NONPARAMETRIC FUNCTIONAL MAPPING OF QUANTITATIVE TRAIT LOCI

By

JIE YANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2006
I dedicate this to my parents, Mingfa Yang and Qiyun Li, my husband, Song Wu
and my son, David D. Wu.
ACKNOWLEDGMENTS

First, I would like to thank my parents, Mingfa Yang and Qiyun Li, my sisters, Jing Yang, Hui Yang and Xiaomin Yang, for all their love and support shown at all stages of my life. I owe the greatest debt of gratitude to my husband, Song Wu, who has been at my side with his tireless love since my day one in the USA. They are great companions during the sunny and cloudy days towards my Ph.D. And I want to thank my son, David, for bring us so much joy and pride.

I express my warmest gratitude to my academic advisor and committee chair, Dr. George Casella. Learning from his intuition and his clear understanding of mathematical statistics has always been a pleasure. As his student, I feel motivated and energized after each research meeting with him, although at times I may not know exactly what to do but I work more efficiently and am more creative following a meeting with him.

My heartfelt thanks go to Dr. Rongling Wu for his generous counsel, help and encouragement during all stages of my Ph.D. study here. I enjoy every talk with him.

I would also like to extend my thanks to the other committee members: Dr. Ramon Littell, Dr. Ronald H. Randles and Dr. Marta L. Wayne. I also thank all other faculty and staff members of the Department of Statistics, University of Florida; whenever I needed advice or feedback, I felt welcome in their offices.

Equally important have been my friends at the University of Florida with their humor, comradeship, and unwavering support. I will always recall with fondness the good times spent here during the graduate work.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td></td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td></td>
<td>xiii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Genetic Basics</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Interval Mapping</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Functional Mapping</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>Splines</td>
<td>12</td>
</tr>
<tr>
<td>1.5</td>
<td>Mixed Models</td>
<td>15</td>
</tr>
<tr>
<td>1.6</td>
<td>Motivation and Outline of the Dissertation</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>DENSE-MAP CASE: WHEN THE QTL IS ON A MARKER</td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>Data Setting</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>The Likelihood Ratio Test at One Marker</td>
<td>21</td>
</tr>
<tr>
<td>2.3</td>
<td>The Likelihood Ratio Test over All Markers</td>
<td>22</td>
</tr>
<tr>
<td>2.4</td>
<td>Examples</td>
<td>25</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Simulation</td>
<td>25</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Poplar Data</td>
<td>27</td>
</tr>
<tr>
<td>2.5</td>
<td>Technical Details: Derivation of the Likelihood Ratio Test</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>NONPARAMETRIC FUNCTIONAL INTERVAL MAPPING</td>
<td>32</td>
</tr>
<tr>
<td>3.1</td>
<td>Data Setting</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Approximate Distribution of Test Statistics at Each $Q$ under $H_0$</td>
<td>35</td>
</tr>
<tr>
<td>3.3</td>
<td>Precision of the Linear Approximation to LRT Statistics under $H_0$</td>
<td>41</td>
</tr>
<tr>
<td>3.4</td>
<td>Alternative Approach to Calculate P-value: Simulation</td>
<td>49</td>
</tr>
<tr>
<td>3.5</td>
<td>Estimation of Consistent Variance-Covariance Matrix with Small Sample Size</td>
<td>56</td>
</tr>
<tr>
<td>3.6</td>
<td>Application to Poplar Data - Revisited</td>
<td>61</td>
</tr>
<tr>
<td>3.6.1</td>
<td>A Chromosome-wise Analysis on Linkage Group 10</td>
<td>61</td>
</tr>
<tr>
<td>3.6.2</td>
<td>A Genome-wise Analysis Using REML Estimate</td>
<td>63</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–1</td>
<td>Conditional probabilities of genotype at a QTL bracketed by markers $\mathcal{M}$ and $\mathcal{N}$ in a backcross population</td>
<td>5</td>
</tr>
<tr>
<td>1–2</td>
<td>Distribution of the phenotypic values for different genotype groups in a backcross population; $\theta_1 = \frac{r_1 r_2}{1-r}$ and $\theta_2 = \frac{r_1 (1-r_2)}{r}$</td>
<td>6</td>
</tr>
<tr>
<td>1–3</td>
<td>Conditional probabilities of genotype at a QTL bracketed by markers $\mathcal{M}$ and $\mathcal{N}$ in a backcross population, when $r_1$ is relatively small</td>
<td>6</td>
</tr>
<tr>
<td>3–1</td>
<td>Analysis of Poplar data using nonparametric functional interval mapping (NPFIM) and parametric functional interval mapping (PFIM). “EB”, the empirical Bayes estimator. “REML”, the REML estimator corresponding to structure $\sigma^2 J + \text{Autoregressive}(\tau^2, \rho)$</td>
<td>64</td>
</tr>
<tr>
<td>3–2</td>
<td>P-values and its standard deviation of nonparametric functional interval mapping (NPFIM) using different variance-covariance estimators under different true $\Sigma$s. “EB” means the shrinkage estimator which is guaranteed to be a consistent estimator. “REMLw” estimator is the REML estimator corresponding to a wrong structure. “REML” estimator is the one selected from SAS Proc MIXED using BIC and assuming each subject has different underlying growth curve. “True” estimator is of course the matrix we actually used to generate data.</td>
<td>66</td>
</tr>
<tr>
<td>3–3</td>
<td>P-values and its standard deviation of nonparametric functional interval mapping (NPFIM) using different variance-covariance estimators for different sample size and $\Sigma = \Sigma_3$. “EB” means the shrinkage estimator which is guaranteed to be a consistent estimator. “REMLw” estimator is the REML estimator corresponding to a wrong structure. “REML” estimator is the one selected from SAS Proc MIXED using BIC and assuming each subject has different underlying growth curve. “True” estimator is of course the matrix we actually used to generate data.</td>
<td>68</td>
</tr>
<tr>
<td>3–4</td>
<td>P-values and standard deviation of nonparametric functional interval mapping (NPFIM) from HIV dynamics data and growth data for different combination of variance-covariance estimators and sample size. “EB” means the shrinkage estimator which is guaranteed to be a consistent estimator. “REML” estimator is obtained from SAS Proc MIXED assuming each subject has a different underlying mean curve. “True” estimator is, of course, the matrix we actually used to generate data.</td>
<td>72</td>
</tr>
</tbody>
</table>
4–1 P-values and its standard deviation of nonparametric functional interval mapping (NPFIM) using case-wise deleted data set and full data set, \[ \Sigma = 0.3 \times J_{11} + \text{Autoregressive}(\tau^2 = 0.1, \rho = 0.8) \] .......................................... 97

4–2 Comparison of critical values from different analyzing methods ............ 98

4–3 Distribution of missing marker conditional on the other observed marker for \( k \)th marker interval for subject \( i \) ................................................................. 103

4–4 Conditional distribution of missing markers of \( k \)th marker interval for subject \( i \) ................................................................. 103

4–5 Distribution of missing marker conditional on the other observed marker for \( k \)th marker interval for subject \( i \) ................................................................. 107

4–6 Conditional distribution of missing markers of \( k \)th marker interval for subject \( i \) ................................................................. 108
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–1</td>
<td>Illustrations of one isolated B-spline and several overlapping ones with degree 1 and 2, respectively.</td>
<td>13</td>
</tr>
<tr>
<td>2–1</td>
<td>Graph of the $p$–value and its 95% confidence interval for data from two underlying flat lines with different distances.</td>
<td>26</td>
</tr>
<tr>
<td>2–2</td>
<td>Graph of $p$–value and its 95% confidence interval for data from two underlying curves: one is flat line the other is in quadratic form.</td>
<td>27</td>
</tr>
<tr>
<td>2–3</td>
<td>Original poplar data profiles.</td>
<td>28</td>
</tr>
<tr>
<td>2–4</td>
<td>The growth curve representing two groups of genotypes at marker CA/CCC-640R on linkage group 10 in the Populus deltoides parent map.</td>
<td>29</td>
</tr>
<tr>
<td>3–1</td>
<td>The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1&lt;i&lt;n_3} S_i \ast 10^5$ when number of subjects $n$ increases under simplest setting that all data from standard normal distribution. Recombination rate between two markers is .2 and $\theta = .5$.</td>
<td>45</td>
</tr>
<tr>
<td>3–2</td>
<td>The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1&lt;i&lt;n_3} S_i$ when number of subjects $n$ increases under simplest setting that all data from standard normal distribution. Recombination rate between two markers is .2 and $\theta = .5$.</td>
<td>46</td>
</tr>
<tr>
<td>3–3</td>
<td>The plot of function $f(x) = f_1(x) - f_2(x)$.</td>
<td>47</td>
</tr>
<tr>
<td>3–4</td>
<td>The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1&lt;i&lt;n_3}</td>
<td>S_i</td>
</tr>
<tr>
<td>3–5</td>
<td>The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1&lt;i&lt;n_3}</td>
<td>S_i</td>
</tr>
<tr>
<td>3–6</td>
<td>The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1&lt;i&lt;n_3}</td>
<td>S_i</td>
</tr>
<tr>
<td>3–7</td>
<td>Two simulated distributions of likelihood ratio test statistic based on same 1000 datasets.</td>
<td>51</td>
</tr>
<tr>
<td>3–8</td>
<td>Two simulated distributions based on same 1000 datasets when treating $\theta$ as unknown parameter.</td>
<td>54</td>
</tr>
</tbody>
</table>
3–9 Two simulated distributions based on different 1000 datasets while treating $\theta$ as unknown parameter.

3–10 Fitted curves from using two different variance-covariance estimator.

3–11 The profile of the likelihood ratio test statistics between the full and reduced (no QTL) subject-specific model for the diameter growth trajectories across the whole *Populus deltoides* parent map. The genomic positions corresponding to the peak of the curve are the MLEs of the QTL localization.

3–12 Heritability curves for the four simulated data sets using different $\Sigma$s.

3–13 The left plot shows the 10 heritability curves of simulated growth data sets. The right plot shows the trend of P-values when heritability on year 4 increases. “EB” and “REML” means the Empirical Bayes estimate and REML estimate of covariance matrix, respectively. “True” represents the covariance matrix used to generate data.

3–14 The HIV dynamics curves used as the mean curves.

3–15 The heritability curve across all observation time points in the HIV simulation study.

3–16 The empirical distribution of the test statistics in the simulation study of Chapter 3. Dataset 1-4 correspond to the four $\Sigma$ matrixes in Table 3–2. “EB” means the Empirical Bayes estimate. “True” represents the covariance matrix used to generate the data.

3–17 The left plot indicates the empirical distribution of the test statistics in the simulated growth data of Chapter 3 with covariance matrix $\Sigma_1$, 61 subjects with 11 observation points. The right one is the one from the simulated HIV dynamics data, 61 subjects with 20 observation points.

3–18 Graph of the $p$–value and its 95% confidence interval for data from two underlying flat lines with different distances using nonparametric functional interval mapping.

3–19 Graph of $p$–value and its 95% confidence interval for data from two underlying curves: one is flat line and the other is in quadratic form using nonparametric functional interval mapping.

3–20 Graph of the likelihood ratio test statistics along the linkage group 10 using nonparametric functional interval mapping.

4–1 Histograms of LRT statistics from different combination of methods handling missing data and methods to estimate $\Sigma$ under $H_0$. 
4–2 The trend of P-values and their 95% confidence intervals (the dotted lines) from using case-wise deleted data (the dashed line) and using full data set (the black line). Underlying phenotypic lines are two flat lines and true covariance matrix is used in analysis.

4–3 This figure shows the histogram of \( \theta \) given \( k = 7 \) from the simulated dataset where QTL is assumed in interval 7. The REML estimate from interval-wise deleted data is used in calculation. “n” is the number of last drawn samples used to plot the histogram.

4–4 The histograms of the interval index \( k \) and \( \theta \) conditional on \( k \) from the simulated dataset under \( H_0 \). The true covariance matrix is used in calculation.

4–5 The histograms of the interval index \( k \) from the last 25,000 iterations (left) and the last 10,000 iterations (right). The REML estimate from interval-wise deleted data is used. Data set is simulated under \( H_0 \).

4–6 The histograms of \( \theta \) from the last 10,000 iterations. The REML estimate from interval-wise deleted data is used. Data set is simulated under \( H_0 \).

4–7 The histograms of \((k, \theta)\) from Gibbs sampler. The left figure shows the histogram of the possible interval where QTL falls in. The right figure shows the histogram of \( \theta \) given \( k = 1 \).

4–8 The underlying four different phenotypic curves controlled by two QTLs together.

4–9 The plot shows the histogram of \( \theta \) given \( k = 2 \). Median of \( \theta \) is .849 while the true \( \theta \) is .917. The REML estimate from interval-wise deleted data is used in calculation. Heritability value at the eight observation point is set to be .4. Two QTLs are assumed to be in interval 2 and interval 8, respectively. Start value for \( k \) is 4. The last 10,000 samples from 50,000 iterations are used to make the plots.

4–10 The histogram of \((k, \theta)\) from the last 10,000 samples of 50,000 iterations. Two QTLs are assumed to be in interval 2 and interval 8, respectively. The left plot shows the histogram of interval index which gives \( \hat{k} = 2 \). The right plot shows the histogram of \( \theta \) given \( k = 2 \). Median of \( \theta \) is .749 while the true \( \theta \) is .917. The REML estimate from interval-wise deleted data is used in calculation. Heritability value at the eight observation point is set to be .1. Start value for \( k \) is 6.
The histogram of \((k, \theta)\) from the last 10,000 samples of 50,000 iterations. Two QTLs are assumed to be in interval 4 and interval 6, respectively. The left plot shows the histogram of interval index which gives \(k = 6\). The right plot shows the histogram of \(\theta\) given \(k = 6\). Median of \(\theta\) is .77 while the true \(\theta\) is .549. The REML estimate from interval-wise deleted data is used in calculation. Heritability value at the eight observation point is set to be .1. Start value for \(k\) is 4.
Functional mapping is a tool for detecting major genes responsible for different phenotypic curves, developed by Ma, Casella and Wu. The methodology uses a parametric functional form, usually derived from a biological law, to drive a maximum-likelihood-based test for a significant QTL (quantitative trait loci). However, in many situations there is no obvious functional form and, in such cases, this strategy will not be optimal.

In this dissertation we propose to use nonparametric function estimation, typically implemented with B-splines, to estimate the underlying functional form of phenotypic trajectories, and then construct a nonparametric test to find evidence of existing quantitative trait loci. Using the representation of a nonparametric regression as a mixed model, we can easily derive a likelihood ratio test statistic. Two situations are considered: one is based on the dense-map assumption that QTL (quantitative trait loci) is on some marker and the other is the situation where the actual genes responsible for different underlying phenotypic trajectories might not just be on a marker locus, which is more realistic. For the dense-map case, after obtaining the joint distribution of all test statistics at every putative
locus (each marker), we can then calculate the p-value directly. Simulation studies show that our method is both powerful and quick. We also provide an application to a real data set.

Nowadays with the cutting-edge biotechnology, more and more marker information can be obtained. Along with this exciting human achievement, the dataset has more possibility to have missing cells. In this dissertation we also take into account this case.
CHAPTER 1
INTRODUCTION

The variations of many quantitative traits in human, plants and animals can be attributed mainly to the segregation of multiple genetic factors. Sax (1923) first investigated the gene affecting seed size in beans with pattern and pigment markers. Since then, the theoretical principle of using markers to analyze quantitative trait loci (QTL) is well established (Sax (1923); Jansen and Stem (1994); Zeng (1993,1994); Kruglyak and Lander (1995); Kao et al. (1999); Ma, Casella and Wu (2002)). In recent years, mapping of QTL is greatly facilitated by fine-scale genetic marker maps constructed through modern molecular biology techniques. After Lander and Botstein’s (1989) pioneering work, many statistical methodologies for mapping Quantitative Trait Loci (QTL) on a high-density linkage map were proposed. These methods have been used to identify major genes controlling different traits important to biomedicine, biology science, agriculture, zoology and so on. In Section 1.2, a brief introduction to interval mapping, first proposed by Lander and Botstein (1989) is given.

Typically these methods treat traits as discrete, even though the trait is an infinite-dimensional characteristic (Kirkpatrick and Heckman 1989) or a functional-valued trait (Pletcher and Geyer 1999), which may be expressed as a smooth function of one continuous variable. In real life, as a consequence of natural selection, an infinite-dimensional characteristic may change its phenotype to be more competitive. Therefore, embedding the underlying biological mechanism and processes into a QTL mapping strategy is a reasonable approach. Functional mapping, developed by Ma, Casella and Wu (2002), implements a parametric
function, usually derived from a universal biological law, to describe a phenotypic curve. For example, a logistic growth curve can be described by a mathematical form \( g(t) = \frac{a}{1 + b \exp(-r_1 t)} \) with parameters \( a, b, r \). Because of its parametric nature, it is straightforward to test numerous biologically important hypotheses by testing some parameters or functions of parameters using the maximum likelihood principle. This method is powerful in providing the estimated position of QTL governing the character process. Section 1.3 provides an introduction of this functional mapping framework. A short description of the comparison between traditional discrete mapping methods and functional mapping methods is also included in Section 1.3.

In this dissertation, we extended parametric functional mapping to non-parametric functional mapping because we propose to use splines to estimate the underlying function and use the spirit of mixed models to characterize the complexity of our beautiful nature. In Section 1.4 and Section 1.5 we briefly explain about splines and mixed models.

1.1 Genetic Basics

The work of Gregor Mendel marked the beginning of modern genetics. (He was an Austrian Monk who developed an interest in heredity, worked with peas and published his work in the 1870’s. His publication was discovered by three independent investigators at the turn of the century, and all three quickly realized the significance of the work. Unfortunately, Mendel died before his work was rediscovered.) Genes are codes of life, which are some particular regions on a chromosome, while the inter-genic materials on chromosomes carry no information. He formed the basis of genetics:

1. **The Law of Segregation:** factors (called alleles) controlling the particular trait separate cleanly one from the other and there is no mixing;

2. **The Law of Independent Assortment:** alleles of one gene segregate while alleles of different genes assort independently.
Many traits of agronomic and horticultural interest are controlled by a single gene and fall into a few distinct phenotypic classes. These classes can be used to predict the genotypes of the individuals. And if we know the genotype we could predict the phenotype. These type of phenotypes are called *discrete traits*. Other traits may have a distribution that resembles the bell-shaped curve for a normal distribution such as human IQ, crop yield, weight gain in animals, fat content of meat etc. These types of traits are called *continuous traits*. Because continuous traits are often given a quantitative value, they are often referred to as quantitative traits and therefore the loci controlling these traits are called *quantitative trait loci* or *QTL*.

The purpose of QTL mapping is to link quantitative phenotypic traits with regions on a chromosome. Two questions are raised: (1) Are there significant QTL effects? and (2) where are the QTLs if they do exist? For mapping a QTL, a cross is arranged between two inbred lines (which means they are homozygotes everywhere) which differ substantially in the quantitative trait of concern. Assume there are two parents having alleles $QQ$ and $qq$ at a certain (unknown) place (named locus) on the chromosome. The offspring of these parents are called the $F_1$ generation which have the allele $Qq$. If the offspring is mated back to one of the parents, say $QQ$, then the new generation produced has alleles $QQ$ and $Qq$. Such population structure is called a *backcross* and the new generation is called $BC_1$.

Since our mapping method is to find the QTL $Q$ or $q$, of course, that locus is unknown. All we know are markers along the chromosome, say $M$ or $m$. Recombination occurs when alleles cross over to another chromosome and recombination rate $r$ can be used as a measure of closeness between two genes or between gene and marker. So we can use markers to locate QTL. If the recombination rate $r$ between marker and QTL is small, with zero being the limiting case of no recombination, then marker $M$ is closely linked to QTL $Q$, that is, when we see $M$ we
hope $Q$ is also there. If $r$ is large with value $1/2$ being the limiting case, then there may be no linkage between $M$ and $Q$. The backcross is the simplest design in which we get enough information to estimate $r$.

A simple statistical model directly follows. Assume $Y$ is a random variable that follows a normal distribution. The variance of $Y$ is $\sigma^2$ and the mean of $Y$ depends on the alleles under consideration, $\mu_{QQ}, \mu_{Qq}$ or $\mu_{qq}$. Clearly in the backcross, parents are $N(\mu_{QQ}, \sigma^2)$ (or $N(\mu_{qq}, \sigma^2)$) and $N(\mu_{Qq}, \sigma^2)$ and generation $BC_1$ is a mixture of normals with means $\mu_{Qq}$ and $\mu_{QQ}$ (or $\mu_{qq}$). If the QTL $Q$ is known, the mixing factor would be $1/2$, but we can only observe the marker genotypes $M$ or $m$. For example, the parents of $BC_1$ have alleles $MQ|MQ$ and alleles $mq|MQ$. When there is no recombination between $M$ and $Q$, we will see alleles $MQ|MQ$ and $mq|MQ$ in generation $BC_1$. But when recombination between $M$ and $Q$ occurs, we will expect four possible genotypes $MQ|MQ$, $Mq|MQ$, $mQ|MQ$ and $mq|MQ$. Now the distribution of genotypes in this backcross population is

$$Y|MM \sim \begin{cases} 
N(\mu_{QQ}, \sigma^2), & \text{with probability } 1 - r \\
N(\mu_{Qq}, \sigma^2), & \text{with probability } r
\end{cases}$$

and

$$Y|Mm \sim \begin{cases} 
N(\mu_{QQ}, \sigma^2), & \text{with probability } r \\
N(\mu_{Qq}, \sigma^2), & \text{with probability } 1 - r
\end{cases}.$$ 

The difference in means of the populations when categorized by the markers is

$$\mu_{MM} - \mu_{Mm} = (1 - 2r)(\mu_{QQ} - \mu_{Qq}).$$

Assuming $\mu_{QQ} - \mu_{Qq} \neq 0$, a test of $H_0 : \mu_{MM} - \mu_{Mm} = 0$ is equivalent to testing $H_0 : r = 1/2$, which means no linkage between $M$ and $Q$. This test can be carried out with something as simple as a t-test or the more popular used likelihood methods. Using this mixture model in likelihood analysis, we can not only test for linkage but also can estimate $r$. 
Table 1–1: Conditional probabilities of genotype at a QTL bracketed by markers \( M \) and \( N \) in a backcross population

<table>
<thead>
<tr>
<th>Marker</th>
<th>QTL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( MN</td>
<td>mn )</td>
<td>((1-r_1)(1-r_2))</td>
</tr>
<tr>
<td>( Mn</td>
<td>mn )</td>
<td>( \frac{1-r}{r_1(1-r_2)} )</td>
</tr>
<tr>
<td>( mN</td>
<td>mn )</td>
<td>( \frac{r}{r_2} )</td>
</tr>
<tr>
<td>( mn</td>
<td>mn )</td>
<td>( \frac{1-r}{1-r} )</td>
</tr>
</tbody>
</table>

1.2 Interval Mapping

The principle of interval mapping was first established by Lander and Botstein (1989). The success of interval mapping as a powerful method for linkage analysis of a complex trait has roots in the rapid development of molecular technologies to detect more and more markers of any organism and improved statistical and computational techniques, such as the EM algorithm (Dempster et al. 1977), which makes it possible to tackle complex genetic and genomic problems.

In an interval mapping procedure we typically use two markers to better locate QTL \( Q \) and hence a somewhat more complex model. Assume the two markers bracketing the unknown QTL are \( M \) (with alleles \( M \) and \( m \)) and \( N \) (with alleles \( N \) and \( n \)). These two markers make four distinct genotype groups in a backcross population

\[
MN|mn, Mn|mn, mN|mn, mn|mn.
\]  

(1–1)

Denote the recombination fraction between \( M \) and \( N \) by \( r \), and that between \( M \) and QTL \( Q \) by \( r_1 \), and that between \( N \) and QTL \( Q \) by \( r_2 \). The numerical relationship between \( r, r_1 \) and \( r_2 \) is \( r = r_1 + r_2 - r_1r_2 \).

In a genetic map, genetic distances \( D \) are used to imply the location of detected markers. Genetic distances and recombination rates are related through a map function (Ott 1991). We usually use the Haldane map function to link the genetic distance \( D \) to corresponding recombination rate \( r : r = (1 - e^{-2D})/2. \)
Table 1–2: Distribution of the phenotypic values for different genotype groups in a backcross population; \( \theta_1 = \frac{r_1 r_2}{1 - r} \) and \( \theta_2 = \frac{r_1 (1 - r_2)}{r} \)

<table>
<thead>
<tr>
<th>Marker</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>mn</td>
</tr>
<tr>
<td>Mn</td>
<td>mn</td>
</tr>
<tr>
<td>mN</td>
<td>mn</td>
</tr>
<tr>
<td>mn</td>
<td>mn</td>
</tr>
</tbody>
</table>

Table 1–3: Conditional probabilities of genotype at a QTL bracketed by markers \( M \) and \( N \) in a backcross population, when \( r_1 \) is relatively small

<table>
<thead>
<tr>
<th>Marker</th>
<th>QTL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MN</td>
<td>mn</td>
</tr>
<tr>
<td>Mn</td>
<td>mn</td>
</tr>
<tr>
<td>mN</td>
<td>mn</td>
</tr>
<tr>
<td>mn</td>
<td>mn</td>
</tr>
</tbody>
</table>

The conditional probability of a QTL genotype, say \( Qq \), given each of the four two-marker genotypes is given in Table 1–1. Let \( \theta_1 = \frac{r_1 r_2}{1 - r} \) and \( \theta_2 = \frac{r_1 (1 - r_2)}{r} \), the corresponding distributions of the phenotypic values for the four marker genotypes can be modeled, respectively, as stated in Table 1–2, where \( \mu_{Qq}, \mu_{qq}, \sigma^2 \) are the unknown parameters contained in the mixture models and can be estimated using likelihood principle.

When \( \theta = r_1/r \) is small, we can use the following approximated conditional probabilities listed in Table 1–3. In the dissertation, we use these approximate conditional probabilities. Suppose that \( N_1, N_2, N_3 \) and \( N_4 \) are the sample sizes of the four marker groups (1–1). If we define \( n_1 = N_1, n_2 = n_1 + N_2, n_3 = n_2 + N_3 \) and \( n = n_3 + N_4 \), the log likelihood of the phenotype data conditional on the marker
information can be written as

$$\ell(\mu_{Qq}, \mu_{qq}, \sigma^2) = -\frac{n}{2} \log \sigma^2$$

$$- \sum_{i=1}^{n_1} \frac{1}{2\sigma^2} (y_i - \mu_{Qq})^2$$

$$+ \sum_{i=n_1+1}^{n_2} \log[(1 - \theta) \exp(-\frac{1}{2\sigma^2} (y_i - \mu_{Qq})^2) + \theta \exp(-\frac{1}{2\sigma^2} (y_i - \mu_{qq})^2)]$$

$$+ \sum_{i=n_2+1}^{n_3} \log[\theta \exp(-\frac{1}{2\sigma^2} (y_i - \mu_{Qq})^2) + (1 - \theta) \exp(-\frac{1}{2\sigma^2} (y_i - \mu_{qq})^2)]$$

$$- \sum_{i=n_3+1}^{n} \frac{1}{2\sigma^2} (y_i - \mu_{qq})^2.$$ 

In Lander and Botstein’s (1989) model for interval mapping of QTL, the profile of likelihood ratio test (LRT) statistics is constructed over the grid of possible QTL locations, which is implied by \(\theta\), in a linkage group or an entire genome. (The genome is the complete set of linkage groups or chromosomes.) The maximum of the LRT statistics is used as a global test statistic. At a given position of the QTL, that is, when \(\theta\) is fixed, the LRT statistic is asymptotically \(\chi^2\)– distributed with degrees of freedom equal to number of associated QTL effects under the null hypothesis that no QTL exists at this \(\theta\). However, under \(H_0: \) no QTL, the QTL position is unidentified and therefore, the final global maximum LRT statistic does not follow the standard \(\chi^2\)-distribution asymptotically. This is exactly why there is no closed form to decide the critical threshold to declare the statistical significance of a QTL. But several authors derived approximate formulas to determine critical thresholds for a particular design based on the results of Davies (1977, 1987) (Rebai et al. (1994); Doerge and Rebai (1996); Piepho (2001)).

To overcome the limitations due to the failure of the test statistic to follow a standard statistical distribution, one can also use simulation studies (Lander and Botstein (1989); Van Ooijen (1992); Chen and Chen (2005)) and permutation tests (see Churchill and Doerge (1994); Doerge and Churchill (1996)) to calculate the
threshold value throughout a genome. Lander and Botstein (1989) established
the finite-dimensional convergence of LOD score process, which arises from the
LRT, to the Orenstein-Uhlenbeck process. Chen and Chen (2005) established
a novel theorem for the asymptotic distribution of the LRT statistic calculated
from one marker interval for backcross model. Then a distribution-free simulation
approach can be used to calculate critical values. Either the simulation-based or
permutation-based approach is a highly computational-demanding approach. For
permutation test, one need perform at least 10,000 permutations for the same
data set to obtain a reasonably accurate estimate of a critical threshold at a
genome-wide type I error rate of 0.01 (Doerge and Rehai (1996)).

Lander and Botstein’s (1989) interval mapping method is based on a simplified
situation, that is, the segregation pattern of all markers strictly agrees with the
Mendalian laws and there is only one QTL on a chromosome controlling a trait
under study. Many authors extended and improved this work by including markers
from other intervals as covariates to control the overall genetic background (Jansen
and Stam (1994); Zeng (1994); Haley et al. (1994); Xu (1996)). Kao et al. (1999)
used multiple marker intervals simultaneously to map multiple QTL of epistatic
interaction throughout a linkage map.

1.3 Functional Mapping

As we know function-valued traits change as a function of some independent
and continuous variable, such as growth trajectories, allometric scalings, and norms
of reaction. A simple approach for mapping such infinite-dimensional character is
to treat it as a discontinuous trait, that is, link markers with phenotypes separately
for different ages, traits, or environments (Cheverud et al. (1996); Nuzhdin et al.
(1997); Verhaegen et al. (1997); Emebiri et al. (1998); Wu et al. (1999)). However,
these separate analyses cannot provide effective estimates because they fail to
implement the information about the covariances of different traits or the same
trait measured at different ages or environments.

Multitrait mapping takes account of covariances among simultaneously
different traits or the same trait measured at different ages or environments. But it
is difficult to produce precise estimate when the number of traits increases (Jiang
and Zeng (1995); Korol et al. (1995); Ronin et al. (1995); Eaves et al. (1996);
Knott and Haley (2000)). To circumvent such difficulties, some authors attempted
to transform the initial trait space into a space of a lower dimension based on
principal component analysis (Mangin et al. (1998); Korol et al. (2001)). However,
they still treat infinite-dimensional characters as discrete traits or eigenvalues and
do not consider the physiological mechanisms predisposing for the phenotypic
variation of functional-valued character.

The functional mapping method first introduced by Ma et al. (2002), incorpo-
rates the underlying physiological or developmental mechanisms of trait variation
into statistical analysis of QTL mapping framework, thus is more likely to produce
more accurate results in terms of biological reality. In their paper, an application
to map the growth of forest trees is used to illustrate their method. So we also use
that example to briefly introduce the functional mapping method.

The simplest backcross design is assumed so that there are only two groups
of genotypes at a locus. Of course, the genotypes of QTL also only have two
possibilities. In practice, instead of a continuum we only observe a finite set of
observations, \( y_i(1), ..., y_i(m) \) for each tree \( i, i = 1...n \). Suppose this finite set of data
can be modelled as the logistic growth curve, which has a mathematical form

\[
g(t) = \frac{a}{1 + be^{-rt}}.\]

For a particular genotype \( j \) (let \( j = 1 \) denote QTL genotype \( Qq \) and \( j = 2 \) for
\( QQ \)), the parameters describing corresponding growth curve are \( a_j, b_j \) and \( r_j \). To
determine whether and how the QTL affects growth trajectories is simply to test the difference or some functions of these parameters.

Similar to interval mapping, the phenotypes of the trait at all time points for each QTL genotype group follows a multivariate normal distribution:

$$f_j(y) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left[-(y - g_j)'\Sigma^{-1}(y - g_j)/2\right],$$

where $g_j$ is the vector of the expected phenotypic value corresponding to QTL genotype group $j$ at $t$ measurement times and $\Sigma$ models the dependence structure between observations from same subject. When we use logistic growth curve to model $y$ then

$$g_j = [g_j(t)]_{1 \times m} = \left[\frac{a_j}{1 + b_je^{-r_jt}}\right]_{1 \times m}.$$  

Ma et al. used an $\Sigma = \text{Autoregressive}(\sigma^2, \rho)$ repeated measurement error structure given by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{m-1} \\ \rho & 1 & \cdots & \rho^{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \cdots & 1 \end{pmatrix}. \quad (1-3)$$

The likelihood function of the backcross progeny with $m$-dimensional measurements can be expressed as a multivariate mixture model

$$L(a_j, b_j, r_j, \theta, \rho, \sigma) = \prod_{i=1}^{n} \left[ \sum_{j=1}^{2} p_{ij} f_j(y_i) \right],$$

where $p_{ij}$, a function of parameter $\theta$ as in Table 1–3, is the conditional probability of subject $i$ having QTL genotype $j$ given observed flanking markers. The maximum-likelihood-based method implemented with the EM algorithm is used to estimate all unknown parameters.

A lot of biologically meaningful hypotheses can be tested using such genetic models. For example, the hypothesis to test the existence of a QTL affecting an
overall growth curve can be stated as

\[ H_0 : a_1 = a_2, b_1 = b_2, r_1 = r_2 \]

against

\[ H_1 : \text{at least one of the equalities above does not hold.} \]

To test if the detected QTL starts to exert or ceases an effect on growth curves at particular time \( t^* \) is essentially to test

\[ H_0 : g_1(t^*) = g_2(t^*) \]

against

\[ H_1 : g_1(t^*) \neq g_2(t^*). \]

The test statistic for testing such hypotheses is calculated as the log-likelihood ratio of the full model over the reduced model. Ma et al. (2002) used 1000 permutation tests to obtain the chromosome-wide empirical estimate of the critical value.

This functional mapping method allows them to successfully detect one QTL underlying poplar stem growth locating on linkage group 10 while traditional interval mapping (Lander and Bostein (1989)) and composite interval mapping (Zeng (1994)) failed to do so, which suggests that by incorporating logistic growth curves functional mapping procedure has greater power to detect a significant QTL than the current methods. The increased detection power of functional mapping results from the simultaneous use of repeated measurements that are correlated to each other due to biological or/and environmental reasons.

Another advantage of functional mapping is that by treating phenotypic values as a function of some variables a large number (theoretically unlimited number) of measurements can be analyzed. An initially high-dimensional mapping model becomes more tractable and the precision of the estimates of QTL parameters
gets improved. Also, a small sample size can have enough power of QTL detection because the function modelled from repeated measurements extracts maximum information about QTL effects and positions.

1.4 Splines

Nowadays, exploratory analysis of data is more common and many datasets are too “rich” to be modelled in parametric ways. So smoothing gains a respectful space in statistics. Its popularity can be demonstrated by the continuous appearance of many papers and a number of books (here to name a few: Silverman (1986); Eubank (1988); Hastie and Tibshirani (1990); Härdle (1990); Wahba (1990); Wand and Jones (1993); Green and Silverman (1994); Ramsay and Silverman (1997); Ruppert, Wand and Carroll (2003)). Smoothing methods can be applied by running statistics like kernel smoothers (Silverman (1986); Härdle (1990)), LOWESS (Cleveland (1979)), splines smoothers and so on. Spline smoothers also have several varieties: smoothing splines, regression splines (Eubank (1988)), B-splines (De Boor (1978); Dierckx (1993)), P-splines (Eilers and Marx (1996)). B-splines are attractive for nonparametric modelling which is embedded in popular statistical softwares such as R/Splus. Actually, the name “nonparametric” is not well-chosen because splines are described by parameters, but these parameters have no particular scientific interpretation. So in this dissertation we use B-splines to model the underlying phenotypic function and still call our method “nonparametric functional mapping”.

Because B-splines are used in our proposed method, we will put emphasis on introducing B-splines. A B-spline consists of polynomial pieces, connected at certain values of $x$, the knots. A very simple example is shown at the upper part of Figure 1–1: one B-spline of degree 1. It consists of two linear pieces: one is from $x = 1$ to $x = 2$, the other is from $x = 2$ to $x = 3$. The knots are $x = 1, 2, 3$. To the left of $x = 1$ and to the right of $x = 3$ the value of this B-spline is zero. In the right
Figure 1–1: Illustrations of one isolated B-spline and several overlapping ones with
degree 1 and 2, respectively.
part of this upper figure in Figure 1–1; three more B-splines of degree 1 are shown with each one based on three knots.

Also in the Figure 1–1, the lower figure illustrates B-splines with degree 2. We can see that the leftmost B-spline consists of three quadratic pieces joined at two knots $x = 2$ and $x = 3$. At the joint knots the values of these quadratic pieces match and so do the first derivatives (but not their second derivatives). In the right part, three overlapping B-splines with degree 2 are shown.

Note a first-degree B-spline can overlap with only at most two neighbors and a second-degree B-spline overlaps with four neighbors. The leftmost and rightmost splines have less overlap. These simple examples shed light on the general properties of a B-spline of degree $p$:

- It consists of $p + 1$ degree $p$ polynomial pieces;
- The polynomial pieces join at $p$ inner knots;
- Derivatives up to order $p - 1$ are continuous at the inner knots;
- B-spline is nonzero only on a domain spanned by $p + 2$ knots while everywhere else is zero;
- A B-spline overlaps with $2p$ neighbor B-splines except at the boundaries;
- At a given $x$, $p + 1$ B-splines are nonzero.

So $n$ knots can determine $n - p - 1$ splines with degree $p$. In the R/Splus software, order $p + 1$ is used in place of degree $p$. A common choice of B-splines is piecewise cubic, that is, degree 3, which can provide a smooth fit.

De Boor (1978) gave an algorithm to compute B-splines of any degree recursively from B-splines of lower degree, which is an simple algorithm for a zero-degree B-spline in just a constant on one interval between two knots. Please see De Boor (1978) for details.

It is straightforward to see that the number and position of knots determine the smoothness and fit. But how to choose the optimal combinations of these
factors is a complex task. Some approaches start with a free choice of knot locations and a rather dense set of knots and then eliminate unneeded knots by an algorithmic procedure similar to variable selection techniques used in multiple regression (see, for example, Friedman and Silverman (1989)). O’Sullivan (1986, 1988) proposed to use a relatively large number of knots but put a penalty on the second derivative to restrict the flexibility of curve fitting. Eilers and Marx (1996) followed this idea and used a simple difference penalty on the coefficients themselves of adjacent B-splines, which they call P-splines. P-splines can be applied in any context where regression on B-splines is useful.

1.5 Mixed Models

Mixed models, or random-effects models have been extensively applied in longitudinal data analysis, which is designed to investigate changes over time in a characteristics measured repeatedly on the same subjects (Laird and Ware (1982); Wolfinger (1993); Davidian and Giltinan (1995); Littell et al. (2000); Verbeke and Molenberghs (2000); Diggle et al. (2002); Davidian and Giltinan (2003); Kowalchuk et al. (2004)). Often, we cannot fully control the circumstances under which the measurements are taken and there may be considerable variation among individuals. A two-stage model is used to analyze such data and a linear mixed model is a result from such a two-stage model. In this formulation, the probability distribution of serial measurements from the same subject has the same form for each individual but the parameters of the distribution vary over individuals, so that we can also view this as a subject-specific model.

Population parameters, individual effects, and within-subject variation are modelled in the first stage while between-subject variation in the second stage. Suppose \( y_i \) are the \( n_i \) serial observations from subject \( i \), \( \beta \) is a \( p \times 1 \) vector of known population parameters, \( b_i \) is a \( q \times 1 \) vector of unknown individual effects. \( X_i \) and \( Z_i \) are known design matrices linking \( \beta \) and \( b_i \) to \( y_i \), respectively.
Stage 1: For individual subject $i$,

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,$$  \hspace{1cm} (1-4)

where $\epsilon_i$ are independent from each other and usually assumed to have a multivariate normal distribution with zero mean vector and variance-covariance matrix $R_i$. $R_i$ is a positive definite matrix with dimension $n_i \times n_i$ and links to subject $i$ only in this way. Each $R_i$ is determined by the same population parameters. Now, both $\beta$ and $b_i$ are considered fixed.

Stage 2: Now $b_i$ are random variables independent from each other and from $\epsilon_i$, which is therefore called random effects. Here we assume $b_i \sim MVN(0, D)$.

Such two-stage models have the following favorable feature:

1. No requirement for balance in data;
2. Explicitly modelling and analysis of within- and between-individual variation;
3. The individual parameters have a natural explanation and therefore facilitate the exploratory analysis.

Unless in a Bayesian framework, the inference is based on least squares and maximum likelihood principles from the marginal distribution of $y_i$. From above modelling marginally

$$y_i \sim MVN(X_i\beta, Z_iDZ'_i + R_i).$$

However, if so, the hierarchical structure of the original model (1–4) is then not taken into account. Indeed, the marginal model is not equivalent to the original hierarchical model. Inferences based on marginal models do not explicitly assume the presence of random effects representing the heterogeneity between subjects.

The purpose of our proposed method is to find the significant statistical evidence of existing QTL which is reflected by the difference of average phenotypic curves. Estimates for random effects show the subject-specific profiles which
deviate from the overall average profile and are needed for prediction of subject-specific evolutions. So inference from the marginal distribution is enough for our study. The selection of variance-covariance structure is a nontrivial step in the model selection process, which will be further addressed in Chapter 3.

Several statistical packages are available for estimation and inference on all parameters in the marginal model. Among these, the most flexible commercially used one is the SAS procedure PROC MIXED. Littell et al. (2000) provided a nice example to how to apply frequently used statements and options in PROC MIXED. More detailed descriptions of all statements and options can be found in SAS manuals.

1.6 Motivation and Outline of the Dissertation

In Section 1.3, we know the basic idea of functional mapping is using a parametric form to model the phenotypic curve. However, in some cases there are many different functions that describe same phenotypic trajectory, for example, there are functions in 3 categories to describe a growth trajectory: exponential, saturating and sigmoidal (Von Bertalanffy (1957), Niklas (1994)). Thus, it may not be clear which one should be used, especially when there are not enough observations for each subject to show obvious characteristics. Moreover, in many situations, there are no obvious functional forms. Nonparametric functional mapping avoids such problems of the original functional mapping method while inheriting all its advantages. In this dissertation we propose a nonparametric functional mapping procedure for different situations.

In Chapter 2 we consider the dense-map situation which assumes that the QTL is on some marker. We use nonparametric function estimation, typically implemented with B-splines, to estimate the underlying functional form of phenotypic trajectories, and then construct a nonparametric test to find evidence of existing quantitative trait loci. Using the representation of a nonparametric regression as a
mixed model, we can easily derive a likelihood ratio test (LRT) statistic. The joint
distribution of LRT statistics at each putative locus (each marker) is derived, so
that P-values can be exactly calculated using an importance sampling method. In
total, one real poplar data set and two simulated data sets are used for application.

In Chapter 3 we develop a nonparametric functional interval mapping proce-
dure for the situation where the actual genes responsible for different underlying
phenotypic trajectories might not just be on a marker locus, which is more real-
istic. LRT statistics are calculated from a nonparametric mixture-mixed model.
The joint distribution of all calculated LRT statistics along a genome or linkage
group can be approximately obtained, but the resulting P-value is a lower bound
of the exact P-value. So a simulation procedure is proposed to calculate the exact
P-value. Analysis from simulated data sets is performed to show the operating
characteristics of the nonparametric functional interval mapping procedure. Two
ways to estimate the covariance matrix are examined through simulation studies
also.

In Chapter 4 we extend the nonparametric functional interval mapping
procedure to the missing genotypic data situation. The test statistic is obtained
from maximum likelihood principle and it critical value can also be determined
by a simulation approach. In this chapter, we also propose a Bayesian mapping
procedure to take into account of missing genotypic data.

In Chapter 5 we suggest to extend the mapping procedure introduced in
Chapter 4 to handle both the missing genotypic and phenotypic data situation.
CHAPTER 2
DENSE-MAP CASE: WHEN THE QTL IS ON A MARKER

In this chapter we consider the simplest case—the dense-map case, which assumes the genotype is known at every point in the genome, that is, the QTL locates on or very near to some marker locus. We use a subject specific model, integrating a functional relationship estimated by B-splines to illustrate the character process. The nonparametric test for evidence of existing QTL is derived from the maximum likelihood approach. The joint distribution of likelihood ratio test statistics at all putative loci is obtained exactly. Then we show how to directly calculate p-values from the joint distribution of likelihood ratio test statistics at every putative QTL locus under $H_0$. Applications to simulated and real data are provided.

2.1 Data Setting

We assume that we have information about both phenotypic values and genotypic values of $N$ subjects. The phenotypic values are from some underlying phenotypic curves, which are functions of some variable $t$. For example, this continuous variable $t$ could be age (corresponding to a growth curve), or body size (corresponding to an allometric law). In practice, we cannot measure continuous phenotypic values but only those values recorded at some fixed $t$ as a repeated measure, $\mathbf{t} = (t_1, t_2, \ldots, t_T)'$. For simplicity, we consider a backcross design from two contrasted homozygous inbred lines, that is, there are only two different phenotypic trajectories, $\mathbf{m}_1(\mathbf{t})$ and $\mathbf{m}_2(\mathbf{t})$. Also, we make the assumption that the QTL only appear at the marker loci. Using the marker information we can detect the major genes responsible for different phenotypic curves.
We consider a subject-specific model, for each subject $i, i = 1, \ldots, N,$

$$y_i(\vec{t}) = m(\vec{t}) + \alpha_i 1_T + \varepsilon_i,$$  \hspace{2cm} (2–1)

where $\vec{t} = (t_1, t_2, \ldots, t_T)'$, $\varepsilon_i$ is a parameter which accounts for the within subject covariance structure of the observations on subject $i$, $\alpha_i 1_T$ models the covariance structure of observations between subjects. The variables $\alpha_i$ and $\varepsilon_i$ are independently distributed as $N(0, \sigma^2)$ and $MVN(0, \tau^2 V_T)$, respectively. So now $y_i(\vec{t}) \sim MVN(m(\vec{t}), \Sigma)$, where $\Sigma = \sigma^2 J_T + \tau^2 V$. Model (2–1) is equivalent to the hierarchical model

$$y_i(\vec{t}) = \mu_i(\vec{t}) + \varepsilon_i$$

$$\mu_i(\vec{t}) = m_i(\vec{t}) + \alpha_i 1_T.$$

or the univariate model,

$$y_{ij} = \mu_j + \varepsilon_{ij},$$  \hspace{2cm} (2–2)

where $i = 1, \ldots, N, j = 1, \ldots, T, \mu_j = m(t_j)$ and $(\varepsilon_{i1}, \ldots, \varepsilon_{iT})' \sim MVN(0_T, \sigma^2 J_T + \tau^2 V)$.

The univariate model (2–2) can also be viewed as a random coefficient model that assumes the random deviation of a single curve with respect to population curve at different time points may not be the same. This characteristic will be determined by the covariance matrix. It is straightforward to see that the test for a difference between two phenotypic curves is a test of the hypothesis $H_0 : m_1 = m_2$ at each marker vs $H_1 : m_1 \neq m_2$ at some marker. We derive a maximum likelihood test for this hypothesis. We assume that we can use $l$ B-spline bases to estimate the underlying functional, where $l \leq T$ (This assures that the expression will be unique). So we can write $m = B\xi$, where $B$ is the basis matrix, and is the same for different functions.
2.2 The Likelihood Ratio Test at One Marker

First we consider the likelihood-ratio test on one particular marker when \( \Sigma = \sigma^2 I_T + \tau^2 V \) is known. Let \( l_1 \) and \( l_2 \) be the total number of subjects in each group, with \( l_1 + l_2 = N \). We assume

\[
m_1(\mathbf{t}) = B\xi_1 \quad \text{and} \quad m_2(\mathbf{t}) \equiv B\xi_2
\]

with

\[
y_i(\mathbf{t}) \sim MVN(B\xi_1, \Sigma), \quad i = 1, \ldots, l_1
\]
\[
y_i(\mathbf{t}) \sim MVN(B\xi_2, \Sigma), \quad i = l_1 + 1, \ldots, N.
\]

The likelihood function for \( \{\xi_1, \xi_2\} \) is

\[
L \propto \exp \left\{ -\frac{1}{2} \sum_{i=1}^{l_1} (y_i - B\xi_1)'\Sigma^{-1}(y_i - B\xi_1) - \frac{1}{2} \sum_{i=l_1+1}^{N} (y_i - B\xi_2)'\Sigma^{-1}(y_i - B\xi_2) \right\}.
\]

To construct the likelihood-ratio test for \( H'_0 \): \( m_1 = m_2 \) on one particular marker, we first need to find the MLEs of the unknown parameters. Denote the common coefficient vector under \( H'_0 \) by \( \xi \). After taking the first derivative of the \( \log(L) \) we get the likelihood equations for \( \xi \):

\[
\sum_{i=1}^{N} \Sigma^{-1}(y_i - B\xi) = 0.
\]

Now it is straightforward to see that the maximum likelihood estimate for \( \xi \) under \( H'_0 \) is \( \hat{\xi} = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\bar{y} \), where \( \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \). Similarly, under \( H_1 \), \( \hat{\xi}_i = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\bar{y}_i \), where \( \bar{y}_1 = \frac{1}{l_1} \sum_{j=1}^{l_1} y_j \) and \( \bar{y}_2 = \frac{1}{l_2} \sum_{j=l_1+1}^{N} y_j \).

The likelihood ratio test statistic is

\[
\lambda = \frac{\max_{H'_0} L}{\max L} = \frac{\exp \left\{ -\frac{1}{2} \sum_{i=1}^{N} (y_i - B\hat{\xi})'\Sigma^{-1}(y_i - B\hat{\xi}) \right\}}{\exp \left\{ -\frac{1}{2} \sum_{i=1}^{l_1} (y_i - B\hat{\xi}_1)'\Sigma^{-1}(y_i - B\hat{\xi}_1) - \frac{1}{2} \sum_{i=l_1+1}^{l_1+l_2} (y_i - B\hat{\xi}_2)'\Sigma^{-1}(y_i - B\hat{\xi}_2) \right\}}.
\]
The LRT rejects $H_0'$ when $\lambda$ is small, which is equivalent to rejecting when

$$\lambda' = \sum_{i=1}^{\tilde{N}} (y_i - B\hat{\xi})' \Sigma^{-1}(y_i - B\hat{\xi}) - \sum_{i=1}^{t_1} (y_i - B\hat{\xi}_1)' \Sigma^{-1}(y_i - B\hat{\xi}_1)$$

$$- \sum_{i=t_1+1}^{t_1+t_2} (y_i - B\hat{\xi}_2)' \Sigma^{-1}(y_i - B\hat{\xi}_2)$$

(2–3)

is large. A straightforward calculation (Section 2.5) shows that this is further equivalent to rejecting $H_0$ when

$$G = \frac{l_1 l_2}{N} (\bar{y}_1 - \bar{y}_2)' A_0 (\bar{y}_1 - \bar{y}_2)$$

(2–4)

is large, where $A_0 = \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1}$.

It is straightforward to see that $A_0 \Sigma = \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1}$ is idempotent, and

$$\bar{y}_1 \sim MVN(m_1, \frac{1}{l_1} \Sigma),$$

$$\bar{y}_2 \sim MVN(m_2, \frac{1}{l_2} \Sigma),$$

$$\bar{y}_1 - \bar{y}_2 \sim MVN(m_1 - m_2, (\frac{1}{l_1} + \frac{1}{l_2}) \Sigma),$$

and hence,

$$\frac{(\bar{y}_1 - \bar{y}_2)' A_0 (\bar{y}_1 - \bar{y}_2)}{\frac{1}{l_1} + \frac{1}{l_2}} \sim \chi^2_n \left( \frac{1}{2} (m_1 - m_2)' A_0 (m_1 - m_2) \right),$$

where $\chi^2_n$ is a noncentral chi squared random variable with degrees of freedom $n = rank(A_0) = rank((B' \Sigma^{-1} B)^{-1} (B' \Sigma^{-1} B)) = rank(B)$ and noncentrality parameter $(1/2)(m_1 - m_2)' A_0 (m_1 - m_2)$. Under $H_0$, $G \sim \chi^2_n$, a central chi squared random variable.

### 2.3 The Likelihood Ratio Test over All Markers

If we write $Y = (y_{11}, \cdots, y_{1T}, y_{21}, \cdots, y_{NT})'$ where the first $l_1 * T$ $y_{ij}s'$ are from group 1, then the test statistic could be written as a quadratic form $G = Y'AY$,
where
\[ A = \frac{l_1 l_2}{N} \left( \begin{array}{cc}
\frac{1}{l_1^2} J_{l_1} & -\frac{1}{l_1 l_2} 1_{l_1} \otimes 1_{l_2} \\
-\frac{1}{l_1 l_2} 1_{l_2} \otimes 1'_{l_1} & \frac{1}{l_2^2} J_{l_2}
\end{array} \right) \otimes A_0 \equiv U \otimes A_0 \quad (2-5) \]
and \( \text{var}(Y) = I_N \otimes \Sigma \equiv \tilde{\Sigma} \). Also, one can easily verify that \( A\tilde{\Sigma} \) is idempotent. At different markers, the test statistics have the same chi squared distribution but are correlated with each other.

Notice that we arrange \( Y \) according to which group \( y_{ij} \) is from, so at different markers, we will have a different vector \( Y \). To make notation uniform, we can use a permutation matrix \( P \). If we denote \( Y_1 \) as the \( Y \) for marker 1, then for other markers, \( k = 2, \ldots, m \),
\[ Y_k = P_k Y_1, \]
where \( P_k \) is the permutation matrix and the matrix (2–5) corresponding to \( k \)th marker is labelled as \( A_k \). Thus, at each marker \( i \), the corresponding test statistic is
\[ Y'_{k} A_k Y_k = Y'_1 P_k' A_k P_k Y_1. \]

Notice that, \( \text{var}(P_k Y_1) = P_k \tilde{\Sigma} P_k' = \tilde{\Sigma} \) because \( \tilde{\Sigma} \) is block diagonal with block size \( T \times T \) and \( P_k \) just interchanges the rows of \( \tilde{\Sigma} \). So under \( H_0 \) we still have
\[ Y'_{k} A_k Y_k = Y'_1 P'_k A_k P_k Y_1 \sim \chi_n^2. \]

Recall that under \( H_0 \), \( m(\bar{t}) = m_1(\bar{t}) = m_2(\bar{t}) \), \( Y_1 \sim MVN(1_N \otimes m(\bar{t}), I_N \otimes \Sigma) \equiv MVN(\mu, \Sigma) \) and we can write
\[
Y'_{k} A_k Y_k = Y'_k \tilde{\Sigma}^{-\frac{1}{2}} \tilde{\Sigma}^{\frac{1}{2}} A_k \tilde{\Sigma}^{\frac{1}{2}} \tilde{\Sigma}^{-\frac{1}{2}} Y_k \\
= Y'_k W_k \tilde{\Sigma}^{-\frac{1}{2}} W_k' \tilde{\Sigma}^{-\frac{1}{2}} Y_k \\
= Y'_1 P'_k \tilde{\Sigma}^{-\frac{1}{2}} W_k W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k Y_1 \\
\equiv Z'_k Z_k.
\]
where $Z_k = W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k Y_1 \sim^d MVN(W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu}, W_k' W_k)$ and $W_k W_k'$ are the spectral decompositions of $\tilde{\Sigma}^{-\frac{1}{2}} A_k \tilde{\Sigma}^{\frac{1}{2}}$. The matrix $W_k$ is composed of the eigenvectors corresponding to the $n$ non-zero eigenvalues, and it has $n$ orthonormal columns and is of order $NT \times n$. Also, under $H_0$ the structure of $A_k$ results in

$$
(W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu})' (W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu}) = \tilde{\mu}' A_k \tilde{\mu} = \tilde{\mu}' (U_k * 1_N) \otimes (A_0 * \bm{m}(\tilde{t})) = \tilde{\mu}' 0 \otimes (A_0 * \bm{m}(\tilde{t})) = 0.
$$

and thus, under $H_0$, $Z_{kn \times 1} = W_k' \tilde{\Sigma}^{-\frac{1}{2}} P_k Y_1 \sim MVN(0, I_n)$.

The entire vector $Z = (Z_1', Z_2', \ldots, Z_m')'$ has distribution

$$Z \sim MVN(\mu_Z, \Delta)$$

where

$$\mu_{Z_i} = W_i' \tilde{\Sigma}^{-\frac{1}{2}} \tilde{\mu}$$

and

$$\Delta = \begin{pmatrix} W_1' \tilde{\Sigma}^{-\frac{1}{2}} \\ W_2' \tilde{\Sigma}^{-\frac{1}{2}} P_2 \\ \vdots \\ W_m' \tilde{\Sigma}^{-\frac{1}{2}} P_m \end{pmatrix} \tilde{\Sigma} \begin{pmatrix} \tilde{\Sigma}^{-\frac{1}{2}} W_1 & P_2 \tilde{\Sigma}^{-\frac{1}{2}} W_2 & \cdots & P_m \tilde{\Sigma}^{-\frac{1}{2}} W_m \end{pmatrix},$$

and under $H_0$, $\mu_Z = 0$.

If we let $B_x$ denote the $n$-dimension ball with radius equal to $x$, then

$$P_0(\max_{1 \leq k \leq m} Y_k' A_k Y_k \leq x) = P_0(\max_{1 \leq k \leq m} Z_k' Z_k \leq x) = P_0(Z_1' Z_1 \leq x, \cdots, Z_m' Z_m \leq x) = \int \cdots \int_{\{Z_i \in B_x\}} \frac{1}{(2\pi)^{mn/2} |\Delta|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \sum_i Z_i' \Delta^{-1} Z_i \right\} dZ_1 \cdots dZ_m.$$
This probability, which is one minus the p-value for $H_0$, can be directly calculated by simulating $Z \sim MVN(0, \Delta)$ many times and counting how many $Z_i$'s fall in the $n$-dimensional ball $B_x$.

All the above derivations are made under the assumption that we know $\Sigma$, which is typically untrue in practice. We suggest substituting a REML estimate of the variance-covariance matrix $\hat{\Sigma}$ instead of $\Sigma$, and, in that case, the above formulae are correct asymptotically (Wolfinger 1993, Littell et.al 2000). (Please refer to Chapter 3 for more details.)

2.4 Examples

In this section we apply our procedure to both simulated and real data. We first use two simulated datasets to illustrate our method, and then analyze the Poplar data of Ma et al. (2002).

2.4.1 Simulation

The purpose of this simulation is to examine how the test statistic behaves as the underlying mean curves move apart from one another. We, of course, expect that the $p$-value gets smaller when we move the underlying curves apart.

The first data set assumes the underlying phenotypic curves are two flat lines. We choose these simple curves because it is easy to measure the distance between them, and to quantify the notion of the curves getting further apart.

The data set has 10 markers and 100 subjects, with measurements generated at 4 observation points using the variance-covariance matrix for each $\mathbf{y}_i$:

\[
\Sigma = \begin{pmatrix}
.35 & .33 & .32 & .31 \\
.33 & .35 & .33 & .32 \\
.32 & .33 & .35 & .33 \\
.31 & .32 & .33 & .35 \\
\end{pmatrix}.
\]

Based on the marker information, 100 data sets are generated and analyzed. Since the underlying curve are the two flat lines $y = a_0$ and $y = a_1$, we can use $|a_0 - a_1|$
Figure 2–1: Graph of the $p$−value and its 95% confidence interval for data from two underlying flat lines with different distances.

as a measure of distance. The $p$−value trend with respect to different distances between two underlying flat lines, and its 95% confidence interval, is given in Figure 2–1. It is clear that the $p$−value decreases as the curve moves apart, which is the behavior we would expect.

The second data set assumes one phenotypic curve is flat ($m_1(t) = 1$) and the other is quadratic ($m_2(t) = ax^2 + bx + c$). In this case, using eight markers, 50 subjects are generated at 12 observation points using the variance-covariance matrix for each $y_i$

\[
\Sigma = .1J_{12} + Autoregressive(\tau^2 = .05, rho = .5),
\]

where $J$ is a matrix with all ones and the structure of matrix $Autoregressive(\tau^2, rho)$ is shown as (1–3). Based on the same marker information, 100 data sets are generated and analyzed. To measure distance, we use the area between the two curves
over the observed time interval. Figure 2–2 shows the $p$–value profile for different distances and its 95% confidence interval. Again, the behavior is what we desire, with the $p$-value decreasing as the curves get further apart.

### 2.4.2 Poplar Data

This data set comes from an experiment of the triple hybridization of Populus (poplar). In the spring of 1988, a total of 450 1-year-old rooted three-way hybrid seedlings were planted at a forest farm near Xuchou City, Jiangsu Province, China. The female parent is a *Populus deltoides* clone and the male parent is an interspecific *P. deltoides × P. nigra* clone. At the end of each of the 11 growing seasons the total diameters were measured (the original data are plotted in Figure 2–3: The left figure shows the original diameter growth trend for every tree. The right one is the growth curve after taking logs.) . A genetic linkage map was constructed which comprises the 19 largest linkage groups for each parental map,
and represents roughly 19 pairs of chromosomes. We used our method to detect QTL affecting diameter growth on linkage group 10 from the *P. deltoides* parent map.

Because our model assumes that all *t* observations have equal variance, we first transformed the original observations by taking the log. Using our method with the variance-covariance structure autoregressive+simple (estimated by PROC MIXED as $\hat{\Sigma} = 0.001236 * J_{11} + \text{Autoregressive}(\tau^2 = 0.06155, \rho = 0.8945)$), we found strong evidence that there is a QTL at a marker on linkage group 10 which controls the growth trajectory of stem diameter in the interspecific hybrids of poplar ($p = .037$, $se = 4.6e - 4$). The biggest likelihood-ratio test statistics $G$ appears at Marker *CA/CCC-640R*. Figure 2–4 shows the estimated growth curves using B-splines for each group indexed by this marker. This finding is consistent with the result in Ma, Casella and Wu (2002).
Figure 2–4: The growth curve representing two groups of genotypes at marker \textit{CA/CCC-640R} on linkage group 10 in the \textit{Populus deltoides} parent map.
2.5 Technical Details: Derivation of the Likelihood Ratio Test

In what follows, we derive the likelihood-ratio test statistics from $\lambda'$ (2.3) to $G$ (2.6). Rewrite $\lambda'$ as

$$\lambda' = \sum_{i=1}^{l_1} (y_i - B\hat{\xi})'\Sigma^{-1}(y_i - B\hat{\xi}) - \sum_{i=1}^{l_1} (y_i - B\hat{\xi}_1)'\Sigma^{-1}(y_i - B\hat{\xi}_1)$$

$$+ \sum_{i=l_1+1}^{N} (y_i - B\hat{\xi})'\Sigma^{-1}(y_i - B\hat{\xi}) - \sum_{i=l_1+1}^{N} (y_i - B\hat{\xi}_2)'\Sigma^{-1}(y_i - B\hat{\xi}_2)$$

$$\equiv \lambda'_1 + \lambda'_2.$$

Notice that

$$\sum_{i=1}^{l_1} (y_i - B\hat{\xi})'\Sigma^{-1}(y_i - B\hat{\xi}) = \sum_{i=1}^{l_1} (y_i - \bar{y}_1)'\Sigma^{-1}(y_i - \bar{y}_1)$$

$$+ \sum_{i=1}^{l_1} (\bar{y}_1 - B\hat{\xi}_1)'\Sigma^{-1}(\bar{y}_1 - B\hat{\xi}_1)$$

$$+ \sum_{i=1}^{l_1} (B\hat{\xi}_1 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi})$$

$$+ 2 \sum_{i=1}^{l_1} (\bar{y}_1 - B\hat{\xi}_1)'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi}).$$

with

$$\sum_{i=1}^{l_1} (y_i - B\hat{\xi}_1)'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi}) = 0.$$

Thus

$$\sum_{i=1}^{l_1} (y_i - B\hat{\xi})'\Sigma^{-1}(y_i - B\hat{\xi})$$

$$= \sum_{i=1}^{l_1} (y_i - \bar{y}_1)'\Sigma^{-1}(y_i - \bar{y}_1)$$

$$+ \sum_{i=1}^{l_1} (\bar{y}_1 - B\hat{\xi}_1)'\Sigma^{-1}(\bar{y}_1 - B\hat{\xi}_1) + \sum_{i=1}^{l_1} (B\hat{\xi}_1 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi}).$$
Similarly, we can get

\[ \sum_{i=1}^{l_1} (y_i - B\hat{\xi}_1)'\Sigma^{-1}(y_i - B\hat{\xi}_1) = \sum_{i=1}^{l_1} (y_i - \bar{y})'\Sigma^{-1}(y_i - \bar{y}) + \sum_{i=1}^{l_1} (\bar{y} - B\hat{\xi}_1)\Sigma^{-1}(\bar{y} - B\hat{\xi}_1). \]

Therefore,

\[ \lambda'_1 = \sum_{i=1}^{l_1} (B\hat{\xi}_1 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi}) = l_1(B\hat{\xi}_1 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi}), \]

and

\[ \lambda'_2 = \sum_{i=l_1+1}^{l_1+l_2} (B\hat{\xi}_2 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_2 - B\hat{\xi}) = l_2(B\hat{\xi}_2 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_2 - B\hat{\xi}). \]

We then have,

\[ \lambda' = \lambda'_1 + \lambda'_2 = l_1(B\hat{\xi}_1 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_1 - B\hat{\xi}) + l_2(B\hat{\xi}_2 - B\hat{\xi})'\Sigma^{-1}(B\hat{\xi}_2 - B\hat{\xi}) \]

\[ = l_1(\bar{y}_1 - \bar{y})'A_0(\bar{y}_1 - \bar{y}) + l_2(\bar{y}_2 - \bar{y})'A_0(\bar{y}_2 - \bar{y}) \]

\[ = l_1(\bar{y}_1 - \frac{l_1\bar{y}_1 + l_2\bar{y}_2}{N})'A_0(\bar{y}_1 - \frac{l_1\bar{y}_1 + l_2\bar{y}_2}{N}) + l_2(\bar{y}_2 - \frac{l_1\bar{y}_1 + l_2\bar{y}_2}{N})'A_0(\bar{y}_2 - \frac{l_1\bar{y}_1 + l_2\bar{y}_2}{N}) \]

\[ = \frac{l_1l_2}{N}(\bar{y}_1 - \bar{y}_2)'A_0(\bar{y}_1 - \bar{y}_2), \]

where \( A_0 = \Sigma^{-1}B(B'S^{-1}B)^{-1}B'S^{-1}T' \).

Therefore, the likelihood ratio test rejects \( H_0 \) when

\[ G = \frac{l_1l_2}{N}(\bar{y}_1 - \bar{y}_2)'A_0(\bar{y}_1 - \bar{y}_2) \quad (2.6) \]

is large.
CHAPTER 3
NONPARAMETRIC FUNCTIONAL INTERVAL MAPPING

In Chapter 2, we introduced a nonparametric functional mapping procedure with the assumption that marker data is everywhere. However nature has such a complicated system that almost all markers in any type of linkage maps are not even genes. So we extend that theoretical framework further to this case in this chapter. The joint distribution of LRT statistics on each putative QTL locus can be obtained approximately. But the resulting P-value from this approximate joint distribution is a lower bound of the exact P-value. So a simulation procedure is proposed to calculate the exact P-value.

3.1 Data Setting

As before, we assume for each subject \( i, i = 1, ..., n \), that we have its information about both phenotypic and genotypic profiles. On every design point \( t_j \) (for example, time point for growth data), \( j = 1, ..., T \), we recorded \( y_{ij} \), the observed value of the underlying function at design point \( t_j \). To be simple, we only consider the population structure from a backcross design, which means there would be two groups with distinct phenotypes if a QTL does exist. Thus, those \( y \)'s are actually from two underlying phenotypic functionals \( m_1(\mathbf{t}) \) and \( m_2(\mathbf{t}) \). \( m_k(t), k = 1, 2 \) could be exactly expanded using \( l \) B-spline basis functions, \( l \leq T \). Let \( \mathbf{B} = \{\beta_j(t_i)\}_{T \times l} \), then \( m_k(\mathbf{t}) = \mathbf{B}\xi_k \). The null hypothesis is \( H_0 : m_1(t) = m_2(t) \) for any \( t \), i.e, \( H_0 : \xi_1 = \xi_2 \).

Denote the putative QTL by \( Q \) with allele \( Q \) and \( q \). Denote its outside markers by \( M \) with allele \( M, m \) and \( N \) with allele \( N, n \). The recombination rate between \( M \) and \( N \) is \( r \) and the recombination rate between \( M \) and \( Q \) is \( r_1 \). Let
\(\theta = \frac{n}{r}\). When \(\theta\) is relatively small, the conditional probabilities of genotypes at a QTL has been shown in Table 1–3. By the combination of two outside markers, arrange the observations \(y^r\)'s such that \(y_1, \cdots, y_n\) is from marker class \(Mm/Nn\), \(y_{n+1}, \cdots, y_{n+2}\) is from marker class \(Mm/nn\), \(y_{n+2+1}, \cdots, y_{n+3}\) is from marker class \(mm/Nn\), and \(y_{n+3+1}, \cdots, y_n\) is from marker class \(mm/nn\).

Again we consider a subject-specific model, for each subject \(i, i = 1, \ldots, n\)

\[y_i(\mathbf{t}) = \mathbf{m}(\mathbf{t}) + \alpha_i \mathbf{1}_T + \varepsilon_i, \tag{3–1}\]

where \(\mathbf{t} = (t_1, t_2, \cdots, t_T)'\). All notations are exactly as before. So now \(y_i(\mathbf{t}) \sim MVN(\mathbf{m}(\mathbf{t}), \Sigma)\), where \(\Sigma = \sigma^2 J_T + \tau^2 \mathbf{V}\).

Denote \(\|y_i\|^2 \equiv y_i^T \Sigma^{-1} y_i\) and \(\Phi(y; \mathbf{B} \xi, \Sigma) \equiv \exp\{-\frac{\|y_i - \mathbf{B} \xi\|^2}{2}\}\). Then under the above assumptions, the log-likelihood function for the unknown coefficient vectors \(\xi_1\) and \(\xi_2\) at each fixed \(\theta\) is

\[
\log L = \sum_{i=1}^{n_1} -\frac{\|y_i - \mathbf{B} \xi_1\|^2}{2} + \sum_{i=n_1+1}^{n_2} \log ((1 - \theta) \Phi(y_i; \mathbf{B} \xi_1, \Sigma) + \theta \Phi(y_i; \mathbf{B} \xi_2, \Sigma)) \\
+ \sum_{i=n_2+1}^{n_3} \log ((1 - \theta) \Phi(y_i; \mathbf{B} \xi_2, \Sigma) + \theta \Phi(y_i; \mathbf{B} \xi_1, \Sigma)) + \sum_{i=n_3+1}^{n} -\frac{\|y_i - \mathbf{B} \xi_2\|^2}{2} \\
+ \text{constant}.
\]

The likelihood equation for \(\xi_1\) is

\[
\frac{\partial \log L}{\partial \xi_1} = \sum_{i=1}^{n_1} \mathbf{B}^T \Sigma^{-1} (y_i - \mathbf{B} \xi_1) + \sum_{i=n_1+1}^{n_2} \frac{(1 - \theta) \Phi(y_i; \mathbf{B} \xi_1, \Sigma) \mathbf{B}^T \Sigma^{-1} (y_i - \mathbf{B} \xi_1)}{(1 - \theta) \Phi(y_i; \mathbf{B} \xi_1, \Sigma) + \theta \Phi(y_i; \mathbf{B} \xi_2, \Sigma)} \\
+ \sum_{i=n_2+1}^{n_3} \frac{\theta \Phi(y_i; \mathbf{B} \xi_1, \Sigma) \mathbf{B}^T \Sigma^{-1} (y_i - \mathbf{B} \xi_1)}{(1 - \theta) \Phi(y_i; \mathbf{B} \xi_2, \Sigma) + \theta \Phi(y_i; \mathbf{B} \xi_1, \Sigma)}.
\]

Denote

\[P(y; t) = \frac{(1 - t) \Phi(y; \mathbf{B} \xi_1, \Sigma)}{(1 - t) \Phi(y; \mathbf{B} \xi_1, \Sigma) + t \Phi(y; \mathbf{B} \xi_2, \Sigma)}.
\]
Then the likelihood equation for $\xi_1$ can be rewritten as
\[
\frac{\partial \log L}{\partial \xi_1} = \sum_{i=1}^{n_1} B'\Sigma^{-1}(y_i - B\xi_1) + \sum_{i=n_1+1}^{n_2} P(y_i; \theta)B'\Sigma^{-1}(y_i - B\xi_1) + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)B'\Sigma^{-1}(y_i - B\xi_1).
\]

Similarly, the likelihood equation for $\xi_2$ can be rewritten as
\[
\frac{\partial \log L}{\partial \xi_2} = \sum_{i=n_3+1}^{n} B'\Sigma^{-1}(y_i - B\xi_2) + \sum_{i=n_1+1}^{n_2} (1 - P(y_i; \theta))B'\Sigma^{-1}(y_i - B\xi_2) + \sum_{i=n_2+1}^{n_3} (1 - P(y_i; 1 - \theta))B'\Sigma^{-1}(y_i - B\xi_2).
\]

Thus, it is straightforward to have the following EM algorithm to numerically find the MLE of $\xi_1$ and $\xi_2$:

**EM Algorithm:** For fixed $\theta$ and known $\Sigma$, repeat the following steps until the convergence criterion is satisfied.

**Step k:** Calculate $P(y_i; 1 - \theta)^{(k)}$ and $P(y_i; \theta)^{(k)}$ using $\hat{\xi}_1^{(k)}$ and $\hat{\xi}_2^{(k)}$.

**Step k + 1:** Calculate
\[
\hat{\xi}_1^{k+1} = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\left(\frac{\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_2} P(y_i; \theta)y_i + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)y_i}{n_1 + \sum_{i=n_1+1}^{n_2} P(y_i; \theta) + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)}\right)
\]
and
\[
\hat{\xi}_2^{k+1} = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\left(\frac{\sum_{i=n_3+1}^{n} y_i + \sum_{i=n_1+1}^{n_2} (1 - P(y_i; \theta))y_i + \sum_{i=n_2+1}^{n_3} (1 - P(y_i; 1 - \theta))y_i}{n_4 - n_3 + \sum_{i=n_1+1}^{n_2} (1 - P(y_i; \theta)) + \sum_{i=n_2+1}^{n_3} (1 - P(y_i; 1 - \theta))}\right).
\]

When the QTL doesn’t exist, that is, under $H_0$, the MLE for the only unknown coefficient vector $\xi$ is $\hat{\xi}_0 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\sum_{i=1}^{n} y_i$. So the likelihood ratio
test statistics at each fixed $\theta$ is

$$-2\log \lambda = -2 \log \frac{\max_{H_0} L(\xi|y)}{\max L(\xi|y)}$$

$$= \sum_{i=1}^{n} ||y_i - B\hat{\xi}_0||^2_{\Sigma} - \sum_{i=1}^{n_1} ||y_i - B\hat{\xi}_1||^2_{\Sigma}$$

$$+ 2 \sum_{i=n_1+1}^{n_2} \log\{(1 - \theta) \exp\left(\frac{-||y_i - B\hat{\xi}_1||^2_{\Sigma}}{2}\right) + \theta \exp\left(\frac{-||y_i - B\hat{\xi}_2||^2_{\Sigma}}{2}\right)\}$$

$$+ 2 \sum_{i=n_2+1}^{n_3} \log\{\theta \exp\left(\frac{-||y_i - B\hat{\xi}_1||^2_{\Sigma}}{2}\right) + (1 - \theta) \exp\left(\frac{-||y_i - B\hat{\xi}_2||^2_{\Sigma}}{2}\right)\}$$

$$- \sum_{i=n_3+1}^{n} ||y_i - B\hat{\xi}_2||^2_{\Sigma}$$

According to classical interval mapping, a profile of likelihood ratio test statistics for an entire linkage map is constructed by calculating it at each putative $Q$ along the map, where the position can be characterized by outside marker $M, N$ and $\theta$. Usually the threshold value to reject $H_0$ is obtained through computationally intense ways: permutation or simulation. In the following sections, we will derive the joint distribution of the likelihood ratio test statistics along the entire linkage map under $H_0$ to calculate the $p-$value directly.

### 3.2 Approximate Distribution of Test Statistics at Each $Q$ under $H_0$

Note that $||y_i - B\hat{\xi}_0||^2_{\Sigma} - ||y_i - B\hat{\xi}_1||^2_{\Sigma} = ||y_i - B\hat{\xi}_1 + B\hat{\xi}_1 - B\hat{\xi}_0||^2_{\Sigma} - ||y_i - B\hat{\xi}_1||^2_{\Sigma} = (\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_1 - \hat{\xi}_0) + 2(\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_1)$. Thus we can rewrite

$$-2\log \lambda = \sum_{i=1}^{n_2} \{ (\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_1 - \hat{\xi}_0) + 2(\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_1) \}$$

$$+ \sum_{i=n_2+1}^{n} \{ (\hat{\xi}_2 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_2 - \hat{\xi}_0) + 2(\hat{\xi}_2 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_2) \}$$

$$+ 2 \sum_{i=n_1+1}^{n_2} \log\{(1 - \theta) + \theta \exp\left(\frac{||y_i - B\hat{\xi}_1||^2_{\Sigma}}{2} - \frac{||y_i - B\hat{\xi}_2||^2_{\Sigma}}{2}\right)\}$$

$$+ 2 \sum_{i=n_2+1}^{n_3} \log\{\theta \exp\left(\frac{||y_i - B\hat{\xi}_2||^2_{\Sigma}}{2} - \frac{||y_i - B\hat{\xi}_1||^2_{\Sigma}}{2}\right) + (1 - \theta)\}$$
When $-2 \log \lambda$ is big enough, $H_0$ can be rejected. By Jensen’s Inequality, we can show that

$$\log \{1 - \theta + \theta \exp(x)\} \geq \theta x.$$  

(A simple proof can be found in Section ??). Let $f(x) = \log \{1 - \theta + \theta \exp(x)\} - \theta x$, then a simple calculation shows that $f(x)$ increases when $x < 0$ and decreases when $x > 0$ with maximum value 0 at $x = 0$. So when $x$ is in the neighborhood of 0, it is reasonable to approximate $\log \{1 - \theta + \theta \exp(x)\}$ by $\theta x$.

Under $H_0: \xi_1 = \xi_2$, we expect $\hat{\xi}_1 \approx \hat{\xi}_2$ so that $\|y_i - B\hat{\xi}_1\|^2_\Sigma - \|y_i - B\hat{\xi}_2\|^2_\Sigma$ is around 0. Therefore, we can approximate the likelihood ratio test statistics $-2 \log \lambda$ with its lower bound

$$-2 \log \lambda \geq \sum_{i=1}^{n_2} \left\{ (\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_1 - \hat{\xi}_0) + 2(\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_1) \right\}$$

$$+ \sum_{i=n_1+1}^{n_2} \left\{ (\hat{\xi}_2 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_2 - \hat{\xi}_0) + 2(\hat{\xi}_2 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_2) \right\}$$

$$+ 2 \sum_{i=n_1+1}^{n_2} \theta \left( \frac{\|y_i - B\hat{\xi}_1\|^2_\Sigma}{2} - \frac{\|y_i - B\hat{\xi}_2\|^2_\Sigma}{2} \right)$$

$$+ 2 \sum_{i=n_2+1}^{n_3} \theta \left( \frac{\|y_i - B\hat{\xi}_2\|^2_\Sigma}{2} - \frac{\|y_i - B\hat{\xi}_1\|^2_\Sigma}{2} \right)$$

$$= \sum_{i=1}^{n_2} \left\{ (\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_1 - \hat{\xi}_0) + 2(\hat{\xi}_1 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_1) \right\}$$

$$+ \sum_{i=n_1+1}^{n_2} \left\{ (\hat{\xi}_2 - \hat{\xi}_0)'B'\Sigma^{-1}B(\hat{\xi}_2 - \hat{\xi}_0) + 2(\hat{\xi}_2 - \hat{\xi}_0)'B'\Sigma^{-1}(y_i - B\hat{\xi}_2) \right\}$$

$$+ \sum_{i=n_1+1}^{n_2} \theta \left\{ (\hat{\xi}_2 - \hat{\xi}_1)'B'\Sigma^{-1}B(\hat{\xi}_2 - \hat{\xi}_1) + 2(\hat{\xi}_2 - \hat{\xi}_1)'B'\Sigma^{-1}(y_i - B\hat{\xi}_2) \right\}$$

$$+ \sum_{i=n_2+1}^{n_3} \theta \left\{ (\hat{\xi}_1 - \hat{\xi}_2)'B'\Sigma^{-1}B(\hat{\xi}_1 - \hat{\xi}_2) + 2(\hat{\xi}_1 - \hat{\xi}_2)'B'\Sigma^{-1}(y_i - B\hat{\xi}_1) \right\}$$

Denote this lower bound by $G$, which is a “good” approximation to $-2 \log \lambda$ under $H_0$. Hence it’s reasonable to use the distribution of $G$ under $H_0$ to approximate the distribution of $-2 \log \lambda$ under $H_0$. 

Under $H_0$, $G$ turns to be a quadratic form $Y'AY$, where $Y$ is a $nT \times 1$ vector with $y_1, ..., y_n$ stacking together and details about $A$ can be found in Section ??.

Furthermore, the matrix $A$ has $\text{rank}(B) = l$ nonzero equal eigenvalues $\beta$. Under $H_0$, $Y$ is distributed as $\text{MVN}(\mathbf{1}_n \otimes \mathbf{B}\xi, I_n \otimes \Sigma) \equiv \text{MVN}(\mu, \Sigma)$, it is straightforward to show that $G \sim \beta \chi^2_l$ (details in Section ??). Thus the distribution of $-2 \log \lambda$ under $H_0$ is approximated by this multiple of a central Chi-square distribution.

Technical Details:

1. **Proof of** $\log\{1 - \theta + \theta \exp(x)\} \geq \theta x$.

   **Proof:** Let a discrete random variable $Y$ with $P(Y = x) = 1 - P(Y = 0) = \theta$, then $E(e^Y) = 1 - \theta + \theta \exp(x)$. Because $\log(.)$ is concave, directly following Jensen’s Inequality we can get $\log\{E(e^Y)\} \geq E\{\log(e^Y)\} = E(Y)$, that is, $\log\{1 - \theta + \theta \exp(x)\} \geq \theta x$.

2. **Distribution of** $G$ **under** $H_0$ *Rewrite $G$ as*

   
   
   $$
   G = n_2(\hat{\xi}_2 - \hat{\xi}_0)' \mathbf{B}' \Sigma^{-1} \mathbf{B}(\hat{\xi}_2 - \hat{\xi}_0) 
   + (n - n_2)(\hat{\xi}_2 - \hat{\xi}_0)' \mathbf{B}' \Sigma^{-1} \mathbf{B}(\hat{\xi}_2 - \hat{\xi}_0) 
   + \theta(n_2 - n_1)(\hat{\xi}_2 - \hat{\xi}_1)' \mathbf{B}' \Sigma^{-1} \mathbf{B}(\hat{\xi}_2 - \hat{\xi}_1) 
   + \theta(n_3 - n_2)(\hat{\xi}_1 - \hat{\xi}_2)' \mathbf{B}' \Sigma^{-1} \mathbf{B}(\hat{\xi}_1 - \hat{\xi}_2) 
   + 2 \sum_{i=1}^{n_2} (\hat{\xi}_1 - \hat{\xi}_0)' \mathbf{B}' \Sigma^{-1} (y_i - \mathbf{B}\hat{\xi}_1) 
   + 2 \sum_{i=n_2+1}^{n} (\hat{\xi}_2 - \hat{\xi}_0)' \mathbf{B}' \Sigma^{-1} (y_i - \mathbf{B}\hat{\xi}_2) 
   + 2\theta \sum_{i=n_1+1}^{n_2} (\hat{\xi}_2 - \hat{\xi}_1)' \mathbf{B}' \Sigma^{-1} (y_i - \mathbf{B}\hat{\xi}_2) 
   + 2\theta \sum_{i=n_2+1}^{n_3} (\hat{\xi}_1 - \hat{\xi}_2)' \mathbf{B}' \Sigma^{-1} (y_i - \mathbf{B}\hat{\xi}_1).
   $$

   


Under $H_0$, the MLE for underlying coefficient vector
\[
\hat{\xi}_0 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\frac{\sum_{i=1}^{n}y_i}{n}
\]
\[
= (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}(1_n \otimes \frac{1}{n} I_T)Y
\]
\[
\equiv (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}(w_0 \otimes I_T)Y
\]
\[
\equiv (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_0 Y,
\]
where $Y$ is a $NT \times 1$ vector with $y_1, ..., y_n$ stacking together.

Under $H_0 : \xi_1 = \xi_2$ everywhere, after enough steps, $P(\cdot; \theta)$ in the EM algorithm should be equal to $1 - \theta$. That is,
\[
\hat{\xi}_1 = (B'\Sigma^{-1}B)^{-1}B'\Sigma \left( \frac{\sum_{i=1}^{n_1} y_i + (1 - \theta) \sum_{i=n_1+1}^{n_2} y_i + \theta \sum_{i=n_2+1}^{n_3} y_i}{n_2 + (n_3 - n_2 - n_2 + n_1)\theta} \right)
\]
\[
\equiv w_1 \otimes I_T
\]
\[
\hat{\xi}_2 = (B'\Sigma^{-1}B)^{-1}B'\Sigma \left( \frac{\sum_{i=n_3+1}^{n} y_i + \theta \sum_{i=n_1+1}^{n_2} y_i + (1 - \theta) \sum_{i=n_2+1}^{n_3} y_i}{n - n^2 - (n_3 - n_2 - n_2 + n_1)\theta} \right)
\]
which now can be represented by a weighted sum of $y_i's$.

Denote $\hat{\xi}_1 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_1(Y) Y$ and $\hat{\xi}_2 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_2(Y) Y$, where
\[
W_1(Y) = 
\left( 1_{n_1} \quad 1 - \theta \quad ... \quad 1 - \theta \quad \theta \quad ... \quad \theta \quad 0_{n - n_3} \right) \otimes \frac{I_T}{n_2 + (n_3 - n_2 - n_2 + n_1)\theta}
\]
\[
\equiv w_1 \otimes I_T
\]
and
\[
W_2(Y) = 
\left( 0_{n_1} \quad \theta \quad ... \quad \theta \quad 1 - \theta \quad ... \quad 1 - \theta \quad 1_{n - n_3} \right) \otimes \frac{I_T}{n - n^2 - (n_3 - n_2 - n_2 + n_1)\theta}
\]
\[
\equiv w_2 \otimes I_T.
\]
Now we can rewrite $G$ as

$$G = n_2 Y'(W_1 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_1 - W_0) Y$$

$$+ (n - n_2) Y'(W_2 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_2 - W_0) Y$$

$$+ \theta(n_2 - n_1) Y'(W_2 - W_1)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_2 - W_1) Y$$

$$+ \theta(n_3 - n_2) Y'(W_1 - W_2)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_1 - W_2) Y$$

$$+ 2Y'(W_1 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{V_1 - n_2 W_1\} Y$$

$$+ 2Y'(W_2 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{V_2 - (n - n_2) W_2\} Y$$

$$+ 2\theta Y'(W_2 - W_1)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{V_3 - (n_2 - n_1) W_2\} Y$$

$$+ 2\theta Y'(W_1 - W_2)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{V_4 - (n_3 - n_2) W_1\} Y,$$

Where

$$V_1 = \begin{pmatrix} 1_{n_2} & 0_{n-n_2} \end{pmatrix} \otimes I_T \equiv v_1 \otimes I_T$$

$$V_2 = \begin{pmatrix} 0_{n_2} & 1_{n-n_2} \end{pmatrix} \otimes I_T \equiv v_2 \otimes I_T$$

$$V_3 = \begin{pmatrix} 0_{n_1} & 1_{n_2-n_1} & 0_{n-n_2} \end{pmatrix} \otimes I_T \equiv v_3 \otimes I_T$$

$$V_4 = \begin{pmatrix} 0_{n_2} & 1_{n_3-n_2} & 0_{n-n_3} \end{pmatrix} \otimes I_T \equiv v_4 \otimes I_T$$

Obviously, $G$ can be expressed as quadratic form $Y' A Y$, where under $H_0$, $Y$ is distributed as $MVN(1_T \otimes B \xi, I_n \otimes \Sigma) \equiv MVN(\overrightarrow{\mu}, \Sigma)$ and $A = (A_1 + A_1')/2$ with

$$A_1 = n_2(W_1 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_1 - W_0)$$

$$+ (n - n_2)(W_2 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_2 - W_0)$$

$$+ \theta(n_2 - n_1)(W_2 - W_1)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_2 - W_1)$$

$$+ \theta(n_3 - n_2)(W_1 - W_2)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} (W_1 - W_2)$$

$$+ 2(W_1 - W_0)' \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{V_1 - n_2 W_1\}$$
\[ + 2(W_2 - W_0)^T \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{ V_2 - (n - n_2)W_2 \} \\
+ 2\theta(W_2 - W_1)^T \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{ V_3 - (n_2 - n_1)W_2 \} \\
+ 2\theta(W_1 - W_2)^T \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \{ V_4 - (n_3 - n_2)W_1 \} \\
= [n_2(w_1 - w_0)'(w_1 - w_0) + (n - n_2)(w_2 - w_0)'(w_2 - w_0) \\
+ \theta(n_2 - n_1)(w_2 - w_1)'(w_2 - w_1) + \theta(n_3 - n_2)(w_1 - w_2)'(w_1 - w_2) \\
+ 2(w_1 - w_0)'(v_1 - n_2w_1) + 2(w_2 - w_0)'(v_2 - (n - n_2)w_2) \\
+ 2\theta(w_2 - w_1)'(v_3 - (n_2 - n_1)w_2) + 2\theta(w_1 - w_2)'(v_4 - (n_3 - n_2)w_1)] \\
\otimes \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \\
\equiv A_2 \otimes \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \\
\]

Since \( Y \sim MVN(\overrightarrow{\mu}, \tilde{\Sigma}) \), the quadratic form \( Y'AY \) can be expressed as
\[
Y'AY = \sum_{i=1}^{(k)} \beta_i W_i, \quad \text{where} \quad W_i \sim \chi^2_{m_i}(\pi_i). \quad \beta_i, \ i = 1, \ldots, k \text{ are those nonzero distinct eigenvalues of } A\tilde{\Sigma}(or \overline{\Sigma^2A\overline{\Sigma}^2}), \quad \text{where multiplicity for each eigenvalue } \lambda_i \text{ is } m_i. \quad \text{Notice that } A\overline{\Sigma} = \frac{1}{2}(A_2 + A_2') \otimes \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \Sigma^{-1} \Sigma = \frac{1}{2}(A_2 + A_2') \otimes \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B'. \quad \text{Because } \Sigma^{-1} B (B' \Sigma^{-1} B)^{-1} B' \text{ is idempotent with rank } l \text{ and } rank(\frac{1}{2}(A_2 + A_2')) = rank(A_2) = 1 \text{ due to the structure of } w_0 \text{ and } w_1, \ A\overline{\Sigma} \text{ has } l \text{ nonzero equal eigenvalues. Thus } k = 1. \text{ Also, } \pi \text{ is the noncentrality parameter } \\
\theta = \frac{1}{2}\overrightarrow{\mu}' \overline{\Sigma}^{\frac{1}{2}} PP' \overline{\Sigma}^{\frac{1}{2}} \overrightarrow{\mu}, \quad \text{where } P \text{ consists of } l \text{ eigenvectors corresponding to } \beta \text{ satisfying } \beta PP' = \overline{\Sigma}^{\frac{1}{2}} A\overline{\Sigma}^{\frac{1}{2}}. \text{ Note that} \\
\theta = \frac{1}{2\beta} \overrightarrow{\mu}' \overline{\Sigma}^{\frac{1}{2}} (\beta PP') \overline{\Sigma}^{\frac{1}{2}} \overrightarrow{\mu} \\
= \frac{1}{2\beta} \overrightarrow{\mu}' A\overrightarrow{\mu} = 0 \\
\]

Thus \( G \sim \beta \chi^2_l \) follows. Furthermore, \( \beta > 0 \) because for all nonzero vectors \( z \), the quadratic form \( z'\overline{\Sigma}^{\frac{1}{2}} A\overline{\Sigma}^{\frac{1}{2}} z \) can finally be written as summation of nonnegative value \( z_i' \Sigma z_i^* \) since \( \Sigma \) is positive definite.
3.3 Precision of the Linear Approximation to LRT Statistics under $H_0$

When the marker map is not dense enough, it is more realistic to adopt the idea of interval mapping, which uses the two flanking markers to locate the QTL position. At each fixed putative QTL position indexed by $\theta$ within the $k$th marker pair, $1 \leq k \leq m$, the actual likelihood ratio test statistic $-2\log \lambda_k^\theta$ has a lower bound $G_k^\theta$, which can be written as a quadratic form $Y_k^A_k\lambda_k^\theta Y_k$. Thus we use the joint distribution of these $G_k^\theta$s at all putative QTL positions to estimate the joint distribution of those actual likelihood ratio test statistics and hence again have a direct way to get P-value estimate. Because $G_k^\theta$ is always lower than $-2\log \lambda_k^\theta$, the resulting P-value is lower than actual P-value. But we believe that under $H_0$, our approximated P-value is “near” enough to the precise P-value. That is,

**Conjecture:** Under $H_0$, i.e., when no QTL exists,

$$P_0(\max_{1 \leq k \leq m; \theta} -2\log \lambda_k^\theta \leq x) \rightarrow P_0(\max_{1 \leq k \leq m; \theta} Y_k^A_k\lambda_k^\theta Y_k \leq x)$$

in probability, when the total number of subjects, $n$, goes to infinity.

Firstly we have the following lemmas which will be part of those steps to prove this tentative theorem.

**Lemma 1:** Under $H_0$, the maximum likelihood estimate of the unknown coefficient vector $\xi$, for smoother matrix $B$, $\hat{\xi}$, is $\sqrt{n}$-consistent for $\xi$ and hence $\hat{\xi} - \xi = o_p(n^{-\rho})$, with $0 < \rho < \frac{1}{2}$.

**Proof:** The $\sqrt{n}$-consistency of $\hat{\xi}$ directly follows from the fact that $\sqrt{n}(|\hat{\xi} - \xi|)$ is asymptotically normal with mean zero and covariance matrix $[I(\xi)]^{-1}$. (See Lehmann and Casella (1998).)

For any $\epsilon > 0$, by the definition of $\sqrt{n}$-consistency, there exists a constant $M_\epsilon$, so that we have

$$P(\sqrt{n}|\hat{\xi}_k - \xi_k| > M_\epsilon) < \epsilon$$

for each $k$, $1 \leq k \leq \text{length}(\xi)$. 

Obviously, for any $\rho$ with $0 < \rho < \frac{1}{2}$,

$$P(n^\rho|\hat{\xi}_k - \xi_k| = \frac{n^{\frac{1}{2}}|\hat{\xi}_k - \xi_k|}{n^{\frac{1}{2} - \rho}} > \frac{M_\epsilon}{n^{\frac{1}{2} - \rho}}) < \epsilon.$$ 

Obviously, when $n \to \infty$, $\frac{M_\epsilon}{n^{\frac{1}{2} - \rho}} \to 0$. So $\hat{\xi} - \xi = o_p(n^{-\rho})$, with $0 < \rho < \frac{1}{2}$. $\triangle$

**Lemma 2:** Under $H_0$, each term in the difference between the true LRT statistics and its approximate contributed by subject $i$,

$$\log\{(1 - \theta) + \theta \exp\left(\frac{||y_i - B\hat{\xi}_1||_2^2}{2} - \frac{||y_i - B\hat{\xi}_2||_2^2}{2}\right)\} - \theta\left(\frac{||y_i - B\hat{\xi}_1||_2^2}{2} - \frac{||y_i - B\hat{\xi}_2||_2^2}{2}\right) \to 0$$

almost surely in $\theta \in [0, 1]$ when the number of subjects, $n \to \infty$. And the convergence rate is $o_p(n^{-\gamma})$ with $1 > \gamma > 0$.

**Proof:** Under $H_0$ with putative QTL position index $\theta$ fixed, we have $\hat{\xi}_1 \to^a.s. \xi_0$ and $\hat{\xi}_2 \to^a.s. \xi_0$. So if we let $x = \frac{||y_i - B\hat{\xi}_1||_2^2}{2} - \frac{||y_i - B\hat{\xi}_2||_2^2}{2}$, then under $H_0$, $x \to^a.s. 0$.

Because $f(x) = \log[1 - \theta + \theta \exp(x)] - \theta x$ is a continuous function, we have $f(x) \to^a.s. 0$ as $x \to^a.s. 0$.

By Taylor series expansion, we have

$$e^x = 1 + x + x_1^2$$

$$\log(1 + x) = x - x_2^2$$

where $|x_1| < |x|$ and $|x_2| < |x|$. Thus,

$$f(x) = \log(1 - \theta + \theta x + \theta x_1^2) - \theta x$$

$$= \log(1 + \theta x + \theta x_1^2) - \theta x$$

$$= \theta x + \theta x_1^2 - x_2^2 - \theta x$$

$$= \theta x_1^2 - x_2^2 \leq \theta x^2$$
where $|x_2| < |\theta x + \theta x_1^2|$. From Lemma 1, the convergence rates for $\hat{\xi}_1$ and $\hat{\xi}_2$ to $\xi$ are both $n^{-\rho}$ with $0 < \rho < 1/2$. That is,

$$n^\rho(\hat{\xi}_1 - \xi) \rightarrow^p 0$$

$$n^\rho(\hat{\xi}_2 - \xi) \rightarrow^p 0.$$

So we have

$$n^{2\rho}x^2 = n^{2\rho}\left[2(\hat{\xi}_2 - \hat{\xi}_1)'B'\Sigma^{-1}y_i + |B\hat{\xi}_1|_2^2 - |B\hat{\xi}_2|_2^2\right]^2$$

$$= \left[2n^\rho(\hat{\xi}_2 - \hat{\xi}_1)'B'\Sigma^{-1}y_i + n^\rho\xi_1'B'\Sigma^{-1}B\hat{\xi}_1 - n^\rho\xi_2'B'\Sigma^{-1}B\hat{\xi}_2\right]^2$$

$$\rightarrow^p \left[2(\xi - \xi)'B'\Sigma^{-1}y_i + n^\rho\xi_1'B'\Sigma^{-1}B\xi - n^\rho\xi_2'B'\Sigma^{-1}B\xi\right]^2 = 0.$$

Therefore, the convergence rate for $x^2$ is $n^{-\gamma} = n^{-2\rho}$. That is, $0 < \gamma < 1$. $\triangle$

Notice that

$$-2 \log \lambda_k^\theta - G_k^\theta =$$

$$2 \sum_{i=n_1+1}^{n_2} \{\log[(1 - \theta) + \theta \exp(\frac{||y_i - B\hat{\xi}_1||_2^2}{2} - \frac{||y_i - B\hat{\xi}_2||_2^2}{2})]$$

$$- \theta(\frac{||y_i - B\hat{\xi}_1||_2^2}{2} - \frac{||y_i - B\hat{\xi}_2||_2^2}{2})\}$$

$$+ 2 \sum_{i=n_2+1}^{n_3} \{\log[(1 - \theta) + \theta \exp(\frac{||y_i - B\hat{\xi}_2||_2^2}{2} - \frac{||y_i - B\hat{\xi}_1||_2^2}{2})]$$

$$- \theta(\frac{||y_i - B\hat{\xi}_2||_2^2}{2} - \frac{||y_i - B\hat{\xi}_1||_2^2}{2})\}$$

$$\equiv \sum_{i=n_1+1}^{n_3} S_i(\leq \sum_{i=n_1+1}^{n_3} \theta x_i^2 = \theta X'X)$$

That is, the difference between the actual LRT statistics $-2 \log \lambda_k^\theta$ and our linear approximation $G_k^\theta$ is a summation of many terms, $S_i$. From these two lemmas we can see that for each $S_i$ goes to 0 when the number of subjects, $n$ gets bigger and bigger. But the number of those terms goes to infinity when the total number of subjects goes to infinity. Is the $S_i$ goes to 0 faster than the summation of all $S_i$? A
simple simulation shows the answer is heartbreaking “NO”, which is implied in the following two figures: Figure 3–1 and Figure 3–2. In the upper part of Figure 3–1 shows the wiggly pattern of the summation of all $S_i$ when the sample size increases, that is, $\sum S_i$ does NOT go to zero even though it is not big. The wiggly pattern again in the upper part of Figure 3–2 resulting from bigger sample sizes than those in Figure 3–1 confirms this. The lower parts in both Figure 3–1 and Figure 3–2 suggest $S_i$ does converge to zero when sample size goes to infinity.

So how about a better approximation in the sense of including higher order terms when we use Taylor approximations to actual LRT statistics $-2 \log \lambda_k^\theta$? That is, how about approximating $f_1(x) = 2 \log(1 + \theta(\exp(x) - 1))$ by $f_2(x) = 2\theta(x + x^2/2) - \theta^2(x + x^2/2)^2 + 2\theta^3(x + x^2/2)^3/3$. We use the simplest setting that all data are from standard normal distributions where the recombination rate between two markers is .2 and $\theta = .5$. Figure 3–3 plots the difference between $f_1(x)$ and $f_2(x)$. We can see that when $x$ is in the neighbor of 0, the difference between two functions is also around 0. Figure 3–4 and Figure 3–5 show a simple simulation result about the summation of the differences, which is the gap between actual LRT and its following approximation, $G^\theta$. Now the gap is small enough with expression:

$$-2 \log \lambda^\theta - G^\theta = \sum_{i=n_1+1}^{n_3} \left\{ 2 \log[(1 - \theta) + \theta \exp(x_i)] - 2\theta(x_i + x_i^2/2) + \theta^2(x_i + x_i^2/2)^2 - 2\theta^3(x_i + x_i^2/2)^3/3 \right\}$$

where $x_i = \frac{\|y_i - B\xi_1\|^2}{2} - \frac{\|y_i - B\xi_2\|^2}{2}$ for $i = n_1 + 1, \ldots, n_2$ and $x_i = \frac{\|y_i - B\xi_2\|^2}{2} - \frac{\|y_i - B\xi_1\|^2}{2}$ for $i = n_2 + 1, \ldots, n_3$.

Figure 3–6 shows the behavior of the other approximation, which also makes the gap goes to 0 visually when $n \to \infty$. But it is hard to find the joint distribution
Figure 3–1: The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1<i<n_3} S_i \ast 10^5$ when number of subjects $n$ increases under simplest setting that all data from standard normal distribution. Recombination rate between two markers is .2 and $\theta = .5$. 
Figure 3–2: The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1<i<n_3} S_i$ when number of subjects $n$ increases under simplest setting that all data from standard normal distribution. Recombination rate between two markers is .2 and $\theta = .5$. 
Figure 3-3: The plot of function $f(x) = f_1(x) - f_2(x)$. 
Figure 3-4: The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1 < i < n_3} |S_i|$ when number of subjects $n$ increases.

Figure 3-5: The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1 < i < n_3} |S_i|$ when number of subjects $n$ increases.
Figure 3-6: The plot of $\sum_{i=n_1+1}^{n_3} S_i$ and $\max_{n_1+1<i<n_3} |S_i|$ when number of subjects $n$ increases using approximation $f_2(x)' = 2\theta(x + x^2/2) - \theta^2(x + x^2/2)^2$ of the approximated test statistics or if we can, very time-consuming. So we have to turn to other methods to find the exact P-value, which is stated in the following section.

### 3.4 Alternative Approach to Calculate P-value: Simulation

Because we finally find that the linear approximation to the actual likelihood ratio test statistic does not have good/acceptable properties, we have to turn to other methods to decide the threshold showing significant evidence of existing QTL. Notice that when $H_0$ is actually true, after enough steps in EM algorithm, the maximum likelihood estimate of two unknown coefficient vectors, $\xi_1$ and $\xi_2$ can be directly written as linear combinations of realized phenotypic vectors. That is, using $Y$ as a $NT \times 1$ vector with $y_1, ..., y_n$ stacking together we can have
\[ \hat{\xi}_1 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_1(Y)Y \] and \[ \hat{\xi}_2 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_2(Y)Y, \] where

\[
W_1(Y) = \left(1_{n_1}, 1 - \theta, \ldots, 1 - \theta, \theta, \ldots, \theta, 0_{n-n_3}\right) \otimes \frac{I_{T_n}}{n_2 + (n_3 - n_2 - n_2 + n_1)^2} \equiv w_1 \otimes I_T
\]

and

\[
W_2(Y) = \left(0_{n_1}, \theta, \ldots, \theta, 1 - \theta, \ldots, 1 - \theta, 1_{n-n_3}\right) \otimes \frac{I_{T_n}}{n - n_2 - (n_3 - n_2 - n_2 + n_1)^2} \equiv w_2 \otimes I_T.
\]

Of course, the MLE of \( \xi_0 \) under \( H_0 \) can also be expressed as a linear combination of \( Y \),

\[
\hat{\xi}_0 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\frac{1}{n} \sum_{i=1}^{n} y_i \equiv (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}(w_0 \otimes I_T)Y \equiv (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_0Y.
\]

We can have following simulation procedure which can determine the cutoff point for detecting QTL from a genome/linkage group:

1. Generate 1000 different data sets from \( H_0 \), that is, having same variance-covariance structure and based on same known marker information. We can easily generate \( y_i - B\hat{\xi}_0, y_i - B\hat{\xi}_1, y_i - B\hat{\xi}_2 \), \( i = 1, \ldots, n \), for each fixed \( \theta \).

2. For each interval calculate the likelihood ratio test statistics at these fixed \( \theta \) based on simulated data; Obtain 1000 different maximum likelihood ratio test statistics through the whole genome(linkage group).
Figure 3–7: Two simulated distributions of likelihood ratio test statistic based on same 1000 datasets.

3. Get the cutoff point value from the simulated distribution of maximum of LRT statistics.

To check the validity of this simulation procedure, we get the distribution of maximum of LRT statistics calculated using MLEs from EM algorithm. Figure 3–7 shows the two histograms resulted from each procedure, which confirms the validity of this simulation procedure. In that figure, under $H_0$, the underlying phenotypic curve is a flat line. 100 subjects are measured at 4 different points. Marker interval is 25 CM and scanning unit is 2 CM. To be simple, we only assume this linkage group has 3 markers, that is, two marker intervals.

By treating $\theta$ as unknown parameter and finding its MLE, we need only calculate one likelihood ratio test statistic for each marker interval, which is already
the maximum LRT statistic for that marker interval. The log-likelihood function for the unknown coefficient vectors $\xi_1, \xi_2$ and $\theta$ is

$$
\log L = \sum_{i=1}^{n_1} \frac{\|y_i - B\xi_1\|^2}{2} + \sum_{i=n_1+1}^{n_2} \log \{(1 - \theta)\Phi(y_i; B\xi_1, \Sigma) + \theta\Phi(y_i; B\xi_2, \Sigma)\}
+ \sum_{i=n_2+1}^{n_3} \log \{(1 - \theta)\Phi(y_i; B\xi_2, \Sigma) + \theta\Phi(y_i; B\xi_1, \Sigma)\} + \sum_{i=n_3+1}^{n} -\frac{\|y_i - B\xi_2\|^2}{2}
+ \text{constant}.
$$

(3–2)

The likelihood equations for $\xi_1$ and $\xi_2$ have been derived before. And using a function defined as

$$
P(y; t) = \frac{(1-t)\Phi(y; B\xi_1, \Sigma)}{(1-t)\Phi(y; B\xi_1, \Sigma) + t\Phi(y; B\xi_2, \Sigma)},
$$

we can easily write out the EM algorithm for finding $\hat{\xi}_1$ and $\hat{\xi}_2$ when $\theta$ is fixed.

Now when $\theta$ is also an unknown parameter, its likelihood equation is

$$
\frac{\partial \log L}{\partial \theta} = \sum_{i=n_1+1}^{n_2} \frac{\Phi(y_i; B\xi_2, \Sigma) - \Phi(y_i; B\xi_1, \Sigma)}{(1 - \theta)\Phi(y_i; B\xi_1, \Sigma) + \theta\Phi(y_i; B\xi_2, \Sigma)}
+ \sum_{i=n_2+1}^{n_3} \frac{\Phi(y_i; B\xi_1, \Sigma) - \Phi(y_i; B\xi_2, \Sigma)}{(1 - \theta)\Phi(y_i; B\xi_2, \Sigma) + \theta\Phi(y_i; B\xi_1, \Sigma)}
= \sum_{i=n_1+1}^{n_2} \left(1 - \frac{P(y_i; \theta)}{\theta} - \frac{P(y_i; \theta)}{1 - \theta}\right)
+ \sum_{i=n_2+1}^{n_3} \left(P(y_i; 1 - \theta) - \frac{1 - P(y_i; 1 - \theta)}{1 - \theta}\right).
$$

Thus, it is straightforward to have this following EM algorithm to numerically find the MLE of $\xi_1, \xi_2$ and $\theta$ simultaneously:

**EM Algorithm 2**: Suppose $\Sigma$ is known. Repeat the following steps until the convergence criterion is satisfied.

**Step k**: Calculate $P(y_i; 1 - \theta)^{(k)}$ and $P(y_i; \theta)^{(k)}$ using $\hat{\xi}_1^{(k)}, \hat{\xi}_2^{(k)}$ and $\hat{\theta}^{(k)}$. 
Step $k + 1$: Calculate

\[
\hat{\xi}_{1}^{(k+1)} = (B^\prime \Sigma^{-1} B)^{-1} B^\prime \Sigma^{-1} \left( \frac{\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_2} P(y_i; \theta)^{(k)} y_i + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)^{(k)} y_i}{n_1 + \sum_{i=n_1+1}^{n_2} P(y_i; \theta)^{(k)} + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)^{(k)}} \right)
\]

\[
\hat{\xi}_{2}^{(k+1)} = (B^\prime \Sigma^{-1} B)^{-1} B^\prime \Sigma^{-1} \left( \frac{\sum_{i=n_3+1}^{n} y_i + \sum_{i=n_1+1}^{n_2} [1 - P(y_i; \theta)^{(k)}] y_i + \sum_{i=n_2+1}^{n_3} [1 - P(y_i; 1 - \theta)^{(k)}] y_i}{n_4 - n_1 - \sum_{i=n_1+1}^{n_2} P(y_i; \theta)^{(k)} - \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)^{(k)}} \right)
\]

\[
\hat{\theta}^{(k+1)} = \frac{n_2 - n_1 + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)^{(k)} - \sum_{i=n_1+1}^{n_2} P(y_i; \theta)^{(k)}}{n_3 - n_1}
\]

Now the simulation procedure to determine the cutoff point of testing if there is a QTL existing in this linkage group (genome) is stated below:

1. Generate 1000 different data sets from $H_0$.
2. For each dataset and each interval, find the MLEs of $\theta$ and hence calculate the maximum of LRT statistic among this interval by easily generating $y_i - B \hat{\xi}_0, y_i - B \hat{\xi}_1, y_i - B \hat{\xi}_2, i = 1, ..., n$.
3. For each dataset, move to each of the left intervals and repeat step 2. Obtain 1000 different maximum LRT statistics through the whole genome (linkage group).
4. Get the cutoff point value from the simulated distribution of maximum of LRT statistics.

Figure 3-8 shows two empirical distributions of the genome-wise maximum LRT statistic. Under $H_0$, the underlying phenotypic curve is a flat line. 100 subjects are measured at 4 different points. Marker interval is 25 CM. To be simple, we only assume this linkage group has 3 markers, that is, two marker intervals. The upper histogram is from the EM algorithm with fixed $\theta$ and the lower histogram is from the simulation procedure with $\hat{\theta}$ from the second EM algorithm. The 95% and 99% cutoff points from the upper procedure are 7.986
Figure 3-8: Two simulated distributions based on same 1000 datasets when treating $\theta$ as unknown parameter.
Figure 3-9: Two simulated distributions based on different 1000 datasets while treating $\theta$ as unknown parameter.

and 11.818, respectively. And the corresponding values from the lower procedure are 7.944 and 11.802. These similar numbers with the similar distribution shapes suggest the correctness of above simulation procedure.

**Remark:**

1. When we generate 1000 data sets from $H_0$, the underlying phenotypic curve under $H_0$ can be randomly picked including $\mu_0 = 0$. Figure 3-9 shows two empirical distributions from generating datasets from two different phenotypic curves.

2. When $H_0$ is true, the EM algorithm for finding MLE of $\theta$ converges slow. But when $H_0$ is not true, the convergence is very fast. So I think for the simulation procedure we can just use fixed $\theta$ which is computationally less time-consuming.
3.5 Estimation of Consistent Variance-Covariance Matrix with Small Sample Size

In a general Gaussian linear model for \( n \) observations \( \mathbf{Y}, \mathbf{Y} \sim N(\mathbf{X}\beta, \Sigma) \), the precision of inference for an unknown vector \( \beta \), which is usually the main interest of investigators, depends on the estimation of the unknown variance-covariance matrix \( \Sigma \). So how to estimate the unknown elements in \( \Sigma \) attracts a lot of attention. We are principally concerned with the situation where the data consist of a collection of independent sets from different subjects, that is, where \( \Sigma \) will be block-diagonal (assume each block is the same, denoted by \( V \)). Obviously this type of model is appropriate for repeated measurements experiments.

Restricted (residual) maximum likelihood (REML) has been well established after being introduced by Patterson and Thompson (1971) to estimate the unknown parameter in \( \Sigma \). The advantage of REML over a standard maximum likelihood procedure is that REML takes account of the loss in degrees of freedom that results from estimating fixed effects and it leads to unbiased estimators for a balanced dataset. The REML estimators are defined as the maximum likelihood estimator (MLE) from the marginal likelihood of a linear transformed set of data \( \mathbf{Z} = A\mathbf{Y} \) such that the distribution of \( \mathbf{Z} \) does not depend on \( \mathbf{Y} \). It turns out that the marginal likelihood of \( \mathbf{Z} \) does not depend on the choices of \( A \) and can be expressed as (assume \( \Theta \) is the vector of unknown parameters in \( \Sigma \))

\[
-2\log L(\Theta) = \log(|\Sigma|) + \log(|\mathbf{X}'\Sigma^{-1}\mathbf{X}|) + \mathbf{Y}'\{\Sigma^{-1} - \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\} \mathbf{Y}.
\]

The REML of \( \Theta \) is the vector, denoted by \( \tilde{\Theta} \), which maximizes the above log-likelihood function, whereas the standard MLE, \( \hat{\Theta} \), maximizes this log-likelihood function

\[
-2\log L(\Theta) = \log(|\Sigma|) + \mathbf{Y}'\{\Sigma^{-1} - \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\} \mathbf{Y}.
\]
The only difference between this two log-likelihood function is the term \( \log\{|X'\Sigma^{-1}X|\} \).

Also, it is straightforward to see the form of the design matrix \( X \) must be correctly specified, otherwise we may not even get consistent estimators for \( \Theta \). So Diggle, Liang and Zeger (1994) recommended to use a saturating model for the mean response profiles to get a guaranteed consistent estimate of variance-covariance matrix.

For experiments where observed values from different subjects are measured at same set of observation points, a robust estimate for \( V \) can be obtained using the REML principle. Suppose for each of \( m_i \) experiment units in \( i \)th of \( g \) experimental treatment units, measurements are made at each of \( T \) observation-points \( t_k \). The complete set of measurements are

\[
y_{ijk}, \ i=1,\ldots,g, \ j=1,\ldots,m_i, \ k=1,\ldots,T.
\]

The saturated model for mean response is \( E(y_{ijk}) = \mu_{ik} \). Thus, we could use the ordinary least-squares fit to get a consistent estimate of \( V \):

\[
\hat{V} = \left( \sum_{i=1}^{g} m_i - g \right)^{-1} \sum_{i=1}^{g} \sum_{j=1}^{m_i} (y_{ij} - \hat{\mu}_i)(y_{ij} - \hat{\mu}_i)',
\]

with \( \hat{\mu}_i = (m_i^{-1} \sum_{j=1}^{m_i} y_{ij1}, \ldots, m_i^{-1} \sum_{j=1}^{m_i} y_{ijT})' \). This approach can be extended to the situation when measurements times are not common to all units by estimating each \( V_i \) in the block diagonal variance-covariance matrix \( \Sigma \). This approach fails for the extreme case where sets of measurements times are essentially unique for each unit.

However, explicit modelling of the covariance structure has been popularly considered for these following reasons. Firstly, when the true covariance matrix has many fewer parameters than the unconstrained variance matrix, the estimate can be made more accurately. Secondly, the objection to estimating \( \frac{1}{2}T(T+1) \) parameters in the covariance matrix gains force when \( T \), the number of observations per
experimental unit is large. Thirdly, this robust approach uses replications across experiment units to estimate the covariance matrix non-parametrically. When there are a lot of missing values or sets of measurements times are essentially unique for all units, a parametric modelling of covariance structure can avoid these problems. Many statistical softwares provides general linear model fitting with a variety of parameterized covariance structures. Littell et.al (2000) provided a tutorial of using the MIXED procedure in SAS, which can model a rich selection of covariance structures through the RANDOM and REPEATED statements.

One can choose the best fitting one using Akaike’s Information Criterion (AIC) or Bayesian Information Criterion (BIC). The formulae for these two criteria are

\[
\begin{align*}
AIC &= L(\Theta) - q \\
BIC &= L(\Theta) - (q/2) \log(N^*)
\end{align*}
\]

where \(L(\Theta)\) is the maximized log-likelihood or restricted log-likelihood, \(q\) is the number of parameters in the covariance matrix, \(p\) is the number of fixed parameters and \(N^*\) is the number of subjects: \(N\) for ML and \(N - p\) for REML. Models with larger AIC or BIC yield a better fit. The difference between these two criterion is that BIC puts more penalty on the number of parameters used to fit covariance matrix.

When the covariance stationarity assumption is likely to fail, one can use some transformation to stabilize the dependence structure. Zimmerman and Núñez-Antón (2001) proposed a structured antedependence (SAD) model to estimate such a nonstationary covariance matrix directly and showed such a model displays many favorable properties. The variances on each observation point are not assumed to be constant in this model, and the correlation between measurements equidistant in time are not assumed equal.
Either as a guide to the formulation of a parametric model or not depending on parametric assumption, non-parametric estimation of covariance structure has also been proposed by a lot of authors. For example, Diggle and Verbyla (1998) proposed to use kernel weighted local linear regression smoothing of sample variogram ordinates and of squared residuals to get a nonparametric estimator for the covariance structure.

For our nonparametric functional mapping procedure, we need find a consistent estimator for the underlying covariance structure that behaves well when the sample size is small. It is known that the sample covariance matrix is guaranteed to a consistent estimator, however it can be very unstable with smallest estimated eigenvalues being too small while largest estimated eigenvalues being too large, especially for small sample sizes. Obviously standard estimators, like REML or ML can gain some stability in estimating the matrix in small samples by assuming some parametric covariance structure that involves estimation of fewer parameters. But such estimators are only consistent when the hypothesized structures are correct. No asymptotical properties of nonparametric estimators are studied by any authors.

Daniels and Kass (1999) provided a better way to estimate the covariance matrix by specifying an appropriate prior for the covariance matrix and choosing an estimator based on a particular loss function. Their method first generically stabilizes an unstructured estimate and then shrinks such estimates toward a parsimonious, structured form of the matrix. How much shrinkage required is decided by the data. Daniels and Kass (2001) extended this method to more easily calculated estimators without using a fully MCMC approach. They called such estimators empirical (or approximate) Bayes estimators. The final estimator includes a combination of shrinking the eigenvalues and then shrinking toward structure. The data decides the amount of shrinkage. These estimators induce
stability over the unstructured estimator of the covariance matrix while providing robustness to misspecification of the structure. These estimators are consistent.

From their simulation study to evaluate the risk in estimating a covariance matrix, Daniels and Kass recommended first shrinking the eigenvalues of the unstructured estimator by replacing it with the Stein estimator and then shrinking the Stein estimator toward a structure using a structured log eigenvalue or correlation shrinkage estimator. This is because when the structure is far from correct, the correlation shrinkage estimator is worse than the structured log eigenvalue estimator. So we will adopt such an estimating procedure into our functional mapping approach:

**Step 1:** Find an unstructured covariance matrix estimator, $\hat{\Sigma}_{un}$ from a saturating model.

**Step 2:** Shrink the eigenvalues of $\hat{\Sigma}_{un}$ to obtain a more stable estimator, Stein’s estimator, $\hat{\Sigma}_{st} = O\Lambda(\hat{\lambda})O^T$, where $O$ is the matrix of normalized eigenvectors, $\hat{\lambda}$ is the vector of sample eigenvalue, and $\Lambda(\hat{\lambda}) = \text{diag}(\lambda_1^*(\lambda), \ldots, \lambda_p^*(\lambda))$ with

$$
\lambda_j^*(\lambda) = n\hat{\lambda}_j/\alpha_j
$$

$$
\alpha_j = n - p + 1 + 2\hat{\lambda}\sum_{i\neq j} \frac{1}{\lambda_j - \lambda_i}
$$

**Step 3:** Fit the saturating model assuming some covariance structures, using AIC or BIC to choose one. Get estimator $\hat{\Sigma}_s$. This step can be easily handled by SAS, proc MIXED.

**Step 4:** Determine the amount of shrinkage $\tau^2$ using $\hat{\Sigma}_{st}$ and $\hat{\Sigma}_s$.

$$
\hat{\tau}^2 = \frac{p}{2} \sum_{i=1}^p (\log(\hat{\lambda}_{i,st}) - \log(\hat{\lambda}_{i,s})/p)^2/(p + 4) - 2/n.
$$

This estimator corresponds to the posterior mode under a uniform shrinkage prior on $\tau^2$, $\pi(\tau) = (2/n)/(2/n + \tau^2)^2$.  

Step 5: Shrink $\hat{\Sigma}_{st}$ toward chosen a structure using structured log eigenvalue estimator $\hat{\Sigma}_{sh} = Odiag\{\hat{\lambda}\}O^T$, which is the final estimator. This step shrinks the log of eigenvalues again toward the structure.

$$\hat{\lambda}_i = \exp\left(\frac{2/n_2}{2/n + \hat{\tau}^2} \sum \log(\hat{\lambda}_{i,s})/p + \frac{\hat{\tau}^2}{2/n + \hat{\tau}^2} \log(\hat{\lambda}_{i,st})\right)$$

This estimator corresponds to the posterior mean of $\lambda_i$ conditional on estimates $\log(\hat{\lambda})$ and $\hat{\tau}^2$. The prior distribution on the logarithm of eigenvalues $\log(\lambda_i), i = 1, ..., p$, is normal: $\log(\lambda_i)|\tau^2 \sim i.i.d N(\log(\lambda), \tau^2)$. The asymptotic distribution of the logarithm of the sample eigenvalues is $\log(\hat{\lambda}_i) \sim N(\log(\lambda_i), 2/n)$. We use $\sum \log(\hat{\lambda}_{i,s})/p$ to estimate $\log(\lambda)$.

3.6 Application to Poplar Data - Revisited

3.6.1 A Chromosome-wise Analysis on Linkage Group 10

We apply this method to the same Poplar Data set as before, where 61 plants all have information from 8 markers of linkage group 10 and 11 observations from the same set of measurement time. To compare with previous analysis using nonparametric functional interval mapping with linear approximation of LRT statistics, we first log-transformed the original observations. We use order 3 B-splines because we need not to have smooth second derivative functions for the underlying phenotypic curves. And the measurement time is at each end of continuous 11 years. So we use equidistant inner knots at $(1, 3, 5, 7, 9, 11)$ to fit the growth curve. As we know in Section 1.4, we need two more knots at each end of the inner knots vector. Thus, we actually use 7 splines to fit the growth curve.

Procedure to find the consistent estimate of $\Sigma$: We will first get a consistent variance-covariance matrix from a saturating model, then using this estimate to proceed following functional interval mapping approach. This estimating procedure of $\hat{\Sigma}$ is adopted from Daniels and Kass(2001).
Step 1: Find the unstructured covariance matrix estimator \( \hat{\Sigma}_{un} \) from a saturating model, \( y_i = \mu_k + \varepsilon_i \), \( i = 1, \ldots, n \) and \( \text{var}(\varepsilon_i) = \Sigma \). \( k = 1, \ldots, 4 \) according to 4 possible combinations from the first two markers. \( \hat{\Sigma}_{un} \) has 11 distinct eigenvalues: 0.328, 0.086, 0.057, 0.011, 0.005, 0.001, 0.0009, 0.0006, 0.0002, 0.0001, 6.6e − 05.

Step 2: Shrink \( \hat{\Sigma}_{un} \) to its Stein’s estimator \( \hat{\Sigma}_{st} \), whose eigenvalues are 0.276, 0.072, 0.055, 0.010, 0.005, 0.0011, 0.0010, 0.0007, 0.0002, 0.00017, 8.3e − 05.

Step 3: Fitting the saturating model assuming some structure. We used the previous estimator having structure autoregressive+simple chosen by SAS procedure PROC MIXED, that is, \( \hat{\Sigma}_s = 0.001236*J_{11} + \text{Autoregressive}(\tau^2 = 0.06155, \rho = 0.8945) \). This estimator has eigenvalues: 0.485, 0.107, 0.037, 0.018, 0.011, 0.008, 0.006, 0.005, 0.004, 0.0037, 0.0035.

Step 4: Determine the amount of shrinkage \( \tau^2 = 6.533 \) using \( \hat{\Sigma}_{st} \) and \( \hat{\Sigma}_s \).

Step 5: Shrink \( \hat{\Sigma}_{st} \) using \( \tau^2 \) to get final estimate \( \hat{\Sigma}_{sh} \), whose eigenvalues now are 0.272, 0.071, 0.054, 0.010, 0.0054, 0.0011, 0.0010, 0.0008, 0.00024, 0.00018, 8.6e − 05.

From the eigenvalues of each variance-covariance estimator we can see the final estimator is closer to the unstructured variance-covariance estimator. Using this empirical Bayes variance-covariance estimator, we find the maximum of LRT appears on the first interval with \( \hat{\theta} = .123 \) (about 3cM from the first marker CA/CCC-640R). The maximum value of \(-2 \log \lambda\) is 15.487 with \( p-value = .143 \). The cutoff point for \( \alpha \)-level at .1 is 16.84. (18.99 for \( \alpha = .05 \) and 23.10 for \( \alpha = .01 \).) Therefore, we could not make a conclusion that a QTL governing stem growth exists in linkage group 10 if we used EB estimator.

If we use the REML estimate of the underlying variance-covariance with the same structure in Section 3.10.2, and use the simulation procedure to find the cutoff points instead of using the approximate joint distribution of all LRT statistics, the p-value is 0 with maximum LRT value 62.79 at \( \hat{\theta} = .32 \) (about 8.5cM
from the first marker CA/CCC-640R). The .01 \( \alpha \)-level cutoff point from simulation is 23.72. (17.26 for \( \alpha = .1 \) and 19.26 for \( \alpha = .05 \).) This finding are consistent with previous analysis (\( p \)-value= \( 5e \) – 4 with standard deviation \( 5e \) – 4).

In Figure 3–10 the black lines are from \( \hat{\Sigma}_{sh} \) while the dashed lines are from \( \hat{\Sigma}_s \). Figure 3–10 shows that the fitted curve does not change much due to different \( \hat{\Sigma} \) but the significance of the difference between two curves caused by the putative QTL depends on how to estimate \( \hat{\Sigma} \).

3.6.2 A Genome-wise Analysis Using REML Estimate

For the Poplar data a genetic linkage map was constructed which comprises the 19 largest linkage groups for each parental map, and represents roughly 19 pairs of chromosomes. AR(1) repeated measurement errors are assumed to model the within subject correlation and a log-transformation is applied to the raw data of
Table 3–1: Analysis of Poplar data using nonparametric functional interval mapping (NPFIM) and parametric functional interval mapping (PFIM). “EB”, the empirical Bayes estimator. “REML”, the REML estimator corresponding to structure $\sigma^2 J + Autoregressive(\tau^2, \rho)$.

<table>
<thead>
<tr>
<th>Variance Estimate</th>
<th>NPFIM</th>
<th>PFIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>REML</td>
<td>&lt;.001(62.69)</td>
<td>&lt;.001(37.06468)</td>
</tr>
<tr>
<td>EB</td>
<td>.143(15.487)</td>
<td>&gt;.5(.858)</td>
</tr>
</tbody>
</table>

stem diameters to stabilize the age-dependent variance heteroscedasticity (Wu et.al 2004). Figure 3–11 shows the likelihood maps resulting from our nonparametric functional mapping approach. The empirical estimate of the critical value is obtained from 1000 simulation tests and we find the threshold value for declaring the genome-wise existence of a QTL is 32.01 at the significance level $P = 0.01$. The QTL candidate positions are the positions corresponding to the peaks of curves higher than the critical value. From Figure 2, there is significant evidence showing that several QTLs exist in linkage group 1, 2, 4, 7, 10, 14 & 18.

3.7 Simulation Study for Nonparametric Functional Interval Mapping

Table 3–1 shows the analysis result when using nonparametric functional interval mapping (NPFIM) and parametric functional interval mapping (PFIM) with two different variance-covariance estimators, respectively. The first number in each cell is the P-value and the number in parenthesis is the maximum LRT statistic along the linkage group 10.

Since the true underlying variance-covariance structure for the Poplar dataset is unknown and the QTL location is also unknown, it is inconclusive regarding the behavior of NPFIM and the behavior of the consistent estimator using procedures introduced in Section 3.5. So we use the following simulated datasets to explore the characteristics of our NPFIM procedure. We use the marker information of linkage group 10 in the Poplar experiment and the true QTL exists in the first marker interval, 8cM away from the first marker. Growth profiles of 61 subjects are from
Figure 3–11: The profile of the likelihood ratio test statistics between the full and reduced (no QTL) subject-specific model for the diameter growth trajectories across the whole *Populus deltoides* parent map. The genomic positions corresponding to the peak of the curve are the MLEs of the QTL localization.
Table 3–2: P-values and its standard deviation of nonparametric functional interval mapping (NPFIM) using different variance-covariance estimators under different true Σs. “EB” means the shrinkage estimator which is guaranteed to be a consistent estimator. “REMLw” estimator is the REML estimator corresponding to a wrong structure. “REML” estimator is the one selected from SAS Proc MIXED using BIC and assuming each subject has different underlying growth curve. “True” estimator is of course the matrix we actually used to generate data.

<table>
<thead>
<tr>
<th>Variance Estimate</th>
<th>NPFIM</th>
<th>Σ₁</th>
<th>Σ₂</th>
<th>Σ₃</th>
<th>Σ₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>EB</td>
<td></td>
<td>.0864(.0108)</td>
<td>.288(.0223)</td>
<td>.384(.0279)</td>
<td>.458(.0293)</td>
</tr>
<tr>
<td>REMLw</td>
<td></td>
<td>.00789(.0023)</td>
<td>.0687(.0099)</td>
<td>.143(.0165)</td>
<td>.329(.0213)</td>
</tr>
<tr>
<td>REML</td>
<td></td>
<td>.0011(.00044)</td>
<td>.103(.0151)</td>
<td>.209(.0238)</td>
<td>.497(.0274)</td>
</tr>
<tr>
<td>True</td>
<td></td>
<td>.00045(.00034)</td>
<td>.0727(.0125)</td>
<td>.197(.0240)</td>
<td>.413(.0285)</td>
</tr>
</tbody>
</table>

The numbers in parentheses are the sampling errors of the P-values.

Two underlying functions: \( \frac{20}{1 + 20 \exp(-.6 \times t)} \) and \( \frac{30}{1 + 27 \exp(-.9 \times t)} \).

Four randomly chosen different variance-covariance matrixes for log-transformed data are

\[
\begin{align*}
\Sigma_1 &= 0.1 \times J_{11} + Autoregressive(\tau^2 = 0.06, \rho = 0.8) \\
\Sigma_2 &= 0.1 \times J_{11} + Autoregressive(\tau^2 = 0.1, \rho = 0.8) \\
\Sigma_3 &= 0.3 \times J_{11} + Autoregressive(\tau^2 = 0.1, \rho = 0.8) \\
\Sigma_4 &= 1 \times J_{11} + Autoregressive(\tau^2 = 0.6, \rho = 0.8),
\end{align*}
\]

where \( J_{11} \) is a square matrix with all ones with dimension 11 and \( Autoregressive(\tau^2, \rho) \) is defined as (1–3). Heritability curves corresponding to these four covariance structures are plotted in Figure 3–12.

Table 3–2 shows the analysis outcome of NPFIM using different variance-covariance structures. From this table, we can see

- The guaranteed consistent estimator is rather conservative. The final empirical Bayes covariance matrix estimator from the procedure stated in Section 3.5 merely shrinks to the true structure a little bit, that is, it is not very stable because of more similarity to unstructured estimators.
Figure 3–12: Heritability curves for the four simulated data sets using different $\Sigma$s.
Table 3–3: P-values and its standard deviation of nonparametric functional interval mapping (NPFIM) using different variance-covariance estimators for different sample size and $\Sigma = \Sigma_3$. “EB” means the shrinkage estimator which is guaranteed to be a consistent estimator. “REMLw” estimator is the REML estimator corresponding to a wrong structure. “REML” estimator is the one selected from SAS Proc MIXED using BIC and assuming each subject has different underlying growth curve. “True” estimator is of course the matrix we actually used to generate data.

<table>
<thead>
<tr>
<th>Variance Estimate</th>
<th>NPFIM $N = 61$</th>
<th>NPFIM $N = 200$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$EB$</td>
<td>.384(.0279)</td>
<td>.0239(.00787)</td>
</tr>
<tr>
<td>$REMLw$</td>
<td>.143(.0165)</td>
<td>.00165(.00056)</td>
</tr>
<tr>
<td>$REML$</td>
<td>.209(.0238)</td>
<td>.00105(.00064)</td>
</tr>
<tr>
<td>$True$</td>
<td>.197(.0240)</td>
<td>.00095(.00058)</td>
</tr>
</tbody>
</table>

The numbers in parentheses are the sampling errors of the P-values.

- When the structure is correct, REML works fine even when the sample size is not very big. The performance of the REML estimate from wrong structure is worse than the performance of REML estimate corresponding to true structure when the phenotypic curves show little variation from the mean curve like the first column in Table 3–2. But when the variation is big, “REMLw” estimator works better than “REML” estimate in terms of smaller P-value. This may due to the small sample size and the variance-covariance structure we chose to simulate the data set.

We also do the same NPFIM analysis for data from underlying variance-covariance matrix $\Sigma_3$ but the sample size has been increased to 200. Table 3–3 lists the change of P-values when the sample size increases. When the sample size increases, because the underlying phenotypic curves are indeed different, the smaller P-value is exactly what we expected. With larger sample sizes, the EB estimate works fine. We can also see that when the sample size is big the REML estimate of true structure outperforms the REML estimate of the wrong structure in terms of smaller P-values.
3.8 Further Evaluation of Empirical Bayes Estimate of Covariance Matrix

From the simulation studies in Section 3.7, we recommend to use the REML estimate selected using AIC or BIC from a saturating model instead of the empirical Bayes (EB) estimate even though it is a guaranteed consistent estimate. Because empirical Bayes estimate is essentially obtained from shrinking an unstructured REML estimate toward some structure, it is rather conservative in the sense that large sample sizes are required to make the EB estimate behaves like a true covariance matrix. The following simulation studies are conducted to further study the performance of the Empirical Bayes estimate.

The first simulation study uses the 61 subjects’ marker information of linkage group 10 in the Poplar data set. The underlying functions are two logistic growth curves: $\frac{20}{1+20e^{-6t}}$ and $\frac{30}{1+27e^{-9t}}$, where $t = 1, \ldots, 11$. Autoregressive correlation is assumed for any two observations. The covariance matrix is determined by letting the heritability on year 4 (The genetic variance is the biggest one in this year.) equal to (.15, .2, .25, .3, .35, .4, .45, .5, .55, .6), respectively. The heritability curves across all 11 years are shown in the left part of Figure 3–13. 100 datasets for each heritability value are generated to perform the nonparametric functional interval mapping (NPFIM) procedure. The average P-values are shown in the right part of Figure 3–13 corresponding to each heritability value. This figure tells us that EB performs better when heritability is bigger. But, in reality, we have no information about how big the heritability is.

The second simulation study also uses the 61 subjects’ marker information of linkage group 10 in the Poplar data set as genotypic data. The two underlying biological trajectories are from the HIV dynamics mechanism, which have double exponential forms $\exp(12 - .7 * t) + \exp(7.5 - .05 * t)$ and $\exp(11 - .4 * t) + \exp(5 - .03 * t)$. The two mean curves are illustrated in Figure 3–14. Assume there are 20
Figure 3-13: The left plot shows the 10 heritability curves of simulated growth data sets. The right plot shows the trend of P-values when heritability on year 4 increases. “EB” and “REML” means the Empirical Bayes estimate and REML estimate of covariance matrix, respectively. “True” represents the covariance matrix used to generate data.
observation point. The covariance matrix is randomly generated without putting any known structure. Figure 3-15 plots the heritability value on each time points. 100 datasets with 200 subjects are generated to do the NPFIM analysis. NPFIM analysis is also conducted for a sub-dataset containing 61 subjects randomly selected from each datasets. The best structure picked by SAS Proc Mixed is autoregressive moving average structure, ARMA(1,1). The result is in Table 3-4. For easy comparison Table 3-3 is also put in this table. From this table we can conclude that

- When the sample size increases, the P-value gets smaller and so does the standard deviation regardless of any covariance matrix estimate.
- If the true covariance matrix has some structure such as autoregressive, Toeplitz and so on, the REML estimate usually outperforms the empirical Bayes estimate, as suggested by the result from the growth data set.
- If the true covariance matrix is actually unstructured, the empirical Bayes estimate is better than the REML estimate.

However, when analyzing a real data sets, there is typically some pattern in the correlations among repeated measurements/longitudinal data. So the REML estimate is still recommended even though the empirical Bayes estimate performs well when the sample size is big.

3.9 Discussion

Given these simulation results in Section 3.7, REML estimates work better than guaranteed consistent empirical Bayes estimates. So, in practice, we recommend using the best REML estimates selected using the AIC or BIC criterion with a saturating model.

When we tried to analyze the simulated data using traditional functional interval mapping method, we found it is very computationally intensive because of the nonlinear nature of the functional forms. For example, for the same data
Table 3–4: P-values and standard deviation of nonparametric functional interval mapping (NPFIM) from HIV dynamics data and growth data for different combination of variance-covariance estimators and sample size. “EB” means the shrinkage estimator which is guaranteed to be a consistent estimator. “REML” estimator is obtained from SAS Proc MIXED assuming each subject has a different underlying mean curve. “True” estimator is, of course, the matrix we actually used to generate data.

<table>
<thead>
<tr>
<th>Variance</th>
<th>HIV dynamics data</th>
<th>Growth data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N = 61$</td>
<td>$N = 200$</td>
</tr>
<tr>
<td>$EB$</td>
<td>0.4018(0.0300)</td>
<td>0.00126(&lt;.0001)</td>
</tr>
<tr>
<td>$REML$</td>
<td>0.1644(0.0169)</td>
<td>0.03346(0.0042)</td>
</tr>
<tr>
<td>$True$</td>
<td>0.0376(0.0167)</td>
<td>&lt;.0001(&lt;.0001)</td>
</tr>
</tbody>
</table>

The numbers in parentheses are the sampling errors of the P-values.
Figure 3–15: The heritability curve across all observation time points in the HIV simulation study.
setting from simulation for Table 3–2, the 1000 simulation tests to decide the
critical value of LRT statistics takes over 10 times more time when PFIM is
used (We used the simplex algorithm to find the MLE of functional parameters)
than the time needed when NPFIM is used. Even worse, the simplex algorithm
sometimes gets stuck in local maxima and results in negative likelihood ratio test
statistics. Therefore, the similarity to simple linear regression of our purposed
nonparametric functional mapping method has computational advantages compared
to the traditional functional mapping method.

As in Chapter 1 Section 1.4, it is pointed out that B-splines are very attractive
for nonparametric regression, but the selection of the optimal number and position
of knots do not have a standard rule. When we analyzed the Poplar dataset and
simulated datasets, we used B-splines with equidistant knots and order 3. If one
needs more control of flexibility, one may used P-splines as Eilers and Marx (1996)
proposed.

Functional mapping is parametric in nature, which allows one to easily test
many different biological hypotheses by testing for equality of parameters. For
example, one could test if a QTL starts or ceases to exert an effect on growth
trajectories by testing hypothesis $H_0 : a_1/(1 + b_1 e^{-r_1 t^*}) = a_2/(1 + b_2 e^{-r_2 t^*})$ at a given
time $t^*$. Testing this hypothesis is equivalent to testing the difference between the
model with no restriction and the model with the restriction: $a_1/(1 + b_1 e^{-r_1 t^*}) =
a_2/(1 + b_2 e^{-r_2 t^*})$. Our method could also be used to test such an hypothesis. In this
example, we can first find the MLE’s of the coefficient vectors under the restriction
$m_1(t^*) = m_2(t^*)$ and compare whether there is a significant difference between the
two coefficient vectors.

Figure 3–16 shows the empirical distributions of the test statistics under
$H_0$ resulting from the simulation procedure proposed in Chapter 3, from which
the critical point showing significant evidence of existing QTL is determined.
From this figure we can easily see that, with all other information is the same (sample size, marker information, underlying function or biological mechanism), the empirical distributions are very similar to each other when the empirical Bayes procedure is used to estimate the four different covariance matrixes. The same finding is observed for the cases using the exact underlying covariance matrix. This interesting phenomena might be due to the orthogonality between the coefficient vectors $\xi$ and the covariance matrix $\Sigma$.

If there is no QTL, that is, $\theta = 0$ or 1, there is no mixture in the likelihood function:

$$L \propto \exp\left\{-\frac{1}{2} \sum_{i=1}^{n_1+n_2} (y_i - B\xi_1)'\Sigma^{-1}(y_i - B\xi_1) - \frac{1}{2} \sum_{i=n_1+n_2+1}^{N} (y_i - B\xi_2)'\Sigma^{-1}(y_i - B\xi_2)\right\}.$$ 

The off-diagonal elements in the Fisher information matrix are (suppose $\sigma$ is the unknown parameter in covariance matrix $\Sigma$):

$$E \frac{\partial^2 (-2 \log L)}{\partial \xi \partial \sigma} = \sum E \frac{\partial \Sigma}{\partial \sigma}(y_i - B\xi) = 0.$$ 

So $\xi_1$, $\xi_2$ and $\Sigma$ are orthogonal to each other. Cox and Reid (1987) stated that the maximum likelihood estimates of $\xi_1$ and $\xi_2$ move slowly with respect to $\Sigma$. (Convergence rate is $O_p(\frac{1}{n})$.)

Furthermore, when the sample size is the same, even for testing QTL controlling different biological trajectories, the empirical distributions of the test statistic are similar, which is illustrated in Figure 3–17 where the two empirical distributions for data sets simulated from different biological trajectories but with equal sample size are plotted. This is because $H_0$ only states that when there is no QTL, the underlying functions are the same but does not specify the functional form. But for data sets with different sample sizes, even if all the other information is the same, there is a big difference among the resulting empirical distributions.
Figure 3–16: The empirical distribution of the test statistics in the simulation study of Chapter 3. Dataset 1-4 correspond to the four $\Sigma$ matrixes in Table 3–2. “EB” means the Empirical Bayes estimate. “True” represents the covariance matrix used to generate the data.
Figure 3–17: The left plot indicates the empirical distribution of the test statistics in the simulated growth data of Chapter 3 with covariance matrix $\Sigma_1$, 61 subjects with 11 observation points. The right one is the one from the simulated HIV dynamics data, 61 subjects with 20 observation points.
3.10 Appendix: Performance of the Approximation $G$ to Actual Likelihood Ratio Test Statistics

3.10.1 Joint distribution of $G$ at each $Q$ under $H_0$

Notice that we arrange $Y$ according to which group $y_{ij}$ is from, so at different pairs of markers, we will have a different vector $Y$. To make notation uniform, we can use a permutation matrix $P$. If we denote $Y_1$ as the $Y$ for first pair of markers, then for other intervals flanked by subsequent markers, $k = 2, \ldots, m$,

$$Y_k = P_k Y_1,$$

where $P_k$ is the permutation matrix. Of course, for each $\theta$ relative to same marker $M_k$, $Y_k$ is the same and so is $P_k$. Suppose the matrix $A$ corresponding to $k$th pair of markers at a fixed $\theta$ is labelled as $A_\theta^k$. Thus, at each $\theta$ among $k$th marker’s pair, the corresponding test statistic is

$$Y_k' A_\theta^k Y_k = Y_1' P_k' A_\theta^k P_k Y_1.$$

Notice that, $\text{var}(P_k Y_1) = P_k \tilde{\Sigma} P_k' = \tilde{\Sigma}$ because $\tilde{\Sigma}$ is block diagonal with block size $T \times T$ and $P_k$ just changes the rows of $\tilde{\Sigma}$. So under $H_0$ at position $\theta$, we still have

$$Y_k' A_\theta^k Y_k = Y_1' P_k' A_\theta^k P_k Y_1 \sim \beta_{\theta, k}^2 \chi^2_l,$$

with $\beta_{\theta, k}^2$ denoting the multiplicity determined by the $k$th markers’ pair and hypothetic QTL indexed by $\theta$. Recall that under $H_0$, $m(\overline{t}) = m_1(\overline{t}) = m_2(\overline{t})$, $Y_1 \sim MVN(1_n \otimes m(\overline{t}), I_n \otimes \Sigma)$ and we can write

$$Y_k' A_\theta^k Y_k = Y_k' \tilde{\Sigma}^{-\frac{1}{2}} \tilde{\Sigma}^{\frac{1}{2}} A_\theta^k \tilde{\Sigma}^{\frac{1}{2}} \tilde{\Sigma}^{-\frac{1}{2}} Y_k$$

$$= \beta_{\theta, k}^2 Y_k' \tilde{\Sigma}^{-\frac{1}{2}} W_k^\theta W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} Y_k$$

$$= \beta_{\theta, k}^2 Y_1' P_k' \tilde{\Sigma}^{-\frac{1}{2}} W_k^\theta W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_k Y_1$$

$$\equiv \beta_{\theta, k}^2 z_k^\theta Z_k^\theta,$$
where  \( Z_k^\theta = W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_k Y_1 \sim^d MVN(W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu}, W_k^\theta W_k^\theta) \) and \( \beta_k^\theta W_k^\theta W_k^\theta \) are the spectral decompositions of \( \tilde{\Sigma}^\frac{1}{2} A_k^\theta \tilde{\Sigma}^\frac{1}{2} \). The matrix \( W_k^\theta \) is composed of the eigenvectors corresponding to the \( l \) non-zero eigenvalues, and it has \( l \) orthonormal columns and is of order \( NT \times l \). Also, under \( H_0 \) the structure of \( A_k^\theta \) results in

\[
(W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu})' W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_k \tilde{\mu} = \tilde{\mu}' A_k^\theta \tilde{\mu} / \beta_k^\theta = 0.
\]

and thus, under \( H_0 \), \( Z_{k_i}^\theta = W_k^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_k Y_1 \sim MVN(0, I_l) \).

At each interval along the entire map, we examine \( n_\theta \) putative positions. The entire vector \( Z = (Z_1^\theta, \ldots, Z_1^{n_\theta}, Z_2^\theta, \ldots, Z_m^{n_\theta})' \) has distribution \( Z \sim MVN(\mu_Z, \Delta) \) where

\[
\mu_{Z_i} = W_i^\theta \tilde{\Sigma}^{-\frac{1}{2}} \tilde{\mu},
\]

\( i = 1, \ldots, m, j = 1, \ldots, n_\theta \) and

\[
\Delta = \begin{pmatrix}
W_1^\theta \tilde{\Sigma}^{-\frac{1}{2}} \\
\vdots \\
W_1^{n_\theta} \tilde{\Sigma}^{-\frac{1}{2}} \\
W_2^\theta \tilde{\Sigma}^{-\frac{1}{2}} P_2 \\
\vdots \\
W_m^{n_\theta} \tilde{\Sigma}^{-\frac{1}{2}} P_m
\end{pmatrix}
\begin{pmatrix}
\tilde{\Sigma}^{-\frac{1}{2}} W_1^\theta & \ldots & \tilde{\Sigma}^{-\frac{1}{2}} W_1^{n_\theta} & P_2^\theta \tilde{\Sigma}^{-\frac{1}{2}} W_2^\theta & \ldots & P_m^\theta \tilde{\Sigma}^{-\frac{1}{2}} W_m^{n_\theta}
\end{pmatrix},
\]

and under \( H_0 \), \( \mu_Z = 0 \).

If we let \( B_k^\theta(x), j = 1, \ldots, n_\theta \) denote the \( l \)-dimension ball with radius equal to \( x/\beta_k^\theta \), then using the joint distribution of \( G_k^\theta \)'s to approximate the joint distribution of every likelihood ratio test statistics at QTL position \( \theta \) within \( k \)th pair of
marker $-2\log \lambda^\theta_k$'s, we can get

$$P_0(\max_{1 \leq k \leq m; \theta} -2\log \lambda^\theta_k \leq x)$$

$$\approx P_0(\max_{1 \leq k \leq m; \theta} Y'_k A^\theta_k Y_k \leq x)$$

$$= P_0(\max_{1 \leq k \leq m; \theta} \beta^\theta_k Z^\theta_k Z^\theta_k \leq x)$$

$$= P_0(Z^\theta_1 Z^\theta_1 \leq x/\beta^\theta_1, \ldots, Z^\theta_m Z^\theta_m \leq x/\beta^\theta_m)$$

$$= \int \ldots \int_{\{Z^\theta_i \in B^\theta_i(x)\}} \frac{1}{\sqrt{2\pi}^{(m+n)l}|\Delta|^{1/2}} \exp \left\{ -\frac{1}{2} \sum \Delta^{-1} Z^\theta_i \right\} dZ^\theta_1 \ldots dZ^\theta_m.$$

This probability, which is one minus the p-value for $H_0$, can be directly calculated by simulating $Z \sim MVN(0, \Delta)$. Again, all the above derivations are made under the assumption that we know $\Sigma$, which is typically untrue in practice. We suggest substituting a REML estimate of the variance-covariance matrix $\hat{\Sigma}$ instead of $\Sigma$, and, in that case, the above formulae are correct asymptotically.

### 3.10.2 Examples

#### Simulation

The purpose of this simulation is to examine how the test statistic behaves as the underlying mean curves move apart from one another when applying this more general method. We, of course, expect that the $p-$value gets smaller when we move the underlying curves apart.

The first data set assumes the underlying phenotypic curves are two flat lines as before. The linkage map has 20 markers totally with 25cM apart from its neighbors. There are 100 subjects with measurements generated at 4 observation points using the variance-covariance matrix for each $y_i$

$$\Sigma = \begin{pmatrix}
0.06 & 0.035 & 0.0225 & 0.01625 \\
0.035 & 0.06 & 0.035 & 0.0225 \\
0.0225 & 0.035 & 0.06 & 0.035 \\
0.01625 & 0.0225 & 0.035 & 0.06
\end{pmatrix}. $$
Based on the marker information, 100 data sets are generated and analyzed. The unit of scanning throughout the linkage map is 2cM. Since the underlying curve are the two flat lines $y = a_0$ and $y = a_1$, we can use $|a_0 - a_1|$ as a measure of distance. The $p$—value trend with respect to different distances between two underlying flat lines, and its 95% confidence interval, is given in Figure 3–18. It is clear that the $P$—value decreases as the curves move apart, which is the behavior we would expect.

The second data set assumes one phenotypic curve is flat ($m_1(t) = 1$) and the other is quadratic ($m_2(t) = ax^2 + bx + c$). In this case, using 25 markers with 20cM apart from nearest neighbors, 50 subjects are generated at 12 observation points.
Figure 3–19: Graph of $p$–value and its 95% confidence interval for data from two underlying curves: one is flat line and the other is in quadratic form using nonparametric functional interval mapping.

using the variance-covariance matrix for each $y_i$

$$
\Sigma = .1J_{12} + \text{Autoregressive}(\tau^2 = .05, \rho = .5).
$$

Based on the same marker information, 100 data sets are generated and analyzed. To measure distance, we use the area between the two curves over the observed time interval. Figure 3–19 shows the $p$–value profile for different distances and its 95% confidence interval. Again, the behavior is what we desire, with the $p$-value decreasing as the curves get further apart.

**Poplar Data**
Figure 3-20: Graph of the likelihood ratio test statistics along the linkage group 10 using nonparametric functional interval mapping.
Again we apply this method to the same Poplar Data set as before. Because our model assumes that all \( t \) observations have equal variance, we first transformed the original observations by taking the log. We use order 3 B-splines with equidistant knots to fit the growth curve and scan the linkage group 10 2cM by 2cM. Using our method with the variance-covariance structure autoregressive+simple (estimated by PROC MIXED as \( \hat{\Sigma} = 0.001236*J_{11} + Autoregressive(\tau^2 = 0.06155, \rho = 0.8945) \)), we found strong evidence that there is a QTL between the first and the second marker on linkage group 10 which controls the growth trajectory of stem diameter in the interspecific hybrids of poplar (\( p \)-value = \( 5e^{-4} \) with standard deviation \( 5e^{-4} \)). The biggest likelihood-ratio test statistics \(-2\log \lambda\) appears in the interval flanked by Marker \( CA/CCC-640R \) and Marker \( CG/CCC-825 \). Figure 3–20 shows likelihood ratio test statistics along the linkage group 10. Again this finding agrees with the conclusion in Ma, Casella and Wu (2002). But the resulting QTL locus, i.e, where the largest likelihood ratio test statistic is, is slightly different from our method and their method. This may due to all the approximations we made.
The dataset used for QTL mapping contains not only genotypic values like markers but also phenotypic values measured from some traits. In practice, missing data situations are usually faced due to many reasons. To handle missing values in the analysis of longitudinal or repeated measurement data, many authors proposed statistical models to incorporate both complete cases and incomplete cases. Their ideas can be similarly adopted to incorporate missing phenotypic cases into QTL mapping. But little literature has appeared to incorporate missing marker cases into QTL mapping. A popular way is to simply discard subjects with missing markers, which directly results in loss of power in mapping QTL because of smaller sample size. So in this chapter we will use full information from all subjects, whether having complete marker information or not, to gain as much power as possible. For simplicity we assume the phenotypic values from each subject are complete.

### 4.1 Simple Methods

In order to incorporate incompleteness into the modelling process, we need to reflect on the nature of the missing value mechanism. For marker information, the loss of data is mainly from technical reasons or just simple recording mistakes. So unlike the complicated missing mechanism in repeated measurement data, we can view such missing mechanisms in marker information as missing completely at random (MCAR) (Little and Rubin 2002).

With this MCAR assumption there are a number of simple and valid techniques for dealing with missing data. The computationally simplest techniques, a complete data analysis, in which the analysis is restricted to the subjects for whom
all markers information has been recorded. A complete data analysis is popular because it maps a ragged data matrix into a rectangle one. But as we stated before, there is nearly always a substantial loss of information so that a loss of detecting power follows.

An alternative way to obtain a data set on which a complete data analysis can be carried out is filling in the missing values, instead of deleting subjects with incomplete sequences. Using observed values to impute values for the missing values is the principle of imputation. For marker data, since we know the genetic distances between the markers, we can impute the missing marker information based on its nearest neighbor and corresponding genetic distance. Then the resulting data set is analyzed as if it represented the true complete data.

A third method is based on the principle of analyzing the incomplete data as such. A simple representative is available data analysis. For the functional interval mapping, we constructed a likelihood ratio test interval by interval. So for each interval, we can use all subjects who have both values of two end markers to derive such LR test. This may cause the number of subjects in each LR test to be different. Obviously, this method is more efficient than the complete case method for more information is used.

4.2 Modelling Incompleteness in Markers

In this section, we directly put the missing markers into the model. The data setting is exactly what has described in Section 3.1. First, we consider the case where at least one of the two neighbors of the missing marker is recorded in the data set, that is, for any marker interval, say, bracketed by $M$ and $N$, all the subjects in the data sets have information of either $M$ or $N$ or both. Those subjects with information of both $M$ and $N$ missing are deleted when we do the likelihood ratio test in this interval.
Without loss of generality, suppose one subject has information on \( \mathcal{M} \) but information about \( \mathcal{N} \) is somehow missed. The likelihood function of one subject with measurement vector \( \mathbf{y} \) when \( \mathcal{M} \) and \( \mathcal{N} \) are both observed is

\[
\begin{align*}
  f(\mathbf{y}|\mathcal{M} = M, \mathcal{N} = N) &= \phi(\mu_{Qq}, \Sigma) \\
  f(\mathbf{y}|\mathcal{M} = M, \mathcal{N} = n) &= (1 - \theta)\phi(\mu_{Qq}, \Sigma) + \theta\phi(\mu_{qq}, \Sigma) \\
  f(\mathbf{y}|\mathcal{M} = m, \mathcal{N} = N) &= \theta\phi(\mu_{Qq}, \Sigma) + (1 - \theta)\phi(\mu_{qq}, \Sigma) \\
  f(\mathbf{y}|\mathcal{M} = m, \mathcal{N} = n) &= \phi(\mu_{qq}, \Sigma),
\end{align*}
\]

where \( \phi(\mu, \Sigma) \) is the probability density function of \( MVN(\mu, \Sigma) \) and \( \theta \) is defined as in Table 1–3. As before, we assume \( \mu_{Qq} = \mathbf{m}_1(\vec{t}) = \mathbf{B}\xi_1 \) and \( \mu_{qq} = \mathbf{m}_2(\vec{t}) = \mathbf{B}\xi_2 \).

Let \( r \) denote the recombination rate between \( \mathcal{M} \) and \( \mathcal{N} \), then we have

\[
\begin{align*}
  P(\mathcal{N} = N|\mathcal{M} = M) &= 1 - r \\
  P(\mathcal{N} = n|\mathcal{M} = M) &= r \\
  P(\mathcal{N} = N|\mathcal{M} = m) &= r \\
  P(\mathcal{N} = n|\mathcal{M} = m) &= 1 - r.
\end{align*}
\]

Thus, we can write the likelihood function of \( \mathbf{y} \) conditional on marker \( \mathcal{M} \) while its neighbor marker \( \mathcal{N} \) is missing as the following:

\[
\begin{align*}
  f(\mathbf{y}|\mathcal{M} = M) \\
  &= f(\mathbf{y}|\mathcal{M} = M, \mathcal{N} = N)P(\mathcal{N} = N|\mathcal{M} = M) \\
  &\quad + f(\mathbf{y}|\mathcal{M} = M, \mathcal{N} = n)P(\mathcal{N} = n|\mathcal{M} = M) \\
  &= (1 - r)\phi(\mu_{Qq}, \Sigma) + r[(1 - \theta)\phi(\mu_{Qq}, \Sigma) + \theta\phi(\mu_{qq}, \Sigma)] \\
  &= (1 - r\theta)\phi(\mu_{Qq}, \Sigma) + r\theta\phi(\mu_{qq}, \Sigma).
\end{align*}
\]
and
\[
f(y|\mathcal{M} = m) \\
= f(y|\mathcal{M} = m, \mathcal{N} = N) P(\mathcal{N} = N | \mathcal{M} = m) \\
+ f(y|\mathcal{M} = m, \mathcal{N} = n) P(\mathcal{N} = n | \mathcal{M} = m) \\
= (1 - r)\phi(\mu_{qq}, \Sigma) + r[(1 - \theta)\phi(\mu_{qq}, \Sigma) + \theta\phi(\mu_{Qq}, \Sigma)] \\
= (1 - r\theta)\phi(\mu_{qq}, \Sigma) + r\theta\phi(\mu_{Qq}, \Sigma) .
\]

Similarly, for subjects who have \( \mathcal{N} \) information but no \( \mathcal{M} \) information, the density function can be obtained with following expressions:
\[
f(y|\mathcal{N} = N) = [1 - r(1 - \theta)]\phi(\mu_{Qq}, \Sigma) + r(1 - \theta)\phi(\mu_{qq}, \Sigma) \\
f(y|\mathcal{N} = n) = [1 - r(1 - \theta)]\phi(\mu_{qq}, \Sigma) + r(1 - \theta)\phi(\mu_{Qq}, \Sigma) .
\]

Suppose all subjects with complete marker information are arranged as before, subjects with only \( \mathcal{M} = M \) or \( \mathcal{M} = m \) are subjects \( n_4 + 1, \ldots, n_5 \) and subjects \( n_5 + 1, \ldots, n_6 \), respectively. Subjects with only \( \mathcal{N} = N \) or \( \mathcal{N} = n \) are subjects \( n_6 + 1, \ldots, n_7 \) and subjects \( n_7 + 1, \ldots, n_8 \), respectively. Then the log likelihood function using all subjects in this interval with end markers \( \mathcal{M} \) and \( \mathcal{N} \) is expressed as
\[
\log L = \sum_{i=1}^{n_1} -\frac{\|y_i - B\xi_1\|^2}{2} \\
+ \sum_{i=n_1+1}^{n_2} \log [(1 - \theta)\Phi(y_i; B\xi_1, \Sigma) + \theta\Phi(y_i; B\xi_2, \Sigma)] \\
+ \sum_{i=n_2+1}^{n_3} \log [(1 - \theta)\Phi(y_i; B\xi_2, \Sigma) + \theta\Phi(y_i; B\xi_1, \Sigma)] \\
+ \sum_{i=n_3+1}^{n_4} -\frac{\|y_i - B\xi_2\|^2}{2}.
\]
we have got the log likelihood function, so we can get the likelihood equation for

Denote

In Section 4.3 Likelihood Ratio Test when Missing Marker is Present

In Section 4.2 we have got the log likelihood function, so we can get the likelihood equation for $\xi_1$ as

$$
\frac{\partial \log L}{\partial \xi_1} = \sum_{i=1}^{n_1} B'S^{-1}(y_i - B\xi_1)
+ \sum_{i=n_1+1}^{n_2} (1 - \theta)\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)
+ \sum_{i=n_2+1}^{n_3} \frac{\theta\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)}{(1 - \theta)\Phi(y_i; B\xi_2, \Sigma) + \theta\Phi(y_i; B\xi_1, \Sigma)}
+ \sum_{i=n_3+1}^{n_4} (1 - r\theta)\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)
+ \sum_{i=n_4+1}^{n_5} \frac{r\theta\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)}{r\Phi(y_i; B\xi_1, \Sigma) + (1 - r\theta)\Phi(y_i; B\xi_2, \Sigma)}
+ \sum_{i=n_5+1}^{n_6} \frac{r(1 - \theta)\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)}{r(1 - \theta)\Phi(y_i; B\xi_1, \Sigma) + [1 - r(1 - \theta)]\Phi(y_i; B\xi_2, \Sigma)}
+ \sum_{i=n_6+1}^{n_7} \frac{[1 - r(1 - \theta)]\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)}{[1 - r(1 - \theta)]\Phi(y_i; B\xi_1, \Sigma) + r(1 - \theta)\Phi(y_i; B\xi_2, \Sigma)}
+ \sum_{i=n_7+1}^{n} \frac{r(1 - \theta)\Phi(y_i; B\xi_1, \Sigma)B'S^{-1}(y_i - B\xi_1)}{r(1 - \theta)\Phi(y_i; B\xi_1, \Sigma) + [1 - r(1 - \theta)]\Phi(y_i; B\xi_2, \Sigma)}
+ \text{constant}.
$$

Denote

$$
P(y; t) = \frac{(1 - t)\Phi(y; B\xi_1, \Sigma)}{(1 - t)\Phi(y; B\xi_1, \Sigma) + t\Phi(y; B\xi_2, \Sigma)}
$$
Then the likelihood equation for $\xi_1$ can be rewritten as

$$
\frac{\partial \log L}{\partial \xi_1} = \sum_{i=1}^{n_1} B' \Sigma^{-1} (y_i - B \xi_1) \\
+ \sum_{i=n_1+1}^{n_2} P(y_i; \theta) B' \Sigma^{-1} (y_i - B \xi_1) + \sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta) B' \Sigma^{-1} (y_i - B \xi_1) \\
+ \sum_{i=n_4+1}^{n_5} P(y_i; r \theta) B' \Sigma^{-1} (y_i - B \xi_1) + \sum_{i=n_5+1}^{n} P(y_i; 1 - r \theta) B' \Sigma^{-1} (y_i - B \xi_1) \\
+ \sum_{i=n_6+1}^{n_7} P(y_i; r (1 - \theta)) B' \Sigma^{-1} (y_i - B \xi_1) + \sum_{i=n_7+1}^{n} P(y_i; 1 - r (1 - \theta)) B' \Sigma^{-1} (y_i - B \xi_1)
$$

Similarly, the likelihood equation for $\xi_2$ can be rewritten as

$$
\frac{\partial \log L}{\partial \xi_2} = \sum_{i=n_3+1}^{n} B' \Sigma^{-1} (y_i - B \xi_2) \\
+ \sum_{i=n_1+1}^{n_2} [1 - P(y_i; \theta)] B' \Sigma^{-1} (y_i - B \xi_2) \\
+ \sum_{i=n_2+1}^{n_3} [1 - P(y_i; 1 - \theta)] B' \Sigma^{-1} (y_i - B \xi_2) \\
+ \sum_{i=n_4+1}^{n_5} [1 - P(y_i; r \theta)] B' \Sigma^{-1} (y_i - B \xi_2) \\
+ \sum_{i=n_5+1}^{n} [1 - P(y_i; 1 - r \theta)] B' \Sigma^{-1} (y_i - B \xi_2) \\
+ \sum_{i=n_6+1}^{n_7} [1 - P(y_i; r (1 - \theta))] B' \Sigma^{-1} (y_i - B \xi_2) + \sum_{i=n_7+1}^{n} [1 - P(y_i; 1 - r (1 - \theta))] B' \Sigma^{-1} (y_i - B \xi_2)
$$

Thus, we have the following EM algorithm to numerically find the MLEs of $\xi_1$ and $\xi_2$:

**EM Algorithm:** For fixed $\theta$ and known $\Sigma$, repeat the following steps until the convergence criterion is satisfied.
Step $k$: Calculate $P(y_i; 1 - \theta)^{(k)}$, $P(y_i; \theta)^{(k)}$, $P(y_i; r\theta)^{(k)}$, $P(y_i; 1 - r\theta)^{(k)}$, $P(y_i; r(1 - \theta))^{(k)}$ and $P(y_i; 1 - r(1 - \theta))^{(k)}$ using $\hat{\xi}_1^k$ and $\hat{\xi}_2^k$.

Step $k + 1$: Calculate

$$S_1 = \sum_{1}^{n_1} y_i + \sum_{n_1+1}^{n_2} P(y_i; \theta)^{(k)} y_i + \sum_{n_2+1}^{n_3} P(y_i; 1 - \theta)^{(k)} y_i$$
$$+ \sum_{n_4+1}^{n_5} P(y_i; r\theta)^{(k)} y_i + \sum_{n_5+1}^{n_6} P(y_i; 1 - r\theta)^{(k)} y_i$$
$$+ \sum_{n_6+1}^{n_7} P(y_i; r(1 - \theta))^{(k)} y_i + \sum_{n_7+1}^{n} P(y_i; 1 - r(1 - \theta))^{(k)} y_i$$

$$S_2 = n_1 + \sum_{n_1+1}^{n_2} P(y_i; \theta)^{(k)} + \sum_{n_2+1}^{n_3} P(y_i; 1 - \theta)^{(k)}$$
$$+ \sum_{n_4+1}^{n_5} P(y_i; r\theta)^{(k)} + \sum_{n_5+1}^{n_6} P(y_i; 1 - r\theta)^{(k)}$$
$$+ \sum_{n_6+1}^{n_7} P(y_i; r(1 - \theta))^{(k)} + \sum_{n_7+1}^{n} P(y_i; 1 - r(1 - \theta))^{(k)}$$

$$\hat{\xi}_1^{(k+1)} = (B^\prime \Sigma^{-1} B)^{-1} B^\prime \Sigma^{-1} \frac{S_1}{S_2}$$

$$S_3 = \sum_{n_3+1}^{n} y_i + \sum_{n_1+1}^{n_2} [1 - P(y_i; \theta)^{(k)}] y_i + \sum_{n_2+1}^{n_3} [1 - P(y_i; 1 - \theta)^{(k)}] y_i$$
$$+ \sum_{n_4+1}^{n_5} [1 - P(y_i; r\theta)^{(k)}] y_i + \sum_{n_5+1}^{n_6} [1 - P(y_i; 1 - r\theta)^{(k)}] y_i$$
$$+ \sum_{n_6+1}^{n_7} [1 - P(y_i; r(1 - \theta))^{(k)}] y_i + \sum_{n_7+1}^{n} [1 - P(y_i; 1 - r(1 - \theta))^{(k)}] y_i$$

$$= \sum_{1}^{n} y_i - S_1$$
\[ S_4 = n_4 - n_3 + \sum_{n_1+1}^{n_2} [1 - P(y_i; \theta)^{(k)}] + \sum_{n_2+1}^{n_3} [1 - P(y_i; 1 - \theta)^{(k)}] \]
\[ + \sum_{n_4+1}^{n_5} [1 - P(y_i; r \theta)^{(k)}] + \sum_{n_5+1}^{n_6} [1 - P(y_i; 1 - r \theta)^{(k)}] \]
\[ + \sum_{n_6+1}^{n_7} [1 - P(y_i; (1 - \theta))^{(k)}] + \sum_{n_7+1}^{n} [1 - P(y_i; 1 - (1 - \theta))^{(k)}] \]
\[ = n - S'2 \]
\[ \hat{\xi}^{(k+1)}_2 = (B\Sigma^{-1}B)^{-1}B\Sigma^{-1}S_3 S_4 \]

When the QTL doesn’t exist, that is, under \( H_0 \), the MLE for the only unknown coefficient vector \( \xi \) is \( \hat{\xi}_0 = (B\Sigma^{-1}B)^{-1}B\Sigma^{-1} \sum_{i=1}^{n} y_i \). So the likelihood ratio test statistic at each fixed \( \theta \) is

\[ -2 \log \lambda = -2 \log \frac{\max_{H_0} L(\xi | y)}{\max L(\xi | y)} \]
\[ = \sum_{i=1}^{n} \|y_i - B\hat{\xi}_0\|^2_\Sigma - \sum_{i=1}^{n_1} \|y_i - B\hat{\xi}_1\|^2_\Sigma \]
\[ + 2 \sum_{i=n_1+1}^{n_2} \log \{ (1 - \theta) \exp(-\|y_i - B\hat{\xi}_1\|^2_\Sigma) + \theta \exp(-\|y_i - B\hat{\xi}_2\|^2_\Sigma) \} \]
\[ + 2 \sum_{i=n_2+1}^{n_3} \log \{ \theta \exp(-\|y_i - B\hat{\xi}_1\|^2_\Sigma) + (1 - \theta) \exp(-\|y_i - B\hat{\xi}_2\|^2_\Sigma) \} \]
\[ - \sum_{i=n_3+1}^{n} \|y_i - B\hat{\xi}_2\|^2_\Sigma \]
\[ + 2 \sum_{i=n_4+1}^{n_5} \log \{ (1 - r \theta) \exp(-\|y_i - B\hat{\xi}_1\|^2_\Sigma) + r \theta \exp(-\|y_i - B\hat{\xi}_2\|^2_\Sigma) \} \]
\[ + 2 \sum_{i=n_5+1}^{n_6} \log \{ r \theta \exp(-\|y_i - B\hat{\xi}_1\|^2_\Sigma) + (1 - r \theta) \exp(-\|y_i - B\hat{\xi}_2\|^2_\Sigma) \} \]
\[ + 2 \sum_{i=n_6+1}^{n_7} \log \{ [1 - r(1 - \theta)] \exp(-\|y_i - B\hat{\xi}_1\|^2_\Sigma) \} \]
\[
+ r(1 - \theta) \exp\left(\frac{-\|y_i - B\hat{\xi}_2\|^2}{2}\right)
\]
\[
+ 2 \sum_{i=n_{T}+1}^{n} \log\left\{ r(1 - \theta) \exp\left(\frac{-\|y_i - B\hat{\xi}_1\|^2}{2}\right) + [1 - r(1 - \theta)] \exp\left(\frac{-\|y_i - B\hat{\xi}_2\|^2}{2}\right) \right\}
\]

The final test statistic is the maximum of these LRT statistics from each marker interval. The critical value to decide if a QTL exists in this linkage group can be similarly obtained from the simulation procedure introduced in Chapter 3.

4.4 Simulation Procedure to Get Threshold Point

As we have seen, when \( H_0 \) is true, after enough steps in the EM algorithm, the maximum likelihood estimate of two unknown coefficient vectors, \( \xi_1 \) and \( \xi_2 \) can be directly written as linear combinations of realized phenotypic vectors. That is, using \( Y \) as an \( NT \times 1 \) vector with \( y_1, ..., y_n \) stacking together, \( (Y \) has been arranged in the same way as in Section 4.2), we can have \( \hat{\xi}_1 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_1 Y \) and \( \hat{\xi}_2 = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_2 Y \), where (let \( n_{ij} = n_j - n_i \))

\[
W_1 = \frac{\begin{pmatrix} 1_{n_1} & (1 - \theta)_{n_{12}} & \theta_{n_{23}} & 0_{n_{34}} & (1 - r\theta)_{n_{45}} & (r\theta)_{n_{56}} & [1 - r + r\theta]_{n_{67}} & [r - r\theta]_{n_{78}} \end{pmatrix}}{n_2 + (n_{23} - n_{12})\theta + n_{45} + (n_{56} - n_{45})r\theta + n_{67} + (n_{78} - n_{67})r(1 - \theta)} \equiv w_1 \otimes I_T,
\]

and

\[
W_2 = \frac{\begin{pmatrix} 0_{n_1} & \theta_{n_{12}} & (1 - \theta)_{n_{23}} & 1_{n_{34}} & (r\theta)_{n_{45}} & (1 - r\theta)_{n_{56}} & [r - r\theta]_{n_{67}} & [1 - r + r\theta]_{n_{78}} \end{pmatrix}}{n_2 - (n_{23} - n_{12})\theta + n_{45} - (n_{56} - n_{45})r\theta + n_{67} - (n_{78} - n_{67})r(1 - \theta)} \equiv w_2 \otimes I_T.
\]
Of course, the MLE of $\xi_0$ under $H_0$ can also be expressed as $(B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}W_0Y$ with $W_0 = 1_n \otimes \frac{1}{n}I_T$.

We can use the same simulation procedure which can determine the cutoff point for detecting QTL from a genome/linkage group:

1. Generate 1000 different data sets from $H_0$, that is, having same variance-covariance structure and based on same known marker information. We can easily generate $y_i - B\hat{\xi}_0, y_i - B\hat{\xi}_1, y_i - B\hat{\xi}_2, i = 1, \ldots, n$, for each fixed $\theta$.
2. For each interval calculate the likelihood ratio test statistics at these fixed $\theta$ based on simulated data; Obtain 1000 different maximum likelihood ratio test statistics through the whole genome(linkage group).
3. Get the cutoff point value from the simulated distribution of maximum of LRT statistics.

If we treat $\theta$ as an unknown parameter, we can estimate it from the solution of

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=n_1+1}^{n_2} \frac{\Phi(y_i; B\xi_2, \Sigma) - \Phi(y_i; B\xi_1, \Sigma)}{(1 - \theta)\Phi(y_i; B\xi_1, \Sigma) + \theta\Phi(y_i; B\xi_2, \Sigma)}$$

$$+ \sum_{i=n_2+1}^{n_3} \frac{\Phi(y_i; B\xi_1, \Sigma) - \Phi(y_i; B\xi_2, \Sigma)}{(1 - \theta)\Phi(y_i; B\xi_2, \Sigma) + \theta\Phi(y_i; B\xi_1, \Sigma)}$$

$$+ \sum_{i=n_4+1}^{n_5} \frac{r[\Phi(y_i; B\xi_2, \Sigma) - \Phi(y_i; B\xi_1, \Sigma)]}{(1 - r\theta)\Phi(y_i; B\xi_1, \Sigma) + r\theta\Phi(y_i; B\xi_2, \Sigma)}$$

$$+ \sum_{i=n_5+1}^{n_6} \frac{r[\Phi(y_i; B\xi_1, \Sigma) - \Phi(y_i; B\xi_2, \Sigma)]}{(1 - r\theta)\Phi(y_i; B\xi_2, \Sigma) + r\theta\Phi(y_i; B\xi_1, \Sigma)}$$

$$+ \sum_{i=n_6+1}^{n_7} \frac{r[\Phi(y_i; B\xi_1, \Sigma) - \Phi(y_i; B\xi_2, \Sigma)]}{[1 - r(1 - \theta)]\Phi(y_i; B\xi_1, \Sigma) + r(1 - \theta)\Phi(y_i; B\xi_2, \Sigma)}$$

$$+ \sum_{i=n_7+1}^{n} \frac{r[\Phi(y_i; B\xi_2, \Sigma) - \Phi(y_i; B\xi_1, \Sigma)]}{[1 - r(1 - \theta)]\Phi(y_i; B\xi_2, \Sigma) + r(1 - \theta)\Phi(y_i; B\xi_1, \Sigma)}$$
\[
\begin{align*}
&= \sum_{i=n_1+1}^{n_2} \left( \frac{1 - P(y_i; \theta)}{\theta} - \frac{P(y_i; \theta)}{1 - \theta} \right) + \sum_{i=n_2+1}^{n_3} \left( \frac{P(y_i; 1 - \theta)}{\theta} - \frac{1 - P(y_i; 1 - \theta)}{1 - \theta} \right) \\
&\quad + \sum_{i=n_4+1}^{n_5} \left( \frac{1 - P(y_i; r\theta)}{\theta} - \frac{r P(y_i; r\theta)}{1 - r\theta} \right) + \sum_{i=n_5+1}^{n_6} \left( \frac{P(y_i; 1 - r\theta)}{\theta} - \frac{r(1 - P(y_i; 1 - r\theta))}{1 - r\theta} \right) \\
&\quad + \sum_{i=n_6+1}^{n_7} \left( \frac{r P(y_i; r(1 - \theta))}{1 - r(1 - \theta)} - \frac{1 - P(y_i; r(1 - \theta))}{1 - \theta} \right) \\
&\quad + \sum_{i=n_7+1}^{n} \left( \frac{r[1 - P(y_i; 1 - r(1 - \theta))]}{1 - r(1 - \theta)} - \frac{P(y_i; 1 - r(1 - \theta))}{1 - \theta} \right)
\end{align*}
\]

\[
= \frac{\sum_{i=n_1+1}^{n_2}[1 - P(y_i; \theta)]}{\theta} + \frac{\sum_{i=n_2+1}^{n_3} P(y_i; 1 - \theta)}{1 - \theta} \\
+ \frac{\sum_{i=n_4+1}^{n_5}[1 - P(y_i; r\theta)]}{\theta} + \frac{\sum_{i=n_5+1}^{n} P(y_i; 1 - r\theta)}{1 - \theta} \\
- \frac{\sum_{i=n_1+1}^{n_2} P(y_i; \theta) + \sum_{i=n_2+1}^{n_3}[1 - P(y_i; 1 - \theta)]}{1 - \theta} \\
- \frac{\sum_{i=n_4+1}^{n_5} P(y_i; r\theta) + \sum_{i=n_5+1}^{n_6}[1 - P(y_i; 1 - r\theta)]}{1 - r\theta} \\
- \frac{\sum_{i=n_6+1}^{n_7} r P(y_i; r(1 - \theta)) + \sum_{i=n_7+1}^{n}[1 - P(y_i; 1 - r(1 - \theta))]}{1 - r(1 - \theta)}
\]

\[
= \frac{A}{\theta} - \frac{B}{1 - \theta} - \frac{C}{1 - r\theta} - \frac{D}{1 - r(1 - \theta)} = 0
\]

Thus, we have this following EM algorithm to numerically find the MLE of \(\xi_1, \xi_2\) and \(\theta\) simultaneously by embedding the likelihood equation of \(\theta\) into the EM algorithm stated in Section 4.3:

**EM Algorithm 2:** Suppose \(\Sigma\) is known. Repeat the following steps until the convergence criterion is satisfied.

**Step k:** Calculate \(P(y_i; 1 - \theta)^{(k)}\), \(P(y_i; \theta)^{(k)}\), \(P(y_i; r\theta)^{(k)}\) and \(P(y_i; 1 - r\theta)^{(k)}\) using \(\hat{\xi}_1^k\) and \(\hat{\xi}_2^k\).
Step $k + 1$: Calculate

\[
S_1 = \sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} P(y_i; \theta^{(k)}) y_i + \sum_{i=1}^{n_3} P(y_i; \theta^{(k)}) y_i
\]
\[
+ \sum_{i=1}^{n_4} P(y_i; \theta^{(k)}) y_i + \sum_{i=1}^{n_5} P(y_i; \theta^{(k)}) y_i
\]
\[
+ \sum_{i=1}^{n_6} P(y_i; \theta^{(k)}) y_i + \sum_{i=1}^{n_7} P(y_i; \theta^{(k)}) y_i
\]
\[
S_2 = n_1 + \sum_{i=1}^{n_2} P(y_i; \theta^{(k)}) + \sum_{i=1}^{n_3} P(y_i; \theta^{(k)})
\]
\[
+ \sum_{i=1}^{n_4} P(y_i; \theta^{(k)}) + \sum_{i=1}^{n_5} P(y_i; \theta^{(k)})
\]
\[
+ \sum_{i=1}^{n_6} P(y_i; \theta^{(k)}) + \sum_{i=1}^{n_7} P(y_i; \theta^{(k)})
\]
\[
\hat{\xi}_1^{(k+1)} = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}S_1 \quad S_2
\]
\[
S_3 = \sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} [1 - P(y_i; \theta^{(k)})] y_i + \sum_{i=1}^{n_3} [1 - P(y_i; \theta^{(k)})] y_i
\]
\[
+ \sum_{i=1}^{n_4} [1 - P(y_i; \theta^{(k)})] y_i + \sum_{i=1}^{n_5} [1 - P(y_i; \theta^{(k)})] y_i
\]
\[
+ \sum_{i=1}^{n_6} [1 - P(y_i; \theta^{(k)})] y_i + \sum_{i=1}^{n_7} [1 - P(y_i; \theta^{(k)})] y_i
\]
\[
S_4 = n_1 - n_3 + \sum_{i=1}^{n_2} [1 - P(y_i; \theta^{(k)})] + \sum_{i=1}^{n_3} [1 - P(y_i; \theta^{(k)})]
\]
\[
+ \sum_{i=1}^{n_4} [1 - P(y_i; \theta^{(k)})] + \sum_{i=1}^{n_5} [1 - P(y_i; \theta^{(k)})]
\]
\[
+ \sum_{i=1}^{n_6} [1 - P(y_i; \theta^{(k)})] + \sum_{i=1}^{n_7} [1 - P(y_i; \theta^{(k)})]
\]
\[
\hat{\xi}_2^{(k+1)} = (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}S_3 \quad S_4
\]

Find $\hat{\theta}^{(k+1)}$ from equation

\[
\frac{A^{(k)}}{\theta} - \frac{B^{(k)}}{1-\theta} - \frac{C^{(k)}}{1-r\theta} - \frac{D^{(k)}}{1-r(1-\theta)} = 0
\]
Table 4–1: P-values and its standard deviation of nonparametric functional interval mapping (NPFIM) using case-wise deleted data set and full data set, $\Sigma = 0.3 \ast J_{11} + \text{Autoregressive}(\tau^2 = 0.1, \rho = 0.8)$

<table>
<thead>
<tr>
<th>Variance Estimate</th>
<th>case-wise deleted data</th>
<th>full data</th>
</tr>
</thead>
<tbody>
<tr>
<td>REML</td>
<td>.209(.0238)</td>
<td>.0258(.0114)</td>
</tr>
<tr>
<td>True</td>
<td>.197(.0240)</td>
<td>.0243(.0116)</td>
</tr>
</tbody>
</table>

The numbers in parentheses are the sampling errors of the P-values.

4.5 Application

In Section 3.7 we created data sets (n=61) with variance-covariance structure set as $\Sigma_3 = 0.3 \ast J_{11} + \text{Autoregressive}(\tau^2 = 0.1, \rho = 0.8)$. Here we add 39 subjects whose marker information is incomplete into those data sets, that is now the data set has $n = 100$ subjects. We use a REML estimate of the underlying dependence matrix $\Sigma$ from a saturating model as recommended in the Chapter 3. We also calculate the P-value using the true underlying variance-covariance matrix $\Sigma_3$. Table 4–1 shows the result from only subjects whose marker information is complete and the full data set including subjects missing markers. The smaller P-values resulted from analysis of full data set obviously suggests the power-boosting advantage of using more information. Again, the performance of the REML estimate of $\Sigma$ is just slightly worse than using the true underlying $\Sigma$.

To study the influence of using more subjects, especially those with partial genotypic information, Figure 4–1 shows the histograms of the LRT statistics under $H_0$ from different combination of methods handling missing data: case-wise deleted data and full data, and methods to estimate $\Sigma$ : REML estimate and true $\Sigma$. Table 4–2 lists the corresponding critical values from these four different analysis procedures. From these we can see that the null distribution of LRT statistic calculated from case-wise deleted data has longer tail than the one calculated using the full information, that is, at the same significance level, the threshold value is higher if we use case-wise deleted data.
Figure 4–1: Histograms of LRT statistics from different combination of methods handling missing data and methods to estimate $\Sigma$ under $H_0$.

Table 4–2: Comparison of critical values from different analyzing methods

<table>
<thead>
<tr>
<th>$\alpha$ level</th>
<th>case-wise deleted data</th>
<th>full data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>REML</td>
<td>True</td>
</tr>
<tr>
<td>.10</td>
<td>12.67454</td>
<td>14.21031</td>
</tr>
<tr>
<td>.05</td>
<td>14.36187</td>
<td>16.02444</td>
</tr>
<tr>
<td>.01</td>
<td>17.77771</td>
<td>19.78819</td>
</tr>
</tbody>
</table>
Figure 4–2: The trend of P-values and their 95% confidence intervals (the dotted lines) from using case-wise deleted data (the dashed line) and using full data set (the black line). Underlying phenotypic lines are two flat lines and true covariance matrix is used in analysis.
It is straightforward to have power-boosting advantage when subjects with partial marker information are included, which is Figure 4–2 demonstrates. In this figure, 100 data sets using same marker file as the one used in Table 4–1 are simulated from two underlying flat phenotypic curves. Variance-covariance matrix is \( \Sigma_3 = 0.3 \ast J_{11} + \text{Autoregressive}(\tau^2 = 0.1, \rho = 0.8) \) and is directly used in NPFIM analysis. When the distance between these two flat lines increases, we get decreasing P-values either using case-wise deleted data only or using full data. But when we use full data to do the analysis, the resulted P-values are consistently smaller than those resulted from using case-wise deleted data only.

4.6 Another Way to Handle All Missing Marker Cases – Gibbs Sampler

In the previous sections, the proposed method assumes that for any marker interval all subjects at most lose the record of one of the two end markers. In other words, subjects whose phenotypic measurements are recorded but marker information for some particular marker interval is lost are not taken into account when we calculate the LRT statistic for that marker interval. In this section, we propose a new procedure to handle all the missing marker cases including those subjects whose both end markers for some interval are not available.

Consider interval mapping procedure based on a backcross design with parents having alleles \( Qq \) and \( qq \) (or 10 and 00). Given \( k \)th marker interval bracketed by markers \( \mathcal{M}_k \) and \( \mathcal{M}_{k+1} \) (suppose both markers are recorded in the data set), subject \( i \) with phenotypic observations \( y_i \) has density function as follows:

\[
y_i|k, \bar{M} \sim \phi_{1,i}I(\mathcal{M}_k\mathcal{M}_{k+1} = 11) + [(1 - \theta)\phi_{1,i} + \theta\phi_{2,i}]I(\mathcal{M}_k\mathcal{M}_{k+1} = 10) + [(1 - \theta)\phi_{2,i} + \theta\phi_{1,i}]I(\mathcal{M}_k\mathcal{M}_{k+1} = 01) + \phi_{2,i}I(\mathcal{M}_k\mathcal{M}_{k+1} = 00) \equiv \sum_{j=1}^{4} f_{ij} I_j^{(k)}
\]
where $\phi_{1,i}$ and $\phi_{2,i}$ are the density function for $y_i \sim MVN(\mu_{Qq}, \Sigma)$ and $y_i \sim MVN(\mu_{qq}, \Sigma)$, $I(x)$ is indicator function with value 1 if $x$ is true, and $\mathcal{M}$ denote all the marker information along the genome/linkage group for subject $i$.

If for every subject all the markers information is available, the likelihood function for unknown QTL indexed by the $k$th interval, $k = 1, \ldots, m$, and $\theta, 0 < \theta < 1$, can be easily derived as

$$l(k, \theta | \mathbf{y}, \mathcal{M}_{\text{comp}}) = \sum_{l=1}^{m} f(y_{\mathcal{M}_{\text{comp}}, k, \theta}) I(k = l)$$

$$= \sum_{l=1}^{m} (\prod_{i=1}^{n} \sum_{j=1}^{4} f_{ij} I_{j}^{(k)}) I(k = l),$$

where $\mathcal{M}_{\text{comp}}$ denotes each subject’s genotypes of all markers. For each fixed $k$, the likelihood function for $\theta$ given all complete data is

$$l(\theta | \mathbf{y}, \mathcal{M}_{\text{comp}}, k) = f(y_{\mathcal{M}_{\text{comp}}, k, \theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{4} f_{ij} I_{j}^{(k)}.$$

(4–1)

If there is missing marker information when we consider $k$th marker interval, the likelihood function for the unknown QTL based on the observed marker information $\mathcal{M}_{\text{obs}}$ can be expressed explicitly. Suppose subjects from $1, \ldots, n_1$ have both marker values for markers $\mathcal{M}_k$ and $\mathcal{M}_{k+1}$, subjects from $n_1 + 1, \ldots, n_2$ only have information about $\mathcal{M}_k$, subjects from $n_2 + 1, \ldots, n_3$ only have information about $\mathcal{M}_{k+1}$ and subjects from $n_3 + 1, \ldots, n$ have no record on either marker. Then
for the fixed $k$th marker interval, the likelihood function for $\theta$ is

$$l(\theta|y, \tilde{M}_{\text{obs}}, k) = f(y|\theta, \tilde{M}_{\text{obs}}, k) = \prod_{i=1}^{n_1} \sum_{j=1}^{4} f_j^{(k)} I(\tilde{M}_{\text{obs}} = (M_k, M_{k+1}))$$

$$\prod_{i=n_1+1}^{n_2} \left[ ((1 - r\theta)\phi_{1,i} + r\theta\phi_{2,i})I(M_k = 1) + [(1 - r\theta)\phi_{2,i} + r\theta\phi_{1,i}]I(M_k = 0) \right]$$

$$\prod_{i=n_1}^{n_1+1} \left[ ((1 - r + r\theta)\phi_{1,i} + r(1 - \theta)\phi_{2,i})I(M_{k+1} = 1) + [(1 - r + r\theta)\phi_{2,i} + r(1 - \theta)\phi_{1,i}]I(M_{k+1} = 0) \right]$$

$$\prod_{i=n_2+1}^{n_3} (\phi_{1,i}/2 + \phi_{2,i}/2) I(\tilde{M}_{\text{obs}} = \emptyset).$$

Thus we can derive the distribution of the missing markers, $\tilde{M}_{\text{mis}}$, in $k$th marker interval conditional on observed data, $(y, \tilde{M}_{\text{obs}})$, by

$$f(\tilde{M}_{\text{mis}}|y, \tilde{M}_{\text{obs}}, k, \theta) = \frac{f(y|\tilde{M}_{\text{comp}}, k, \theta)f(\tilde{M}_{\text{mis}}|\tilde{M}_{\text{obs}}, k, \theta)}{f(y|\tilde{M}_{\text{obs}}, k, \theta)}. \quad (4-2)$$

Notice that $\tilde{M}_{\text{comp}} = (\tilde{M}_{\text{obs}}, \tilde{M}_{\text{mis}})$, so equation $(4-2)$ is just two different ways to factorize the joint distribution of $y$ and $\tilde{M}_{\text{mis}}$ given $\tilde{M}_{\text{obs}}$ and other parameters.

And $f(\tilde{M}_{\text{mis}}|\tilde{M}_{\text{obs}}, k, \theta)$ actually does not depend on $\theta$, which has expression as following:

$$f(\tilde{M}_{\text{mis}}|\tilde{M}_{\text{obs}}, k)$$

$$= 1I(\tilde{M}_{\text{obs}}, k = (M_k, M_{k+1}))$$

$$+ (1 - r)I(\tilde{M}_{\text{mis}}, k = M_k = 1|\tilde{M}_{\text{obs}}, k = M_{k+1} = 1)$$

$$+ rI(\tilde{M}_{\text{mis}}, k = M_k = 0|\tilde{M}_{\text{obs}}, k = M_{k+1} = 1)$$

$$+ (1 - r)I(\tilde{M}_{\text{mis}}, k = M_k = 0|\tilde{M}_{\text{obs}}, k = M_{k+1} = 0)$$

$$+ rI(\tilde{M}_{\text{mis}}, k = M_k = 1|\tilde{M}_{\text{obs}}, k = M_{k+1} = 0)$$

$$+ (1 - r)I(\tilde{M}_{\text{mis}}, k = M_{k+1} = 1|\tilde{M}_{\text{obs}}, k = M_k = 1)$$

$$+ rI(\tilde{M}_{\text{mis}}, k = M_{k+1} = 0|\tilde{M}_{\text{obs}}, k = M_k = 1) \quad (4-4)$$
Using equation (4-2), it is easy to get the conditional probabilities of missing marker types for each subject $i$ whose marker information is incomplete in the $k$th marker interval conditional on all observed phenotypic and genotypic data, which are listed in the Table 4–3 and Table 4–4.
Table 4–3 and Table 4–4 give the conditional distribution \( f(\tilde{M}_{\text{mis}}|k, \theta, y, M_{\text{obs}}) \). Once we have \( f(k|\theta, \tilde{M}_{\text{mis}}, y, M_{\text{obs}}) \) and \( f(\theta|k, \tilde{M}_{\text{mis}}, y, M_{\text{obs}}) \), we can use the following Gibbs sampler algorithm to find \((\hat{k}, \hat{\theta})\), which specifies possible positions of the QTL.

**Gibbs Sampler:** Given \((k^{(t)}, \theta^{(t)}, \tilde{M}_{\text{mis}}^{(t)})\), generate

1. \( \tilde{M}_{\text{mis}}^{(t+1)} \sim f(\tilde{M}_{\text{mis}}|k^{(t)}, \theta^{(t)}, y, M_{\text{obs}}) \)
2. \( k^{(t+1)} \sim f(k|\theta^{(t)}, \tilde{M}_{\text{mis}}^{(t+1)}, y, M_{\text{obs}}) \)
3. \( \theta^{(t+1)} \sim f(\theta|k^{(t+1)}, \tilde{M}_{\text{mis}}^{(t+1)}, y, M_{\text{obs}}) \)

Suppose the prior distribution for \( k \) is discrete uniform distribution with probability \( P(k = l) = \frac{1}{m}, l = 1...m \). Then we have

\[
P(k = k^*|\theta, y, \tilde{M}_{\text{comp}}) = \frac{f(k = k^*, y|\tilde{M}_{\text{comp}}, \theta)}{m(y|\tilde{M}_{\text{comp}}, \theta)} = \frac{f(y|k = k^*, \tilde{M}_{\text{comp}}, \theta) \frac{1}{m}}{\sum_{l=1}^{m} f(y|k = l, \tilde{M}_{\text{comp}}, \theta) \frac{1}{m}}
\]

where

\[
S(l) = \left( \prod_{i=1}^{n_1} \phi_{1,i} \prod_{i=n_1+1}^{n_2} [(1 - \theta) \phi_{1,i} + \theta \phi_{2,i}] \prod_{i=n_2+1}^{n_3} [(1 - \theta) \phi_{2,i} + \theta \phi_{1,i}] \prod_{i=n_3+1}^{n} \phi_{2,i} \right).
\]

The posterior distribution of \( k \) given \((\theta, y, \tilde{M}_{\text{comp}})\) is also a multinomial distribution, which is rather easy to generate random variables from.

Naturally we let the prior distribution of \( \theta \) be uniformly distributed in the interval \([0, 1] \). It is straightforward to write down the posterior distribution
\[ f(\theta|k, \tilde{M}_{\text{comp}}, y) \] as follows:

\[
\frac{f(\theta|k, \tilde{M}_{\text{comp}}, y)}{m(y|k, \tilde{M}_{\text{comp}})} = \frac{f(y|k, \theta, \tilde{M}_{\text{comp}})\pi(\theta)}{\int_0^1 f(y|k, \theta, \tilde{M}_{\text{comp}})\pi(\theta)d\theta} \propto n_k 2 \prod_{i=n_k^1+1}^{n_k^2} [(1 - \theta)\phi_{1,i} + \theta\phi_{2,i}] \prod_{i=n_k^2+1}^{n_k^3} [(1 - \theta)\phi_{2,i} + \theta\phi_{1,i}].
\]

The main part of this distribution is actually a polynomial function of \( \theta \). We could use Accept-Reject algorithm or Metropolis-Hastings algorithm to generate random variables from it.

But notice that the coefficient vectors \( \xi_1 \) and \( \xi_2 \) are unknown and when there does exist a QTL, we can not directly express them through observed \( Y \)s and markers. So instead of getting joint distribution of \( (k, \theta, \tilde{M}_{\text{mis}}) \), in the following we propose another Gibbs Sampler algorithm to find possible QTL from the joint distribution of \( (k, \theta, \tilde{M}_{\text{mis}}, \tilde{Z}_{\text{mis}}) \), where \( \tilde{Z} = (Z_1, Z_2, ..., Z_n) \) represents the allele vector of QTL for subject 1 to subject \( n \). Define \( Z_i = 1 \) for allele \( Qq \) and \( Z_i = 0 \) for allele \( qq \). For the \( k \)th interval bracketed by \( \mathcal{M}_k \mathcal{M}_{k+1} \), only when \( \mathcal{M}_k \mathcal{M}_{k+1} \) is observed as 11 or 00 \( Z \) is known as 1 or 0, respectively. So \( \tilde{Z}_{\text{mis}} = (Z_{n_1+1}, ..., Z_{n_3}) \). Naturally when \( \mathcal{M}_k \mathcal{M}_{k+1} \) is observed as 10 or 01, \( Z \) is defined to be distributed as \( \text{Bernoulli}(1 - \theta) \) and \( \text{Bernoulli}(\theta) \), respectively. That is,

\[
Z_i|\tilde{M}_{\text{comp}}, k, \theta = 1 I(\mathcal{M}_k \mathcal{M}_{k+1} = 11) \\
\sim \text{Bernoulli}(1 - \theta) I(\mathcal{M}_k \mathcal{M}_{k+1} = 10) \\
\sim \text{Bernoulli}(\theta) I(\mathcal{M}_k \mathcal{M}_{k+1} = 01) \\
= 0 I(\mathcal{M}_k \mathcal{M}_{k+1} = 00).
\]
Of course for observed phenotypic values \( y \) we have

\[
f(y_i | \tilde{Z}, \tilde{M}_{\text{comp}}, k, \theta) = \phi_{1,i} I(M_k M_{k+1} = 11)
+ \phi_{2,i}^1 Z_k^i I(M_k M_{k+1} = 10)
+ \phi_{1,i}^1 Z_k^i I(M_k M_{k+1} = 01)
+ \phi_{2,i} I(M_k M_{k+1} = 00),
\]

where \( \phi_1 \) and \( \phi_2 \) are two multivariate normal distributions corresponding to two underlying phenotypic curves in backcross design. Therefore, we have the joint distribution of \((y_i, Z_i)\) given \((k, \theta, \tilde{M}_{\text{comp}})\) as the following

\[
f(y_i, Z_i | k, \theta, \tilde{M}_{\text{comp}}) = \phi_{1,i} I(M_k M_{k+1} = 11)
+ \phi_{2,i}^1 Z_k^i (1 - \theta)^Z_k^i I(M_k M_{k+1} = 10)
+ \phi_{1,i}^1 Z_k^i (1 - \theta) I(M_k M_{k+1} = 01)
+ \phi_{2,i} I(M_k M_{k+1} = 00),
\]

The following algorithm is our new Gibbs Sampler procedure.

**Gibbs Sampler 2:** Given \((k^{(t)}, \theta^{(t)}, \tilde{M}_{\text{mis}}^{(t)}, \tilde{Z}_{\text{mis}}^{(t)})\), generate

1. \( \tilde{M}_{\text{mis}}^{(t+1)} \sim f(\tilde{M}_{\text{mis}}^{(t)} | k^{(t)}, \theta^{(t)}, \tilde{Z}_{\text{mis}}^{(t)}, y, \tilde{M}_{\text{obs}}) \)
2. \( k^{(t+1)} \sim f(k^{(t)} | \theta^{(t)}, \tilde{M}_{\text{mis}}^{(t+1)}, \tilde{Z}_{\text{mis}}^{(t)}, y, \tilde{M}_{\text{obs}}) \)
3. \( \theta^{(t+1)} \sim f(\theta^{(t)} | k^{(t+1)}, \tilde{M}_{\text{mis}}^{(t+1)}, \tilde{Z}_{\text{mis}}^{(t)}, y, \tilde{M}_{\text{obs}}) \)
4. \( \tilde{Z}_{\text{mis}}^{(t+1)} \sim f(\tilde{Z}_{\text{mis}}^{(t)} | k^{(t+1)}, \theta^{(t+1)}, \tilde{M}_{\text{mis}}^{(t+1)}, y, \tilde{M}_{\text{obs}}) \)

Below we derive the posterior distributions in the above Gibbs Sampler algorithm.
Table 4–5: Distribution of missing marker conditional on the other observed marker for kth marker interval for subject i.

<table>
<thead>
<tr>
<th>Observed marker</th>
<th>Missed marker</th>
<th>$M_{k+1} = 1$</th>
<th>$M_{k+1} = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_k = 1$</td>
<td>(1-$r$)$\phi_{1,i}$</td>
<td>$r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
<td>(1-$r$)$\phi_{1,i}+r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
</tr>
<tr>
<td>$M_k = 0$</td>
<td>$r(\phi_{1,i}\theta)^{Z_i}(\phi_{2,i}(1-\theta))^{1-Z_i}$</td>
<td>(1-$r$)$\phi_{1,i}+r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
<td>(1-$r$)$\phi_{1,i}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observed marker</th>
<th>Missed marker</th>
<th>$M_k = 1$</th>
<th>$M_k = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{k+1} = 1$</td>
<td>(1-$r$)$\phi_{1,i}$+r[\phi_{2,i}(1-\theta)]^{Z_i}(\phi_{1,i}\theta)^{2-Z_i}$</td>
<td>$r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
<td>(1-$r$)$\phi_{1,i}+r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
</tr>
<tr>
<td>$M_{k+1} = 0$</td>
<td>$r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
<td>(1-$r$)$\phi_{2,i}$+r[\phi_{1,i}(1-\theta)]^{Z_i}(\phi_{2,i}\theta)^{1-Z_i}$</td>
<td>(1-$r$)$\phi_{2,i}$</td>
</tr>
</tbody>
</table>

1. $f(\tilde{M}_{mis}|k, \theta, \tilde{Z}, \tilde{M}_{obs}, y)$: It is straightforward to have

\[
f(\tilde{M}_{mis}|k, \theta, \tilde{Z}, \tilde{M}_{obs}, y) = f(\tilde{M}_{mis}, y, \tilde{Z}|k, \theta, \tilde{M}_{obs}) \cdot m(y, \tilde{Z}|k, \theta, \tilde{M}_{obs})
\]

\[
= \frac{f(y, \tilde{Z}|k, \theta, \tilde{M}_{comp})f(\tilde{M}_{mis}|\tilde{M}_{obs}, k, \theta)}{\sum_{\tilde{M}_{mis}} f(y, \tilde{Z}|k, \theta, \tilde{M}_{comp})f(\tilde{M}_{mis}|\tilde{M}_{obs}, k, \theta)}
\]

$f(\tilde{M}_{mis}|\tilde{M}_{obs}, k, \theta)$ actually does not depend on $\theta$ with expression stated before.

The conditional probabilities of missing marker types for each subject $i$ whose marker information is incomplete in $k$th marker interval conditional on all observed phenotypic and genotypic data in addition to the latent variable $Z_i$ are listed in the Table 4–5 and Table 4–6.

2. $f(k|\theta, \tilde{Z}, \tilde{M}_{comp}, y)$:
As before suppose prior distribution for \( k \) is discrete uniform distribution with probability \( P(k = l) = \frac{1}{m}, l = 1...m \). Then we have

\[
P(k = k^*|\theta, \mathbf{y}, \mathbf{Z}, \tilde{\mathcal{M}}_{\text{comp}}) = \frac{f(k = k^*, \mathbf{y}, \mathbf{Z}|\theta, \tilde{\mathcal{M}}_{\text{comp}})}{m(\mathbf{y}, \mathbf{Z}|\theta, \tilde{\mathcal{M}}_{\text{comp}})} \]

\[
= \frac{f(\mathbf{y}, \mathbf{Z}|k = k^*, \tilde{\mathcal{M}}_{\text{comp}}, \theta) \frac{1}{m}}{\sum_{l=1}^{m} f(\mathbf{y}, \mathbf{Z}|k = l, \tilde{\mathcal{M}}_{\text{comp}}, \theta) \frac{1}{m}}
\]

\[
= \frac{S(k^*)}{\sum_{l=1}^{m} S(l)}
\]

where

\[
S(l) = \prod_{i=1}^{n_1} \phi_{1,i} \prod_{i=n_1+1}^{n_2} [\phi_{1,i}^{Z_{1,i}} (1 - \theta)^{Z_{1,i}}(1 - Z_{1,i})^{Z_{1,i}} \prod_{i=n_1+1}^{n_3} [\phi_{2,i}^{Z_{2,i}} (1 - \theta)^{Z_{2,i}}(1 - Z_{2,i})^{Z_{2,i}} \prod_{i=n_1+1}^{n} \phi_{2,i}.
\]

The posterior distribution of \( k \) given \((\theta, \mathbf{y}, \mathbf{Z}, \tilde{\mathcal{M}}_{\text{comp}})\) is also a multinomial distribution, which is rather easy to generate random variables from.
3. $f(\theta|k, \tilde{Z}, \tilde{M}_{\text{comp}}, y)$ : Again let $\pi(\theta)$ be $\text{uniform}(0, 1)$, then

$$f(\theta|k, \tilde{Z}, \tilde{M}_{\text{comp}}, y) = \frac{f(\theta, y, \tilde{Z}|k, \tilde{M}_{\text{comp}})}{m(y, \tilde{Z}|k, \tilde{M}_{\text{comp}})} = \frac{f(y, \tilde{Z}|k, \tilde{M}_{\text{comp}})\pi(\theta)}{\int_0^1 f(y, \tilde{Z}|k, \tilde{M}_{\text{comp}})\pi(\theta)d\theta}$$

$$\propto \theta^{\sum_{i=n_k^1+1}^{n_k^2} (1-Z_i^k) + \sum_{i=n_k^2+1}^{n_k^3} Z_i^k} (1-\theta)^{\sum_{i=n_k^1+1}^{n_k^2} Z_i^k + \sum_{i=n_k^2+1}^{n_k^3} (1-Z_i^k)}.$$

Define $a = \sum_{i=n_k^1+1}^{n_k^2} (1-Z_i^k) + \sum_{i=n_k^2+1}^{n_k^3} Z_i^k + 1$ and $b = \sum_{i=n_k^1+1}^{n_k^2} (1-Z_i^k) + \sum_{i=n_k^2+1}^{n_k^3} Z_i^k + 1$, then the posterior distribution of $\theta$ given all other information a $\beta$ distribution with parameter $a$ and $b$. When $a = 1$ and $b = 1$, which happens when there is no recombination of two bracketed markers, $\theta$ is actually non-identifiable. In this case, we can let $\theta \sim \text{uniform}(0, 1)$.

4. $f(\tilde{Z}|k, \theta, \tilde{M}_{\text{comp}}, y)$ :

$$f(\tilde{Z}|k, \theta, \tilde{M}_{\text{comp}}, y) = \frac{f(\tilde{Z}, y|k, \theta, \tilde{M}_{\text{comp}})}{m(y|k, \theta, \tilde{M}_{\text{comp}})} = \frac{f(\tilde{Z}, y|k, \theta, \tilde{M}_{\text{comp}})}{\sum_{\tilde{Z}} f(\tilde{Z}, y|k, \theta, \tilde{M}_{\text{comp}})}.$$

Therefore, for $i = n_k^1 + 1, ..., n_k^2$,

$$Z_i|k, \theta, \tilde{M}_{\text{comp}}, y = Z_i^k|\theta, y \sim \text{Bernoulli} \left( \frac{\phi_{1,i}(1-\theta)}{[\phi_{1,i}(1-\theta) + \phi_{2,i}\theta]} \right)$$

and for $i = n_k^2 + 1, ..., n_k^3$,

$$Z_i|k, \theta, \tilde{M}_{\text{comp}}, y = Z_i^k|\theta, y \sim \text{Bernoulli} \left( \frac{\phi_{1,i}\theta}{[\phi_{1,i}\theta + \phi_{2,i}(1-\theta)]} \right).$$

Notice that in all the above formulae we actually assume that the two coefficient vectors $\xi_1$ and $\xi_2$ corresponding to two underlying phenotypic curves are
known. But notice that if we put flat prior on $\xi_1$, then

$$f(\xi_1, \xi_2 | \tilde{Z}, k, \theta, \tilde{M}_{comp}, y)$$  \hspace{1cm} (4-5)$$

$$\propto f(y | \xi_1, \xi_2, \tilde{Z}, k, \theta, \tilde{M}_{comp})$$

$$\propto f(y | \xi_1, \xi_2, \tilde{Z}, k)$$

$$\propto \exp \left( -\frac{1}{2} \sum_{(Z^k_i = 1)} ||y - B\xi_1||^2_\Sigma - \frac{1}{2} \sum_{(Z^k_i = 0)} ||y - B\xi_2||^2_\Sigma \right).$$

Suppose there are $n_{Z^k}$ subjects with $Z^k_i = 1$ and $y$ is rearranged so that for $i = 1, \ldots, n_{Z^k}$, $y_i$ has value 1 for $Z^k_i$ while for $i = n_{Z^k} + 1, \ldots, n$, $y_i$ has value 0 for $Z^k_i$. Thus from (4-5) it is straightforward to see that the posterior distributions of $\xi_1$ and $\xi_2$ are independent to each other with the following multivariate normal distributions:

$$\xi_1 | \tilde{Z}, k, \theta, \tilde{M}_{comp}, y \sim MVN \left( (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\sum_{i=1}^{n_{Z^k}}y_i, \frac{(B'\Sigma^{-1}B)^{-1}}{n_{Z^k}} \right)$$

$$\xi_2 | \tilde{Z}, k, \theta, \tilde{M}_{comp}, y \sim MVN \left( (B'\Sigma^{-1}B)^{-1}B'\Sigma^{-1}\sum_{i=n_{Z^k}+1}^{n}y_i, \frac{(B'\Sigma^{-1}B)^{-1}}{n - n_{Z^k}} \right).$$

Therefore, practically we need the following Gibbs Sampler algorithm:

Given $(k^{(t)}, \theta^{(t)}, \tilde{M}_{mis}^{(t)}, \tilde{Z}_{mis}^{(t)}, \xi_1^{(t)}, \xi_2^{(t)})$, generate:

1. $\tilde{M}_{mis}^{(t+1)} \sim f(M_{mis} | k^{(t)}, \theta^{(t)}, \tilde{Z}_{mis}^{(t)}, \xi_1^{(t)}, \xi_2^{(t)}, y, \tilde{M}_{obs})$
2. $k^{(t+1)} \sim f(k | \theta^{(t)}, \tilde{M}_{mis}^{(t+1)}, \tilde{Z}_{mis}^{(t)}, \xi_1^{(t)}, \xi_2^{(t)}, y, \tilde{M}_{obs})$
3. $\theta^{(t+1)} \sim f(\theta | k^{(t+1)}, \tilde{M}_{mis}^{(t+1)}, \tilde{Z}_{mis}^{(t)}, \xi_1^{(t)}, \xi_2^{(t)}, y, \tilde{M}_{obs})$
4. $\tilde{Z}_{mis}^{(t+1)} \sim f(\tilde{Z}_{mis} | k^{(t+1)}, \theta^{(t+1)}, \tilde{M}_{mis}^{(t+1)}, \xi_1^{(t)}, \xi_2^{(t)}, y, \tilde{M}_{obs})$
5. $\xi_1^{(t+1)} \sim f(\xi_1 | k^{(t+1)}, \theta^{(t+1)}, \tilde{M}_{mis}^{(t+1)}, \tilde{Z}_{mis}^{(t+1)}, \xi_2^{(t)}, y, \tilde{M}_{obs})$
6. $\xi_2^{(t+1)} \sim f(\xi_2 | k^{(t+1)}, \theta^{(t+1)}, \tilde{M}_{mis}^{(t+1)}, \tilde{Z}_{mis}^{(t+1)}, \xi_1^{(t+1)}, y, \tilde{M}_{obs})$

4.6.1 Simulation

We use a simulated data to test the performance of this Gibbs sampler procedure. Assume the studied linkage group has 10 markers, that is, there are nine intervals with distance 20cM. QTL is located on the 7th interval, 18cM from
the seventh marker. 100 subjects with 15 continuous seasons’ growth datas are generated and the heritability on year 8 is set to be .4. The mathematical forms for the two underlying growth curves are \(20/[1+20 \exp(-.6t)]\) and \(30/[1+27 \exp(-.9t)]\). For each marker, we randomly cancel out 10 subjects’ marker types. As suggested in Chapter 3, when we consider \(k\)th interval, we use interval-wise deleted REML estimate of the covariance structure, that is, the REML estimate from all subjects who have both information of marker \(k\) and marker \(k + 1\).

The Gibbs sampler sampled 50,000 times. The algorithm quickly finds the 7th interval and stays there. The histograms of \(\theta\) based on the last \(n\) samples are shown in Figure 4–3, \(n = 10,000, 20,000, 30,000, 40,000\). As we can see that the algorithm arrives the stationary distribution rather quickly. The median and mean for \(\theta\) is .915 while the actual QTL is assumed at .917.

When there is no QTL, the histograms of the interval index \(k\) and \(\theta\) conditional on \(k\) are illustrated in Figure 4–4. There are no obvious peaks in any of the histogram which indicates that the algorithm does not find any QTL. In this Figure, the underlying true covariance matrix is used. If we use the REML estimate from interval-wise deleted data, the outcome is shown in Figure 4–5 and Figure 4–6. Figure 4–5 suggests \(\hat{k} = 1\) and in Figure 4–6 the histogram of \(\theta\) corresponding to \(k = 1\) is very similar to uniform distribution.

4.6.2 Application

We also applied this Gibbs sampler to the poplar data set. In the Chapter 3, we found that in linkage group 10, there is only one single QTL candidate which is located in the first interval. So here, we only use the marker information from linkage group 10 also. There are 8 markers in that linkage group. There are 78 trees in the dataset, and 61 of them have information of all markers recorded. For the covariance matrix, as before we use the REML estimate for the log-transformed data.
Figure 4–3: This figure shows the histogram of $\theta$ given $k = 7$ from the simulated dataset where QTL is assumed in interval 7. The REML estimate from interval-wise deleted data is used in calculation. “n” is the number of last drawn samples used to plot the histogram.
Figure 4–4: The histograms of the interval index $k$ and $\theta$ conditional on $k$ from the simulated dataset under $H_0$. The true covariance matrix is used in calculation.
Figure 4–5: The histograms of the interval index $k$ from the last 25,000 iterations (left) and the last 10,000 iterations (right). The REML estimate from interval-wise deleted data is used. Data set is simulated under $H_0$. 
Figure 4–6: The histograms of $\theta$ from the last 10,000 iterations. The REML estimate from interval-wise deleted data is used. Data set is simulated under $H_0$. 
Figure 4–7: The histograms of \((k, \theta)\) from Gibbs sampler. The left figure shows the histogram of the possible interval where QTL falls in. The right figure shows the histogram of \(\theta\) given \(k = 1\).

We do Gibbs sampler for 10,000 times, and we use the samples from last 5,000 iterations to get the joint distribution of \((k, \theta)\). (After about 1500 times, the algorithm falls in the 1st interval and never gets out of it.) The histograms are in Figure 4–7. We can directly see that the QTL candidate is in the first interval. The median and mean of \(\theta\) given \(k = 1\) is .606 and .603, respectively, which corresponds to about 7.8cM from the first marker. The 95% confidence interval for \(\theta\) from this empirical posterior distribution is [.4, .78], that is, [5cM, 10.3cM] from the first marker. This result is consistent with previous analysis.

4.7 Discussion

In this chapter, we proposed two quite different methods to include all subjects with incomplete markers into mapping procedure: one is in frequentist way and the other is in Baysian framework. Because of more information used our methods
naturally have more power. The first method explicitly expresses the conditional distribution of phenotypic observations $y$ given on all observed markers so that it is easy to use the likelihood principle to find the MLE’s of the unknown coefficient vectors and construct the final test statistic for finding significant evidence of existing QTL. The likelihood map across the whole linkage group/genome can be obtained and hence all QTL candidates implied by the peaks over threshold values are detected very easily. The second method uses a Gibbs sampler algorithm which can provide the posterior distribution of QTL specified by interval $k$ and location in that interval $\theta$. The posterior variance and the empirical confidence interval about $\theta$ is straightforward to calculate. The significance test is also embedded in the posterior distribution of the parameters and hence it is not necessary to obtain the critical value for a test statistic to declare significance. But since we only assume that there is one QTL in the whole genome, this algorithm might not find all QTLs if there do exist several QTLs together controlling the different phenotypic curves.

In the following simulation studies, we test the performance of our proposed Gibbs sampler algorithm when there are two QTLs together controlling the phenotypic trajectories. The four curves corresponding to four possible combination of these two QTLs are shown in Figure 4–8. We assume there are 9 intervals with equal distance 20cM. 15 observations from 100 subjects are recorded. The heritability at the 8th measurement points is set as .4 in Figure 4–9 and .1 in Figure 4–10, Figure 4–11. The two QTLs are set to be 18cM and 10cM from their left bracketed marker, respectively. In Figure 4–10 and Figure 4–9 the two QTLs are located in the second interval and the eighth interval but the algorithm only suggests that in the second interval there exists a QTL. Figure 4–11 where the two QTLs are assumed to be in the fourth interval and the sixth interval implies QTL locates in the sixth interval even though the algorithm starts from interval 4.
Figure 4–8: The underlying four different phenotypic curves controlled by two QTLs together.
Figure 4–9: The plot shows the histogram of $\theta$ given $k = 2$. Median of $\theta$ is .849 while the true $\theta$ is .917. The REML estimate from interval-wise deleted data is used in calculation. Heritability value at the eight observation point is set to be .4. Two QTLs are assumed to be in interval 2 and interval 8, respectively. Start value for $k$ is 4. The last 10,000 samples from 50,000 iterations are used to make the plots.
Figure 4–10: The histogram of \((k, \theta)\) from the last 10,000 samples of 50,000 iterations. Two QTLs are assumed to be in interval 2 and interval 8, respectively. The left plot shows the histogram of interval index which gives \(\hat{k} = 2\). The right plot shows the histogram of \(\theta\) given \(k = 2\). Median of \(\theta\) is .749 while the true \(\theta\) is .917. The REML estimate from interval-wise deleted data is used in calculation. Heritability value at the eight observation point is set to be .1. Start value for \(k\) is 6.
Figure 4–11: The histogram of \((k, \theta)\) from the last 10,000 samples of 50,000 iterations. Two QTLs are assumed to be in interval 4 and interval 6, respectively. The left plot shows the histogram of interval index which gives \(\hat{k} = 6\). The right plot shows the histogram of \(\theta\) given \(k = 6\). Median of \(\theta\) is .77 while the true \(\theta\) is .549. The REML estimate from interval-wise deleted data is used in calculation. Heritability value at the eight observation point is set to be .1. Start value for \(k\) is 4.
CHAPTER 5  
FUTURE STUDY

In practice, usually we are faced with a missing data situation for a lot of reasons. Some researcher or statistical package simply discards the incomplete cases and uses only those cases with full observations to carry out a complete-case analysis. Clearly this method is inefficient because it is based on the strong assumption that those complete cases are a random subsample of all possible cases. Thus many statistical methods that incorporate both complete cases and incomplete cases have been proposed. These statistical models are mainly from two model families: pattern-mixture models and selection models. The main difference between these two models is the different way to factor the joint distribution of the observed data $Y$ and the unobserved data indicator $M$.

A selection model specifies the joint distribution of the available measurement and missing mechanism by the product of the marginal measurement distribution and the missing mechanism distribution, conditional on measurement, that is, $f(Y, M) = m(Y)f(M|Y)$ where $m(Y)$ represents the complete-data model for $Y$. The pattern mixture model factors the joint distribution of $Y$ and $M$ into the distribution of the missing data pattern $M$ and the conditional distribution of measurement on specific missing pattern, $f(Y, M) = m(M)f(Y|M)$. The selection model is more appealing when the marginal distribution of measurements is of interest. However, the selection model should be applied with caution especially in the context of nonrandom missing mechanism even though it is identifiable (Glynn et al. (1986)). So the pattern mixture model attracts more interest.

When $M$ is independent of $Y$, it is called missing complete at random (MCAR). Less restriction is assumed in missing at random (MAR), where the conditional
distribution of $\mathbf{M}$ does not depend on the unobserved $\mathbf{Y}$, that is, $f(\mathbf{M}|\mathbf{Y}) = f(\mathbf{M}|\mathbf{Y}_{obs})$. It is straightforward to get the more general category: missing not at random (MNAR). Kenward et al. (2003) further split MNAR to non-future-dependent and future-dependent. In non-future dependent mechanisms missingness is allowed to be dependent on present possibly non-response but not on future measurements. This class contains the important subfamily of MAR models while putting a sensible restriction for MNAR. Kenward et al. (2003) argued that non-future missing values family is not equivalent to the interior family named by Thijs et al. (2002), but they both have MAR models as their intersection.

The pattern mixture model is nonidentifiable by construction because the missing data provides no information about the distribution of the measurements conditional on the incomplete data pattern. To solve this problem, Little (1993, 1994) proposed identifying restrictions by setting the inestimable parameters of incomplete data pattern equal to those of complete data cases, which was called the complete-case missing value restriction. Other restrictions include the neighbouring-case missing values restriction which borrows information from closest available pattern, and the available-case missing values restriction which needs to consider a particular linear combination and is shown to be equivalent to missing at random in a selection model (Molenberghs et al. (1998)). All these restrictions are in the 'interior' family of identifying-restrictions since they all use the observable distribution to study the unobservable distribution (Kenward et al. (2003)). Thijs et al. (2002) studied these three restrictions in detail.

The missing mechanism for non-responses is not identified from the observed data. So we have to make some unverified assumptions. Many researchers (e.g. Rosenbaum and Rubin (1983, 1985); Nordheim (1984); Little and Rubin (1987); Scharfstein et al. (1999)) have pointed out the importance of conducting sensitivity analyses for a range of plausible assumptions. Birmingham et al. (2003) conducted
a sensitivity analysis for three pattern-mixture-type unverified restrictions under the assumption of sequential ignorability when data shows monotone missing patterns and showed that only one type of restrictions allows the examination of the sensitivity.

It is worthwhile to integrate both missing genotypic values situation and missing phenotypic values situation into our nonparametric function mapping framework. Naturally the methods which may be helpful to make inference include EM algorithm, Gibbs sampler, multiple imputation or a combination of these methods.

The pattern mixture model is more attractive for us because those missing patterns may be highly correlated to, if not directly caused by, underlying QTL especially in monotone drop-out patterns and pattern mixture model allows us model the conditional distribution of missing mechanisms directly. There may exist a link between this pattern mixture model and the selection model with some prior distribution. To identify pattern mixture model, unverified identification restrictions are assumed so sensitivity studies will be conducted. Inference about estimates of parameters will be studied too.
REFERENCES


Davies, R.B. (1977). Hypothesis test when a nuisance parameter is present only under the alternative. *Biometrika* **64**: 247-254.

Davies, R.B. (1987). Hypothesis test when a nuisance parameter is present only under the alternative. *Biometrika* **74**: 33-43.


BIOGRAPHICAL SKETCH

Jie Yang was born on June 5, 1981, in Lianyungang, Jiangsu Province, P.R.China, as the eldest daughter of Mr. Mingfa Yang and Mrs. Qiyun Li. When she was 15 years old, she entered the Special Class of Gifted Young, University of Science and Technology of China. From there she gained a lot of interest in statistics field even though she graduated with a Bachelor of Engineering in computer science.

Eager to further intensify her knowledge in statistics, Jie decided to pursue a Ph.D. degree in statistics at the University of Florida. She joined as a research assistant in the Maternal Child Health and Education Research Data Center at University of Florida. In the summer of 2003, she was awarded the Merck BARDS Graduate Fellowship in Biostatistics and started her dissertation work with Dr. George Casella in nonparametric functional mapping of quantitative trait loci. Aside from completing the standard curriculum, she worked as an summer intern in Summer 2004 at Merck Research Laboratories, Rahway, NJ, and the paper from the research she did at Merck won the 2005 ENAR student award for paper competition. In Spring 2005 she volunteered to be a lab instructor for STA2023, An Introduction to Statistics. From Jan, 2004, till May, 2005, she organized the weekly student seminar in the Department of Statistics, University of Florida.