STATISTICAL METHODS IN LINKAGE ANALYSIS

by

Jun Teng

TECHNICAL REPORT NO. 7
August 1996

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Statistical Methods in Linkage Analysis

by

Jun Teng
Department of Statistics
Stanford University

TECHNICAL REPORT NO. 7
July 1996

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Abstract

This dissertation discusses several statistical problems associated with locating trait loci using maps of markers spanning the whole genome. Such maps are becoming readily available and can be especially useful in mapping traits that are non Mendelian.

Affected relative pairs provide an important tool for genetic analysis of complex traits. This method is based on the notion that the trait gene, if present, will likely be found in the region of the genome that is common to most pairs of affected relatives. Tests based on a continuous specification of identity by descent (IBD) between pairs of affected relatives or based on a dense set of completely polymorphic markers has been proposed (cf. Feingold, Brown and Siegmund, 1993). In practice, the situation is more complicated because one often can not determine the IBD state unambiguously due to incomplete polymorphism of the markers. In such situations, we define tests based on multipoint analysis. We study the p-value for such tests and provide a conservative threshold. The effectiveness of such tests is also discussed.

Allele-sharing methods can also be applied to quantitative traits. A popular approach is the regression analysis for sib pairs proposed by Haseman and Elston (1972). Since only differences in the phenotype between sibling pairs are used, their method fails to take advantage of the information of the trait value for both siblings. We study new approaches that make use of the information more efficiently.

In the last part of this dissertation, we study the problem of statistical significance level related to identifying quantitative trait loci in experimental organisms.
Acknowledgements

I am very grateful to my advisor, Professor David O. Siegmund for guiding my research during the past four years at Stanford. His warmth, patience, and encouragement have been indispensable in my completing this dissertation.

I wish to thank Professor Neil Risch for giving me advice on my research and job searching. I am especially grateful to him for never being selfish with his time. I would also like to thank Professor Tze Leung Lai for reading this dissertation.

I thank the entire statistics department for the supportive environment they have provided me. I am very lucky to have worked with and become acquainted with the many esteemed professors and the many brilliant and collaborative students in this department.

I would like to thank my husband and best friend, Xiaolei, for his unwavering love. I could not have finished without his generous support and encouragement.
Contents

Abstract iv

Acknowledgements v

1 Introduction 1
   1.1 Background in Genetics 2
   1.2 Linkage Analysis 3
   1.3 Genome Scan 4
      1.3.1 Gaussian Model 5
      1.3.2 Significance Level and Power 6
   1.4 Outline 7

2 Multipoint Analysis 10
   2.1 Introduction 10
   2.2 Tests and P-values 11
      2.2.1 Grandparent/Grandchild Pairs 12
      2.2.2 Half-Siblings 21
      2.2.3 Other Unilineal Relatives 22
      2.2.4 Sibling Pairs 27
   2.3 Efficiency 27

3 QTLs Based on Sib Pairs 36
   3.1 Introduction 36
   3.2 Haseman and Elston's Method 37
3.3 Likelihood Approach ........................................ 38
3.4 Generalized Regression ...................................... 42
3.5 Test Statistics .................................................. 46
3.6 Comparing with Haseman-Elston's Method ............... 49
3.7 Genome Wide Scan ............................................ 50

4 Quantitative Trait Loci (QTLs) .............................. 56
  4.1 Introduction .................................................. 56
  4.2 Backcross ..................................................... 56
  4.3 Testing for the genetic effect .............................. 57
    4.3.1 Test statistic ........................................... 57
    4.3.2 P-value .................................................. 58
    4.3.3 Numerical examples .................................. 67

Bibliography ......................................................... 69
List of Tables

2.1 Threshold for type-1 error of 0.05 for Grandparent/Grandchild . . . . . 22
2.2 Threshold for type-1 error of 0.05 for Half-siblings . . . . . . . . . . 22
4.1 Approximation for n=100 and b=4 . . . . . . . . . . . . . . . . . . . . . . 68
4.2 Approximation for n=25 and b=4 . . . . . . . . . . . . . . . . . . . . . . 68
List of Figures

2.1 Relative efficiency for trait locus exact at marker locus I ................. 33
2.2 Relative efficiency for trait locus midway between markers I ............ 33
2.3 Relative efficiency for trait locus exact at marker locus II ............... 34
2.4 Relative efficiency for trait locus midway between markers II ........... 34
2.5 Sample size for trait locus exact at marker locus .......................... 35
2.6 Sample size for trait locus midway between markers ....................... 35

3.1 Comparison of different linear combinations ................................. 54
3.2 Comparison with Haseman and Elston's statistic ............................. 55
Chapter 1

Introduction

One of the goals of genetic mapping is to discover the location of the genes causing inherited diseases, for example, Cystic fibrosis and Huntington's disease. The method of genetic mapping, by which one compares the inheritance pattern of a trait with the inheritance patterns of chromosomal regions, allows one to find where a disease predisposing gene is without any prior biological clues as to how it functions. The inheritance pattern in human pedigrees is traced by a set of genetic markers on the genome, and linkage analysis is used to investigate the relationships between disease genes and genetic markers. Thus a great deal of research effort is focused on expanding the set of available markers.

The development of numerous DNA polymorphic markers over the last decade has led to rapid progress in localizing genes associated with simple Mendelian diseases. More than 400 such diseases have been genetically mapped in this manner, and nearly 40 have been positionally cloned. Now human geneticists are exploring traits, which do not follow simple Mendelian monogenic inheritance, such as diabetes, hypertension, etc.

A brief introduction providing some genetic background follows. It will be accompanied by a description of recent work in the area of genetic linkage analysis and an outline of the contents of this thesis.
CHAPTER 1. INTRODUCTION

1.1 Background in Genetics

This section gives a brief introduction to a very much simplified genetics background. Good references for more detailed summary include Cavalli-Sforza and Bodmer (1986) and Crow (1989).

Heritable characteristics are determined by genes, and different genes are responsible for the expression of different characteristics. A gene may occur in different forms or states called alleles, each potentially having a different physical expression. A chromosome is a single thread-like molecule of DNA, along which genes are arranged linearly. The number of different chromosomes is characteristic of each species. Human beings have twenty-three pairs of chromosomes, and each gene has two copies, not necessarily the same alleles. The pair of alleles in an individual constitutes that individual’s genotype, and the expression of a particular genotype is called a phenotype.

During reproduction, the chromosome pairs are split in half during a biological process called meiosis, which results in the production of gametes (sperms or eggs) that carry a single set of chromosomes. The sets from a sperm and egg unite at fertilization to form a zygote, from which the individual develops. Therefore each individual receives one copy of its chromosomes from its mother and one from its father. In meiosis, homologous chromosomes pair up. They may exchange genetic material between them during a process called crossing over and produce gametes that contain mixture of the homologous chromosomes in each pair.

A chromosome in a gamete, which is a mixture of the two homologous chromosomes in the parent, can be modeled in the following way. It starts with either homologous chromosome randomly, moves a random distance along this chromosome, and then switches to the other chromosome. It moves another random distance, and switches again. This process continues until the end of the chromosome is reached. A standard model for the random distance between the crossovers is the exponential distribution (called the Haldane model), which we will assume throughout the dissertation. This implies the number of crossovers on each chromosome has a Poisson distribution. The genetic distance between two genes is defined to be the expected
number of crossovers between them. The unit for genetic distance is Morgan or centimorgan (cM). If two alleles on the same parental chromosome are passed to the offspring together, one says that there is no recombination between them, otherwise one says that there is recombination. Under the assumption of the Haldane model, the probability, $\theta$, of a recombination between two loci with a genetic distance $x$ Morgan is

$$\theta = (1/2)[1 - \exp(-2x)].$$

This function relating the genetic distance $x$ to the recombination probability $\theta$ is called the Haldane mapping function.

1.2 Linkage Analysis

The classical statistical method used in linkage analysis is the Lod score method which involves proposing a model for the inheritance pattern and computing the likelihood ratio as a measure of evidence for linkage (Ott 1991). It is conducted by first finding families in which multiple members are affected. Then a discrete set of markers is typed for each individual in the families. The joint likelihood of the trait status and genotype information at a selected marker for members of each family are calculated under the proposed model $M_1$ and a null hypothesis $M_0$. The model $M_1$ typically involves the recombination fraction between the marker and the gene causing the trait as an unknown parameter, and the maximum likelihood estimate of this parameter can be found. Then the Lod score defined as the log of the ratio of the maximum likelihood under $M_1$ and likelihood under $M_0$ is used to test $M_1$ against $M_0$.

The Lod score method has been used successfully in mapping simple Mendelian traits because the allowable models are few and number of parameters involved is small. This method is very powerful if the model is correct, but it's highly sensitive to misspecification of the linkage model (Clerget-Darpoux et al., 1986). It is much more difficult to apply for mapping complex traits since it's hard to model the inheritance pattern precisely.

An alternative approach is the allele-sharing method, usually based on pairs of
CHAPTER 1. INTRODUCTION

affected relatives (e.g. Risch, 1990). The intuitive idea of this method is that two affected relatives would be expected to share alleles from a common ancestor (identical by descent) at the disease locus more often than would be expected under random segregation. We can use the increase in identity by descent at markers located close to the disease locus to test for linkage. The allele-sharing method is nonparametric in the sense that it assumes no model for the inheritance of the trait, and therefore it tends to be more robust than the Lod score method. The allele-sharing methods can also be applied to quantitative traits based on the notion that the phenotypic similarity between two relatives should be correlated with the number of alleles shared identity by descent at a locus affecting the trait (Haseman and Elston, 1989).

Methods for typing a dense set of polymorphic markers spanning the whole genome are now approaching feasibility (Botstein et al., 1980 and Nelson et al., 1993). This has raised some new statistical questions and has motivated the research by Feingold, Brown and Siegmund (1993), Dupuis (1994). In the next section, we will briefly review some of the work by these authors.

1.3 Genome Scan

Motivated by the development of Genomic Mismatch Scanning (GMS, Nelson et al., 1993) Feingold (1993) and Feingold, Brown and Siegmund (1993) developed statistical methods to study the data that can provide an essentially continuous specification for regions of identity by descent. Feingold (1993) has developed a Markov-chain model to obtain the p-value and power for testing genetic linkage based on GMS data. A Gaussian approximation to the Markov-chain model is discussed by Feingold, Brown and Siegmund (1993). The Gaussian approximation requires a large sample size of affected relatives to be valid but is substantially simpler to analyze than the Markov-chain models. We will review some of their results based on Gaussian models. We review the simplest possible context, where the data are derived entirely from a number of independent grandparent/grandchild pairs.
1.3.1 Gaussian Model

Let $N$ be the total number of affected grandparent-grandchild pairs. It's convenient to let $p$ denote the probability of identity by descent at an arbitrary locus. Note $p = 1/2$ for grandparent-grandchild pairs. Let $X_t$ be the total number of pairs identical by descent at locus $t$ and let

$$Z_t = (X_t - Np)/N^{1/2}.$$ 

Then, on an unlinked chromosome, it follows from the central limit theorem and the Haldane model that, for large values of $N$, $Z_t$ is approximately a stationary Ornstein-Uhlenbeck process with mean zero and covariance function $\text{cov}(Z_{s+t}, Z_s) = \sigma^2 \exp(-\beta|t|)$, where $\sigma^2 = p(1 - p) = 1/4$ and $\beta = 2\lambda = 0.02$.

On the linked chromosome, we assume there is a single locus $r$, at which one or more alleles confers susceptibility to the trait. Each affected pair has an increased probability $(1 + \alpha)/2$ of identity by descent at locus $r$. The parameter $\alpha$ can be interpreted in terms of the increased risk of the trait appearing in a relative of a person who has the trait of interest (Risch, 1990). For a monogenic trait, $\alpha = (\lambda_O - 1)/(\lambda_O + 1)$, where $\lambda_O$ is the relative risk of the offspring of an affected individual to be affected. Therefore on the linked chromosome the expected value of $Z_t$ is

$$E(Z_t) = \xi \exp(-\beta|t - r|),$$

where $\xi = N^{1/2}\alpha p$.

The log-likelihood function of the observed process $\{Z_t, 0 \leq t \leq l\}$ as a function of the unknown parameters $r, \xi$ equals

$$\sigma^{-2}[\xi Z_r - \xi^2/2].$$

And the log-likelihood ratio test for testing the null hypothesis $H_0 : \xi = 0$, against the one-sided alternative, $H_1 : \xi > 0$ is

$$\max_{1 \leq c \leq C} \max_{0 \leq t \leq l} Z_t/\sigma.$$ 

The first maximum is over all chromosomes and the second maximum is over all locations on each chromosome. We suppress the chromosome index $c$ in our description of the statistic $Z_t$ for convenience.
1.3.2 Significance Level and Power

In the statistical literature there are several simple approximations for the significance level, \( P_0\{\max_{0 \leq t \leq l} Z_t/\sigma > b\} \), where the subscript 0 denotes that the probability is evaluated under the null hypothesis. The approximations suggested by Feingold, Brown and Siegmund (1993) is

\[
P_0\{\max_{0 \leq t \leq l} Z_t/\sigma > b\} \approx 1 - \Phi(b) + \beta b \phi(b),
\]

(1.1)

where \( l \) is the length of the chromosome in centimorgans and \( \Phi \) and \( \phi \) are the standard normal cumulative distribution and density functions, respectively. For an overall significance level when the test is applied to each chromosome, one simply sums approximation (1.1) over all chromosomes.

To approximate the power for the test, Feingold, Brown and Siegmund (1993) give the following approximation by arguing along the lines of James et al. (1987):

\[
P_\xi\{\max_{0 \leq t \leq l} Z_t/\sigma > b\} \approx 1 - \Phi(b - \xi/\sigma) + \phi(b - \xi/\sigma)[2(\xi/\sigma)^{-1} - (\xi/\sigma + b)^{-1}].
\]

(1.2)

The above results were extended to include the case that the IBD information is not continuous but rather from a set of equispaced genetic markers. The significance level becomes

\[
P_0\{\max_{0 \leq t \leq l} Z_t/\sigma > b\} \approx 1 - \Phi(b) + \beta b \phi(b) \nu(b(2\Delta)^{1/2}),
\]

where \( \Delta \) is the distance between adjacent markers and \( \nu(x) \) is a special function which can be evaluated numerically and is reasonably well approximated by \( \exp(-0.583x) \) (Siegmund 1985, chap. 4).

The power in the discrete case is a little more complicated since it depends on the distance between the trait locus \( r \) and its nearest markers. For the case where there is zero recombination between the trait locus and its closest marker, the power becomes

\[
P_\xi\{\max_{0 \leq t \leq l} Z_t/\sigma > b\} \approx 1 - \Phi(b - \xi/\sigma) + \phi(b - \xi/\sigma) \\
\times[2(\xi/\sigma)^{-1} \nu(b(2\Delta)^{1/2}) - (\xi/\sigma + b)^{-1} \nu^2(b(2\Delta)^{1/2})].
\]
Dupuis (1994) gives an approximation for the case that the trait locus \( r \) is not exactly at a marker. Let \( t_1 \) and \( t_2 \) be the two flanking markers of \( r \), then

\[
P_{\xi}\left\{ \max_{0 \leq \Delta \leq \xi} Z_\Delta > b \right\} \approx 1 - \int_0^\infty \phi(h(x, t_1, r)) \Phi \left( \frac{g(x, t_2, t_1, r)}{d(\Delta)} \right) dx
+ \nu(b\sqrt{2\beta\Delta}) \int_0^\infty e^{-bx} \phi(h(x, t_2, r)) \Phi \left( \frac{g(x, t_1, t_2, r)}{d(\Delta)} \right) dx
+ \nu(b\sqrt{2\beta\Delta}) \int_0^\infty e^{-bx} \phi(h(x, t_1, r)) \Phi \left( \frac{g(x, t_2, t_1, r)}{d(\Delta)} \right) dx
- \nu^2(b\sqrt{2\beta\Delta}) \int_0^\infty e^{-bx} \phi(h(x, t_1, r)) G(x, t_1, t_2) dx
\]

where

\[
d(\Delta) = (1 - \exp(-2\beta\Delta))^{1/2}
\]

\[
h(x, t, r) = b - x - \left( \xi / \sigma \right) \exp(-\beta|t - r|)
\]

\[
g(x, s, t, r) = b - \left( \xi / \sigma \right) \exp(-\beta|t - r|) - \exp(-\beta\Delta)[b - x - \left( \xi / \sigma \right) \exp(-\beta|t - r|)]
\]

\[
G(x, s, t, r) = \exp[-bx - bg(x, t_2, t_1, r) + b^2d^2(\Delta)/2] \Phi \left( \frac{g(x, t_2, t_1, r)}{d(\Delta)} - bd(\Delta) \right)
\]

The above approximations can be applied to other types of unilineal relatives and with slight modification can be applied to sibling pairs also. Dupuis (1994) generalized the results to multigenic diseases, i.e. diseases for which susceptibility is increased by mutant alleles acting separately or in conjunction at more than one locus. Dupuis, Brown and Siegmund (1995) compared three statistical methods for detecting linkage of such diseases.

### 1.4 Outline

Chapter 2 contains extension of the previous authors' work to the more complicated situation that the IBD state cannot be unambiguously determined due to incomplete polymorphism of the markers. In such situations, the exact IBD has to be estimated from the marker information. We define tests based on multipoint analysis using the information from all genetic markers. We study the p-value for such tests and provide a conservative threshold. The effectiveness of multipoint analysis is evaluated
CHAPTER 1. INTRODUCTION

cmpared to the ideal case, i.e. the exact IBD is known at markers. The noncentrality parameter of the score statistic is calculated as a function of the polymorphism and density of the markers. Simple simulation studies of the power are also conducted. For dense markers, tests based on multipoint analysis can be almost as powerful as the ideal case. The tests based on multipoint analysis are consistently better than the tests based on single point analysis, and for most cases, the differences are substantial.

In chapter 3, we present new tests for mapping quantitative trait loci (QTLs) in humans. Haseman and Elston (1972) introduced a method based on sibling pairs to test for genetic linkage to a quantitative trait. Their test is based on regression analysis of the square of sib pair phenotypic difference on the number of alleles shared IBD for the sib pair. Since only differences of sibling pair phenotypes are used, this approach fails to take advantage of the information of the trait value for both siblings. This led us to extend the method proposed by Haseman and Elston in the following way. We first assume that the trait values of sib pairs have a bivariate normal distribution. The log likelihood is computed based on this assumption. A test based on the score statistic is derived. Since this method depends on the rather restrictive assumption of normality, a more general approach is also proposed. Two regression equations are constructed. A test based on multivariate regression analysis is proposed, and is shown to be asymptotically as efficient as the test based on the assumption of bivariate normality when that assumption is true.

In chapter 4, we study the problem of statistical significance level related to identifying a QTL in experimental organisms. Lander and Botstein (1989) developed a statistical test for the backcross design. They noticed that as the sample size increases, the statistic tends in distribution to the maximum of an Ornstein-Uhlenbeck process, and they calculated the false positive rate for the test with continuous markers. For discrete maps with markers equispaced at every Δ centimorgens, Dupuis gives a similar approximation with a discreteness correction factor. We derive a more accurate approximation to the false positive rate. Our approximation follows the method proposed by Woodroofe (1976). The problem of evaluating the tail probability is then reduced to calculating two probabilities, a marginal probability and a more complicated conditional probability. We improve the approximation by computing both
probabilities more accurately. The marginal probability is rather straightforward to calculate. The difficult part is to calculate the conditional probability, which can be formulated as a boundary crossing problem for a sum of compound Poisson variables, which can be approximated using techniques in sequential analysis (cf. Siegmund 1985).
Chapter 2

Multipoint Analysis

2.1 Introduction

The previous chapter describes the allele-sharing method for locating a disease susceptibility gene. The results are derived assuming the availability of a continuous map of identity-by-descent (IBD) between the relatives through the use of Genomic Mismatch Scanning (GMS, Nelson et al. (1993)) and can be extended to the situation that the IBD information is available at a set of equally spaced markers. Conceptually, this method is rather straightforward – one simply looks for regions that are excessively shared IBD among affected relatives.

In practice, the situation is more complicated because one cannot always unambiguously determine whether two relatives have identity by descent or not at every position along the genome. For example, if two relatives have the same alleles at a genetic marker, we cannot tell whether these two alleles are actually inherited from a common ancestor (IBD) or not, especially when this is a common allele. Various solutions have been proposed to cope with this important practical difficulty. They basically fall into two categories. One approach is to focus solely on whether affected relatives happen to show the same alleles at a locus (i.e., identity-by-state or IBS). The extent of IBS sharing among all pairs of affected members of the pedigree is compared to Mendelian expectation (Weeks and Lange 1988 and 1992). Another approach amounts to inferring IBD sharing on the basis of the marker data, which
usually involves the use of genotype information for additional members of the pedigree to resolve whether or not shared alleles are actually identical by descent (Risch 1990c). However, most studies have focused on studying individual markers one-at-a-time and fail to take advantage of information available from multiple markers.

To fully exploit the power of a map, a complete multipoint analysis using the information from all genetic markers to infer the probability distribution of the IBD status at each point along the genome is proposed by Karuglyak and Lander (1995) to analyze sib-pair data. And a computer program GENEHUNTER is implemented for such analysis.

The next section describes a mathematical framework for multipoint analysis for affected relative pairs. It involves calculating the joint likelihood of the observed genotype data at all markers. The score statistic is used for testing the null hypothesis of no trait locus on the chromosome. It is asymptotically equivalent to the maximum likelihood ratio test, but much easier to compute. We propose a threshold that will control the total false positive rate by taking into account the multiple tests involved. Finally, through a combination of theoretical analysis and simulation, we try to estimate the efficiency of the multipoint analysis.

2.2 Tests and P-values

First we introduce some notation. Consider a chromosome of length \( l \) cM's containing at most one trait susceptibility gene at locus \( \tau \). Assume there are \( M \) markers at loci \( t_1, t_2, \ldots, t_M \) on the chromosome. For each pedigree \( j \), let \( X^j(t) \) be the number of alleles shared IBD by the affected relative pairs at locus \( t \), let \( G^j(t_i) \) denote the genotype information for all the members at marker \( t_i \). Members in each pedigree could be only the affected relative pairs or could include other additional relatives. Let \( \mathbf{G}^j = (G^j(t_1), G^j(t_2), \ldots, G^j(t_M)) \) be the total genotype information for the \( j^{th} \) pedigree.
2.2.1 Grandparent/Grandchild Pairs

We adopt the statistical model described by Feingold (1993) for the IBD process $X^j(t)$. At the trait locus $\tau$, the pairs have an excess probability of sharing identity by descent, say $(1 + \alpha)/2$. Risch (1990a,b) presents a model that allows us to relate $\alpha$ to epidemiological parameters. He defines $\lambda_0$ as the relative risk of the offspring of an affected individual to be affected. Then under certain assumptions, $\alpha = (\lambda_0 - 1)/(\lambda_0 + 1)$. So at the trait locus $\tau$, $X^j(\tau)$ has the Bernoulli$(1 + \alpha)/2$ distribution. For $t$ moving in both directions from $\tau$, $X^j(t)$ is Markov chain on the states 0 and 1. The transition rates can be described by the $Q$-matrix

$$
\begin{pmatrix}
-\lambda & \lambda \\
\lambda & -\lambda
\end{pmatrix},
$$

where $\lambda$ is the crossover rate per unit length, .01/centimorgan(cM).

The likelihood function for a family is the probability of its observed marker data. So for $j^{th}$ family, the likelihood can be written as,

$$
\Pr(G^j) = \Pr(G^j|X^j(\tau) = 0) \Pr(X^j(\tau) = 0) + \Pr(G^j|X^j(\tau) = 1) \Pr(X^j(\tau) = 1)
= \Pr(G^j|X^j(\tau) = 0)(1 - \alpha)/2 + \Pr(G^j|X^j(\tau) = 1)(1 + \alpha)/2.
$$

The calculation of $\Pr(G^j|X^j(\tau))$ can be separated into two factors. The first factor is the probability of the observed genotype data given the unobserved IBD process, i.e.

$$
\Pr(G^j|X^j(t_1), X^j(t_2), \cdots, X^j(t_M)).
$$

If we assume the genotype information consists of only grandparents, parents and the child (missing values are allowed), this value, under linkage equilibrium, equals

$$
\Pr(G^j(t_1)|X^j(t_1)) \Pr(G^j(t_2)|X^j(t_2)) \cdots \Pr(G^j(t_M)|X^j(t_M)),
$$

and does not depend on the parameter $\alpha$. The second factor is the probability for the IBD process given the IBD information at the trait locus, i.e.

$$
\Pr(X^j(t_1), X^j(t_2), \cdots, X^j(t_M)|X^j(\tau)).
$$
CHAPTER 2. MULTIPOINT ANALYSIS

This factor, under our model just described, does not depend on $\alpha$ either. Therefore the conditional probability $\Pr(G^j|X^j(\tau))$ does not depend on $\alpha$. So the score statistic is

\[
\frac{\partial}{\partial \alpha} \sum_j \log[\Pr(G^j)]|_{\alpha=0}
= \sum_j \frac{\Pr(G^j|X^j(\tau) = 1)/2 - \Pr(G^j|X^j(\tau) = 0)/2}{\Pr(G^j|X^j(\tau) = 1)/2 + \Pr(G^j|X^j(\tau) = 0)/2}
= \sum_j \frac{\Pr_0(G^j, X^j(\tau) = 1) - \Pr_0(G^j, X^j(\tau) = 0)}{\Pr_0(G^j)}
= \sum_j [\Pr_0(X^j(\tau) = 1|G^j) - \Pr_0(X^j(\tau) = 0|G^j)]
= \sum_j [2\Pr_0(X^j(\tau) = 1|G^j) - 1].
\]

For simplicity, we introduce the notation

\[Y^j(t) = \Pr_0(X^j(t) = 1|G^j).\]

If the trait location $\tau$ is known, the score test for the null hypothesis $\alpha = 0$ is

\[Z_2(t) = \sum_j (Y^j(t) - 1/2)/(\sqrt{n}\sigma_Y(t))\]

where $n$ is total number of pedigrees and $\sigma_Y(t)$ is the variance of $Y^j(t)$ under $H_0$. When the trait locus is unknown, the test statistic for the global search is

\[\max_{1 \leq i \leq M} Z_2(t_i).\]

Here we only test for linkage at the markers, because we think testing for linkage between markers (interval mapping) will not improve the power by a lot. (For a detailed discussion, see Dupuis (1994).) But our analysis can be easily extended to the interval mapping situation.

Compared to the statistic used by Feingold (1993), we have replaced the exact number of IBD $X^j(t_i)$ at each marker locus by its estimator $Y^j(t_i)$. Since we have changed the test statistic, we have to re-examine the proper threshold $b$ for declaring linkage.
CHAPTER 2. MULTIPOINT ANALYSIS

We use the approximate Gaussian framework, i.e., assuming $Z_2(t)$ is Gaussian process. The Gaussian models have the limitation of needing large numbers of affected relative pairs to be valid, but they allow the rich literature of normal theory to be applied and give useful insight into the problems at hand.

Our statistic $Z_2(t)$ was standardized so marginally it has approximately a standard normal distribution under the null hypothesis. So to find the threshold defining the critical region for our new test, one needs to find the correlation function for $Z(t)$. For the ideal case that the exact IBD is known, the calculation is rather straightforward and the covariance is of the form $R(t) = e^{-2|t|}$. However, when the exact information is not available and the IBD status has to be estimated from the genotype data, the calculation for the covariance function becomes much more difficult if not impossible. So instead of calculating the exact covariance function for our new statistic, we establish an inequality between the covariance functions of our new statistic and the one based on exact IBD. Based on this inequality, we can show that the thresholds for the process using estimated IBD will be smaller than those based on the exact IBD. This can be formally stated by the following theorem.

**Theorem**

Let

$$Z_1(t_i) = \sum_j (2X^j(t_i) - 1)/n^{1/2}$$

$$Z_2(t_i) = \sum_j (Y^j(t_i) - 1/2)/(\sqrt{n}\sigma_Y(t_i)).$$

Under the null hypothesis of $\alpha = 0$,

$$\text{Pr}\{\max_{1 \leq i \leq M} Z_1(t_i) \geq b\} \geq \text{Pr}\{\max_{1 \leq i \leq M} Z_2(t_i) \geq b\}.$$

The following three lemmas will be used in the proof of the theorem.

**Lemma 1** ("Slepian's Lemma")

Let $\{\zeta_1(t)\}$ and $\{\zeta_2(t)\}$ be normal processes (possessing continuous sample functions but not necessarily stationary). Suppose that they are standardized so that $E(\zeta_1(t)) = E(\zeta_2(t)) = 0$, and $E(\zeta_1(t)^2) = E(\zeta_2(t)^2) = 1$, and write $\rho_1(t,s)$ and $\rho_2(t,s)$ for their
covariance functions. Suppose that for some \( c > 0 \) we have \( \rho_1(t, s) \geq \rho_2(t, s) \) when \( 0 \leq t, s \leq c \). Then the respective maxima \( M_1(t) \) and \( M_2(t) \) satisfy
\[
P\{M_1(T) \leq u\} \geq P\{M_2(T) \leq u\}
\]
when \( 0 \leq T \leq c \).

This is the well known Slepian's Inequality and its proof can be found in Leadbetter, Lindgren and Rootzen (1983) p.156.

Lemma 2

\( Z_1(t_i) \) and \( Z_2(t_i) \) are defined as before, then under the null hypothesis,
\[
\text{cov}(Z_1(t_i), Z_1(t_j)) \leq \text{cov}(Z_2(t_i), Z_2(t_j))
\]

Because the distribution of \( \{X^i(t_i), Y^j(t_i), G^j(t_i), 1 \leq i \leq M\} \) does not depend on the particular family \( j \) chosen, it is convenient to drop the subscript \( j \) in the following arguments.

Lemma 3

\[
E_0[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \geq E_0[Y(t_i)|X(t_i) = \delta]
\]

(2.1)

where \( \delta = 0 \) or \( 1 \).

Proof of Lemma 3.

Without loss of generality, we may assume \( j > i \). To prove (2.1), I'm going to show the following inequality first.

\[
E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1] \geq E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta']
\]

(2.2)

where \( \delta' = 0 \) or \( 1 \).

Proof of (2.2)

(2.2) is equivalent to
\[
E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1] \\
\geq E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 0].
\]

(2.3)
And
\[
E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1] \\
- E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 0] \\
= \sum_G P(X(t_i) = 1|G)[P(G|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1) \\
- P(G|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 0)] \\
= \sum_G P(X(t_i) = 1|G)P(G(t_1), \cdots, G(t_{j-1})|X(t_i) = \delta, X(t_{j-1}) = \delta') \\
\left[P(G(t_j), \cdots, G(t_M)|X(t_j) = 1) - P(G(t_j), \cdots, G(t_M)|X(t_j) = 0)\right] \\
= \sum_{G(t_1), \cdots, G(t_{j-1})} P(X(t_i) = 1|G)[P(G(t_j), \cdots, G(t_M)|X(t_j) = 1) \\
- P(G(t_j), \cdots, G(t_M)|X(t_j) = 0)]
\]

So it's enough to show, for any fixed \(G(t_1), \cdots, G(t_{j-1}),\)
\[
\sum_{G(t_j), \cdots, G(t_M)} P(X(t_i) = 1|G)[P(G(t_j), \cdots, G(t_M)|X(t_j) = 1) \\
- P(G(t_j), \cdots, G(t_M)|X(t_j) = 0)] \geq 0 \quad (2.4)
\]

To prove (2.4), I will use induction which consists of the following two parts (a) and (b).

Part (a)
\[
\sum_{G(t_j), \cdots, G(t_M)} P(X(t_j) = 1|G)[P(G(t_j), \cdots, G(t_M)|X(t_j) = 1) \\
- P(G(t_j), \cdots, G(t_M)|X(t_j) = 0)] \geq 0
\]

Proof of (a)
\[
P(X(t_j) = 1|G) - P(X(t_j) = 1|G(t_1), \cdots, G(t_{j-1})) \\
= \frac{P(X(t_j) = 1, G) - P(X(t_j) = 1, G(t_1), \cdots, G(t_{j-1}))}{P(G)} \\
= \frac{P(X(t_j) = 1, G(t_1), \cdots, G(t_{j-1}))P(G(t_j), \cdots, G(t_M)|X(t_j) = 1)}{P(G)}
\]
\[ P(X(t_j) = 1, G(t_1), \ldots, G(t_{j-1})) \\
= \frac{P(X(t_j) = 1, G(t_1), \ldots, G(t_{j-1}))}{P(G(t_1), \ldots, G(t_{j-1}))} \\
= \frac{P(G(t_1), \ldots, G(t_{j-1}))P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) - P(G)}{P(G)P(G(t_1), \ldots, G(t_{j-1}))} \\
= \frac{P(G(t_1), \ldots, G(t_{j-1}))}{P(G)P(G(t_1), \ldots, G(t_{j-1}))} \\
\times [P(G(t_1), \ldots, G(t_{j-1}))P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) \cdot (1)P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) \\
- P(G(t_1), \ldots, G(t_{j-1}), X(t_j) = 1)P(G(t_j), \ldots, G(t_M)|X(t_j) = 0) \\
- P(G(t_1), \ldots, G(t_{j-1}), X(t_j) = 0)P(G(t_j), \ldots, G(t_M)|X(t_j) = 0)] \\
= \frac{P(X(t_j) = 1, G(t_1), \ldots, G(t_{j-1}))P(X(t_j) = 0, G(t_1), \ldots, G(t_{j-1}))}{P(G)P(G(t_1), \ldots, G(t_{j-1}))} \\
\times [P(G(t_1), \ldots, G(t_M)|X(t_{j-1}) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 0)] \\
\geq 0. \\
\]

Therefore

\[
\sum_{G(t_1), \ldots, G(t_M)} (P(X(t_j) = 1|G) - P(X(t_j) = 1|G(t_1), \ldots, G(t_{j-1}))) \\
\quad \quad \quad [P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 0)] \\
= \sum_{G(t_1), \ldots, G(t_M)} \frac{P(X(t_j) = 1, G(t_1), \ldots, G(t_{j-1}))P(X(t_j) = 0, G(t_1), \ldots, G(t_{j-1}))}{P(G)P(G(t_1), \ldots, G(t_{j-1}))} \\
\quad \quad \quad [P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 0)]^2 \\
\geq 0. \\
\]

But

\[
\sum_{G(t_1), \ldots, G(t_M)} P(X(t_j) = 1|G(t_1), \ldots, G(t_{j-1})) \\
\quad \quad \quad [P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_{j-1}) = 0)] \\
= 0. \\
\]

So (a) is true.

Part (b)

Assume

\[
\sum_{G(t_1), \ldots, G(t_M)} P(X(t_i) = 1|G)[P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) \\
- P(G(t_j), \ldots, G(t_M)|X(t_j) = 0)] \geq 0, \\
(2.5)
\]
for some $i \leq j$. Then

$$
\sum_{G(t_j), \ldots, G(t_M)} P(X(t_{i-1}) = 1|G) [P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_j) = 0)] \geq 0,
$$

(2.6)

**Proof of (b)**

$$
\sum_{G(t_j), \ldots, G(t_M)} P(X(t_{i-1}) = 1|G) [P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_j) = 0)]
$$

$$
= \sum_{G(t_j), \ldots, G(t_M)} P(X(t_{i-1}) = 1|G(t_1), \ldots, G(t_{i-1}))
$$

$$
\sum_{\varepsilon=0}^{1} \left[ \frac{P(X(t_i) = \varepsilon|G) P(X(t_i) = \varepsilon|X(t_{i-1}) = 1) + P(X(t_i) = 0|G) P(X(t_i) = 0|X(t_{i-1}) = 1)}{P(X(t_i) = 1|G(t_1), \ldots, G(t_{i-1}))} \right]
$$

$$
[P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_j) = 0)]
$$

$$
= P(X(t_{i-1}) = 1|G(t_1), \ldots, G(t_{i-1}))
$$

$$
\sum_{G(t_j), \ldots, G(t_M)} \left[ \frac{P(X(t_i) = 1|X(t_{i-1}) = 1) - P(X(t_i) = 0|X(t_{i-1}) = 1)}{P(X(t_i) = 1|G(t_1), \ldots, G(t_{i-1}))} \right]
$$

$$
[P(X(t_i) = 1|G)|P(G(t_j), \ldots, G(t_M)|X(t_j) = 1) - P(G(t_j), \ldots, G(t_M)|X(t_j) = 0)]
$$

$$
\geq 0.
$$

We have showed that the inequality (2.4) is true, and consequently inequalities (2.2) and (2.3) are true.

Now we are going to prove (2.1). Again, we use an induction argument.

**Part (a)**

$$
E[Y(t_i)|X(t_i) = \delta, X(t_{i+1}) = 1] \geq E[Y(t_i)|X(t_i) = \delta]
$$
CHAPTER 2. MULTIPOINT ANALYSIS

Proof of (a)

This is a direct consequence of inequality (2.2).

Part (b)

Assume

\[ E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1] \geq E[Y(t_i)|X(t_i) = \delta]. \]

Then

\[ E[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \geq E[Y(t_i)|X(t_i) = \delta]. \]

Proof of (b)

\[
E[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \\
= \sum_{\delta'} \sum_{t' = 0} E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1] \\
\times P(X(t_{j-1}) = \delta'|X(t_i) = \delta, X(t_j) = 1) \\
\geq \sum_{\delta'} \sum_{t' = 0} E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta'] P(X(t_{j-1}) = \delta'|X(t_i) = \delta, X(t_j) = 1) \\
= \sum_{\delta'} \sum_{t' = 0} E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta'] P(X(t_{j-1}) = \delta'|X(t_i) = \delta) \\
+ E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 0](P(X(t_{j-1}) = 0|x(t_i) = \delta, X(t_j) = 1) \\
- P(X(t_{j-1}) = 0|x(t_i) = \delta)) \\
+ E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1](P(X(t_{j-1}) = 1|x(t_i) = \delta, X(t_j) = 1) \\
- P(X(t_{j-1}) = 1|x(t_i) = \delta)) \\
= E[Y(t_i)|X(t_i) = \delta] \\
+ (E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1] - E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 0]) \\
\times (P(X(t_{j-1}) = 1|x(t_i) = \delta, X(t_j) = 1) - P(X(t_{j-1}) = 1|x(t_i) = \delta))
\]

Since

\[ (E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1] - E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 0]) \geq 0 \]

by induction assumption and

\[ (P(X(t_{j-1}) = 1|x(t_i) = \delta, X(t_j) = 1) - P(X(t_{j-1}) = 1|x(t_i) = \delta)) \geq 0 \]
from straightforward calculation,

\[ E[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \geq E[Y(t_i)|X(t_i) = \delta]. \]

We will use Lemma 3 to prove Lemma 2.

**Proof of Lemma 2.**

\[
\text{cov}(Y(t_i), Y(t_j)) \\
= (1/2)E[(Y(t_i) - 1/2)Y(t_j)] + (1/2)E[(Y(t_j) - 1/2)Y(t_i)] \\
= (1/2)E[(Y(t_i) - 1/2)E[X(t_j)|G]] + (1/2)E[(Y(t_j) - 1/2)E[X(t_i)|G]] \\
= (1/2)E[(Y(t_i) - 1/2)X(t_j)] + (1/2)E[(Y(t_j) - 1/2)X(t_i)]
\]

However,

\[
E[(Y(t_i) - 1/2)X(t_j)] \\
= \sum_{\delta=0}^{1} E[Y(t_i) - 1/2|X(t_i) = \delta, X(t_j) = 1]P(X(t_i) = \delta, X(t_j) = 1) \\
\geq \sum_{\delta=0}^{1} E[Y(t_i) - 1/2|X(t_i) = \delta]P(X(t_i) = \delta, X(t_j) = 1) \\
= \sum_{\delta=0}^{1} E[(Y(t_i) - 1/2)I_{X(t_i) = \delta}]P(X(t_j) = 1|X(t_i) = \delta) \\
= E[(Y(t_i) - 1/2)X(t_i)](P(X(t_j) = 1|X(t_i) = 1) - P(X(t_j) = 1|X(t_i) = 0)] \\
= E[(Y(t_i) - 1/2)Y(t_i)][(1 + \exp(-2\lambda|t_i - t_j|))/2 - (1 - \exp(-2\lambda|t_i - t_j|))] \\
= \sigma_Y^2(t_i) \exp(-2\lambda|t_i - t_j|).
\]

Similarly,

\[ E[(Y(t_j) - 1/2)X(t_i)] \geq \sigma_Y^2(t_j) \exp(-2\lambda|t_i - t_j|). \]

Therefore,

\[
\text{cov}(Y(t_i), Y(t_j)) \geq \exp(-2\lambda|t_i - t_j|)(\sigma_Y^2(t_i) + \sigma_Y^2(t_j))/2 \\
\geq \exp(-2\lambda|t_i - t_j|)\sigma_Y(t_i)\sigma_Y(t_j) \quad (2.7)
\]
CHAPTER 2. MULTIPOINT ANALYSIS

Since

\[ \text{cov}(Z_1(t_i), Z_1(t_j)) = n\text{cov}(X(t_i), X(t_j))/(1/4) = n \exp(-2\lambda|t_i - t_j|) \]

and

\[ \text{cov}(Z_2(t_i), Z_2(t_j)) = n\text{cov}(Y(t_i), Y(t_j))/(\sigma_Y(t_i)\sigma_Y(t_j)) \]

Lemma 2 follows from (2.7).

**Proof of Theorem.**

It's a consequence of Lemma 1 and 2.

Remark: The $\sigma_Y(t)$ is in general difficult to compute exactly, it can be replaced by

\[ \hat{\sigma}_Y(t) = \left[ \sum_{j=1}^n (Y^j(t) - p)^2/n \right]^{1/2} \]

in practice, where $p = E_0(Y^j(t)) = E_0(X^j(t))$ which equals 0.5 for grandparent/grandchild. The asymptotic properties of the process $Z(t)$ will not change.

This Theorem gives an upper bound of the total false positive rate for our multi-point analysis and consequently we can derive a conservative threshold for the test. This threshold can be checked against simulations.

In the simulation study, we assume each marker to be either fully informative or noninformative, which corresponds to case that other relevent relatives, i.e. spouse of the grandparent and the intervening parents, are typed (Risch 1990c). The simulation was done for the case that there are $n = 100$ pairs, markers are equally spaced with $\Delta$ centimorgan apart and each marker has 0.5 probability to be informative. The simulation was repeated 100,000 times and for each iteration the value of the test statistic was computed. The 95th quantile was then calculated and is presented in Table 2.1. In all the numerical calculation we assume human genome consist of 23 chromosomes of length 140 centimorgans each.

2.2.2 Half-Siblings

The stochastic process for half-siblings has exactly the same structure as the grandparent/grandchild case, but the underlying IBD process has transition rate $2\lambda$. The
CHAPTER 2. MULTIPOINT ANALYSIS

<table>
<thead>
<tr>
<th>Δ</th>
<th>simulation</th>
<th>nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 cM</td>
<td>3.886</td>
<td></td>
</tr>
<tr>
<td>1 cM</td>
<td>3.662</td>
<td>3.780</td>
</tr>
<tr>
<td>5 cM</td>
<td>3.520</td>
<td>3.643</td>
</tr>
<tr>
<td>10 cM</td>
<td>3.445</td>
<td>3.551</td>
</tr>
<tr>
<td>20 cM</td>
<td>3.332</td>
<td>3.435</td>
</tr>
</tbody>
</table>

Table 2.1: Threshold for type-1 error of 0.05 for Grandparent/Grandchild

rest of the development for the half-sibling case follows the grandparent/grandchild case exactly.

A simulation was done under the same conditions as grandparent/grandchild case and the result is given in Table 2.2.

<table>
<thead>
<tr>
<th>Δ</th>
<th>simulation</th>
<th>nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 cM</td>
<td>4.076</td>
<td></td>
</tr>
<tr>
<td>1 cM</td>
<td>3.799</td>
<td>3.913</td>
</tr>
<tr>
<td>5 cM</td>
<td>3.613</td>
<td>3.726</td>
</tr>
<tr>
<td>10 cM</td>
<td>3.537</td>
<td>3.604</td>
</tr>
<tr>
<td>20 cM</td>
<td>3.377</td>
<td>3.461</td>
</tr>
</tbody>
</table>

Table 2.2: Threshold for type-1 error of 0.05 for Half-siblings

2.2.3 Other Unilineal Relatives

For other unilineal relative pairs, for example, aunt/niece pairs, cousins, or second cousins, the IBD process $X(t)$ is no longer a Markov chain. However, we can still use the hidden Markov chain model. But the underlying chain is not the IBD process as in the grandparent/grandchild or half-sibling case. Here, we use the inheritance vector $\psi(t) = (p_1, m_1, p_2, m_2, \ldots, p_n, m_n)$ (Kruglyak et. al, 1996) as the underlying chain. The coordinates of $\psi(t)$ describe the outcome of the paternal and maternal meioses
giving rise to the \( n \) non-founders in the pedigree \( j \). Specifically, \( p_i = p_i^j(t) = 0 \) or 1 according to whether the grandpaternal or grandmaternal allele was transmitted in the paternal meiosis giving rise to the \( i \)-th non-founder; \( m_i = m_i^j(t) \) carries the same information for the corresponding maternal meiosis. So \( v^j(t) \) completely specifies the inheritance pattern at each point \( t \). Let \( v^j = (v^j(t_1), v^j(t_2), \ldots, v^j(t_m)) \). Now the likelihood of the observed genotypes for the \( j \)th pedigree can be written as

\[
Pr(G^j) = \sum_{v^j} Pr(G^j, v^j) \\
= \sum_{v^j} Pr(G^j | v^j) Pr(v^j | X^j(\tau) = 1) Pr(X^j(\tau) = 1) \\
+ \sum_{v^j} Pr(G^j | v^j) Pr(v^j | X^j(\tau) = 0) Pr(X^j(\tau) = 0).
\]

We consider the specific alternative hypothesis described by Feingold (1993), which assigns alternative probabilities to the states of the underlying chain. She sets \( Pr(X^j(\tau) = 1) \) equal to \( p_\alpha = p_0 + \alpha \sqrt{p_0(1 - p_0)} \), and for any vector \( i \) describing specific meiotic outcomes, she assigns probabilities \( Pr(v^j(\tau) = i) = a_i \) for all \( i \) such that \( v^j(\tau) = i \Rightarrow X^j(\tau) = 1 \) and \( Pr(v^j(\tau) = i) = b_i \) for \( i \) such that \( v^j(\tau) = i \Rightarrow X^j(\tau) = 0 \), where \( \sum a_i = p_\alpha \) and \( \sum b_i = 1 - p_\alpha \). She makes the further assumption that \( a_i = a \) for all \( i \) and \( b_i = b \) for all \( i \), and obtains the important result that the distribution of \( v^j \) given \( X^j(\tau) \) is independent of \( \alpha \). The likelihood becomes,

\[
Pr(G^j) = \sum_{v^j} Pr_0(G^j | v^j) Pr_0(v^j | X^j(\tau) = 1)p_\alpha \\
+ \sum_{v^j} Pr_0(G^j | v^j) Pr_0(v^j | X^j(\tau) = 0)(1 - p_\alpha) \\
= Pr_0(G^j | X^j(\tau) = 1)p_\alpha + Pr_0(G^j | X^j(\tau) = 0)(1 - p_\alpha).
\]

The subscript 0 indicates that the probability is taken under the null hypothesis of random segregation. Since the likelihood depends on \( \alpha \) only through \( p_\alpha \) and \( 1 - p_\alpha \), the calculation of the score statistic is quite straightforward.

\[
\frac{\partial}{\partial \alpha} \sum_j \log[Pr(G^j)]|_{\alpha=0} \\
= \sum_j \frac{Pr_0(G^j | X^j(\tau) = 1) - Pr_0(G^j | X^j(\tau) = 0)}{Pr_0(G^j)} \sqrt{p_0(1 - p_0)}
\]
CHAPTER 2. MULTIPoint ANALYSIS

\[ \sum_j \frac{\Pr_0(G^j, X^j(\tau) = 1)/p_0 - \Pr_0(G^j, X^j(\tau) = 0)/(1 - p_0)}{\Pr_0(G^j)} \sqrt{p_0(1 - p_0)} \]

\[ = \sum_j (Y^j(\tau)/p_0 - (1 - Y^j(\tau))/(1 - p_0)) \sqrt{p_0(1 - p_0)} \]

\[ = \sum_j (Y^j(\tau) - p_0)/\sqrt{p_0(1 - p_0)}. \]

If the trait location \( \tau \) is known, the score test for the null hypothesis \( \alpha = 0 \) is

\[ Z_2(\tau) = \sum_j (Y^j(\tau) - p_0)/\sqrt{n \sigma_Y(\tau)}. \]

and the test statistic for the global search when the trait locus is unknown is

\[ \max_{1 \leq i \leq M} Z_2(t_i). \]

Again we obtain a statistic closely represented the statistic used by Feingold (1993), but the exact number of IBD is replaced by its estimator. The calculation of this estimated IBD in principle can be solved by using a hidden Markov chain model (HMM) described by Lander and Green (1987). Kruglyak et al (1995) (1996) have devised substantial accelerations of the HMM.

To study the p-value for this new statistics, we once again use the normal approximation. We try to give an upper bound for the p-value by establishing an inequality between the correlation of our new statistic and the one based on the exact IBD process. In the following lemma, all the expectations are taken under the null hypothesis, we omit the subscript 0 for convenience.

**Lemma 4**

Let

\[ R(t) = \text{cov}(X(s), X(s + t)). \]

Then

\[ \text{cov}(Z_2(t_i), Z_2(t_j)) \geq R(t_j - t_i) - (1 - R(t_j - t_i))(c_{i,j} - 1), \]

where

\[ c_{i,j} = [\sigma^2_X(t_i) + \sigma^2_X(t_j)]/[\sigma^2_Y(t_i) + \sigma^2_Y(t_j)]. \]
Proof.

\[
\text{cov}(Y(t_i), Y(t_j)) = \frac{1}{2} \left[ \sigma_Y^2(t_i) + \sigma_Y^2(t_j) \right] - \frac{1}{2} E(Y(t_i) - Y(t_j))^2 \\
\geq \frac{1}{2} \left[ \sigma_Y^2(t_i) + \sigma_Y^2(t_j) \right] - \frac{1}{2} E(X(t_i) - X(t_j))^2 \\
= \frac{1}{2} \left[ \sigma_Y^2(t_i) + \sigma_Y^2(t_j) \right] - \frac{1}{2} (1 - R(t_j - t_i))[\sigma_X^2(t_i) + \sigma_X^2(t_j)].
\]

Therefore,

\[
\text{cov}(Z_2(t_i), Z_2(t_j)) = \text{corr}(Y(t_i), Y(t_j)) \\
= \text{cov}(Y(t_i), Y(t_j))/\sigma_Y(t_i)\sigma_Y(t_j) \\
\geq \text{cov}(Y(t_i), Y(t_j))/[(\sigma_Y^2(t_i) + \sigma_Y^2(t_j))/2] \\
\geq 1 + (1 - R(t_j - t_i))c_{i,j} \\
= R(t_j - t_i) - (1 - R(t_j - t_i))(c_{i,j} - 1).
\]

Feingold and Siegmund (1993) have shown that the covariance function \(R(t)\) of the exact IBD process has the following form,

\[
R(t) = 1 - \beta|t| + o(t),
\]

where \(\beta\) is a parameter measuring the rate of recombination for the relative pairs involved. For example, when genetic distance \(t\) is measured in centimorgans (cM), for aunt/niece pairs, \(\beta = 0.05\). So the right hand side of (8) can be written in the following form,

\[
R(t_j - t_i) - (1 - R(t_j - t_i))(c_{i,j} - 1) \\
= (1 - \beta|t_j - t_i|) - \beta|t_j - t_i|(c_{i,j} - 1) + o(|t_j - t_i|) \\
= 1 - c_{i,j}\beta|t_j - t_i| + o(|t_j - t_i|).
\]

Remarks: \(c_{i,j}\) is always greater or equal to 1, and the equal holds when the exact IBD can be recovered from the data, which can be achieved when the markers are 100 percent polymorphic or they are infinitely dense. Based on this upper bound for the covariance function, we see the basic structure of the covariance - linear decay,
CHAPTER 2. MULTIPOINT ANALYSIS

does not change. However, for the single point analysis, i.e. making inference about
the IBD configuration based on the single marker alone, there is a substantial change
in the covariance structure. Let’s define

\[ Y^*_j(t_i) = E_0[X^j(t_i)|G^j(t_i)] \]

and

\[ Z^*_i(t_i) = \sum_j(Y^*_j(t_i) - p_0)/\sqrt{n\sigma^*_{Y^*}(t_i)}. \]

If, for example, all the markers have the same polymorphism rate, then

\[ \text{cov}(Z^*_i(t_i), Z^*_j(t_j)) = \begin{cases} 1 & \text{if } t_i = t_j \\ f(1 - \beta|t_i - t_j| + o(|t_i - t_j|) & \text{if } t_i \neq t_j \end{cases} \]

where \( f \) is a constant less than 1. Because the process is substantially less correlated
for single point analysis, we need a significantly larger threshold to control the total
false positive rate.

If the markers are equally spaced at \( \Delta \) cM apart, and have the same polymorphic
information content (PIC), then \( \sigma^2_Y(t_i) \) would be approximately equal across markers,
say \( \sigma^2_Y(t_i) \approx \sigma^2_Y \), and hence \( c_{i,j} \approx c = \sigma^2_X/\sigma^2_Y \). Then using Lemma 4 and Slepian’s
inequality, we can get the following approximate upper bound for the p-value:

\[ \Pr(\max_{1 \leq i \leq M} Z_2(t_i) \geq b) \leq 1 - \Phi(b) + c\beta lb\Phi(b)\nu(b\sqrt{2c\beta\Delta}). \tag{2.9} \]

The threshold based on this upper bound is larger than the threshold derived for
the exact IBD case. However, since the threshold is very robust in the sense that a
small perturbation of the threshold will result in a large change of the p-value, the
discrepancy between these two thresholds tends to be small unless we have very large
value of \( c \).

If the markers are not homogeneous in the sense that they don’t have the same
PIC value or they are not equally spaced or both, we can still get an upper bound like
(2.9) with \( c \) defined as \( \min_{i,j}\{c_{i,j}\} \) and \( \Delta \) as the minimum distance between adjacent
markers. However this upper bound may be too crude for practical use. We propose
to choose \( c \) as the average of \( c_{i,j} \) and \( \Delta \) as the average distance between adjacent
markers. However no theoretical results have been derived to show (9) will still be
an upper bound.
2.2.4 Sibling Pairs

Sibling pairs can share zero, one or two alleles IBD, so they are more complicated than the unilineal pair, but in principle can be treated in the same manner. The underlying Markov chain is \((X^j_p, X^j_m)\), where \(X^j_p\) indicates whether the \(j^{th}\) pair have IBD on their paternal chromosome or not, and \(X^j_m\) indicates whether they have IBD on their maternal chromosome or not. Using the hidden Markov chain argument, we can write down the likelihood as

\[
Pr(G^j) = \sum_{\delta_p=0}^{1} \sum_{\delta_m=0}^{1} Pr_0(G^j | (X^j_p(\tau), X^j_m(\tau)) = (\delta_p, \delta_m)) z_{\delta_p \delta_m},
\]

where \(z_{\delta_p \delta_m} = Pr((X^j_p(\tau), X^j_m(\tau)) = (\delta_p, \delta_m))\).

If an additive model is assumed, then \(z_{00} = (1 - \alpha)/4\), \(z_{10} = z_{01} = 1/4\) and \(z_{11} = (1 + \alpha)/4\). The score test for \(H_0 : \alpha = 0\) is

\[
Z_2(\tau) = \sum_j (Y^j(\tau) - 1)/\left(\sqrt{n} \sigma_Y(\tau)\right),
\]

where \(Y^j(t) = E_0(X^j(t)|G^j)\) and \(X^j(t) = X^j_p(t) + X^j_m(t)\) is the number of alleles shared IBD by the \(j^{th}\) pair.

The result of Lemma 4 still holds. As the unilineal case, we can derive an approximate upper bound for the total false positive rate based on this lemma.

2.3 Efficiency

In the previous sections, we described the multipoint analysis for affected relative pairs, and derived upper bounds for \(p\)-values. In this section, we try to evaluate the effectiveness of multipoint analysis in recovering the information about the status of identity by decent for different values of the polymorphism and density of the markers.

We first calculate the expectation of \(Z(t)\) under the alternative hypothesis. For unilineal relative pairs,

\[
E_\alpha[Z_2(t)] = \frac{\sqrt{n}}{\sigma_Y(t)} E_\alpha[Y(t) - p_0]
\]
and

\[
E_{\alpha}[Y(t)] = \sum_G \Pr_0(X(t) = 1|G)\Pr_\alpha(G) \\
= \frac{\Pr_0(G|X(t) = 1)p_0}{\Pr_0(G|X(t) = 1)p_0 + \Pr_0(G|X(t) = 0)(1 - p_0)} \times [\Pr_0(G|X(\tau) = 1)p_\alpha + \Pr_0(G|X(\tau) = 0)(1 - p_\alpha)] \\
= p_0 + (p_\alpha - p_0) \sum_G \frac{\Pr_0(G|X(t) = 1)p_0}{\Pr_0(G|X(t) = 1)p_0 + \Pr_0(G|X(t) = 0)(1 - p_0)} \times [\Pr_0(G|X(\tau) = 1) - \Pr_0(G|X(\tau) = 0)] \\
= p_0 + (p_\alpha - p_0) \sum_G \Pr_0(X(t) = 1|G) \\
\times [\Pr_0(X(\tau) = 1|G)/p_0 - \Pr_0(X(\tau) = 0|G)/(1 - p_0)] \Pr_0(G) \\
= p_0 + \frac{p_\alpha - p_0}{p_0(1 - p_0)} \cdot E_0[Y(t)(Y(\tau) - p_0)] \\
= p_0 + \frac{p_\alpha - p_0}{p_0(1 - p_0)} \cdot \text{cov}_0(Y(t), Y(\tau)) \\
= p_0 + \frac{\alpha}{\sqrt{p_0(1 - p_0)}} \cdot \text{cov}_0(Y(t), Y(\tau)) \\
= p_0 + \frac{\alpha}{\sqrt{p_0(1 - p_0)}} \cdot \text{corr}_0(Y(t), Y(\tau)) \cdot \sigma_Y(t) \sigma_Y(\tau).
\]

Therefore,

\[
E_{\alpha}[Z_2(t)] = \sqrt{n_\alpha} \frac{\sigma_Y(\tau)}{\sqrt{p_0(1 - p_0)}} \cdot \text{corr}_0(Y(t), Y(\tau)) \\
= \sqrt{n_\alpha} \frac{\sigma_Y(\tau)}{\sigma_X(\tau)} \cdot \text{corr}_0(Y(t), Y(\tau)),
\]

and in particular

\[
E_{\alpha}[Z_2(\tau)] = \sqrt{n_\alpha} \frac{\sigma_Y(\tau)}{\sigma_X(\tau)}.
\]

We first assume that the trait locus \(\tau\) is at exactly one of the marker loci. The above calculations suggest that if instead of using the statistic \(Z_1\) based on the exact IBD, we use the statistic \(Z_2\) based on the estimated IBD, \(Y(t)\), we would have
approximately the same power if we have $\sigma_X^2(\tau)/\sigma_Y^2(\tau)$ many observations. A crude analysis goes as follows: If we only test for one candidate gene locus $\tau$, then this is the usual definition of asymptotic relative efficiency of $Z_2$ with respect to $Z_1$, which says to obtain the same power, asymptotically $\sigma_Y^2(\tau)/\sigma_X^2(\tau)$ as many observations are required with $Z_1$ as with $Z_2$. However, we are in a more complicated situation. Our tests are based on $\max_{1 \leq i \leq M} Z_1(t_i) \geq b$ and $\max_{1 \leq i \leq M} Z_2(t_i) \geq b$, so the usual asymptotic arguments are changed in two ways. First although marginally $Z_1(t)$ and $Z_2(t)$ are the same under the null hypothesis, they have different covariance functions, and hence different thresholds $b$ are required for a given false positive rate. As we noticed earlier, there are usually minor changes in $b$, so this difference plays an almost insignificant role. Secondly, the power of the test can be written as

$$\Pr(\max_{1 \leq i \leq M} Z_j(t_i) \geq b) = \Pr(Z_j(\tau) \geq b) + \Pr(\max_{1 \leq i \leq M} Z_j(t_i) \geq b, Z_j(\tau) \leq b).$$

The first term accounts for the probability that the statistic will exceed the threshold at the trait locus $\tau$ and is a simple function of the noncentrality parameter $E_\alpha[Z_j(\tau)]$. The second term is the probability that the process is below the threshold at the trait locus $\tau$ but, because of random variation, exceeds the threshold at some nearby locus. This depends not only on the noncentrality parameter, but also on the covariance function. However, the first term makes the major contribution to the power for most interesting parameter values. So the ratio of the square of the noncentrality parameters, or, equivalently, $\sigma_Y^2(\tau)/\sigma_X^2(\tau)$, could still be used as a rough measure of relative efficiency in our situation.

Some numerical examples are given for the half-sibling cases. We assume that the markers are equally spaced and each marker has the same number of alleles, all of which occur with equal frequency. Also in calculating the noncentrality, we assume the chromosome is infinitely long so that we can ignore end effects. Figure 2.1 gives the relative efficiency for the case that only half-sibling pairs are typed, while Figure 2.3 gives the relative efficiency for the case that other relevant relatives are also typed.

For the case that the trait locus is not exactly at one of the marker loci, it’s less clear what should be used as a measure of efficiency. Dupuis (1994) suggested that $E_\alpha[\max\{Z_2(t_i), Z_2(t_{i+1})\}]$, where $t_i$ and $t_{i+1}$ are the two flanking markers, is an
appropriate measure, and we are going to use it here. The relative efficiency is defined to be \( E_\alpha^2 [\max\{Z_2(t_i), Z_2(t_{i+1})\}] / E_\alpha^2 [\max\{Z_1(t_i), Z_1(t_{i+1})\}] \).

For the numerical examples, we assume the same conditions for markers as before. We further assume that the trait locus is midway between two markers. Under such conditions, we have

\[
E_\alpha [\max\{Z_2(t_i), Z_2(t_{i+1})\}]
= E_\alpha [(Z_2(t_i) + Z_2(t_{i+1}))/2] + E_\alpha [\sqrt{\frac{1 - \rho(t_i, t_{i+1})}{\pi}}]
= \mu(t_i) + \sqrt{\frac{1 - \rho(t_i, t_{i+1})}{\pi}},
\]

(2.10)

where \( \mu(t) = E_\alpha [Z(t)] \) and \( \rho(t_i, t_{i+1}) = \text{corr}(Z_2(t_i), Z_2(t_{i+1})) \), and for the last equality, we have assumed that \( Z_2(t_i), Z_2(t_{i+1}) \) are jointly normally distributed.

Remarks: 1) The relative efficiency depends on \( \xi = \sqrt{n} \alpha \), so we cannot calculate the relative efficiency without specifying the value of \( \xi \). We use \( \xi = 5 \) in the numerical examples, which corresponds to power of about 90% with infinitely dense completely polymorphic markers. 2) \( \text{corr}_\alpha(Z_2(t_i), Z_2(t_{i+1})) \) could depend on \( \alpha \), but for small value of \( \alpha \),

\[
\text{corr}_\alpha(Z_2(t_i), Z_2(t_{i+1})) \approx \text{corr}_0(Z_2(t_i), Z_2(t_{i+1}))
\]

and the latter is what we used in the numerical examples.

We notice that the first term of (2.10) is proportional to \( \sqrt{n} \), but the second term does not depend on \( n \), so overall \( E_\alpha [\max\{Z_2(t_i), Z_2(t_{i+1})\}] \) is not proportional to \( \sqrt{n} \). Hence we cannot draw a conclusion about the relative sample sizes needed for using statistic \( Z_1 \) based on the exact IBD and using statistic \( Z_2 \) based on the estimated IBD directly from the relative efficiency defined above. To calculate approximately the sample size for a specific value of \( \alpha \) and power, we can first calculating the sample size \( n_0 \) for the ideal case using the approximations by Dupuis (1994), and then determine the sample size \( n \) for the incomplete polymorphism case by solving the equation

\[
(2.10) = E_\alpha [\max\{Z_1(t_i), Z_1(t_{i+1})\}]
= \sqrt{n_0} \alpha \exp(-\beta \Delta/2) + \left(\frac{1 - \exp(-\beta \Delta)}{\pi}\right)^{1/2}.
\]
CHAPTER 2. MULTIPoint ANALYSIS

We did simulations for a very simple case to check the above approximations for sample sizes. For the simulation, we assume each marker could be either fully informative or noninformative with equal probability. This is corresponding to the case that parents of the half sibling pairs also typed and marker PIC value approximately equals to 0.5. The value of \( \alpha \) is taken to be 0.25. In conjunction with \( \xi = 5 \), this requires a sample size of about \( n = 400 \) in the case of completely polymorphic markers. The simulation results are given in Figure 2.5 and 2.6. From this simulation, we see the approximation for sample sizes is quite reasonable, although simulations for more complicated situations needed to be carried out in the future.
CHAPTER 2. MULTIPONT ANALYSIS

Figure Legends

Figure 2.1, 2.3 Relative efficiencies for multipoint analysis for two typing schemes assuming there is zero recombination between the trait locus and marker. The first one assumes only the half-sibling pairs are typed and allele frequencies are used to infer the true identity by descent. The second one assumes the parents of the half-sibling are typed to infer the true identity by descent. Horizontal axis gives the number of alleles at each marker.

Figure 2.2, 2.4 Relative efficiencies for multipoint analysis for two typing schemes assuming the trait locus is midway between two consecutive markers.

Figure 2.5 Comparison of sample sizes required for 90% power to detect linkage assuming there is zero recombination between the trait locus and marker. Horizontal axis gives the distance between markers. Vertical axis gives (i) single point analysis “,” (ii) multipoint analysis based on simulation “x”, (iii) multipoint analysis based on approximation “+”, and (iv) ideal cases “o”.

Figure 2.6 Comparison of sample sizes required for 90% power to detect linkage assuming the trait locus is midway between two consecutive markers.
Figure 2.1: Relative efficiency for trait locus exact at marker locus I

Figure 2.2: Relative efficiency for trait locus midway between markers I
CHAPTER 2. MULTIPoint ANALYSIS

Figure 2.3: Relative efficiency for trait locus exact at marker locus II

Figure 2.4: Relative efficiency for trait locus midway between markers II
Figure 2.5: Sample size for trait locus exact at marker locus

Figure 2.6: Sample size for trait locus midway between markers
Chapter 3

QTLs Based on Sib Pairs

3.1 Introduction

An important class of traits for study in human genetics are quantitative ones. By quantitative traits we mean characters that can be measured on a continuous scale such as height, weight, blood pressure etc. Quantitative traits are usually influenced by many genes and environmental factors. Haseman and Elston (1972) proposed a linear regression method to detect linkage from data on independent sib pairs. Recently this method was extended to allow analysis of pedigree relationships other than siblings (Amos and Elston 1989; Olson and Wijsman 1993).

We begin this chapter by describing the work of Haseman and Elston (1972). Then we develop more powerful tests for linkage. We considered two approaches. One is based on likelihood analysis which requires assumption of normality, but yields efficient tests. So it provides an ideal against which to evaluate the performance of the other statistics when the assumption is ture. Another is based on regression approach which does not require assumption of normality, so can be applied more generally. Finally, sample sizes needed to detect linkage, calculated using asymptotic results, are compared for the usual Haseman-Elston method and our new approach.
3.2 Haseman and Elston’s Method

If $x_{1j}$, $x_{2j}$ represent the trait values for the first and second sibs, respectively, in the $j$th sib pair, then a general model describing the trait values is

\[
\begin{align*}
    x_{1j} &= \mu + g_{1j} + e_{1j} \\
    x_{2j} &= \mu + g_{2j} + e_{2j}
\end{align*}
\]

where $\mu$ is the overall mean, $g_{ij}$ is a fixed major gene component, and $e_{ij}$ is the residual effect which could include both the effect of other genes and the environmental effect. Without loss of generality, $E(g_{ij}) = E(e_{ij}) = 0$ can be assumed, since any average effects can be incorporated into the overall mean of the data. Hardy-Weinberg equilibrium will be assumed throughout.

Let $u_{ij}$, $v_{ij}$ be the alleles inherited from the mother and father, respectively, at the major gene locus. Then the genetic effect $g_{ij}$ is a function of $u_{ij}$ and $v_{ij}$, which can be written as

\[
\begin{align*}
    g_{1j} &= \alpha_{u_{1j}} + \alpha_{v_{1j}} + d_{u_{1j}v_{1j}} \\
    g_{2j} &= \alpha_{u_{2j}} + \alpha_{v_{2j}} + d_{u_{2j}v_{2j}}
\end{align*}
\]

The first two moments describing the data are

\[
E(x_{ij} | \pi_j) = \mu
\]

and

\[
\text{cov}(x_{i1j}, x_{i2j} | \pi_j) = \begin{cases} 
\sigma_d^2 + \sigma_e^2 + \sigma_e^2 & \text{if } i_1 = i_2 \\
\pi_j \sigma_a^2 + \sum_{\pi_j=1} \sigma_a^2 + r \sigma_e^2 & \text{if } i_1 \neq i_2
\end{cases}
\]

where $\pi_j$ is the proportion of alleles that are identical by descent at the trait locus, $\sigma_e^2$ is the residual component of variance, and $r$ is correlation between residual effect which could be due to the polygenic effect or common environmental effect, while $\sigma_a^2 = 2E(\alpha^2)$ and $\sigma_d^2 = E(d^2)$ are, respectively, the additive and dominant components of genetic variance due to the major gene locus.

The linkage test of Haseman and Elston (1972) is based on the idea that the greater the proportion $\pi$ of alleles at the trait locus that are identical by descent for a pair of sibs, the smaller should be the squared difference between the sibs’ trait
values. Let \( Y_j = (x_{1j} - x_{2j})^2 \) be the squared difference for sib pair \( j \). Then based on the above calculation, the expected value of \( Y_j \) conditional on \( \pi_j \) is

\[
E(Y_j|\pi_j) = \beta_0 - \beta_1 \pi_j + \beta_2 I_{\pi_j=1/2},
\]

(3.3)

where

\[
\begin{align*}
\beta_0 &= 2(1 - \tau) \sigma_e^2 + 2 \sigma_a^2 + 2 \sigma_d^2 \\
\beta_1 &= 2(\sigma_e^2 + \sigma_d^2) \\
\beta_2 &= \sigma_d^2.
\end{align*}
\]

Rather than testing the null hypothesis that both \( \beta_1 \) and \( \beta_2 \) are 0, Haseman and Elston (1972) ignore \( \beta_2 \) in the model and, assuming a simple regression of the form \( \beta_0 + \beta_1 \pi \), use standard normal theory to test the hypothesis \( H_0 : \beta_1 = 0 \) against the one-sided alternative \( H_1 : \beta_1 \leq 0 \). Although this is strictly correct only when the dominance component \( \sigma_d^2 \) is zero, their test appears to be satisfactory for nonzero \( \sigma_d^2 \) also (Blackwelder and Elston, 1982).

### 3.3 Likelihood Approach

Instead of analyzing the original bivariate data \((x_{1j}, x_{2j})\) directly, Haseman and Elston first summarized the phenotype data on the sibling pair by a single variable \( Y_j \), then used it as the input data for a regression analysis. A natural question is “Is there any loss of information in doing so?” To explore fully the information in the data, we first consider a maximum-likelihood approach. For this approach, we assume that conditional on \( \pi_j \), the trait values of siblings are multivariate normally distributed. Let

\[
\begin{align*}
\sigma^2 &= \text{var}(x_{ij}|\pi_j) \\
\rho_{\pi j} &= \text{corr}(x_{1j}, x_{2j}|\pi_j)
\end{align*}
\]

The log likelihood of the data is

\[
\log(L) = c - n \log(2\pi \sigma^2) - \frac{1}{2} \sum_{j=1}^{n} \log(1 - \rho_{\pi j}^2) \\
&\quad - \frac{1}{2\sigma^2} \sum_{j=1}^{n} \frac{1}{1 - \rho_{\pi j}^2} \left[ (x_{1j} - \mu)^2 + (x_{2j} - \mu)^2 - 2 \rho_{\pi j} (x_{1j} - \mu)(x_{2j} - \mu) \right],
\]

(3.4)
where $c$ does not depend on any of parameters $(\mu, \sigma^2, \rho_0, \rho_{1/2}, \rho_1)$.

It's convenient to introduce a transformation to new variables by writing

$$
\begin{align*}
\left\{
\begin{array}{l}
u_j = x_{1j} - x_{2j} \\
u_j = x_{1j} + x_{2j} - 2\mu
\end{array}
\right.
\end{align*}
$$

Then the log likelihood can be rewritten as

$$
\log(L) = c - n \log(2\pi\sigma^2) - \frac{1}{2} \sum_{j=1}^{n} \log(1 - \rho_{\pi_j}^2)
$$

$$
- \frac{1}{4\sigma^2} \sum_{j=1}^{n} \frac{1}{1 - \rho_{\pi_j}} \nu_j^2 - \frac{1}{4\sigma^2} \sum_{j=1}^{n} \frac{1}{1 + \rho_{\pi_j}} \nu_j^2.
$$

(3.5)

If we transform the parameters as follow

$$
\begin{align*}
\rho_0 &= \rho - \alpha \\
\rho_{1/2} &= \rho - \delta \\
\rho_1 &= \rho + \alpha
\end{align*}
$$

then according to (3.2), these parameters can be related to the variance components by the equations

$$
\alpha = \frac{1}{2} \frac{\sigma_a^2 + \sigma_d^2}{\sigma_a^2 + \sigma_d^2 + \sigma_e^2},
$$

(3.6)

$$
\delta = \frac{1}{2} \frac{\sigma_d^2}{\sigma_a^2 + \sigma_d^2 + \sigma_e^2}.
$$

(3.7)

So to test if there is any genetic variance component is equivalent to test whether $\alpha = \delta = 0$.

Instead of using the maximum likelihood ratio test, which involves numerical maximization, we will use the score test since it's asymptotically equivalent to the maximum likelihood ratio test and can be carried out analytically. The score statistic can be computed by first fitting $H_0 : \alpha = \delta = 0$, followed by taking the derivative of the log likelihood with respect to $\alpha$ and $\delta$. Both the derivatives and the Fisher information are computed under the MLE of $H_0$.

Define $\psi = (\alpha, \delta)$ to be the parameters we are interested, and $\lambda = (\mu, \sigma^2, \rho)$ to be the nuisance parameters. To carry out the score test, we first try to fit the null
hypothesis. Under $H_0$, the log likelihood (3.4) can be simplified to

$$
\log(L) = c - n \log(2\pi \sigma^2) - \frac{1}{2} n \log(1 - \rho^2) \\
- \frac{1}{2(1 - \rho^2)\sigma^2} \sum_{j=1}^{n} [(x_{1j} - \mu)^2 + (x_{2j} - \mu)^2 - 2\rho(x_{1j} - \mu)(x_{2j} - \mu)].
$$

(3.8)

Maximizing $\log(L)$ over $\mu$, $\rho$ and $\sigma^2$, we get

$$
\hat{\mu} = \frac{1}{2n} \sum_{j=1}^{n} (x_{1j} + x_{2j}), \\
\hat{\sigma}^2 = \frac{1}{2n} \sum_{j=1}^{n} ((x_{1j} - \hat{\mu})^2 + (x_{2j} - \hat{\mu})^2), \\
\hat{\rho} = \frac{\sum_{j=1}^{n} (x_{1j} - \hat{\mu})(x_{2j} - \hat{\mu})}{n\hat{\sigma}^2}.
$$

(3.9)

Then we take the derivatives of the log likelihood (3.5) under the full model with respective to $\alpha$ and $\delta$ and evaluate them at $\psi = 0 = (0, 0)$ and $\lambda = \hat{\lambda} = (\hat{\mu}, \hat{\sigma}^2, \hat{\rho})$. We have

$$
U_\alpha = \frac{\partial}{\partial \alpha} \log(L) = \frac{1}{4(1 - \hat{\rho})^2\hat{\sigma}^2} \sum_{j=1}^{n} (\tilde{u}^2 - u_j^2)(\pi_j - \frac{1}{2}) \\
+ \frac{1}{4(1 + \hat{\rho})^2\hat{\sigma}^2} \sum_{j=1}^{n} (-\tilde{v}^2 + \tilde{v}_j^2)(\pi_j - \frac{1}{2})
$$

(3.10)

and

$$
U_\delta = \frac{\partial}{\partial \delta} \log(L) = \frac{1}{4(1 - \hat{\rho})^2\hat{\sigma}^2} \sum_{j=1}^{n} (\tilde{u}^2 - u_j^2)I_{\pi_j = 1/2} \\
+ \frac{1}{4(1 + \hat{\rho})^2\hat{\sigma}^2} \sum_{j=1}^{n} (-\tilde{v}^2 + \tilde{v}_j^2)I_{\pi_j = 1/2},
$$

(3.11)

where

$$
\tilde{v}_j = x_{1j} + x_{2j} - 2\hat{\mu}, \\
\tilde{u}^2 = \frac{1}{n} \sum_{i=1}^{n} u_i^2, \\
\tilde{v}^2 = \frac{1}{n} \sum_{i=1}^{n} \tilde{v}_i^2.
$$
Finally, we compute the Fisher information matrix and evaluate it at $\psi = 0$ and $\lambda = \lambda$. We have

$$i_{\psi\psi} = \begin{pmatrix} i_{\alpha \alpha} & i_{\alpha \delta} \\ i_{\alpha \delta} & i_{\delta \delta} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{n}{4} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right) & 0 \\ 0 & \frac{n}{4} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right) \end{pmatrix},$$

$$i_{\psi \lambda} = \begin{pmatrix} i_{\alpha \mu} & i_{\alpha \sigma^2} & i_{\alpha \rho} \\ i_{\delta \mu} & i_{\delta \sigma^2} & i_{\delta \rho} \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 \\ 0 & \frac{n}{4 \delta^2} \left( \frac{1}{1-\rho} - \frac{1}{1+\rho} \right) & \frac{n}{4} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right) \end{pmatrix},$$

and

$$i_{\lambda \lambda} = \begin{pmatrix} i_{\mu \mu} & i_{\mu \sigma^2} & i_{\mu \rho} \\ i_{\mu \sigma^2} & i_{\sigma^2 \sigma^2} & i_{\sigma^2 \rho} \\ i_{\mu \rho} & i_{\sigma^2 \rho} & i_{\rho \rho} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{2n}{(1+\rho)\delta^2} & 0 & 0 \\ 0 & \frac{n}{\delta^2} & -\frac{n}{2 \delta^2} \left( \frac{1}{1-\rho} - \frac{1}{1+\rho} \right) \\ 0 & -\frac{n}{2 \delta^2} \left( \frac{1}{1-\rho} - \frac{1}{1+\rho} \right) & \frac{n}{2} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right) \end{pmatrix}.$$ 

Then

$$i_{\psi \psi} - i_{\psi \lambda} i_{\lambda \lambda}^{-1} i_{\psi \lambda}^T = \begin{pmatrix} \frac{n}{4} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right) & 0 \\ 0 & \frac{n}{8} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right) \end{pmatrix}.$$ (3.12)

Let

$$Z_1 = \frac{U_\alpha}{\sqrt{\frac{n}{4} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right)}},$$

$$Z_2 = \frac{U_\delta}{\sqrt{\frac{n}{8} \left( \frac{1}{(1-\rho)^2} + \frac{1}{(1+\rho)^2} \right)}}.$$
CHAPTER 3. QTLS BASED ON SIB PAIRS

According to asymptotic theory, under $H_0$, $Z_1$ and $Z_2$ are approximately independent standard normal random variables, and under local alternatives $\psi = \xi/\sqrt{n} = \left(\xi_1/\sqrt{n}, \xi_2/\sqrt{n}\right)$, $Z_1$ and $Z_2$ are approximately independent unit variance normal random variables with mean

$$\sqrt{\frac{1}{4} \left(\frac{1}{(1 + \rho)^2} + \frac{1}{(1 - \rho)^2}\right)} \xi_1$$

and

$$\sqrt{\frac{1}{8} \left(\frac{1}{(1 + \rho)^2} + \frac{1}{(1 - \rho)^2}\right)} \xi_2$$

respectively.

The likelihood analysis is a systematic procedure that in some sense provides good approximate solutions. But in order to carry out the likelihood analysis, the distribution of the data needs to be explicitly specified, in our case, we have to assume that $(x_{1j}, x_{2j})$ are bivariate normally distributed which is a rather restrictive assumption. However, although we don’t expect the normality assumption holds exactly, in many cases, it’s reasonable to expect it to hold approximately. So the above analysis will provide us some insight about the problem. Especially, we notice that the information can be approximately divided into two parts, one from $u_j$ and one from $v_j$. By ignoring $v_j$ in their analysis, Haseman and Elston’s analysis loses some information. This observation also leads us to the following generalized regression approach.

### 3.4 Generalized Regression

Haseman-Elston’s method is based on a linear relationship between $u_j^2$ and $\pi_j$. A straightforward calculation based on (3.2) shows that there is also a similar relationship between $v_j^2$ and $\pi_j$ which can be formally written as,

$$E(v_j^2|\pi_j) = \beta'_0 + \beta_1 \pi_j - \beta_2 I_{\pi_j=1/2},$$
where $\beta'_0 = 4\mu^2 + 2(1 + r)\sigma^2 + 2\sigma_d^2 + 2\sigma_a^2$ and $\beta_1, \beta_2$ are defined the same as before. This result, combined with (3.3) leads to the following multireponse linear model.

\[
\begin{align*}
  u_j^2 &= \beta_0 - \beta_1 \pi_j + \beta_2 I_{\pi_j=1/2} + \varepsilon_{1j} \\
  v_j^2 &= \beta'_0 + \beta_1 \pi_j - \beta_2 I_{\pi_j=1/2} + \varepsilon_{2j}
\end{align*}
\]

(3.15) (3.16)

where

\[
E(\varepsilon_j) = E\left(\begin{array}{c}
  \varepsilon_{1j} \\
  \varepsilon_{2j}
\end{array}\right) = 0
\]

and

\[
cov(\varepsilon_j) = cov\left(\begin{array}{c}
  \varepsilon_{1j} \\
  \varepsilon_{2j}
\end{array}\right) = \begin{pmatrix}
  \sigma_{1}^2 & s\sigma_1\sigma_2 \\
  s\sigma_1\sigma_2 & \sigma_2^2
\end{pmatrix} = \Sigma
\]

Remark: In general, $cov(\varepsilon_j)$ may depend on $\pi_j$ as well. However, it's reasonable to expect it to be constant under null hypothesis and approximately constant under local alternatives.

A standard general least square procedure can be used. First, find values of $\beta_0$, $\beta'_0$, $\beta_1$ and $\beta_2$ which minimize

\[
\sum_{i=1}^{n} \left(\begin{array}{c}
  u_i^2 - \beta_0 + \beta_1 \pi_j - \beta_2 I_{\pi_j=1} \\
  v_i^2 - \beta'_0 - \beta_1 \pi_j + \beta_2 I_{\pi_j=1}
\end{array}\right) \Sigma^{-1} \left(\begin{array}{c}
  u_i^2 - \beta_0 + \beta_1 \pi_j - \beta_2 I_{\pi_j=1} \\
  v_i^2 - \beta'_0 - \beta_1 \pi_j + \beta_2 I_{\pi_j=1}
\end{array}\right)
\]

Straightforward calculations give us the following result,

\[
\hat{\beta}_1 = w_1 \hat{\beta}_{11} + w_2 \hat{\beta}_{12} \\
\hat{\beta}_2 = w_1 \hat{\beta}_{21} + w_2 \hat{\beta}_{22}
\]

where

\[
\begin{align*}
  \hat{\beta}_{11} &= \frac{\sum_{j=1}^{n}(u_j^2 - \bar{u}^2)(-\pi_j + \bar{\pi})}{\sum_{j=1}^{n}(\pi_j - \bar{\pi})^2} \\
  \hat{\beta}_{12} &= \frac{\sum_{j=1}^{n}(v_j^2 - \bar{v}^2)(\pi_j - \bar{\pi})}{\sum_{j=1}^{n}(\pi_j - \bar{\pi})^2}
\end{align*}
\]

are the least square estimators of $\beta_1$ based on (3.15) and (3.16) respectively.
\[ \hat{\beta}_{21} = \frac{\sum_{j=1}^{n}(v_j^2 - \bar{v}^2)(I_{i_j=1} - \bar{I}_{i=1})}{\sum_{j=1}^{n}(I_{i_j=1} - \bar{I}_{i=1})^2} \]

\[ \hat{\beta}_{22} = \frac{\sum_{j=1}^{n}(v_j^2 - \bar{v}^2)(-I_{i_j=1} + \bar{I}_{i=1})}{\sum_{j=1}^{n}(I_{i_j=1} - \bar{I}_{i=1})^2} \]

are the least square estimators of \( \beta_2 \) based on (3.15) and (3.16) respectively. And

\[ w_1 = \frac{\sigma_1^2 + s\sigma_2 \sigma_2}{\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2} \]

\[ w_2 = \frac{\sigma_1^2 + s\sigma_2 \sigma_2}{\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2} \]

Here we did not write down the estimator for \( \beta_0 \) or \( \beta'_0 \) since we are only interested in testing \( \beta_1 = \beta_2 = 0 \).

It can be shown that

\[ \text{var}(\hat{\beta}_1) = \frac{(1 - s^2)\sigma_1^2 \sigma_2^2}{(\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2) \sum_{j=1}^{n}(i_j - \bar{i})^2} \]

and

\[ \text{var}(\hat{\beta}_2) = \frac{(1 - s^2)\sigma_1^2 \sigma_2^2}{(\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2) \sum_{j=1}^{n}(I_{i_j=1} - \bar{I}_{i=1})^2} \]

If we define

\[ Z'_1 = \frac{\hat{\beta}_1}{\sqrt{\text{var}(\hat{\beta}_1)}} \]

\[ Z'_2 = \frac{\hat{\beta}_2}{\sqrt{\text{var}(\hat{\beta}_2)}} \]

Then according to asymptotic theory, under \( H_0 \), \( Z'_1 \) and \( Z'_2 \) are approximately standard normal random variables and under local alternatives given by \( \beta_1 = \xi'_1/\sqrt{n} \) and \( \beta_2 = \xi'_2/\sqrt{n} \), \( Z'_1 \) and \( Z'_2 \) are approximately independent unit variance normal random variables with means

\[ \frac{1}{2\sqrt{2}} \sqrt{\frac{\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2}{(1 - s^2)\sigma_1^2 \sigma_2^2}} \xi'_1 \]

and

\[ \frac{1}{2\sqrt{2}} \sqrt{\frac{\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2}{(1 - s^2)\sigma_1^2 \sigma_2^2}} \xi'_2 \]
CHAPTER 3. QTLS BASED ON SIB PAIRS

respectively.

However, in practice, \( \mu \) and \( \Sigma \) are usually unknown. So a usable form of \( Z_1' \) and \( Z_2' \) is obtained by replacing \( \mu \) by \( \hat{\mu} \) and \( \Sigma \) by its consistent estimator \( \hat{\Sigma} \) which will be defined later. It can be shown this substitution adds only \( o_p(1) \) terms. So the asymptotic properties of \( Z_1' \) and \( Z_2' \) will not change.

A consistent estimator of \( \Sigma \) can be derived by the following procedure. First fit regression (3.15) and (3.16) separately, we get the least square estimators \( (\hat{\beta}_0, \hat{\beta}_{11}, \hat{\beta}_{12}) \) and \( (\hat{\beta}_0', \hat{\beta}_{21}, \hat{\beta}_{22}) \) respectively. Let

\[
\hat{\varepsilon}_{1j} = u_j^2 - \hat{\beta}_0 - \hat{\beta}_{11} \pi_j - \hat{\beta}_{12} I_{\pi_j = 1/2}
\]

\[
\hat{\varepsilon}_{2j} = v_j^2 - \hat{\beta}_0' - \hat{\beta}_{21} \pi_j - \hat{\beta}_{22} I_{\pi_j = 1/2}
\]

and

\[
\hat{\sigma}_{i_1i_2} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_{i_1j} \hat{\varepsilon}_{i_2j} \ for \ i_1 = 1, 2 \ and \ i_2 = 1, 2
\]

Let's consider the special case that \( (x_{1j}, x_{2j}) \) are bivariate normal conditional on \( \pi_j \). It can be easily shown that under \( H_0 \) and local alternative,

\[
\Sigma = \begin{pmatrix}
8(1 - \rho)^2 \sigma^4 & 0 \\
0 & 8(1 + \rho)^2 \sigma^4
\end{pmatrix}
\]

where \( \rho \) and \( \sigma^2 \) are the same as in section 2.1. Then the noncentralities of \( Z_1' \) and \( Z_2' \) become

\[
\frac{1}{8} \sqrt{\frac{1}{(1 - \rho)^2} + \frac{1}{(1 + \rho)^2} \frac{\xi_1'}{\sigma^2}}
\]

and

\[
\frac{1}{4\sqrt{2}} \sqrt{\frac{1}{(1 - \rho)^2} + \frac{1}{(1 + \rho)^2} \frac{\xi_2'}{\sigma^2}}
\]

Based on (3.6) and (3.7), we have

\[
\beta_1/\sigma^2 = 4\alpha \ and \ \beta_2/\sigma^2 = 2\delta
\]

and consequently,

\[
\xi_1'/\sigma^2 = 4\xi_1 \ and \ \xi_2'/\sigma^2 = 2\xi_2.
\]

\( (Z_1', Z_2') \) and \( (Z_1, Z_2) \) have the same noncentralities.
Since the regression approach does not depend on the normality assumption in general, but is asymptotically equivalent to the likelihood approach when the normality assumption is true, it seems preferable to the likelihood approach. We’ll focus on it in the rest of this chapter.

3.5 Test Statistics

So far, we have not described the test explicitly. Based on the results of the previous section, the log likelihood ratio function of \( Z'_1, Z'_2 \) can be approximated by

\[
\xi'_1 Z'_1 + \xi'_2 Z'_2 - \left( \frac{\xi'_1^2 + \xi'_2^2}{2} \right)
\]

where

\[
\xi'_1 = \frac{1}{2\sqrt{2}} \sqrt{\frac{\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2}{(1 - s^2)\sigma_1^2 \sigma_2^2}} \xi'_1
\]

and

\[
\xi'_2 = \frac{1}{2} \sqrt{\frac{\sigma_1^2 + 2s\sigma_1 \sigma_2 + \sigma_2^2}{(1 - s^2)\sigma_1^2 \sigma_2^2}} \xi'_2
\]

are the noncentralities of \( Z'_1 \) and \( Z'_2 \) respectively. This implies that if the ratio \( \xi'_1/\xi'_2 \) is known, or equivalently, \( \sigma_1^2/\sigma_2^2 \) is known, our test statistic would be the linear combination

\[
Z' = \frac{\xi'_1 Z'_1 + \xi'_2 Z'_2}{\sqrt{\xi'_1^2 + \xi'_2^2}}
\]

\[
= \frac{\sqrt{2}(\sigma_1^2 + \sigma_2^2)}{\sqrt{\sigma_1^2 + 2(\sigma_1^2 + \sigma_2^2)^2}} Z'_1 + \frac{\sigma_2^2}{\sqrt{\sigma_1^2 + 2(\sigma_1^2 + \sigma_2^2)^2}} Z'_2
\]  

which has standard normal distribution under the null hypothesis. Under the local alternative, it has noncentrality

\[
\sqrt{\xi'_1^2 + \xi'_2^2}.
\]

When the ratio \( \sigma_1^2/\sigma_2^2 \) is unknown, we can still consider linear combinations of \( Z'_1 \) and \( Z'_2 \) as the the test statistics, but the optimal weights for \( Z'_1 \) and \( Z'_2 \) can’t be
determined. However, we hope we may be able to find weights that have high relative efficiency in most important cases. Suppose we use the linear combination

\[ c_1 Z'_1 + c_2 Z'_2, \text{ where } c_1^2 + c_2^2 = 1. \] (3.18)

Then the noncentrality of this statistic equals \( c_1 \xi_1^* + c_2 \xi_2^* \).

So the relative efficiency of statistic (3.18) and (3.17) is

\[ \frac{(c_1 \xi_1^* + c_2 \xi_2^*)^2}{\xi_1^{*2} + \xi_2^{*2}} \]

For example, the linear combination \(0.953Z'_1 + 0.303Z'_2\) is at least 91% efficient for all cases. The linear combination \(0.986Z'_1 + 0.169Z'_2\) is at least 81% efficient for all cases and at least 97% efficient for cases such that \(\sigma_d^2 \geq \sigma_a^2\). Another important test statistic is \(Z'_1\). This corresponds to the optimal linear combination under the additive model, i.e. \(\sigma_d^2 = 0\). In general, its asymptotic relative efficiency is

\[ \frac{\xi_1^{*2}/8}{\xi_1^{*2}/8 + \xi_2^{*2}/4} = \frac{2(\sigma_a^2 + \sigma_d^2)^2}{\sigma_d^2 + 2(\sigma_a^2 + \sigma_d^2)^2} \]

which is at least 89% if \(\sigma_d^2 \leq \sigma_a^2\), and would be between 67% and 89% if \(\sigma_d^2 \geq \sigma_a^2\). The asymptotic relative efficiencies against the optimal weight for the above three linear combinations as a function of the ratio \(\sigma_d^2/\sigma_a^2\) are plotted in Figure 3.1.

Once \(c_1\) and \(c_2\) are determined, a standard normal test can be used for testing a single marker for linkage, i.e.

\[ \text{reject } H_0, \text{ if } Z' \geq Z_{1-\alpha}. \]

where \(Z_{1-\alpha}\) is the \(1 - \alpha\)th quantile of normal distribution.

A different strategy would be to choose the value of \(\xi_1^*\) and \(\xi_2^*\) maximizing the log likelihood. If we don’t put any constraint on \(\xi_1^*, \xi_2^*\), this would induce the chi-square test

\[ \text{reject } H_0, \text{ if } Z_1^2 + Z_2^2 \geq \chi_2^2(1 - \alpha). \]

where \(\chi_2^2(1 - \alpha)\) is the \(1 - \alpha\)th quantile of chi-square with 2 degree of freedom.
CHAPTER 3. QTLS BASED ON SIB PAIRS

However we have some natural constraints on this parameter space, i.e. $\xi_1^* \geq 0$, $\xi_2^* \geq 0$ and $\xi_1^*/\xi_2^* \geq \sqrt{2}$. If we take maximum of the log likelihood over this subset, we get the following test

$$\text{reject } H_0, \text{ if } Z^* \geq c_\alpha.$$  

where

$$Z^* = \begin{cases} 
0 & \text{if } Z'_1 \leq 0 \\
Z'_1 & \text{if } Z'_1 \geq 0, \ Z'_2 \leq 0 \\
\sqrt{Z'_1^2 + Z'_2^2} & \text{if } Z'_1 \geq \sqrt{2} Z'_2 \geq 0 \\
\sqrt{2/3} Z'_1 + \sqrt{1/3} Z'_2 & \text{if } \sqrt{2} Z'_2 \geq Z'_1 \geq 0
\end{cases}$$

and $c_\alpha$ satisfies

$$\Pr(Z^* \geq c_\alpha) = \alpha, \quad \text{under } H_0.$$

Regarding $Z'_1$ and $Z'_2$ as independent normal variables, we have

$$\Pr(Z^* \geq c_\alpha) = 1 - \Phi(c_\alpha) + [\tan^{-1}(1/\sqrt{2})/(2\pi)] \Pr(\chi_2^2 \geq c_\alpha^2)$$

$$\approx 1 - \Phi(c_\alpha) + 0.1 \Pr(\chi_2^2 \geq c_\alpha^2).$$

where $\Phi(x)$ is the cumulative distribution function of standard normal variable and $\chi_2^2$ is a chi-squared random variable with 2 degree of freedom. It's obvious that

$$Z_{1-\alpha} \leq c_\alpha \leq \sqrt{\chi_2^2(1 - \alpha)}.$$

It's simple to verify that

$$Z^* = \max_{\xi'_1 \geq \sqrt{2}\xi'_2 \geq 0} (\xi'_1 Z'_1 + \xi'_2 Z'_2)/\sqrt{\xi'_1^2 + \xi'_2^2}$$

Thus this statistic is equivalent to the most significant value among all possible optimal linear combination test statistics.

Neither the linear combination statistics nor the maximum likelihood ratio statistic is uniformly superior to the other. A formal comparison of these two tests depends on the significance value and the noncentralities, which we are not going to pursue here. Feingold and Siegmund (1995) discussed the comparison for a similar situation.
3.6 Comparing with Haseman-Elston’s Method

Haseman-Elston’s method seems inadequate in two aspects: 1) It only tests $\beta_1 = 0$ and ignores $\beta_2$; 2) It only uses $u_j$ in the analysis and ignores $v_j$. The effect of 1) corresponds to using $Z'_1$ as the test statistic as we discussed in the previous section. It performs quite well when the trait is reasonably additive ($\sigma_d^2 \leq \sigma_a^2$), but may lose substantial efficiency otherwise. We are going to focus on the effect of 2) in this section. To do that, we will compare the Haseman-Elston’s statistic with $Z'_1$.

The test statistic based on the simple regression $E(u_j^2 | \pi_j) = \beta_0 - \beta_1 \pi_j$ is

$$Z = \frac{\sum_{j=1}^n (u_j^2 - \bar{u}^2)(\pi_j - \bar{\pi})}{\sqrt{\sum_{j=1}^n (\pi_j - \bar{\pi})^2}}.$$  

It’s simple to verify that $Z$ is asymptotically $N(0, 1)$ under $H_0$, and has noncentrality

$$\frac{\xi'_1}{2\sqrt{2}\sigma_a}$$

under the local alternative of section 2.4. Thus the asymptotic relative efficiency of $Z$ to $Z'_1$ is

$$\frac{(1 - s^2)\sigma_a^2}{\sigma_1^2 + 2s\sigma_1\sigma_2 + \sigma_2^2},$$

which is always between 0 and 1. For the special case of the bivariate normal distribution, this relative efficiency becomes

$$\frac{(1 + \rho)^2}{(1 - \rho)^2 + (1 + \rho)^2},$$

which is a monotone increasing function of $\rho$. It has maximum of 1 when $\rho = 1$ and minimum of 0.5 when $\rho = 0$.

If the environmental contributions to the phenotypes of sibling pairs are uncorrelated, then

$$\rho = \frac{1}{2} h^2 + \frac{1}{4} D^2,$$

where $h^2$ and $D^2$ are the proportions of the phenotypic variance due to purely additive gene action and dominance, respectively. These measure how important the genes are in determining the trait value. Our method gains most when the value of $\rho$ is small,
or equivalently, the genetic effect is small. This is the least favorable situation for detecting linkage, where we would be particularly concerned about loss of efficiency. But when the value of $\rho$ increases, the relative gain by using our method will become small. Figure 3.2 gives the asymptotic relative efficiency of Haseman and Elston's statistics against $Z_1'$ as a function of $\rho$. For examples, if $\rho = 0.2$, a test based on $Z_1'$ could save about 31% of the sample size required by Haseman-Elston's method, if $\rho = 0.4$, a test based on $Z_1'$ could save about 16% of the sample size.

Combining Figure 3.1 and 3.2, we can compare Haseman-Elston's statistic $Z$ with statistics based on other linear combinations of $Z_1'$ and $Z_2'$. For example, if $\rho = 0.2$ and $\sigma_2^2 / \sigma_a^2 = 0$, $Z$ is 69% efficient compared to $Z_1'$, while the later is 100% efficient compared to the optimal linear combination of $Z_1'$ and $Z_2'$, so $Z$ is 69% as efficient as the optimal linear combination. Under the same condition, $0.986Z_1' + 0.169Z_2'$ is about 97% efficient. Therefore the relative efficiency of $Z$ to $0.986Z_1' + 0.169Z_2'$ is $(69/97) \times 100\% = 71\%$. However, if we keep $\rho$ unchanged, and let $\sigma_2^2 / \sigma_a^2 = 0.8$, then the relative efficiency of $Z$ to the optimal linear combination of $Z_1'$ and $Z_2'$ becomes $69\% \times 91\% = 63\%$, and its relative efficiency to $0.986Z_1' + 0.169Z_2'$ is $63/98 = 64\%$.

### 3.7 Genome Wide Scan

In previous sections, we assume that the trait locus is known, which is usually unrealistic in practice. We may have to search the whole genome for the possible trait locus. We calculate the test statistic (3.18) for each locus that the IBD status is known. Then we get a process

$$Z_t' = c_1Z_{1,t} + c_2Z_{2,t}, \ t \in T$$

Our overall test statistic is the maximum value of the process over the index set $T$. In order to define a critical region for the test, we need to find the distribution of this statistic under $H_0$. We assume our data come from a dense set of polymorphic markers, or from a method such as Genomic Mismatch Scanning (Nelson et al. 1993) that gives whole-genome IBD information.

Direct calculations show that as $n \to \infty$, $Z_{1,t}$ and $Z_{2,t}$ are Gaussian processes with
mean 0 and covariance function $e^{-4\lambda|t|}$ and $e^{-8\lambda|t|}$, respectively. Moreover, $Z_{1,t}$ and $Z_{2,t}$ are asymptotically independent. So the covariance function of $Z_t'$ is

$$c_1^2 e^{-4\lambda|t|} + c_2^2 e^{-8\lambda|t|} \approx 1 - \beta|t|,$$

where

$$\beta = 4\lambda(1 + c_2^2).$$

An approximation to the tail probability (Feingold et al. 1993) is

$$\Pr\{\max_{t \in T} Z_t' \geq b\} \approx 1 - \exp[-n(1 - \Phi(b)) - \beta lb\phi(b)],$$

where $\phi$ and $\Phi$ are the standard normal density and distribution, respectively, and $l$ is the total genetic length of the region of the genome being searched, and $n$ is the number of chromosome being searched.

Under the alternative hypothesis, assuming there is only one gene on the chromosome associated with the trait and the trait locus is $r$, the expected value of $Z_t'$ is

$$c_1\xi_1^* e^{-4\lambda|t-r|} + c_2\xi_2^* e^{-8\lambda|t-r|},$$

where $\xi_1^*, \xi_2^*$ are defined in section (2.5). We can rewrite this in the form of $\xi R_1(t-r)$, where $\xi = c_1\xi_1^* + c_2\xi_2^*$ and

$$R_1(x) = \frac{c_1\xi_1^*}{c_1\xi_1^* + c_2\xi_2^*} e^{-4\lambda|x|} + \frac{c_2\xi_2^*}{c_1\xi_1^* + c_2\xi_2^*} e^{-8\lambda|x|}$$

$$\approx 1 - \beta_1|x|,$$

where

$$\beta_1 = 4\lambda \left(1 + \frac{c_2\xi_2^*}{c_1\xi_1^* + c_2\xi_2^*}\right).$$

An approximation to the power (Feingold et al. 1993) is

$$1 - \Phi(b - \xi) + \phi(b - \xi) \left\{ \frac{2}{\beta_1} \frac{1}{\xi} - \left[ \frac{2}{\beta_1} - 1 \right] + b \right\}^{-1}. $$
CHAPTER 3. QTLS BASED ON SIB PAIRS

For the likelihood ratio statistic $Z^*$, the overall p-value and power can be approximated by an arguments similar to those used in Feingold et al. (1993). Under $H_0$,\
\[
\Pr(\max Z_i^* \geq b) \approx 1 - \exp[-I \exp(-b^2/2)\{Cb^2/(2\pi) + b(14/3\lambda)/(2\pi)^{1/2}\}],
\]
where\
\[
C = \int_0^{\tan^{-1} 2^{-1/2}} (4\lambda \cos^2 w + 8\lambda \sin^2 w)dw
\]
\[
= 0.0275.
\]

The approximation to the power has two slightly different formulas depends on whether the the alternative is on the boundary, i.e. $\xi_2^* = 0$ or $\xi_1^*/\xi_2^* = \sqrt{2}$, or not. In the former case, the power can be approximated by\
\[
1 - \Phi(b - \xi^*) + \phi(b - \xi^*) \left\{ \frac{1}{4\xi^*} + \left[ \frac{b}{\xi^*} \right]^{1/2} + 1 \right\} \left[ \frac{1}{\xi^*} - \frac{1}{2(b + \xi^*)} \right]
\]
while in the latter it is\
\[
1 - \Phi(b - \xi^*) + \phi(b - \xi^*) \left\{ \frac{1}{2\xi^*} + \left[ \frac{b}{\xi^*} \right]^{1/2} \right\} \left[ \frac{2}{\xi^*} - \frac{1}{(b + \xi^*)} \right]
\]
where\
\[
\xi^* = \xi_1^* + \xi_2^*
\]

Feingold and Siegmund (1995) give a detailed discussion for the qualitative trait for sib pairs, and our results are completely parallel to theirs.
Figure Legends

Figure 3.1  Relative efficiencies of the three linear combinations of $Z'_1$ and $Z'_2$ compared to the optimal linear combination. Horizontal axis gives the ratio of $\sigma_0^2/\sigma_a^2$. (i) The solid line is for the linear combination $0.953Z'_1 + 0.303Z'_2$, (ii) The dash line is for the linear combination $0.986Z'_1 + 0.169Z'_2$, (iii) The dotted line is for the linear combination $Z'_1$.

Figure 3.2  Relative efficiencies of Haseman and Elston's statistics to $Z'_1$. Horizontal axis gives the correlation $\rho$. 
CHAPTER 3. QTLS BASED ON SIB PAIRS

Figure 3.1: Comparison of different linear combinations

Figure 3.2: Comparison with Haseman and Elston's statistic
Chapter 4

Quantitative Trait Loci (QTLs)

4.1 Introduction

The detection of genes that control quantitative characters in plants and animals is an important problem with a broad range of applications.

The purpose of this paper is to discuss the methodology for mapping QTLs in experimental organisms. The methods are for the particular case of controlled breeding and do not apply as is to human data. A breeding experiment consists of arranging a cross between two parental strains that differ substantially in the quantitative trait of interest. The parental lines are often “pure” breeding lines obtained through in-breeding or simply two different strains of the same organism with differing mean phenotype. Typically, the progeny are produced by an intercross or a backcross. We will focus on the relative simple case of backcross in this chapter.

4.2 Backcross

Let A and B be inbred strains differing for a quantitative trait of interest. An backcross is generated by mating the two lines that produced the first generation of offspring (generation $F_1$). The $F_1$ generation is then mated with A as the recurrent parent to produce the second generation ($F_2$). The offspring from the backcross will have 0 or 1 alleles from strain B at any locus on their genome. The model proposed in
CHAPTER 4. QUANTITATIVE TRAIT LOCI (QTLS)

Lander and Botstein (1989) is as follows. Let \( L \) be the map length of the chromosome, and a position on the chromosome is denoted by a value \( t \in [0, L] \). The genotype for Backcross individual \( i \) at a location \( t \) is coded by

\[
g_i(t) = \begin{cases} 
-1 & \text{no alleles from strain B} \\
1 & \text{one allele from strain B} 
\end{cases}
\]

Let \( y_i \) be the value of the quantitative trait for individual \( i \). If there exists only one gene located at \( r \) that influence the traits, then the phenotype can be modeled as

\[
y_i = \mu + \alpha g_i(r) + e_i,
\]

where the \( e_i \)'s are usually assumed to be normally distributed with mean 0 and variance \( \sigma^2 \).

4.3 Testing for the genetic effect

4.3.1 Test statistic

The log likelihood ratio for testing \( H_0 : \alpha = 0 \) versus \( H_1 : \alpha \neq 0 \) for \( r \) and \( g_i(r) \) known is

\[
LR(r) = \frac{n}{2} \log \left( 1 + \frac{\hat{\alpha}^2(r) \sum (g_i(r) - \bar{g}(r))^2}{n\hat{\sigma}_e^2} \right),
\]

which is equivalent to the t-test based on statistic

\[
\frac{\hat{\alpha}(r)(\sum (g_i(r) - \bar{g}(r))^2)^{1/2}}{n\hat{\sigma}_e/(n-2)}.
\]

where

\[
\hat{\alpha}(r) = \frac{\sum_{i=1}^n [y_i - \bar{y}] [g_i(r) - \bar{g}(r)]}{\sum_{i=1}^n [g_i(r) - \bar{g}(r)]^2},
\]

and

\[
\hat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y} - \hat{\alpha}(r)(g_i(r) - \bar{g}(r)))^2.
\]

When \( r \) in unknown, the log likelihood ratio statistic is calculated for every marker location \( d \), and the test statistic becomes

\[
\max_d Z(d)^{1/2}
\]
where
\[ Z(d) = [2LR(d)]^{1/2} \]

### 4.3.2 P-value

When an entire genome is tested for the presence of QTLS, one important issue is: What threshold \( b \) should be used in order to maintain an acceptably low rate of false positives? So we need to find an appropriate value \( b \) such that
\[ \Pr\{\max_d Z(d) \geq b\} = \alpha, \]
for some predetermined significant level \( \alpha \), which is usually taken to equal 0.05.

Assuming \( d \) is continuous, Lander and Botstein (1989) showed that as \( n \to \infty \), \( \max_d \{Z(d)\} \) has the same distribution as the extrema of the absolute value of an Ornstein-Uhlenbeck process and provided an approximation for the tail probability. Dupuis generalized their results to the cases that \( g_i(d) \) are only known at equispaced distance of \( \Delta \) centimorgans and gave the following approximation
\[ \Pr\{\max_d Z(d) \geq b\} \sim 2[1 - \Phi(b) + 2\lambda Lb\phi(b)\nu(2b\sqrt{\lambda\Delta})], \]
(4.1)
where \( L \) is the length of the chromosome and \( \nu(x) \approx \exp[-0.583x] \). The exact definition of \( \nu \) is given by Siegmund (1985, p.82).

The proof of (4.1) followed the method by Woodroofe (1976). The total probability was broken down according to the last index for which \( X(k\Delta) \) exceeds \( b \) so that,
\[ \Pr(\max_k Z(k\Delta) \geq b) = \sum_{k=0}^{L/\Delta} \int_0^\infty \Pr(Z(k\Delta) \in b + dx) \Pr(\max_{k<b\Delta} Z(l\Delta) < b|Z(k\Delta) = b + x). \]

Then (4.1) was obtained by approximating the marginal distribution of \( Z(k\Delta) \) by a normal random variable and the conditional distribution of \( \{Z((k - 1)\Delta), Z((k - 2)\Delta), \ldots\} \) given \( Z(k\Delta) \) by a Gaussian random walk.

(4.1) gives an approximation to the tail probability for the limiting case of \( n \to \infty \), which is not especially accurate for moderately large \( n \). This is due to the following
two reasons: (i) The marginal distribution is not exactly a normal distribution. Since we are considering extreme tail probabilities, this approximation may not be very accurate unless the sample size is very large. (ii) The conditional distribution is not a Gaussian process unless \( n \Delta \) is very large, which is not always the case. We have tried to address these two problems to give a more accurate approximation for the tail probability of \( \Pr\{ \max_d Z(d) \geq b \} \).

First we consider a simpler case that \( \mu \) is known. In this case, the least square estimator of \( \alpha \) is

\[
\hat{\alpha} = \frac{\sum_{i=1}^{n} (y_i - \mu) g_i(t)}{\sum_{i=1}^{n} g_i^2(t) / n},
\]

and the t-statistic for testing \( H_0 : \alpha = 0 \) versus \( H_1 : \alpha \neq 0 \) is

\[
Z'(r) = \frac{\sqrt{n} \hat{\alpha}(r)}{\sum_{i=1}^{n} (y_i - \mu - \hat{\alpha} g_i(t))^2 / (n - 1)}. \tag{4.2}
\]

Both the numerator and denominator of (4.2) are changing with the location \( t \). To simplify, we define a new statistic,

\[
Z''(t) = \left\{ \frac{\sum_{i=1}^{n} (y_i - \mu) g_i(t)}{\sqrt{n} \hat{\sigma}} \right\}, \tag{4.3}
\]

where

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n - 1}.
\]

The test based on \( Z''(t) \) is equivalent to the one based on \( Z'(t) \) since \( Z''(t) \) is a monotone function of \( Z'(t) \) with

\[
Z''(t) = \frac{Z'(t)}{\sqrt{1 + Z'(t)^2 / (n - 1)}}.
\]

So we need to calculate

\[
\Pr( \max_{0 \leq k \Delta \leq L} Z''(k \Delta) \geq b). \tag{4.4}
\]

Using the same idea as before, we partitioned the set \( \{ \max Z''(k \Delta) \geq b \} \) into disjoint sets according to the last time that the process \( Z''(t) \) crosses the boundary \( b \). Therefore,

\[
\Pr(\max \sum_{k=0}^{L/\Delta} \int_{-1}^{\infty} \Pr(Z''(k \Delta) \in b + dx) \Pr(\max_{k \Delta < t < (k + 1) \Delta} Z''(k \Delta) < b | Z''(k \Delta) = b + x).)
\]
Ignoring the boundary effect due to \( L \) which is negligible when \( L/\Delta \) is large, we see that the integrand does not dependent on \( k \), so

\[
\Pr(\max_k Z''(k\Delta) \geq b) \sim \frac{L}{\Delta} \int_0^\infty f_{(n-1)}(b+x) \Pr(\max_{k>1} Z''(k\Delta) < b | Z''(0) = b+x) dx. \tag{4.5}
\]

Here \( f_n \) is the density function of

\[
\frac{t_n}{\sqrt{1+t_n^2/n}},
\]

where \( t_n \) is a t-statistic with \( n \) degree of freedom.

By developing approximations for the conditional probability and the marginal density in the above integral, we obtain the following theorem.

**Theorem 1.** Assume \( b/\sqrt{n} \to c \) and \( n\Delta \to d \) as \( n \to \infty \). Then as \( n \to \infty \),

\[
\Pr(\max_k Z''(k\Delta) \geq b) \sim f_{(n-1)}(b)(2\lambda Lb) \times \exp \left( -(2\pi)^{-1} \int_{-\infty}^{\infty} \left( \frac{1}{\alpha + i\theta} - \frac{1}{i\theta} \right) \left[ \log \left( \frac{1}{1-g(\theta)} \right) + \log(-i\mu\theta) \right] d\theta \right), \tag{4.6}
\]

where \( \alpha = c/(1-c^2) \), \( \mu = 2\lambda cd \), and \( g(\theta) = \exp \left( \lambda d(e^{-2\theta^2(1-c^2)+2i\theta c} - 1) \right) \).

**Proof:** To evaluate the tail probability, we first consider the marginal density

\[
f_{(n-1)}(b+x).
\]

Straightforward calculations show that

\[
f_{(n-1)}(b+x) \sim f_{(n-1)}(b)e^{-bx/(1-c^2)}. \tag{4.7}
\]

Substituting into (4.5), we get

\[
\Pr(\max_k Z''(k\Delta) \geq b) \sim \frac{L}{\Delta} f_{(n-1)}(b) \int_0^\infty e^{-bx/(1-c^2)} \Pr(\max_{k \geq 1} Z''(k\Delta) < b | Z(0) = b+x) dx \tag{4.8}
\]
To calculate the conditional probability, let

\[ X(t) = \sum_{i=1}^{n} g_i(t)(y_i - \mu). \]

Then

\[ Z''(t) = \frac{X(t)}{\hat{\sigma}/\sqrt{n}}. \]

Under the null hypothesis of no linkage, \((y_i - \mu)/\hat{\sigma}\) and \(g_i(t)\) are independent of \(\hat{\sigma}\), and so \(Z''(t)\) is independent of \(\hat{\sigma}\). Therefore,

\[
\Pr(\max_{k \geq 1} Z''(k\Delta) < b | Z''(0) = b + x) = \Pr(\max_{k \geq 1} X(k\Delta) < b\sqrt{n} | X(0) = (b + x)\sqrt{n}, \hat{\sigma} = 1) = \Pr(\max_{k \geq 1} (X(k\Delta) - X(0)) < -y | X(0) = b\sqrt{n} + y, \hat{\sigma} = 1),
\]

where \(y = \sqrt{n}x\).

\[
X(k\Delta) - X(0) = \sum_{i=1}^{n} (g_i(k\Delta) - g_i(0))(y_i - \mu) = -2 \sum_{i=1}^{n} g_i(0)(y_i - \mu)I_i(0, k\Delta),
\]

where

\[ I_i(j\Delta, j'\Delta) = I\{g_i(j\Delta) \neq g_i(j'\Delta)\}. \]

It’s easy to show that

\[
\Pr\{I_i(0, k\Delta) \neq I_i(0, \Delta) + I_i(\Delta, 2\Delta) + \cdots + I_i((k-1)\Delta, k\Delta)\} = o(\Delta).
\]

Let

\[ W_j = 2 \sum_{i=1}^{n} g_i(0)(y_i - \mu)I_i((j-1)\Delta, j\Delta). \]

Then under the condition of the theorem,

\[
\Pr\{X(k\Delta) - X(0) \neq -\sum_{j=1}^{k} W_j\} \to 0.
\]

Let \(T_j\) be independent copies of compound Poisson random variable \(\sum_{j=1}^{N} U_j\) where \(U_1, U_2, \ldots\) are i.i.d \(N(c, 1 - c^2)\) random variables and \(N\) is a Poisson random variable.
with parameter \( \lambda d \), and let \( S_j = \sum_{i=1}^j T_i \). From Lemma 1 below, one obtains that conditioning on \( X(0) = b\sqrt{n} + y \), and \( \hat{\sigma} = 1 \), \( \{W_1, W_2, \ldots, W_k\} \) converges in law to \( \{T_1, T_2, \ldots, T_k\} \) and therefore

\[
\Pr(\max_k Z''(k\Delta) \geq b) \\
= f_{(n-1)}(b) \frac{L}{\Delta\sqrt{n}} \int_0^\infty e^{-by/(1-c^2)\sqrt{n}} \Pr(\max_{k \geq 1} (-S_k) < -y) dy \\
= f_{(n-1)}(b) \frac{L}{\Delta\sqrt{n}} \int_0^\infty e^{-cy/(1-c^2)} \Pr(\min_{k \geq 1} S_k > y) dy \\
= f_{(n-1)}(b) \frac{L}{\Delta\sqrt{n}} \mu [\mathbb{E}(S_{r+})]^{-1} \int_0^\infty e^{-cy/(1-c^2)} \Pr(S_{r+} > y) dy \\
= f_{(n-1)}(b) 2\lambda L \sqrt{n} c(\alpha \mu)^{-1} \exp\left[-\sum_{i=1}^\infty \frac{1}{n_i} E e^{-\alpha S_i^+}\right]
\]

The last two equations follow from Siegmund (1985) problem 8.13 and Corollary 8.45. The proof of Theorem 1 is completed by applying a computational procedure given by Siegmund (1985) Theorem 8.51.

**Lemma 1**: \( \mathcal{L}(g_{i_1}(0)(Y_{i_1} - \mu), \ldots, g_{i_k}(0)(Y_{i_k} - \mu)|X(0) = b\sqrt{n} + y, \hat{\sigma} = 1) \) converges to the law of \( k \) independent, identically distributed \( N(c, 1 - c^2) \) random variables.

**Proof**: Let \( X_1, \ldots, X_n \) be i.i.d. \( N(0, 1) \), \( S_n = \sum_{i=1}^n X_i \) and \( U_n = \sum_{i=1}^n X_i^2 \). Then the joint density of \( X_1, \ldots, X_k, S_n \) and \( U_n \) is

\[
f_{X_1, \ldots, X_k, S_n, U_n}(x_1, \ldots, x_k, cn + y, n - 1) \\
= \phi(x_1)\phi(x_2)\cdots\phi(x_k) \cdot \frac{1}{\sqrt{n-k}} \phi\left(\frac{cn + y - (x_1 + \cdots + x_k)}{\sqrt{n-k}}\right) \\
\times \frac{1}{2^{n-k-1} \Gamma\left(\frac{n-k-1}{2}\right)} \left(n - 1 - \sum_{j=1}^k x_j^2 - \frac{(cn + y - (x_1 + \cdots + x_k))^2}{n-k}\right)^{\frac{n-k-1}{2}-1} \\
\times \exp\left(-\frac{1}{2[n - 1 - \sum_{j=1}^k x_j^2 - \frac{(cn + y - (x_1 + \cdots + x_k))^2}{n-k}]}\right) \\
= \left(\frac{1}{2\pi}\right)^{k/2} \frac{1}{\sqrt{2\pi(n-k)}} \frac{1}{2^{n-k-1} \Gamma\left(\frac{n-k-1}{2}\right)} \\
\times \left(n - 1 - \sum_{j=1}^k x_j^2 - \frac{(cn + y - (x_1 + \cdots + x_k))^2}{n-k}\right)^{\frac{n-k-1}{2}-1}.
\]
The joint density of $S_n, U_n$ is
\[
f_{S_n, U_n}(cn + y, n - 1) = \frac{1}{\sqrt{2\pi n}} \frac{1}{2^{(n-1)/2}\Gamma\left(\frac{n-1}{2}\right)} \left(n - 1 - \frac{(cn + y)^2}{n}\right)^{\frac{n-1}{2}-1},
\]
so the conditional distribution of $X_1, X_2, \ldots, X_k$ given $S_n = cn + y$ and $U_n = n - 1$ is
\[
f_{X_1, X_2, \ldots, X_k}(x_1, x_2, \ldots, x_k | S_n = cn + y, U_n = n - 1) \sim \left(\frac{1}{2\pi}\right)^{k/2} \frac{2^{k/2}\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-k-1}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)} \exp\left(-\frac{1}{2(1-c^2)}[1 + \sum_{i=1}^{k} x_i^2 + c^2 k + 2cy - 2c(x_1 + x_2 + \cdots + x_k)]\right) \exp\left(-\frac{1}{2(1-c^2)}[1 + 2cy]\right)
\]
\[
= \left(\frac{1}{2\pi}\right)^{k/2} \frac{2^{k/2}\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-k-1}{2}\right)}{\Gamma\left(\frac{n-k-1}{2}\right)} \frac{(1 - c^2)^{k/2}}{\Gamma\left(\frac{n-k-1}{2}\right)} \exp\left(-\frac{1}{2(1-c^2)} \sum_{i=1}^{k} (x_i - c)^2\right).
\]

This completes the proof of the lemma.

Now we go back to case that $\mu$ is unknown. The least square estimator of $\alpha$ becomes
\[
\hat{\alpha} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(g_i(t) - \bar{g}(t))}{\sum_{i=1}^{n} (g_i(t) - \bar{g}(t))^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(g_i(t) - \bar{g}(t))}{n(1 - \bar{g}(t)^2)}.
\]

And the test statistic becomes
\[
Z''(t) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})g_i(t)}{\hat{\sigma} \sqrt{n(1 - \bar{g}(t)^2)}}
\]
where
\[
\hat{\sigma}^2 = \frac{1}{n - 2} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

**Theorem 1'.** The result of Theorem 1 is still true with $f_{(n-1)}$ replaced by $f_{(n-2)}$.

**Proof:** It's easy to show the marginal distribution of $Z''(t)$ is the same as
\[
\frac{t_{(n-2)}}{\sqrt{1 + t_{(n-2)}^2/(n - 2)}}.
\]
And by the similar argument as in Theorem 1 that

\[ \Pr(\max_k Z''(k\Delta) \geq b) \]

\[ \approx f_{(n-2)}(b) \frac{L}{\Delta \sqrt{n}} \]

\[ \times \int_0^\infty e^{-\alpha y} \Pr(\max_{k \geq 1} (\sqrt{n}Z''(k\Delta) - \sqrt{n}Z''(0)) < -y|\sqrt{n}Z''(0) = nc + y) dy. \]

Let

\[ Z^*(k\Delta) = \frac{\sum_{i=1}^n (y_i - \bar{y})g_i(k\Delta)}{\hat{\sigma} \sqrt{n(1 - \hat{y}(0))^2}}. \]

Then

\[ Z^*(k\Delta) - Z''(k\Delta) = Z''(k\Delta) \left( \left( \frac{1 - \hat{g}^2(k\Delta)}{1 - \hat{g}^2(0)} \right)^{1/2} - 1 \right) \]

\[ \approx Z''(k\Delta)(1/2)(-\hat{g}^2(k\Delta) + \hat{g}^2(0)) \]

\[ = Z''(k\Delta)(\hat{g}(0) - \hat{g}(k\Delta))(\hat{g}(0) + \hat{g}(k\Delta))/2 \]

Since

\[ \hat{g}(0) - \hat{g}(k\Delta) = O_p \left( \frac{1}{n} \right), \]

\[ \hat{g}(0), \hat{g}(k\Delta) = O_p \left( \frac{1}{\sqrt{n}} \right), \]

\[ Z''(k\Delta) \sim Z''(0) \sim c \sqrt{n}, \]

we have

\[ Z^*(k\Delta) - Z''(k\Delta) = O_p \left( \frac{1}{n} \right). \]

Therefore

\[ \Pr(\max(\sqrt{n}Z''(k\Delta) - \sqrt{n}Z''(0)) < -y|\sqrt{n}Z''(0) = nc + y) \]

\[ \approx \Pr(\max(\sqrt{n}Z^*(k\Delta) - \sqrt{n}Z''(0)) < -y|\sqrt{n}Z''(0) = nc + y). \]

\[ \sqrt{n}Z^*(k\Delta) - \sqrt{n}Z''(0) = \frac{\sum_{i=1}^n (y_i - \bar{y})(g_i(k\Delta) - g_i(0))}{\hat{\sigma} \sqrt{1 - \hat{y}^2(0)}} \]

\[ = -\sum_{j=1}^k W_j \]
where
\[
W_j = 2 \sum g_i(0) (y_i - \bar{y}) I_i((j - 1)\Delta, j\Delta) / \bar{\sigma} \sqrt{1 - \bar{g}^2(0)}
\]

Then Theorem 1’ follows by applying Lemma 1’ and essentially the same arguments in Theorem 1.

**Lemma 1’**. \( \mathcal{L}(\sum g_i(0)(Y_i - \bar{Y})) / (\bar{\sigma} \sqrt{1 - \bar{g}^2(0)}) \), \( \cdots, (g_k(0)(Y_k - \bar{Y})) / (\bar{\sigma} \sqrt{1 - \bar{g}^2(0)}) \mid \sqrt{n}Z(0) = nc + y \) converges to the law of \( k \) independent, identically distributed \( N(c, 1 - c^2) \) random variables.

**Proof**: Basu’s Theorem (Lehmann (1986), Theorem 5.2) tells us that under the null hypothesis \( (Y_i - \bar{Y}) / \bar{\sigma} \) is independent of the complete sufficient statistics \( \bar{Y} \) and \( \bar{\sigma} \). Therefore,
\[
\mathcal{L} \left( \frac{g_i(0)(Y_i - \bar{Y})}{\bar{\sigma} \sqrt{1 - \bar{g}^2(0)}}, \cdots, \frac{g_k(0)(Y_k - \bar{Y})}{\bar{\sigma} \sqrt{1 - \bar{g}^2(0)}} \mid \sum g_i(0)Y_i / \sqrt{1 - \bar{g}^2(0)} = nc + y, \bar{Y} = 0, \bar{\sigma} = 1 \right)
\]

Since
\[
\bar{g}(0) = O_p\left(\frac{1}{\sqrt{n}}\right),
\]
it’s enough to show
\[
\mathcal{L}(g_i(0)Y_i, \cdots, g_k(0)Y_k \mid \sum g_i(0)Y_i = nc + x, \bar{Y} = 0, \bar{\sigma} = 1) \Rightarrow \mathcal{L}(k \text{ i.i.d } N(c, 1 - c^2)).
\]

Without loss of generality, we may assume \( l_i = i \) for \( i = 1, \cdots, k \) and \( Y_1, Y_2, \cdots, Y_n \) are i.i.d \( N(0,1) \). Let \( X_i = g_i(0)Y_i \), then \( X_1, \cdots, X_n \) are i.i.d \( N(0,1) \) also. For fixed \( g \),
\[
\mathcal{L}(g_i(0)Y_i, \cdots, g_k(0)Y_k \mid \sum g_i(0)Y_i = nc + x, \bar{Y} = 0, \bar{\sigma} = 1) \Rightarrow \mathcal{L}(X_1, \cdots, X_k \mid S_n^- = \frac{nc + x}{2}, S_n^+ = \frac{nc + x}{2}, U_n = n - 1),
\]
where
\[
S_n^+ = \sum_{i=1} g_i X_i,
\]

\[
S_n^- = \sum_{i=1} g_i X_i.
\]
\[ S_n^- = \sum_{g_i = -1} X_i, \]
\[ U_n = \sum_{i=1}^n X_i^2. \]

Let
\[ n^+ = \#\{g_i = 1, \ i \in \{1, \ldots, n\}\}, \quad n^- = \#\{g_i = -1, \ i \in \{1, \ldots, n\}\}, \]
\[ k^+ = \#\{g_i = 1, \ i \in \{1, \ldots, k\}\}, \quad n^- = \#\{g_i = -1, \ i \in \{1, \ldots, k\}\}, \]
\[ x_i^+ = x_i I_{g_i = 1}, \quad x_i^- = x_i I_{g_i = -1}. \]

Straightforward calculations show that
\[ f_{x_1, \ldots, x_k}(x_1, \ldots, x_k | S_n^+ = \frac{nc + x}{2}, S_n^- = \frac{nc + x}{2}, U_n = n - 1) \]
\[ = \left(\frac{1}{2\pi}\right)^{k/2} \sqrt{\frac{n^+}{n^+ - k^+}} \sqrt{\frac{n^-}{n^- - k^-}} 2^{k/2} \Gamma((n - 2)/2) \]
\[ \times \left( n - 1 - \frac{(nc + x)/2}{n^+} - \frac{(nc + x)/2}{n^-} \right)^{-k/2} \left(\frac{A}{B}\right)^{(n-4)/2}, \quad (4.13) \]

where
\[ A = n - 1 - \sum_{j=1}^k x_j^2 - \frac{(cn + x)/2 - (x_1^+ + \cdots + x_k^+))^2}{n^+ - k^+} \]
\[ - \frac{(cn + x)/2 - (x_1^- + \cdots + x_k^-))^2}{n^- - k^-}, \]
\[ B = n - 1 - \frac{(cn + x)/2}{n^+} - \frac{(cn + x)/2}{n^-}. \]

Obviously
\[ \sqrt{\frac{n^+}{n^+ - k^+}} \sim 1, \]
\[ \sqrt{\frac{n^-}{n^- - k^-}} \sim 1, \]
\[ \left( n - 1 - \frac{(nc + x)/2}{n^+} - \frac{(nc + x)/2}{n^-} \right)^{-k/2} \sim (n(1 - c^2))^{-k/2}, \]
\[ \left(\frac{A}{B}\right)^{(n-4)/2} \sim \exp\left(\frac{A - B}{2B/n}\right). \]
And

\[ A - B \sim - \sum_{j=1}^{k} (x_j - c)^2 \]

\[ \frac{B}{n} \sim 1 - c^2 \]

Substitution into (4.13) yields the lemma.

It's easy to show that

\[ Z = \left( n \log \frac{n - 2}{n - 2 - Z''} \right)^{1/2}. \]

So we have

\[ \Pr\{\max_k Z(k\Delta) \geq b\} = \Pr\{|Z''| \geq ((n - 2)(1 - \exp(-b^2/n)))^{1/2}\}. \]

Then plugging in the result of Theorem 1', we can get a proximation for

\[ \Pr\{\max_k Z(k\Delta) \geq b\}. \]

Remark: We think the above argument can also be applied to the intercross data. However, the computations are more complicated, and need to be studied carefully in the future.

### 4.3.3 Numerical examples

We calculated some numerical examples to see the discrepancy between the normal approximation and ours. We assume that \( \mu \) is unknown. We calculated the p-values of one chromosome for two sample sizes \( n = 100 \) and \( n = 25 \), \( b \) is fixed at 4. The length of the chromosome is taken to be 100 centimorgans. The results are given in Table 4.1 and Table 4.2. The correction term corresponds to the \( \nu \) function in the normal approximation and the exponential term in the our approximation.
<table>
<thead>
<tr>
<th>Δ</th>
<th>our approx</th>
<th>normal approx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>correction</td>
</tr>
<tr>
<td>10 cM</td>
<td>$0.657 \times 10^{-3}$</td>
<td>0.230</td>
</tr>
<tr>
<td>1 cM</td>
<td>$1.535 \times 10^{-3}$</td>
<td>0.537</td>
</tr>
<tr>
<td>0.1 cM</td>
<td>$1.757 \times 10^{-3}$</td>
<td>0.615</td>
</tr>
<tr>
<td>0.01 cM</td>
<td>$1.781 \times 10^{-3}$</td>
<td>0.624</td>
</tr>
</tbody>
</table>

Table 4.1: Approximation for n=100 and b=4

<table>
<thead>
<tr>
<th>Δ</th>
<th>our approx</th>
<th>normal approx</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value</td>
<td>correction</td>
</tr>
<tr>
<td>10 cM</td>
<td>$1.614 \times 10^{-3}$</td>
<td>0.194</td>
</tr>
<tr>
<td>1 cM</td>
<td>$2.887 \times 10^{-3}$</td>
<td>0.347</td>
</tr>
<tr>
<td>0.1 cM</td>
<td>$3.090 \times 10^{-3}$</td>
<td>0.371</td>
</tr>
<tr>
<td>0.01 cM</td>
<td>$3.112 \times 10^{-3}$</td>
<td>0.374</td>
</tr>
</tbody>
</table>

Table 4.2: Approximation for n=25 and b=4
Bibliography


BIBLIOGRAPHY


