as

\[ \phi(\mathbf{x}) \cdot \phi(\mathbf{y}) = \sum_{i=1}^{n} a_r \phi_r(\mathbf{x}) \cdot \phi_r(\mathbf{y}) = K(\mathbf{x}, \mathbf{y}) \]

where \( K(\mathbf{x}, \mathbf{y}) \) is a symmetric function that satisfies the Mercer’s Condition

\[ \sum_{i=1}^{n} \sum_{j=1}^{n} K(\mathbf{x}, \mathbf{y}) h_i(\mathbf{x}) h_j(\mathbf{y}) \geq 0 \]

for all \( h_i, h_j \), and natural number \( M \). This means that for any kernel function \( K(\mathbf{x}, \mathbf{y}) \) satisfying the Mercer’s condition there exists a feature map \( \mathbf{g} \) such that the inner product in the feature space is generated by \( \mathbf{g} \).
By above property, the decision surface now can be written as

\[ D(x) = \sum_{i \in S} \alpha_i y_i K(x_i, x) + b \]

where, \( x \) is a test point and \( x_i \) is a support vector obtained from training set. The decision rule is

\[ x \in \begin{cases} 
  \text{Class 1, if } D(x) > 0 \\
  \text{Class 2, if } D(x) < 0 \\
  \text{undecided if } x = 0.
\end{cases} \]

Hence, we see that, the points are linearly separable in the feature space although they are not linearly separable in the input space. For this we need to know the kernel functions, instead of the actual transformation. Some of the important kernels are:

- **dth Degree Polynomial Kernel**, \( K(X, Y) = (1 + \langle X, Y \rangle)^d \)
- **Radial Basis Kernel**, \( K(X, Y) = \exp\left(-\frac{||X-Y||^2}{c}\right) \)
- **Neural Network Kernel**, \( K(X, Y) = \tanh(\kappa_1 \langle X, Y \rangle + \kappa_2) \)

Note that each random variable \( X \) represents a sample. So, in the case of gene expression, the number of components of \( X \) are very large.

### 3.7 Nearest Shrunken Centroids Method

The purpose of discriminant analysis or classification is to assign samples to one of the several (G) classes based on a set of measurements \( x = (x_1, x_2, ..., x_p) \) measured from samples. In the case of supervised learning, the classes are predetermined from a set of samples, called *training samples*. These training samples are used to build a
classifier. The classifier is then used to determine the class of a new sample. When a sample is misclassified, then an error is said to be incurred. The cost or loss associated with such an error is defined as

\[ L(k, \hat{k}) = \begin{cases} 
0, & \text{if } k = \hat{k} \\
1, & \text{otherwise.} 
\end{cases} \tag{3.7.5} \]

where \( k \) is the correct group of the sample and \( \hat{k} \) is the assignment made to that sample by the classifier. If the class conditional densities, \( f_k(x) \), and the class priors, \( \pi_k \), of class \( k \) are known, then Bayesian optimal rule of classification to classify a new sample \( x^* \) is to minimize the risk

\[ R(\hat{k}|x) = \sum_{k=1}^{G} L(k, \hat{k}) \Pr(G = k|X = x) \tag{3.7.6} \]

Then the classification rule is:

\[ \hat{k} = \text{argmax}_k f_k(x^*)\pi_k. \]

But, the problem is we do not know the class conditional densities, \( f_k(x) \), of each of the classes. So, many researchers assume that these densities are multivariate normal with densities,

\[ f_k(x) = (2\pi)^{-n/2}|\Sigma_k|^{-1/2} \exp\left[-\frac{1}{2} (x - \mu_k)'\Sigma_k^{-1}(x - \mu_k)\right] \]

Then assuming the equal variance-covariance matrix of each class, \( \Sigma_k = \Sigma \), the linear discriminant score

\[ D^l_k(x) = x'\Sigma^{-1}\mu_k - \frac{1}{2} \mu_k'\Sigma^{-1}\mu_k + \log \pi_k \tag{3.7.7} \]

and assuming classwise covariance matrices unequal, the quadratic discriminant score

\[ D^q_k(x) = -\frac{1}{2} \log|\Sigma_k| - \frac{1}{2}(x - \mu_k)'\Sigma_k^{-1}(x - \mu_k) + \log \pi_k \tag{3.7.8} \]
Generally, the maximum likelihood estimate are used to estimate the population mean and population variance. Furthermore, the empirical probability is used to estimate the class priors.

In the case of DNA microarray data, the number of covariates (genes), \( p \), which are in the order of several thousands, are much greater than the number of samples, \( n_k \), generally within hundreds, in each class. So, the sample covariance matrix is singular and this gives the unreliable estimate of covariance matrix because of high variability. Friedman [14] introduced the regularized discriminant analysis in which the the unequal variances were shrunked towards the common variance using the regularization, thus increasing the performance of the RDA classifier. The microarray problem is thus unique and challenging. Since the expression level of most of the genes are same in two different treatment samples, those genes contribute little in the case of classification. Thus it is important to identify the genes that actually contribute for the classification. Assuming that those genes which have common class means do not contribute for the classification, Tibshirani et. al. [47] proposed the Nearest Shrunken Centroids (NSC) method. They used the shrinkage parameter, \( \Delta \), for thresholding and declared those genes as non-contributing genes if the shrunken centroids for gene \( g \) in class \( k \),

\[
\bar{x}'_{gk} \text{ shrinks to the overall mean } \bar{x}_g
\]

when the

\[
|d_{gk}| - \Delta \leq 0
\]

where,

\[
s_g^2 = \frac{1}{n - G} \sum_{k=1}^{G} \sum_{j \in C_k} (x_{gj} - \bar{x}_{gk})^2, \quad d_{gk} = \frac{\bar{x}_{gk} - \bar{x}_g}{\sqrt{1/n_k + 1/n} \cdot s_g^2}
\]

The non-contributing genes are removed from the data, thus reducing the dimensionality of the gene-matrix. The discriminant score of the NSC classifier was defined as
\[
D_k(\mathbf{x}^*) = \sum_{g=1}^{p} \frac{(x_g^* - \bar{x}'_{gk})^2}{s_g^2} - 2 \log \pi_k \tag{3.7.9}
\]

In the classification process, the genes \( g \) which have each of the shrunken class-means \( \bar{x}'_{gk} \) shrinks towards the overall class means \( \bar{x}_g \) in each class \( k = 1, 2, \ldots, G \) have the same \( (x_g^* - \bar{x}'_{gk})^2 \) values. So, the numerator of the above discriminant score (3.7.9) can be replaced by the square of differences of only those genes \( g \) for which

\[
\bar{x}'_{gk} \neq \bar{x}_g, \quad \forall \ k = 1, 2, \ldots, G
\]

The optimal value of the shrinkage parameter, \( \Delta \), is chosen by the cross validation that minimizes the cross-validation error. The idea of cross-validation is to obtain the unbiased estimate of future prediction error associated with a particular observation and is obtained by removing it from the model. This gives the genes that are useful for classification.

### 3.8 Weighted Voting Method

Here we briefly review the methods of classification that are used by Golub et al. [16]. To identify the genes which are truly expressed in the new samples, one can use the weighted voting scheme (WVS) method. This uses a weighted linear combination of the "marker" or "relevant" genes obtained in the training set to classify the new sample. In this method, the correlation between the expression values of a gene \( g \) in two classes is defined as

\[
w_g = \frac{\mu_{g1} - \mu_{g2}}{\sigma_{g1} + \sigma_{g2}} \tag{3.8.10}
\]

where \( \mu_{gi} \) and \( \sigma_{gi} \) are the mean and standard deviations of gene \( g \) in the class \( i, \ i = 1, 2 \). The larger the absolute value \( |w_g| \) is the more important the gene \( g \) is for prediction. The genes are ranked by their \( |w_g| \)'s and top ones are selected. These top selected genes are the marker or informative genes.
For each informative gene $g$ in the training sample, let $\mu_{g1}$ and $\mu_{g2}$ be the means and $\sigma_{g1}$ and $\sigma_{g2}$ be the standard deviations respectively. Then the weight of gene $g$ is determined by

$$w_g = \frac{\mu_{g1} - \mu_{g2}}{\sigma_{g1} + \sigma_{g2}}$$

(3.8.11)

This measure is also called the signal-to-noise ratio. This weighting factor reflects the correlation between the expression level of gene $g$ and class distinction. The parameter $b_g$ is calculated as

$$b_g = \frac{\mu_{g1} + \mu_{g2}}{2},$$

(3.8.12)

which is the average mean of expression levels of two classes. Hence we define the parameters $(w_g, b_g)$ for each informative gene in the training set. For a new sample $x^*$ with $x^*_g$ being the normalized log expression level of the gene $g$, we calculate the votes casted by each of the genes in the "informative" set. The vote of a gene $g$ is

$$v_g = w_g(x^*_g - b_g) = \frac{\mu_{g1} - \mu_{g2}}{\sigma_{g1} + \sigma_{g2}} [x^*_g - \frac{\mu_{g1} + \mu_{g2}}{2}]$$

(3.8.13)

A positive vote indicates that the sample belongs to class 1 and negative vote indicates it being in class -1. Then the total vote for the sample to be in class 1 is obtained by adding $V_1 = \sum_g max(v_g, 0)$ and the total vote for sample to be in class -1 is $V_2 = \sum_g max(-v_g, 0)$. Then the sample is assigned to that class corresponding to the higher total vote. Generally, we take the 5% of most positive and 5% most negative genes as the "informative" genes in the training set. But this number is a free parameter and depends on the user.
3.9 Dudoit’s Multi-class Classification Method

Several proposals have been made for ranking the genes for multiclass classification. Dudoit et al. (2002) used the ratio of between-sum-squares to within-sum-squares of each gene for the multiclass classification. Explicitly, let there are $G$ classes and the number of samples be $n$ each of dimension $p$. Then, the samples in each class

$$n = n_1 + n_2 + \ldots + n_G$$

Let $\bar{x}_g$ be the mean of gene $g$ over all classes. For each gene $g$, $g = 1, 2, \ldots, p$, let $\bar{x}^{(k)}_g$ be the mean in class $k$, $k = 1, 2, \ldots, G$. Then the ranking of genes are done using the ratio

$$\rho_g = \left| \frac{\sum_{k=1}^{G} n_k (\bar{x}^{(k)}_g - \bar{x}_g)^2}{(n - G)\sigma_g^2} \right|$$

where, $\sigma_g$ is the pooled within-class standard deviation of gene $g$:

$$\sigma_g^2 = \frac{1}{(n - G)} \sum_{k=1}^{G} (n_k - 1)\sigma^{(k)}_g^2$$

A new sample $x^* = (x^*_1, x^*_2, \ldots, x^*_p)$ is then classified into class $k$, if

$$k = \min_{k'} ||x^* - \bar{x}^{(k')}||$$

where, $||.||$ is the Euclidean norm, and $x^*$ and $\bar{x}^{(k')}$ are the component vector and mean of class $k'$ of only those component genes selected by the ranking procedure.

3.10 Gene Selection by Behrens-Fisher Statistic

In this section we are going to discuss briefly how we have developed the method to chose the genes used for the classification. Suppose there are $G$ different classes in the population. Let $n_k$ be the number of samples in class $k$, $(k = 1, 2, \ldots, G)$. Let $x_{gk} = (x_{g1}, x_{g2}, \ldots, x_{g_{nk}}) \overset{iid}{\sim} N(\mu_{gk}, \sigma_{gk}^2)$ be the expression level (possibly log
transformed) of the gene $g$ in class $k$. For $k=1,2,\ldots,G$, the density of $x_{gk}$ can be written as

$$f(x_{gk}) = \frac{1}{(\sigma_{gk})^{n_k}(2\pi)^{n_k}} \exp\left[-\frac{1}{2\sigma_{gk}^2} (n_k - 1)s_{gk}^2 + n_k(\bar{x}_{gk} - \mu_{gk})^2\right]$$

(3.10.14)

where, $\bar{x}_{gk}$ and $s_{gk}^2$ are sample mean and sample variance of gene $g$ in class $k$ respectively. Assuming the independency of location parameter $\mu_{gk}$ and scale parameter $\sigma_{gk}^2$, the joint prior for $\mu_{gk}$ and $\sigma_{gk}^2$ can be written as

$$p(\mu_{gk}, \sigma_{gk}^2) = p(\mu_{gk})p(\sigma_{gk}^2)$$

(3.10.15)

Assume that the priors for $\mu_{g1}$ and $\mu_{g2}$ are flat priors and the priors for $\sigma_{g1}^2$ and $\sigma_{g2}^2$ are scaled inverse $\chi^2$ distributions, i.e. $p(\sigma_{g1}^2) = I(\sigma_{g1}^2; \nu_0, \sigma_0^2)$ and $p(\sigma_{g2}^2) = I(\sigma_{g2}^2; \eta_0, \tau_0^2)$, where, $\alpha = (\nu_0, \eta_0, \sigma_0^2, \tau_0^2)$ is the hyper-parameters that should be estimated from the data.

Let $\Delta \mu_g = \mu_{g2} - \mu_{g1}$. Then the statistic, called the BF-statistic

$$B = \frac{\Delta \mu_g - (\bar{x}_{g2} - \bar{x}_{g1})}{(\frac{\sigma_{g1}^2}{n_1} + \frac{\sigma_{g2}^2}{n_2})^{\frac{1}{2}}}$$

(3.10.16)

$$= B_{x_2}\cos\theta - B_{x_1}\sin\theta$$

where,

$$\tan\theta = \frac{\sigma_{g1}/\sqrt{n_1}}{\sigma_{g2}/\sqrt{n_2}}, \quad 0 \leq \theta \leq \frac{\pi}{2}$$

$$B_{x_1} = \frac{(\mu_{g1} - \bar{x}_{g1})}{\sigma_{g1}/\sqrt{n_1}}$$

$$B_{x_2} = \frac{(\mu_{g2} - \bar{x}_{g2})}{\sigma_{g2}/\sqrt{n_2}}$$

and, $B_{x_1}$ and $B_{x_2}$ are independently distributed as $t$- statistics with $v_{n_1}$ and $v_{n_2}$ degrees of freedom respectively. Hence, the statistic $B$ is distributed as the Behrens- Fisher distribution with
\[ v_{n_1} = n_1 + \nu_0 - 1, \quad \text{and} \quad v_{n_2} = n_2 + \eta_0 - 1 \]
degrees of freedom [42]. That is,

\[ B \sim BF(v_{n_1}, v_{n_2}, \theta) \]

with pdf

\[
f(\beta|\mu_{g_1}, \mu_{g_2}, \sigma_{g_1}^2, \sigma_{g_2}^2) = k \int_{-\infty}^{\infty} \left[ 1 + \frac{(\alpha \cos \theta - \beta \sin \theta)^2}{v_{n_1}} \right]^{-\frac{v_{n_1} + 1}{2}} \left[ 1 + \frac{(\alpha \sin \theta + \beta \cos \theta)^2}{v_{n_2}} \right]^{-\frac{v_{n_2} + 1}{2}} d\alpha
\]

where,

\[
\alpha = B_{x_2} \sin \theta + B_{x_1} \cos \theta, \quad \beta = B_{x_2} \cos \theta - B_{x_1} \sin \theta
\]

This can be further approximated by scaled t-statistic [31]:

\[ \frac{B}{a} \sim t(b) \quad (3.10.17) \]

where,

\[
f_1 = \left( \frac{v_{n_1}}{v_{n_2} - 2} \right) \cos^2 \theta + \left( \frac{v_{n_1}}{v_{n_1} - 2} \right) \sin^2 \theta
\]

\[
f_2 = \frac{v_{n_2}^2}{(v_{n_2} - 2)^2(v_{n_2} - 4)} \cos^4 \theta + \frac{v_{n_1}^2}{(v_{n_1} - 2)^2(v_{n_1} - 4)} \sin^4 \theta
\]

\[ a^2 = \frac{(b-2)}{b} f_1 \]

\[ b = 4 + \frac{f_2}{f_1} \]

\[
\cos^2 \theta = \frac{\sigma_{g_2}^2}{\frac{\sigma_{g_2}^2}{v_{n_2}} + \frac{\sigma_{g_1}^2}{v_{n_1}}}, \quad \sin^2 \theta = 1 - \cos^2 \theta.
\]

That is, \( B \) has approximately \( t \)-distribution with \( b \) degrees of freedom \( (b \geq 1) \) and scale parameter \( a \). This statistic \( B \) can also be denoted as \( B \sim t(0, a^2, b) \) and is valid only for \( v_{n_1}, v_{n_2} \geq 5 \).
3.11 Choosing the number of genes required for classification

For this work we select those genes that are uniformly expressed in each of the classes. Since leave-one-out cross validation (LOOCV) error is almost unbiased estimate of generalization error [49], we use the leave-one-out cross validation on the training samples. The genes that are common in each of the cross validated training samples are the preliminary set of genes that have the power to discriminate between different classes. From this preliminary gene set, only those genes are selected that do not further decrease the cross validation error. This final set is the optimal set that is useful for the classification. Since the genes chosen by the t-test are the marker genes for the two different conditions, these can be used to classify samples into any one of the two classes. The genes selected by BF method may not be appropriate for the multi-class classification. To fit with the multi-class classification, we want to choose the marker genes that are useful. For this, we can use the one-versus-all method. In this method, we take one of the class as from normal (condition 1) and combine the rest of classes and take that combined classes as from diseased (condition 2). Then we use the leave-one out method in the training set to choose the genes that are useful for the classification. We simply leave one of the training sample and find the genes that are differentially expressed by the BF method in the rest of training samples that contains samples from both conditions. We repeat this procedure for all the samples in the training set. Then the differentially expressed genes that are common in each of the leave-one-out training set is taken as the marker genes, which we choose as the genes useful for classification. We repeat the same procedure for the rest of the classes (leaving the classes that were used) and get the differentially expressed genes. Finally, the informative genes are those common genes that is found expressed, thus uniformly expressed, in all the classes and leave one out training sets. Using this gene set, we classify one sample with the rest. In the above example, we classify the ALL sample with the rest. Then, this is repeated for all remaining classes. This method is the one-versus-all method.
### 3.12 Simulation and Results

To evaluate how well the genes selected by the Behrens-Fisher Statistic mentioned above performs the classification tasks, we have simulated four classes each with 10,000 genes as in the following:

1. Out of 10,000 genes, simulate 98,000 genes from a normal distribution with mean $\mu \sim U[0, 1]$ and variance $\sigma^2 \sim \text{Inv} - \chi^2(df = 20, \text{scale} = 2)$.

2. Without loss of generality, simulate first 100 genes from normal distribution with mean $\mu \sim U[5, 10] + \frac{1}{2}$ and variance $\sigma^2 \sim \text{Inv} - \chi^2(df = 20, \text{scale} = 4)$ in the control condition and mean $\mu \sim U[5, 10] - \frac{1}{2}$ and variance $\sigma^2 \sim \text{Inv} - \chi^2(df = 18, \text{scale} = 6)$ in the control condition.

3. Simulate another 100 genes from normal distribution with mean $\mu \sim U[5, 10] - \frac{1}{2}$ and variance $\sigma^2 \sim \text{Inv} - \chi^2(df = 18, \text{scale} = 6)$ in the control condition and mean $\mu \sim U[5, 10] + \frac{1}{2}$ and variance $\sigma^2 \sim \text{Inv} - \chi^2(df = 18, \text{scale} = 4)$ in the control condition.

4. Repeat steps 1 to 3 to simulate another two classes, but replace the scales by 3 by and 6 respectively.

5. The combined simulated data will have 10,000 genes and 4 classes.

In the 4 classes, we have $15 - 10, 13 - 5, 30 - 12, \text{and } 23 - 12$ training and test samples respectively. The BF statistic was used to select the genes. It selected most of the genes that were marked as differentially expressed in the simulation. We used these selected genes in the SVM classifier and in the Weighted Voting (WV) methods.

The accuracy rate (AR) of a classifier is defined as

$$AR = \frac{\text{No. of Samples Accurately Classified by the Classifier}}{\text{Total no. of Samples}}$$

The support vector machines misclassified 3.1 samples on average, whereas the WV method misclassified 0.037 of the samples incorrectly on average. The multiclass
algorithm of Dudoit made the most mistakes in the test samples although it classified the samples correctly in the training set. On average, it misclassified 0.15 of the test samples incorrectly. Finally, the RDA performed more accurately. It misclassified 1 samples out of 39 test samples when all 10,00 genes were used. For this, $\alpha = 0.11$ and $\Delta = 2.33$. But, when we use the genes selected by the proposed Behrens-Fisher method, the following Table 3.1 was obtained.

<table>
<thead>
<tr>
<th>Method</th>
<th>no. of Genes used</th>
<th>No. of Classes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>203</td>
<td>4</td>
<td>.92</td>
</tr>
<tr>
<td>WV</td>
<td>203</td>
<td>4</td>
<td>.97</td>
</tr>
<tr>
<td>Dudoit</td>
<td>203</td>
<td>4</td>
<td>.84</td>
</tr>
<tr>
<td>SCRDA</td>
<td>203</td>
<td>4</td>
<td>.84</td>
</tr>
</tbody>
</table>

Table 3.1: Accuracy Rate for the Simulated Test data

3.13 Datasets Pre-processing and Filtering

All the data sets used in this paper are oligonucleotide microarray data and was pre-processed as in Dudoit et. al.[10] (2002). The threshold was set with floor of 100 units and ceiling of 16,000 units. A ceiling of 16,000 units was chosen because it is at this level that we observe the fluorescence saturation of the scanner; values above this can’t be reliable measure. Similarly a floor of 100 units was chosen to minimize the noise and maximise the interpretation of marker genes due to the correlation of genes. We have filtered out (excluded) those low quality genes that have ratio $(\max/\min) < 5$ and $(\max - \min) < 500$ across all of the samples. To make the data somewhat symmetrical, base-10 logarithm has been used for the transformation.

3.14 MLL Leukemia Data

MLL Leukemia data is Affymetrix oligonucleotide data and consists of 72 samples and 12,582 genes. There are 3 different classes - ALL, MLL, and AML. ALL has 20 training sample and 4 testing samples, MLL has 17 training samples and 3 test samples and AML has 20 training samples and 8 testing samples. After the preprocessing and filtering the low quality data, we are left with 8,681 genes.
Figure 3.1: Histogram of Pre-processed AML Data
Figure 3.2: Histogram of ALL, MLL, and AML Training and Testing Data after Pre-processing
Table 3.2: Comparison of Classification Performance on MLL Leukemia Data.

Table 3.2 shows the comparison and performance of different methods for this data. The nearest shrunken centroid (NSC) method chooses only 12 genes but the performance of the model in the testing samples are not as good as in the other method. It makes four errors when classifying the training samples. The genes chosen by the Beherens-Fisher statistic have more discriminating power, as seen by these genes used in weighted voting, Dudoit and Support Vector Machines methods. In all methods, misclassification occurs only in the training samples. The SVM method seems to be one of the perfect classifiers. Since the SVM has solid mathematical derivation that is derived by optimizing the margin between two classes, one can expect that it is one of the best classifiers. We see from the classification performance of Golub data that it makes no error in training and testing samples.

3.15 Golub Leukemia Data

Golub Leukemia data consists of 7,129 genes and 72 samples. These samples are from two classes: Acute Lymphoid Leukemia (ALL) and Acute Myeloid Leukemia (AML). We have chosen 38 training samples: 27 ALL and 11 AML, and 34 testing samples: 20 ALL and 14 AML as in Golub et. al.[16]. First of all, the classification of the samples are done using all the genes.

Here from the Table 3.3, we see that, taking all the genes into consideration, there is one AML sample in that was misclassified as ALL sample by the SVM method. Similarly, the Weighted Vote and Multiclass methods misclassified two AML samples-the one is the same misclassified by the SVM method, into ALL samples.
Figure 3.3: Cross-Validation Error of Nearest Shrunken Centroid Method for MLL data
Next we use only those genes selected by the Behrens Fisher method. We have used the selected genes in three classification methods: Support Vector Machines method, Weighted Vote Method and Dudoit Multiclass Method. Using the training samples (27 ALL and 11 AML), we got 639 genes as differentially expressed. We used these 639 genes for the classification. The result was shown below:

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Errors</th>
<th>Test Errors</th>
<th>Total Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALL</td>
<td>AML</td>
<td>ALL</td>
</tr>
<tr>
<td>W. Vote</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SVM</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multi-class</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison Classification of Golub Data using all the genes

In the Table 3.4, we have seen that the genes selected by the BF method actually are useful for the classification. There is no training errors by each classification methods, all of the 38 samples are classified correctly. To see whether the classifier build by using the selected genes actually work for the unseen samples, it has been found that the the SVM and Weighted Vote misclassifies 1 samples each out of 34 samples and Multiclass method misclassifies two samples. More specifically, SVM method misclassifies one AML sample as ALL. This is the same sample as misclassified by the SVM and multiclass method. The Weighted Vote misclassifies one AML sample as ALL. Multiclass method misclassifies 2 test samples: one ALL sample as AML and one AML sample as ALL. This could be because the BF distribution selects the genes for classifying 2 samples and the multiclass method is generally for the multi-classes.

So, from the comparison of the above two tables, we see that, even though we use all the genes for the classifiers, the classification performance is not improved. In fact, the performance was improved if we take the genes selected by the BF method.
Next, we have selected 33 genes by the BF method using LOOCV error. These genes were used for the classification of ALL and AML samples. The classification error are shown in Figure 3.4. The genes selected by BF method seems optimal set in the sense that it makes very few error in classifying the samples. It makes no error while using the SVM method, whereas it makes 1 error out of 38 training samples and 1 error out of 34 testing samples. As the number of genes increased, the error does not decrease when we select 80 genes. By taking 8 genes, it has been found that 3 errors were made in the training samples and 2 errors were made on the testing sample.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Errors</th>
<th>Test Errors</th>
<th>No.of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALL</td>
<td>AML</td>
<td>ALL</td>
</tr>
<tr>
<td>W. Vote</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SVM</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NSC</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.5: Comparison of Classification Performance on Golub Data.

3.16 Subtypes of Pediatric Acute Lymphoblastic Leukemia

This data consists of 12,625 genes and 327 samples. There are 7 classes and the samples are separated as training and test samples. For the comparison purposes, we have selected the same 215 training and 112 test samples as in Pomeroy et. al. (2001) [33]. These are: BCR (9 train, 6 test), E2A (18 train, 9 test), HYP (42 train, 22 test), MLL (14 train, 6 test), T.ALL (28 train, 15 test), TEL.AML (52 train, 27 test), Others (52 train, 27 test). After applying the pre-processing and filtering steps, there are 12,061 genes. For the sake of convenience, we call this data as ALL-7 data.

In this 7 classes case, the genes selected by the BF method has shown the promising result over the nearest shrunken centroid (NSC) method and shrunken centroid regularized discriminant analysis (SCRDA) method of Guo et. al. (2007). In the NSC and SCRDA methods, the overall mean of each classes are shrunk by a shrinking fac-
Comparison of Misclassification

Figure 3.4: Error Comparison for Golub data
We have used the genes selected by the proposed BF statistic for the Weighted Voting method. For the Dudoit method, we have used the 150 genes selected by the BF statistic. The classification performed by both weighted vote and Dudoit method are better than the rest two methods. The overall error for the weighted vote method in the training samples is 0.0093 whereas it is 0.1209 and 0.0651 in NSC and SCRDA methods respectively. Similarly, the overall test errors in the weighted voting method is 0.0625, and that is 0.3482 and 0.1160 in NSC and SCRDA methods respectively. Another important fact to note is that the number of genes selected for the classification by BF method is also comparable to both of these methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Tr</th>
<th>W. Vote</th>
<th>Dudoit</th>
<th>NSC</th>
<th>SCRDA</th>
<th>W. Vote</th>
<th>Dudoit</th>
<th>NSC</th>
<th>SCRDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>Genes</td>
<td>370</td>
<td>150</td>
<td>185</td>
<td>543</td>
<td>543</td>
<td>543</td>
<td>543</td>
<td>543</td>
<td>543</td>
</tr>
</tbody>
</table>

Table 3.6: Comparison of Classification Performance on ALL-7 Data.

The confusion matrix, as described in the Section 5.2 of Chapter 5, for the different methods are shown in Table 3.7 and in Table 3.8. It is the matrix of number of samples classified by the method, and shows explicitly how many samples are misclassified and in which class they were assigned by the classifier. For the ideal classifier, the matrix has all the non-diagonal elements are zero. The sum of the all the elements of a confusion matrix is the number of samples used for classification.

<table>
<thead>
<tr>
<th>Class</th>
<th>MLL Data</th>
<th>Golub Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted</td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td>ALL</td>
<td>AML</td>
</tr>
<tr>
<td>True</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>AML</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>MLL</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.7: Confusion matrix for the MLL-3 training data by Dudoit method, and Golub Test Data by NSC Method.
In the Table 3.7, we see that the NSC method misclassifies 2 samples out of 57 samples in the MLL data with 3 classes. There is no error in the ALL samples. But, one AML sample was misclassified as MLL sample and one MLL sample was misclassified as AML sample. On the right side of the table we have classified the Golub data using the NSC method. There we see that one ALL sample was misclassified as AML sample.

<table>
<thead>
<tr>
<th>Class</th>
<th>BCR</th>
<th>E2A</th>
<th>HYP</th>
<th>MLL</th>
<th>T.ALL</th>
<th>TEL.AML</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCR</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E2A</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HYP</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>MLL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T.ALL</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TEL.AML</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Others</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 3.8: Confusion matrix for the ALL-7 test data by SCRDA method.

Similarly from Table 3.8, we see that while the SCRDA method was used for the classification, 4 BCR samples were misclassified as others, 3 HYP samples were misclassified as others, 3 TEL.AML samples were misclassified as others and one sample which was on the others class was misclassified as BCR.

3.17 Conclusion

In this chapter we have used the genes selected by the proposed Behrens-Fisher distribution for the classification of samples. For the classification, we have used the Weighted Vote, Multiclass, Neares Shrunk Centroid and Shrunk Centroid Regularized Discriminant Analysis for classifying binary and multiclass classification as seen on Sections 3.14, 3.15, 3.16. We have also shown the classification performance of classifiers using the confusion matrix in Section 3.16.
Chapter 4

Selection of Differentially Expressed Genes in Correlated Statistics

4.1 Summary

When testing the large number of hypotheses in the microarray data simultaneously, one need to assess the evidence against the global null hypothesis that none of the hypotheses are false [11]. Such evidence is typically based on the test statistics obtained from the data using, for example, $t$-test in the control and experimental samples. For the large number of hypotheses to be tested, the reliability of the outcome depends on the correlation between the corresponding test statistics as seen in Section 4.3. This can severely affect the accuracy of the decision in the applications of popular multiple testing methods such as false discovery rate as seen in Section 4.4 which does not require independence of the test statistics. In this chapter, the effect of the correlated test statistics in the spread of the histogram of standardized test statistics are discussed and the effect of the correlation is analyzed using the real and simulated data.

4.2 Introduction

Let us consider that we have the gene expression $x_{ij}$ microarrays from two conditions: control ($m$ samples) and experimental ($n$ samples). In each sample there are $p$ genes. The test statistics from these two microarray experiment can be obtained using the Behrens-Fisher (BF) Statistics as described in Chapter 3. For each gene $i$, the corresponding BF statistic is denoted by $t_i$. For easier analysis, these statistics
are converted into the standard normal score

\[ z_i = \Phi^{-1}(G_0(t_i)) \]  (4.2.1)

where, \( G_0 \) is a believed null cumulative distribution function (cdf) of the \( t_i \) values and \( \Phi \) is the standard normal cdf. Since microarray experiments usually presupposes that most of the genes are not differentially expressed between two conditions, we expect that the \( N(0, 1) \) density curve fits the center of the histogram of \( z_i \) values. If there is no correlation among the test statistics, \( t_i \), then after the transformation we expect that the \( z_i \) values follow standard normal density. If the correlation among \( t_i \) is present, then it has effect on the widening or narrowing the standard normal density curve of the \( z_i \) values. This has serious effect on the tail counts, which are the differentially expressed genes. Because this is the data driven method, it has no other assumption except that the data are independent and identically distributed and the test statistics are independent.

4.3 Effects of Correlations

To know the effect of correlation, we discretise the transformed test statistics, \( z_i \) by partitioning into bins or intervals. Generally the number of intervals, \( K \), is chosen by \( K = 1 + 3.3 \log(n) \). Let \( y_k \) = no. of \( z_i \) in the \( k \)th bin, \( k = 1, 2, ..., K \). This makes the partition of the sample space, \( Z \), obtained by the \( z \)-scores into \( K \) bins: \( Z_1, Z_2, ..., Z_K \) such that \( Z = Z_1 \cup Z_2 \cup ... \cup Z_K \) and the center point of the \( k \)th bin, \( z[k] \). Defining

\[ \pi_k(i) = \Pr(z_i \in Z_k), \quad \pi_k = \sum_{i=1}^{p} \frac{\pi_k(i)}{p} \]  (4.3.2)

and

\[ \gamma_{kl}(i, j) = \Pr(z_i \in Z_k \text{ and } z_j \in Z_l), \quad \gamma_{kl} = \sum_{i \neq j} \frac{\gamma_{kl}(i, j)}{p(p - 1)} \]  (4.3.3)

all the \( \pi_k(i) \) values are determined by the standard normal density, \( \phi(z) \), around the point \( z[k] \). The following result was proved by Efron [11].
**Lemma**: Let \( y = (y_1, y_2, ..., y_K) \). Then

\[
\text{Cov}(y) = C_0 + C_1 \tag{4.3.4}
\]

where, \( C_0 \) is the multinomial covariance matrix that would apply if the \( z \)-values were independent,

\[
C_0 = \text{diag}(\nu) - \nu \nu' / p = p[\text{diag}(\pi.) - \pi. \pi'] \tag{4.3.5}
\]

and

\[
C_1 = \left( 1 - \frac{1}{p} \right) \text{diag}(\nu) \delta \text{ diag}(\nu) \quad \text{with} \quad \delta_{kl} = \gamma_{kl} / \pi_k \pi_l - 1 \tag{4.3.6}
\]

where, \( \text{"diag"} \) represents \( p \times p \) diagonal matrix \( \delta \) whose elements are \( \delta_{kl} \) and \( \pi = (\pi_1, \pi_2, ..., \pi_K)' \).

**Proof**: 

When \( k \neq l \), then

\[
E(y_k y_l) = \sum_{i \neq j} \gamma_{kl}(i, j) = p(p - 1) \gamma_{kl} \tag{4.3.7}
\]

and,

\[
\text{Cov}(y_k, y_l) = E(y_k y_l) - E(y_k)E(y_l) = p(p - 1) \gamma_{kl} - p^2 \pi_k \pi_l. \tag{4.3.8}
\]

and,

\[
\text{Var}(y_k) = \text{Cov}(y_k, y_k) = p(\pi_k - \pi_k^2) + p(p - 1)(\gamma_{kk} - \pi_k^2) \tag{4.3.9}
\]
where, $\gamma_{kk}(i, j)$ is given by

$$
\gamma_{kl}(i, j) = \frac{\Delta^2}{2\pi \sqrt{1 - \rho^2}} \exp \left[ -\frac{1}{2} \frac{z[k]^2 - 2\rho_{ij}z[k]z[l] + z[l]^2}{1 - \rho^2} \right] (4.3.10)
$$

From the above lemma, if $z_i$ and $z_j$ are independent, then $\gamma_{kl}(i, j) = \pi_k(i)\pi_l(j)$. This means all the elements of the $\delta$ in (4.3.6) are zero, thus $C_1 = 0$. So, $Cov(y) = C_0$ if the statistics are not correlated. Furthermore, the amount of correlation between the $t$-scores determines the size of $\delta$ and increases the $Cov(y)$ above $C_0$. The elements $\delta_{kl} \delta$ can be estimated by

$$
\delta_{kl} = \int_{-1}^{1} \left[ \frac{1}{\sqrt{1 - \rho^2}} \exp \left( \frac{\rho}{2(1 - \rho^2)} \{2z[k]z[l] - \rho(z[k]^2 + z[l]^2)\} \right) - 1 \right] g(\rho) d\rho (4.3.11)
$$

where, $\rho$ is the correlation matrix and $g(\rho)$ is the empirical density of $p(p-1)/2$ correlations, $\rho_{ij}$, of genes.

Let us define

$$
Y_0 = \text{No. of } \{z_i \in [-a, a]\} \quad \text{and} \quad Y_1 = \text{No. of } \{z_i \geq b\} \quad (4.3.12)
$$

where, $a$ and $b$ are suitable positive numbers (cutoffs) which covers the central and tail part of the density of $z_i$. We will analyze the effect of correlations of $z_i$ on the $sd(Y_0)$ and $sd(Y_1)$; and on the $Cor(Y_0, Y_1)$ in the application section.

We will use the above lemma to find the $sd(Y_0)$ and $sd(Y_1)$, $Cor(Y_0, Y_1)$.

4.4 Application

The golub data described in the chapter 3 was used to test the correlation of genes and its effects on testing simultaneous multiple hypotheses. Because of thousands of hypotheses are tested simultaneously, there is a greater chance of correlation among
the test statistics. To know whether the statistics are correlated or not, the first visual method is to transform the test statistics using the equation (4.2.1) and see whether the transformed standard z - scores are actually fit the standard normal curve. In the golub data, the histogram was produced for the transformed standard z - values. The Behrens Fisher (BF) test statistics was obtained by the proposed method Section 3.10 of Chapter 3. The BF test statistics were transformed into z-scores. The distribution of the Behrens Fisher statistic was found to fit the Cauchy distribution with mean = −0.111142 and standard deviation = 1.733, Kolmogorov’s Goodness of Fit test = 0.0533, with rank 1. This means that Cauchy distribution fitted best to the BF test statistics. The cumulative distribution function G in equation (4.2.1) is the cumulative distribution function of Cauchy distribution with mean = −0.111142 and standard deviation = 1.733. The table for the goodness-of-fit test was given in the Table 4.1.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Kolmogorov-Smirnov Statistic</th>
<th>Rank</th>
<th>Anderson-Darling Statistic</th>
<th>Rank</th>
<th>Chi-Square Statistic</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cauchy</td>
<td>0.05333</td>
<td>1</td>
<td>12.205</td>
<td>1</td>
<td>117.45</td>
<td>3</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>0.05534</td>
<td>2</td>
<td>12.285</td>
<td>2</td>
<td>81.852</td>
<td>1</td>
</tr>
<tr>
<td>Johnson (SU)</td>
<td>0.06687</td>
<td>3</td>
<td>14.46</td>
<td>3</td>
<td>98.901</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 4.1: Goodness-of-Fit test of the test-statistics obtained by the BF method for the Golub Data

The density of the correlations is given in the Figure 4.3. It is approximately normal, \( N(0, 0.26^2) \).

Let us take, in particular, \( a = 0.75 \) and \( b = 3.0 \) in equation (4.3.12) as suggested by Efron, 2007. Since out of all the genes the probability to lie in the interval of \([−0.75, 0.75]\) can be considered as success, we model that \( Y_0 \) is approximated by the binomial distribution with \( n = \text{No.} \) of genes, and probability of \( Y_0 \) lying in the \([−0.75, 0.75] = \Phi(0.75) − \Phi(−0.75) \). Similarly, \( Y_1 \) is approximated by the binomial with probability of \( Y_1 \) lying in \([3, \infty] = 1 − \Phi(3) \). The covariance between the genes is obtained by using the Lemma 1,
Histogram of Test Statistics

Figure 4.1: Histogram and the fitted Cauchy density curve for the BF test statistics
Figure 4.2: Histogram of transformed test-statistics
Figure 4.3: Density, $g(\rho)$, of correlations between the genes
\[
Cov(Y_0, Y_1) = E(Y_0 Y_1) - E(Y_0)E(Y_1)
\]
\[
= p(p - 1)\gamma_{kl} - p^2\pi_k\pi_l.
\]  
(4.4.13)

where, \(\gamma_{kl}\) is given by (4.3.3) and \(\pi_k\) and \(\pi_l\) are given by (4.3.2) with suitable modification to cover the intervals. In particular, if \(z_i\) and \(z_j\) are independent, then \(\gamma_{kl} = \pi_k\pi_l = [\Phi(0.75) - \Phi(-0.75)][1 - \Phi(3)]\). There are \(p = 1761\) genes remaining after preprocessing, normalization and transforming into normality.

\[
\text{sd}(Y_0) = \sqrt{p[\Phi(0.75) - \Phi(-0.75)][1 - (\Phi(0.75) - \Phi(-0.75))]}
\]
\[
= \sqrt{1761(0.5467453)(1 - 0.5467453)}
\]
\[
= 20.89
\]  
(4.4.14)

\[
\text{sd}(Y_1) = \sqrt{p[1 - \Phi(3)]\Phi(3)}
\]
\[
= \sqrt{1761(1 - 0.9986501)(0.9986501)}
\]
\[
= 1.54
\]  
(4.4.15)

\[
Cor(Y_0, Y_1) = \frac{Cov(Y_0, Y_1)}{\text{sd}(Y_0)\cdot\text{sd}(Y_1)}
\]
\[
= \frac{p(p - 1)[\Phi(0.75) - \Phi(-0.75)][1 - \Phi(3)] - p^2[\Phi(0.75) - \Phi(-0.75)][1 - \Phi(3)]}{(20.89)(1.54)}
\]
\[
= \frac{1761(1760)(0.5467453)(1 - 0.9986501) - (1761)^2(0.5467453)(1 - 0.9986501)}{(20.89)(1.54)}
\]
\[
= \frac{-1.2997}{32.1706} = -0.04
\]  
(4.4.16)

Now, let the transformed statistics \(z_i\) and \(z_j\) are correlated. Then the average of all correlations between the genes is 0 as seen from the Figure 4.3. Using the results of above Lemma for the variable \(Y_0\), we get
4.5 Computation of Effect of Correlations

The following Table 4.2 summarizes the effect of correlation on the spread of the histogram. Because of the correlation among the test statistics in multiple testing the standard error multiplies by the several fold. We can see that the standard deviations of the transformed test statistics are multiplied by the several folds when we consider the correlation among the test statistics as compared to when they are independent. Furthermore, the correlation coefficient, which is almost negligible when the test statistics are assumed independent, has tended toward large negative value.
<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th>Correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{sd}(Y_0)$</td>
<td>20.89</td>
<td>217.56</td>
</tr>
<tr>
<td>$\text{sd}(Y_1)$</td>
<td>1.54</td>
<td>21.54</td>
</tr>
<tr>
<td>$\text{cor}(Y_0, Y_1)$</td>
<td>-0.04</td>
<td>-0.488</td>
</tr>
</tbody>
</table>

Table 4.2: Effect of Correlation on the standard deviations of the middle and tail count of the transformed test statistics

### 4.6 Conditional and Unconditional p - values

One of the ways [37] to alleviate the problem arising from the correlated test statistics is to condition the central portion $Y_0$ of the normalized test statistics, $z_i$. If $X = \max_i |z_i|$, $i = 1, 2, ..., p$, and the $p$ statistics, $z_i$, were independent, then the overall $p$-value associated with the observed value $X = x$ is given by

$$p_x = 1 - [1 - 2\Phi(-x)]^p = 2p\Phi(-x)$$  \hspace{1cm} (4.6.20)

where, $p$ is the number of hypotheses to be tested. Because the shape of the histogram of $z$ values varies with the correlated test statistics than it does for the independent test statistics, the $p$-value to test the null hypotheses should be smaller than the actual $p$-value obtained from the assumption of independent test statistics. The larger variation among the histograms of correlated test statistics can cause misleading tail counts and thus more genes are declared to be differentially expressed than actual number of differentially expressed genes.

The conditional $p$-value is estimated by permuting the samples and calculating the number of $z_i$'s that lie inside a small interval $(C - \delta, C + \delta)$ around $C$, where,

$$C = \frac{\#(z_i : |z_i| < 1)}{p}$$  \hspace{1cm} (4.6.21)

In particular, we find the values of $X = \max_i |z_i|$, $i = 1, 2, ..., p$, say $x$, and $C$ from the original non-permuted samples. Permutation of samples is repeated many times, say $B$. In each permutation $b$, the following are calculated: (1) the proportion of genes, $C_b$, lying within $(C - 0.05, C + 0.05)$ (2) $Y_b = 1$ if $x_b > x$ and $Y_b = 0$
otherwise (3) coefficients $\hat{\alpha}$ and $\hat{\beta}$ by fitting the logistic regression, $\log \left( \frac{P(Y_b=1|C_b)}{1-P(Y_b=1|C_b)} \right) = \exp(\alpha + \beta C_b)$. Finally, the estimate of conditioned $p$-value at the point $C = c$ is given by

$$\hat{p}_x(c) = \frac{\exp(\hat{\alpha} + \hat{\beta}c)}{1 + \exp(\hat{\alpha} + \hat{\beta}c)} \quad (4.6.22)$$

Similarly, the unconditional $p$-value, $P_x$, is estimated as the proportion of all permutations whose maximum test statistic exceeds the observed value $x$. Those genes with $p$-value less than or equal to the adjusted $p$-value are taken as the differentially expressed genes. In the Golub data, the permutation was repeated 5000 times to calculate conditional $p$-value and unconditional $p$-value.

$$C = \frac{\#(z_i: |z_i| < 1)}{p} = 0.28393$$
$$X = \max_i |z_i| = 2.078$$
$$E(C) = \Phi(1) - \Phi(-1) = 0.6827$$
$$Var(C) = 0.00016$$

Unconditional $p$-value, $\hat{P}_x = 0.0286$ [for 5000 permutations]
Conditional $p$-value, $\hat{p}_x(c) = 0.0577$ [for 5000 permutations]

Table 4.3: Effect of correlation on $p$-value on test statistics in Golub data

In the Table 4.3, we see that the $X = \max_i |z_i| = 2.078$ and $C = 0.28393$. Since this value of $C$ is smaller than the $E(C) = 0.6827$, the histogram of test statistics is more wider than expected under the assumption of independence. The magnitude of difference between these two conditional and unconditional $p$-value is $|0.0286 - 0.0577| = 0.0291$. Hence the $p$-values should be adjusted accordingly. Table 4.2 gives some insight about the actual effect of correlations. The correlation among the genes increases the standard deviation of $Y_0$ by almost 10-fold, the standard deviation of $Y_1$ by almost 14-fold, and the correlation by almost 12-fold toward the negative side.

### 4.7 Dispersion Parameter

The central peak of the histogram depends on the dispersion parameter $A$. If there is no correlation between the test statistics, then $A = 0$. Positive value of $A$ widens the central peak whereas the negative value of $A$ narrows the central peak. This is reflected in the Figure 4.2. Taking $x_0 = 2.0$ (which is almost 95% confidence interval
by the normal theory) for the cut points of the test statistics, $z_i$, one can obtain the dispersion parameter $A$ as in Efron [11] which conditions the FDR on $A$

$$\text{FDR}\{x|A\} = \frac{E\{Y(x)|A\}}{T(x)}.$$  \hspace{1cm} (4.7.23)

Let us define

$$Y_0 = \text{number of } z_i \in [-x_0, x_0]$$

$$P_0 = 2\Phi(x_0) - 1$$

$$Q_0 = 2\sqrt{2}\phi(x_0)$$

$$\hat{P}_0 = \frac{Y_0}{p}$$

$$\hat{A} = \frac{P_0 - \hat{P}_0}{Q_0}$$ \hspace{1cm} (4.7.24)

where, $\phi$ is the standard normal density function, and $\Phi$ is the standard normal cumulative distribution function. For the Golub data with $x_0 = 2$, we got the following:

$$Y_0 = 1760$$

$$P_0 = 0.9545$$

$$Q_0 = 0.1527$$

$$\hat{P}_0 = \frac{1261}{1761} = 0.716$$

$$\hat{A} = 1.561$$

In the Figure 4.2, the cut-off value of the $z$ statistics correspond to the Behrens-Fisher test statistics of $\alpha = 0.05$ is given by 1.24. But because of the correlation we need to adjust the value of $z$, so that the correlation among the test statistics affect the inference as little as possible. Hence, taking the conditional $p$-value for the transformed $z_i$ data, the cut-off values is 1.574. Hence, we declare those genes as differentially expressed if the $|z_i| < 1.574$. Using this as a cut-off value, we got $32 + 64 = 96$ genes as differentially expressed. Hence, there are 96 genes that are
found to be differentially expressed after taking into account of correlation among the test statistics.

4.8 Conclusion

In this chapter we have studied the effect of correlation on the test statistics in Section 4.5. The effect of the correlated test statistics on the selection of genes have been studied in Section 4.7. The cut-off of z-values should be adjusted so that the effect of correlation is as small as possible. Furthermore, a method of how to find the cutoff using the p-value obtained by conditioning the central part of the histogram to account for the correlation has been studied in Section 4.6. So, the possible interval in which we accept the null hypotheses have been determined.
Chapter 5

Performance of a Classifier taking into account of Correlated Genes

Summary: In this chapter, we introduce the correlation between the genes and study its effect on the performance of different classifiers. Furthermore, we have introduced how to set up the cut-off value while filtering the highly correlated genes. We have used McNemar’s test to test the hypothesis of whether the two classifiers performance are same. The assumptions for this test is the data are independent and they are classified into two classes. Since the confusion matrix is generally used for the performance of classifiers in the multiclass classification task, we have calculated confusion matrices for different data and we have used these matrices for measuring the performance of classifiers.

5.1 Performance of Classifiers

Despite one’s best efforts to remove the misclassification classifier does not perform well, particularly with the new data. Understanding the potential causes of poor performance can be useful guide to possible alterations that might improve performance. Classifiers fail to make sufficiently accurate predictions in a number of circumstances.

1. The form of classifier is too complex and over-fits the data. This tends to arise when the ratio of parameter to cases exceeds some desirable limit and the classifier begins to fit the random noise in the training data. This will lead to poor generalization. This is why simple models often outperform complex models on novel data.
2. The form of classifier is too simple or has an inappropriate structure. For example, classes may not be linearly separable or important predictors have been excluded. This will reduce accuracy.

3. If some of the training class labels are incorrect or 'fuzzy' there will be problems. Most classifiers assume that class membership in the training data is known without error. Obviously if class definitions are ambiguous, it becomes more difficult to apply a classification procedure, and any measure of accuracy is likely to be compromised. More importantly, mislabeled cases can influence the structure of the classifier, leading to bias or classifier degradation. For example, a misclassified sample may be an outlier for some predictors, but its influence will depend on the classifier. In discriminant analysis outliers may have large effects because of their contribution to the covariance matrix.

4. Training samples may be unrepresentative. If they are, this leads to bias and poor performance when the classifier is applied to new samples. Careful sampling designs should avoid this problem but such bias may be unavoidable if there is significant unrecognised regional and temporal variability.

5. Unequal class-sizes in training and test sample also may affect the classifier’s performance.

5.2 McNemar’s Test and Confusion Matrix

This test is generally used to determine the significance of the two methods used for the classification. So, it is used to access the classification performance of binary classifiers. The entries in the Table 5.1 are numbers of misclassifications. In $e_{ij}$, the subscript $i$ represents classification by method I and subscript $j$ represents the classification by method II. Furthermore, if the subscript is 0 then correct classification, and if it is 1 then misclassification. Under the null hypothesis that the classifiers have same error rates, the McNemar’s test statistic, $M$, is given by

$$M = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \sim \chi^2(1)$$  (5.2.1)
When we classify the samples in two or more than two classes, there are some error since no classifier is perfect. The classification performance of the classifier can be easily seen from the confusion matrix, see Table 5.2. The confusion matrix is a matrix in which the entries are the no. of samples misclassified. The columns represent the true class of the samples and the rows represent the predicted class. Thus it is easy to see that how many samples are misclassified and the assignment of the misclassified samples in the classes.

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Pred</td>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>c</td>
</tr>
</tbody>
</table>

Table 5.2: Confusion Matrix for misclassification

5.3 Filtering the Highly Correlated Genes

For a subset $S$ of genes obtained from the Behrens-Fisher Distribution, the discrimination ability and the correlation of genes can be found by computing the score of each of the genes as in Liu et al. [27]. Let $g$ be any gene in $S$. If there are 2-classes, $G = 2$ with class labels $k = 1$ and 2, we determine the score of each of the genes. Let $\mu_k^g$ be the mean and $\sigma_k^g$ be the standard deviation of gene $g$ in class $k$. the discriminant ability of the gene $g$ is described by [27]

$$D_g = \frac{\left| \mu_1^g - \mu_2^g \right|}{\sigma_1^g + \sigma_2^g}$$  \hspace{1cm} (5.3.2)

From (5.3.2), it can be seen that the larger the $D_g$, the further the classes are separated. Therefore, the better the discriminant ability the gene $g$ has. The correlation
between the genes $g$ and $h$ can be calculated by

$$r_{g,h} = \prod_{k=1}^{2} \frac{Cov[S_k(g), S_k(h)]}{\sqrt{Var(S_k(g)) \cdot Var(S_k(h))}}$$  \hspace{1cm} (5.3.3)$$

where, $S_k(g) = (x_{g1}, x_{g2}, ..., x_{gk})$ is the vector of components of gene $g$ in the class $k$ only. Based on $r_{g,h}$, the correlation coefficient between a single gene $g$ and the subset $S$ is determined as

$$r_{g,S} = \max_{h \in S} |r_{g,h}|.$$ \hspace{1cm} (5.3.4)

A high value of $r_{g,S}$ indicates that $g$ is highly correlated with certain gene in $S$, and therefore it carries the redundant information.

Finally, it is desirable to select the genes that can individually separate the classes well and has small correlation with the genes in the subset $S$. The final score assigned to each of the gene $g$ is defined as

$$R_{g,S} = \frac{D_g}{\max_{g \in S} D_g} - |r_{g,S}|.$$ \hspace{1cm} (5.3.5)

where, $D_g$ is normalized such that it is in the same range as $|r_{g,S}|$. So, the final score given by (5.3.5) depends not only on feature $g$ but depends on the discriminant ability $D_g$ and correlation score $r_{g,S}$. We take only those genes for which the final absolute correlation score, $R$, is greater than 0.5.

### 5.4 Classification using SVM

The support vector machine (SVM) is one of the methods that has been successfully applied to the cancer diagnosis problem in previous studies (Lee and Lee, 2002; Mukherjee et al.[28]). Let $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n)\}$ be the training samples, where $y_i = +1$ or $-1$ according as $x_i$ is in class 1 or class 2. The points $x_i$, $i = 1, 2, ..., n$ are $p$-dimensional points. This data set is called the *training set*. We are interested whether we can separate these $d$-dimensional points $x_i$ by a $(p-1)$-dimensional hyper-plane, just as points in a plane by a straight line on the plane.
This is a form of linear classifier. Although there are many linear classifiers but there is just one which maximizes the separation between the points of two classes. Such a hyper-plane is known as the maximum margin hyper-plane (optimal hyper-plane) and such a linear classifier is known as a maximum margin classifier.

In two-class classification, the linear SVM fits a model

\[ f(x) = b_0 + \sum_{j=1}^{p} b_j x_j \]

that minimizes

\[ \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \frac{\lambda}{2} (||b_1||^2 + \ldots + ||b_p||^2) \]  

The classification decision is then made according to \( \text{sign}[f(x)] \). The weakness of the SVM is that it only estimates \( \text{sign}[p(x) - \frac{1}{2}] \), where the probability \( p(x) = P(C = 1|X = x) \) is the conditional probability of a point being in class 1 given that \( X = x \).

In multi-class classification, the one-vs-rest scheme is often used: given \( K \) classes, the problem is divided into a series of \( K \) one-vs-rest problems, and each one-vs-rest problem is addressed by a different class-specific SVM classifier (e.g. class 1 vs. not class 1); then a new sample takes the class of the classifier with the largest real valued output \( c = \arg\max_{k=1,2,\ldots,K} f_k \), where \( f_k \) is the real valued output of the \( k \)th SVM classifier. Recently, Lee and Lee (2002) extends the two-class SVM to the multiclass case in a more direct way and estimates \( \arg\max_{k} P(C = k|X = x) \) directly, but it still lacks the estimates of \( P(C = k|X = x) \) themselves.

5.5 Application

Binary Classification

In this section we will apply the logistic regression method for the binary classification of the Golub data. The genes were selected by using the Behrens-Fisher method as proposed in chapter 2. The logistic discrimination highly suffered from the high-dimension of the data, so we could not perform the logistic discrimination using all
the genes selected by the BF method. Instead, we filtered the highly correlated genes as in Liu [27]. The number of genes selected was reduced in number from 639 when not taking the correlation among the genes to 25 after taking the correlation among the genes into account.

We have used Weighted Vote method, Support Vector Machines and the Multi-class method of Dudoit for assigning the class labels of the Golub data. In each of these methods, we used the same 25 genes obtained as above after filtering the highly correlated genes. The result has been shown in the Table 5.3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Errors</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>Total</td>
<td>% Error</td>
</tr>
<tr>
<td></td>
<td>ALL</td>
<td>AML</td>
<td>ALL</td>
<td>AML</td>
</tr>
<tr>
<td>Wt. Vote</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SVM</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Multiclass</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.3: Classification of Golub Data taking into Correlation among the genes

From the above table we see that, in the case of two class classification, the filtering of highly correlated genes shows the better performance than its non-filtering of correlated genes counterpart. The above classification result is almost similar to the method used by [55] in which they have found that 3 samples misclassified using the 22 genes.

**Multiclass classification using SVM**

This section discusses the application of the support vector machines in classifying the samples from real leukemia data. The data is taken from Yeoh et. al (2002). There are 12,625 genes and 7 classes. The total number of training samples are 215 and test samples are 112. Hence the data is 12625 \( \times \) 327 matrix. After pre-processing steps of the training samples as mentioned in Dudoit et al [10], 11,960 genes are remaining. We selected those genes in the new test samples which were remaining after the pre-processing steps. We have used these remaining genes for determining the marker genes that are helpful for classifying the one kind of sample versus the rest.
samples.

In all the genes selection process, I have used the Behrens-Fisher method to select the differentially expressed genes between one kind of sample versus the rest samples. In the multiclass classification, there are two approaches:

- One versus All Method
- One by One Method

The idea of binary classification rule can be extended to multiclass classification. In the one versus all method, the samples are classified by the classifier in the two classes. Let there are $C$ classes. We construct $C$ binary classifiers. The $k$th classifier is trained to classify the $k$th class from the remaining classes. If any sample is classified in two or more than two classes, then the majority of vote is taken and assign the class of that sample. In the One by One method, we construct the all possible binary classifiers. as in the One versus All method, if some samples are classified in more than two classes, the majority of vote is taken and assign the class of that sample.

<table>
<thead>
<tr>
<th>Quantiles</th>
<th>Score($R$)</th>
<th>No. of misclassified Samples(train, test)</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>-0.1389128</td>
<td>0.2</td>
<td>639</td>
</tr>
<tr>
<td>55%</td>
<td>-0.0628013</td>
<td>0.2</td>
<td>639</td>
</tr>
<tr>
<td>60%</td>
<td>0.01331021</td>
<td>0.2</td>
<td>625</td>
</tr>
<tr>
<td>65%</td>
<td>0.089421</td>
<td>0.2</td>
<td>526</td>
</tr>
<tr>
<td>70%</td>
<td>0.165532</td>
<td>0.2</td>
<td>418</td>
</tr>
<tr>
<td>75%</td>
<td>0.2416447</td>
<td>0.2</td>
<td>300</td>
</tr>
<tr>
<td>80%</td>
<td>0.317756</td>
<td>0.2</td>
<td>161</td>
</tr>
<tr>
<td>85%</td>
<td>0.393867</td>
<td>0.2</td>
<td>89</td>
</tr>
<tr>
<td>87%</td>
<td>0.425267</td>
<td>0.2</td>
<td>47</td>
</tr>
<tr>
<td>88%</td>
<td>0.431419</td>
<td>0.2</td>
<td>32</td>
</tr>
<tr>
<td>89%</td>
<td>0.458178</td>
<td>0.2</td>
<td>27</td>
</tr>
<tr>
<td>90%</td>
<td>0.469979</td>
<td>0.2</td>
<td>25</td>
</tr>
<tr>
<td>91%</td>
<td>0.514972</td>
<td>1.2</td>
<td>23</td>
</tr>
<tr>
<td>95%</td>
<td>0.5460907</td>
<td>2.3</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 5.4: Comparison of Classification Performance on Golub Data after taking into correlation between the genes. Different percentage points were used to determine the optimal cut-point.
When taking account of correlation between the genes, we select those genes which have high discriminating power and that have no redundancy information due to correlation. So, we will determine the score of these genes. From the Table 5.4 we see that when we choose the score above 90%, then the classifiers has the best performance. So, we take 90% score as the threshold for determining the number of genes in the correlated data for classification. It has also seen from the Table 5.4 that the larger the value of discrimination score, $R$, the better the performance. The discrimination power increases until it reaches some value then begin to decrease. All the genes selected by the proposed BF method also does not have much discriminating power, unless we take the correlation of the genes into account. These selected genes before taking correlation is just adding the noise for the classifiers. The percentage points are calculated by dividing the difference of maximum and minimum final scores into equal 100 parts.

When selecting the genes, I selected the genes using the proposed Beherns Fisher (BF) Statistic. I took one training sample from the rest of the training samples and selected the genes controlling the FDR at the level 0.05. Then, I took account of the correlation between the selected genes as in Liu et al. [27]. This procedure was repeated for each of the 7 classes. Finally, the common genes selected between BCR vs. rest and E2A vs. rest was used to classify BCR test samples with the E2A test samples, and so on.

<table>
<thead>
<tr>
<th>Classes</th>
<th>No. of genes</th>
<th>No. of Samples</th>
<th>No. of Cost</th>
<th>No. of Mis-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Train</td>
<td>Test</td>
<td>SV</td>
<td>classifications</td>
</tr>
<tr>
<td>BCR vs. E2A</td>
<td>6</td>
<td>18,9</td>
<td>9,6</td>
<td>5</td>
</tr>
<tr>
<td>E2A vs. HYP</td>
<td>12</td>
<td>42,18</td>
<td>22,9</td>
<td>4</td>
</tr>
<tr>
<td>HYP vs. MLL</td>
<td>2</td>
<td>14,42</td>
<td>6,22</td>
<td>5</td>
</tr>
<tr>
<td>MLL vs. T.ALL</td>
<td>7</td>
<td>28,14</td>
<td>15,6</td>
<td>5</td>
</tr>
<tr>
<td>T.ALL vs. TEL.AML</td>
<td>10</td>
<td>28,52</td>
<td>15,27</td>
<td>5</td>
</tr>
<tr>
<td>TEL.AML vs. Others</td>
<td>9</td>
<td>52,52</td>
<td>27,27</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 5.5: Comparison of Classification Performance on ALL-7 Data using the genes Selected by the BF method and taking into correlation between the genes. The classes were compared using one versus another.
<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Weighted Vote</th>
<th>SVM</th>
<th>Dudoit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True class</td>
<td>True class</td>
<td>True class</td>
</tr>
<tr>
<td>ALL</td>
<td>ALL 19</td>
<td>ALL 19</td>
<td>ALL 20</td>
</tr>
<tr>
<td>AML</td>
<td>AML 1</td>
<td>AML 2</td>
<td>AML 1</td>
</tr>
</tbody>
</table>

Table 5.6: Confusion Matrices of Test samples of Golub Data by Weighted Vote, SVM and Dudoit methods

<table>
<thead>
<tr>
<th>Pred. Class</th>
<th>Training Sample</th>
<th>Test Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Class</td>
<td>True Class</td>
</tr>
<tr>
<td>ALL</td>
<td>ALL 20</td>
<td>ALL 2</td>
</tr>
<tr>
<td>MLL</td>
<td>MLL 3</td>
<td>MLL 1</td>
</tr>
<tr>
<td>AML</td>
<td>AML 0</td>
<td>AML 8</td>
</tr>
</tbody>
</table>

Table 5.7: Confusion Matrix of MLL data applying the Dudioit Classification method after taking correlation

### 5.6 Confusion Matrices

The confusion matrices for the Golub data is shown in the Table 5.6. In the table, only the confusion matrices of test samples have been shown. The classification methods made no error in the training samples. Each method used the same set of 25 genes for the classification. These genes were selected after the taking account of correlation. The classification performance was improved comparing to before correlation, at least, in terms of cost. In the previous chapter 3, the number of genes used for the classification was 639, but here is just 25 genes. This mens that if we take just these 25 genes, then the classification performance is almost equal to that of taking 639 genes.

The following Table 5.7 shows the confusion matrix for the MLL 3 classes data as described in Chapter 2.

From the tables of confusion matrices, we see that the classification performance of the classifiers is greatly increased even with small number of genes. In each of the methods, the genes used for classification is less than 10. So, in terms of cost, the methods are suitable for classification.

Now we are going to calculate the assessments of different classifiers. If two clas-
Table 5.8: Confusion Matrix of MLL data applying the Weighted Vote Classification method after taking correlation

<table>
<thead>
<tr>
<th>Pred. Class</th>
<th>Training Sample</th>
<th>Test Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Class</td>
<td>True Class</td>
</tr>
<tr>
<td>ALL</td>
<td>ALL  20 MLL 1 AML 0</td>
<td>ALL 3 MLL 1 AML 0</td>
</tr>
<tr>
<td>MLL</td>
<td>0 15 0</td>
<td>1 2 0</td>
</tr>
<tr>
<td>AML</td>
<td>0 1 20</td>
<td>0 0 8</td>
</tr>
</tbody>
</table>

Table 5.9: Confusion Matrix of MLL data applying the SVM Classification method after taking correlation

<table>
<thead>
<tr>
<th>Pred. Class</th>
<th>Training Sample</th>
<th>Test Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True Class</td>
<td>True Class</td>
</tr>
<tr>
<td>ALL</td>
<td>ALL 19 MLL 1 AML 1</td>
<td>ALL 3 MLL 1 AML 0</td>
</tr>
<tr>
<td>MLL</td>
<td>0 12 0</td>
<td>0 1 0</td>
</tr>
<tr>
<td>AML</td>
<td>1 4 9</td>
<td>1 1 8</td>
</tr>
</tbody>
</table>

Classifiers are used to classify the test samples after they were trained on the same training sets, the number of misclassifications by each methods (called errors) are calculated and put as in the contingency table. For comparing the performance of Dudoit multi-class method and Support Vector Machines method on the classification of Golub test samples, we get the contingency table 5.10. The McNemars statistic, $M = 0.5 < \chi^2(1), 0.05 = 3.84$. So, we conclude that the two methods have same error rates.

Table 5.10: Contingency table for misclassification of Golub Test Data

<table>
<thead>
<tr>
<th></th>
<th>Dudoit Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>1</td>
</tr>
<tr>
<td>Method</td>
<td>0</td>
</tr>
</tbody>
</table>

Similarly the performance of the classifiers on the MLL data can be determined from the following Contingency Table 5.11. The McNemars statistic $M = 0 < \chi^2(1), 0.05 = 3.84$. So, we conclude that these two methods- Wt. Vote and SVM Method, have same error rates.

The performance of classifiers are given in Tables 3.8 and Table 3.7 of Chapter 3. In the chapter 3, we had trained classifiers using leave one out method.
Table 5.11: Contingency table for misclassification of MLL Test Data

<table>
<thead>
<tr>
<th>Method</th>
<th>SVM</th>
<th>Wt. Vote</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

5.7 Binary Accuracy Measures

Different measures of accuracy of binary classifiers are obtained from the confusion matrix of Table 5.2. The various measure of prediction accuracy from the confusion matrix of binary classifier are given in the Table 5.12:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Classification Rate (CCR)</td>
<td>$(a + d)/N$</td>
</tr>
<tr>
<td>Misclassification Rate</td>
<td>$(b + c)/N$</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>$a/(a + c)$</td>
</tr>
<tr>
<td>Specificity</td>
<td>$d/(b + d)$</td>
</tr>
<tr>
<td>False positive Rate</td>
<td>$b/(b + d)$</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>$c/(a + c)$</td>
</tr>
</tbody>
</table>

Table 5.12: Confusion matrix derived accuracy measures, $N = a + b + c + d$

The accuracy measures have been calculated for the Golub data for the classifiers: Wt. Vote, SVM, Logistic Discrimination and Dudoit Multiclass methods. These are shown in the following Table 5.13.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Wt. Vote</th>
<th>SVM</th>
<th>Dudoit</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Classification Rate (CCR)</td>
<td>0.941</td>
<td>0.912</td>
<td>0.971</td>
<td>0.853</td>
</tr>
<tr>
<td>Misclassification Rate</td>
<td>0.059</td>
<td>0.089</td>
<td>0.029</td>
<td>0.147</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.95</td>
<td>0.95</td>
<td>1</td>
<td>0.85</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.929</td>
<td>0.857</td>
<td>0.928</td>
<td>0.857</td>
</tr>
<tr>
<td>False positive Rate</td>
<td>0.071</td>
<td>0.143</td>
<td>0.072</td>
<td>0.143</td>
</tr>
<tr>
<td>False Negative Rate</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 5.13: Accuracy measures for the Wt. Vote, SVM, Logistic Discrimination and Dudoit Multi-class Classifiers

From the Table 5.13, we can see that the Dudoit multi-class method has the highest correct classification rate (CCR) compared to other three classifiers. The sensitivity and specificity are the probabilities of predicting the same classes of the samples if they are actually in that class. So, those classifiers whose sensitivity and specificity alongwith the CCR are higher are the best classifiers. From the table we see that all
the classifiers perform superiorly if we take the correlation of the genes into account.

5.8 Classification using the genes selected by the Correlated test Statistics

The classification was also done using the genes selected by the correlated test statistics method. In the correlated test statistic method, we have used only those genes selected by the proposed BF method. We selected 100 genes as in chapter 4. These genes are used to classify the Golub data. The classification table is shown below:

<table>
<thead>
<tr>
<th>Method</th>
<th>Errors</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>Total</td>
<td>% Error</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL AML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL AML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wt. Vote</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SVM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Multiclass</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8.82</td>
</tr>
</tbody>
</table>

Table 5.14: Classification of Golub Data taking into Correlated test statistics

From the Table 5.14 we see that the 100 genes selected by the correlated test statistic method classifies the sample better than any other method. Although there is a training error, there is no test error for the Golub data. So, this set of genes are found extremely important for classification since it gave the best result for classification compared to all other previous methods.

5.9 Conclusion

In this chapter we have introduced the effect of correlation on the classification. We have selected the genes such that the effect of correlation is minimized and these genes have more discriminating power. The result of this has been studied from the performance of different classifiers. We have found that after taking the correlation into account, there is an improvement in discriminating samples. Instead of using all the genes, only the small number of genes were found useful for improving classification performance.
Chapter 6

Future Research

In this dissertation, we have introduced the Behrens-Fisher distribution to select the differentially expressed genes and used those genes to classify the different types of cancers. Although we have identified the expressed genes, but there may be some genes which are not statistically significant but may work in a group. My future work will be to find the cluster of genes which are functionally similar.

In the microarray data, outliers are also present because of the measurement error and other sources. The expression levels of such outliers have not been accounted. Before analyzing such a data, one has to identify and remove such an outliers. Presence of such an outlier might distort the result. So, I will work to identify the outliers and remove them in my future work.

Finally, since the dimension reduction is extremely important for correct classification of samples using the microarray data, I will work on this aspect as well in my future research.
References


About the Author

The author was born in Nepal on Saturday, August 28, 1971. He is the first child of Mrs. Shree Devi Manandhar and Mr. Jyoti Prasad Shrestha. He completed his undergraduate degree in sciences from the Tribhuvan University, Nepal, in 1993. He completed his Master’s degree in mathematics also from the Tribhuvan University, Nepal, in 1996. After that, he worked as an Assistant University Professor of Mathematics at the Institute of Engineering, Pokhara and at the Prithvi Narayan Campus, Pokhara, for 4 years. He also taught Higher Secondary Mathematics at the SOS Hermann Gmeiner School Gandaki, Pokhara, for 3 years. He came to the University of South Alabama, Mobile, Alabama for the higher study as a graduate teaching assistant in December 2002 and completed his Master’s degree in mathematics from there. He joined the University of South Florida, Tampa, Florida, as a graduate student and as a graduate teaching assistant in the Department of Mathematics and Statistics in August 2004.

The author is interested in applying the statistical knowledge in biological and genetic data, especially in microarray data analysis. He is currently working in finding the differentially expressed genes and classification of different kind of cancers using microarray data.