2012

Application of order restricted statistical inference and hidden Markov modeling to problems in biology and genomics

Heng Wang
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Biostatistics Commons, and the Mathematics Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Graduate College at Digital Repository @ Iowa State University. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
Application of order restricted statistical inference
and hidden Markov modeling to problems in biology and genomics

by

Heng Wang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Dan Nettleton, Major Professor
Song Xi Chen
Kenneth J. Koehler
Peng Liu
Huaiqing Wu

Iowa State University
Ames, Iowa
2012

Copyright © Heng Wang, 2012. All rights reserved.
DEDICATION

To my family
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1. General Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Order Restricted Statistical Inferences</td>
<td>1</td>
</tr>
<tr>
<td>1.3 the Intersection-Union Test</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Gene Set Testing</td>
<td>4</td>
</tr>
<tr>
<td>1.5 Copy Number Variation Detection Using Next Generation Sequencing Data</td>
<td>5</td>
</tr>
<tr>
<td>1.5.1 Hidden Markov Models</td>
<td>6</td>
</tr>
<tr>
<td>1.5.2 The Expectation-Maximization Algorithm</td>
<td>7</td>
</tr>
<tr>
<td>1.6 Organization</td>
<td>7</td>
</tr>
<tr>
<td>References</td>
<td>9</td>
</tr>
<tr>
<td>CHAPTER 2. Testing a Union-of-Cones Null Hypothesis for the Identification of Heterosis</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>13</td>
</tr>
<tr>
<td>2.2 The Transformation</td>
<td>16</td>
</tr>
<tr>
<td>2.3 The Likelihood Ratio Test</td>
<td>18</td>
</tr>
<tr>
<td>2.4 The Intersection-Union Test</td>
<td>19</td>
</tr>
<tr>
<td>2.5 More Powerful Tests</td>
<td>22</td>
</tr>
<tr>
<td>2.5.1 Uniformly More Powerful Tests when the Alternative Parameter Space is a Single Cone</td>
<td>23</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.5.2 Uniformly More Powerful Tests when the Alternative Parameter Space is a Union of Two Cones</td>
<td>25</td>
</tr>
<tr>
<td>2.5.3 Comments on the Uniformly More Powerful Tests</td>
<td>26</td>
</tr>
<tr>
<td>2.6 An Example</td>
<td>27</td>
</tr>
<tr>
<td>2.7 Summary and Conclusions</td>
<td>30</td>
</tr>
<tr>
<td>2.8 Appendix</td>
<td>31</td>
</tr>
<tr>
<td>2.8.1 Derivation of the LRT</td>
<td>31</td>
</tr>
<tr>
<td>2.8.2 Application of Zelterman’s Test on Our Testing Problem</td>
<td>38</td>
</tr>
<tr>
<td>References</td>
<td>42</td>
</tr>
<tr>
<td>CHAPTER 3. Testing for Monotonic Changes in Multivariate Gene Expression Distributions</td>
<td>44</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>44</td>
</tr>
<tr>
<td>3.1.1 Background and Motivation</td>
<td>44</td>
</tr>
<tr>
<td>3.1.2 A Formal Description of the Problem of Interest</td>
<td>45</td>
</tr>
<tr>
<td>3.1.3 Previous Studies</td>
<td>46</td>
</tr>
<tr>
<td>3.1.4 Organization</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Proposed Methods</td>
<td>47</td>
</tr>
<tr>
<td>3.2.1 Method I – Index Weighted Sum Permutation Test (IWSPT)</td>
<td>47</td>
</tr>
<tr>
<td>3.2.2 Methods Based on Combining P-Values from Individual Genes within a Gene Set</td>
<td>49</td>
</tr>
<tr>
<td>3.3 Other Methods that Do Not Focus on Monotonic Alternatives</td>
<td>53</td>
</tr>
<tr>
<td>3.3.1 Multiresponse Permutation Procedure (MRPP)</td>
<td>53</td>
</tr>
<tr>
<td>3.3.2 Energy Statistic (ES)</td>
<td>54</td>
</tr>
<tr>
<td>3.3.3 Permutation Based ANOVA Test (PANOVA)</td>
<td>54</td>
</tr>
<tr>
<td>3.4 Simulation Studies</td>
<td>55</td>
</tr>
<tr>
<td>3.4.1 The First Simulation Study</td>
<td>55</td>
</tr>
<tr>
<td>3.4.2 The Second Simulation Study</td>
<td>57</td>
</tr>
<tr>
<td>3.4.3 The Third Simulation Study</td>
<td>57</td>
</tr>
<tr>
<td>3.5 An Example with Real Data Analysis</td>
<td>58</td>
</tr>
<tr>
<td>3.6 Discussion and Conclusions</td>
<td>59</td>
</tr>
<tr>
<td>3.7 Appendix</td>
<td>61</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Normalized estimated means for genes A through F, and comparisons of p-values using LRT, IUT and overall F-test.</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>Monotonic gene set detections using different statistics with normal random variables. Each gene set contains 6 genes.</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Monotonic gene set detections using different statistics with normal random variables. Each gene set contains 20 genes.</td>
<td>56</td>
</tr>
<tr>
<td>3.3</td>
<td>Monotonic Gene Detections Using Different Statistics with Normal Random Variables</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>Monotonic Gene Detections Using Different Statistics using Gamma Random Variables</td>
<td>58</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparison between the m-HMM and the original HMM</td>
<td>81</td>
</tr>
<tr>
<td>4.2</td>
<td>Comparison between the m-HMM and the SegSeq</td>
<td>81</td>
</tr>
<tr>
<td>4.3</td>
<td>Absent State in Chromosome 6</td>
<td>88</td>
</tr>
<tr>
<td>4.4</td>
<td>Large CNV Identifications</td>
<td>91</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The null and alternative spaces in the dose-response example</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>The null and alternative spaces in the heterosis example</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>The null (unshaded) and alternative (shaded) parameter spaces in terms of $\mu$ and $\theta$</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Rejection region (shaded) of the LRT</td>
<td>19</td>
</tr>
<tr>
<td>2.3</td>
<td>The LRT rejection regions for testing (a) $H_{0I}$ vs. $H_{1I}$ and (b) $H_{0D}$ vs. $H_{1D}$</td>
<td>21</td>
</tr>
<tr>
<td>2.4</td>
<td>The Null (unshaded) and Alternative (shaded) Parameter Spaces for the Case where the Alternative Parameter Space is an Acute Cone</td>
<td>24</td>
</tr>
<tr>
<td>2.5</td>
<td>The Null (unshaded) and Alternative (shaded) Parameter Spaces for the Case where the Alternative Parameter Space is a Right Angle Cone</td>
<td>24</td>
</tr>
<tr>
<td>2.6</td>
<td>The Null (unshaded) and Alternative (shaded) Parameter Spaces for the Case where the Alternative Parameter Space is an Obtuse Cone</td>
<td>25</td>
</tr>
<tr>
<td>2.7</td>
<td>The Null (unshaded) and Alternative (shaded) Parameter Spaces for the Case where the Alternative Parameter Space is a Union of Two Acute Cones</td>
<td>26</td>
</tr>
<tr>
<td>2.8</td>
<td>The Rejection Region (shaded) of Berger’s (1989) Test for the Alternative Parameter Space being a Union of Two Acute Cones</td>
<td>28</td>
</tr>
<tr>
<td>2.9</td>
<td>(a) A projection of estimated means for genes A through F, along with rejection regions of the overall F-test, the IUT, and LRT; (b) Boxplots of normalized expression distributions by gene and genotype.</td>
<td>29</td>
</tr>
<tr>
<td>2.10</td>
<td>A schematic plot of a rejection region using Zelterman’s method</td>
<td>39</td>
</tr>
<tr>
<td>3.1</td>
<td>Gene Sets Identified by All the Old Tests But None of the New Tests</td>
<td>60</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.2</td>
<td>Gene Sets Identified by All the Old Tests But None of the New Tests</td>
<td>61</td>
</tr>
<tr>
<td>3.3</td>
<td>Examples for the Gene Sets Identified by All the Proposed Tests (Genes with permutation p-values smaller than 0.05 in LRPT single gene data analysis are shown in red color.)</td>
<td>62</td>
</tr>
<tr>
<td>3.4</td>
<td>Examples for the Gene Sets Identified by All the Proposed Tests (Genes with permutation p-values smaller than 0.05 in IEPT single gene data analysis are shown in red color.)</td>
<td>63</td>
</tr>
<tr>
<td>3.5</td>
<td>Examples for the Gene Sets Identified by All the Proposed Tests (Genes with permutation p-values smaller than 0.05 in ILRPT single gene data analysis are shown in red color.)</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>Real Change Point and the Estimated Change Point</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>log ratio between the copy numbers of Mo17 and B73 on chromosome 1</td>
<td>83</td>
</tr>
<tr>
<td>4.3</td>
<td>log ratio between the copy numbers of Mo17 and B73 on chromosome 3</td>
<td>83</td>
</tr>
<tr>
<td>4.4</td>
<td>log ratio between the copy numbers of Mo17 and B73 on chromosome 6</td>
<td>84</td>
</tr>
<tr>
<td>4.5</td>
<td>log ratio between the copy numbers of Mo17 and B73 on chromosome 10</td>
<td>84</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Dan Nettleton for his guidance, support, encouragement and tremendous effort to my research projects during my Ph.D studies. His knowledge and intuition on research topics and his patient on revising and reading manuscripts of my thesis is of great value in helping me finish this thesis. I am also grateful to my committee members: Dr. Song Xi Chen, Dr. Ken Koehler, Dr. Peng Liu and Dr. Huaiqing Wu, for their efforts and advices to my work. I also thank Kai Ying for providing me the next generation sequencing data set and introducing the biological background. I appreciate my family for their great support and encouragement. Finally, I would like to thank my friends, classmates for various help during my PhD studies.
ABSTRACT

Statistics is a powerful tool in different scientific fields by providing statistical supports in experimental designs, data processing and statistical inference. In this thesis, we conduct theoretical and methodological statistical research with applications in biological and genomic areas.

In Chapter 2, we study the statistical testing problems with order-restricted null hypothesis, where the null parameter space is a union of two disjoint convex cones. We derive the likelihood ratio test and the intersection-union test, and show that the likelihood ratio test is uniformly more powerful than the intersection-union test. We also discuss the possibility of developing a test uniformly more powerful than the likelihood ratio test.

In Chapter 3, we propose four testing procedures for detecting the monotonic changes in multivariate gene expression distributions. We consider cases in which the treatment factor is ordinal and can be naturally ordered. The proposed procedures focus the detection powers to genes with monotonic departures from mean equality. Also, the proposed methods are able to deal with small sample sizes and high-dimensional distributions.

In Chapter 4, we propose a new methodology, based on a hidden Markov model with emission probabilities that are governed by mixture distributions, to detect copy number variations between different genomics using next generation sequencing read counts. This method demonstrates an improvement comparing to existing methods. We use this method to identify copy number variations between two maize genotypes, and the result is concordant to previous genomic studies using microarray data.

This thesis concludes in Chapter 5, which provides a discussion of future research directions.
CHAPTER 1. General Introduction

1.1 Introduction

Statistics is a powerful tool in various scientific areas, providing theoretical and methodological support in experimental designs, data processing and statistical inference. Conversely, questions raised in scientific studies motivate new research topics in statistical theories and methodologies. In this thesis, we focus on applications of statistical methodology in biological research and genomic studies. In this chapter, we briefly introduce several basic concepts and ideas that are covered in this thesis.

1.2 Order Restricted Statistical Inferences

In many scientific experiments, only parts of the parameter spaces are possible or conform to scientific limitations. For example, some treatment factors are quantitative or ordinal that can be naturally ordered. As the treatment level increased, it is reasonable to assume that the response values change monotonically if it is associated with the treatment factor. In these cases, it is of greater practical interest to study the statistical inference with parameters constrained to a sub-space with order restrictions than over the whole space.

For a more specific example, as the dose of a drug increased within some range the plasma concentrate of the drug is increased, and the drug effect may be strengthened. To search for the molecular genetic mechanisms that govern this phenomenon, gene expression values under different experiment levels are measured. We may wish to find genes with monotonic changes in mean expressions. In this example, suppose $\mu_i$, $i = 1, 2, 3$ is a mean gene expression value for the $i$th treatment level. The null hypothesis associated with homogeneity is

$$\mu_1 = \mu_2 = \mu_3,$$
and the alternative hypothesis associated with the monotonic alternative is

\[ \mu_1 \leq \mu_2 \leq \mu_3 \quad \text{or} \quad \mu_1 \geq \mu_2 \geq \mu_3 \]

with at least one inequality hold. Figure 1.1 shows the null and the alternative parameter spaces projected onto a two-dimensional sub-space in this example. The point in the center denotes the null space, and the shaded area denotes the alternative space.

Another example is the heterosis phenomenon in hybridization. One kind of heterosis is also called hybrid vigor. It occurs when the mean trait value of the offspring is more extreme than both of its parents. The maize F1 hybrid offspring from a cross of the parental inbred lines B73 and Mo17 has a series of improved trait characteristics, such as larger crop size, lower maturation time, and higher grain yield (Hallauer and Miranda, 1981). Suppose \( \theta_i, i = 1, 2, 3 \) is the mean measurement of a trait of interest, with \( i = 1 \) for one parent, \( i = 3 \) for another parent, and \( i = 2 \) for the offspring. The null hypothesis of no heterosis can be expressed as:

\[ \theta_1 \leq \theta_2 \leq \theta_3 \quad \text{or} \quad \theta_1 \geq \theta_2 \geq \theta_3, \]
The null and alternative spaces in the heterosis example

and the alternative hypothesis of heterosis is:

$$\theta_2 < \min\{\theta_1, \theta_3\} \quad \text{or} \quad \theta_2 > \max\{\theta_1, \theta_3\}.$$  

Figure 1.2 shows the null and the alternative parameter spaces of the heterosis example projected onto a two-dimensional sub-space. The unshaded area is the null space, and the shaded area is the alternative space.

There is a long history of research on univariate testing problems involving order restricted alternatives. Some of the earliest work was conducted by Bartholomew (1959a, 1959b, 1961a, 1961b). A large amount of research work in this area is summarized in Robertson, Wright and Dykstra (1988) and reviews for recent work can be found in Silvapulle and Sen (2005). However, most of the existent work is related to the null or alternative space restricted to a single cone. There is little work with parameter constrained to a union of cones.

In Chapter 2, we consider the heterosis example discussed above, where the null hypothesis space is a union of two cones. With the assumption that the random variables

Figure 1.2: The null and alternative spaces in the heterosis example
independently follow normal distributions, we derive the likelihood ratio test, and discuss the situations in which the uniformly more powerful tests can be constructed and their practical applicabilities.

In Chapter 3, we consider the example of monotonic gene expressions under ordinal treatment levels, with the condition that genes are grouped together into gene sets so that we have large random variable dimensions but small sample sizes. We propose several nonparametric procedures that have good performance on detecting multivariate gene sets with monotonic changes in gene expression distributions.

1.3 the Intersection-Union Test

In the heterosis example mentioned above, the null hypothesis is a union of two cones: \( \Theta_I = \{ \theta : \theta_1 \leq \theta_2 \leq \theta_3 \} \), \( \Theta_{II} = \{ \theta : \theta_1 \geq \theta_2 \geq \theta_3 \} \). As a result, it is natural to consider an intersection-union test (IUT). The intersection-union test was first described by Lehmann (1952) with the name coined by Gleser (1973). In this heterosis example, the null hypothesis \( H_0 \) can be expressed as \( \theta \in \Theta_0 = \Theta_I \cup \Theta_{II} \), and the alternative \( H_1 \) can be expressed as \( \theta \in \Theta_1 = \Theta_I^c \cap \Theta_{II}^c \). Let \( R_I \) be the rejection region for a test of

\[
H_{0I} : \theta \in \Theta_I \quad \text{vs.} \quad H_{1I} : \mu \in \Theta_I^c,
\]

and \( R_{II} \) be the rejection region for a test of

\[
H_{0II} : \theta \in \Theta_{II} \quad \text{vs.} \quad H_{1II} : \mu \in \Theta_{II}^c,
\]

then the intersection-union test (IUT) based on \( R_I \) and \( R_{II} \) rejects \( H_0 \) in favor of \( H_1 \) if and only if the test for \( H_{0I} \) vs. \( H_{1I} \) and the test for \( H_{0II} \) vs. \( H_{1II} \) both reject the null hypotheses. Berger (1982) showed that if each of the tests corresponding to \( R_I \) and \( R_{II} \) is of level-\( \alpha \), then the IUT corresponding to \( R = R_I \cap R_{II} \) is also of level no larger than \( \alpha \). By this means, the size of the IUT is no larger than \( \alpha \), but it might be very conservative.

1.4 Gene Set Testing

In gene expression studies, we have to deal with thousands of response variables corresponding to gene expressions for thousands of genes. However it is difficult to extract meaningful biological interpretations from single genes. To improve the interpretability,
genes are grouped into gene sets, based on the prior knowledge that some genes are associated with each other. The Gene Ontology (GO) is such a database that provides biological annotations for genes, by using the knowledge of some genes are associated with each other in aspects of cellular components, molecular function, and biological process, which provides a basis to group genes into gene sets.

There has been a large body of work in the area of detecting gene sets with joint expression distributions changing across treatment levels. Among many well-known gene set testing methods and softwares, some use gene specific statistics. These include the GSEA by Subramanian et al. (2005), the SAFE by Barry et al. (2005), and the GSA by Efron and Tibshirani (2007), among others. Some other methods use global and multivariate gene set testing, which include Goeman’s Global Test (Goeman et al, 2004), the ANCOVA global test by Mansmann and Meister (2005), the Multiresponse Permutation Procedure (MRPP) described by Mielke and Berry (2001) and used by Nettleton et al. (2008) in gene set testing, and the HMM aproach using the directed acyclic graph (DAG) by Liang and Nettleton (2010), among others.

1.5 Copy Number Variation Detection Using Next Generation Sequencing Data

A copy number variation (CNV) is a variation between different genomics in the number of copies in a genomic region for at least 1000 DNA bases or larger (Banerjee et al., 2011). Maize exhibits extensive genotypic and phenotypic variations. Since the maize genotype was fully sequenced using whole-genome sequencing technologies in 2009 (Schnable et al. 2009), there is a large body of research work on CNVs detections among various maize genotypes using array comparative genomic hybridization (aCGH) (Swanson-Wagner et al., 2009; Springer et al., 2009; Belo et al., 2010). Although aCGH has served as a robust and effective approach for CNV identification since it was proposed in 1997 (Solinas-Toldo et al, 1997), it is expensive and has limited resolution and accuracy. Recently developed next generation sequencing (NGS) technologies provide a sensitive and accurate approach for genomic variation detections.

However, the use of NGS for CNV detection has been limited by a lack of available and effective statistical approaches. Most of the existing methods for CNV identification
using NGS data can be classified into two categories. One category contains the sliding window methods. The common idea of the sliding window methods is to divide the whole genome into small windows, then to carry out tests of significance for copy number variance between any adjacent windows. The category of sliding window methods includes the Segseq by Chiang et al. (2009), Event-wise testing by Yoon et al. (2009), the rSW-seq by Kim et al. (2010) and the JointSLM by Magi et al. (2011), among others. The sliding window methods face the problem of a huge number of simultaneous tests they heavily rely on the determination of the critical values. The other category includes methods using Hidden Markov Models. Ivakhno et al. (2010) proposed the CNAseg HMM based method, using multiple sequencing samples for each of the genotypes in the CNV detection.

Chapter 4 proposes a new CNV detection methodology for NGS data originating from two different genomes, using a Hidden Markov Model approach. We use the Expectation-Maximization (EM) algorithm to obtain estimates for the parameters in the model. An introduction to some basic ideas of HMMs and the EM algorithm is provided in Sub-sections 1.5.1 and 1.5.2.

1.5.1 Hidden Markov Models

A Hidden Markov Model (HMM) is a probabilistic model based on a Markov process. The theory of the HMM was published in a series of paper by Baum et al. (1966, 1967, 1970, 1972). A HMM is constructed from a bivariate random process with discrete time, \( \{S_t, X_t\}, \ t = 1, 2, \ldots \). Here \( S_t, \ t = 1, 2, \ldots, T \) is an invisible Markov chain with finite states, where the next state and past states are independent given the current state \( S_t \), i.e.,

\[
P(S_{t+1}|S_1, \ldots, S_t) = P(S_{t+1}|S_t).
\]

\( X_t \) is a visible observation generated by the hidden state \( S_t, \ t = 1, 2, \ldots, T \). Depending on the current state \( S_t \), the observation \( X_t \) is independent of all other states and observations. Each state has a probability distribution in generating the observations. The information about the HMM is provided by the observations \( X_t, \ t = 1, 2, \ldots, T \).

A HMM is defined by three probabilities. First is the transition probability \( P(S_{t+1}|S_t) \), which is the conditional probability of the next state given the current state. The transition probabilities determine the associations among the hidden states. Another is the conditional probability of having an observation \( X_t \) given the current state \( S_t \), which is called the
emission probability. The emission probabilities determine the intrinsic relation between the hidden states and the visible observations. The third is the initial probability of the hidden state at time one, $P(S_1)$. The transition probability, the emission probability and the initial probability are the three essential probabilities through which we can reveal a HMM.

1.5.2 The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is an iterative procedure searching for the parameter values that maximize the likelihood function, when part of the data set is missing. With this feature, the EM algorithm is also an effective method in estimating the parameters and predicting the hidden state in HMMs, by assuming $X = (X_1, X_2, \ldots, X_T)'$ as the incomplete data and the hidden states $S = (S_1, S_2, \ldots, S_T)'$ as the missing data. The complete data is $(X, S)$. The target likelihood to be maximized is represented as $\mathcal{L}(\theta|X)$. Since $X$ is the incomplete data, $\mathcal{L}(\theta|X)$ is also referred to as the incomplete data likelihood. If the value of $S$ were known, then the complete data likelihood would be $\mathcal{L}(\theta|S, X)$. With some initial estimation of the parameter value $\theta^{(0)}$ using the simplest assumptions, the first step of the EM algorithm is to evaluate the expectation of the complete data log-likelihood, with respect to the conditional distribution of the missing data $S$, given the observed data $X$ and the current parameter value $\theta^{(0)}$: $E_{S|X, \theta^{(0)}}(\ell(\theta|X, S))$. This step is called the E-step of the EM algorithm, and $E_{S|X, \theta^{(0)}}(\ell(\theta|X, S))$ is a function of $\theta$. The second step of the EM algorithm is to maximize $E_{S|X, \theta^{(0)}}(\ell(\theta|X, S))$ with respect to $\theta$, which gives an updated parameter value $\theta^{(1)}$. This step is called the M-step of the EM algorithm. By iteratively applying the E-step and the M-step, a sequence of parameter values $\{\theta^{(m)}\}_{m \geq 0}$ is obtained, where each $\theta^{(m)}$ increases the value of $\mathcal{L}(\theta|X)$ with respect to the previous step $\theta^{(m-1)}$. The iteration stops when the amount of enlargement of the likelihood is sufficiently small.

1.6 Organization

The main idea throughout this thesis is statistical theory and methodology research with motivations from biological and genomic areas. Chapter 2 studies statistical inference with an order-restricted null hypothesis, where the null parameter space is a union of two cones. This test can be applied to detect heterosis in trait values between the parents in-
bred lines and their hybrid offsprings. Chapter 3 studies monotonic alternative hypothesis testing with high-dimensional random variables and small sample sizes, which can be applied to multivariate gene set detections using microarray gene expression data. Chapter 4 proposes a new methodology based on a Hidden Markov Model with mixture emission probabilities, which can be employed in copy number variation detections using next generation sequencing data.

Chapter 2 derives the likelihood ratio test (LRT) and the intersection-union test (IUT) when the null parameter space is a union of two cones and the alternative space is the complementary set of the null space. We prove that the likelihood ratio test is uniformly more powerful than the intersection-union test. We also demonstrate the existence of the tests that are uniformly more powerful than the likelihood ratio test in some related problems, and discuss the applicability of these uniformly more powerful tests in real data analyses. An analysis of maize gene expression data of two inbred lines and their hybrid F1 offspring is used to illustrate the application of union-of-cone null hypothesis testing in heterosis identifications.

Chapter 3 considers the case when the null hypothesis of homogeneity of distributions and the alternative hypothesis of monotonic order-restricted of means, with random variables with relatively large dimensions and small sample sizes. The hypotheses are motivated by multivariate gene set testing with expression levels that change monotonically with an ordinal treatment factor. We propose four nonparametric tests based on permutation procedures. A simulation study is carried out to compare the proposed tests with tests not focusing on monotonic alternatives, and it is shown that the proposed tests provide better detection power for gene sets with monotonic changing trends.

Chapter 4 explores a new statistical methodology, based on a Hidden Markov Model with mixture emission distributions (m-HMM), to detect copy number variation change points in comparing next generation DNA sequences between a target genome and a reference genome. The Expectation-Maximization algorithm is used to estimate the parameters and predict the hidden states in the m-HMM.

Conclusions and additional discussion are provided in Chapter 5, with an overview of future work.
References


Bartholomew, D. J. (1959b). A Test of Homogeneity for Ordered Alternatives. II. *Biometrika*, 46, 328-35.


CHAPTER 2. Testing a Union-of-Cones Null Hypothesis for the Identification of Heterosis

Abstract

High-parent or low-parent heterosis is a genetic phenomenon that occurs when the mean trait value of offspring is more extreme than that of either parent. We discuss statistical tests that can be used to detect heterosis of this type. The null hypothesis of no high-parent or low-parent heterosis constrains a vector of mean parameters to a union of two closed convex cones. This unusual null parameter space leads to a non-standard testing problem for which we derive both the likelihood ratio test and the intersection-union test. Although these tests are often equivalent, we show that the likelihood ratio test is preferred over the intersection-union test in this case. Moreover, we also discuss the possibility of developing a test uniformly more powerful than the likelihood ratio test. Although such tests are of theoretical interest, we ultimately recommend the likelihood ratio test for use in practical applications.

KEY WORDS: Cone; Intersection-Union Test; Likelihood Ratio Test; Order-restricted Statistical Inference; Uniformly More Powerful Tests

2.1 Introduction

One form of heterosis, also known as hybrid vigor, occurs when the mean trait value of offspring is more extreme than that of either parent. For example, the maize F1 hybrid offspring produced from a cross of parental inbred lines B73 and Mo17 is larger in size, has lower maturation time, and has higher grain yield than both of its parents (Hallauer and Miranda, 1981). The hybrid vigor phenomenon was first documented in the late 1800s and is the basis of the multibillion dollar seed industry today.

When two inbred parental lines are crossed to produce hybrid offspring, it is natural to
ask whether there is evidence of heterosis for a trait of interest. The purpose of this paper is to develop a method that can be used to test for heterosis using measures of a trait of interest from multiple parent and offspring replications. We illustrate our proposed method in Section 2.6 using measures of gene expression levels in B73, Mo17, and F1 hybrid maize lines originally studied by Swanson-Wagner et al. (2006).

Formally, the testing problem that we consider can be described as follows. Suppose for \( i = 1, 2, 3 \) and \( j = 1, \ldots, n \),

\[
y_{ij} = \mu_i + \epsilon_{ij},
\]

where \( \mu_1, \mu_2 \) and \( \mu_3 \) are unknown parameters in \( \mathbb{R} \) and \( \epsilon_{11}, \ldots, \epsilon_{3n} \overset{iid}{\sim} N(0, \sigma^2) \). We wish to test

\[
H_0 : \quad \mu_1 \leq \mu_2 \leq \mu_3 \quad \text{or} \quad \mu_1 \geq \mu_2 \geq \mu_3 \\
\text{vs.} \quad H_1 : \quad \mu_2 < \min\{\mu_1, \mu_3\} \quad \text{or} \quad \mu_2 > \max\{\mu_1, \mu_3\}.
\]

(2.1)

Here, \( i \) indexes genotypes with \( i = 1 \) for one parent, \( i = 2 \) for the hybrid, and \( i = 3 \) for the other parent. The response \( y_{ij} \) is the measure of a trait of interest for the \( j \)th replication of genotype \( i \). The parameter vector \( \mu = (\mu_1, \mu_2, \mu_3)' \) contains the genotype means.

The alternative hypothesis in (2.1) implies the existence of heterosis for the trait of interest. More precisely, the alternative hypothesis indicates either low-parent heterosis (hybrid mean lower than the lowest of the two parental means) or high-parent heterosis (hybrid mean higher than the highest of the two parental means). Low-parent and high-parent heterosis can be considered special cases of a third type of heterosis known as mid-parent heterosis where the mean of the hybrid differs from the average of the two parental means. Because mid-parent heterosis in general is not as interesting as the special cases of low-parent and high-parent heterosis from a scientific, economic, and statistical standpoint, we focus in this paper on the problem of detecting either low-parent or high-parent heterosis and for simplicity use the term heterosis to mean low-parent or high-parent heterosis. With this terminology, the null hypothesis in (2.1) is the “no heterosis” null hypothesis.
The hypotheses in (2.1) may be written as

\[ H_0 : b_1' \mu \geq 0, b_2' \mu \geq 0 \text{ or } b_1' \mu \leq 0, b_2' \mu \leq 0 \]

vs.

\[ H_1 : b_1' \mu > 0, b_2' \mu < 0 \text{ or } b_1' \mu < 0, b_2' \mu > 0, \]

where \( b_1 = (-1, 1, 0)' \) and \( b_2 = (0, -1, 1)' \). The null hypothesis in (2.1) constrains \( \mu \) to a union of two closed, convex polyhedral cones in \( \mathbb{R}^3 \). The boundary of this region is defined by the two intersecting planes \( \{ \mu \in \mathbb{R}^3 : \mu_1 = \mu_2 \} = \{ \mu \in \mathbb{R}^3 : b_1' \mu = 0 \} \) and \( \{ \mu \in \mathbb{R}^3 : \mu_2 = \mu_3 \} = \{ \mu \in \mathbb{R}^3 : b_2' \mu = 0 \} \). Panel (a) in Figure 2.1 shows a two-dimensional projection of the null (unshaded) and the alternative (shaded) parameter spaces. The point in the center of the figure corresponds to the line \( \{ \mu \in \mathbb{R}^3 : \mu_1 = \mu_2 = \mu_3 \} \), the intersection of the two region-defining planes.

Robertson, Wright and Dykstra (1988) and Silvapulle and Sen (2005) describe a long history of research on problems where either the null or the alternative hypothesis constrains a parameter vector to a single cone. There is considerably less work on problems that involve union-of-cones constraints. One notable exception is the problem of testing for qualitative or cross-over interactions, which are of great importance in clinical trials. Gail and Simon (1985) derived the asymptotic distribution of the likelihood ratio test (LRT) for qualitative interactions, and Silvapulle (2001) derived the exact distribution for the LRT statistic for the finite sample case. In order to obtain a test more powerful than the likelihood ratio test, Zelterman (1990) showed how to expand the LRT rejection region without changing the size of the test. Zelterman’s (1990) more powerful test is calculated using the constraint that, for any parameters on the boundary of the null and alternative parameter spaces, the uniformly more powerful test rejects the null hypothesis with probability equal to the size \( \alpha \). Berger (1989) constructed a size-\( \alpha \) uniformly more powerful test by adding a series of cubic areas to the LRT rejection region for problems where the alternative parameter space is a single acute-angled cone or a union of two acute-angled cones. For the case where the alternative space is an obtuse-angled cone or a union of two obtuse-angled cones, Liu and Berger (1995) showed how to expand the LRT rejection area by adding a spindle shaped region without increasing the size of the test. Despite the theoretical advantages of these uniformly more powerful tests, Perlman and Wu (1999) have criticized their development.
and questioned their value in practical data analysis.

In this paper, we develop tests of the null and alternative hypotheses defined in (2.1) and (2.2). We begin by introducing a transformation that reduces our problem involving three mean parameters to a two-dimensional problem. The details of this transformation are provided in Section 2.2. We provide the LRTs for the cases of known and unknown variance $\sigma^2$ in Section 2.3. In Section 2.4, we present the Intersection-Union Test (IUT) and show that it is uniformly less powerful than the LRT for testing the hypotheses in (2.1). In Section 2.5, we discuss the possibility of constructing a test uniformly more powerful than the LRT. We illustrate the application of our proposed LRT by testing for gene expression heterosis in Section 2.6. Summary and conclusions are provided in Section 2.7.

2.2 The Transformation

The hypotheses in (2.1) involve the three mean parameters $\mu_1$, $\mu_2$, and $\mu_3$. However, based on the expression of the constraints in (2.2), only two dimensions are relevant to the problem. Thus, we can develop a transformation that projects the testing problem onto a two-dimensional space.

For the case of known $\sigma^2$, consider the transformation $X^0 = (X_1, X_2, X_3)' = TY$ and $\theta^0 = (\theta_1, \theta_2, \theta_3)' = T\mu$, where $Y = \left( \frac{\sum Y_{ij}}{n}, \frac{\sum Y_{2i}}{n}, \frac{\sum Y_{3i}}{n} \right)' = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)'$, and

$$T = \frac{\sqrt{n}}{\sigma} \begin{pmatrix} -\frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{3} \end{pmatrix}. \quad (2.3)$$

Then $X^0$ follows a three-dimensional normal distribution with mean $\theta^0 = (\theta_1, \theta_2, \theta_3)'$ and variance matrix $I_{3 \times 3}$.

In the case that $\sigma^2$ is unknown, the transformation becomes $X^0 = T(\hat{\sigma})Y$, with $\sigma$ in (2.3) being replaced by the square root of the estimate $\hat{\sigma}^2 = \text{MSE} = \frac{\sum_{i=1}^{3} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2}{3n-3}$. In this case, $X^0$ follows Kshirsagar’s non-central multivariate $t$ distribution with degrees of freedom $3n - 3$ and parameter $\theta^0$ (Kshirsagar, 1961) with joint pdf

$$f_{ncmt}(x^0) = \exp \left( -\frac{1}{2} \theta^0' \theta^0 \right) \frac{\Gamma\left(\frac{3n}{2}\right)}{\{(3n - 3)\pi\}^{3n/4} \Gamma\left(\frac{3n-3}{2}\right)} \cdot \left(1 + \frac{x^0'x^0}{3n-3}\right)^{-\frac{3n}{2}} \cdot \sum_{r=0}^{\infty} \frac{\Gamma\left(\frac{3n+r}{2}\right)}{r! \Gamma\left(\frac{3n}{2}\right)} \left(\frac{\sqrt{2}x^0'\theta^0}{\sqrt{3n - 3 + x^0'x^0}}\right)^r.$$
For each $i \in \{1, 2, 3\}$, $X_i$ has a non-central $t$ distribution with $3n - 3$ degrees of freedom and non-centrality parameter $\theta_i$. Note that $X_1$, $X_2$ and $X_3$ are not independent (Siotani, 1976; Kotz and Nadarajah, 2004). Define $X$ to be $(X_1, X_2)'$ and $\theta$ to be $(\theta_1, \theta_2)'$.

The hypothesis in (2.2) can be stated in terms of the two parameters $\theta_1$ and $\theta_2$ as follows:

$$H_0 : \quad h_k' \theta \leq 0 \quad k = 1, 2$$

or

$$H_0 : \quad h_k' \theta \geq 0 \quad k = 1, 2$$

vs.

$$H_1 : \quad h_1' \theta > 0 \text{ and } h_2' \theta < 0$$

or

$$H_1 : \quad h_1' \theta < 0 \text{ and } h_2' \theta > 0,$$

where $h_1 = (\frac{\sqrt{3}}{2}, \frac{1}{2})'$ and $h_2 = (-\frac{\sqrt{3}}{2}, \frac{1}{2})'$. Figure 2.1 shows a two-dimensional projection of the null and alternative parameter spaces in terms of $\mu$ (a) and the parameter spaces in terms of $\theta$ (b). In both panels, the unshaded area represents the null space, and the shaded area is the alternative space. Both the null and alternative spaces are unions of two vertical cones.

Figure 2.1: The null (unshaded) and alternative (shaded) parameter spaces in terms of $\mu$ and $\theta$
2.3 The Likelihood Ratio Test

In this section, we will present the LRT for testing the hypotheses in (2.1) or, equivalently, (2.4). We will separately discuss the cases when $\sigma^2$ is known and unknown. A detailed derivation of the LRT is provided in Section 2.8.1 of the appendix.

For the case when $\sigma^2$ is known, the LRT rejects the null hypothesis if $\chi^2 = \frac{n \sum_{i=1}^{3}(\bar{Y}_i - \bar{\mu})^2}{\sigma^2}$ is large, where $\bar{\mu} = (\bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3)'$ is the MLE of the parameters under the restriction that $\mu_1$, $\mu_2$ and $\mu_3$ are either increasingly or decreasingly ordered (i.e., $\mu_1 \leq \mu_2 \leq \mu_3$ or $\mu_1 \geq \mu_2 \geq \mu_3$). If $\bar{Y}_1 \leq \bar{Y}_2 \leq \bar{Y}_3$ or $\bar{Y}_1 \geq \bar{Y}_2 \geq \bar{Y}_3$, then $\bar{\mu}_i = \bar{Y}_i$ and $\chi^2 = 0$. If $\bar{Y}_2$ is the smallest group mean, then $\chi^2 = n(\bar{Y}_2 - \max(\bar{Y}_1, \bar{Y}_3))^2/2\sigma^2$. If $\bar{Y}_2$ is the largest group mean, then $\chi^2 = n(\bar{Y}_2 - \min(\bar{Y}_1, \bar{Y}_3))^2/2\sigma^2$.

The size-$\alpha$ LRT rejects $H_0$ if and only if $\chi^2 \geq \chi^2_{1,2\alpha}$, where $\chi^2_{1,2\alpha}$ is the upper $2\alpha$ quantile of a chi-square distribution with 1 degree of freedom ($\chi^2_1$).

Equivalently, we can express the rejection region in terms of $X$, as

$$R_L = \left\{ x : -\sqrt{3}x_1 + 2z_{\alpha} \leq x_2 \leq \sqrt{3}x_1 - 2z_{\alpha} \right\}$$

$$\cup \left\{ x : \sqrt{3}x_1 + 2z_{\alpha} \leq x_2 \leq -\sqrt{3}x_1 - 2z_{\alpha} \right\}, \quad (2.5)$$

where $z_{\alpha}$ is the upper $\alpha$ quantile of a standard normal distribution. The shaded area in Figure 2.2 shows the LRT rejection region $R_L$ in the $X$ coordinate system. $R_L$ is a union of two cones. The faces of the two cones are parallel to $h'_k X = 0$, $k = 1, 2$. The distance between the two vertices is $\frac{4\sqrt{3}}{3}z_{\alpha}$.

For the case when $\sigma^2$ is unknown, the LRT rejects the null if

$$\tilde{F} = \frac{n \sum_{i=1}^{3}(\bar{Y}_i - \bar{\mu})^2}{\text{MSE}} > F_{1,3n-3,2\alpha},$$

where $F_{1,3n-3,2\alpha}$ is the upper $2\alpha$ quantile of $F$ distribution with degree of freedom 1 and $3n - 3$. The rejection region expressed in terms of $X$ is,

$$R_L(\tilde{\sigma}) = \left\{ x : -\sqrt{3}x_1 + 2t_{3n-3,\alpha} \leq x_2 \leq \sqrt{3}x_1 - 2t_{3n-3,\alpha} \right\}$$

$$\cup \left\{ x : \sqrt{3}x_1 + 2t_{3n-3,\alpha} \leq x_2 \leq -\sqrt{3}x_1 - 2t_{3n-3,\alpha} \right\}, \quad (2.6)$$

where $t_{3n-3,\alpha}$ is the upper $\alpha$ quantile of a central $t$ distribution with $3n - 3$ degrees of freedom.
2.4 The Intersection-Union Test

Because the parameter space associated with the null hypothesis in (2.1) is a union of the two cones $\Theta_I = \{\mu : \mu_1 \leq \mu_2 \leq \mu_3\}$ and $\Theta_D = \{\mu : \mu_1 \geq \mu_2 \geq \mu_3\}$, it is natural to consider an intersection-union test (IUT). The IUT was first described by Lehmann (1952), and the name was coined by Gleser (1973). We will discuss the IUT in this section and show that the LRT is more powerful than the IUT.

In our special case, a level-$\alpha$ IUT based on individual LRTs will reject $H_0 : \mu \in \Theta_I \cup \Theta_D$ in favor of $H_1 : \mu \in \Theta_I^c \cap \Theta_D^c$ if and only if both the size-$\alpha$ LRT of $H_0 : \mu \in \Theta_I$ vs. $H_1 : \mu \in \Theta_I^c$ and the size-$\alpha$ LRT of $H_0 : \mu \in \Theta_D$ vs. $H_1 : \mu \in \Theta_D^c$ reject their null hypotheses.

When $\sigma^2$ is unknown, the LRT statistic for testing $H_{0I} : \mu \in \Theta_I$ vs. $H_{1I} : \mu \in \Theta_I^c$ is
\( F^I = n \sum_{i=1}^{3} (\tilde{y}_i - \tilde{\mu}_I^2)^2 / \text{MSE} \), where \( \tilde{\mu}_I^2 = (\tilde{\mu}_1^2, \tilde{\mu}_2^2, \tilde{\mu}_3^2) \) is the MLE of \( \mu \) under the restriction \( \mu_1 \leq \mu_2 \leq \mu_3 \). The value of the statistic \( F^I \) is shown in the following equations:

\[
F^I = \begin{cases} 
0, & \tilde{Y}_1. \leq \tilde{Y}_2. \leq \tilde{Y}_3. \\
\frac{n(\tilde{Y}_1 - \tilde{Y}_2)^2}{2 \text{MSE}}, & \tilde{Y}_2. \leq \tilde{Y}_1. \leq \tilde{Y}_3. \text{ or } \\
\frac{n(\tilde{Y}_2 - \tilde{Y}_3)^2}{2 \text{MSE}}, & \tilde{Y}_1. \leq \tilde{Y}_3. \leq \tilde{Y}_2. \text{ or } \\
\frac{n \sum_{i=1}^{3} (\tilde{Y}_i - \tilde{Y}_1 + \tilde{Y}_2 + \tilde{Y}_3)^2}{\text{MSE}}, & \tilde{Y}_3. \leq \tilde{Y}_1. \leq \tilde{Y}_2. \text{ and } \tilde{Y}_1. \leq \tilde{Y}_2 + \tilde{Y}_3. \\
0, & X_2 \geq \sqrt{3}X_1 \cdot \text{sign}(X_1) \\
\left( \frac{\sqrt{3}}{2}X_1 + \frac{1}{2}X_2 \right)^2, & \frac{1}{\sqrt{3}}X_1 \leq X_2 < -\sqrt{3}X_1, X_1 < 0 \\
\left( \frac{\sqrt{3}}{2}X_1 - \frac{1}{2}X_2 \right)^2, & -\frac{1}{\sqrt{3}}X_1 \leq X_2 < \sqrt{3}X_1, X_1 > 0 \\
X_1^2 + X_2^2, & X_2 < -\frac{1}{\sqrt{3}}X_1 \cdot \text{sign}(X_1). 
\end{cases}
\]

For any \( c \geq 0 \),

\[
\sup_{\mu \in \Theta} \left\{ P(F^I \geq c | \mu_1, \mu_2, \mu_3) : \mu_1 \leq \mu_2 \leq \mu_3 \right\} 
\geq \frac{1}{3} \left( P(F_{1,3n-3} \geq \tilde{c}) + P(F_{2,3n-3} \geq \tilde{c}) \right) 
\geq \frac{1}{2} \left( P(F_{1,3n-3} \geq \tilde{c}) + \frac{1}{3} P(F_{2,3n-3} \geq \tilde{c}) \right).
\]

Thus, the size-\( \alpha \) LRT rejection region for testing \( H_{0I} \) vs. \( H_{1I} \) is \( R_I = \{ Y : F^I \geq c_{IU} \} \) with

\[
\frac{1}{2} P(F_{1,3n-3} \geq c_{IU}) + \frac{1}{3} P(F_{2,3n-3} \geq c_{IU}) = \alpha.
\]

Likewise, for \( H_{0D} \) vs. \( H_{1D} \), the LRT test statistic and corresponding size-\( \alpha \) rejection region are \( F^D = n \sum_{i=1}^{3} (\tilde{y}_i - \tilde{\mu}_I^D)^2 / \text{MSE} \) and \( R_D = \{ Y : F^D \geq c_{IU} \} \), where \( \tilde{\mu}_1^D, \tilde{\mu}_2^D, \tilde{\mu}_3^D \) are
the MLEs of $\mu_1, \mu_2, \mu_3$ subject to the restriction $\mu_1 \geq \mu_2 \geq \mu_3$. The IUT rejection region is $R_I \cap R_D$.

In Figure 2.3, the unshaded cone near the top of panel (a) is the null parameter space $\Theta_I$ projected onto the X coordinate system, and the shaded area is the projection of the LRT rejection region $R_I$ onto this same coordinate system; the unshaded cone near the bottom of panel (b) is the null parameter space $\Theta_D$ projected onto a the X coordinate system, and the shaded area is the corresponding projection of the LRT rejection region $R_D$. Then the IUT based on $R_I$ and $R_D$ rejects $H_0$ with the rejection region $R_{IU} = R_I \cap R_D$. By (2.8), it is clear that $c_{IU} > F_{1,3n-3,2\alpha}$. So the LRT is uniformly more powerful than the IUT.

In a more general case, if the null and alternative hypotheses can be expressed as $H_0 : \mu \in \Theta_0 = \bigcup_{m=1}^{M} \Theta_m$ vs. $H_1 : \mu \in \Theta_1 = \cap_{m=1}^{M} \Theta_m^c$, with $M \geq 2$. The level $\alpha$ IUT is rejected if and only if all the $M$ individual tests of $H_{0m} : \mu \in \Theta_m$ vs. $H_{1m} : \mu \in \Theta_m^c$ are rejected at level $\alpha$. Let $\Lambda(y)$ and $R = \{y : \Lambda(y) < c\}$ be the test statistic and the size-\(\alpha\) LRT rejection region for $H_0$ vs. $H_1$. Also, let $\Lambda_m(y)$ and $R_m = \{y : \Lambda_m(y) < c_{m\alpha}\}$ be the test statistic and the size-\(\alpha\) LRT rejection region for $H_{0m} : \mu \in \Theta_m$ vs. $H_{1m} : \mu \in \Theta_m^c$, $m = 1, \ldots, M$. Berger (1997) proved that $R = \cap_{m=1}^{M} R_m$ under the following conditions:

1. $c_{1\alpha} = \ldots = c_{M\alpha}$, and

2. for some $m_0 \in \{1, \ldots, M\}$, there exists a sequence of parameter points $\{\mu_Q, Q = 1, 2, \ldots\} \subset \Theta_{m_0}$
such that \( \lim_{Q \to \infty} P_{\mu_Q}(Y \in R_m) = \alpha \) and \( \lim_{Q \to \infty} P_{\mu_Q}(Y \in R_m) = 1 \) for all \( m \in \{1, \ldots, M\} \backslash m_0 \).

In the special case of testing for heterosis, the first condition of Berger’s theorem holds because \( c_{IU} \) is the common critical value for the size-\( \alpha \) LRTs of both \( H_{0I} \) vs. \( H_{1I} \) and \( H_{0D} \) vs. \( H_{1D} \). However, the second condition does not hold. For both \( H_{0I} \) vs. \( H_{1I} \) and \( H_{0D} \) vs. \( H_{1D} \), the least favorable parameter set within each null parameter space is the intersection of the two cones, i.e., \( \Theta_I \cap \Theta_D = \{\mu \in \mathbb{R}^3 : \mu_1 = \mu_2 = \mu_3\} \). Thus, for any \( \mu \in \Theta_I \setminus \{\mu \in \mathbb{R}^3 : \mu_1 = \mu_2 = \mu_3\} \), \( P(\hat{F}^I \geq c_{IU}|\mu) < \alpha \). For any \( \{\mu_Q, Q = 1, 2, \ldots\} \subset \Theta_I \) such that \( \lim_{Q \to \infty} P_{\mu_Q}(Y \in R_I) = \alpha \), the parameter \( \mu_Q \) must approach the set \( \{\mu \in \mathbb{R}^3 : \mu_1 = \mu_2 = \mu_3\} \) as \( Q \to \infty \). By continuity of \( P_{\mu_Q}(Y \in R_D) \) as a function of \( \mu_Q \), \( \lim_{Q \to \infty} P_{\mu_Q}(Y \in R_D) = \alpha < 1 \). Reversing the roles of \( I \) and \( D \) and applying the same basic argument shows that \( \lim_{Q \to \infty} P_{\mu_Q}(Y \in R_D) = \alpha \) implies \( \lim_{Q \to \infty} P_{\mu_Q}(Y \in R_I) = \alpha < 1 \). Therefore, Berger’s condition 2 does not hold for our test for heterosis.

### 2.5 More Powerful Tests

Berger (1989), Zelterman (1990) and Liu and Berger (1995) constructed tests that are uniformly more powerful than size-\( \alpha \) LRTs for testing problems that share some similarities with the problem we consider in this paper. The common feature of the testing problems that makes it possible to construct tests uniformly more powerful than size-\( \alpha \) likelihood ratio tests were discussed by Perlman and Wu (1999). In particular, for our problem and others considered previously, the least favorable null distribution is not attained at any point in the null parameter space, and the level \( \alpha \) is only approached asymptotically as the parameter moves on the boundary of the null parameter space away from the intersection of cones whose union is the null parameter space. Such a situation makes it possible to expand the rejection region of the LRT near the intersection of cones while maintaining size \( \alpha \). By such expansions, the rejection probabilities of Berger’s (1989), Zelterman’s (1990) and Liu and Berger’s (1995) tests are larger than the rejection probability of size-\( \alpha \) LRT for any points in the alternative space. Thus, these new tests are uniformly more powerful than the size-\( \alpha \) LRT. In this section, we will discuss the situation in which the uniformly more powerful tests are constructed, and discuss their applicability in real data analyses.
2.5.1 Uniformly More Powerful Tests when the Alternative Parameter Space is a Single Cone

Let $\mathbf{T} = (T_1, \ldots, T_p)$ be a $p$-dimensional ($p \geq 2$) random variable such that $\mathbf{T} \sim N_p(\boldsymbol{\xi}, \mathbf{I})$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)'$ is a $p$-dimensional unknown parameter. Berger (1989) constructed a test uniformly more powerful than the LRT for the test of

$$H_0(B) : \quad b_k^{(B)}' \boldsymbol{\xi} \leq 0 \text{ for some } k = 1, \ldots, K$$

vs.

$$H_1(B) : \quad b_k^{(B)}' \boldsymbol{\xi} > 0 \text{ for all } k = 1, \ldots, K,$$

for the case in which for each $k = 1, \ldots, K$ there is a $k_0 \in \{1, \ldots, K\}$ such that $b_k^{(B)}' b_{k_0}^{(B)} \leq 0$.

Berger’s uniformly more powerful test was constructed by adding a series of diamond shaped areas between the origin $\mathbf{T} = \mathbf{0}_p$ and the vertex of the cone of the size-$\alpha$ LRT rejection region. If $K = 2$, Berger’s (1989) hypotheses can be expressed as

$$H_0(B) : \quad b_1^{(B)}' \boldsymbol{\xi} \leq 0 \text{ or } b_2^{(B)}' \boldsymbol{\xi} \leq 0$$

vs.

$$H_1(B) : \quad b_1^{(B)}' \boldsymbol{\xi} > 0 \text{ and } b_2^{(B)}' \boldsymbol{\xi} > 0$$

with $b_1^{(B)}' b_2^{(B)} \leq 0$. In this case, the alternative parameter space is a cone with angle no larger than 90 degrees, and the null parameter space is a union of the cone $b_1^{(B)}' \boldsymbol{\xi} \leq 0$ and the cone $b_2^{(B)}' \boldsymbol{\xi} \leq 0$. Figure 2.4 shows the null and the alternative parameter spaces of Berger’s (1989) test with $K = 2$ and $b_1^{(B)}' b_2^{(B)} < 0$, where the alternative space is an acute cone. Figure 2.5 shows the special case that $b_1^{(B)}' b_2^{(B)} = 0$, in which case the alternative space is a right angle cone.

Liu and Berger (1995) pointed out that Berger’s (1989) test is not a size-$\alpha$ test if $b_1^{(B)}' b_2^{(B)} > 0$, in which case the alternative parameter space is an obtuse cone (Figure 2.6). Liu and Berger (1995) constructed a uniformly more powerful test by adding a spindle shaped rejection region to the size-$\alpha$ LRT rejection region. This test can be applied to cases with alternative parameter spaces with any angles smaller than 180 degrees.
Figure 2.4: The Null (unshaded) and Alternative (shaded) Parameter Spaces for the Case where the Alternative Parameter Space is an Acute Cone

Figure 2.5: The Null (unshaded) and Alternative (shaded) Parameter Spaces for the Case where the Alternative Parameter Space is a Right Angle Cone
2.5.2 Uniformly More Powerful Tests when the Alternative Parameter Space is a Union of Two Cones

Berger (1989) also constructed a test that can be applied when the alternative parameter space is a union of two vertical cones with angles no larger than 90 degrees. The null and alternative hypotheses can be expressed as follows:

\[ H_0(B) : \quad b_k^{(B)'} \xi \leq 0 \text{ for some } k = 1, \ldots, K \]

\[ \text{and } b_k^{(B)'} \xi \geq 0 \text{ for some } k = 1, \ldots, K \]

vs.

\[ H_1(B) : \quad b_k^{(B)'} \xi > 0 \text{ for all } k = 1, \ldots, K, \]

\[ \text{or } b_k^{(B)'} \xi < 0 \text{ for all } k = 1, \ldots, K. \]

with the condition that for each \( k = 1, \ldots, K \) there is a \( k_0 \in \{1, \ldots, K\} \) such that \( b_k^{(B)'} b_{k_0}^{(B)} \leq 0 \). Figure 2.7 shows the null and the alternative parameter spaces for one case where \( K = 2 \).
Zelterman (1990) constructed an “approximate” size-\(\alpha\) test for

\[ H_0(Z) : \quad \xi \geq 0 \text{ for all } k = 1, \ldots, K \]

and \( \xi \leq 0 \text{ for all } k = 1, \ldots, K \)

vs.

\[ H_1(Z) : \quad \xi > 0 \text{ for some } k = 1, \ldots, K, \]

and \( \xi < 0 \text{ for some } k = 1, \ldots, K \)

by obtaining the boundary of the rejection region with the size of the test as close to \(\alpha\) as possible. Because the rejection region constructed using Zelterman’s technique contains the size-\(\alpha\) LRT rejection region, this test is also uniformly more powerful than the LRT. In Section 2.8.2 in the Appendix, we utilize Zelterman’s strategy to construct an approximate size-\(\alpha\) test that is uniformly more powerful than our LRT of Section 2.3. However, because the rejection region is approximated using numerical techniques, the size of the test using Zelterman’s method may slightly exceed \(\alpha\).

2.5.3 Comments on the Uniformly More Powerful Tests

The null and the alternative hypotheses we considered in this paper are shown in (2.4), with the alternative parameter space being a union of two obtuse cones. To our knowledge, no tests have been constructed in an analytical way that are uniformly more powerful than the size-\(\alpha\) LRT when the alternative parameter space is a union of two obtuse cones. Even if it is possible to construct a uniformly more powerful test in this case, it is questionable if
such a test would be “superior” to the LRT. First of all, all the uniformly more powerful tests have been constructed under the condition that the variances are known. Without known variances, the uniformly more powerful tests encounter various problems in the construction procedures or in maintaining the test levels. Thus, in real data analyses when the variances are not known, the uniformly more powerful test cannot be applied. Secondly, Perlman and Wu (1999) pointed out that, because the LRT rejection region is a subset of the uniformly more powerful tests, for any true parameters in the null parameter space, the probability for the uniformly more powerful tests to reject the null hypothesis is greater than the LRT. This means, the uniformly more powerful tests have larger Type-I error than the LRT for any parameters in the null parameter space. Cohen, Gatsonis and Marden (1983) showed that the LRT is the uniformly most powerful test among all monotone tests with level-\( \alpha \). Laska and Meisner (1989) showed that because the new constructed tests are not monotone, they might give testing results that are counterintuitive.

We use an example to illustrate Laska and Meisner’s (1989) point. Figure 2.8 shows the uniformly more powerful test constructed by Berger (1989), for the hypothesis in (2.9). The shaded area \( \bigcup_{v=1}^{6} R_v \) is Berger’s test rejection region, and \( R_1 \cup R_6 \) is the size-\( \alpha \) LRT rejection region. We can see that there are several facts that are counterintuitive. First, Berger’s test rejects the null hypothesis when the MLE of \( \xi \) (say \( \hat{\xi} \)) lies in the red area. However, the red area is actually a part of the null parameter space. Second, the LRT statistic is smaller when \( \hat{\xi} = a \) than when \( \hat{\xi} = b \). However, by using Berger’s test, the null hypothesis is rejected when \( \hat{\xi} = a \) but is not rejected when \( \hat{\xi} = b \). Both of these two facts demonstrate that the uniformly more powerful test constructed using Berger’s strategy may yield testing result that are counterintuitive.

In general, although the uniformly more powerful tests might of some theoretical interest, we ultimately recommend the likelihood ratio test in practical applications.

2.6 An Example

Approximately 95% of U.S. maize is produced from hybrids (Swanson-Wagner, 2006). One example of phenotypic heterosis in maize is the F1 hybrid offspring with parent inbred lines B73 and Mo17. F1 has taller plants, shorter maturation time, and higher yields than both its parents (Hallauer and Miranda, 1981). In this example, we compare the gene
expressions of B73, Mo17 and F1 using the LRT with unknown $\sigma^2$ introduced in Section 2.3, the IUT discussed in Section 2.4, and the method employed by Swanson-Wagner et al. (2006). The method used by Swanson-Wagner et al. (2006) first conducts the overall F-test. Then for the genes that show significance in the overall test, a follow-up t-test is carried out to detect genes with significant expression heterosis. The experiment described in Swanson-Wagner (2006) has a randomized complete block design. So the estimate of the MSE and its corresponding degrees of freedom need to be concordant with the design, which is $(\text{the number of genotypes} - 1) \times (\text{the number of blocks} - 1)$. Other than that, the procedures of our LRT and the IUT are not affected by the design.

Figure 2.9 shows some examples of genes with and without significant evidence of expression heterosis. The left panel gives six example genes with their estimated mean expression values of the two inbred lines B73 and Mo17, and the hybrid F1 offspring, projected in a two-dimensional sub-space of the $\mu$ parameter space. The left panel also gives the projections of the rejection regions of the LRT (shaded), the IUT (smaller cones with boundaries parallel to the boundary of the LRT), and the overall F-test (outside of the circle). The right panel gives box-plots of the genotype-specific expression distributions for these six genes.

Gene A and gene E are identified by the LRT with low-parent heterosis, where the mean gene expression of F1 smaller than the mean gene expressions of either B73 or Mo17.
Gene B and gene F are identified by the LRT with high-parent heterosis, where the mean gene expression of F1 greater than the mean gene expressions of either B73 or Mo17. The IUT does not identify genes B, E, or F with significant heterosis. Gene C and gene D are two genes without significant evidence of heterosis. The mean expression values for gene C follow the pattern B73 < F1 < Mo17, and the mean expression values for gene D follow the pattern Mo17 < F1 < B73. The estimated means of standardized expressions for gene A through F are provided in the first three columns of Table 2.1, and the p-values from tests for evidence of gene expression heterosis using the size-α LRT and the IUT are also given in this table (the fourth column and the fifth column).

<table>
<thead>
<tr>
<th>Gene</th>
<th>B73</th>
<th>F1</th>
<th>Mo17</th>
<th>LRT</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.657</td>
<td>-1.371</td>
<td>-0.651</td>
<td>1.098 × 10^{-4}</td>
<td>4.115 × 10^{-4}</td>
</tr>
<tr>
<td>B</td>
<td>2.291</td>
<td>2.472</td>
<td>2.295</td>
<td>0.022</td>
<td>0.064</td>
</tr>
<tr>
<td>C</td>
<td>1.305</td>
<td>1.560</td>
<td>1.600</td>
<td>0.5</td>
<td>0.833</td>
</tr>
<tr>
<td>D</td>
<td>0.292</td>
<td>0.164</td>
<td>-0.098</td>
<td>0.5</td>
<td>0.833</td>
</tr>
<tr>
<td>E</td>
<td>-1.007</td>
<td>-1.140</td>
<td>-1.007</td>
<td>0.032</td>
<td>0.091</td>
</tr>
<tr>
<td>F</td>
<td>-0.152</td>
<td>0.067</td>
<td>-0.564</td>
<td>0.040</td>
<td>0.109</td>
</tr>
</tbody>
</table>

Table 2.1: Normalized estimated means for genes A through F, and comparisons of p-values using LRT, IUT and overall F-test.
The null and alternative hypotheses in the first step test of Swanson-Wagner et al. (2006) approach are \( \mu_1 = \mu_2 = \mu_3 \) and \( \mu_i \)'s not all equal, respectively. The preliminary test rejects the null hypothesis if \( F_{\text{test}} = \frac{\sum_{i=1}^{3} n(Y_i - \bar{Y})^2}{\text{MSE}} / (3-1) \geq F_{2,18, \alpha} \). In terms of \( X \), the test rejects the null if \( X_1^2 + X_2^2 \geq 2F_{2,18, \alpha} \). The circle on the left panel in Figure 2.9 shows the boundary of the overall F-test rejection region. Gene expression means that project within the circle will show no significant difference among the genotypes, so no further study will be done on these genes. We can see that there are two areas inside both the circle and the shaded region. Data with means that project within the overlapping areas will be identified as having significant evidence of heterosis by the LRT, but will fail to be rejected by the overall F-test. Thus, pre-screening with the overall F test can fail to identify heterotic traits that the LRT can detect. For example, gene B and gene E in Figure 2.9 are not rejected by the overall test. On the other hand, gene D does have a significant overall F-test even though it shows no evidence of heterosis. The p-values using the overall F-test are listed in the last column of Table 2.1.

2.7 Summary and Conclusions

In this paper, we considered a hypothesis that is motivated by the problem of detecting heterosis for a trait of interest. The null parameter space of no heterosis is a union of two convex cones. The alternative parameter space is the complement of the null and is also a union of two convex cones. We derived the size-\( \alpha \) LRT with and without known sample variance \( \sigma^2 \). Because the null parameter space is a union of two cones, we also considered the intersection-union test (IUT). We derived the testing procedure of the IUT, compared it with the LRT, and drew the conclusion that the LRT is uniformly more powerful than the IUT for detecting heterosis.

We also found that, for any parameters in the null space, the type-I error for the size-\( \alpha \) LRT is smaller than \( \alpha \), and the least favorable distribution is only achieved in the limiting case as the parameter moves on the boundary of the null space away from the intersection point of the null cones. We discuss the situations in which tests that are uniformly more powerful than the LRT can be constructed. We also derived an approximate test that is uniformly more powerful than the size-\( \alpha \) LRT, by using the strategy proposed by Zelterman (1990), for the case that \( \sigma^2 \) is known. However, there are concerns about such uniformly...
In the example of detecting heterosis using maize gene expression data, we compared the result of the LRT, the IUT and the method by Swanson-Wagner (2006). We found examples of genes that showed significant evidence of heterosis at the 0.05 level with the LRT but not with the IUT or the methods used by Swanson-Wagner et al. (2006). Although uniformly more powerful tests shows larger detection power than the LRT in theory, they are not applicable in the real data analyses. In general, we recommend the use of the LRT in practical data analyses.

2.8 Appendix

2.8.1 Derivation of the LRT

For the case when $\sigma^2$ is known, the ratio of the maximized likelihoods under $H_0$ and $H_0 \cup H_1$ is

$$\Lambda = e^{-\frac{\sum_{i=1}^{3} \sum_{j=1}^{n} (Y_{ij} - \tilde{\mu}_i)^2 - (Y_{ij} - \bar{Y}_i)^2)}{2\sigma^2}},$$

where $\bar{Y}_i = \frac{\sum_{j=1}^{n} Y_{ij}}{n}$ ($i = 1, 2, 3$), and $\tilde{\mu} = (\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\mu}_3)'$ is the MLE of the parameters with the restriction that $\mu_1 \leq \mu_2 \leq \mu_3$ or $\mu_1 \geq \mu_2 \geq \mu_3$. Equivalently, the LRT rejects the null hypothesis if $\chi^2 = \frac{n \sum_{i=1}^{3} (Y_{ij} - \bar{Y}_i)^2}{\sigma^2}$ is too large. We evaluate $\chi^2$ as follows:

$$\chi^2 = \begin{cases} 0, & \text{if } \min(\bar{Y}_1, \bar{Y}_3) \leq \bar{Y}_2 \leq \max(\bar{Y}_1, \bar{Y}_3); \\ \frac{n(\bar{Y}_2 - \min(\bar{Y}_1, \bar{Y}_3))^2}{2\sigma^2}, & \text{if } \bar{Y}_2 < \min(\bar{Y}_1, \bar{Y}_3); \\ \frac{n(\bar{Y}_2 - \max(\bar{Y}_1, \bar{Y}_3))^2}{2\sigma^2}, & \text{if } \bar{Y}_2 > \max(\bar{Y}_1, \bar{Y}_3). \end{cases}$$

The size-$\alpha$ LRT is to reject $\chi^2 \geq c$ with some relevant values of $c$, such that

$$\sup \{ P(\chi^2 \geq c | \mu_1, \mu_2, \mu_3): \mu_1 \leq \mu_2 \leq \mu_3 \text{ or } \mu_1 \geq \mu_2 \geq \mu_3 \} = \alpha. \quad (2.10)$$

Expressing $\chi^2$ using $X$, we have

$$\chi^2 = \begin{cases} 0, & \text{if } -X_2 \cdot \text{sign}(X_2) \leq \sqrt{3}X_1 \leq X_2 \cdot \text{sign}(X_2); \\ \left(\frac{\sqrt{3}}{2}X_1 + \frac{1}{2}X_2 \cdot \text{sign}(X_2)\right)^2, & \text{if } \sqrt{3}X_1 < -X_2 \cdot \text{sign}(X_2); \\ \left(\frac{\sqrt{3}}{2}X_1 - \frac{1}{2}X_2 \cdot \text{sign}(X_2)\right)^2, & \text{if } \sqrt{3}X_1 > X_2 \cdot \text{sign}(X_2). \end{cases}$$
With some relevant values of $c$, the size-$\alpha$ LRT rejection region can be expressed as

$$R_L = \{x: -\sqrt{3}X_1 + 2c \leq X_2 \leq \sqrt{3}X_1 - 2c\}$$

$$\cup \{x: \sqrt{3}X_1 + 2c \leq X_2 \leq -\sqrt{3}X_1 - 2c\},$$  \hspace{1cm} (2.11)

such that

$$\sup_{\theta \in H_0} \{P(x \in R_L)\} = \alpha. \hspace{1cm} (2.12)$$

By Sasabuchi (1980), for the case when the null parameter space is a union of two vertical cones, and the alternative parameter space being the complement of the null parameter space is also a union of two vertical cones, the least favorable cases in the null parameter space can be determined only when each cone of the alternative parameter space is an acute cone. In this paper, we prove that, with some relevant values of $c$ not too small, the supremum of the left side of (2.10) achieves with any one of the following four cases:

$$\theta_1 \to -\infty \text{ and } h'_1 \theta = 0,$$

or

$$\theta_1 \to \infty \text{ and } h'_1 \theta = 0,$$

or

$$\theta_1 \to -\infty \text{ and } h'_2 \theta = 0,$$

or

$$\theta_1 \to \infty \text{ and } h'_2 \theta = 0. \hspace{1cm} (2.13)$$

**Proof.** First we prove that for any $\theta = (\theta_1, \theta_2)' \in H_0$, there exists a $\theta_1$ on the boundary of the null parameter space, i.e., $\theta_1 \in H_0^0 \triangleq \{\theta: h'_1 \theta = 0 \text{ or } h'_2 \theta = 0\}$, such that $P_{\theta}(X \in R_L) \leq P_{\theta_1}(X \in R_L)$. So the least favorable cases only happen on the boundary of the null parameter space.

For random variable $X = (X_1, X_2)'$ with mean $\theta = (\theta_1, \theta_2)'$, we have if $(x_1, x_2)' \in R_L$, then $(x_1, -x_2)' \in R_L$. Consider the transformation $\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = \Delta \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with $\Delta = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$. Then the mean for $(T_1, T_2)'$ is $\eta = (\eta_1, \eta_2)' = \Delta(\theta_1, \theta_2)'$. The LRT rejection region in the $(T_1, T_2)'$ coordinate system is $\Delta R_L = \{t: \Delta^{-1}t \in R_L\}$. We have that if $(t_1, t_2)' \in \Delta R_L$, then $(t_2, t_1)' \in \Delta R_L$. So $\Delta R_L$ satisfies:

For any $(t_{11}, t_{12})' \in \Delta R_L$, if $(t_{21}, t_{22})'$ is majorized by $(t_{11}, t_{12})'$, then we have $(t_{21}, t_{22})' \in \Delta R_L$. Here $(t_{21}, t_{22})'$ is majorized by $(t_{11}, t_{12})'$ means $\max\{t_{21}, t_{22}\} \leq \max\{t_{11}, t_{12}\}$, and $t_{11} + t_{12} = t_{21} + t_{22}$, which is denoted as $(t_{21}, t_{22})' \prec (t_{11}, t_{12})'$. 

Also, for any \((t_{21}, t_{22})'\) and \((t_{11}, t_{12})'\), if \((t_{21}, t_{22})' < (t_{11}, t_{12})'\), then \(\phi(t_{21})\phi(t_{22}) \geq \phi(t_{11})\phi(t_{12})\), i.e., \(\phi(t_1)\phi(t_2)\) is Schur-Concave function of \((t_1, t_2)'\). By Theorem 2.1 of Marshall and Olkin (1974), \(\int_{\Delta R_L + \eta} \phi(t_1)\phi(t_2) \, dt_1 \, dt_2\) is a Schur-Concave function of \(\eta\), where \(\int_{\Delta R_L + \eta} \phi(t_1)\phi(t_2) \, dt_1 \, dt_2 = P_{(0,0)'}(\{T_1, T_2)' \in \Delta R_L + \eta\} = P_{\eta}(\{T_1, T_2)' \in \Delta R_L\).

Consequently, for any \(\theta \in H_0\) and its corresponding \(\eta = \Delta \theta\), there exists an

\[
\eta_1 = \begin{cases} 
(\sqrt{\frac{3}{2}} - \sqrt{\frac{3}{2}}, \sqrt{\frac{3}{2}} + \sqrt{\frac{3}{2}}) \theta_1 & \text{if } \theta_1 \theta_2 \geq 0, \\
(\sqrt{\frac{3}{2}} + \sqrt{\frac{3}{2}}, \sqrt{\frac{3}{2}} - \sqrt{\frac{3}{2}}) \theta_1 & \text{if } \theta_1 \theta_2 < 0,
\end{cases}
\]  

(2.14)

with corresponding \(\theta_1 = (\theta_{11}, \theta_{12})' = \Delta^{-1} \eta_1 \in H_0^0\), such that \(\eta_1 \prec \eta\) and

\[
\int_{\Delta R_L + \eta} \phi(t_1)\phi(t_2) \, dt_1 \, dt_2 \leq \int_{\Delta R_L + \eta_1} \phi(t_1)\phi(t_2) \, dt_1 \, dt_2.
\]

So we have

\[
P_{\theta}(\{X_1, X_2)' \in R_L\) = \int_{R_L + \theta} \phi(x_1)\phi(x_2) \, dx_1 \, dx_2 \\
= \int_{\Delta R_L + \eta} \phi(t_1)\phi(t_2) \, dt_1 \, dt_2 \\
\leq \int_{\Delta R_L + \eta_1} \phi(t_1)\phi(t_2) \, dt_1 \, dt_2 \\
= \int_{R_L + \theta_1} \phi(x_1)\phi(x_2) \, dx_1 \, dx_2 \\
= P_{\theta_1}(\{X_1, X_2)' \in R_L),
\]

which means the least favorable cases only happen on the boundary of the null parameter space, \(H_0^0\).

Now we are going to prove that under certain conditions of \(c\), the least favorable cases in \(H_0^0\) are the cases in (2.13).

\[
\sup_{\theta \in H_0^0} \{ P(X \in R_L) \} = \sup_{\theta \in H_0^0} \left\{ P\left(\frac{\sqrt{3}}{2} X_1 + \frac{1}{2} X_2 \geq c, \frac{\sqrt{3}}{2} X_1 - \frac{1}{2} X_2 \geq c \right) \right\} \\
+ P\left(\frac{\sqrt{3}}{2} X_1 + \frac{1}{2} X_2 \leq -c, \frac{\sqrt{3}}{2} X_1 - \frac{1}{2} X_2 \leq -c \right).
\]

Let \(W_1 = (\sqrt{\frac{3}{2}}, \frac{1}{2})(X - \theta)\), \(W_2 = (\sqrt{\frac{3}{2}}, -\frac{1}{2})(X - \theta)\). Then \(W = (W_1, W_2)' \sim \mathcal{N}_2(0, R),\)
where $R = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$. So

$$\sup_{\theta \in H_0^c} \{ P(X \in R_L) \}$$

$$= \sup_{\theta \in H_0^c} \left\{ P(W_1 \geq c - \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\theta, W_2 \geq c - \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)\theta) + P(W_1 \leq -c - \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\theta, W_2 \leq -c - \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)\theta) \right\}.$$

Let $\tau_1 = \left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)\theta$, $\tau_2 = \left(\frac{\sqrt{3}}{2}, -\frac{1}{2}\right)\theta$, and

$$\zeta(\tau_1, \tau_2; c)$$

$$= P(W_1 \geq c - \tau_1, W_2 \geq c - \tau_2) + P(W_1 \leq -c - \tau_1, W_2 \leq -c - \tau_2)$$

$$= \int_{c-\tau_1}^{\infty} \int_{c-\tau_2}^{\infty} n_{0,R}(w_1, w_2) \, dw_1 \, dw_2 + \int_{c+\tau_1}^{\infty} \int_{c+\tau_2}^{\infty} n_{0,R}(w_1, w_2) \, dw_1 \, dw_2,$$

where $n_{0,R}(w_1, w_2)$ denotes the two-dimensional multivariate normal density function with mean $0 = (0, 0)'$ and variance matrix $R$. Notice that $\tau_1 = 0$ is equivalent to $h_1'\theta = 0$, $\tau_2 = 0$ is equivalent to $h_2'\theta = 0$. Taking derivative of $\zeta(\tau_1, \tau_2; c)$ with respect to $\tau_1$ and evaluating it at $\tau_2 = 0$, we get

$$\frac{d}{d\tau_1} \zeta(\tau_1, \tau_2; c) \bigg|_{\tau_2=0}$$

$$= \int_{c}^{\infty} n_{0,R}(c - \tau_1, w_2) - n_{0,R}(c + \tau_1, w_2) \, dw_2$$

$$= \frac{1}{2\pi |R|^{1/2}} \int_{c}^{\infty} \exp \left( - \frac{2}{3}(c - \tau_1)^{2} + \frac{2}{3}(c - \tau_1)w_2 - \frac{2}{3}w_2^{2} \right)$$

$$- \exp \left( - \frac{2}{3}(c + \tau_1)^{2} + \frac{2}{3}(c + \tau_1)w_2 - \frac{2}{3}w_2^{2} \right) \, dw_2. \tag{2.15}$$

Consider

$$\left( - \frac{2}{3}(c - \tau_1)^{2} + \frac{2}{3}(c - \tau_1)w_2 - \frac{2}{3}w_2^{2} \right)$$

$$- \left( - \frac{2}{3}(c + \tau_1)^{2} + \frac{2}{3}(c + \tau_1)w_2 - \frac{2}{3}w_2^{2} \right)$$

$$= \frac{4}{3} \tau_1(2c - w_2),$$

if $\tau_1 > 0$, we have $\int_{c}^{2c} n_{0,R}(c - \tau_1, w_2) - n_{0,R}(c + \tau_1, w_2) \, dw_2 > 0$, and $\int_{-\infty}^{c} n_{0,R}(c - \tau_1, w_2) - n_{0,R}(c + \tau_1, w_2) \, dw_2 \leq 0$; if $\tau_1 < 0$, $\int_{c}^{2c} n_{0,R}(c - \tau_1, w_2) - n_{0,R}(c + \tau_1, w_2) \, dw_2 < 0$, and $\int_{-\infty}^{c} n_{0,R}(c - \tau_1, w_2) - n_{0,R}(c + \tau_1, w_2) \, dw_2 \geq 0.$
We have the conclusion that under the condition that \( c \) is not too small, for \( \tau_1 > 0 \),

\[
\left. \frac{d}{d\tau_1} \zeta(\tau_1, \tau_2; c) \right|_{\tau_2=0} > 0,
\]
and \( f(\tau_1, \tau_2; c) \) is increasing as \( \tau_1 \) increasing; for \( \tau_1 < 0 \),

\[
\left. \frac{d}{d\tau_1} \zeta(\tau_1, \tau_2; c) \right|_{\tau_2=0} < 0,
\]
and \( \zeta(\tau_1, \tau_2; c) \) is increasing as \( \tau_1 \) decreasing. So we proved that with \( c \) not too small, the maximum value of the power function attains at \( \theta_1 \to \infty, h'_2 \theta = 0, \) or \( \theta_1 \to -\infty, h'_2 \theta = 0. \)

Similarly, by taking derivative of \( \zeta(\tau_1, \tau_2; c) \) with respect to \( \tau_2 \) and evaluating it at \( \tau_1 = 0 \), we can prove that with the same condition of \( c \), the maximum value of the power function also attains at \( \theta_1 \to \infty, h'_1 \theta = 0, \) or \( \theta_1 \to -\infty, h'_1 \theta = 0. \)

Equivalently, with the same condition on \( c > 0 \), the least favorable cases in the null parameter space is expressed in terms of \( Y \) and \( \mu \) are

\[
\mu_1 \to -\infty \text{ and } \mu_2 = \mu_3, \\
\text{or } \mu_1 \to \infty \text{ and } \mu_2 = \mu_3, \\
\text{or } \mu_3 \to -\infty \text{ and } \mu_1 = \mu_2, \\
\text{or } \mu_3 \to \infty \text{ and } \mu_1 = \mu_2. \tag{2.16}
\]

So we have

\[
\sup \left\{ P(\chi^2 \geq c|\mu_1, \mu_2, \mu_3) : \mu_1 \leq \mu_2 \leq \mu_3 \text{ or } \mu_1 \geq \mu_2 \geq \mu_3 \right\} = \lim_{\mu_1 \to -\infty} P(\chi^2 \geq c|\mu_1, \mu_2 = \mu_3) = \lim_{\mu_1 \to -\infty} P(\chi^2 \geq c|\mu_1, \mu_2 = \mu_3)
\]

\[
= \lim_{\mu_3 \to -\infty} P(\chi^2 \geq c|\mu_1 = \mu_2, \mu_3) = \lim_{\mu_3 \to -\infty} P(\chi^2 \geq c|\mu_1 = \mu_2, \mu_3). \tag{2.17}
\]

Taking \( \lim_{\mu_1 \to -\infty} P(\chi^2 \geq c|\mu_1, \mu_2 = \mu_3) \) as an example,

\[
\sup \left\{ P(\chi^2 \geq c|\mu_1, \mu_2, \mu_3) : \mu_1 \leq \mu_2 \leq \mu_3 \text{ or } \mu_1 \geq \mu_2 \geq \mu_3 \right\} = \lim_{\mu_1 \to -\infty} P(\chi^2 \geq c|\mu_1, \mu_2 = \mu_3)
\]

\[
= \left( \lim_{\mu_1 \to -\infty} P(\bar{Y}_1 < \bar{Y}_2 < \bar{Y}_3 | \mu_1, \mu_2 = \mu_3) \right) \mathbb{I}_{\{0 \geq c\}}
\]

\[
+ \left( \lim_{\mu_1 \to -\infty} P(\bar{Y}_1 < \bar{Y}_3 < \bar{Y}_2 | \mu_1, \mu_2 = \mu_3) \right) P\left( \frac{(n\bar{Y}_2 - \bar{Y}_3)^2}{2\sigma^2} \geq c \right)
\]

\[
= \frac{1}{2} \mathbb{I}_{\{0 \geq c\}} + \frac{1}{2} P(\chi^2_1 \geq c),
\]
we have \( c = \chi^2_{1,2\alpha} \), which is the upper \( 2\alpha \) quantile of \( \chi^2 \). So the size-\( \alpha \) LRT with known \( \sigma^2 \) is to reject \( H_0 \) with \( \chi^2 \geq \chi^2_{1,2\alpha} \). Or in terms of \( X \) and \( \theta \), the size-\( \alpha \) LRT rejection region is

\[
R_L = \{ x : -\sqrt{3}x_1 + 2z_\alpha \leq x_2 \leq \sqrt{3}x_1 - 2z_\alpha \} \\
\cup \{ x : \sqrt{3}x_1 + 2z_\alpha \leq x_2 \leq -\sqrt{3}x_1 - 2z_\alpha \}.
\]

For the case when \( \sigma^2 \) is unknown,

\[
\Lambda = \frac{1}{(2\pi \hat{\sigma}^2)^{\frac{3n}{2}}} \exp\left(-\frac{\sum_{i=1}^{3n} (Y_{ij} - \bar{Y}_i)^2}{2\hat{\sigma}^2}\right),
\]

where \( \hat{\sigma}^2 = \sum_{i=1}^{3} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2 / (3n - 3) = \text{MSE}, \hat{\sigma}^2 = \sum_{i=1}^{3} \sum_{j=1}^{n} (Y_{ij} - \bar{\mu}_i)^2 / (3n - 3) \).

So

\[
\Lambda = \frac{\left(\frac{\sum_{i=1}^{3} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2}{\sum_{i=1}^{3} \sum_{j=1}^{n} (Y_{ij} - \bar{\mu}_i)^2}\right)^{\frac{3n}{2}}}{(3n - 3)\text{MSE} + n \sum_{i=1}^{3} (Y_i - \bar{\mu}_i)^2)^{\frac{3n}{4}}}.
\]

\( H_0 \) is rejected when \( \Lambda \) is small, which means rejecting if

\[
F = \frac{n \sum_{i=1}^{3} (Y_i - \bar{\mu}_i)^2}{\text{MSE}}
\]

is too large.

\[
F = \begin{cases} 
0, & \text{if } \min(\bar{Y}_1, \bar{Y}_3) \leq \bar{Y}_2 \leq \max(\bar{Y}_1, \bar{Y}_3); \\
\frac{n(\bar{Y}_2 - \min(\bar{Y}_1, \bar{Y}_3))^2}{2\text{MSE}}, & \text{if } \bar{Y}_2 < \min(\bar{Y}_1, \bar{Y}_3); \\
\frac{n(\bar{Y}_2 - \max(\bar{Y}_1, \bar{Y}_3))^2}{2\text{MSE}}, & \text{if } \bar{Y}_2 > \max(\bar{Y}_1, \bar{Y}_3).
\end{cases}
\]

In order to generalize to the case with \( \sigma^2 \) unknown that, for some values of \( c \) not too small, the least favorable cases in the null parameter space is \((2.13)\) in terms of \( X \) and \( \theta \), first we prove the least favorable cases only happen on the boundary of the null parameter space, \( H_0^0 \). Here we only need to prove that the central multivariate \( t \) distribution

\[
f_{cmvt}(t_1, t_2) = \frac{\Gamma \left(\frac{3n-1}{2}\right)}{(3n-3)\pi \Gamma \left(\frac{3n-2}{2}\right)} \left(1 + \frac{t_1^2 + t_2^2}{3n-3}\right)^{-\frac{3n-1}{2}}
\]
is a Schur-Concave function of \((t_1, t_2)'\). This is true because, if \((t_{21}, t_{22})' < (t_{11}, t_{12})'\) then 
\[ t_{21}^2 + t_{22}^2 \leq t_{11}^2 + t_{12}^2. \]
We have \(f_{c\text{met}}((t_{11}, t_{12})') \leq f_{c\text{met}}((t_{21}, t_{22})')\), i.e., \(f_{c\text{met}}((t_{11}, t_{12})')\) is a Schur-Concave function of \((t_{11}, t_{12})'\). By Theorem 2.1 of Marshall and Olkin (1974), 
\[ \int_{\Delta R_L+\eta} f_{c\text{met}}((t_{11}, t_{12})')dt_1dt_2 \] is a Schur-Concave function of \((t_{11}, t_{12})'\). By similar process, for any \(\theta \in H_0\), there exists an \(\theta_1 \in H_0^0\), and \(P_{\theta}(X_1, X_2)' \in R_L\) \(\leq P_{\theta_1}(X_1, X_2)' \in R_L\). So when \(\sigma^2\) is unknown, the least favorable cases also only happen on the boundary of the null parameter space, \(H_0^0\).

The next step is to prove with the same condition of \(c\) (where \(c\) not too small), the least favorable cases in \(H_0^0\) are the cases in (2.13). With the transformation \(W_1 = (\sqrt[3]{2}, \frac{1}{2})(X - \theta)\) and \(W_2 = (\sqrt[3]{2}, -\frac{1}{2})(X - \theta)\), \(W = (W_1, W_2)' \sim c\text{met}_3\text{sn-3}(0, R)\) with probability function 
\[ f_{c\text{met}_3\text{sn-3}}(w_1, w_2) = \frac{\sqrt[3]{2}}{2} \Gamma \left( \frac{3n-1}{2} \right) \left( 1 + \left( \frac{\sqrt[3]{2}}{3} w_1 + \frac{\sqrt[3]{2}}{3} w_2 \right)^2 + \left( w_1 - w_2 \right)^2 \right)^{-\frac{3n-1}{2}}. \]
So
\[
\zeta_c(\tau_1, \tau_2; c) \]
\[
= \begin{cases} 
P(W_1 > c - \tau_1, W_2 > c - \tau_2) + P(W_1 \leq c - \tau_1, W_2 \leq c - \tau_2) 
\end{cases}
\]
\[
= \int_{c-\tau_1}^{\infty} \int_{c-\tau_2}^{\infty} f_{c\text{met}_3\text{sn-3}}(0, R)(w_1, w_2) dw_1 dw_2 + \int_{c+\tau_1}^{\infty} \int_{c+\tau_2}^{\infty} f_{c\text{met}_3\text{sn-3}}(0, R)(w_1, w_2) dw_1 dw_2.
\]

\[
\frac{d}{d\tau_1} \zeta_c(\tau_1, \tau_2; c) \bigg|_{\tau_2=0} = \int_{c}^{\infty} f_{c\text{met}_3\text{sn-3}}(0, R)(c - \tau_1, w_2) - f_{c\text{met}_3\text{sn-3}}(0, R)(c + \tau_1, w_2) dw_2
\]
\[
= \frac{\sqrt[3]{2}}{2} \Gamma \left( \frac{3n-1}{2} \right) \int_{c}^{\infty} \left( 1 + \frac{(c - \tau_1 + w_2)^2}{3} + \frac{(c + \tau_1 - w_2)^2}{3} \right)^{-\frac{3n-1}{2}} dw_2.
\]

Consider \((c - \tau_1 + w_2)^2/3 + (c - \tau_1 - w_2)^2 - (c + \tau_1 + w_2)^2/3 + (c + \tau_1 - w_2)^2 = \frac{8}{3} \tau_1 (w_2 - 2c).\)

So we draw the conclusion that with the same condition of \(c\) that \(c\) is not too small, for \(\tau_1 > 0, \frac{d}{d\tau_1} \zeta_c(\tau_1, \tau_2; c) \bigg|_{\tau_2=0} > 0;\) for \(\tau_1 < 0, \frac{d}{d\tau_1} \zeta_c(\tau_1, \tau_2; c) \bigg|_{\tau_2=0} < 0.\) And the least favorable cases in the null parameter space is (2.13) in terms of \(\theta\) or (2.16) in terms of \(\mu.\)
Consequently,
\[
\sup \{ P(\bar{F} \geq c|\mu_1, \mu_2, \mu_3) : \mu_1 \leq \mu_2 \leq \mu_3 \text{ or } \mu_1 \geq \mu_2 \geq \mu_3 \} = \lim_{\mu_1 \to -\infty} P(\bar{F} \geq c|\mu_1, \mu_2 \leq \mu_3) = \lim_{\mu_1 \to -\infty} P(\bar{F} \geq c|\mu_1, \mu_2 = \mu_3) = \frac{1}{2} \mathbb{I}_{\{0 \geq c\}} + \frac{1}{2} P(F_{1,3n-3,2} \geq c).
\]

So the critical value for the size-\(\alpha\) LRT is \(F_{1,3n-3,2\alpha}\), the upper \(2\alpha\) quantile of \(F\) distribution with degree of freedom 1 and \(3n - 3\). Expressing the rejection region in terms of \(X\),

\[
R_L(\hat{\sigma}) = \{ x : -\sqrt{3}x_1 + 2t_{3n-3,\alpha} \leq x_2 \leq \sqrt{3}x_1 - 2t_{3n-3,\alpha} \} \]

\[
\cup \{ x : \sqrt{3}x_1 + 2t_{3n-3,\alpha} \leq x_2 \leq -\sqrt{3}x_1 - 2t_{3n-3,\alpha} \}.
\]

### 2.8.2 Application of Zelterman’s Test on Our Testing Problem

We construct the uniformly more powerful test to our testing problem, following Zelterman’s strategy. The main idea of Zelterman’s method is to enlarge the size-\(\alpha\) LRT rejection region with the constraint that the rejection probability at the boundary of the new test rejection region strictly equals \(\alpha\). Denote \(d(x)\) as the rejection rule for \(H_0\) vs. \(H_1\), i.e., \(d(x) = 1\) if \(H_0\) is rejected with observation \(x\) and \(d(x) = 0\) otherwise. The power function is denoted as \(P_\theta(d(X) = 1)\). Suppose \(q(x_1)\) is a non-negative function symmetric with respect to \(X_1\), which defines the rejection region as follows,

\[
d(x) = \begin{cases} 
1, & \text{if } |x_2| \leq q(x_1), \\
0, & \text{otherwise},
\end{cases}
\]

where \(q(x_1) \geq 0\) and \(q(-x_1) = q(x_1)\).

By Zelterman’s construction, the expression of \(q(x_1)\) is as follows

\[
q(x_1) = \begin{cases} 
\Phi^{-1}\left(\sum_{j=0}^{\infty} c_{2j} H_{2j}(x_1)\right), & \text{if } |x_1| < x_{1\alpha}, \\
|x_1| - 2^{1/2} \Phi^{-1}(1-\alpha), & \text{if } |x_1| \geq x_{1\alpha},
\end{cases}
\]

where \(c_{2j} \ (j = 1, 2, \ldots)\) are the coefficients, and \(H_n(x_1)\) are the probabilists’ Hermite polynomials defined as follows:

\[
\left( \frac{\partial}{\partial x_1} \right)^n \phi(x_1) = (-1)^n H_n(x_1) \phi(x_1), \quad n = 1, 2, \ldots, \quad \text{and } H_0(x_1) = 1.
\]
Figure 2.10: A schematic plot of a rejection region using Zelterman’s method

By defining $q(x_1)$ using this piecewise function, the new test coincides with the LRT when $x_1$ is greater than $x_{1\alpha}$. The values of $x_{1\alpha}$ and all $c_{2j}$ ($j = 1, 2, \ldots$) are determined numerically by the constraint

$$\int_{-\infty}^{\infty} |P_{(\theta_1, \sqrt{3}\theta_1)}(d(X) = 1) - \alpha| d\theta_1 = 0,$$  \hspace{1cm} (2.20)

where

$$P_{\theta}(d(X) = 1) = \int_{-\infty}^{\infty} \phi(x_1 - \theta_1)\phi(x_2 - \theta_2) dx_2 dx_1$$

$$= \int_{-\infty}^{\infty} \phi(x_1 - \theta_1)\left(\Phi(q(x_1) - \theta_2) + \Phi(q(x_1) + \theta_2)\right) dx_1 - 1.$$ \hspace{1cm} (2.21)

Figure 2.10 is a schematic plot of a rejection region for our problem derived using Zelterman’s method, analogous to the plot shown in Zelterman’s (1990) paper.

In the case where $\sigma^2$ is unknown, the values of $x_{1\alpha}$ and all $c_{2j}$, $j = 1, 2, \ldots$ are also determined by the constraint (2.20), while the power function in (2.21) needs to be cal-
culated using the non-central multivariate t distribution density function, i.e., \( P_\theta(d(X) = 1) = \int \int d(X) \int_{-\infty}^{\infty} f_{ncmvt}(\mathbf{z}^0) \, dx_3 \, dx_2 \).

### 2.8.2.1 Notes on Zelterman’s Test

There are some notes worth to be made here. First, all the unknown parameters are determined by the constraint in (2.20) with the numeric solutions. No analytical solution can be made here, which is not easy in real data analyses. It is even more complicated when \( \sigma^2 \) is unknown due to the dependency among the variables \( X_1, X_2 \) and \( X_3 \).

The more powerful test rejection region is a union of two areas: the LRT rejection region, which is a union of two cones, and the expanded area between the two cones. By Liu and Berger (1995), if this expanded area touches the LRT rejection region at more than the vertices of the LRT rejection region cones, then the size of the new test will be greater than \( \alpha \). So for test discussed in this paper, \( x_{1\alpha} \) should be no larger than \( \frac{2\sqrt{3}}{3} z_\alpha \) with known \( \sigma^2 \); and for the testing problem in Zelterman’s (1990) paper, \( x_{1\alpha} \) should be be no larger than \( \sqrt{2} z_\alpha \). However, this value shown on Zelterman’s paper is \( x_{1\alpha} = 2.7689 \) at \( \alpha = 0.05 \), which is greater than \( \sqrt{2} z_{0.05} \). This contradiction might be caused by a mistake in the derivation in Zelterman’s paper, which will be discussed in the following paragraph.

We use the notation of Zelterman’s paper within this paragraph, for the sake of the consistency. We will discuss the equation (2.8) in Zelterman’s paper, where

\[
\int_{-\infty}^{\infty} h_2(r)\phi(r)\Phi(q(r)) \, dr = \frac{1}{2} \alpha.
\]

By Zelterman’s statement, this is derived from the equations in Zelterman’s paper that, \( \int_{-\infty}^{\infty} \Phi(q(r))\phi(r) \, dr = \frac{1}{2}(\alpha + 1) \), and that \( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d(x)x_1x_2\phi(x_1)\phi(x_2) \, dx_1 \, dx_2 = -\alpha \) with
\( s = \frac{1}{\sqrt{2}}(x_1 + x_2) \) and \( r = \frac{1}{\sqrt{2}}(x_1 - x_2) \). We will see that this is not correct:

\[
-\alpha = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d(x) x_1 x_2 \phi(x_1) \phi(x_2) \, dx_1 \, dx_2 \\
= \int_{-\infty}^{\infty} \int_{-q(r)}^{q(r)} \frac{1}{2} \left( s^2 - r^2 \right) \phi(s) \phi(r) \, ds \, dr \\
= \int_{-\infty}^{\infty} \frac{1}{2} \phi(r) \int_{-q(r)}^{q(r)} (s^2 - r^2) \phi(s) \, ds \, dr \\
= \frac{1}{2} \int_{-\infty}^{\infty} \phi(r) \left( \int_{-q(r)}^{q(r)} s^2 \phi(s) \, ds - 2r^2 \Phi(q(r)) + r^2 \right) \, dr \\
= \int_{-\infty}^{\infty} \phi(r) \int_{0}^{q(r)} s^2 \phi(s) \, ds \, dr - \int_{-\infty}^{\infty} r^2 \phi(r) \Phi(q(r)) \, dr + \frac{1}{2} \\
= \int_{-\infty}^{\infty} \phi(r) \int_{0}^{q(r)} s^2 \phi(s) \, ds \, dr - \int_{-\infty}^{\infty} h_2(r) \phi(r) \Phi(q(r)) \, dr - \frac{1}{2} (\alpha + 1) + \frac{1}{2}, \\
= \int_{-\infty}^{\infty} \phi(r) \int_{0}^{q(r)} s^2 \phi(s) \, ds \, dr - \int_{-\infty}^{\infty} h_2(r) \phi(r) \Phi(q(r)) \, dr - \frac{1}{2} \alpha,
\]

which means

\[
\int_{-\infty}^{\infty} h_2(x) \phi(x) \Phi(q(r)) \, dr = \int_{-\infty}^{\infty} \phi(r) \int_{0}^{q(r)} s^2 \phi(s) \, ds \, dr + \frac{1}{2} \alpha
\]

with \( \int_{-\infty}^{\infty} \phi(r) \int_{0}^{q(r)} s^2 \phi(s) \, ds \, dr \geq 0 \). \( \int_{-\infty}^{\infty} \phi(r) \int_{0}^{q(r)} s^2 \phi(s) \, ds \, dr = 0 \) if and only if \( q(r) = 0 \) except sets with measure zero, so \( \int_{-\infty}^{\infty} h_2(r) \phi(r) \Phi(q(r)) \, dr > \frac{1}{2} \alpha \) with probability 1.

This mistake in Zelterman’s paper results in the incorrect values for \( x_{1\alpha} \) and \( c_{2j} \ (j = 1, 2, \ldots) \) obtained in Zelterman’s paper.
References


CHAPTER 3. Testing for Monotonic Changes in Multivariate Gene Expression Distributions

Abstract

We develop four procedures that can be used to test for association between a response vector and a treatment factor whose levels are quantitative or ordinal. The testing procedures are applied to the problem of identifying gene sets comprised of genes whose expression levels tend to change monotonically with the level of treatment. The proposed procedures are able to deal with small sample sizes and high-dimensional distributions, and simulation studies demonstrate that the proposed methods have higher power for detecting monotonic associations than traditional testing procedures that spread power more evenly over the whole alternative parameter space.

KEY WORDS: Gene Sets; Large \( p \), Small \( n \); Monotonic Alternative Hypothesis; Permutation Test; Order-restricted Statistical Inference.

3.1 Introduction

3.1.1 Background and Motivation

Many experimental factors are quantitative or ordinal and thus have levels that can be naturally ordered. As the level of such a factor is increased, it is often reasonable to assume that a mean response will change monotonically if the factor and response are associated with each other. For example, within certain dose ranges, as the dose of a drug is increased, the plasma concentration of the drug increases, and the effect or side effects of the drug may be strengthened. To study the molecular genetic mechanisms underlying phenomena like dose-response relationships, researchers measure the expression of genes in response to varying levels of quantitative or ordinal treatment factors. In these studies, finding genes with monotonic changes in mean expression is of great interest. To increase power
for detecting such genes, we propose four statistical tests that focus power on monotonic
departures from mean equality. We compare these four tests to three other standard testing
approaches that spread power more uniformly over the alternative parameter space.

The testing procedures that we propose are useful for detecting association between a
quantitative or ordinal factor and a single response variable. However, in gene expression
studies, we must deal simultaneously with thousands of response variables corresponding to
the expression levels of thousands of genes. To improve interpretability and to increase bio-
logical understanding, genes are often grouped into sets based on information obtained from
past research. Gene Ontology (GO) is one database that provides biological annotations
for genes, based on the knowledge that some genes are associated with each other in the
same cellular component or collaborate together to implement biological processes or com-
plete molecular functions (Ashburner, et al., 2000). Each GO category contains from under
10 to thousands of genes. Thus, in addition to considering tests of association between a
quantitative or ordinal factor and a single response variable, we also develop methods for
finding sets of genes (as defined by GO, for example) with expression levels that tend to be
monotonically associated with a quantitative or ordinal treatment factor.

3.1.2 A Formal Description of the Problem of Interest

The problem discussed in this paper can be described as follows. For $i = 1, \ldots, T$
and $j = 1, \ldots, n_i$, let $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijG})'$ denote the $j^{th}$ replication under the $i^{th}$
treatment of an expression vector corresponding to $G$ genes in a gene set of interest. Here
$Y_{ijg}$ represents the expression associated with treatment $i$, replication $j$, and gene $g$ in the
gene set. We assume that all the $n_i$ replicates under the $i^{th}$ treatment share one common
multivariate distribution $F_i$, i.e.;

$$Y_{i1}, Y_{i2}, \ldots, Y_{in_i} \sim F_i, \text{ for } i = 1, \ldots, T.$$ 

Furthermore, we assume that all $N \equiv \sum_{i=1}^{T} n_i$ random vectors are independent and that
$E(Y_{ij}) = (\mu_{i1}, \ldots, \mu_{iG})'$, $i = 1, \ldots, T$, $j = 1, \ldots, n_i$.

We wish to test

$$H_0 : F_1 = \ldots = F_T,$$
vs. the alternative that $H_0$ does not hold. We are especially interested in detecting departures from $H_0$ in which the means $\mu_{1g}, \ldots, \mu_{Tg}$ tend to change in a monotonic manner across the ordered levels of the treatment factor for one or more genes $g \in \{1, \ldots, G\}$. Thus, we wish to develop testing procedures with enhanced power for detecting departures from $H_0$ when a gene set contains genes whose mean expression levels change monotonically with treatment.

3.1.3 Previous Studies

There is a long history of research on univariate testing problems involving order restricted alternatives. Some of the earliest work closely related to the problems we consider was conducted by Bartholomew (1959a, 1959b, 1961a, 1961b). With the assumption that random variables follow univariate normal distributions, Bartholomew successively derived the likelihood ratio tests (LRTs) for homogeneity of means against monotonic alternatives (1959a), against two-sided monotonic alternatives (1959b), against more general order restrictive alternatives with known variances (1961a), and against order restrictive alternatives with unknown variances (1961b). The large body of subsequent related work is summarized in Robertson, Wright, and Dykstra (1988) with more recent developments chronicled by Silvapulle and Sen (2005). Little work has been done to connect order restricted inference to modern genomics applications. One exception is the work of Peddada et al. (2003) who proposed a gene clustering algorithm using time-course or dose-response gene expression profiles.

Although order restricted inference has not played a role, considerable work has been conducted on the problem of identifying gene sets whose joint expression distributions differ across treatments. For the most part, these methods can be grouped into two categories. Methods in one category begin by computing a statistic value for each individual gene; then the gene-specific values of the test statistic are combined for the genes in a gene set to determine the significance of the gene set. Methods belonging to this category include the GSEA by Subramanian et al. (2005), SAFE by Barry et al. (2005), and GSA by Efron and Tibshirani (2007).

Methods in the other category use a global and multivariate approach to gene set testing. Tests for differences in multivariate expression distributions across different experimental
conditions are used to identify gene sets of interest. Such methods include the Global Test by Goeman (2004), global ANCOVA by Mansmann and Meister (2005), PLAGE by Tomfohr et al. (2005), DEA by Liu et al. (2007), the Multiresponse Permutation Procedure (MRPP) described by Mielke and Berry (2001) and used in gene set testing by Nettleton et al. (2008), the HMM approach incorporating directed acyclic graphs (DAG) by Liang and Nettleton (2010), the Shrinkage Approach to Gene-set Analysis (SAGA) by Parks et al. (2011), and the multiple testing considered by Sohn (2011), among others.

In the remainder of this paper, we develop methods of both types that use order restricted inference to identify gene sets whose gene expression levels tend to change monotonically with a quantitative or ordinal treatment factor.

3.1.4 Organization

In Section 3.2, we propose four tests for detecting multivariate gene sets comprised of genes with expression levels that change monotonically as treatment levels increase. We refer to these tests as the Index Weighted Sum Permutation Test (IWSPT), the Linear Regression Permutation Test (LRPT), the Isotonic Estimator Permutation Test (IEPT) and the Isotonic Likelihood Ratio Permutation Test (ILRPT). We also provide the asymptotic null distribution of the isotonic likelihood ratio test statistic for the case of normally distributed data, based on Bartholomew’s work in late 1950’s and early 1960’s (Bartholomew, 1959a, 1959b, 1961a, 1961b). In Section 3.3, we introduce three tests that do not focus on monotonic alternatives. We carry out a simulation study to compare our proposed methods to the other three methods in Section 3.4. In Section 3.5, we apply these procedures to analyze data reported by Thomas et al. (2007). Results are discussed in Section 3.6.

3.2 Proposed Methods

3.2.1 Method I – Index Weighted Sum Permutation Test (IWSPT)

The first test we propose is based on pairwise Euclidean distances between gene expression vectors from different treatment groups. If the alternative hypothesis is true, the observations from within any treatment group should be, relatively speaking, close to each other and farther from the observations of a different treatment group. Moreover, under
the alternatives of greatest interest, the larger the difference between the indices \( |i - i'| \),
the farther the distance between the observations from these groups is expected to be. Let
\( \Lambda(i, i') \) be the average of all the Euclidean distances between pairs of data vectors between
group \( i \) and group \( i' \), i.e.,
\[
\Lambda(i, i') = \frac{1}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{i'}} \|Y_{ij} - Y_{i'k}\|.
\]
(3.1)

Our proposed test is a weighted sum of all \( \Lambda(i, i') \) values given by,
\[
S_I = \sum_{i=1}^{T-1} \sum_{i'=i+1}^{T} w(i, i') \Lambda(i, i'),
\]
(3.2)
where \( w(i, i') \) is a non-decreasing function of \( |i - i'| \). Thus, the contribution of \( \Lambda(i, i') \) to
\( S_I \) is more heavily weighted when \( |i - i'| \) is large. Throughout this paper, we simply use
\( w(i, i') = |i - i'| \). However, when the \( i^{th} \) treatment group is associated with a quantitative
value \( x_i \) (\( i = 1, \ldots, T \)), other choices like \( w(i, i') = |x_i - x_{i'}| \) are simple to implement. For
all such choices of \( w(i, i') \), the null hypothesis is rejected if \( S_I \) is sufficiently large.

A permutation test is used to obtain a p-value as follows. The treatment labels are
permuted \( M \) times relative to the gene set expression vectors, and the test statistic value
for the \( m^{th} \) permutation is calculated and denoted \( S_{Im} \) (\( m = 1, \ldots, M \)). The permutation
p-value is the proportion of the \( M + 1 \) test statistic values (the statistic computed for the
original data and the \( M \) permutations) no less than the test statistic value observed for the
original data: \( p = \sum_{m=1}^{M} I(S_{Im} \geq S_I) / M + 1 \). Although the total number of permutations is \( N! \),
these \( N! \) permutations give rise to \( \prod_{i=1}^{N} n_i! \) distinct test statistic values. When \( \prod_{i=1}^{N} n_i! \) is
too large for the available time and computing resources, a random sample of the \( \prod_{i=1}^{N} n_i! \)
relevant permutation test statistic values may be selected and combined with the observed
statistic to form a valid permutation reference distribution.

It is known that different genes have different variances. The genes with larger variances
may have dominant effects over the genes with smaller variances. In order to adjust the
heterogeneity of variance among genes in a gene set, a Euclidean commensuration approach
was proposed by Mielke and Berry (2001) and utilized by Nettleton et al. (2008) in mul-
tivariate gene set testing. For the problem we consider in this paper, each observation of
gene \( g \), \( Y_{ijg} \), is standardized with
\[
\left\{\sum_{i=1}^{T} \sum_{i'=i}^{T} \sum_{j=1}^{n_i} \sum_{k=j}^{n_{i'}} (Y_{ijg} - Y_{i'kg})^2\right\}^{-\frac{1}{2}}, \quad g = 1, \ldots, G,
\]
so that data from each gene have the same sample variance after the standardization. Throughout this paper, we assume all the genes within a gene set have a same variance, or have been standardized using Euclidean commensuration.

3.2.2 Methods Based on Combining P-Values from Individual Genes within a Gene Set

In Sections 3.2.2.1 through 3.2.2.3, we introduce three testing procedures that begin by computing a test statistic for each individual gene within a gene set. The gene-specific test statistic differs for each method, but the way the test statistic values are used to draw a conclusion about the entire gene set is the same for all three methods. We now describe this approach in general terms before introducing the specific test statistics in Section 3.2.2.1 through 3.2.2.3.

First, the observed test statistics for the $G$ genes in the gene set (denoted by $S_1, \ldots, S_G$) are converted to permutation p-values. As described in Section 2.1, the treatment labels are randomly permuted relative to the gene expression vectors $M$ times. Let $S_{gm}$ denote the value of the test statistic computed for gene $g$ and permutation $m$. This yields $G \times (M + 1)$ values that can be organized in the matrix

\[
\begin{pmatrix}
S_1 & S_{11} & \cdots & S_{1M} \\
S_2 & S_{21} & \cdots & S_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
S_G & S_{G1} & \cdots & S_{GM}
\end{pmatrix}.
\]

The first column of this matrix contains the test statistics calculated using the original data, and the $m + 1$st column $(S_{1m}, \ldots, S_{Gm})$ contains the test statistics calculated using the $m$th permutation, $m = 1, \ldots, M$. Then the permutation p-values for the original data and the permutation data also construct a $G \times (M + 1)$ matrix

\[
\begin{pmatrix}
p_1 & p_{11} & \cdots & p_{1M} \\
p_2 & p_{21} & \cdots & p_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
p_G & p_{G1} & \cdots & p_{GM}
\end{pmatrix},
\]

where the first column are the permutation p-values for gene $g$ using the original data, and the $m + 1$st column contains the corresponding permutation p-values for the $m$th permuta-
If the null hypothesis is rejected with large statistic values, then \( p_g = \frac{\sum_{m=1}^{M} 1(S_{gm} \geq S_g) + 1}{M + 1} \)
with \( g = 1, \ldots, G \), and \( p_{gm} = \frac{\sum_{m=1}^{M} 1(S_{gm} \geq S_{gm}) + 1(S_g \geq S_{gm})}{M + 1} \), \( g = 1, \ldots, G \), \( m = 1, \ldots, M \). If the null hypothesis is rejected with small statistic values, then the direction of the sign is changed from “\( \geq \)” to “\( \leq \)” in the above two formulas.

For each column of the p-value matrix, we combine the \( G \) elements into one statistic. There are many different ways to combine p-values: the maximum, the minimum, the mean or the median of the \( G \) p-values, etc. Here we use Fisher’s combined probability method (1925):
\[
X^2 = -2 \sum_{g=1}^{G} \ln(p_g)
\]
for the original data and
\[
X^2_m = -2 \sum_{g=1}^{G} \ln(p_{gm})
\]
for the permutation data, \( m = 1, \ldots, M \). For a level \( \alpha = 0.05 \) test, the null hypothesis is rejected if
\[
X^2_m > X_{0.05}^2
\]
for each column of the p-value matrix, we combine the \( G \) elements into one statistic.

In the following three testing methods described in subsections 3.2.2.1 to 3.2.2.3, we first compute a test statistic for each gene in a gene set and then combine the result across genes within a gene set. To simplify notations in these subsections, for gene \( g \) \((g = 1, \ldots, G)\), we use \( Y_{ij} \) instead of \( Y_{ijg} \) to denote the gene expression value for the \( j \)th replication under the \( i \)th treatment.

### 3.2.2.1 Method II – Linear Regression Permutation Test (LRPT)

The second method is motivated by simple linear regression models. Consider a simple linear regression for each gene in the gene set as follows:
\[
Y_{ij} = \beta_0 + \beta_1 i + \epsilon_{ij} \quad i = 1, \ldots, T, \quad j = 1, \ldots, n_i,
\]
where the \( \epsilon_{ij} \) terms are assumed to be independent, zero-mean random variables with constant variance. Let \( |t_1| = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \) denote the absolute value of the t-statistic for testing \( \beta_1 = 0 \) vs. \( \beta_1 \neq 0 \), where \( \hat{\beta}_1 = \frac{\sum_{i=1}^{T} i \cdot n_i \cdot Y_i - (\sum_{i=1}^{T} i \cdot n_i) \cdot \bar{Y}}{\sum_{i=1}^{T} i \cdot n_i - \frac{(\sum_{i=1}^{T} i \cdot n_i)^2}{\sum_{i=1}^{T} n_i^2}} \), and
\[
s_{\hat{\beta}_1} = \left( \frac{1}{\sum_{i=1}^{T} n_i} \cdot \sum_{i=1}^{T} (Y_{ij} - \bar{Y})^2 \right)^{\frac{1}{2}}.\]
If the alternative hypothesis of interest is true for this gene, \( |t_1| \) will tend to be large. Similar testing ideas can be found in Armitage (1955), and Bartholomew (1961a). The Linear Regression Permutation Test statistic is defined as \( S_{II} = |t_1| \). The null hypothesis is rejected with large \( S_{II} \) and the permutation p-value is
\[
p = \frac{\sum_{m=1}^{M} 1(|t_{1m}| \geq |t_1|) + 1}{M + 1},
\]
where \( |t_{1m}| \) is the test statistic calculated using the \( m \)th permuted data (\( m = 1, \ldots, M \)).
Similar to the IWSPT statistic, using the treatment indices in the LRPT statistic \( S_{II} \) makes it applicable in general cases in which only the treatment indices are available to use. In the calculation of \( S_{II} \), it is straightforward to use the real values of the treatment levels as the regressor variable instead of using the treatment indices for the cases where the treatment factor is numerical.

### 3.2.2.2 Method III – Isotonic Estimator Permutation Test (IEPT)

The basic idea of IEPT is as follows. If the alternative hypothesis is true, the distance between the mean gene expression under the first treatment index and the mean gene expression under the last treatment index should be the larger than the distances between the mean gene expressions for any other pair of treatments. This idea is motivated by Peddada et al.’s (2003) method for clustering genes using order-restricted inference.

Suppose \( \hat{\mu}^{(I)} = (\hat{\mu}_1^{(I)}, \ldots, \hat{\mu}_T^{(I)})' \) and \( \hat{\mu}^{(D)} = (\hat{\mu}_1^{(D)}, \ldots, \hat{\mu}_T^{(D)})' \) are the projections of \( (\hat{Y}_1, \ldots, \hat{Y}_T) \) onto the increasing order restricted profile \( C_I = \{ \mu \in \mathbb{R}^T : \mu_1 \leq \mu_2 \leq \cdots \leq \mu_T \} \) and the decreasing order restricted profile \( C_D = \{ \mu \in \mathbb{R}^T : \mu_1 \geq \mu_2 \geq \cdots \geq \mu_T \} \), respectively, using isotonic estimation. Our proposed test statistic for gene \( g \) is defined as

\[
S_{III} = \ell_\infty = \max\{\ell_\infty^{(I)}, \ell_\infty^{(D)}\},
\]

where \( \ell_\infty^{(I)} = |\hat{\mu}_1^{(I)} - \hat{\mu}_T^{(I)}| \) and \( \ell_\infty^{(D)} = |\hat{\mu}_1^{(D)} - \hat{\mu}_T^{(D)}| \). The null hypothesis is rejected for large \( \ell_\infty \) values. The permutation p-values is

\[
p = \frac{\sum_{m=1}^M \mathbb{1}(\ell_m^{\infty} \geq \ell_\infty)+1}{M+1},
\]

where \( \ell_m^{\infty} \) is the test statistic calculated from the \( m \)th permutation \( (m = 1, \ldots, M) \).

### 3.2.2.3 Method IV – Isotonic Likelihood Ratio Permutation Test (ILRPT)

Consider \( \hat{\mu}^{(I)} = (\hat{\mu}_1^{(I)}, \hat{\mu}_2^{(I)}, \ldots, \hat{\mu}_T^{(I)})' \) and \( \hat{\mu}^{(D)} = (\hat{\mu}_1^{(D)}, \hat{\mu}_2^{(D)}, \ldots, \hat{\mu}_T^{(D)})' \) defined in Section 3.2.2.2. The corresponding sums of squares are: \( SS_{Total} = \sum_{i=1}^T \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_.)^2 \), \( SSE_I = \sum_{i=1}^T \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_1^{(I)})^2 \), \( SSE_D = \sum_{i=1}^T \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_1^{(D)})^2 \), \( SST_1 = \sum_{i=1}^T n_i (\hat{\mu}_i^{(I)} - \bar{Y}_.)^2 \), \( SST_D = \sum_{i=1}^T n_i (\hat{\mu}_i^{(D)} - \bar{Y}_.)^2 \). Define the Isotonic Likelihood Ratio Permutation Test (ILRPT) statistic as \( S_{IV} = \hat{B}^* = \max\{\hat{B}^{(I)}, \hat{B}^{(D)}\} \), where \( \hat{B}^{(I)} = \frac{SST_I}{SS_{Total}} \) and \( \hat{B}^{(D)} = \frac{SST_D}{SS_{Total}} \).

Permute the treatment labels relative to gene set expression vectors for \( M \) times, and the test statistic using the \( m \)th (\( m = 1, \ldots, M \)) permuted data \( \hat{B}_{m}^* \) is also calculated. The null
hypothesis is rejected when the statistic value is large enough. The permutation p-values for ILRPT is \( p = \frac{\sum_{m=1}^{M} I(B^*_m \geq B^*) + 1}{M + 1} \).

Consider the assumption that gene expressions follow normal distributions:

\[
Y_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, T, \quad j = 1, \ldots, n_i,
\]

with independence among all \( N = \sum_{i=1}^{T} n_i \) random variables. Also, denote the parameter space as

\[
\theta = (\mu_1, \mu_2, \ldots, \mu_T, \sigma^2)' \in \Theta = \mathbb{R}^T \times \mathbb{R}^+.
\]

Under this assumption, \( \hat{\mu}^I \) is the MLE of \( \mu \) under restriction \( \mu_1 \leq \mu_2 \leq \ldots \leq \mu_T \), and \( \hat{\mu}^D \) is the MLE of \( \mu \) under restriction \( \mu_1 \geq \mu_2 \geq \ldots \geq \mu_T \). Also, rejecting the null hypothesis with \( \bar{B}^* \) large enough is equivalent to rejecting the null hypothesis with the LRT statistic small enough for testing the hypotheses

\[
H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1 = (\Theta_I \cup \Theta_D) \setminus \Theta_0,
\]

where

\[
\Theta_0 = \{ \theta \in \Theta : \mu_1 = \mu_2 = \ldots = \mu_T \},
\]

\[
\Theta_I = \{ \theta \in \Theta : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_T \} = C_I \times \mathbb{R}^+,
\]

and

\[
\Theta_D = \{ \theta \in \Theta : \mu_1 \geq \mu_2 \geq \ldots \geq \mu_T \} = C_D \times \mathbb{R}^+.
\]

The derivation is provided in Section 3.7.1 of the Appendix.

Moreover, with the normal distribution assumption in (3.5), \( \bar{B}^I \) and \( \bar{B}^D \) follow weighted sums of Beta distributions (Bartholomew, 1961b) under the null hypothesis, i.e.,

\[
P(\bar{B}^I \geq c) = P(\bar{B}^D \geq c) = \sum_{q=2}^{T} w(q,T) P(B_{(q-1)/2}(N-q)/2 \geq c),
\]

with weights \( w(q,T) \), \( q = 2, \ldots, T \). We present the derivation in Section 3.7.2 of the Appendix.

For the “two-sided” order restricted testing problem discussed in this paper, Bartholomew (1959b) proposed an approximation method to obtain the critical value:

\[
P(\max\{\bar{B}^I, \bar{B}^D\} \geq C_\alpha) \approx 2P(\bar{B}^I \geq C_\alpha) = 2P(\bar{B}^D \geq C_\alpha).
\]

So the critical value \( C_\alpha \) is approximated by the upper \( \frac{\alpha}{2} \) quantile of the distribution of \( \bar{B}^I \) or \( \bar{B}^D \). For balanced data, Table 1 in Bartholomew’s paper (1959a) gives the exact
values of \( w(q,T) \) for \( q = 2, \ldots, T \) for \( T \leq 5 \). For unbalanced data or \( T > 5 \), \( w(q,T) \) can be approximated using Plackett’s (1954) formula. With modern computers, we are able to obtain the distribution of \( \bar{B}^{(I)} \) (or \( \bar{B}^{(D)} \)) and \( \bar{B}^* \) directly using simulation.

### 3.3 Other Methods that Do Not Focus on Monotonic Alternatives

In this section, we introduce three methods for testing the hypotheses

\[
H_0 : F_1 = F_2 = \ldots = F_T \quad \text{vs.} \quad H_1 : \text{not all } F_i \text{'s are equal,}
\]

that do not focus on monotonic trends in gene expression. Again, we use a permutation procedure to obtain p-values.

#### 3.3.1 Multiresponse Permutation Procedure (MRPP)

The first test we present is the Multiresponse Permutation Procedure (MRPP) introduced by Mielke and Berry (2001). The basic idea of the MRPP is as follows. If the alternative hypothesis is true, i.e., gene expression vectors under different treatments follow different distributions, then the observation vectors from the same treatment group will tend to be relatively close together and be farther away from the observation vectors from a different treatment group. So randomly permuting labels relative to the \( N \) gene set expression vectors is likely to produce greater average pairwise within-group distances than obtained with the original treatment labels. The MRPP based on this idea is to reject the null hypothesis for sufficiently small mean pairwise within-group distances. The test statistic can be written as

\[
\bar{D} = \frac{\sum_{i=1}^{T} n_i D_i}{N},
\]

where

\[
D_i = \frac{1}{n_i(n_i - 1)/2} \sum_{j_1=1}^{n_i-1} \sum_{j_2=j_1+1}^{n_i} \| Y_{ij_1} - Y_{ij_2} \|
\]

is the average pairwise within-group distance.

The treatment labels are permuted \( M \) times and the \( m^{th} \) permutation test statistic value is calculated \( \bar{D}_m \) \((m = 1, \ldots, M)\). The permutation p-value is the proportion of the permutation statistic values no larger than the observed statistic value:

\[
p = \frac{\sum_{m=1}^{M} \mathbb{1}(D_m \leq D) + 1}{M + 1}.
\]
3.3.2 Energy Statistic (ES)

Another method is based on the concept of an “energy statistic” proposed by Zech and Aslan (2005). This test was originally applied in multivariate two sample tests. The statistic is defined as

\[
\Phi_{n_1n_2} = \frac{1}{n_1^2} \sum_{j=1}^{n_1-1} \sum_{j'=j+1}^{n_1} R(\|Y_{1j} - Y_{1j'}\|) + \frac{1}{n_2^2} \sum_{j=1}^{n_2-1} \sum_{j'=j+1}^{n_2} R(\|Y_{2j} - Y_{2j'}\|)
\]

\[
- \frac{1}{n_1n_2} \sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} R(\|Y_{1j} - Y_{2j'}\|),
\]

(3.10)

where \(n_1\) and \(n_2\) are the sample sizes for the two groups, \(R(\cdot)\) is a continuous, monotonically decreasing function. \(R(r) = -\ln(r)\) is suggested by the authors. The three terms in the formula refer to the energies for the two groups and the interaction energy between them.

When the treatment number is greater than 2, the statistic is generalized as follows:

\[
E = \sum_{i=1}^{T} \frac{1}{n_i^2} \sum_{j=1}^{n_i-1} \sum_{j'=j+1}^{n_i} R(\|Y_{ij} - Y_{ij'}\|)
\]

\[
- \sum_{i>i'}^{n_i} \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_i} R(\|Y_{ij} - Y_{i'j'}\|), \quad T \geq 2.
\]

(3.11)

P-values are obtained using a permutation procedure. Gene sets with large statistic values will be identified as having significant differences between treatment groups. So for a procedure with \(M\) permutations, the permutation p-value is

\[
p = \sum_{m=1}^{M} I(E_m \geq E) + 1 \over M + 1,
\]

where \(E_m\) is the test statistic calculated from the \(m^{th}\) \((m = 1, \ldots, M)\) permutation.

3.3.3 Permutation Based ANOVA Test (PANOVA)

One-way ANOVA provides a good default test in testing overall homogeneity of means for one-dimensional data satisfying the assumption of normality. When the data fails to satisfy the assumption of normality and the sample size is too small, the F null distribution for the test statistic may not be reliable. In this case, we conduct a permutation test for each gene using the ANOVA F-statistic as the test statistic. For multi-dimensional data analyses, such as the gene set detections in this paper, we follow the procedure discussed in Section 2.2, so as to use Fisher’s (1925) probability combination method to obtain the permutation p-value for the whole gene set.
3.4 Simulation Studies

In this section we use three simulation studies to compare the four tests that consider monotonic alternatives and the three methods that do not focus on monotonic alternatives.

3.4.1 The First Simulation Study

In the first study, we generate data from multivariate normal distributions with dimensions = 6 and 20 under: 1. the null case for which the simulated gene expression distribution remains the same across different treatments; 2. the alternative for which the mean values increase linearly, exponentially or logarithmically as a function of treatment level in each dimension, which is the first monotonic case; 3. another alternative case for which the mean values increase in an arc-tangential manner in all dimensions, which is the second monotonic case; and 4. the case for which the mean values change with different treatment levels but do not change monotonically as the treatment levels increase. We use the first-order autoregressive structure as the correlation structure for all the simulated gene expressions within a independent simulation replication. The number of treatment levels is $T = 5$, the number of replicates under each treatment level is $n_i = 3$ ($i = 1, \ldots, 5$). A total of 1,000 independent simulation replications are simulated under each simulation setting. We do tests using our proposed four test statistics, compared with the MRPP, the Energy Statistic (ES), and the PANOVA. We did 1,000 permutations including the original data to obtain one p-value. A gene set is identified if its permutation p-value is smaller than 0.05. Because we focus on single gene set detections, we do not discuss challenges in multiple testing that arise when many gene sets are tested simultaneously. Recent studies in this area include Liang and Nettleton (2010) and Sohn et al. (2011). Estimated detection rates from this simulation study using 6 genes in a gene set are shown in Table 3.1. Estimated detection rates from this simulation study using 20 genes in a gene set are shown in Table 3.2.

<table>
<thead>
<tr>
<th>1000 gene sets in each case</th>
<th>ES</th>
<th>PANova</th>
<th>MRPP</th>
<th>IWSP</th>
<th>LRPT</th>
<th>IEPT</th>
<th>ILRPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>38</td>
<td>42</td>
<td>39</td>
<td>45</td>
<td>53</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td>Monotonic case 1</td>
<td>725</td>
<td>732</td>
<td>720</td>
<td>931</td>
<td>962</td>
<td>923</td>
<td>946</td>
</tr>
<tr>
<td>Monotonic case 2</td>
<td>723</td>
<td>735</td>
<td>721</td>
<td>924</td>
<td>946</td>
<td>921</td>
<td>940</td>
</tr>
<tr>
<td>Not Monotonic</td>
<td>723</td>
<td>709</td>
<td>721</td>
<td>11</td>
<td>3</td>
<td>26</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 3.1: Monotonic gene set detections using different statistics with normal random variables. Each gene set contains 6 genes.
From Table 3.1, we can see that under both monotonic cases, the proposed test statistics have more identifications than the other three methods when the gene expressions change monotonically as the treatment level increases. In both cases where gene expressions change monotonically, all the four proposed tests have identification rates over 90%, while the three existing tests have identification rates around 79%. Among all the tests, LRPT and ILRPT have the best power. When gene expression changes across the treatment labels but not in a monotonic pattern, ES, PANOVA and MRPP have identification rates similar to the two monotonic cases, while none of the four proposed test statistics reject more than 4% of the gene sets. We can see that our proposed tests focus detection power on monotonically changing expression patterns, which is a sub-space of the complement of the null parameter space, while the other three methods spread detection power over the complement of the null parameter space. Thus the proposed tests have more detections for gene sets that contain genes whose expression levels tend to change monotonically across treatments than other methods.

From Table 3.2, we can see that when the number of genes in a gene set increases to 20, all the testing methods have an increased number of identifications in both monotonic cases. The proposed IWSPT, IRPT and IEPT identify all of the gene sets with monotonic changing patterns. The ILRPT identifies all the gene sets with monotonic changing pattern 1, and 999 gene sets of the total 1,000 gene sets with monotonic changing pattern 2. Although other tests also identify most of the gene sets in the two monotonic cases, their numbers of identifications are still significantly less than the proposed tests. In the case that the gene expression means are different across treatment levels but not monotonically related to the treatment index, most of the gene sets are not rejected by the proposed tests. Actually, the numbers of gene sets rejected by the proposed tests are even slightly less than the number of gene sets rejected in the non-monotonic case when the dimension is 6. On the contrary,
the numbers of rejections by the tests that do not focus on monotonic change patterns are similar to the number of rejections in both monotonic cases.

### 3.4.2 The Second Simulation Study

In the second simulation study, we compare ILRPT with the test using test statistic $\bar{B}^*$ and the critical value from Bartholomew’s approximation, and the test using $\bar{B}^*$ and the critical value approximated by simulation. We focus exclusively on univariate analysis for a single gene and consider the case where $T = 5$ and $n_i = 3$ for $i = 1, \ldots, 5$.

For the Bartholomew’s critical value approximation, we use the upper 2.5\% percentile of the null distribution of $\bar{B}(I)$ (or $\bar{B}(D)$) as an approximation of the upper 5\% percentile of the null distribution of $\bar{B}^* = \max\{\bar{B}(I), \bar{B}(D)\}$. By Bartholomew (1959a, 1961b),

$$P_{H_0}(\bar{B}(I) \geq c) = P_{H_0}(\bar{B}(D) \geq c) = \sum_{q=2}^{5} w(q,5)P\left(B_{(q-1)/2,(15-q)/2} \geq c\right),$$

where $w(2,5) = \frac{50}{120}$, $w(3,5) = \frac{35}{120}$, $w(4,5) = \frac{10}{120}$, $w(5,5) = \frac{1}{120}$. Thus we obtain the critical value $C_1 = 0.41962$.

For the critical value directly obtained by simulation, we simulate the empirical distribution of $\bar{B}^*$ under the null hypothesis using normal data. The upper 5\% percentile of the empirical distribution is an approximation of the exact critical value. We simulated 3,000,000 null data set to obtain the critical value $C_2 = 0.4198589$. We can see that with $\alpha = 0.05$, the two critical value approximations are actually very close.

We compare ILRPT with tests using $C_1$ and $C_2$ based on a simulation involving 10,000 normally distributed data sets for each of four cases. The results are presented in Table 3.3. From Table 3.3, we can see that all four proposed test statistics show larger powers and better specificities relative to the methods that do not focus power on ordered alternatives. The ILRPT results are similar to the results for which the two critical values are obtained by approximations.

### 3.4.3 The Third Simulation Study

In the third simulation study, we replicate the second simulation study except that normal distributions are replaced with gamma distributions. The results using the gamma distribution with $\alpha$-level 0.05 are shown in Table 3.4. We found that the tests that do not
10000 genes in each case | ES | PANova | MRPP | IWSP | LRPT | IEPT | ILRPT | $C_1$ | $C_2$
---|---|---|---|---|---|---|---|---|---
Null | 452 | 454 | 466 | 470 | 521 | 490 | 498 | 490 | 487
Monotonic case 1 | 6508 | 6805 | 6525 | 8984 | 9388 | 8816 | 9069 | 9077 | 9076
Monotonic case 2 | 6485 | 6831 | 6521 | 8761 | 8975 | 8625 | 8916 | 8916 | 8933
Not Monotonic | 6454 | 6819 | 6471 | 119 | 54 | 358 | 406 | 418 | 417

Table 3.3: Monotonic Gene Detections Using Different Statistics with Normal Random Variables

focus power on monotonic patterns have low identification rates in the second monotonic case, while the proposed tests have good identification rates for both cases in which gene expression changes monotonically as the treatment level increases. Moreover, the proposed tests have fewer rejections in the case for which changes in gene expression are not monotonic. Also, the results using critical values $C_1$ and $C_2$ are very similar to the result of ILRPT.

10000 genes in each case | ES | PANova | MRPP | IWSP | LRPT | IEPT | ILRPT | $C_1$ | $C_2$
---|---|---|---|---|---|---|---|---|---
Null | 530 | 509 | 511 | 440 | 516 | 520 | 507 | 503 | 500
Monotonic case 1 | 5075 | 4588 | 5076 | 7215 | 7472 | 6850 | 7023 | 7064 | 7059
Monotonic case 2 | 3581 | 3781 | 3564 | 5926 | 6472 | 5922 | 6202 | 6228 | 6221
Not Monotonic | 8379 | 7553 | 8478 | 21 | 27 | 20 | 22 | 14 | 14

Table 3.4: Monotonic Gene Detections Using Different Statistics using Gamma Random Variables

### 3.5 An Example with Real Data Analysis

In this section, we apply all the methods introduced in this paper to analyze mouse gene expression data from Thomas et al. (2007). The aim is to detect gene sets that have expressions that monotonically change with exposure to increasing formaldehyde, with level of concentration 0 (control group, exposed to filtered air), 0.7, 2.0, 6.0 and 15.0 ppm. The experiment used 8 mice for the control group and 4 mice for each non-zero formaldehyde concentration. After exposure for 6 hours, the mice were sacrificed, and the gene expression values were obtained (with log-2 transformation) and stored in Biotechnology Information Gene Expression Omnibus. The information of Gene Ontology annotation is obtained by using `rat2302.db` and `annotate` R Bioconductor packages. We use the four proposed testing procedures to identify multivariate monotonic gene expressions, and compare the results with the three tests that do not focus on monotonic changes. We identify gene sets
with permutation p-values smaller than 0.05 as significant.

The gene sets identified by all of the MRPP, ES and PANova tests but none of the proposed tests are shown in Figure 3.1 and Figure 3.2. Each panel is for genes from a single gene set. The labels 1 through 5 on the horizontal axis in each panel denote the five formaldehyde concentrations, i.e., the five treatment levels. The standardized log-2 mean gene expression levels for each individual gene are connected by line segments. We can see that the gene sets shown in Figure 3.1 and 3.2 have different mean gene expression levels across the treatments, but the changing patterns are not monotonic. One example is Gene Set 1 shown in the upper left panel of Figure 3.1, which has an “up – down – up” pattern. When the formaldehyde concentration changes from 0 ppm to 0.7 ppm, the mean gene expression levels of genes in Gene Set 1 tend to increase. This increase is followed by a decreasing change when the concentration increases from 0.7 ppm to 6.0 ppm. Gene expression tends to increase again when the formaldehyde concentration increases to 15.0 ppm. Some other gene sets have a “down – up – down” pattern, such as Gene Sets 4, 10 and 13. There are also some gene sets with more than one gene expression trend. Some examples are Gene Set 5 and 9, which have an “up – down – up” pattern and a “down – up – down” pattern in the same gene set. None of these gene sets have significant monotonic gene expression patterns.

Figure 3.3 through Figure 3.5 show five examples of the gene sets detected by all four of the proposed tests. The gene sets are the same from Figure 3.3 to Figure 3.5. The genes in red color in Figure 3.3 have permutation p-values less than 0.05 when doing single gene analysis using LRPT, and the genes in red color in Figure 3.4 and Figure 3.5 have permutation p-values less than 0.05 when doing single gene analysis using IEPT and ILRPT, respectively. In Gene Set D, it is clear that all the genes in red color have a predominantly decreasing pattern. Other gene sets, for example Gene Set C and Gene Set E, have some genes with a decreasing pattern and other genes with an increasing pattern. We can see that differences among Figures 3.3 through 3.5 are very slight.

3.6 Discussion and Conclusions

In this paper, we consider the case where an experimental treatment factor is numerical or ordinal and the response variable is a measure of gene expression. We propose four tests:
Figure 3.1: Gene Sets Identified by All the Old Tests But None of the New Tests

the Index Weighted Sum Permutation Test (IWSPT), the Linear Regression Permutation Test (LRPT), the Isotonic Estimator Permutation Test (IEPT) and the Isotonic Likelihood Ratio Permutation Test (ILRPT), to detect multivariate gene sets that contain genes with monotonic changes in means when increasing the treatment level. Other multivariate tests, such as the Multiresponse Permutation Procedure (MRPP), the Energy Statistic (ES) and the Permutation Based ANOVA Test (PANOVA), spread power over the entire complement of the null parameter space, and do not focus more power on the monotonic subspace.

By simulation, we demonstrate that the proposed tests have higher detection powers when the gene expression levels in multivariate gene sets change monotonically in mean values, compared to tests that do not focus on monotonic changes. In the application of all the testing methods to a real data set, we found that the gene sets detected by all of the proposed tests contain genes with monotonic changes in their mean gene expression levels, while the gene sets detected by the other three tests have gene expression patterns that are not monotonic.
Figure 3.2: Gene Sets Identified by All the Old Tests But None of the New Tests

3.7 Appendix

3.7.1 Derivation of the Isotonic Likelihood Ratio Test Statistic

Let \( Y_i = (Y_{i1}, \ldots, Y_{im})' \), \( i = 1, \ldots, T \). The ratio of the supremum of likelihoods under the null hypothesis parameter space and the whole parameter space is

\[
\Lambda = \frac{\sup_{\theta \in \Theta_0} \prod_{i=1}^T f(Y_i; \theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} \prod_{i=1}^T f(Y_i; \theta)} \cdot \min \left\{ \frac{e^{-N \frac{2 \pi \text{SS}_{\text{Total}}}}}{N^2}, \frac{e^{-N \frac{2 \pi \text{SS}_{\text{I}}}}}{N^2} \right\}.
\]

Thus, the LRT statistic is

\[
-2 \log \Lambda = -2 \left\{ \frac{N}{2} \log \min \{\text{SSE}_I, \text{SSE}_D\} - \frac{N}{2} \log \text{SS}_{\text{Total}} \right\}
\]

\[
= -N \min \left\{ \frac{\text{SSE}_I}{\text{SS}_{\text{Total}}}, \frac{\text{SSE}_D}{\text{SS}_{\text{Total}}} \right\}.
\]

3.7.2 The null distribution of \( \bar{B}^{(I)} \)

In this section, we review the derivation of the null distribution of \( \bar{B}^{(I)} \), derived by Bartholomew (1959a, 1961b). The null distribution of \( \bar{B}^{(D)} \) is the same as \( \bar{B}^{(I)} \).
Figure 3.3: Examples for the Gene Sets Identified by All the Proposed Tests (Genes with permutation p-values smaller than 0.05 in LRPT single gene data analysis are shown in red color.)

First, we obtain the formulas of $SST_I$ and $SSE_I$. For group means $\{\bar{Y}_1, \ldots, \bar{Y}_T\}$, there is a unique subset $S = \{0 = i_0 < i_1 < \ldots < i_q = T\} \subset \{0, 1, \ldots, T\}$, such that

$$\bar{Y}_{[i_1,i_1]} = \bar{Y}_{[i_0+1,i_1]} < \bar{Y}_{[i_1,i_2]} < \bar{Y}_{[i_2+1,i_3]} < \ldots < \bar{Y}_{[i_{q-1}+1,i_q]} = \bar{Y}_{[i_{q-1}+1,T]},$$

where $\bar{Y}_{[i_r-1+1,i_r]} = (\sum_{i=i_r-1+1}^{i_r} n_i \bar{Y}_i) / (\sum_{i=i_r-1+1}^{i_r} n_i)$ is the mean of the gene expressions from group $i_r-1+1$ to group $i_r$. The number of samples from group $i_r-1+1$ to group $i_r$ is denoted as $n_{[i_r-1+1,i_r]} = \sum_{i=i_r-1+1}^{i_r} n_i$. Also, $\bar{Y}_{[i_r-1+1,i_r]}$ is no less than any of the group means with index smaller than $i_r-1+1$, and is no larger than any of the group means with index larger than $i_r$, i.e.,

$$\max\{\bar{Y}_1, \ldots, \bar{Y}_{i_r-1}\} \leq \bar{Y}_{[i_r-1+1,i_r]} \leq \min\{\bar{Y}_{i_r+1}, \ldots, \bar{Y}_T\}.$$

So $\mu^{(f)}_i = \bar{Y}_{[i_r-1+1,i_r]}$, $i_r-1+1 \leq i \leq i_r$, $r = 0, \ldots, q$, and

$$SST_I = \sum_{r=1}^{q} n_{[i_r-1+1,i_r]} (\bar{Y}_{[i_r-1+1,i_r]} - \bar{Y}_c)^2,$$

$$SSE_I = \sum_{r=1}^{q} \sum_{i=i_r-1+1}^{i_r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{[i_r-1+1,i_r]})^2.$$

For example, if $\bar{Y}_1 < \ldots < \bar{Y}_T$, then $q = T$, $i_r = i$ ($r = 1, \ldots, T$). So $\mu^{(f)}_i = \bar{Y}_i$, $SST_I = SST$ and $SSE_I = SSE$. For another example, if $\bar{Y}_1 \geq \ldots \geq \bar{Y}_T$, then $q = 1$, so $\mu^{(f)}_1 = \ldots = \mu^{(f)}_T = \bar{Y}_c$, $SST_I = 0$. 
Once the formulas for $SST_I$ and $SSE_I$ are obtained, we begin to derive the asymptotic null distributions for $SST_I$ and $SSE_I$. According to the statement above, given group means $\bar{Y}_1, \ldots, \bar{Y}_T$, there is a unique subset $S = \{i_0, i_1, \ldots, i_{q-1}, i_q\}$. Based on that, the distribution of $SSE_I$ is

$$P(SSE_I \geq c \mid \bar{Y}_{[1,i_1]} \leq \ldots \leq \bar{Y}_{[q-1+1,T]})$$

$$= P\left( \sum_{r=1}^{q} \sum_{i=i_{r-1}+1}^{i_r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{[i_{r-1}+1,i_r]})^2 \geq c \mid \bar{Y}_{[1,i_1]} \leq \ldots \leq \bar{Y}_{[q-1,T]} \right).$$

Since the within group sum of square $\sum_{r=1}^{q} \sum_{i=i_{r-1}+1}^{i_r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{[i_{r-1}+1,i_r]})^2$ is irrelevant with the order of group means $\bar{Y}_{[1,i_1]} \leq \ldots \leq \bar{Y}_{[q-1,T]}$, the conditional probability is equal to the unconditional probability, i.e.,

$$P\left( \sum_{r=1}^{q} \sum_{i=i_{r-1}+1}^{i_r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{[i_{r-1}+1,i_r]})^2 \geq c \mid \bar{Y}_{[1,i_1]} \leq \ldots \leq \bar{Y}_{[q-1,T]} \right)$$

$$= P\left( \sum_{r=1}^{q} \sum_{i=i_{r-1}+1}^{i_r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{[i_{r-1}+1,i_r]})^2 \geq c \right)$$

$$= P(\chi^2_{N-q} \geq c).$$

So $SSE_I$ is a $\chi^2$ random variable with degree of freedom $N - q$.

For the distribution of $P(SST_I \geq c \mid \bar{Y}_{[1,i_1]} \leq \ldots \leq \bar{Y}_{[q-1,T]})$, consider the following
Figure 3.5: Examples for the Gene Sets Identified by All the Proposed Tests (Genes with permutation p-values smaller than 0.05 in ILRPT single gene data analysis are shown in red color.)

\[
\sum_{i=1}^{T} W_i (Z_i - \bar{Z})^2 = \sum_{i<j} W_i W_j (Z_i - Z_j)^2 / \sum_{i=1}^{T} W_i,
\]

where \( \bar{Z} = \frac{\sum_{i=1}^{T} W_i Z_i}{\sum_{i=1}^{T} W_i} \). So we have

\[
P(SST_i \geq c \mid \bar{Y}_{[1,i_1]} \leq \cdots \leq \bar{Y}_{[q-1+1,T]}) \]
\[
\begin{align*}
= & \ P \left( \sum_{r=1}^{q} n_{[r-1+1,i_r]} (\bar{Y}_{[r-1+1,i_r]} - \bar{Y}_{[i,1,i_1]})^2 \geq c \mid \bar{Y}_{[1,i_1]} \leq \cdots \leq \bar{Y}_{[q-1,T]} \right) \\
= & \ P \left( \sum_{1 \leq s < r \leq q} n_{[r-1+1,i_r]} n_{[s-1+1,i_s]} (\bar{Y}_{[r-1+1,i_r]} - \bar{Y}_{[s-1+1,i_s]})^2 / \sum_{r=1}^{q} n_{[r-1+1,i_r]} \geq c \mid \right. \\
& \left. \bar{Y}_{[1,i_1]} \leq \cdots \leq \bar{Y}_{[q-1,T]} \right).
\]

Because \((\bar{Y}_{[r-1+1,i_r]} - \bar{Y}_{[s-1+1,i_s]})^2\) is the square of a normal random variable with mean 0,
which has distribution irrelevant with the order of \( \bar{Y}_{[1,i]} \) through \( \bar{Y}_{[q-1,T]} \), we have

\[
P\left( \sum_{1 \leq s < r \leq q} n_{[i_{r-1}+1,i_{r}]} n_{[i_{s-1}+1,i_{s}]} (\bar{Y}_{[i_{r-1}+1,i_{r}]} - \bar{Y}_{[i_{s-1}+1,i_{s}]])^2 / \sum_{r=1}^{q} n_{[i_{r-1}+1,i_{r}]} \geq c \middle| \bar{Y}_{[1,i]} \leq \ldots \leq \bar{Y}_{[q-1,T]} \right)
\]

\[
= P\left( \sum_{1 \leq s < r \leq q} n_{[i_{r-1}+1,i_{r}]} n_{[i_{s-1}+1,i_{s}]} (\bar{Y}_{[i_{r-1}+1,i_{r}]} - \bar{Y}_{[i_{s-1}+1,i_{s}]})^2 / \sum_{r=1}^{q} n_{[i_{r-1}+1,i_{r}]} \geq c \right)
\]

\[
= P\left( \sum_{r=1}^{q} n_{[i_{r-1}+1,i_{r}]} (\bar{Y}_{[i_{r-1}+1,i_{r}]} - \bar{Y}_{..})^2 \geq c \right)
\]

\[
= P(\chi_{q-1}^2 \geq c).
\]

Consequently, for a given \( q \), the conditional distribution of \( SST_I \) is \( \chi_{q-1}^2 \). By Bartholomew (1959a), the conditional distribution of \( \bar{B}^{(I)} \) is Beta with parameters \((q-1)/2\) and \((N-q)/2\), i.e., \( B_{(q-1)/2,(N-q)/2} \). Due to the different orders among \( \bar{Y}_1, \ldots, \bar{Y}_T \), \( q \) goes from 1 to \( T \), with probability \( P(\#S = q + 1) = w(q,T) \), \( q = 1, \ldots, T \). So we have \( P(\bar{B}^{(I)} \geq c) = \sum_{q=2}^{T} w(q,T) P(B_{(q-1)/2,(N-q)/2} \geq c) \). By the same reasoning, \( P(\bar{B}^{(O)} \geq c) = \sum_{q=2}^{T} w(q,T) P(B_{(q-1)/2,(N-q)/2} \geq c) \).
References


CHAPTER 4. Copy Number Variation Detection Using Next Generation Sequencing Read Counts

Abstract

A copy number variation (CNV) is a difference between genotypes in the number of copies of a genomic region. Next generation sequencing (NGS) technologies provide sensitive and accurate tools for detecting genomic variations that include CNVs. We propose a new methodology for detecting CNVs using NGS data. This method (henceforth denoted by m-HMM) is based on a hidden Markov model with emission probabilities that are governed by mixture distributions. We use the Expectation-Maximization (EM) algorithm to estimate the parameters in the model. A simulation study demonstrates that our proposed m-HMM approach improves upon existing methods. We apply the m-HMM method to NGS data from the two maize inbred lines B73 and Mo17 to identify CNVs that may play a role in creating phenotypic differences between these inbred lines. We show that the results of our m-HMM analysis are concordant with previous array-based efforts to identify CNVs.

KEY WORDS: Count data; Gamma-Poisson mixture; Hidden Markov model; Plant genomics; Poisson mixture model.

4.1 Introduction

A copy number variation (CNV) is a variation between genomes in the number of copies of a genomic region that is 1,000 DNA bases (1Kb) or larger (Banerjee et al., 2011). CNVs are also known as structural variations (SVs) because they affect relatively large regions in a DNA molecule. Structural genomic duplications or deletions correspond to copy number gains or losses, respectively. CNVs play an important role in human hereditary illnesses (Gokcumen et al., 2009) and in plant breeding and agricultural improvement (Schnable et al., 2009).
Maize exhibits extensive variation in both genotype and phenotype relative to the variation seen in humans (Buckler et al., 2006). The genotypic diversity in maize species permits a variety of uses, such as human and animal food and fuel. The maize genotype B73 was sequenced in 2009 (Schnable et al., 2009). This accomplishment allows a further comparison and understanding in different types of maize. Swanson-Wagner et al. (2009) and Belo et al. (2010) compared a variety of maize inbreds with B73 using array comparative genomic hybridization (aCGH) and identified a considerable number of CNVs along the genome. Springer et al. (2009) also analyzed the structural variance between the two maize genotypes B73 and Mo17 using aCGH.

Array comparative genomic hybridization, first proposed in 1997 (Solinas-Toldo et al., 1997), has served as a robust and effective approach for CNV screening (Feuk, 2006). Statistical methods for analyzing aCGH data are readily available and are described in review articles such as Wineinger et al. (2008) and Medvedev et al. (2009). However, aCGH is expensive and has limited resolution and accuracy. Nowadays, rapidly developing next generation sequencing (NGS) technologies provide a sensitive and accurate alternative approach for accessing genomic variations. The quality, speed, and affordability give NGS a significant advantage over microarrays (Hurd et al., 2009; Su et al., 2011).

Despite the advantages of NGS over aCGH, the use of NGS for CNV identification has been limited by a lack of available and effective statistical approaches. The well-developed aCGH data analysis methods cannot take the full advantage of NGS data, and thus, new statistical analysis methods for NGS data are needed. Most of the existing methods for CNV detection using NGS data can be classified into two categories: sliding window methods and Hidden Markov model (HMM) methods. Sliding window methods include Segseq by Chiang et al. (2009), Event-wise testing by Yoon et al. (2009), rSW-seq by Kim et al. (2010) and JointSLM by Magi et al. (2011), among others. This category of methods must simultaneously deal with a large number of tests of significance, and the results of such methods are highly dependent on the determination of critical values. One example of the methods using HMMs is CNAseg by Ivakhno et al. (2010), which requires multiple sequencing samples for each of the genotypes in the comparison.

In this paper, we propose a new CNV change point detection methodology for use with NGS data originating from two different genomes. To understand the data and our model
for the data, it is necessary to introduce some NGS data collection details and terminology. First of all, we use reference genome to describe the genome of the genotype that has been fully sequenced using whole-genome sequencing technologies. In contrast, the target genome is the genome of a genotype of interest that has not been fully sequenced. The goal is to use NGS data from the reference and target genotypes to identify regions of copy number variation between the reference and target genomes. We say that a genomic region in the target genotype where the number of copies is amplified relative to the reference genotype has a copy number gain. A target genomic region present but at a reduced copy number relative to the reference genome, is said to have a copy number loss. A region that is present in the reference genome but absent in the target genome is described as absent. These three states (copy number gain, copy number loss, and absent) represent copy number variations in the target genome relative to the reference. A region with no difference in the number of copies between the target and the reference genotypes is said to be normal in state. A genomic location where there is a change from one copy number state to another is called a copy number change point.

To identify copy number change points and copy number states, thousands of genomes from both the target and the reference genotypes are obtained. The DNA strands are fragmented into 100 to 1,000 base segments, and a sample of these segments is obtained separately for each genotype. At one end of every sampled segment, a sequence of 30 to 70 bases is determined and recorded. Such a sequence of bases is called a read. Each of the reads is then aligned to the reference genome to determine its origin in the genome. The location of the first base of the read on the reference genome is recorded as the position of the read. The numbers of reads for the target genome and the reference genome are recorded as the target read counts and the reference read counts. If a location has a positive target read count or a positive reference read count, it is called a site. Thus, data from NGS technologies are small integer counts with associated site positions on the reference genome.

In this paper, we proposed a statistical methodology, involving a Hidden Markov model with mixture emission distributions (m-HMM), to identify and classify regions of CNV between a target and a reference genome. The proposed m-HMM makes a segmentation according to the reference genomic locations and identifies regions in the target genome
where the copy numbers are different from the reference genome. A formal description of
the data set to be analyzed in presented in Section 4.2. In Section 4.3, we introduce our
proposed m-HMM in full detail. The computational algorithms and analysis details are
explained in Sections 4.4 and 4.5. In Section 4.6, our proposed m-HMM is compared with
other methods using a simulation study. The application of the m-HMM is demonstrated
in Section 4.7 with an analysis of two maize inbred lines: B73 and Mo17, where B73 is the
reference genome and Mo17 is the target genome. Conclusions are provided in Section 4.8.

4.2 The Data Set in this Study

The data to be analyzed can be described as follows. Suppose $o_i^{[c]}$ and $o_i^{[t]}$ are the
observed read counts for the reference genome B73 and the target genome Mo17 at genomic
position $b_i$, $i = 1, \ldots, I$. The read counts take small non-negative integer values, including
a large number of zeros. Because it is difficult to carry out accurate modeling and inference
using such data, it is common to work with sums of counts rather than the original individual
counts.

A common way to group individual counts is to define windows with a specific width and
calculate the sum of target and reference read counts within each window, so as to obtain
shorter series of larger target and reference read counts. Kim et al. (2010) defined windows
using a fixed number of read counts in the reference genome. Chiang et al. (2009), Xie at
al. (2009) and Ivakhno et al. (2010) defined windows using a fixed genomic distance. These
methods have an underlying assumption that the sites within a window share the same
copy number state. Such an assumption may be reasonable because a CNV is a somewhat
rare type of genomic mutation, and the closer any two sites are located on the genome, the
less likely there is a CNV change point between these sites. For example, if a site is in a
genomic region that has been duplicated in the target genome relative to the reference, it
is likely that a nearby site falls within that same region of duplication. However there are
also some problems in implementing these methods. Sites are randomly located along a
genomic sequence, with a high density in some parts of the genome and a low density in
other parts of the genome. Rigidly defining windows with a fixed number of read counts
has the potential to put sites physically far away with each other into one window, which
increases the risk of including copy number change points in a window. Rigidly defining
the windows with a fixed genomic distance can produce high variation in the number of sites and in read counts across windows. This can lead to decreased accuracy for identifying CNVs.

In this paper, we propose a new grouping method, which uses K-means clustering on the physical site positions to group the sites into windows. For each chromosome, we first divide the chromosome into $M + 1$ parts, by defining breakpoints that correspond to the $M$ largest distances between adjacent sites. In practice, we use $M = 20$. Then we perform the K-means clustering for each of the $M + 1$ parts, where $K$ is chosen for any particular part as the number of reference genomic sites in the part divided by 40. Thus, each cluster will have 40 reference genome sites on average. Finally, we obtain $W$ windows, where each window is defined by a collection of sites in one cluster. For $w = 1, \ldots, W$, let $g_w$ denote the set of indices corresponding to sites in window $w$. The target and reference read count for window $w$ are the sum of the target read counts and the sum of the reference read counts within that window: $u^t_w = \sum_{i \in g_w} o^t_i$, $u^r_w = \sum_{i \in g_w} o^r_i$. We use the median position of sites within a window as the position for that window. Using this method, the sites that are closest together have a larger chance to be grouped into a single window, which results in a more reasonable grouping than previously used approaches.

4.3 Mixture-Hidden Markov Model (m-HMM)

In this section, we describe the proposed mixture-hidden Markov model (m-HMM) that we use to estimate the copy number change points along the genome. This CNV detecting methodology is carried out separately on each chromosome. In the hidden Markov model, we assume that the unobserved copy number states for the windows along the genome follow a Markov chain, with some unknown transition probability matrix that specifies the conditional probabilities of transitioning to each state given the current state. Given each copy number state, the target and the reference read counts are generated according to a distribution known as the emission distribution.

4.3.1 Transition Probability

For each window $w$ ($w \in \{1, \ldots, W\}$), we wish to predict one of the four hidden states $S_w \in \{1, 2, 3, 4\}$, where
1 = gain: copy number gain / amplification in the target relative to the reference,
2 = normal: no difference in copy number between the target and the reference,
3 = loss: region present in the target genome but at reduced copy number relative to the reference,
4 = absent: region absent in the sample but present in the reference.

For the copy number states $k$ and $l$, the transition probability $a_{kl} \ (k, l = 1, \ldots, 4)$ is defined as, the conditional probability of the next window $w + 1$ taking copy number state $l$, given the copy number state $k$ for the current window $w$. Motivated by Marioni et al. (2006), we define the transition matrix for each window $w$ as a function of $w$ given by $A_w = [a_{kl}(w)]_{4 \times 4}$, where

$$
a_{kl}(w) = P(S_{w+1} = l | S_w = k, \theta) =
\begin{cases}
p_{kl}(1 - e^{-\rho d_w}) & l \neq k \\
1 - (\sum_{j \neq k} p_{kj})(1 - e^{-\rho d_w}) & l = k
\end{cases},
$$

for $k, l = 1, \ldots, 4$ and $w = 1, \ldots, W - 1$. Here $d_w$ denotes the physical distance on the reference genome between the location of window $w$ and the location of window $w + 1$. The parameter $p_{kl}$ affects the transition probabilities from state $k$ to state $l$, regardless of the distances between adjacent windows. We constrain $p_{kl} \in (0, 1)$ and $\sum_{l \neq k} p_{kl} < 1$ for each $k$. The parameter $\rho$ is a positive-valued parameter that determines the effect of distance on the transition matrix. The larger the value of $\rho$, the larger the impact of distance on the transition probabilities. $\theta$ represents all the parameters in the model, including $p_{kl} \ (k, l = 1, \ldots, 4; l \neq k)$, $\rho$ and all the parameters in the emission distribution. The advantage of this transition matrix is that it takes the relative positions of windows into consideration.

As the distance between adjacent windows increases, the transition probability from state $k$ to state $l \neq k$ increases and approaches $p_{kl}$ as $d_w \to \infty \ (k, l = 1, \ldots, 4; l \neq k)$. On the other hand, as the distance between adjacent windows decreases, the probability of a difference in copy number states between windows diminishes, i.e., $A_w \to I_{4 \times 4}$ as $d_w \to 0$.

### 4.3.2 The Emission Distributions

The emission distributions define emission probabilities, which are the conditional probabilities of reference and target read counts, given the copy number state of the window.
Each window \( w \) has two observations, a reference aligned read count \( U_w^{[r]} = u_w^{[r]} \) and a target aligned read count \( U_w^{[t]} = u_w^{[t]} \). We model the reference read count using a Poisson distribution with mean \( \lambda_w^{[r]} \), i.e., \( U_w^{[r]} | \lambda_w^{[r]} \sim \text{Poisson}(\lambda_w^{[r]}) \), where \( \lambda_w^{[r]} \) follows a Gamma distribution with parameters \( \alpha \) and \( \beta \).

For the target aligned read count, one natural choice for the emission distribution is

\[
U_w^{[t]} | (\lambda_w^{[r]}, S_w = k) \sim \text{Poisson}(K_k c_0 \lambda_w^{[r]}),
\]

(4.2)

where \( K_1 = 2, K_2 = 1, K_3 = 0.5, K_4 = 0, \) and \( c_0 \) is a normalization factor that can account for any discrepancy between the total number on reference and target reads in normal regions. However in real data sets, background noise and mis-alignments along the genomic sequence data are inevitable. For example, within a normal genomic segment, we often see some windows with target and reference read count ratios significantly higher or lower than 1; within a segment of copy number gain (or loss), we also find windows that have target and reference read count ratios significantly lower than 2 (or higher than 0.5). The original HMM introduced above will not only identify real copy number variation signals, but will also pick local variations caused by random error. In simulation studies, modeling with the emission distributions as defined above leads to more state changes than true CNV signals.

The problem of identifying too many CNV change points is also pointed out by Ivakhno et al. (2010). To address this problem, Ivakhno et al.’s CNAseg employs a merging adjustment procedure on the outcomes of the original HMM segmentations using Pearson’s \( \chi^2 \) statistics. However, CNAseg segmentation depends heavily of the determination on the merging threshold. In the method we propose, instead of using (4.2) to model the target aligned read count distribution, we use a Poisson mixture model for each of the four copy number states \( k = 1, 2, 3, 4 \):

\[
U_w^{[t]} | (\lambda_w^{[r]}, S_w = k) \sim \sum_{j=1}^{4} q_{kj} \text{Poisson}(v_{kj} c_0 \lambda_w^{[r]}),
\]

(4.3)

where \( Q = [q_{kj}]_{4 \times 4} = \begin{pmatrix}
q_{11} & 1 - q_{11} & 0 & 0 \\
1 - q_{22} & q_{22} & 1 - q_{22} & 0 \\
0 & 1 - q_{33} & q_{33} & 1 - q_{33} \\
0 & 0 & 1 - q_{44} & q_{44}
\end{pmatrix} \) with \( q_{kk} \in (0.5, 1) \) for \( k = \)
As a consequence, we multiply the

\[ v_{k,j} \]

\( j = 1, \ldots, 4 \), and 

\[ V = [v_{k,j}]_{4 \times 4} = \begin{pmatrix} 2 & v_{12} & 0 & 0 \\ v_{21} & 1 & v_{23} & 0 \\ 0 & v_{32} & 0.5 & v_{34} \\ 0 & 0 & v_{43} & 0 \end{pmatrix} \]

with \( v_{12}, v_{21}, v_{23}, v_{32} \in (0.5, 1) \), \( v_{34}, v_{43} \in (0, 0.5) \). The parameters \( q_{k,j} (k, j = 1, \ldots, 4) \) are the weights for the Poisson components in the mixture distribution, \( v_{kk} (k = 1, \ldots, 4) \) denote the effects of the copy number variations, and \( v_{kj} (k, j = 1, \ldots, 4; k \neq j) \) denote the effects of local variations due to random error.

Based on the model as specified so far, the joint distribution for the target and the reference read counts at window \( w \), conditional on the hidden state for window \( w \) being \( k \), is

\[
P(s) (U_w^t = u_w^t, U_w^r = u_w^r | S_w = k, \theta) = \frac{\sum_j q_{kj} \Gamma(u_w^t + u_w^r + \alpha) (v_{kj}c_0) u_w^t \beta^\alpha}{\Gamma(\alpha) u_w^t! u_w^r! (v_{kj}c_0 + 1 + \beta)^{u_w^t + u_w^r + \alpha}}
\]

\[
= \sum_j q_{kj} NB_{U_w^t} (\alpha, \frac{1}{\beta + 1}) NB_{U_w^r} (\alpha, \frac{v_{kj}c_0}{v_{kj}c_0 + 1 + \beta + 1})
\]

(4.4)

with \( u_w^t \in Z \setminus Z^- \) and \( u_w^r \in Z \setminus Z^- \). Here \( NB_{y,\xi} (\eta, \xi) = \frac{\Gamma(y+\eta)}{\eta^y (\xi^\eta)} \frac{\Gamma(\frac{1}{\xi+\eta})^\xi}{(\xi+\eta)^y} \) is the probability mass function of the negative binomial distribution with parameters \( \eta \) and \( \xi \). Note that

\[
P(s) (u_w^t = 0, u_w^r = 0 | S_w = k, \theta) = \sum_j q_{kj} \left( \frac{\beta}{v_{kj}c_0 + 1 + \beta} \right)^\alpha,
\]

which is greater than 0. We know that it is actually not possible for a window with the target and the reference read counts to both be zero because sites with both zero reference and target read counts are not part of the data set. Thus, the joint distribution of \( u_w^t \) and \( u_w^r \) is truncated in the sense that

\[
P(u_w^t = u_w^r = 0 | \theta) = 0.
\]

As a consequence, we multiply the \( j^{\text{th}} \) component of (4.4) by a constant \( \frac{(v_{kj}c_0 + 1 + \beta)^{\alpha}}{(v_{kj}c_0 + 1 + \beta)^{\alpha} - \beta^\alpha} \)

(\( j = 1, 2, 3, 4 \)), and obtain the joint probability for the target and reference reads at window \( w \) given \( S_w = k \):

\[
P(u_w | S_w = k, \theta) = P(u_w | S_w = k, \alpha, \beta, c_0, q, v) = \sum_j q_{kj} \frac{(v_{kj}c_0)^{u_w^t} \beta^\alpha}{u_w^t! u_w^r! \Gamma(\alpha) (v_{kj}c_0 + 1 + \beta)^{u_w^t + u_w^r} ((v_{kj}c_0 + 1 + \beta)^\alpha - \beta^\alpha)}
\]

(4.5)
where \( u_w = \begin{pmatrix} u_w^t \\ u_w^r \end{pmatrix} \in (\mathbb{Z} \setminus \mathbb{Z}^{-}) \times (\mathbb{Z} \setminus \mathbb{Z}^{-}) \setminus \{ (0,0) \} \).

### 4.4 Parameter Estimation Using the EM Algorithm

We use the Expectation-Maximization (EM) algorithm to estimate the parameters in our model, \( \theta = \{ p, \rho, \alpha, \beta, c_0, q, v \} \), where \( p \) and \( \rho \) are transition probability parameters, and \( \alpha, \beta, c_0, q = (q_{11}, q_{22}, q_{33}, q_{44})' \) and \( v = (v_{12}, v_{21}, v_{23}, v_{32}, v_{34}, v_{43})' \) are emission probability parameters. The aim is to obtain the value of \( \theta \) that maximizes the likelihood \( L(\theta | u) \) where \( u = (u_1', \ldots, u_W')' \).

#### 4.4.1 Characterizing the E and M Steps

The observed data likelihood function is

\[
L_{\text{obs}}(\theta | u) = \sum_{s} \pi_{s_1} \prod_{w=1}^{W-1} a_{s_ws_{w+1}}(w) \prod_{w=1}^{W} P(u_w | s, \alpha, \beta, c_0, q, v). \tag{4.6}
\]

\( s = (s_1, \ldots, s_W)' \in \{1, 2, 3, 4\}^W \) is a vector of unobserved states for all the windows along the chromosome, and \( \pi_{s_1} = P(S_1 = s_1) \). If the hidden states are known, the complete data likelihood function can be written as

\[
L_{\text{comp}}(\theta | u, s) = \pi_{s_1} \prod_{w=1}^{W-1} a_{s_ws_{w+1}}(w) \prod_{w=1}^{W} P(u_w | s, \alpha, \beta, c_0, q, v). \tag{4.7}
\]

In these two likelihoods, \( \pi_{s_1} \prod_{w=1}^{W-1} a_{s_ws_{w+1}}(w) \) is the probability of that the hidden states for all the windows on the chromosome are \( s_1, \ldots, s_W \), respectively, and \( \prod_{w=1}^{W} P(u_w | s, \alpha, \beta, c_0, q, v) \) is the probability of the observed read counts \( u \), given the hidden states \( s_1, \ldots, s_W \).

Given the parameter estimates \( \hat{\theta}^{(m)} \) from the iteration \( m \) of the EM algorithm, the E-step is to evaluate \( E_{S|u, \hat{\theta}^{(m)}}(\ell(\theta | u, s)) \) the expectation of the complete data log-likelihood with respect to the conditional distribution of the hidden states \( S \) given the observed data \( u \), with \( \theta = \hat{\theta}^{(m)} \). The M-step is to find \( \hat{\theta}^{(m+1)} \), the value of \( \theta \) that maximizes \( E_{S|u, \hat{\theta}^{(m)}}(\ell(\theta | u, s)) \). A detailed look at both the E and M steps is provided in Section 4.9.1 of the Appendix.
4.4.2 Initialization, Convergence, and Prediction of Hidden States

The initial values for all the parameters are defined as follows. In the transition probabilities, we define the initial values $p^{(0)} = (p_{12}, \ldots, p_{43})' = 0.1 \cdot 1_{1 \times 9}$ and $\rho^{(0)} = 0.5$. For the parameters in the emission probabilities, we define $q_{11}^{(0)} = q_{22}^{(0)} = q_{33}^{(0)} = q_{44}^{(0)} = 0.5$, $v_{12}^{(0)} = v_{21}^{(0)} = 1.5$, $v_{23}^{(0)} = v_{32}^{(0)} = 0.75$, $v_{34}^{(0)} = v_{43}^{(0)} = 0.25$, and use the maximum likelihood estimates of $\alpha$ and $\beta$ with all sites assigned with “normal” state (state 2) as the initial values $\alpha^{(0)}$ and $\beta^{(0)}$.

To obtain the initial value for $c_0$, we proceed as follows. First we calculate the ratio between the target aligned read count and the reference aligned read count for all the windows where both $u_t[w]$ and $u_r[w]$ are positive through the whole genome across different chromosomes. Then we classify these ratios into three groups using K-means clustering and get three group means denoted $M_1$, $M_2$ and $M_3$. Suppose $M_1 < M_2 < M_3$, then $x_t^{(0)}$ is the total target aligned reads located in the windows belonging to the cluster with mean $M_2$, and $x_r^{(0)}$ is the corresponding total control aligned reads located in the windows belonging to the cluster with mean $M_2$. After obtaining $x_t^{(0)}$ and $x_r^{(0)}$, we calculate $c_0^{(0)}$ using $\frac{x_t^{(0)}}{x_r^{(0)}}$.

The probabilities of the four hidden states for the first window $\pi_{s_1}$ ($s_1 = 1, 2, 3, 4$ and $\sum_{s_1=1}^4 \pi_{s_1} = 1$) are determined by the target and the reference read counts in the first window. We calculate

$$R_1 = \frac{u_t^{[t]} x_t^{(0)}}{u_r^{[r]} x_r^{(0)}}.$$  

If $R_1 > 1.5$, then $\pi_1 = 1$, and we assign copy number gain as the initial copy number state of the first window; if $0.75 < R_1 \leq 1.5$, then $\pi_2 = 1$, and we assign normal as the initial copy number state to the first window; if $0 < R_1 \leq 0.75$, then $\pi_3 = 1$, and we assign copy number loss as the initial state to the first window; if $R_1 = 0$, then $\pi_4 = 1$, and we assign absent as the initial state to the first window.

For all the other windows $S_w$ ($w \geq 2$), we use normal state (state 2) as the initial state. All the parameters are updated using the EM algorithm described in Section 4.1. The iteration stops if the difference between $\hat{\theta}^{(m)}$ and $\hat{\theta}^{(m+1)}$ is very small, and we obtain the estimate $\hat{\theta} = \hat{\theta}^{(m+1)}$ of the parameter $\theta$.

After convergence, the conditional probability of each of the four states $P_k(w) = P(S_w = k|\mathbf{u}, \hat{\theta})$ ($k = 1, 2, 3, 4$) is calculated, and the conditional hidden state prediction for window
w is given by the state that has the largest conditional probability, i.e., \( \hat{S}_w = \arg \max_k \hat{P}_k(w) \), \( w = 1, \ldots, W \).

### 4.5 Adjustments on the Segmentation Result

Aligned sites are grouped into windows at the beginning, because we wish to have larger read counts to better capture the copy number variation signals in the m-HMM. We initially assume that the copy number state remains the same within a window. In reality, it is possible that the copy number change points happen within windows. Figure 4.1 demonstrates the relationship between the estimated changes point before adjustment and the real change point. Suppose window \( w_b \) is the location identified by the algorithm from where the copy number state change from state 1 to state 2 with genomic sites grouped into windows. We have \( g_{w_b} = \{i_{II}, i_{II}+1, \ldots, i_{III}-1\} \), so \( i_{II} \) is the first site in window \( w_b \), which is initially identified as a change point. Window \( w_b-1 \) and \( w_b+1 \) are the two windows next to window \( w_b \). With \( g_{w_b-1} = \{i_I, i_{I}+1, \ldots, i_{II}-1\} \) and \( g_{w_b+1} = \{i_{III}, i_{III}+1, \ldots, i_{IV}-1\} \), \( i_I \) and \( i_{III} \) are the first sites of \( w_b-1 \) and \( w_b+1 \), respectively. The real change point \( i_{(real)} \) may happen between sites \( i_I \) and \( i_{III} - 1 \). In order to obtain a more accurate result, the following algorithm makes the adjustment.

1. For each site \( i \) between site \( i_I \) and site \( i_{III} - 1 \), we obtain the total target and reference counts from \( i_I \) to \( i-1 \), denoted as \( z^{[t]}_{i(I)} \) and \( z^{[r]}_{i(I)} \); also obtain the total target and reference counts from \( i \) to \( i_{III} - 1 \), denoted as \( z^{[t]}_{i(R)} \) and \( z^{[r]}_{i(R)} \).

2. Calculate the Pearson’s \( \chi^2 \) test statistic using \( \{z^{[t]}_{i(I)}, z^{[r]}_{i(I)}, z^{[t]}_{i(R)}, z^{[r]}_{i(R)}\} \).

3. Do step 1 and 2 for each candidate change point between \( i_I \) and \( i_{III} - 1 \) and the adjusted breakpoint \( i_{(adj)} \) is the one with the largest Pearson’s \( \chi^2 \) test statistic value.

### 4.6 Simulation Study and Method Comparison

To test the m-HMM methodology, we conducted a simulation study based on real DNA sequencing data from chromosome 4 of the lung cancer cell line NCI-H2347 from Chiang et al. (2009). Simulation based on real data can best maintain the characteristic of the data including noise and errors that exist in real data. We first randomly simulated the
genomic positions along the target and reference genomes, using a uniform distribution. Then we generated the reference and target genome read counts by shuffling the reference genome read counts in chromosome 4 of NCI-H2347. After that, we randomly picked 90 CNV segments on the simulated target genome. We considered three sizes for the CNV segments: 10 kb, 50 kb and 100 kb and generated 10 segments for each CNV type. We doubled the read counts for the segments with copy number gains, halved the read counts for the segments with copy number losses and set read counts to 0 for segments with no copies. In this simulation data set, we have 15668 sites with copy number gains (state 1), 13311 sites with copy number losses (state 3), 6971 sites with no copies (state 4), and 1813984 sites in normal regions (state 2). We compare the result of the m-HMM with mixture Poisson emission probability in (4.5), the result using the original HMM with Poisson emission probability in (4.2), and the result using Segseq by Chiang et al. (2009).

Table 4.1 is the comparison between the original HMM (red) and the proposed m-HMM (blue). The real number of sites in each states are listed in the last cell of each row. We can see that the m-HMM identifies 15926 sites with copy number gain state, among which there are 14375 correct identifications and 1551 false identifications. On the contrary, the original HMM identifies 54261 sites with copy number gain state, with 13528 correct identifications and 40733 false identifications. For copy number loss state and absent state, the m-HMM also identifies more sites with correct CNV states.

Among all the sites identified with CNVs by the m-HMM, 82% of them are correct identifications and 18% of them are false identifications. In contrast, among all the sites
identified with CNVs by the original HMM, 23% of them are correct identifications and 77% of them are false identifications. This means, the m-HMM is less affected by the noise and errors and can capture the true CNV signals better. Also, the m-HMM identifies over 90% of all the sites simulated with CNVs, comparative to 83% for the original HMM. Among all the sites in normal regions in the simulation data, the m-HMM has 99.6% identifications compared to 94.5% correct identifications of the original HMM. These results highlight the advantage of the m-HMM approach over the HMM approach.

Table 4.2 lists a comparison between our m-HMM approach (blue) and SegSeq (red). Segseq classifies the copy number states using only three categories: normal, copy number gain, and copy number loss, and it does not distinguish between the copy number loss state and the absent state. In addition, Segseq does not identify any sites with copy number gain state in this simulation study. Moreover, among the 4754 sites identified by Segseq with copy number loss or no copies, only 1532 sites are loss or absent sites. Also, Segseq failed to identify 18750 loss / absent sites.

<table>
<thead>
<tr>
<th># Sites</th>
<th>Gains</th>
<th>Normals</th>
<th>Losses</th>
<th>Absents</th>
<th>Real # Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gains</td>
<td>14375</td>
<td>1293</td>
<td>0</td>
<td>0</td>
<td>15668</td>
</tr>
<tr>
<td>Normals</td>
<td>1551</td>
<td>1807204</td>
<td>5157</td>
<td>72</td>
<td>1813984</td>
</tr>
<tr>
<td>Losses</td>
<td>0</td>
<td>1884</td>
<td>11427</td>
<td>0</td>
<td>13311</td>
</tr>
<tr>
<td>Absents</td>
<td>0</td>
<td>194</td>
<td>85</td>
<td>6692</td>
<td>6971</td>
</tr>
<tr>
<td>Gains</td>
<td>14375</td>
<td>1293</td>
<td>0</td>
<td>0</td>
<td>15668</td>
</tr>
<tr>
<td>Normals</td>
<td>1551</td>
<td>1807204</td>
<td>5157</td>
<td>72</td>
<td>1813984</td>
</tr>
<tr>
<td>Losses</td>
<td>0</td>
<td>1884</td>
<td>11427</td>
<td>0</td>
<td>13311</td>
</tr>
<tr>
<td>Absents</td>
<td>0</td>
<td>194</td>
<td>85</td>
<td>6692</td>
<td>6971</td>
</tr>
</tbody>
</table>

Table 4.1: Comparison between the m-HMM and the original HMM

<table>
<thead>
<tr>
<th># Sites</th>
<th>Gains</th>
<th>Normals</th>
<th>Losses or Absents</th>
<th>Real # Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gains</td>
<td>14375</td>
<td>1293</td>
<td>0</td>
<td>15668</td>
</tr>
<tr>
<td>Normals</td>
<td>1551</td>
<td>1807204</td>
<td>5229</td>
<td>1813984</td>
</tr>
<tr>
<td>Losses or Absents</td>
<td>0</td>
<td>2078</td>
<td>18204</td>
<td>20282</td>
</tr>
</tbody>
</table>

Table 4.2: Comparison between the m-HMM and the SegSeq
4.7 Application

We apply the proposed m-HMM to compare the sequence data from two maize genotypes: B73 and Mo17. We use the B73 genome as the reference sequence and the Mo17 genome as the target sequence. In the data preparation stage, both the B73 genome and the Mo17 genome are sampled with 44 bases segments and aligned to the reference B73 genome. We obtained 4.3 million aligned reads from B73 and 1.54 million aligned reads from Mo17, and we have 2.3 million genomic positions on the reference B73 genome with positive alignments. Using the m-HMM method, we found 1096 segments of 2000 bases or longer that have copy number variations that are at least 2 fold increasing/decreasing, among which, Mo17 has 14 segments with copy number gain state, 835 segments with copy number loss state and 247 segments with absent state, compared to the B73 inbred line.

The most significant copy number loss or absent state detected on Mo17 is in chromosome 6 (Figure 4.4), located from 42.2 million base to 46.2 million base on the reference genome. Within this segment, the reference genome B73 has 8152 total read counts, and the sample genome Mo17 has 542 total read counts. The copy number ratio is about 0.066, or 0.13 after taking into account the normalization factor $c_0$. Table 4.3 of the Appendix Section 4.9.2 shows the m-HMM segmentation result between 35.3 million bases and 57.0 million bases. Two other long segments with copy number loss or absent state in chromosome 6 are from 26.5 million to 28.8 million, and from 47.8 million to 49.7 million. These identifications are concordant with the result from other studies, in comparing DNA sequences between B73 and Mo17 using aCGH data (Springer et al. 2009; Belo et al. 2010).

There are several large segments with few or no copy number variations between Mo17 and B73. They are 121.3 Mb ~ 130 Mb on chromosome 1 (Figure 4.2), 69.2 Mb ~ 82.0 Mb and 84.8 Mb ~ 95.9 Mb on chromosome 3 (Figure 4.3) and 49.5 Mb ~ 61.5 Mb on chromosome 10 (Figure 4.5). These results are also concordant with previous results using aCGH data.

Table 4.4 in the Appendix Section 4.9.2 lists the detected CNV segments that are longer than 2 Mb and have CNV greater than 2-fold changes between B73 and Mo17.
4.8 Conclusion

In this paper, we proposed a new methodology to detect CNVs in the DNA sequences from two genotypes using next generation sequencing data. We used a hidden Markov model incorporating the mixture emission probability model to identify copy number variation change points. The simulations study suggests that the m-HMM has better sensitivities and specificities in CNV identifications compared to other methodologies.

The proposed m-HMM was applied to compare NGS data between the two maize inbred lines B73 and Mo17, and identified CNV change points of the target sequence Mo17 relative to the reference sequence B73. The result of the m-HMM is concordant with previous genomic studies using aCGH data by Springer et al. (2009) and Belo et al. (2010).

In addition, the m-HMM can be applied in comparing NGS of two genotypes when only one target genome sequence and one reference genome sequence are available, while many
other existing methods require multiple sequences in both target and reference genotypes. With less data and financial requirement, m-HMM has broader applications in genotype comparisons using NGS data.

4.9 Appendix

4.9.1 More Details in the EM Algorithm

The EM algorithm is an iterative procedure to find the parameter value \( \theta \) that maximizes the likelihood function \( \mathcal{L}(\theta|u) \). The observed data likelihood function is

\[
\mathcal{L}_{(\text{obs})}(\theta|u) = \sum_s \pi_{s_1} \prod_{w=1}^{W-1} a_{s_ws_{w+1}}(w) \prod_{w=1}^{W} P(u_w|s, \alpha, \beta, c_0, q, v).
\]
The complete data likelihood function is

\[
L^{(\text{comp})} = L(\theta | u, s) = \pi_s \prod_{w=1}^{W-1} a_{s_w s_{w+1}}(w) \prod_{w=1}^{W} P(u_w | s, \alpha, \beta, c_0, q, v).
\]

1. the E step:

The E step is to evaluate the expectation of the complete data log-likelihood with respect to the conditional distribution of the hidden states \(S\), given the observation \(u\) and current parameter value \(\hat{\theta}^{(m)}\): \(E_{S|u,\hat{\theta}^{(m)}}(\ell(\theta|u, s))\). The complete data log-likelihood function is

\[
\ell(\theta | u) = \log \pi_s + \sum_{w=1}^{W-1} \log (a_{s_w s_{w+1}}(w)) + \sum_{w=1}^{W} \log P(u_w | s, \alpha, \beta, c_0, q, v).
\]

The expectation with respect to the conditional probability of the hidden states, given the observations \(u\) and the parameters \(\hat{\theta}^{(m)}\) obtained from the \(m\)th iteration is

\[
E_{S|u,\hat{\theta}^{(m)}}(\ell(\theta|u, s))
= \sum_s P(S = s | u, \hat{\theta}^{(m)}) \log \pi_s + \sum_{w=1}^{W-1} \sum_s P(S = s | u, \hat{\theta}^{(m)}) \log (a_{s_w s_{w+1}}(w)) + \sum_{w=1}^{W} \sum_s P(S = s | u, \hat{\theta}^{(m)}) \log P(u_w | s, \alpha, \beta, c_0, q, v).
\]

We prove the last “=” as follows:

\[
(4.8) = \sum_s P(S = s | u, \theta) \log \pi_s
= \sum_s \sum_{k=1}^{4} P(S = s | u, \theta) I[S_1 = k] \log(\pi_k)
= \sum_{k=1}^{4} P_k(1) \log(\pi_k),
\]

where \(P_k(w) = P(S_w = k | u, \theta)\), and \(P_k(w) = P(S_w = k, S_{w+1} = l | u, \theta)\).
The numeric values of $P_{k} = \sum_{s} P(S = s | u, \theta) \sum_{w=1}^{W-1} \log (a_{w, s_{w+1}}(w))$

$$ = \sum_{s} P(S = s | u, \theta) \sum_{w=1}^{W-1} \sum_{k=1}^{4} \sum_{l=1}^{4} I_{S_{w}=k} I_{S_{w+1}=l} \log (a_{kl})$$

$$ = \sum_{w=1}^{W-1} \sum_{k=1}^{4} \sum_{l=1}^{4} P(S = s | u, \theta) I_{S_{w}=k} I_{S_{w+1}=l} I_{k \neq l} \log (a_{kl})$$

$$ = \sum_{w=1}^{W-1} \sum_{k=1}^{4} \sum_{l=1}^{4} P_{kl}(w) \log (a_{kl}).$$

The numeric values of $P_{k}(w) = \frac{P(S_{w}=k, u|\theta)}{P(u|\theta)}$ and $P_{kl}(w) = \frac{P(S_{w}=k, S_{w+1}=l | u|\theta)}{P(u|\theta)}$ can be evaluated using the forward-backward algorithm, which was introduced by Rabiner and Juang (1986). The forward probability $f_{k}(w)$ is defined as the probability of having state $k$ at time $w$, and having the observations $\{u_{1}, \ldots, u_{w}\}$ from window 1 to window $w$, given the parameter $\theta$, i.e., $f_{k}(w) = P(u_{1}, \ldots, u_{w}, S_{w} = k | \theta)$. The backward probability $b_{k}(w)$ is defined as the probability of having the observations $\{u_{w+1}, \ldots, u_{W}\}$ from window $w + 1$ to window $W$, given the state $k$ at window $w$, the observations from window 1 to window $w$, and the parameter $\theta$, i.e., $b_{k}(w) = P(u_{w+1}, \ldots, u_{W} | S_{w} = k, u_{1}, \ldots, u_{w}, \theta)$. The forward and backward probabilities can be obtained using recursions: $f_{k}(1) = \pi_{k} P(u_{1} | S_{1} = k, \theta)$, $b_{k}(W) = 1$, $f_{k}(w) = \sum_{l=1}^{4} f_{l}(w-1) a_{lk}(w-1) P(u_{w} | S_{w} = k, \theta)$ for $w = 2, \ldots, W$, and $b_{k}(w) = \sum_{l=1}^{4} a_{kl}(w) P(u_{w+1} | S_{w+1} = l, \theta) b_{l}(w+1)$ for $w = W-1, \ldots, 1$.

Consequently,

$$P(S_{w} = k, u|\theta) = P(u_{1}, \ldots, u_{W}, S_{w} = k | \theta) = P(u_{1}, \ldots, u_{w}, S_{w} = k | \theta) P(u_{w+1}, \ldots, u_{W} | S_{w} = k, u_{1}, \ldots, u_{w}, \theta) = f_{k}(w) b_{k}(w),$$
and

\[ P(S_w = k, S_{w+1} = l, u|\theta) = P(u_1, \ldots, u_w, S_w = k|\theta) \cdot P(u_{w+1}, S_{w+1} = l|S_w = k, u_1, \ldots, u_w, \theta) \]

\[ \cdot P(u_{w+2}, \ldots, u_W|S_w = k, S_{w+1} = l, u_1, \ldots, u_w+1, \theta) \]

\[ = f_k(w)P(u_{w+1}, S_{w+1} = l|S_w = k, \theta) b_l(w+1) \]

\[ = f_k(w) a_{kl}(w) P(u_{w+1}|S_{w+1} = l, \theta) b_l(w+1). \]

2. the M step:

The M step of EM algorithm is to find the value of \( \theta \), to make \( E_{S|u, \theta^{(m)}}(\ell(\theta|u, s)) \) obtain the maximum. And this optimum value is the updated parameter \( \theta^{(m+1)} \) for the \((m+1)\)th iteration.

\[ E_{S|u, \theta^{(m)}}(\ell(\theta|u, s)) \]

\[ = \sum_s P(S|u, \theta) \left[ \log \pi_{s_1} + \sum_{w=1}^{W-1} \log \left( a_{s_w s_{w+1}}(w) \right) + \sum_{w=1}^{W} \log P(u_w|s, \alpha, \beta, c_0, q, v) \right] \]

\[ = \sum_{k=1}^{4} P_k(1) \log(\pi_k) + \sum_{w=1}^{W-1} \sum_{k=1}^{4} \sum_{l=1}^{M} P_{kl}(w) \log(a_{kl}) \]

\[ + \sum_{t=1}^{T} \sum_{k=1}^{4} P_k(t) \log P(u_w|s, \alpha, \beta, c_0, q, v) \]

\[ = \sum_{k=1}^{4} P_k(1) \log(\pi_k) + \sum_{w=1}^{W-1} \sum_{k=1}^{4} P_{kk}(w) \log \left( 1 - \left( \sum_{l \neq k} P_{kl} \right) \left( 1 - e^{-\rho d_w} \right) \right) \]

\[ + \sum_{w=1}^{W-1} \sum_{k=1}^{4} \sum_{l \neq k} P_{kl}(w) \log \left( p_{kl} \left( 1 - e^{-\rho d_w} \right) \right) \]

\[ + \sum_{w=1}^{W} P_k(w) \log P(u_w|s, \alpha, \beta, c_0, q, v) \]

\[ \triangleq G_1(\pi_k) + G_2(p, \rho) + G_3(p, \rho) + G_4(s, \alpha, \beta, c_0, q, v). \]

We take derivative of \( g_2 \) and \( g_3 \) with respect to \( p_{kl} \):

\[ \frac{\partial G_2(p, \rho)}{\partial p_{kl}} + \frac{\partial G_3(p, \rho)}{\partial p_{kl}} \triangleq 0 \quad (k, l = 1, \ldots, 4; l \neq k) \]

\[ \Rightarrow \sum_{w=1}^{W-1} \frac{(1 - e^{-\rho d_w}) P_{kk}(w)}{1 - (1 - e^{-\rho d_w}) \sum_{l \neq k} P_{kl}} = \sum_{w=1}^{W-1} \frac{P_{kl}(w)}{p_{kl}} \quad (k, l = 1, \ldots, 4; l \neq k). \]
Then we get
\[ \sum_{w=1}^{W-1} \frac{P_{kl}(w)}{p_{kl}} = \ldots = \sum_{w=1}^{W-1} \frac{P_{k4}(w)}{p_{k4}} = \frac{\sum_{w=1}^{W-1} (1 - e^{-\rho d_w})P_{kk}(w)}{1 - (1 - e^{-\rho d_w})\sum_{l\neq k} p_{kl}} \]
for \( k, l = 1, \ldots, 4 \) and \( l \neq k \).

We ask \( \sum_{w=1}^{W-1} \frac{P_{kl}(w)}{p_{kl}} = h_k \), \( k = 1, \ldots, 4 \), and find the value of \( h_k \) that maximizes
\[
\sum_{w=1}^{W-1} P_{kk}(w) \log \left( 1 - \left( \frac{\sum_{l'\neq k} \sum_{w=1}^{W-1} P_{kl'}(w)}{h_k} \right) \left( 1 - e^{-\rho d_w} \right) \right)
+ \sum_{w=1}^{W-1} \sum_{l=1;l\neq k}^{4} P_{kl}(w) \log \left( \frac{\sum_{w=1}^{W-1} P_{kl}(w)}{h_k} \left( 1 - e^{-\rho d_w} \right) \right)
\]
for each \( k \). Then a new \( p \) value can be obtained by \( p_{kl}(w) = \frac{\sum_{w=1}^{W-1} P_{kl}(w)}{h_k} \), \( k, l = 1 \ldots, 4 \), \( l \neq k \). And \( \rho \) can be obtained by directly maximize \( G_2(p, \rho) + G_3(p, \rho) \) with respect to \( \rho \), using the new \( p \) value.

After obtaining a pair of values of \( p \) and \( \rho \) that maximize \( G_2(p, \rho) + G_3(p, \rho) \), we estimate the values for \( \alpha, \beta, q \) and \( v \) by maximizing \( G_4 \). Iteration continues until all parameter values getting converged and we obtain an updated \( \theta^{(m+1)} \) value.

### 4.9.2 CNV Identifications

Table 4.3: The m-HMM Segmentation Result in Chromosome 6 from 35.3 Mb to 57.0 Mb

<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Mo17 Read counts</th>
<th>B73 Read counts</th>
<th>CN Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>35,274,511</td>
<td>36,847,009</td>
<td>1,511</td>
<td>2291</td>
<td>1.33</td>
</tr>
<tr>
<td>36,847,028</td>
<td>36,986,070</td>
<td>81</td>
<td>242</td>
<td>0.67</td>
</tr>
<tr>
<td>36,988,155</td>
<td>39,227,681</td>
<td>2,375</td>
<td>3850</td>
<td>1.21</td>
</tr>
<tr>
<td>39,227,711</td>
<td>39,392,612</td>
<td>53</td>
<td>294</td>
<td>0.36</td>
</tr>
<tr>
<td>39,392,816</td>
<td>39,998,180</td>
<td>605</td>
<td>810</td>
<td>1.51</td>
</tr>
<tr>
<td>39,998,210</td>
<td>40,070,876</td>
<td>114</td>
<td>95</td>
<td>2.43</td>
</tr>
<tr>
<td>40,071,001</td>
<td>42,219,831</td>
<td>1,978</td>
<td>2751</td>
<td>1.45</td>
</tr>
</tbody>
</table>

Continued on Next Page...
Table 4.3 – Continued

<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Mo17 Read counts</th>
<th>B73 Read counts</th>
<th>CN Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>42,221,884</td>
<td>42,272,394</td>
<td>3</td>
<td>305</td>
<td>0.01</td>
</tr>
<tr>
<td>42,273,061</td>
<td>42,279,935</td>
<td>0</td>
<td>37</td>
<td>0.00</td>
</tr>
<tr>
<td>42,281,369</td>
<td>42,467,432</td>
<td>6</td>
<td>484</td>
<td>0.02</td>
</tr>
<tr>
<td>42,467,722</td>
<td>42,510,527</td>
<td>0</td>
<td>199</td>
<td>0.00</td>
</tr>
<tr>
<td>42,520,362</td>
<td>42,588,228</td>
<td>10</td>
<td>134</td>
<td>0.15</td>
</tr>
<tr>
<td>42,589,445</td>
<td>42,592,668</td>
<td>0</td>
<td>16</td>
<td>0.00</td>
</tr>
<tr>
<td>42,600,850</td>
<td>42,949,605</td>
<td>41</td>
<td>734</td>
<td>0.11</td>
</tr>
<tr>
<td>42,957,788</td>
<td>43,020,977</td>
<td>0</td>
<td>213</td>
<td>0.00</td>
</tr>
<tr>
<td>43,021,808</td>
<td>43,021,808</td>
<td>1</td>
<td>0</td>
<td>50.00</td>
</tr>
<tr>
<td>43,021,875</td>
<td>43,034,674</td>
<td>0</td>
<td>20</td>
<td>0.00</td>
</tr>
<tr>
<td>43,050,012</td>
<td>43,156,464</td>
<td>7</td>
<td>214</td>
<td>0.06</td>
</tr>
<tr>
<td>43,159,602</td>
<td>43,239,996</td>
<td>0</td>
<td>110</td>
<td>0.00</td>
</tr>
<tr>
<td>43,240,069</td>
<td>43,240,069</td>
<td>4</td>
<td>1</td>
<td>8.11</td>
</tr>
<tr>
<td>43,240,496</td>
<td>43,253,531</td>
<td>0</td>
<td>25</td>
<td>0.00</td>
</tr>
<tr>
<td>43,270,468</td>
<td>43,367,135</td>
<td>27</td>
<td>396</td>
<td>0.13</td>
</tr>
<tr>
<td>43,382,243</td>
<td>43,436,720</td>
<td>0</td>
<td>211</td>
<td>0.00</td>
</tr>
<tr>
<td>43,437,634</td>
<td>43,437,634</td>
<td>1</td>
<td>0</td>
<td>50.00</td>
</tr>
<tr>
<td>43,437,664</td>
<td>43,438,342</td>
<td>0</td>
<td>19</td>
<td>0.00</td>
</tr>
<tr>
<td>43,440,089</td>
<td>43,466,314</td>
<td>8</td>
<td>146</td>
<td>0.11</td>
</tr>
<tr>
<td>43,466,499</td>
<td>43,502,281</td>
<td>0</td>
<td>183</td>
<td>0.00</td>
</tr>
<tr>
<td>43,502,441</td>
<td>43,502,441</td>
<td>1</td>
<td>0</td>
<td>50.00</td>
</tr>
<tr>
<td>43,504,419</td>
<td>43,516,326</td>
<td>0</td>
<td>3</td>
<td>0.00</td>
</tr>
<tr>
<td>43,532,757</td>
<td>43,802,344</td>
<td>29</td>
<td>815</td>
<td>0.077</td>
</tr>
<tr>
<td>43,802,387</td>
<td>43,809,486</td>
<td>0</td>
<td>214</td>
<td>0.00</td>
</tr>
<tr>
<td>43,809,952</td>
<td>43,809,952</td>
<td>4</td>
<td>0</td>
<td>50.00</td>
</tr>
<tr>
<td>43,809,982</td>
<td>43,811,684</td>
<td>0</td>
<td>47</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Continued on Next Page...
Table 4.3 – Continued

<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Mo17 Read counts</th>
<th>B73 Read counts</th>
<th>CN Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>43,812,791</td>
<td>43,828,685</td>
<td>1</td>
<td>128</td>
<td>0.01</td>
</tr>
<tr>
<td>43,833,620</td>
<td>43,862,804</td>
<td>0</td>
<td>121</td>
<td>0.00</td>
</tr>
<tr>
<td>43,863,025</td>
<td>43,863,025</td>
<td>1</td>
<td>0</td>
<td>50.00</td>
</tr>
<tr>
<td>43,866,525</td>
<td>43,959,900</td>
<td>0</td>
<td>164</td>
<td>0.00</td>
</tr>
<tr>
<td>43,959,976</td>
<td>43,960,274</td>
<td>2</td>
<td>10</td>
<td>0.40</td>
</tr>
<tr>
<td>43,960,449</td>
<td>43,984,663</td>
<td>0</td>
<td>95</td>
<td>0.00</td>
</tr>
<tr>
<td>43,991,670</td>
<td>44,242,537</td>
<td>24</td>
<td>498</td>
<td>0.09</td>
</tr>
<tr>
<td>44,247,904</td>
<td>44,297,744</td>
<td>0</td>
<td>50</td>
<td>0.00</td>
</tr>
<tr>
<td>44,297,961</td>
<td>44,297,961</td>
<td>1</td>
<td>0</td>
<td>50.00</td>
</tr>
<tr>
<td>44,299,769</td>
<td>44,341,066</td>
<td>0</td>
<td>252</td>
<td>0.00</td>
</tr>
<tr>
<td>44,341,857</td>
<td>46,215,928</td>
<td>371</td>
<td>2308</td>
<td>0.32</td>
</tr>
<tr>
<td>46,215,958</td>
<td>46,834,832</td>
<td>572</td>
<td>895</td>
<td>1.29</td>
</tr>
<tr>
<td>46,835,472</td>
<td>47,130,192</td>
<td>74</td>
<td>545</td>
<td>0.27</td>
</tr>
<tr>
<td>47,130,202</td>
<td>47,668,306</td>
<td>642</td>
<td>833</td>
<td>1.56</td>
</tr>
<tr>
<td>47,681,491</td>
<td>47,739,876</td>
<td>18</td>
<td>72</td>
<td>0.50</td>
</tr>
<tr>
<td>47,740,273</td>
<td>47,740,273</td>
<td>0</td>
<td>2</td>
<td>0.00</td>
</tr>
<tr>
<td>47,746,723</td>
<td>48,880,371</td>
<td>283</td>
<td>1356</td>
<td>0.42</td>
</tr>
<tr>
<td>48,880,568</td>
<td>48,923,056</td>
<td>0</td>
<td>51</td>
<td>0.00</td>
</tr>
<tr>
<td>48,923,160</td>
<td>49,722,664</td>
<td>213</td>
<td>1128</td>
<td>0.38</td>
</tr>
<tr>
<td>49,722,844</td>
<td>50,132,724</td>
<td>231</td>
<td>298</td>
<td>1.57</td>
</tr>
<tr>
<td>50,132,929</td>
<td>50,239,093</td>
<td>28</td>
<td>119</td>
<td>0.47</td>
</tr>
<tr>
<td>50,260,756</td>
<td>52,852,085</td>
<td>2,487</td>
<td>3567</td>
<td>1.41</td>
</tr>
<tr>
<td>52,852,089</td>
<td>52,958,397</td>
<td>155</td>
<td>130</td>
<td>2.41</td>
</tr>
<tr>
<td>52,958,403</td>
<td>53,950,690</td>
<td>918</td>
<td>1453</td>
<td>1.28</td>
</tr>
<tr>
<td>53,951,899</td>
<td>56,137,418</td>
<td>895</td>
<td>3196</td>
<td>0.56</td>
</tr>
<tr>
<td>56,142,451</td>
<td>56,472,218</td>
<td>296</td>
<td>570</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Continued on Next Page...
Table 4.3 – Continued

<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Mo17 Read counts</th>
<th>B73 Read counts</th>
<th>CN Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>56,472,614</td>
<td>56,863,324</td>
<td>102</td>
<td>570</td>
<td>0.36</td>
</tr>
<tr>
<td>56,866,117</td>
<td>57,022,366</td>
<td>123</td>
<td>212</td>
<td>1.17</td>
</tr>
</tbody>
</table>

Table 4.4: Large CNV Identifications

<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Copy Number Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7,674,194</td>
<td>10,130,535</td>
<td>0.47</td>
</tr>
<tr>
<td>73,051,891</td>
<td>76,150,311</td>
<td>0.49</td>
</tr>
<tr>
<td>78,765,566</td>
<td>80,973,937</td>
<td>0.40</td>
</tr>
<tr>
<td>106,020,503</td>
<td>108,085,310</td>
<td>0.46</td>
</tr>
<tr>
<td>183,346,370</td>
<td>185,514,363</td>
<td>0.34</td>
</tr>
<tr>
<td>185,561,332</td>
<td>187,744,738</td>
<td>0.34</td>
</tr>
<tr>
<td>206,016,332</td>
<td>208,795,988</td>
<td>0.49</td>
</tr>
<tr>
<td>chr 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14,327,370</td>
<td>17,154,616</td>
<td>0.49</td>
</tr>
<tr>
<td>43,421,667</td>
<td>45,665,032</td>
<td>0.47</td>
</tr>
<tr>
<td>53,360,846</td>
<td>57,062,635</td>
<td>0.48</td>
</tr>
<tr>
<td>101,464,135</td>
<td>103,592,856</td>
<td>0.37</td>
</tr>
<tr>
<td>173,361,552</td>
<td>175,485,424</td>
<td>0.46</td>
</tr>
<tr>
<td>183,250,632</td>
<td>186,406,359</td>
<td>0.46</td>
</tr>
<tr>
<td>209,163,585</td>
<td>211,227,635</td>
<td>0.48</td>
</tr>
<tr>
<td>211,597,012</td>
<td>214,704,617</td>
<td>0.48</td>
</tr>
<tr>
<td>chr 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3,653,572</td>
<td>5,790,383</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Continued on Next Page...
<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Copy Number Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>45,465,224</td>
<td>48,749,064</td>
<td>0.40</td>
</tr>
<tr>
<td>82,020,005</td>
<td>84,053,318</td>
<td>0.48</td>
</tr>
<tr>
<td>176,248,175</td>
<td>179,361,747</td>
<td>0.48</td>
</tr>
<tr>
<td>chr 4</td>
<td>645</td>
<td>2,454,064</td>
</tr>
<tr>
<td>23,025,833</td>
<td>25,854,581</td>
<td>0.48</td>
</tr>
<tr>
<td>136,792,400</td>
<td>140,387,061</td>
<td>0.36</td>
</tr>
<tr>
<td>140,502,832</td>
<td>143,245,529</td>
<td>0.49</td>
</tr>
<tr>
<td>146,672,844</td>
<td>148,733,733</td>
<td>0.44</td>
</tr>
<tr>
<td>239,597,061</td>
<td>242,255,833</td>
<td>0.48</td>
</tr>
<tr>
<td>243,186,165</td>
<td>245,433,834</td>
<td>0.49</td>
</tr>
<tr>
<td>chr 5</td>
<td>753,941</td>
<td>4,311,600</td>
</tr>
<tr>
<td>5,201,605</td>
<td>7,723,585</td>
<td>0.48</td>
</tr>
<tr>
<td>11,328,383</td>
<td>14,574,771</td>
<td>0.49</td>
</tr>
<tr>
<td>16,611,265</td>
<td>19,032,650</td>
<td>0.44</td>
</tr>
<tr>
<td>68,583,278</td>
<td>71,949,086</td>
<td>0.40</td>
</tr>
<tr>
<td>156,325,001</td>
<td>158,697,338</td>
<td>0.47</td>
</tr>
<tr>
<td>168,069,569</td>
<td>170,816,075</td>
<td>0.39</td>
</tr>
<tr>
<td>176,121,726</td>
<td>179,598,151</td>
<td>0.43</td>
</tr>
<tr>
<td>191,459,521</td>
<td>195,013,031</td>
<td>0.48</td>
</tr>
<tr>
<td>chr 6</td>
<td>25,898,308</td>
<td>28,476,332</td>
</tr>
<tr>
<td>70,440,134</td>
<td>72,858,938</td>
<td>0.46</td>
</tr>
<tr>
<td>76,981,074</td>
<td>80,570,101</td>
<td>0.46</td>
</tr>
<tr>
<td>85,620,416</td>
<td>87,659,132</td>
<td>0.48</td>
</tr>
<tr>
<td>102,350,617</td>
<td>104,422,628</td>
<td>0.49</td>
</tr>
<tr>
<td>154,917,250</td>
<td>159,407,299</td>
<td>0.47</td>
</tr>
<tr>
<td>162,488,577</td>
<td>165,722,159</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Continued on Next Page...
Table 4.4 – Continued

<table>
<thead>
<tr>
<th>Start Position</th>
<th>End Position</th>
<th>Copy Number Ratio (Mo17/B73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18,304,500</td>
<td>22,673,932</td>
<td>0.49</td>
</tr>
<tr>
<td>25,498,467</td>
<td>30,547,169</td>
<td>0.47</td>
</tr>
<tr>
<td>78,344,496</td>
<td>81,198,663</td>
<td>0.40</td>
</tr>
<tr>
<td>110,502,115</td>
<td>113,262,773</td>
<td>0.49</td>
</tr>
<tr>
<td>117,913,056</td>
<td>119,946,052</td>
<td>0.49</td>
</tr>
<tr>
<td>120,044,241</td>
<td>122,233,553</td>
<td>0.49</td>
</tr>
<tr>
<td>chr 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20,108,771</td>
<td>23,447,168</td>
<td>0.43</td>
</tr>
<tr>
<td>160,011,805</td>
<td>166,514,647</td>
<td>0.45</td>
</tr>
<tr>
<td>167,430,533</td>
<td>169,880,960</td>
<td>0.45</td>
</tr>
<tr>
<td>chr 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38,523,842</td>
<td>40,732,642</td>
<td>0.41</td>
</tr>
<tr>
<td>46,734,889</td>
<td>49,382,812</td>
<td>0.38</td>
</tr>
<tr>
<td>61,499,201</td>
<td>64,577,206</td>
<td>0.39</td>
</tr>
<tr>
<td>83,491,462</td>
<td>88,345,837</td>
<td>0.46</td>
</tr>
<tr>
<td>118,555,633</td>
<td>121,557,270</td>
<td>0.49</td>
</tr>
<tr>
<td>143,196,239</td>
<td>145,630,853</td>
<td>0.45</td>
</tr>
<tr>
<td>145,813,631</td>
<td>150,079,829</td>
<td>0.46</td>
</tr>
<tr>
<td>chr 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20,262,176</td>
<td>24,734,243</td>
<td>0.48</td>
</tr>
<tr>
<td>39,046,734</td>
<td>41,852,436</td>
<td>0.46</td>
</tr>
<tr>
<td>72,370,857</td>
<td>74,553,412</td>
<td>0.42</td>
</tr>
<tr>
<td>111,247,775</td>
<td>116,238,001</td>
<td>0.46</td>
</tr>
<tr>
<td>117,013,634</td>
<td>120,679,932</td>
<td>0.49</td>
</tr>
<tr>
<td>121,011,209</td>
<td>124,203,304</td>
<td>0.49</td>
</tr>
<tr>
<td>130,590,451</td>
<td>139,841,271</td>
<td>0.46</td>
</tr>
</tbody>
</table>
References


CHAPTER 5. Summary

5.1 General Conclusion

This thesis studied several topics in statistical theories and methodologies related to biological and genomic areas.

In Chapter 2 and Chapter 3 we focused on statistical inference with order-restricted hypotheses. Chapter 2 derived the likelihood ratio test (LRT) and the intersection-union test (IUT) for the hypothesis with the null parameter space constructed by a union of two cones. The tests can be applied to heterosis identifications, in detecting the case where the mean trait value of the hybrid offspring more extreme than its inbred parental lines. We studied the asymptotic power of the LRT and the IUT, and shown that the LRT is uniformly more powerful than the IUT. We also discuss the situations in which tests that are uniformly more powerful than the LRT can be constructed. We concluded that, although the uniformly more powerful test are of interest in a theoretical aspect, these tests face difficulties in real practical data analyses.

Chapter 3 proposed four procedures to detect multivariate gene sets with changes in gene expression distributions under quantitative or ordinal treatment factors. The proposed tests focus more detecting power focusing on monotonic changes patterns. The four proposed tests were compared to other tests, which spread power over the entire alternative space, with the same power on the monotonic subspace and the non-monotonic subspace. We found that the proposed tests have higher detection powers when the gene expressions in multivariate gene set change monotonically. In order to deal with the problem of high dimensional random vectors and small sample sizes, we proposed to use permutation procedures to obtain non-parametric statistical inferences in all the testing procedures.

Chapter 4 explored a new methodology to identify copy number variations (CNVs) between two different genotype lines using next generation sequencing data. The method
uses a Hidden Markov Model with mixture emission probabilities to detect the CNV change points. We demonstrate that our proposed methodology improves some existing methods, with higher sensitivities and specificities.

5.2 Future work

5.2.1 Testing a union-of-cones null hypothesis with an extension on parameter dimensions

In Chapter 2, we studied tests with null parameter space being a union of two cones, where each of the cones is defined by two linear combinations with coefficients without positive relations. In future study, one possible extension that could be made is to test a union-of-cones null hypothesis with each cone defined by more linear combinations, which results in two multidimensional polyhedral cones with more parameters involved. When the number of dimension increased, the null and the alternative parameter spaces are no longer be able to project on a plate, but the extension study of the likelihood ratio test and the asymptotic null distribution of the test statistic can be conducted.

5.2.2 copy number variation detections using next generation sequencing data with corporations of single nucleotide polymorphism (SNP) information

In Chapter 4, we proposed a new methodology in copy number variation detection with next generation sequencing read counts. Based on genomic knowledge, copy number variations usually occur accompanied by single nucleotide polymorphisms (SNPs). SNP is a variation of DNA sequence on single nucleotide, Adenine (A), Cytosine (C), Guanine (G), or Thymine (T), which is the basic building block. In future work, we can make use of this SNP information, for a further improvement on the accuracies and the efficiencies of the CNV detections.