# Multivariate Statistical Methods for Testing a Set of Variables Between Groups with Application to Genomics

# MULTIVARIATE STATISTICAL METHODS FOR TESTING A SET OF VARIABLES BETWEEN GROUPS WITH APPLICATION TO GENOMICS

BY

HUDA ALSULAMI, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2013) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

| | |
|---|---|
| TITLE: | Multivariate Statistical Methods for Testing a Set of Variables Between Groups with Application to Genomics |
| AUTHOR: | Huda Alsulami<br>B.Sc., (Statistics/Computer Science)<br>King Abdualziz University, Saudi Arabia |
| SUPERVISOR: | Dr. Joseph Beyene |
| NUMBER OF PAGES: | x, 67 |

# Abstract

The use of traditional univariate analyses for comparing groups in high-dimensional genomic studies, such as the ordinary t-test that is typically used to compare two independent groups, might be suboptimal because of methodological challenges including multiple testing problem and failure to incorporate correlation among genes. Hence, multivariate methods are preferred for the joint analysis of a group or set of variables. These methods aim to test for differences in average values of a set of variables across groups. The variables that make the set could be determined statistically (using exploratory methods such as cluster analysis) or biologically (based on membership to known pathways).

In this thesis, the traditional One-Way Multivariate Analysis of Variance (MANOVA) method and a robustified version of MANOVA (Robustified MANOVA) are compared with respect to Type I error rates and power through a simulation study. We generated data from multivariate normal as well as multivariate gamma distributions with different parameter settings. The methods are illustrated using a real gene expression data.

In addition, we investigated a popular method known as Gene Set Enrichment

Analysis (GSEA), where sets of genes (variables) that belong to known biological pathways are considered jointly and assessed whether or not they are "enriched" with respect to their association with a disease or phenotype of interest. We applied this method to a real genotype data.

# Acknowledgements

My greatest gratitude goes to my supervisor Dr. Joseph Beyene who has helped me a lot with his expertise, encouragement and guidance during my journey in this thesis.

I would like to thank my thesis committee members: Dr. Peter Macdonald and Dr. Roman Viveros-Aguilera for their comments, questions and feedback.

I am also very grateful to King Abdulaziz University, Jeddah, Saudi Arabia for giving me the opportunity to study abroad and supporting me financially.

I also would like to acknowledge the supportive and stimulating environment of Dr. Beyene's Statistics for Integrative Genomics and Methods Advancement (SIGMA) research group. I would like to thank my fellow lab members for helpful feedback, encouragement and friendship throughout my graduate studies.

Last but not least, my parents and siblings were very supportive and I really thank them for their love and care. Also, many thanks to my lovely husband in this journey "Mohammed" who has helped me to continue my studies. His care, love and support are the keys for my success.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction and Background

A few years ago, the candidate gene approach had been widely used in genomic research and the study of complex diseases. This approach focuses on a selected gene (possible candidate gene) that may affect a disease or a phenotype and try to assess the association between this gene and a particular disease or a phenotype. The limited technologies that were used in the past did not allow researchers to study the whole genome and involve more genes in their studies. However, nowadays advance in molecular biotechnologies has played a major role in the study of genes and diseases. These technologies have enabled scientists to extract information on tens of thousands of genes simultaneously.

The completion of The Human Genome Project in 2003 helped in estimating the number of genes which are between 20,000 to 25,000 genes in the human genome. Other projects such as Genome-wide association studies (GWAS) and hapmap project

have been done to identify genetic variations that are associated with complex diseases as well as other environmental factors. These projects have enabled researchers and scientists to conduct important studies about complex diseases such as cancer, type 2 diabetes, asthma, etc (McCarthy and Zeggini, 2009; Fanale et al., 2012).

Each human cell consists of 46 chromosomes (23 pairs of chromosomes) except the gametes (the sex cells) which have only 23 chromosomes. The chromosome is made up of two DNA (Deoxyribonucleic acid) strands which carry all of the genetic information; the two strands are made up of sequences of chemical bases known as A (Adenine), T (Thymine), C (Cytosine), and G (Guanine) and the two strands are linked together through these bases; they are known as Single Nucleotide Polymorphisms (SNPs) and they are repeated in pairs along the chromosome. There are approximately 3 billion base pairs in the human genome (Francis et al., 2001). Variations in the DNA sequences between individuals affect diseases, for example, Genome-wide association studies (GWAS) have focused on common SNPs (variants) within a population to study associations with complex diseases or traits. A minor allele frequency (MAF) of a SNP is the frequency of the less common variant within a population and it also has been linked to diseases. Moreover, each chromosome is divided into segments that represent genes and those genes are distributed along the chromosomes.

Microarray technology (DNA chip) measures the expression levels of tens of thousands of genes within a cell at a time. According to the so-called the Central Dogma of Molecular Biology, DNA is transcribed into messenger RNA (mRNA), which represents gene expression, and then (mRNA) is translated into proteins that perform

the cells' functions (Crick, 1970):

$$\text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{proteins}$$

Obtaining the gene expressions using Microarrays enables scientists to track changes in gene expression levels and hence helps them to study the effect of the variations on different phenotypes or diseases of interest. A probeset on microarray is multiple probes that represent a single gene expression (Malone and Oliver, 2011).

In gene expression microarray studies, the aim is to detect significant differences in genes, i.e. which genes are differentially expressed, that are associated with a phenotype or a disease. Usually, the number of samples ($n$) is small (less than 100) while the number of variables (genes) ($p$) is very large (Efron and Tibshirani, 2007). Thus, this high dimensional data pose some statistical challenges and has led to enhanced and new statistical methods.

The traditional univariate analyses used in genomic studies test SNP by SNP associations or probeset by probeset for gene expression data which cause some challenges. Testing a large number of variables (tens of thousands of tests) poses the multiple hypothesis testing problem if this multiplicity of the tests is not taken into account. This problem increases the probability of false positives among the variables or sets that have been declared to be significant. As an example, if the significance level is 0.05 and we have 10 thousand hypotheses that we want to test and if all of the null hypotheses are true, then we expect that 500 tests will be rejected. Many different procedures of correction for multiple testing have been introduced in the statistical literature such as Bonferroni correction, The Family-wise Error Rate

(FWER) and False Discovery Rate (FDR)(Hung et al., 2012; Dudoit et al., 2003).

Another potential limitation of using univariate analyses is ignoring the correlations between variables when they should be incorporated. In genomic studies, genes are not expected to work independently. Some genes interact with each other or with some environmental factors (gene-gene or gene-environment interactions) to influence a phenotype. Hence, these interactions should be taken into account.

The study of a set of variables jointly falls under the framework of multivariate analysis. Traditional multivariate methods such as **M**ultivariate **AN**alysis **O**f **VA**riance (MANOVA) are used to test for significant differences of mean vectors of a set of variables between different groups. This method tests the differences of mean vectors jointly under the assumption of multivariate normality. In the small $n$, large $p$ situation, finding the inverse of the covariance matrix is difficult. Schäfer and Strimmer (2005) addressed this issue using accurate and reliable estimate of the covariance matrix called the shrinkage covariance matrix estimator.

The parametric assumptions such as homogeneity and normality may be violated. Gene expression data, for instance, is very noisy and in general genetic data is not normally distributed. Hence analyses based on Gaussian assumption are inappropriate in this situation (Stewart and Excoffier, 1996). Xu and Cui (2008) proposed a robustified version of MANOVA that does not require any assumptions regarding the distribution of the dataset. Their method is based on a new test statistic and a permutation-based approach to estimate the null distribution.

Pathway-based analysis methods are other multivariate statistical approaches that are especially used for genomic data. These methods evaluate the association

of pre-defined biologically related gene sets/pathways with a phenotype or between different phenotypes. Gene Set Enrichment Analysis (GSEA) is one of the most popular methods that has been used in the context of gene expression. It was originally proposed by Mootha et al. (2003), and then Subramanian et al. (2005) refined it; this method has gained its popularity after the validation of their results. Mootha et al. (2003), were able to identify a group of related genes (oxidative phosphorylation genes) that were differentially expressed between diabetic and non-diabetic patients (these genes were less expressed in diabetics group). Afterwards, their results were validated by independent laboratory studies published in the *New England Journal of Medicine* (Subramanian et al., 2005; Shi and Walker, 2007). Many extensions and modifications to this method have been introduced since then.

This thesis explores applications of multivariate approaches applied to genomic data. It is organized as follows: In Chapter 2, we present an overview of clustering techniques, in particular the k-means clustering algorithm, and provide details of three different multivariate methods: MANOVA, robustified MANOVA, and GSEA. In Chapter 3, we present results from a simulation study where we evaluate the performance of the MANOVA and robustified MANOVA methods. We apply the multivariate methods to a real gene expression and SNP (genotype) data in Chapter 4 and provide summary and future directions in Chapter 5.

# Chapter 2

# Methods

This chapter starts with an overview of cluster analysis then proceeds to different multivariate methods. The first section presents different methods of cluster analysis with a focus on the $k$-means algorithm. It also introduces Positive Root Eigenvalue Plot (PRE plot) to determine the number of clusters believed to exist in a given data set. The second section introduces the three multivariate methods MANOVA, robustified MANOVA, and GSEA.

## 2.1 Clustering High-Dimensional Data

Dimension reduction techniques such as the classical clustering methods and Principal Component Analysis (PCA) are useful tools to capture meaningful data patterns especially from a complex high-dimensional data structure. The large number of variables (either for gene expression data or genome-wide association study i.e. SNPs) as well as the correlations between variables increase the difficulty of interpreting

and analysing such data. Clustering is a widely used technique that helps reduce the complexity of the relationships between variables. Results from cluster analysis facilitate visualizing the data and to extract useful information. Unlike classification, which is a technique that optimally assigns new objects to previously known groups, clustering aims to group variables based on some similarity or distance measures (Johnson and Wichern, 2007).

### 2.1.1 Clustering Methods

Clustering is an exploratory tool that aims to group variables into disjoint clusters such that variables within a cluster are more similar across samples, than they are from others in different clusters. It can be performed on variables, samples, or both variables and samples simultaneously which is called bi-clustering. Clustering techniques are broadly divided into two different methods: the hierarchical and non-hierarchical (or partition-based) clustering methods (Johnson and Wichern, 2007). In both methods, the variables are clustered on the basis of some similarity measures where maximizing the similarity of variables within a cluster and the dissimilarity between clusters is the main objective (Fallah et al., 2008).

**Similarity measures**

There are different similarity and distance measures that can be used in clustering. For example, gene expression (GE) data set is represented by a $p \times n$ matrix, where $n$ is the number of samples (observations) and $p$ is the number of variables (probesets). As described in the introduction, several probesets represent one gene and practically probesets are mapped to their corresponding genes using different

biological databases. Each gene can be represented as a vector $\boldsymbol{x}_i = \{x_{ij},\ 1 \leq j \leq n\ \}$, where $1 \leq i \leq p$ and $x_{ij}$ is the GE measurement for gene $i$ in individual $j$. Proximity measurement between a pair of genes $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$, where $1 \leq i,\ l \leq p$, can be estimated using a variety of similarity measures including absolute or squared correlation measures or measures that depend on some distance measures such as the Euclidean distance (Fallah et al., 2008). Pearson's correlation coefficient ($r$) is a commonly-used measure of similarity. Given a pair of genes $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$ the Pearson's correlation coefficient between them is defined as:

$$r(\boldsymbol{x}_i, \boldsymbol{x}_l) = \frac{\sum_{j=1}^{n}(x_{ij} - \mu_{\boldsymbol{x}_i})(x_{lj} - \mu_{\boldsymbol{x}_l})}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \mu_{\boldsymbol{x}_i})^2}\sqrt{\sum_{j=1}^{n}(x_{lj} - \mu_{\boldsymbol{x}_l})^2}}, 1 \leq i, l \leq p \tag{2.1}$$

where $\mu_i$ and $\mu_l$ are the means of the gene expression measurements $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$, respectively. Also, similarity measures can be constructed from distance measures such as the Euclidean distance defined as:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_l) = \sqrt{\sum_{j=1}^{n}(x_{ij} - x_{lj})^2} \ = \ \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_l)'(\boldsymbol{x}_i - \boldsymbol{x}_l)} \tag{2.2}$$

This distance function satisfies the following properties:

1. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{x}_l) = \mathrm{d}(\boldsymbol{x}_l,\boldsymbol{x}_i)$

2. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{x}_l) > 0$

3. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{x}_l) = 0$ if and only if $\boldsymbol{x}_i = \boldsymbol{x}_l$

4. $\mathrm{d}(\boldsymbol{x}_i,\boldsymbol{x}_l) \leqslant d(\mathbf{x}_i,\boldsymbol{x}_u) + \mathrm{d}(\boldsymbol{x}_u,\boldsymbol{x}_l)$, where $u$ is any other intermediate point.

**Hierarchical clustering methods**

This type of clustering can be represented graphically as a dendrogram that shows the structure of the hierarchical clusters. Agglomerative hierarchical methods start with as many clusters as the variables; then the most similar variables are first clustered and progressively these clusters are merged according to their similarities until one cluster contains all the variables is formed. On the contrary, divisive hierarchical methods start with a single cluster with all variables; then variables are divided into two smaller clusters such that these clusters are dissimilar. This procedure is repeated until each variable forms a cluster. The number of clusters in this type of methods is not specified a priori (Johnson and Wichern, 2007).

**Non-hierarchical clustering methods**

These methods are also known as partition-based clustering methods. Initially, variables are partitioned into $k$ non-overlapping clusters where $k$ is fixed a priori; then variables are re-assigned to the $k$ clusters such that variables within a cluster are more homogeneous and between clusters are more dissimilar.

### 2.1.2 $K$-means Algorithm

$K$-means is a popular partition-based clustering method and was introduced by MacQueen (1967). It is one of the non-hierarchical clustering algorithms that aims to assign each variable to the cluster that has the nearest centroid (mean). The main objective of this algorithm is to minimize the sum of the squared differences between variables within a cluster and the mean of that cluster. This algorithm is summarized as follows (Johnson and Wichern, 2007):

1. **Determine the number of clusters $k$:**

   Different methods of estimating the number of clusters ($k$), $k \leq n$, have been discussed and we use a method called Positive Root Eigenvalue Plot which is presented in the next section.

2. **Partition the variables into $k$ clusters:**

   The variables are initially randomly partitioned into $k$ clusters; each cluster is a set $S_i$ where $1 \leq i \leq k$. Or instead, this procedure could be replaced by a randomly selected $k$ seed points from the variables and these are considered as the initial means of the $k$ sets.

3. **Reassign variables to the closest cluster:**

   Calculate the squared distance, using Euclidean distance in equation 2.2, between each variable and the means of each cluster $S_i$. Then reassign each variable to the closest cluster but before proceeding to the next variable, update the means of the clusters (unless the variable has not been moved to any cluster).

4. **Repeat step 3:**

   Repeat reassigning the variables and updating the means of the clusters until all the variables are allocated to the nearest clusters and then the process ends.

### 2.1.3 Positive Root Eigenvalue Plot

In non-hierarchical clustering methods, the number of clusters $k$ should be specified initially. Estimating $k$ is not an easy problem; however, several approaches have been

proposed and different methods have been developed to solve this issue. Fallah et al. (2008), developed a graphical method called PRE plot (Positive Root Eigenvalue Plot) that is based on eigenvalues of the similarity matrix. This plot determines the number of clusters by plotting the positive square roots of eigenvalues against their ranks. This method is based on the block diagonal structure of the similarity matrix $M$ when ordering the variables using spectral properties. The similarity matrix $M$ can be obtained by calculating the Euclidean distance as defined in (2.2) between variables and then consider this transformation $1-\frac{distance}{max(distance)}$ as a similarity measure. Importantly, comparisons between different candidate numbers of clusters by redoing the clustering analysis many times is not required. To illustrate this plot, Figure 2.1 shows the PRE plot of Melanoma data which is presented by Fallah et al. (2008). They used $p$=3,613 genes and $n$=38 arrays (samples) with the aim of clustering the samples. This plot suggested three to four clusters.

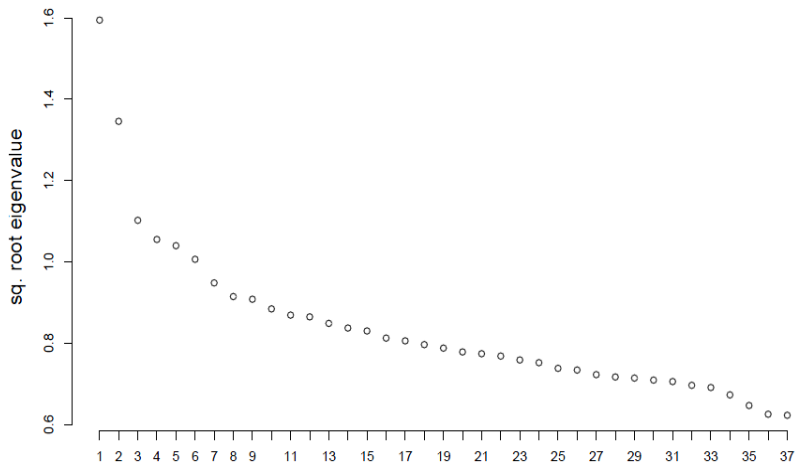



Figure 2.1: The PRE plot of Melanoma data by Fallah et al. (2008).

After determining the number of clusters $k$ by using PRE plot, $k$-means algorithm can be applied.

## 2.2 Multivariate Statistical Methods

Multivariate analysis techniques are used when more than one variable are of interest and needed to be analyzed simultaneously. There are many applications of these techniques to real-world problems depending on the researchers' interests. Some of these techniques are:

**Canonical correlation analysis:** It is used to find the correlation between two groups of variables. Each group consists of a linear combination of some variables; for instance, someone is interested in finding the correlation between a group of psychological variables and a group of academic variables.

**Principal component analysis (PCA) and clustering techniques:** They are used for dimension reduction purposes and grouping variables.

**Multivariate regression analysis:** Often, it is used to predict values of multiple dependent (response) variables from more than one independent (predictor) variables. For example, predicting values of factors affecting sales of a product from several factors that represent the quality of that product.

However, in this thesis we focus on multivariate methods that can be used to test

for significant differences between vectors of means among two or more independent groups. In gene expression context, for example, we aim to identify a group of genes that are differentially expressed (DE) among distinct groups. A group of variables can be selected statistically using cluster analysis or PCA or it can be determined biologically. We present Multivariate Analysis of Variance (MANOVA) and a robustified version of MANOVA, and then present another method called Gene Set Enrichment Analysis (GSEA).

### 2.2.1 Multivariate Analysis of Variance - MANOVA

Multivariate Analysis of Variance (MANOVA) is a generalization of univariate Analysis of Variance (ANOVA). In ANOVA, we test whether the means of a variable of several groups differ significantly or not; two-sample t-test is a special case of ANOVA when we have only two groups. The purpose of MANOVA is to test vectors of means of more than two dependent variables jointly and test if they are significantly different between two or more groups. MANOVA is useful since using multiple ANOVA causes multiple testing problems as discussed in the introduction.
So why is it called method of analysis of variance? This is because the overall variance is divided into variability between samples and variability within the samples.

Suppose there are $p$ dependent variables and $g$ groups (populations):

| Population 1 | Population 2 | $\cdots$ | Population g |
|---|---|---|---|
| $\boldsymbol{x}_{11}$ | $\boldsymbol{x}_{21}$ | $\cdots$ | $\boldsymbol{x}_{g1}$ |
| $\boldsymbol{x}_{12}$ | $\boldsymbol{x}_{22}$ | $\cdots$ | $\boldsymbol{x}_{g2}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $\boldsymbol{x}_{1n_1}$ | $\boldsymbol{x}_{2n_2}$ | $\cdots$ | $\boldsymbol{x}_{gn_g}$ |

Each population $\ell$, $\ell = 1, 2, ..., g$, has a sample of size $n_\ell$, mean vector $\boldsymbol{\mu}_\ell$ and variance-covariance matrix $\boldsymbol{\Sigma}_\ell$ that represents the correlation between the $p$ dependent variables.

$$\boldsymbol{\mu}_\ell = \begin{pmatrix} \mu_{\ell 1} \\ \mu_{\ell 2} \\ \vdots \\ \mu_{\ell p} \end{pmatrix}_{p\times 1} \quad , \quad \boldsymbol{\Sigma}_\ell = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_p^2 \end{pmatrix}_{p\times p}$$

In order to apply MANOVA for testing the equality of the mean vectors across groups, there are three main assumptions:

1. **Independence:** The samples are random and independent.

2. **Normality:** The variables in each group are distributed as multivariate normal $\boldsymbol{X_\ell} \sim \mathrm{N}_p(\boldsymbol{\mu_\ell}, \boldsymbol{\Sigma_\ell})$

3. **Homogeneity of variance-covariance matrices:** The variance-covariance matrices of the dependent variables are homogeneous across the groups.

MANOVA tests the hypothesis:

$$H_0\text{: } \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$$

$H_a$: At least one $\boldsymbol{\mu}_\ell$ $(\ell = 1, ..., g)$ is different.

Table 2.1 summarizes the calculations of the One-way MANOVA (taken from Johnson and Wichern, 2007).

| Source of variation | Matrix of sum of squares and cross products (SSP) | Degrees of freedom |
|---|---|---|
| Treatment(groups) | $\boldsymbol{B}=\sum_{\ell=1}^{g} n_\ell(\bar{\boldsymbol{x}}_\ell - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_\ell - \bar{\boldsymbol{x}})'$ | $g$-1 |
| Residual | $\boldsymbol{W}=\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell} (\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell)(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell)'$ | $\sum_{\ell=1}^{g} n_\ell - g$ |
| Total | $\boldsymbol{B}+\boldsymbol{W}=\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell} (\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}})'$ | $\sum_{\ell=1}^{g} n_\ell - 1$ |

Table 2.1: One-way MANOVA table.

From Table 2.1, $\bar{\boldsymbol{x}}$ is the overall sample mean vector for all treatments, $\bar{\boldsymbol{x}}_\ell$ is the $\ell^{th}$ sample mean vector and $g$ is the number of groups.

The test statistic $\Lambda^* = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}$, called Wilks' Lambda and proposed by Wilks (Wilks, 1946), is commonly used to test for differences between mean vectors across groups and it ranges from 0 to 1. It is the ratio of the within sum of squares to the total sum of squares, where $|\cdot|$ is the determinant of the matrices. In addition, when we have only two groups to compare, Wilks' Lambda $\Lambda^*$ reduces to the Hotelling $T^2$ test. Also, $\Lambda^*$ can be calculated based on the eigen values of $\boldsymbol{W}^{-1}\boldsymbol{B}$.

Table 2.2, (taken from Johnson and Wichern, 2007), shows the exact distribution of $\Lambda^*$ for different numbers of groups and variables:
Bartlett (Bartlett, 1954) proved that if $H_0$ is true and if $n = \sum_{\ell=1}^{g} n_\ell$ is large then,

$$-(n - 1 - \tfrac{(p+g)}{2}) \ln (\Lambda^*)$$

| No. of variables | No. of groups | Sampling distribution for multivariate normal data |
|---|---|---|
| $p = 1$ | $g \geq 2$ | $(\frac{\sum n_\ell - g)}{g-1})(\frac{1-\Lambda^*}{\Lambda^*}) \sim \mathrm{F}_{g-1, \sum n_\ell - g}$ |
| $p = 2$ | $g \geq 2$ | $(\frac{\sum n_\ell - g - 1)}{g-1})(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}) \sim \mathrm{F}_{2(g-1), 2(\sum n_\ell - g - 1)}$ |
| $p \geq 1$ | $g = 2$ | $(\frac{\sum n_\ell - p - 1)}{p})(\frac{1-\Lambda^*}{\Lambda^*}) \sim \mathrm{F}_{p, \sum n_\ell - p - 1}$ |
| $p \geq 1$ | $g = 3$ | $(\frac{\sum n_\ell - p - 2)}{p})(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}) \sim \mathrm{F}_{2p, 2(\sum n_\ell - p - 2)}$ |

Table 2.2: Distribution of Wilks' Lambda $\Lambda^*$ for different number of variables ($p$) and groups ($g$).

has approximately a chi-square distribution with $p(g-1)$ degree of freedom.

**Violations of MANOVA assumptions**

It was proved that Wilks' Lambda test statistic is not robust to normality assumption (Todorov and Filzmoser, 2010). Alternative test statistics have been proposed by many authors; Nath and Pavur (1985) proposed one called the rank transformed Wilks' Lambda which is based on ranking the observations. Todorov and Filzmoser (2010) proposed the robust Wilks' Lambda which is based on the Minimum Covariance Determinant (MCD) estimator; Van Aelst and Willems (2011) proposed several robust test statistics using multisample multivariate S-estimators or MM-estimators.

### 2.2.2 Robustified MANOVA

A novel multivariate method called robustified MANOVA for one-way and two-way cases was proposed by Xu and Cui (2008) and it was applied to a gene expression dataset. In this method, no assumptions are required regarding the distribution of

the dataset. However, the samples are assumed to be independent and identically distributed (i.i.d) and the variance-covariance matrices between groups are homogeneous. They used the same concept of MANOVA, but using robust matrices of within-treatment and between-treatment variations; also they constructed two test statistics based on these matrices and then used permutation-based method in order to estimate the null distribution of the test statistics.

Let us have $\ell = 1, 2, ..., g$ groups, and for each group we have $\boldsymbol{x}_{j(\ell)}$ $p$-dimensional vectors (samples) where $j = 1, 2, ..., n_\ell$ and let $\boldsymbol{\mu}_\ell$ be the mean vector of group $\ell$ and $\boldsymbol{\mu}$ be the overall mean vector across all groups. Let $\tilde{\boldsymbol{\mu}}_\ell$ and $\tilde{\boldsymbol{\mu}}$ be robust estimates of $\boldsymbol{\mu}_\ell$ and $\boldsymbol{\mu}$, respectively, where:

$$\tilde{\boldsymbol{\mu}}_\ell = \text{median}\{\boldsymbol{x}_{j(\ell)},\ j = 1, 2, ..., n_\ell\}$$

$$\tilde{\boldsymbol{\mu}} = \text{median}\{\boldsymbol{x}_{j(\ell)},\ j = 1, 2, ..., n_\ell,\ \ell = 1, 2, ...g\}$$

The median is taken component-wise across all samples within a treatment $\ell$ and for all samples across all treatments. The two robust within-treatment and between-treatment variation matrices are defined respectively as:

$$\tilde{\boldsymbol{W}} = \sum_{\ell=1}^{g} n_\ell \underset{j}{\text{median}}\{(\boldsymbol{x}_{j(\ell)} - \tilde{\boldsymbol{\mu}}_{(\ell)})(\boldsymbol{x}_{j(\ell)} - \tilde{\boldsymbol{\mu}}_{(\ell)})'\} \tag{2.3}$$

$$\tilde{\boldsymbol{B}} = \sum_{\ell=1}^{g} n_\ell (\tilde{\boldsymbol{\mu}}_{(\ell)} - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_{(\ell)} - \tilde{\boldsymbol{\mu}})' \tag{2.4}$$

Since $\tilde{\boldsymbol{B}}$ and $\tilde{\boldsymbol{W}}$ are both singular matrices, instead of using the definition of Wilks'

Lambda which is based on the determinant of the matrices, Xu and Cui proposed the following two test statistics:

$$G = \frac{tr(\tilde{\boldsymbol{B}})}{tr(\tilde{\boldsymbol{W}})} \tag{2.5}$$

and

$$\tilde{G} = \frac{\text{median}(\text{diag}\tilde{\boldsymbol{B}})}{\text{median}(\text{diag}\tilde{\boldsymbol{W}})} \tag{2.6}$$

These test statistics are the ratio of between-treatment to within-treatment variations. To estimate the null distribution of the test statistics $G$ or $\tilde{G}$, permutation approach is used.

This method works as follow:

1. **Relabel all treatments $\ell$:**

   Randomly shuffle all the samples ($n = \sum_{i=1}^{\ell} n_i$) to form $\ell$ new treatments; each treatment consists of $n_1, n_2, ..., n_\ell$ samples. Then compute the test statistics from this new dataset.

2. **Repeat step 1:**

   Repeat step 1 ($s$) for $M = 1000$ times, and each time denote the test statistics as $\{G^{(s)}$ or $\tilde{G}^{(s)}, \ \ s = 1, 2, ..., M\}$.

3. **Calculate the empirical $p$-value:**

   Compare the observed test statistic to the shuffled test statistics (the permutation distribution) to obtain the empirical $p$-value. For the test statistics $G$, for example, this can be calculated as:

$$p\text{-value} = \frac{\sum_{s=1}^{M} I(G^{(s)} > G)}{M}, \qquad \text{where } I \text{ is the indicator function.}$$

A $p$-value for $\tilde{G}$ is obtained similarly.

### 2.2.3 Gene Set Enrichment Analysis - GSEA

Gene Set Enrichment Analysis (GSEA) is a popular method in genomics which combines information from multiple genes that belong to a well-defined set or pathway. In general, pathway-based approaches consider genes that are believed to be related and act in concert in various biological processes. These pathways are pre-defined gene-sets from biological databases "that share common biological function, chromosomal location, or regulation" (Subramanian et al., 2005). For instance, "the sets of genes representing biological pathways in the cell or sets of genes whose DNA sequences are close together on the cell's chromosomes" are said to belong to the same pathway (Efron and Tibshirani 2007). The goal is to evaluate the association of these gene sets/pathways with a phenotype or disease.

GSEA was first proposed by Mootha et al. (2003) and improved by Subramanian et al. (2005). It was originally applied to gene expression data, but later Wang et al. (2007) developed a modified version of GSEA and was used to analyze data from a genome-wide association study (GWAS) of complex diseases (using SNPs "genotype" data instead of gene expression). The first GSEA application was able to identify a group of genes (oxidative phosphorylation genes) that were differentially expressed between diabetic and non-diabetic groups. Independent laboratory studies,

published in the *New England Journal of Medicine*, confirmed this result (Shi and Walker, 2007).

The primary aim of GSEA is to assess the significance of the gene sets/pathways by evaluating the enrichment of genes within a pathway at the top (or bottom) of a list of ranked genes (Beyene et al., 2009; Subramanian et al., 2005; Wang et al.,2007). A gene set is deemed to be related to a phenotype (thus enriched) if the genes within that set are not randomly distributed among the ranked list ($L$); instead, they will likely be at the top or bottom of $L$ (Subramanian et al., 2005). Subramanian et al. (2005) provided GSEA-P software as well as an initial database of 1,325 gene sets/pathways (the Molecular Signatures Database MSigDB 1.0). There are several options for GSEA software including R package and GSEA desktop application (javaGSEA). The Molecular Signatures Database (MSigDB) is a collection of gene sets/pathways that are freely available through GSEA/MSigDB web site which is maintained by the GSEA team and developed at the Broad Institute of MIT and Harvard. Currently, MSigDB v3.1 (updated on Sep 27, 2012) consists of 8,513 gene sets.

**GSEA Algorithm**

Suppose we have $k$ samples and $N$ genes $(g_1, g_2, ..., g_N)$ in the dataset. For a given gene set/pathway $(S)$, with $N_H$ pre-defined genes, we want to identify if the genes within the set/pathway $(S)$ show statistically significant differences (i.e. differentially expressed) between two or three phenotypes. In general, GSEA follows the following steps:

**Step 1. Generating a list of ranked genes $L$:**

Compute a gene-level statistic for each gene using any association score. There are different gene-level statistics that can be used to determine if a gene is significantly different (i.e., differentially expressed) among groups such as t-statistic, Wilcoxon rank sum test, or signal-to-noise ratio (Shi and Walker, 2007; Efron and Tibshirani, 2007). We can use ANOVA for three or more groups. Hence, we have $N$ association scores $r(g_j) = r_j,\ j = 1, ..., N$ and then sort the genes by the most significant ones.

**Step 2. Calculating the enrichment score ($ES$) for each set/pathway $S$:**

Calculate the enrichment score ($ES$) for each set/pathway ($S$) using a weighted Kolmogorov-Smirnov-like running-sum statistic. For a given position $i$ in $L$:

$$ES(S) = \max_{1 \leqslant i \leqslant N} \left\{ P_{hit}(S,i) - P_{miss}(S,i) \right\}, \qquad \text{where}$$

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \leqslant i}} \frac{\mid r_j \mid^p}{N_R}, \qquad N_R = \sum_{g_j \in S} \mid r_j \mid^p$$

$$P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \leqslant i}} \frac{1}{(N - N_H)}$$

Where $P_{hit}(S,i)$ is the fraction of genes in a gene set ($S$) weighted by the association scores of the genes and $P_{miss}(S,i)$ is the fraction of genes not in $S$. $N_H$ is the number of genes in a set ($S$). The enrichment score ($ES$) is the maximum deviation from zero of $P_{\text{hit}} - P_{\text{miss}}$. If $ES$ is small then the genes within a set ($S$) are randomly distributed throughout the list, otherwise $ES$ will be high. If the genes that belong to a gene set/pathway ($S$) are randomly distributed across the ranked list $L$, then ($S$) does not

show statistical significance. Otherwise, if these genes within $(S)$ are distributed at the top (or bottom) of $L$ then this gene set $(S)$ is considered significantly enriched. Subramanian et al. (2005) suggested $p$ to be equal to 1 in order to give weight to the genes within $S$, so genes with higher association scores would contribute more in the set/pathway $S$. Figure 2.2 shows a schematic representation of the steps involved in GSEA (Subramanian et al., 2005).

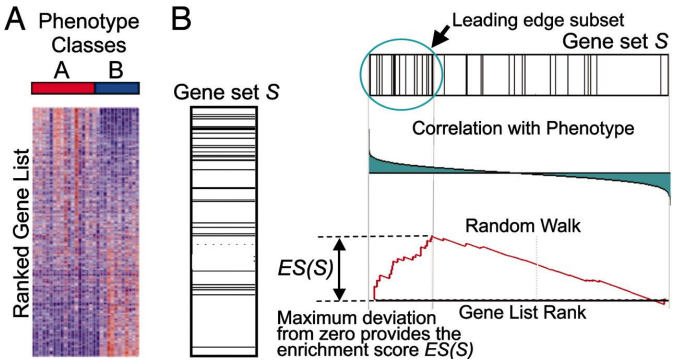



Figure 2.2: GSEA illustration by Subramanian et al. (2005). A) represents a heatmap of a list of ranked genes on the order of differential expression. B) shows different genes from a list (A) which are included in a gene set/pathway $S$, and a plot of the enrichment score $(ES)$ of $S$.

**Step 3. Estimating the significance of the enriched score $(ES)$:**

Phenotype-based permutation test is used to obtain the null distribution of the enrichment score $(ES)$ of the set $S$. We randomly shuffle the phenotype labels of the $k$ samples and then rank the genes based on their significance to produce $L$. Then, we compute the $ES$ of the set $S$ for the shuffled phenotypes. Thus, for gene set/pathway $S$ in permutation $\pi$ we have $ES(S, \pi)$. We repeat the permutation procedure a large

number of times, say $M$= 1000 times, and for each permuted data set we calculate $ES(S, \pi)$ and then compute the empirical $p$-value for the observed $ES(S)$ from the permuted distribution. Depending on the sign of the observed $ES(S)$, we can determine its $p$-value using the negative or positive portion of the permuted distribution. If $ES(S)$ is positive then:

$$p\text{-value}(ES(S)) = \frac{\sum_{\pi=1}^{M} I(ES(S,\pi) > ES(S))}{M}, \qquad \text{where } I \text{ is the indicator function.} \tag{2.7}$$

**Step 4. Adjusting for multiple hypothesis testing:**

If we have many gene sets/pathways that we want to test simultaneously, we have to correct for multiple hypothesis testing and adjust the estimated $p$-value accordingly. First, we adjust for different member of genes within sets by normalizing the $ES(S)$ and $ES(S, \pi)$ for a given $S$. The normalized enrichment scores $NES(S)$ and $NES(S, \pi)$ are obtained as:

$$NES(S) = \frac{ES(S)}{\text{mean}(ES(S,\pi))}$$

$$NES(S,\pi) = \frac{ES(S,\pi)}{\text{mean}(ES(S,\pi))}$$

Adjustment for multiple testing is applied to control the proportion of the tests that are declared to be significant when they are not (false positives) using the false discovery rate (FDR)(Holmans, 2010; Subramanian et al., 2005). The FDR

(Benjamini and Hochberg, 1995) is the ratio of the expected proportion of false positives among the rejected hypotheses. We calculate FDR $q$-value instead of the nominal $p$-value to determine the significance of the test.

In this thesis we illustrated this method and applied it to genotype data. This data set was provided by the organizers of the Genetic Analysis Workshop 18 (GAW 18) which focused on analysis of whole genome sequence data. The aim is to test the effect of both rare and common genetic variants (SNPs) in a blood pressure study. We use a pathway-based approach, gene set enrichment analysis, to search for related genes affecting four phenotypes: systolic blood pressure, diastolic blood pressure, the difference between them and mean arterial pressure, which is a weighted linear combination of systolic and diastolic blood pressure.

Using the GAW 18 data, we consider both rare and common variants in our analysis and incorporated other covariates by using a recently proposed test statistic VW-TOW (Variable Weight Test for testing the effect of an Optimally Weighted combination of variants)(Sha et al., 2012). We followed the same step of GSEA Algorithm but we used VW-TOW as an association score and we used a gene-based permutation instead of phenotype-based permutation because we have only one phenotype at a time, which means we permute the genes and their test statistics to obtain the null distribution of the enrichment score $ES$.

### 2.2.4 A Brief Description of The VW-TOW Test

VW-TOW test statistics and the corresponding $p$-values are obtained by assuming that we have $n$ individuals who have been genotyped at $M$ variants and $y_i$, $i = 1, ..., n$

is the trait of interest for the $i^{th}$ individual. Each individual has a genotypic score $X_i$=$(x_{i1}, x_{i1}, ..., x_{iM})^T$, where $x_{im} \in \{0,1,2\}$ denotes the number of copies of the minor allele for the $m^{th}$ variant of the $i^{th}$ individual. We used a minor allele frequency (MAF) threshold of less than 1% to define rare variants. To Test the effect of the Optimally Weighted combination (TOW) of variants $x_i^0 = \sum_{m=1}^{M}(w_m^0 x_{im})$, the following statistic is used:

$$T_T = \sum_{i=1}^{n}(y_i - \bar{y})(x_i^0 - \bar{x}^0)$$

where $w_m^0 = \sum_{i=1}^{n}(y_i - \bar{y})(x_{im} - \bar{x}_m)/\sum_{i=1}^{n}(x_{im} - \bar{x}_m)^2$ are the optimal weights. TOW is used for rare variants and it loses power when testing the effect of both rare and common variants because it gives small weight for the common variants. In order to test the effect of both rare and common variants, we applied TOW to each of them separately; $T_r$ and $T_c$ denote the statistics for the 'rare' and 'common' variants, respectively. The test statistic for VW-TOW is given by:

$$T_{VW\_T} = \min_{0\leq\lambda\leq 1} p_\lambda$$

where $p_\lambda$ is the $p$-value of the test $T_\lambda$, $T_\lambda = \lambda\frac{T_r}{\sqrt{\text{var}(T_r))}} + (1-\lambda)\frac{T_c}{\sqrt{\text{var}(T_c))}}$. To obtain a $p$-value for $T_{VW\_T}$, we used a permutation approach. Other covariates $(z_{i1}, ..., z_{ip})^T$ for each individual $i$ are incorporated by adjusting $y_i$ and $x_{im}$ using linear regression:

$$y_i = \alpha_0 + \alpha_1 z_{i1} + ... + \alpha_p z_{ip} + \epsilon_i \qquad \text{and} \tag{2.8}$$

$$x_{im} = \alpha_{0m} + \alpha_{1m} z_{i1} + \ + \alpha_{pm} z_{ip} + \tau_{im}$$

And by using the residuals $\tilde{y}_i$ and $\tilde{x}_{im}$, the following TOW and VW-TOW are used:

$$T_{TOW} = T_{T|y_i=\tilde{y}_i, x_{im}=\tilde{x}_{im}} \qquad \text{and}$$

$$T_{VW\_TOW} = T_{VW-T|y_i=\tilde{y}_i, x_{im}=\tilde{x}_{im}}$$

The following linear model is used to model the relationship between trait values, covariates and genotypes which is equivalent to adjusting trait values and genotypic scores for the covariates in the linear model in equatinon (2.9):

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 z_{i1} + ... + \alpha_p z_{ip} + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_M x_{iM} + \epsilon_i \\ &= \alpha^T Z_i + \beta^T X_i + \epsilon_i \end{aligned} \tag{2.9}$$

The score test statistic to test the effect of weighted combination of variants, $x_i = \sum_{m=1}^{M} (w_m x_{im})$, is given by:

$$SC(w_1, ..., w_M) = \frac{n(\sum_{i=1}^{n} \tilde{y}_i \tilde{x}_i)^2}{(\sum_{i=1}^{n} \tilde{y}_i^2 \sum_{i=1}^{n} \tilde{x}_i^2)} \tag{2.10}$$

# Chapter 3

# Simulation Studies

In this chapter, we assess the performance of the Wilks' Lambda ($\Lambda^*$) and the two robustified MANOVA test statistics ($G$ and $\tilde{G}$) with respect to Type I error rates and power using a Monte Carlo simulation study.

## 3.1 Simulation Design

In our simulation we considered the multivariate normal and multivariate gamma distributions. We generated $p$ dependent variables and $n$ observations from each distribution and considered the case when $n > p$. Due to the singularity of the covariance matrix in the high dimensional case, i.e., $n < p$, and hence difficulty calculating Wilks' Lambda in MANOVA we did not consider this case in our simulation scenarios.

We varied the following parameters in the simulation: the number of groups ($g$), the sample size for each group ($n_i, i = 1, 2, ..., g$), the number of variables ($p$), the

correlation between the $p$ variables and the effect size. The number of groups that we considered in the simulation settings is two and three, and allowed both balanced and unbalanced samples. We also varied the correlation between the $p$ variables ($r$); we considered zero, weak ($r$=0.2), moderate ($r$=0.5) and strong ($r$=0.8) correlations. We assumed the same variance-covariance matrix $\Sigma$ across groups. We generated 1000 simulated datasets for each setting and 1000 permutation to estimate the null distribution for the two test statistics $G$ and $\tilde{G}$. Table 3.1 shows the parameters that we varied during the simulation and the different settings.

| **Distributions** | **Parameters** | |
|---|---|---|
| Multivariate Normal | $g$ | 2, 3 |
| | $n$ when $g$=2 | (10,10),(50,50),(100,100),(20,25),(20,30),(20,50) |
| | $n$ when $g$=3 | (10,10,10),(50,50,50),(100,100,100),(20,22,25), (20,25,30),(20,30,50) |
| | $p$ | 5, 10, 15 |
| | $r$ | 0, 0.2, 0.5, 0.8 |
| Multivariate Gamma | (shape, rate) | (2, 0.5) |

Table 3.1: Parameters that varied in the simulation.

We implemented the simulation procedure in the R statistical package. We simulated multivariate normal distribution using the mvrnorm() function from the 'MASS' package (Venables & Ripley ,2002). We generated multivariate gamma distribution using a mixture approach. To estimate the null distribution for $G$ and $\tilde{G}$, we used the 'permute' package (Gavin, 2012) to permute the samples between groups. Dr. John R. Stevens provided us with the R function of the test statistics $G$ and $\tilde{G}$, since the original author (Xu and Cui, 2008) provided Matlab function (Stevens et al., 2010).

### 3.1.1 Multivariate Distributions

Simulating data from a multivariate normal distribution is straightforward and convenient in R, but simulating multivariate non-normal distributions needs some theories and techniques in order to obtain such data.

We used a mixture approach to simulate the $p$-variate gamma distribution. This method is proposed by Minhajuddin et al. (2004). The method depends on the Bayesian conjugate prior families of distributions (i.e., the posterior and prior distributions are from the same family); the gamma family is conjugate for the Poisson family and hence we can simulate $p$-variate gamma distribution using this property. There are two steps in this approach. First we simulate $G = g$ from the marginal probability distribution function $m(g;\theta)$. Then conditional on $G = g$, we independently simulate $p$ random variables $X_i, i = 1, ..., p$ from a posterior $p(x_i|G = g, \eta)$, where $\eta$ determines the correlation between any two variables $X_i$ and $X_j, i \neq j$.

In order to simulate a $p$-variate gamma distribution with shape parameter $r$ and rate parameter $\lambda$, we follow the two steps outlined above:

1. Simulate $G = g$ from the negative binomial distribution with parameters $r$ and $\pi$, where $\pi = \frac{\lambda}{\lambda+\theta}$, $\theta = \frac{\lambda\rho}{1-\rho}$, and $\rho$ is the correlation coefficient between any two variables $X_i$ and $X_j$, $i \neq j$.

2. Conditional on the same value of $G = g$, we independently simulate $p$ random variables from gamma $(r + g,\lambda + \theta)$ with pdf:

$$p(x_i|g; r, \lambda, \theta) = \frac{(\lambda + \theta)^{r+g}}{\Gamma(r + g)} x_i^{r+g-1} e^{-(\lambda+\theta)x_i}$$

The R code for generating data from a multivariate gamma distribution was provided by Dr. Ahmed Hossain (Hossain et al., 2013). We used a Multivariate kernel density estimation to illustrate the skewness of the multivariate gamma distribution. This method is used to estimate the probability density functions. Figure 3.1 shows the density estimate of a multivariate gamma distribution when we have $p$=2 variables, shape parameter = 2 and rate parameter = 0.5.

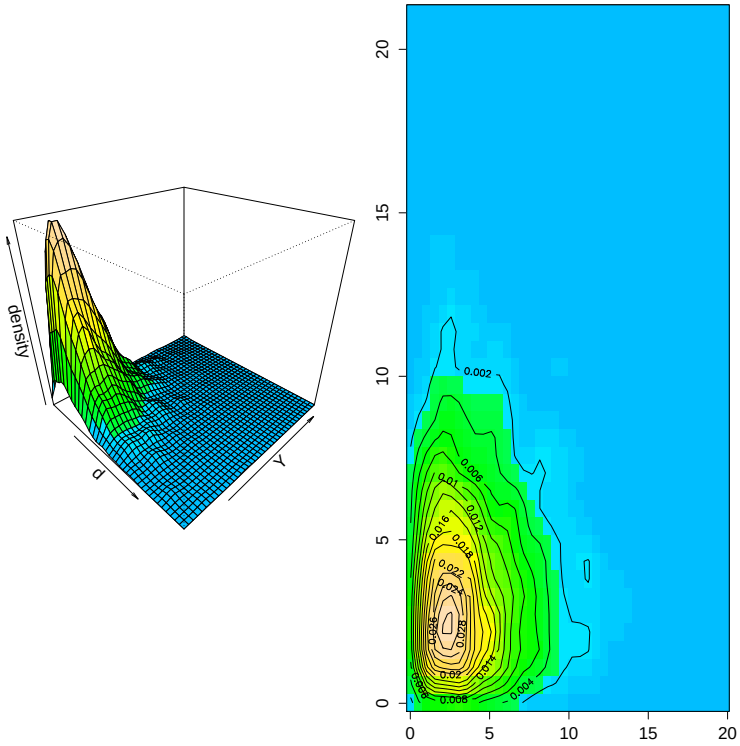



Figure 3.1: The density estimate of a bivariate Gamma distribution with shape parameter = 2 and rate parameter = 0.5. On the left side, is the perspective plot and on the right side is the contour plot corresponding to the perspective plot.

## 3.2 Simulation Results

### 3.2.1 Type I Error Rate

Type I error occurs when a true null hypothesis is rejected. To calculate Type I error rates for different scenarios (Table 3.1), we simulated data under the null hypothesis. We generated data from multivariate normal distribution with $\boldsymbol{\mu}$=**0** and $\Sigma$=**I** and from multivariate gamma distribution with shape parameter = 2 and rate parameter = 0.5. The results are summarized in Tables 3.2 and 3.3.

| | | | Multivatriate Normal | | | | | | Multivatriate gamma | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\alpha$=0.05 | | | $\alpha$=0.01 | | | $\alpha$=0.05 | | | $\alpha$=0.01 | | |
| $p$ | $n1$ | $n2$ | $\Lambda^*$ | $G$ | $\tilde{G}$ | $\Lambda^*$ | $G$ | $\tilde{G}$ | $\Lambda^*$ | $G$ | $\tilde{G}$ | $\Lambda^*$ | $G$ | $\tilde{G}$ |
| 5 | 10 | 10 | 0.037 | 0.05 | 0.057 | 0.009 | 0.01 | 0.016 | 0.05 | 0.054 | 0.044 | 0.005 | 0.01 | 0.01 |
| 5 | 50 | 50 | 0.051 | 0.05 | 0.05 | 0.009 | 0.011 | 0.007 | 0.051 | 0.05 | 0.041 | 0.007 | 0.008 | 0.012 |
| 5 | 100 | 100 | 0.053 | 0.048 | 0.048 | 0.012 | 0.009 | 0.008 | 0.058 | 0.049 | 0.04 | 0.004 | 0.011 | 0.006 |
| 5 | 20 | 25 | 0.042 | 0.043 | 0.052 | 0.009 | 0.007 | 0.007 | 0.049 | 0.056 | 0.05 | 0.011 | 0.012 | 0.017 |
| 5 | 20 | 30 | 0.049 | 0.04 | 0.041 | 0.008 | 0.008 | 0.008 | 0.049 | 0.052 | 0.052 | 0.012 | 0.009 | 0.01 |
| 5 | 20 | 50 | 0.046 | 0.044 | 0.048 | 0.004 | 0.004 | 0.004 | 0.049 | 0.057 | 0.068 | 0.005 | 0.012 | 0.017 |
| 10 | 50 | 50 | 0.056 | 0.043 | 0.034 | 0.011 | 0.006 | 0.004 | 0.048 | 0.049 | 0.048 | 0.009 | 0.009 | 0.01 |
| 10 | 100 | 100 | 0.05 | 0.055 | 0.055 | 0.006 | 0.011 | 0.018 | 0.053 | 0.041 | 0.039 | 0.008 | 0.007 | 0.007 |
| 10 | 20 | 25 | 0.048 | 0.048 | 0.062 | 0.01 | 0.012 | 0.006 | 0.044 | 0.057 | 0.048 | 0.008 | 0.011 | 0.008 |
| 10 | 20 | 30 | 0.051 | 0.049 | 0.047 | 0.008 | 0.009 | 0.004 | 0.057 | 0.042 | 0.04 | 0.011 | 0.012 | 0.005 |
| 10 | 20 | 50 | 0.046 | 0.039 | 0.036 | 0.005 | 0.007 | 0.009 | 0.049 | 0.048 | 0.055 | 0.007 | 0.015 | 0.013 |
| 15 | 50 | 50 | 0.049 | 0.054 | 0.045 | 0.01 | 0.004 | 0.01 | 0.047 | 0.048 | 0.057 | 0.01 | 0.012 | 0.006 |
| 15 | 100 | 100 | 0.05 | 0.052 | 0.061 | 0.014 | 0.01 | 0.017 | 0.044 | 0.048 | 0.05 | 0.009 | 0.011 | 0.011 |
| 15 | 20 | 25 | 0.044 | 0.048 | 0.054 | 0.008 | 0.008 | 0.01 | 0.037 | 0.051 | 0.042 | 0.007 | 0.01 | 0.005 |
| 15 | 20 | 30 | 0.036 | 0.054 | 0.034 | 0.001 | 0.007 | 0.007 | 0.047 | 0.044 | 0.044 | 0.006 | 0.008 | 0.007 |
| 15 | 20 | 50 | 0.04 | 0.044 | 0.043 | 0.006 | 0.005 | 0.009 | 0.04 | 0.051 | 0.048 | 0.006 | 0.016 | 0.009 |

Table 3.2: Type I error rates for the test statistics $\Lambda^*$, $G$, and $\tilde{G}$ at $\alpha$=0.05 and 0.01 for multivariate normal and multivariate gamma distributions in the case of **two groups**.

| | | | | Multivatriate Normal | | | | | | Multivatriate gamma | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 0.05 | | | 0.01 | | | 0.05 | | | 0.01 | | |
| $p$ | $n1$ | $n2$ | $n3$ | $\Lambda^*$ | $G$ | $\tilde{G}$ | $\Lambda^*$ | $G$ | $\tilde{G}$ | $\Lambda^*$ | $G$ | $\tilde{G}$ | $\Lambda^*$ | $G$ | $\tilde{G}$ |
| 5 | 10 | 10 | 10 | 0.039 | 0.05 | 0.052 | 0.008 | 0.014 | 0.015 | 0.064 | 0.058 | 0.055 | 0.012 | 0.013 | 0.011 |
| 5 | 50 | 50 | 50 | 0.047 | 0.038 | 0.041 | 0.013 | 0.007 | 0.008 | 0.043 | 0.056 | 0.038 | 0.004 | 0.011 | 0.006 |
| 5 | 100 | 100 | 100 | 0.05 | 0.042 | 0.054 | 0.007 | 0.009 | 0.011 | 0.044 | 0.05 | 0.041 | 0.003 | 0.009 | 0.007 |
| 5 | 20 | 22 | 25 | 0.045 | 0.044 | 0.05 | 0.006 | 0.009 | 0.013 | 0.039 | 0.062 | 0.057 | 0.009 | 0.007 | 0.01 |
| 5 | 20 | 25 | 30 | 0.045 | 0.039 | 0.041 | 0.008 | 0.01 | 0.008 | 0.056 | 0.052 | 0.053 | 0.014 | 0.013 | 0.015 |
| 5 | 20 | 30 | 50 | 0.05 | 0.055 | 0.048 | 0.007 | 0.01 | 0.009 | 0.054 | 0.062 | 0.047 | 0.012 | 0.007 | 0.007 |
| 10 | 50 | 50 | 50 | 0.059 | 0.037 | 0.042 | 0.009 | 0.006 | 0.007 | 0.037 | 0.057 | 0.051 | 0.005 | 0.015 | 0.011 |
| 10 | 100 | 100 | 100 | 0.049 | 0.045 | 0.057 | 0.008 | 0.012 | 0.014 | 0.045 | 0.048 | 0.043 | 0.005 | 0.009 | 0.009 |
| 10 | 20 | 22 | 25 | 0.057 | 0.055 | 0.036 | 0.008 | 0.011 | 0.011 | 0.047 | 0.045 | 0.053 | 0.006 | 0.008 | 0.012 |
| 10 | 20 | 25 | 30 | 0.038 | 0.054 | 0.049 | 0.011 | 0.013 | 0.011 | 0.037 | 0.054 | 0.051 | 0.007 | 0.007 | 0.009 |
| 10 | 20 | 30 | 50 | 0.044 | 0.044 | 0.037 | 0.009 | 0.007 | 0.002 | 0.055 | 0.047 | 0.041 | 0.012 | 0.01 | 0.009 |
| 15 | 50 | 50 | 50 | 0.056 | 0.044 | 0.048 | 0.008 | 0.005 | 0.009 | 0.037 | 0.058 | 0.047 | 0.008 | 0.012 | 0.01 |
| 15 | 100 | 100 | 100 | 0.05 | 0.054 | 0.056 | 0.007 | 0.009 | 0.014 | 0.041 | 0.052 | 0.042 | 0.004 | 0.007 | 0.005 |
| 15 | 20 | 22 | 25 | 0.047 | 0.047 | 0.039 | 0.009 | 0.008 | 0.008 | 0.047 | 0.05 | 0.053 | 0.008 | 0.008 | 0.011 |
| 15 | 20 | 25 | 30 | 0.037 | 0.046 | 0.048 | 0.007 | 0.007 | 0.006 | 0.03 | 0.045 | 0.047 | 0.006 | 0.007 | 0.012 |
| 15 | 20 | 30 | 50 | 0.037 | 0.033 | 0.034 | 0.009 | 0.011 | 0.009 | 0.045 | 0.048 | 0.049 | 0.008 | 0.008 | 0.01 |

Table 3.3: Type I error rates for the test statistics $\Lambda^*$, $G$, and $\tilde{G}$ at $\alpha$=0.05 and 0.01 for multivariate normal and multivariate gamma distributions in the case of **three groups**.

The probability of making a Type I error is called the significance (nominal) level and denoted as $\alpha$. The 95% confidence intervals around the two significance levels ($\alpha$) 0.05 and 0.01 are: (0.037 , 0.064) and (0.004 , 0.016), respectively. This intervals are calculated using the formula:

$$\alpha \pm 1.96\sqrt{\tfrac{\alpha(1-\alpha)}{1000}}$$

where 1000 is the number of simulations.

The results for the two groups case (Table 3.2), show that the empirical Type I error rates for the three test statistics (Wilks' Lambda $\Lambda^*$, robustified MANOVA

$G$, and $\tilde{G}$) for all scenarios in the two different distributions are around the nominal levels. Similarly, the results for the three groups case (Table 3.3) show that most of the time the Type I error rates for all three test statistics are around the nominal levels for the multivariate normal as well as the multivariate gamma distributions.

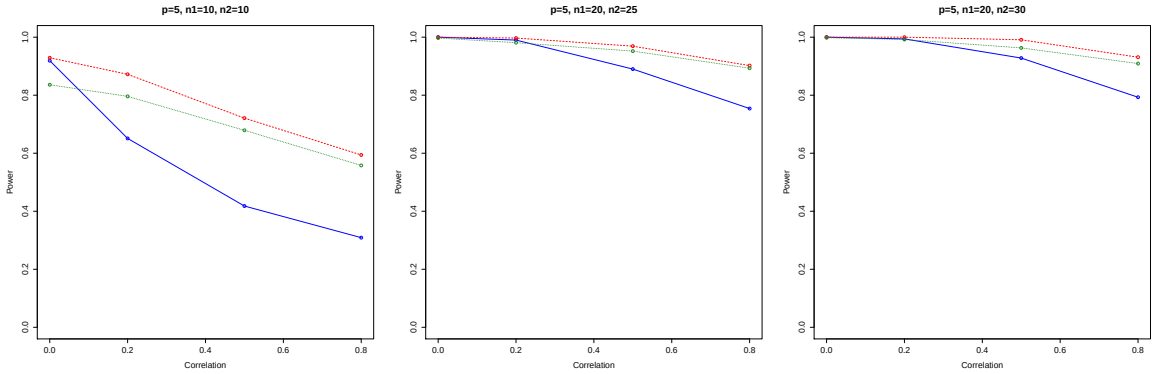
### 3.2.2 Power

The power of a test statistic is defined as the probability of rejecting a false null hypothesis i.e., making the right decision by rejecting a false null hypothesis. In order to assess the power of the three test statistics, we generated data under an alternative hypothesis.

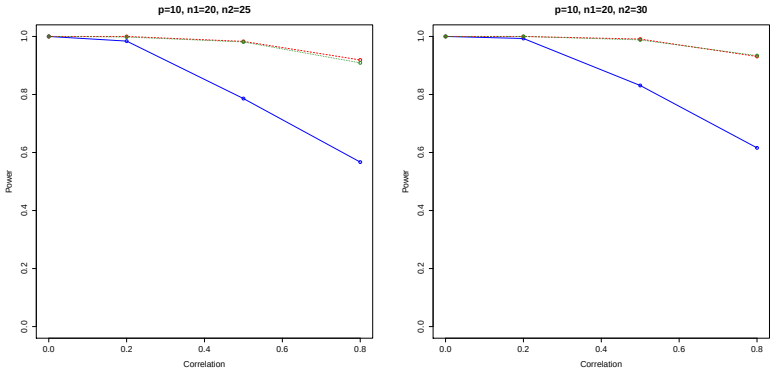
$$H_a\text{: at least one } \boldsymbol{\mu}_\ell \text{ is different, } \ell = 1,2,...,g.$$

Here, we considered the same parameters summarized in Table 3.1 and we also varied the correlation between the $p$ variables ($r$); we considered zero, weak, moderate and strong correlations (0, 0.2, 0.5, 0.8), respectively.

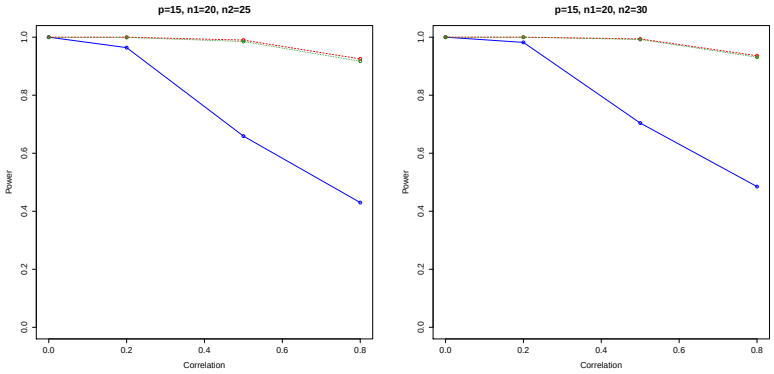
In general, we noticed that as the number of sample sizes, variables and effect size increase, the power increases to 1 for the three test statistics. Below, we first present the power results for multivariate normal distribution for the two and three groups scenarios. In the case of two groups, we generated data where the effect size is equal to one and we investigated whether the correlation between the dependent variables affects the power of the three test statistics. Figure 3.2 shows the power of the three test statistics for some scenarios in the case of two groups.




(a) $p$=5, $n_1$=10, $n_2$=10 (b) $p$=5, $n_1$=20, $n_2$=25 (c) $p$=5, $n_1$=20, $n_2$=30




(d) $p$=10, $n_1$=20, $n_2$=25 (e) $p$=10, $n_1$=20, $n_2$=30




(f) $p$=15, $n_1$=20, $n_2$=25 (g) $p$=15, $n_1$=20, $n_2$=30

Figure 3.2: The power of three test statistics for different number of variables ($p$), correlation between variables ($r$) and different sample sizes in the case of **two groups** form multivariate normal distribution. The blue line is for $\Lambda^*$, the red dashed line is for $G$ and the green dashed line corresponds to $\tilde{G}$.

For all the three test statistics ($\Lambda^*$, $G$ and $\tilde{G}$), as the correlation between the $p$ variables increases, the power decreases. Also, we noticed that as the number of sample sizes increase (50 or 100) the power of the three test statistics increases to as high as 100%. $G$ and $\tilde{G}$ have higher power than $\Lambda^*$ in all the scenarios.

In the case of three groups, we generated data from multivariate normal with mean vectors ($\boldsymbol{\mu}_1 = 1$, $\boldsymbol{\mu}_2 = 2$ and $\boldsymbol{\mu}_3 = 4$), we did not present the results because the power of all the test statistics are approximately equal to one.

For the multivariate gamma distribution, the effect sizes of the shape parameters are varied as shown in Table 3.4 below in the case of two and three groups. We set the rate parameter at 0.5 in all cases. The results of the power when we considered

| | Number of groups | | | | |
|---|---|---|---|---|---|
| Effect size (shape parameter) | $g$=2 | | $g$=3 | | |
| | $g1$ | $g2$ | $g1$ | $g2$ | $g3$ |
| 1 | 2 | 2.5 | 2 | 2.5 | 3 |
| 2 | 2 | 4 | 2 | 4 | 2.4 |

Table 3.4: The shape parameters that we considered for multivariate gamma distribution in the case of two and three groups.

effect size 2 in Table 3.4 are all close to 1 so we only represented the results for effect size 1. The results for the power of $\Lambda^*$, $G$ and $\tilde{G}$ when we considered multivariate gamma in the case of two and three groups are summarized in Tables 3.5 and 3.6.

From Tables 3.5 and 3.6, we notice that $G$ and $\tilde{G}$ have higher power than $\Lambda^*$ when the correlations between variables are weak, moderate and strong. Also, for the same number of variables $p$, the power increases when the number of samples increases and for the same sample size, as the number of variables increases the power of the two test statistcs $G$ and $\tilde{G}$ also increases.

| | | | | Power | | |
|---|---|---|---|---|---|---|
| $p$ | $n1$ | $n2$ | $r$ | $\Lambda^*$ | $G$ | $\tilde{G}$ |
| 5 | 10 | 10 | 0 | 0.162 | 0.197 | 0.154 |
| | | | 0.2 | 0.113 | 0.186 | 0.17 |
| | | | 0.5 | 0.074 | 0.154 | 0.153 |
| | | | 0.8 | 0.069 | 0.128 | 0.122 |
| 5 | 50 | 50 | 0 | 0.807 | 0.724 | 0.634 |
| | | | 0.2 | 0.546 | 0.64 | 0.563 |
| | | | 0.5 | 0.326 | 0.53 | 0.493 |
| | | | 0.8 | 0.251 | 0.432 | 0.422 |
| 10 | 50 | 50 | 0 | 0.961 | 0.933 | 0.858 |
| | | | 0.2 | 0.496 | 0.807 | 0.755 |
| | | | 0.5 | 0.226 | 0.611 | 0.574 |
| | | | 0.8 | 0.189 | 0.461 | 0.444 |
| 15 | 50 | 50 | 0 | 0.995 | 0.982 | 0.92 |
| | | | 0.2 | 0.459 | 0.869 | 0.816 |
| | | | 0.5 | 0.187 | 0.643 | 0.607 |
| | | | 0.8 | 0.14 | 0.466 | 0.453 |
| 5 | 100 | 100 | 0 | 0.991 | 0.968 | 0.921 |
| | | | 0.2 | 0.856 | 0.917 | 0.873 |
| | | | 0.5 | 0.587 | 0.815 | 0.793 |
| | | | 0.8 | 0.447 | 0.71 | 0.7 |
| 10 | 100 | 100 | 0 | 1 | 1 | 0.996 |
| | | | 0.2 | 0.866 | 0.984 | 0.979 |
| | | | 0.5 | 0.506 | 0.868 | 0.858 |
| | | | 0.8 | 0.35 | 0.753 | 0.741 |
| 15 | 100 | 100 | 0 | 1 | 1 | 1 |
| | | | 0.2 | 0.853 | 0.992 | 0.985 |
| | | | 0.5 | 0.426 | 0.902 | 0.894 |
| | | | 0.8 | 0.294 | 0.7775 | 0.754 |
| 5 | 20 | 25 | 0 | 0.384 | 0.349 | 0.262 |
| | | | 0.2 | 0.225 | 0.318 | 0.273 |
| | | | 0.5 | 0.151 | 0.272 | 0.262 |
| | | | 0.8 | 0.107 | 0.222 | 0.208 |

| | | | | Power | | |
|---|---|---|---|---|---|---|
| $p$ | $n1$ | $n2$ | $r$ | $\Lambda^*$ | $G$ | $\tilde{G}$ |
| 10 | 20 | 25 | 0 | 0.514 | 0.484 | 0.401 |
| | | | 0.2 | 0.19 | 0.41 | 0.393 |
| | | | 0.5 | 0.11 | 0.305 | 0.301 |
| | | | 0.8 | 0.072 | 0.225 | 0.211 |
| 15 | 20 | 25 | 0 | 0.553 | 0.603 | 0.485 |
| | | | 0.2 | 0.165 | 0.479 | 0.433 |
| | | | 0.5 | 0.09 | 0.32 | 0.305 |
| | | | 0.8 | 0.056 | 0.222 | 0.225 |
| 5 | 20 | 30 | 0 | 0.391 | 0.344 | 0.289 |
| | | | 0.2 | 0.227 | 0.324 | 0.288 |
| | | | 0.5 | 141 | 0.278 | 0.274 |
| | | | 0.8 | 0.101 | 0.212 | 0.215 |
| 10 | 20 | 30 | 0 | 0.534 | 0.509 | 0.449 |
| | | | 0.2 | 0.184 | 0.434 | 0.404 |
| | | | 0.5 | 0.094 | 0.322 | 0.328 |
| | | | 0.8 | 0.065 | 0.215 | 0.223 |
| 15 | 20 | 30 | 0 | 0.614 | 0.635 | 0.519 |
| | | | 0.2 | 0.162 | 0.477 | 0.458 |
| | | | 0.5 | 0.071 | 0.338 | 0.334 |
| | | | 0.8 | 0.058 | 0.223 | 0.23 |
| 5 | 20 | 50 | 0 | 0.47 | 0.377 | 0.337 |
| | | | 0.2 | 0.251 | 0.325 | 0.332 |
| | | | 0.5 | 0.143 | 0.3 | 0.307 |
| | | | 0.8 | 0.092 | 0.221 | 0.234 |
| 10 | 20 | 50 | 0 | 0.66 | 0.531 | 0.539 |
| | | | 0.2 | 0.21 | 0.427 | 0.459 |
| | | | 0.5 | 0.096 | 0.343 | 0.364 |
| | | | 0.8 | 0.055 | 0.236 | 0.247 |
| 15 | 20 | 50 | 0 | 0.786 | 0.684 | 0.615 |
| | | | 0.2 | 0.164 | 0.522 | 0.52 |
| | | | 0.5 | 0.077 | 0.363 | 0.382 |
| | | | 0.8 | 0.04 | 0.227 | 0.242 |

Table 3.5: The power of $\Lambda^*$, $G$ and $\tilde{G}$ for multivariate gamma distribution in the case of **two groups** using the effect size 1 in Table 3.4.

| | | | | | Power | | |
|---|---|---|---|---|---|---|---|
| $p$ | $n1$ | $n2$ | $n3$ | $r$ | $\Lambda^*$ | $G$ | $\tilde{G}$ |
| 5 | 10 | 10 | 10 | 0 | 0.408 | 0.435 | 0.36 |
| | | | | 0.2 | 0.247 | 0.425 | 0.362 |
| | | | | 0.5 | 0.174 | 0.335 | 0.318 |
| | | | | 0.8 | 0.135 | 0.259 | 0.246 |
| 5 | 50 | 50 | 50 | 0 | 1 | 1 | 0.992 |
| | | | | 0.2 | 0.952 | 0.984 | 0.958 |
| | | | | 0.5 | 0.777 | 0.942 | 0.907 |
| | | | | 0.8 | 0.633 | 0.866 | 0.842 |
| 10 | 50 | 50 | 50 | 0 | 1 | 1 | 1 |
| | | | | 0.2 | 0.953 | 0.997 | 0.996 |
| | | | | 0.5 | 0.674 | 0.973 | 0.963 |
| | | | | 0.8 | 0.458 | 0.88 | 0.872 |
| 15 | 50 | 50 | 50 | 0 | 1 | 1 | 1 |
| | | | | 0.2 | 0.936 | 0.999 | 0.998 |
| | | | | 0.5 | 0.59 | 0.975 | 0.968 |
| | | | | 0.8 | 0.374 | 0.885 | 0.884 |
| 5 | 100 | 100 | 100 | 0 | 1 | 1 | 1 |
| | | | | 0.2 | 0.999 | 1 | 0.999 |
| | | | | 0.5 | 0.985 | 0.998 | 0.995 |
| | | | | 0.8 | 0.934 | 0.992 | 0.99 |
| 10 | 100 | 100 | 100 | 0 | 1 | 1 | 1 |
| | | | | 0.2 | 1 | 1 | 1 |
| | | | | 0.5 | 0.968 | 1 | 0.999 |
| | | | | 0.8 | 0.85 | 0.994 | 0.995 |
| 15 | 100 | 100 | 100 | 0 | 1 | 1 | 1 |
| | | | | 0.2 | 1 | 1 | 1 |
| | | | | 0.5 | 0.942 | 1 | 1 |
| | | | | 0.8 | 0.769 | 0.997 | 0.998 |
| 5 | 20 | 22 | 25 | 0 | 0.88 | 0.823 | 0.699 |
| | | | | 0.2 | 0.607 | 0.731 | 0.657 |
| | | | | 0.5 | 0.368 | 0.632 | 0.588 |
| | | | | 0.8 | 0.236 | 0.504 | 0.486 |

| | | | | | Power | | |
|---|---|---|---|---|---|---|---|
| $p$ | $n1$ | $n2$ | $n3$ | $r$ | $\Lambda^*$ | $G$ | $\tilde{G}$ |
| 10 | 20 | 22 | 25 | 0 | 0.976 | 0.97 | 0.92 |
| | | | | 0.2 | 0.548 | 0.873 | 0.816 |
| | | | | 0.5 | 0.274 | 0.699 | 0.703 |
| | | | | 0.8 | 0.16 | 0.525 | 0.514 |
| 15 | 20 | 22 | 25 | 0 | 0.99 | 0.996 | 0.976 |
| | | | | 0.2 | 0.472 | 0.92 | 0.882 |
| | | | | 0.5 | 0.216 | 0.731 | 0.706 |
| | | | | 0.8 | 0.129 | 0.531 | 0.528 |
| 5 | 20 | 25 | 30 | 0 | 0.916 | 0.844 | 0.723 |
| | | | | 0.2 | 0.626 | 0.766 | 0.696 |
| | | | | 0.5 | 0.395 | 0.65 | 0.636 |
| | | | | 0.8 | 0.263 | 0.517 | 0.506 |
| 10 | 20 | 25 | 30 | 0 | 0.987 | 0.976 | 0.946 |
| | | | | 0.2 | 0.586 | 0.897 | 0.857 |
| | | | | 0.5 | 0.284 | 0.724 | 0.727 |
| | | | | 0.8 | 0.168 | 0.548 | 0.55 |
| 15 | 20 | 25 | 30 | 0 | 0.998 | 0.997 | 0.98 |
| | | | | 0.2 | 0.517 | 0.993 | 0.899 |
| | | | | 0.5 | 0.198 | 0.752 | 0.738 |
| | | | | 0.8 | 0.131 | 0.557 | 0.544 |
| 5 | 20 | 30 | 50 | 0 | 0.979 | 0.917 | 0.845 |
| | | | | 0.2 | 0.742 | 0.847 | 0.792 |
| | | | | 0.5 | 0.465 | 0.738 | 0.705 |
| | | | | 0.8 | 0.287 | 0.597 | 0.58 |
| 10 | 20 | 30 | 50 | 0 | 1 | 0.998 | 0.98 |
| | | | | 0.2 | 0.683 | 0.847 | 0.928 |
| | | | | 0.5 | 0.321 | 0.811 | 0.804 |
| | | | | 0.8 | 0.168 | 0.606 | 0.606 |
| 15 | 20 | 30 | 50 | 0 | 1 | 1 | 0.996 |
| | | | | 0.2 | 0.613 | 0.976 | 0.956 |
| | | | | 0.5 | 0.227 | 0.837 | 0.839 |
| | | | | 0.8 | 0.124 | 0.626 | 0.618 |

Table 3.6: The power of $\Lambda^*$, $G$ and $\tilde{G}$ for multivariate gamma distribution in the case of **three groups** using the effect size 1 in Table 3.4.

# Chapter 4

# Real Data Applications

In this chapter, we use two real genomic data sets to illustrate the multivariate methods presented in Chapter 2.

## 4.1 Application to Real Gene Expression Data

In this Section we applied MANOVA and robustified MANOVA to a real gene expression data.

### 4.1.1 Data Description

The dataset was provided by Dr. Yigal Dror from the Hospital for Sick Children, Toronto, ON. The dataset consists of 54,675 probesets from 22 patients. The patients are divided into three groups. The first group has 9 patients who have a condition known as Shwachman-Diamond Syndrome (SDS). The second group has 5 patients

with Fanconi Anemia (FA), and the other 8 patients have no conditions and they are considered controls.

Shwachman-Diamond Syndrome and Fanconi Anemia are both very rare genetic disorders and they lead to bone marrow failure. Both Fanconi Anemia and Shwachman Diamond syndrome are genetic diseases that are transmitted mainly as autosomal recessive pattern that means both parents have to carry the mutated allele in order to transmit it to their children. In Fanconi Anemia there is a mutation in DNA repair proteins so the patient is prone to bone marrow failure and leukemia. Shwachman-Diamond syndrome also causes bone marrow failure beside exocrine pancreatic insufficiency. The 54,675 probesets represent gene expression levels which are continuous measures; multiple probesets measure the expression level of each gene. Detecting sets of genes that are differentially expressed (DE) among the three groups would help researchers to link genes to diseases and do further investigations.

### 4.1.2 Pre-Processing Step

Usually before primary statistical analysis, a pre-processing step is necessary; The data were provided on the Affymetrix Human Genome U133 plus2 Array. We used some of the R Bioconductor packages (*http* : *//www.bioconductor.org/*) to pre-process the data. First, normalization is done using 'RMA' (Robust Multi-array Average) method implemented in the 'limma' package (Smyth, 2005). Then, we mapped the 54,675 probesets to their corresponding HUGO gene symbols (Human Genome Organisation) using 'biomaRt' package (Steffen et al., 2009; Steffen et al., 2005). We removed probesets that do not have corresponding gene symbols and

used the average expression of the probesets as the expression level for each gene. Different summarization approaches for each gene can be used (Liu and Zhang, 2010; Miller et al., 2011).

After mapping probesets to genes, the resulting data had 19,732 genes. We then filtered out genes that do not vary much across groups using some variation criteria; we filtered out genes based on interquartile range (IQR) and retained genes with the highest 1000 IQR values ($\approx 5\%$ of the total number of genes). Our analyses focus on these 1000 genes.

### 4.1.3 Clustering Data to Obtain a Group of Genes

Standardizing the data is often used before applying cluster analysis. We standardized each gene (row) to have a mean of zero and a variance of one. Then, we applied cluster analysis to the resulting dataset which has $p$= 1000 genes and $n$= 22 patients. First, we determined the number of clusters using the PRE plot method that is discussed in Chapter 2. We used 'cluspect' R-package which can be downloaded from $http://www.uhnres.utoronto.ca/labs/tritchler/$ (Fallah et al., 2008). We estimated the number of clusters to be 4 based on the PRE plot. This plot is represented in Figure 4.1. We applied $k$-means cluster analysis with $k = 4$ clusters using 'cluster' R-package (Maechler et al., 2012). The resulting clusters contain 359, 135, 271 and 235 genes. Since we are interested in providing an illustrative example for the MANOVA and robustified MANOVA, we selected the cluster with the fewest number of genes.

We repeated the PRE Plot and $k$-means analysis only for the second cluster

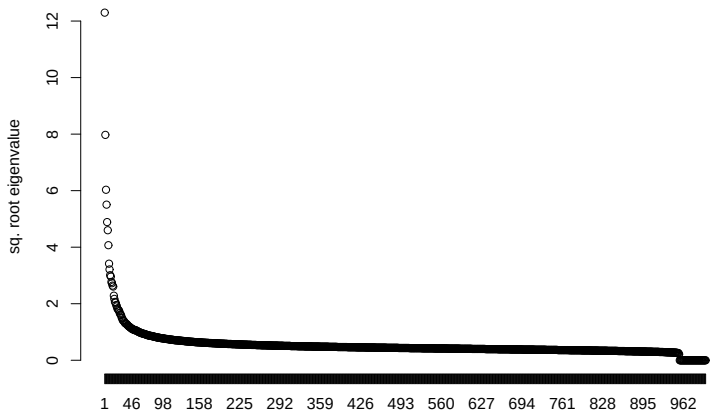



Figure 4.1: The PRE plot of 1000 genes from the gene expression dataset of SickKids Hospital.

which has 135 genes. The PRE plot suggested 10 clusters (see Figure 4.2). Then, we applied $k$-means with $k$=10 on the 135 genes, the resulting clusters have 21, 4, 15, 18, 4, 22, 15, 11, 13 and 12 genes.

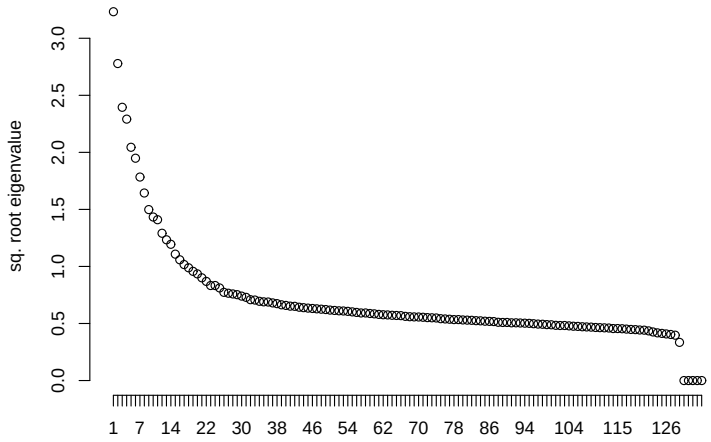



Figure 4.2: The PRE plots of 135 genes.

### 4.1.4 MANOVA and Robustified MANOVA

To identify differentially expressed genes among the three groups (SDS, FA, and Control), we applied MANOVA and robustified MANOVA. We considered the clusters that have 4 and 11 genes and the results are summarized in Table 4.1.

| Cluster size | Genes Names | $\Lambda^*$ | $G$ | $\tilde{G}$ |
|---|---|---|---|---|
| (Cluster#2) 4 | KIF18B, NEURL1B, MCM2, FOS | 0.089 | 0.629 | 0.463 |
| (Cluster#5) 4 | NRGN, ANKRD10-IT1, TLN1, FAM118A | 0.073 | 0.212 | 0.287 |
| (Cluster#8) 11 | VPREB1, TCL1A, MME, SPIB, IGLL1, PCDH9, FCRL1, TOX2, FAM129C, AKAP12, PBK | 0.272 | 0.012 | 0.031 |

Table 4.1: $P$-values of $\Lambda^*$, $G$ and $\tilde{G}$ test statistics for three different clusters.

From Table 4.1, we observe that the average gene expression of the 11 genes corresponding to cluster 8 are significantly different across the three groups when the robustified MANOVA tests are applied, but not with the standard Wilks' Lambda. For the smaller clusters (i.e., clusters 2 and 5 which contain each 4 genes), the test based on Wilks' Lambda appears to suggest a marginal significance, but the robust tests both show clear lack of significance.

Figures 4.3, 4.4 and 4.5 visualize the data in clusters number 2, 5 and 8, respectively, using heatmap.

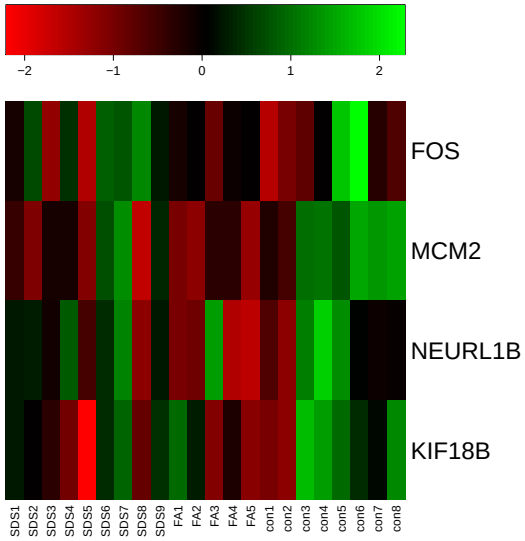



Figure 4.3: A heatmap for cluster # 2 where rows correspond to genes in the same cluster and columns correspond to subjects. The red color represents under expression and the green corresponds to over expression.

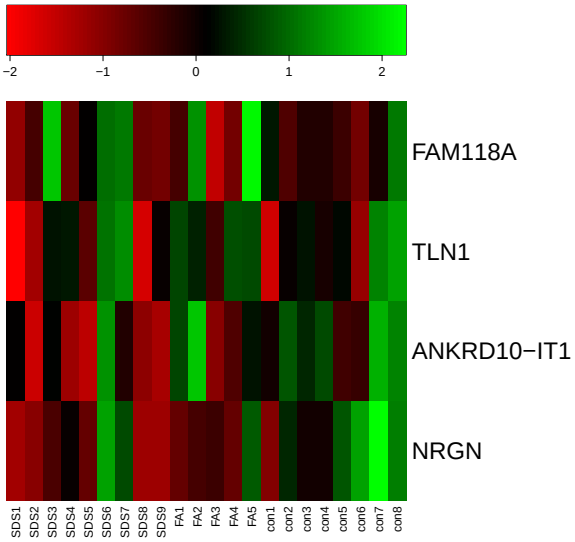



Figure 4.4: A heatmap for cluster # 5 where rows correspond to genes in the same cluster and columns correspond to subjects. The red color represents under expression and the green corresponds to over expression.

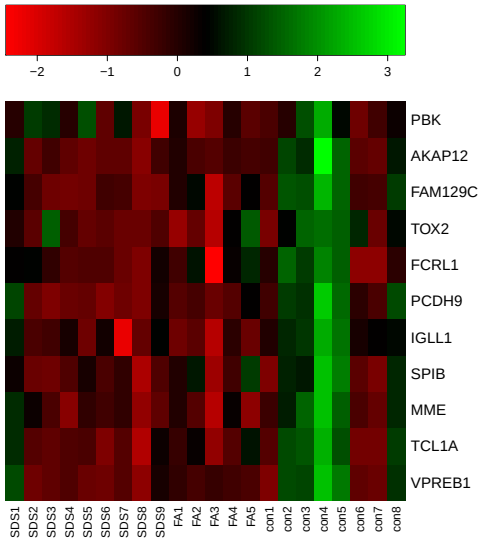



Figure 4.5: A heatmap for cluster # 8 where rows correspond to genes in the same cluster and columns correspond to subjects. The red color represents under expression and the green corresponds to over expression.

## 4.2 Application to Real Genotype Data

In this Section, we applied GSEA to a real high-dimensional genotype data.

### 4.2.1 Background

The data set was provided by the organizers of the Genetic Analysis Workshop 18 (GAW 18) which focused on analysis of whole genome sequence data in a blood pressure study. Worldwide, hypertension contributes to over 10 million deaths and it affects one-third of the adult population per year (Levy et al., 2009). It was predicted that the incidence of hypertension among adults in 2025 will reach 1.56 billion and contribute to about 54% of strokes and 47% of ischemic heart disease.

Furthermore, it is a major risk factor for cardiovascular disease (Lin et al., 2011). Several factors including genetic, environmental, and demographics play major role in the development of hypertension. However, it is believed that 30-60% of the variability in blood pressure (BP) is inherited (Levy et al., 2009).

Many genome wide association studies (GWAS) have been conducted to identify single-nucleotide polymorphism (SNPs) that are significantly associated with systolic and diastolic blood pressure (SBP, DBP) and/or hypertension. Meta-analysis findings of the Global BPgen (Global Blood Pressure Genetics) consortium ($n$=34,433) and CHARGE (The Cohorts for Heart and Aging Research in Genome Epidemiology) consortium ($n$=29,136) based on populations of European ancestry identified four loci significantly associated with SBP (ATP2B1, CYP17A1, PLEKHA7, SH2B3), six for DBP (ATP2B1, CACNB2, CSK-ULK3, SH2B3, TBX3-TBX5, ULK4) and one for hypertension (ATP2B1) (Levy et al., 2009).

However, a genome-wide association study by Adeyemo et al. (Adeyemo et al., 2009) based on a population of African Americans ($n$=1,017) identified significant loci for SBP in or near the genes: PMS1, SLC24A4, YWHA7, IPO7, and CACANA1H while no significantly loci were discovered to be associated with DBP or hypertension.

Our main focus is to test the effects of both rare and common variants on systolic blood pressure (SBP), diastolic blood pressure (DBP), the pulse pressure which is the difference between SBP and DBP, and mean arterial pressure (defined as MAP = (2/3) DBP + (1/3) SBP) by applying GSEA.

### 4.2.2 Phenotype, Covariate and Genotype Data Description

Phenotypes were taken at four time points and included systolic blood pressure (SBP) and diastolic blood pressure (DBP) measurements and hypertension. The following covariates were also provided: age, smoking status, antihypertensive medications usage, and gender. In our analysis, we used the baseline data. Among 157 unrelated individuals that were provided, we had 129 individuals who had been genotyped. Table 4.2 provides descriptive statistics for the outcome variables as well as the covariates.

| **Variable** | **Summary measure** |
|---|---|
| **SBP** | 128.4 ±21.8 |
| **DBP** | 71.8 ±9.2 |
| **MAP** | 90.7 ±11.6 |
| **SBP-DBP** | 56.6 ±19 |
| **Hypertension (Yes, No)** | 129 (29.5) |
| **AGE** | 52.9 ±15.6 |
| **SEX (FEMALE)** | 129 (60.5) |
| **Medications use (Yes, No)** | 129 (20.2) |
| **Smoking status (Yes, No)** | 129 (24.8) |

Table 4.2: Descriptive statistics for GAW 18 data. Summary measure refers to mean ± SD for continuous variables; $n(\%)$ for categorical variables

Genotype data were provided only for odd-numbered autosomal chromosomes and we focused only on variants on chromosome 3 (as suggested by the GAW 18 organizers to allow comparisons of findings with other GAW 18 contributions).

### 4.2.3 Pathway-Based Analysis

We considered four phenotypes of interest SBP, DBP, SBP-DBP and MAP and we performed pathway-based analysis based on GSEA method. We followed the following steps (Wang et al., 2007; Beyene et al., 2009):

1. **Mapping SNPs to genes:**

   We only considered the SNPs that are located on known genes and discarded all the other SNPs on the intergenic regions. Among the 1,215,296 SNPs on chromosome three, 523,147 SNPs were mapped to 1224 known genes using NCBI2R (which is an R package that maps SNPs to genes based on physical distance).

2. **Obtaining test statistics for genes:**

   We considered both rare and common variants and other covariates (age, smoking status, medications use and gender) to assign a test statistic for each gene. VW-TOW (Variable Weight Test for testing the effect of an Optimally Weighted combination of variants)(Sha et al., 2012) was used to construct test statistics and their $p$-values. A brief description of this method is given in Chapter 2 of this thesis.

3. **Pathway analysis:**

   We ranked all the genes ($N$) based on their significance level. From step 2, we had N=1187 genes for SBP and DBP and $N$=1188 genes for (SBP-DBP) and MAP. Using the GSEA method (Subramanian et al., 2005), we evaluated

the significance of predefined gene sets/pathways obtained from online pathway databases (The Molecular Signatures Database (MSigDB)). We used the c2 curated gene sets (v3.1) which are compiled from online pathway databases, publications in PubMed, and knowledge of domain experts and it can be downloaded from

$http://www.broadinstitute.org/gsea/msigdb/collection_details.jsp\#C2$.

This database consists of 4,850 gene sets but we only considered 3,638 sets that had at least one gene from chromosome three and at least 10 genes in total. 69.1% of the 3,638 pathways contain between 1 to 5 genes on chromosome 3, while 24.1% of the pathways have between 6 to 20 genes and 6.8% have between 21 to 140 genes on the same chromosome.

We calculated the enrichment score ($ES$) for each gene set/pathway using a weighted Kolmogorov-Smirnov-like running-sum statistic. Then we adjusted for different sizes of genes using 1000 gene-based permutations ($\pi$) and calculated the normalized enrichment score ($NES$) for each set ($S$). Similar to the approach used by Beyene et al. (2009), we estimated the significance level of NES for each gene set/pathway using the gene-based permutation approach and obtained the empirical p-values of the NES. We used 1000 gene-set permutations and then we considered the gene set/pathway to be significantly enriched if its FDR $q$-value is less than 0.05. We implemented the analysis using the GseaPreranked tool included in the GSEA software (Subramanian et al., 2005; Mootha et al., 2003).

### 4.2.4 Results and Discussion

Considering common and rare variants from chromosome three with other covariates, and applying GSEA to our data, we ranked the top 10 gene sets/pathways based on their FDR $q$-values for each phenotype. These ranked genes are listed in Tables 4.3, 4.4, 4.5 and 4.6 for MAP, (SBP-DBP), SBP and DBP phenotypes respectively. We found that no gene sets were enriched when we considered SBP or DBP separately. However, we were able to identify one significantly enriched pathway from c2 curated gene sets (see Table 4.3) for mean arterial pressure (MAP). Interestingly, the same pathway was declared to be significantly enriched for the difference between SBP and DBP phenotype (see Table 4.4).

We identified the same gene pathway (KOYAMA_SEMA3B_TARGETS_DN) to be significantly enriched for both phenotypes, and this pathway had been shown to be related to different kinds of cancer (Marsit et al., 2005; Joseph et al., 2010). In this pathway, 12 out of 18 genes on chromosome three contributed to the enrichment score and the most interesting gene in this pathway is CD47. Several articles (Isenberg et al., 2009; Bauer et al., 2010) reported that this gene regulates blood pressure.

Since our pathway-based analysis is restricted to genes on chromosome three, the number of pathways that was used for analysis exceeded the number of genes, which can have important implications in interpreting our findings. The results from our analyses should be interpreted cautiously.

| PATHWAY NAME | # genes | ES | NES | FDR $q$-val |
|---|---|---|---|---|
| KOYAMA_SEMA3B_TARGETS_DN | 18 | 0.611 | 2.226 | 0.04 |
| HUANG_GATA2_TARGETS_UP | 11 | 0.559 | 1.769 | 0.903 |
| ONO_FOXP3_TARGETS_DN | 5 | 0.724 | 1.757 | 0.929 |
| BENPORATH_ES_2 | 4 | 0.859 | 1.938 | 0.967 |
| CARDOSO_RESPONSE_TO_GAMMA_RADIATION_AND_3AB | 3 | 0.865 | 1.77 | 0.98 |
| ZHANG_TLX_TARGETS_60HR_UP | 22 | 0.423 | 1.625 | 0.981 |
| LAIHO_COLORECTAL_CANCER_SERRATED_DN | 1 | 0.891 | 1.202 | 0.992 |
| PID_INTEGRIN2_PATHWAY | 1 | 0.915 | 1.203 | 0.992 |
| LIU_TARGETS_OF_VMYB_VS_CMYB_DN | 5 | 0.501 | 1.203 | 0.992 |
| LIM_MAMMARY_STEM_CELL_DN | 21 | 0.317 | 1.203 | 0.992 |

Table 4.3: The top 10 gene sets/pathways from c2 curated gene sets ranked by FDR $q$-values for MAP. # genes is the number of genes on chromosome 3.

| PATHWAY NAME | # genes | ES | NES | FDR $q$-val |
|---|---|---|---|---|
| KOYAMA_SEMA3B_TARGETS_DN | 18 | 0.614 | 2.227 | 0.042 |
| ZHAN_MULTIPLE_MYELOMA_CD1_AND_CD2_UP | 4 | 0.83 | 1.867 | 0.822 |
| TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_UP | 7 | 0.691 | 1.888 | 0.823 |
| BREDEMEYER_RAG_SIGNALING_NOT_VIA_ATM_DN | 4 | 0.811 | 1.844 | 0.871 |
| LI_INDUCED_T_TO_NATURAL_KILLER_DN | 7 | 0.693 | 1.907 | 0.892 |
| BENPORATH_ES_2 | 4 | 0.858 | 1.941 | 0.917 |
| TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_SUSTAINED_IN_GRANULOCYTE_UP | 3 | 0.869 | 1.762 | 0.958 |
| CHANG_CORE_SERUM_RESPONSE_DN | 21 | 0.29 | 1.091 | 0.976 |
| WONG_ENDMETRIUM_CANCER_UP | 2 | 0.625 | 1.082 | 0.976 |
| NOUZOVA_TRETINOIN_AND_H4_ACETYLATION | 18 | 0.299 | 1.09 | 0.976 |

Table 4.4: The top 10 gene sets/pathways from c2 curated gene sets ranked by FDR $q$-values for the difference between SBP and DBP. # genes is the number of genes on chromosome 3.

| PATHWAY NAME | # genes | ES | NES | FDR $q$-val |
|---|---|---|---|---|
| GOTTWEIN_TARGETS_OF_KSHV_MIR_K12_11 | 6 | 0.588 | 1.554 | 0.623 |
| SMIRNOV_RESPONSE_TO_IR_6HR_UP | 6 | 0.583 | 1.554 | 0.628 |
| PID_SHP2_PATHWAY | 3 | 0.738 | 1.542 | 0.629 |
| JIANG_VHL_TARGETS | 6 | 0.588 | 1.55 | 0.63 |
| SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A12 | 18 | 0.408 | 1.555 | 0.631 |
| PLASARI_TGFB1_SIGNALING_VIA_NFIC_10HR_UP | 5 | 0.628 | 1.543 | 0.632 |
| CHEN_PDGF_TARGETS | 4 | 0.668 | 1.545 | 0.632 |
| PID_IGF1_PATHWAY | 3 | 0.759 | 1.542 | 0.633 |
| NAKAMURA_TUMOR_ZONE_PERIPHERAL_VS_CENTRAL_DN | 32 | 0.349 | 1.555 | 0.633 |
| PID_BCR_5PATHWAY | 3 | 0.764 | 1.557 | 0.633 |

Table 4.5: The top 10 gene sets/pathways from c2 curated gene sets ranked by FDR $q$-values for SBP. # genes is the number of genes on chromosome 3.

| PATHWAY NAME | # genes | ES | NES | FDR $q$-val |
|---|---|---|---|---|
| PHONG_TNF_RESPONSE_VIA_P38_COMPLETE | 13 | 0.583 | 1.743 | 0.58 |
| DELYS_THYROID_CANCER_DN | 11 | 0.612 | 1.748 | 0.599 |
| CORRE_MULTIPLE_MYELOMA_DN | 3 | 0.897 | 1.733 | 0.603 |
| SHEPARD_BMYB_MORPHOLINO_UP | 10 | 0.609 | 1.712 | 0.617 |
| LI_INDUCED_T_TO_NATURAL_KILLER_UP | 17 | 0.537 | 1.724 | 0.617 |
| WILCOX_PRESPONSE_TO_ROGESTERONE_UP | 6 | 0.705 | 1.7 | 0.621 |
| WAMUNYOKOLI_OVARIAN_CANCER_LMP_DN | 13 | 0.562 | 1.674 | 0.632 |
| KEGG_RENIN_ANGIOTENSIN_SYSTEM | 3 | 0.932 | 1.761 | 0.632 |
| OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_DN | 11 | 0.605 | 1.751 | 0.633 |
| REACTOME_POST_TRANSLATIONAL_PROTEIN_MODIFICATION | 16 | 0.535 | 1.685 | 0.644 |

Table 4.6: The top 10 gene sets/pathways from c2 curated gene sets ranked by FDR $q$-values for DBP. # genes is the number of genes on chromosome 3.

# Chapter 5

# Summary and Future Directions

## 5.1 Summary

Multivariate methods are very useful especially in genomic applications due to the complexity of the data. Using simulations, we compared the performance of three test statistics (Wilks' Lambda and two robust test statistics) in this thesis with respect to Type I error rate and power. Although Type I error rates did not show any remarkable difference between the three test statistics, the two robustified MANOVA test statistics resulted in higher power when data was generated from multivariate gamma distribution. We applied the methods to gene expression data.

We also applied a method called Gene Set Enrichment Analysis (GSEA) to a genotype data. Gene-set enrichment analysis considers multiple genes that are related biologically. In our data, we identified one enriched gene set/pathway

with two clinically important blood pressure related phenotypes: 1) the mean arterial pressure (MAP), and 2) the difference between systolic blood pressure (SBP) and diastolic blood pressure (DBP). In our analysis, we included only 129 unrelated individuals. Sample size plays a major role in identifying enriched gene sets/pathways and this could explain the lack of significant pathways in our analysis.

## 5.2 Future Directions

In the future, it would be interesting to consider more realistic scenarios in order to compare the power of the three test statistics. For example, one could vary the correlations between pairs of variables in each group as well as varying the variances. Xu and Cui (2008) modified the permutation procedure by applying quantile normalization to the permuted distribution, which we did not use in this thesis. In future work we can use this method to estimate the null distribution of the two test statistics in the robustified MANOVA and compare it with the results presented in this thesis. One also could compare other non parametric tests such as the rank transformed Wilks' Lambda (Nath and Pavur, 1985) and the robust Wilks' Lambda which is based on the Minimum Covariance Determinant (MCD) estimator (Todorov and Filzmoser, 2010). Future studies can be done by applying GSEA on large family-based genotype data where incorporating both rare and common variants, taking into account the correlations between individuals and increasing the sample size may lead

to new discoveries.

# Appendices

## R - Code

Function to generate $P$-variate gamma distribution

```
# This function was provided by Dr. Ahmed Hossain.

rmvgamma=function(n,p,shape,rate,rho,scale)
{
  stopifnot(length(rho)==1,length(shape)==1,length(rate)==1,
            length(scale)==1,length(n)==1,length(p)==1,
            rate>0, shape>0, rho>=0, rho<=1, scale>0, n>=0, p>0
  )
  n=round(n);p=ceiling(p)
  theta=rate*rho/(1-rho)
  k=rnbinom(n,shape,rate/(rate+theta))
  matrix(rgamma(p*n,shape+k,rate+theta),n)
}
```

Function for MANOVA

```
#function to do MANOVA, it returns Wilk's lambda test statistic and its pvalue
doMANOVA<-function(groupsmatrix,data)
{
  group<-data[,1]
```

```
  mantable<-manova(groupsmatrix~group,data=data)
  resultWilks<-summary(mantable,test="Wilks")
  resultPillai<-summary(mantable,test="Pillai")
  wilkstest<-resultWilks$stats[1,2]
  Wpvalue<-resultWilks$stats[1,6]
  return(c(wilkstest,Wpvalue))
}
```

Function for Robustified MANOVA

# This function was provided by Dr. John R. Steven.

```
# Define rMANOVA function; this takes three arguments:
# Y is matrix of PM values for a given gene
# trt1 is an index vector of which columns of Y correspond to treatment 1
# trt2 is an index vector of which columns of Y correspond to treatment 2
# The function returns the test statistic G (both the trace and median results)

DorMANOVA <- function(X,trt1,trt2)
{
  X<-t(X)   # we take the transpose because this function considers the rows as
  mu1 <- apply(X[,trt1],1,median)
  mu2 <- apply(X[,trt2],1,median)

  mum <- apply(X,1,median)
  v1 <- X[,trt1]-mu1
  v2 <- X[,trt2]-mu2
  V1 <- array(dim=c(nrow(X),nrow(X),length(trt1)))
  for(a in 1:length(trt1))
  {
    V1[,,a] <- (v1[,a]-mu1)%*%t(v1[,a]-mu1)
  }
  W1 <- apply(V1,c(1,2),median)
  V2 <- array(dim=c(nrow(X),nrow(X),length(trt2)))
  for(a in 1:length(trt2))
  {
    V2[,,a] <- (v2[,a]-mu2)%*%t(v2[,a]-mu2)
  }
```

```
  W2 <- apply(V2,c(1,2),median)

  n1=length(trt1)
  n2=length(trt2)
  W <- n1*W1 + n2*W2
  B <- n1*(mu1-mum)%*%t(mu1-mum) + n2*(mu2-mum)%*%t(mu2-mum)
  Test1<-sum(diag(B))/sum(diag(W))
  Test2<-median(diag(B))/median(diag(W))
  return( c(  Test1 ,Test2  ))
}
```

Function for VW-TOW test statistic

# This function can be downloaded from:

*http* : *//www.math.mtu.edu/ shuzhang/software.html*

```
##  TOW is an association test to test association   ##
## between a phenotype and a optimal weighted      ##
## combination of variants (can be rare and common)  ##
## in a genomic region.             ##
## y represents trait values(a vector). x represents  ##
## genotype (number of minor allele) matrix. Each row ##
## represents an individual and each column represents ##
## a SNP. xc represents the covariate matrix. Each row ##
## represents an individual and each column represents ##
## a covriate.                      ##
TOW=function(y,x,xc,numper)
{
  q=colMeans(x)/2
  a=which(q>0)
  x=x[,a]
  y=lm(y~xc)$resid
  x=lm(x~xc)$resid
  T1per=rep(0,numper)
  m=ncol(x)
  aa=apply(x,2,var)
  T1=sum(cov(y,x)^2/aa)
  for(i in 1:numper)
```

```
  {
    yper=sample(y)
    T1per[i]=sum(cov(yper,x)^2/aa)
  }
  p<-sum(T1<T1per)/numper
c(T1,p)
}

## VW-TOW is a variable weight version of TOW.     ##
## y represents trait values(a vector). x represents  ##
## genotype (number of minor allele) matrix. Each row ##
## represents an individual and each column represents ##
## a SNP. xc represents the covariate matrix. Each row ##
## represents an individual and each column represents ##
## a covriate. q0 is the threshold of MAF of rare      ##
## variants.                        ##
#######################################################################

VW_TOW=function(y,x,xc,q0,numper)
{
  q=colMeans(x)/2
  a=which(q>0)
  x=x[,a]
  m=ncol(x)
  n=nrow(x)
  q=colMeans(x)/2
  a0=which(q<q0)
  y=lm(y~xc)$resid
  x=lm(x~xc)$resid
  x0=x
  ############################
  T1per=rep(0,numper)
  x1=x=x0[,a0]
  aa1=aa=apply(x,2,var)
  T1=sum(cov(y,x)^2/aa)
  ############################
  T2per=rep(0,numper)
  x=x0[,-a0]
```

```
  nn=ncol(x0)-length(a0)
  if(nn==1) aa=var(x)
  else aa=apply(x,2,var)
  T2=sum(cov(y,x)^2/aa)
  for(i in 1:numper)
  {
    yper=sample(y)
    T2per[i]=sum(cov(yper,x)^2/aa)
    T1per[i]=sum(cov(yper,x1)^2/aa1)
  }
  ##############################
  mm=20
  sd1=sd(T1per)
  sd2=sd(T2per)
  T1=c(T1,T1per)/sd1
  T2=c(T2,T2per)/sd2
  lambda=c(0:mm)/mm
  T=T1%*%t(1-lambda)+T2%*%t(lambda)
  T0=apply(T,2,rank)
  T=apply(T0,1,max)
  pv=sum(T[2:(numper+1)]>T[1])/numper+sum(T[2:(numper+1)]==T[1])/numper/2
  c(T[1],pv)
}
```

# Bibliography


[1] Adeyemo, A. et al. (2009). A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genetics* ,**5**, No. 7, e1000564.

[2] Bartlett, M.S. (1954). A note on multiplying factors for various $\chi 2$ approximations. *Journal of the Royal Statistical Society. Series B (Methodological)* **16**, No. 2, 296-298.

[3] Bauer, E. M. et al. (2010). Thrombospondin-1 supports blood pressure by limiting eNOS activation and endothelial-dependent vasorelaxation. *Cardiovascular Research* **88**, No. 3, 471-481.

[4] Benjamini, Y. and Hochberg, Y. (1995).Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B.* **57**, 289-300.

[5] Beyene, J. et al. (2009). Pathway-based analysis of a genome-wide case-control association study of rheumatoid arthritis. *BMC Proceedings* **3**, No. 7, S128.

[6] Crick, F. (1970). Central dogma of molecular biology. *Nature* **227**, 561-563.

[7] Dudoit, S. et al. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, No. 1, 71-103.

[8] Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics* **1**, No. 1, 107-129.

[9] Fallah, S., Tritchler, D. and Beyene, J. (2008). Estimating number of clusters based on a general similarity matrix with application to microarray data. *Statistical Applications in Genetics and Molecular Biology* **7**, No. 1, Article 24.

[10] Fanale, D. et al. (2012). Breast cancer genome-wide association studies: there is strength in numbers. *Oncogene* **31**, No. 17, 2121-2128.

[11] Francis, S. et al. (2001). Implications of the Human Genome Project for Medical Science. *The Journal of the American Medical Association* **285**, No. 5, 540-544.

[12] Gavin, L. S. (2012). permute: Functions for generating restricted permutations of data. R package version 0.7-0. http://CRAN.R-project.org/package=permute.

[13] Holmans, P. (2010). Statistical methods for pathway analysis of genome wide data for association with complex genetic traits. *Computational Methods For Genetics Of Complex Traits* **72**, 141-179.

[14] Hossain, A., Willan, A. and Beyene, J. (2013). A flexible nonparametric approach to find candidate genes associated to disease in microarray experiments. *Journal of Bioinformatics and Computational Biology* **11**, No.2, 1250021.

[15] Hung, JH. et al. (2012). Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics* **13**, No. 3, 281-91.

[16] Isenberg, J. S. et al. (2009). Thrombospondin-1 and CD47 regulate blood pressure and cardiac responses to vasoactive stress. *Matrix Biology* **28**, No. 2, 110-119.

[17] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pear- son Prenctice Hall, New Jersey. Sixth Edition.

[18] Joseph, D. et al. (2010). Hormonal regulation and distinct functions of semaphorin-3B and semaphorin-3F in ovarian cancer. *Molecular Cancer Therapeutics* **9**, No. 2, 499-509.

[19] Levy, D. et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics* **41**, No. 6, 677-687.

[20] Lin, Y. et al. (2011). Genetic variations in CYP17A1, CACNB2 and PLEKHA7 are associated with blood pressure and/or hypertension in She ethnic minority of China. *Atherosclerosis* **219**, No. 2, 709-714.

[21] Liu, ZP. and Zhang, XS. (2010). Effects of Multiple Probesets in Affymetrix GeneChips on Identifying Differentially Expressed Genes in

iPS Cells. The Fourth International Conference on Computational Systems Biology. pp 187-195.

[22] Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**. pp 281-297.

[23] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2012). cluster: Cluster Analysis Basics and Extensions. R package version 1.14.2.

[24] Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology* **9**, No. 34.

[25] Marsit, C. J. et al. (2005). The race associated allele of Semaphorin 3B (SEMA3B) T415I and its role in lung cancer in African-Americans and Latino-Americans. *Carcinogenesis* **26**, No. 8, 1446-1449.

[26] McCarthy, M. I. and Zeggini, E. (2009). Genome-wide association studies in type 2 diabetes. *Current Diabetes Reports* **9**, No. 2, 164-171.

[27] Miller, J. A., et al. (2011). Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics* **12**, No. 322.

[28] Minhajuddin, A.T.M, Harris, I. R. and Schucany, W. R. (2004). Simulating multivariate distributions with specific correlations. *Journal of Statistical Computation and Simulation* **74**, No. 8, 599-607.

[29] Mootha, V. K. et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, No. 3, 267-73.

[30] Nath, R. and Pavur, R. (1985). A new statistic in the one-way multivariate analysis of variance. *Computational Statistics and Data Analysis* **2**, No. 4, 297-315.

[31] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**, No. 1, Article 32.

[32] Sha, Q. et al. (2012). Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genetic Epidemiology* **36**, No. 6, 561-571.

[33] Shi, J. and Walker, M. G. (2007). Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *Current Bioinformatics* **2**, No. 2.

[34] Smyth, GK (2005). Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pages 397-420.

[35] Steffen, D., Yves, M., Arek, K., Sean, D., Bart, D. M., Alvis, B. and

Wolfgang, H. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.*Bioinformatics* **21**, 3439-3440.

[36] Steffen, D., Paul, T. S., Ewan, B. and Wolfgang, H. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.*Nature Protocols* **4**, 1184-1191.

[37] Stevens, J. R., Bell, J. L., Aston, K. I. and White, K. L. (2010). A comparison of probe-level and probeset models for small-sample gene expression data. *BMC Bioinformatics* **11**, 281.

[38] Stewart, C. N. and Excoffier, L. (1996). Assessing population genetic structure and variability with RAPD data: application to vaccinium macrocarpon (American Cranberry). *Evolutionary Biology* **9**, No. 2, 153-171.

[39] Subramanian, A. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Briefings in Bioinformatics* **102**, No. 43, 15545-15550.

[40] Todorov, V. and Filzmoser, P. (2010). Robust statistic for the one-way MANOVA. *Computational Statistics and Data Analysis* **54**, No. 1, 37-48.

[41] Van Aelst, S. and Willems, G. (2011). Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association* **106**, No. 494, 706-718.

[42] Venables, W. N. Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

[43] Wang, K., Li, M. and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics* **81**, No. 6, 1278-1283.

[44] Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics* **17**, No. 3, 257-281.

[45] Xu, J. and Cui, X. (2008). Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. *Bioinformatics* **24**, No. 8, 1056-1062.