STATISTICAL PROBLEMS ASSOCIATED WITH
MAPPING COMPLEX AND QUANTITATIVE TRAITS
FROM GENOMIC MISMATCH SCANNING DATA

by

Josée Dupuis

TECHNICAL REPORT NO. 2
September 1994

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 Problems Associated with 
Mapping Complex and Quantitative Traits 
From Genomic Mismatch Scanning Data 

by 

Josée Dupuis 

TECHNICAL REPORT NO. 2 
September 1994 

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

DEPARTMENT OF STATISTICS 
STANFORD UNIVERSITY 
STANFORD, CALIFORNIA
Chapter 1

Introduction

Recent advances in genetics have led to the discovery of genes responsible for certain diseases such as cystic fibrosis, Huntington disease and more. The localization of disease-susceptibility genes can help improve diagnosis and, eventually, treatment for the disease. Much effort has been put into genetic research, the human genome project being a prime example, and is giving rise to a wealth of new statistical problems.

What motivated this thesis is the increasing availability of dense sets of genetic markers and a new method for determining continuous segments of Identity-by-Descent between relatives called Genomic Mismatch Scanning (GMS, Nelson et al. (1993)). GMS is currently being developed and, as of this writing, has yet to be applied to humans, although preliminary results on yeast make it look promising. On the other hand, dense sets of markers are readily available and the number of genetic markers that can be used for genetic research is increasing rapidly. This thesis extends work by Feingold (1993) and Feingold, Brown and Siegmund (1993) to more complex diseases. It also extends the work of Lander and Botstein (1989) for finding quantitative trait loci in experimental organisms.

A brief introduction to genetic linkage follows. It will be accompanied by a description of recent work in the area of genetic linkage and an outline of the contents
of this thesis.

1.1 Background

Humans have twenty-three pairs of chromosomes, forming the genome. Other organisms of interest to us (such as the tomato or the mouse) also have pairs of chromosomes. Genes are thought of as being located along chromosomes, the words locus and gene often being used interchangeably to mean the location on the chromosome where a gene lies. A gene can take different forms which are called alleles. Humans have two alleles for each gene, one for each chromosome in the pair. The specific alleles that an organism has for a certain gene is called the genotype. The genotype of an individual cannot be observed directly, but can sometimes be determined through a laboratory experiment. The outward manifestation of a specific genotype, such disease/non-disease, eye colors or fruit size, is called the phenotype.

![Diagram](image)

Figure 1.1: Mating between two organisms, producing one offspring.

At reproduction, an embryo is formed from two gametes, one from each parent. A gamete is a cell such as an egg or a sperm that contains a single copy of the chromosomes. This implies that an offspring receives one copy of its chromosomes from its mother and one from its father. Gametes are formed during a process called
meiosis. At meiosis, chromosome pairs are split in half, producing cells (gametes) with half the chromosomes of the original organism. The splitting of chromosomes occurs in a special way. The gamete does not contain one whole chromosome from the pair, but rather a mix of the two chromosomes in the pair. The mixing occurs in the following way (see Figure 1). The splitting process starts with either chromosome with probability 1/2, goes a random distance along this chromosome and then switches to the other chromosome. The process goes another random distance along the new chromosome before the next switch and so on until the end of the chromosome is reached. The switches are called crossovers. The random distance is often modeled as having an exponential distribution (called the Haldane mapping function). Therefore, alleles that are located close together on a chromosome will tend to be grouped in the same gamete. Genes that tend to be inherited together are said to be linked. Two alleles on a chromosome will be passed on together if and only if an even number of crossovers occurs between them. If an odd number of crossovers occurs, the alleles are said to have recombined. The genetic distance between two genes is defined to be the expected number of crossovers that occurs at meiosis between the two genes on a single strain of the chromosome. The genetic distance unit is the Morgan (or centimorgan). So one Morgan (or 100 centimorgans) is equivalent to a genetic distance where, on average, one crossover occurs at meiosis.

The Haldane map function assumes that the crossovers are independent, an assumption that seems to be violated in most organisms, including humans. The lack of independence of the crossovers is called interference. Other map functions that take into account varying degrees of interference can be found in Ott (1991 p. 14); of necessity they are more complicated than the Haldane map function.

When we mention that a gene increases the susceptibility to a disease, it means that the presence of an allele or of a pair of alleles at a specific locus increases the odds of having/developing the disease. The identification of such genes is often done
CHAPTER 1. INTRODUCTION

through linkage analysis.

To perform a linkage analysis study, families where the disease is present in more than one individual are collected. Then a marker is selected. A marker is a gene or a portion of DNA that can be observed (i.e. one in which we can determine which alleles a certain organism has), and that has a known location on the genome. The marker is examined to determine if it is passed along with the disease more often than would be expected by chance. If so, the disease locus is probably linked to the marker. If not, other markers can be selected and the process is repeated.

An alternative to the classical way of doing linkage analysis, is the use of pairs of affected relatives (Risch's (1990b)). A definite advantage of this latter method is that only affected individuals need to be studied, and they are often eager to cooperate. The rationale behind the affected relative pair method is that if there exists a gene that influences the disease, it will most probably be found in the region of the genome of the relatives that comes from a common ancestor (the region that is identical-by-descent). With this method, a marker is selected and examined to see if the pairs of affected relatives share alleles at the marker more often than would be expected by chance. If so, we expect the disease locus to be linked to the marker.

Dense sets of markers covering the genome are becoming more available (Botstein et al. (1980)). The construction of dense maps of markers has prompted some authors (Lander and Botstein (1989), Feingold (1993), Feingold, Brown and Siegmund (1993)) to look at sets of genetic markers together instead of one at a time in order to increase the power of detecting genetic linkage. The next section gives a brief description of the work of the cited authors.
1.2 Review

Feingold (1993) and Feingold, Brown and Siegmund (1993) developed statistical methods to find disease susceptibility genes using continuous maps of identity-by-descent between pairs of affected relatives. They used data coming from Genomic Mismatch Scanning, a method to determine whether two relatives share alleles from a common ancestor at any location on their genome. They define tests to detect genetic linkage at any point on the genome, and give p-values and approximations to the power of these tests. They develop their method for many different types of relatives: siblings, half-siblings, cousins, grandparent-grandchild and avuncular (uncle-nephew) pairs. They also provide methods of combining different pairs of relatives, and extend their method to sibling trios. A more detailed description of their work will be provided in the chapter 3.

Other authors that have studied methods of using a dense set of markers to do genetic linkage include Lander and Botstein (1989). Their goal was to identify quantitative trait loci in experimental organisms. By a quantitative trait they mean any observable characteristic that is measured on a continuous scale such as the size of a fruit, the height of a plant etc. They use an analysis of variance model to explain the variation in the quantitative trait (phenotype), as a function of the alleles (genotype) an individual has at a large number of markers. Their method has already been used for many purposes. Examples include the identification of genes that influence mass, pH and soluble solid concentration in tomatoes (Patterson et al. (1991)) and genes that increase blood pressure in rats (Jacob et al. (1991)). The last chapters of this thesis will be devoted to the presentation and extension of their methods to more general models.
CHAPTER 1. INTRODUCTION

1.3 Outline

Chapter 2 has methods to analyze the output from Genomic Mismatch Scanning (GMS) data. GMS is a method designed to identify continuous regions of Identity-by-Descent (IBD) between a pair of relatives. The analysis of GMS data can be formulated as a multiple change-point problem. Likelihood methods and a modified version of an algorithm suggested by Churchill (1989) are shown to work quite well with the yeast data of Nelson, et al. (1993).

Chapter 3 is a description of how to use the continuous map of IBD between affected relatives to find the location of genes that increase susceptibility to a specific disease. Feingold (1993) and Feingold, Brown and Siegmund (1993) provide a thorough development of the monogenic case that will be summarized in this chapter.

Chapter 4 contains extensions of the previous authors' work to more complex diseases such as the case of polygeny i.e., diseases that are influenced by more than one gene. We define tests for the complex case situations and provide approximations to the p-values and power of these tests. Simulations are used to show that the approximations are quite accurate. We deal with both the case where the continuous map of IBD is available and the case where the data come from a discrete set of markers. The powers from using a map of markers one centimorgan apart and a continuous map are equivalent. However, a slight decrease in power can be observed from using markers 5 centimorgan apart when the loci are located between markers. When the loci are located at markers, sparser maps perform better because they require lower thresholds than the continuous map.

A method of interpolating between markers called interval mapping is presented in Chapter 5. Interval mapping was introduced by Lander and Botstein (1986) in the context of a set of markers spanning the whole genome. We show through a simulation study that the gain in power from using interval mapping is not substantial.
CHAPTER 1. INTRODUCTION

In chapter 6, we describe methods of constructing confidence intervals for the location of a disease susceptibility gene once it has been established that there is such a gene. Lod-support intervals, likelihood confidence intervals and Bayesian credibility sets can all be used to make confidence statements about the location of a gene for data coming from GMS. When the data are from a discrete set of markers, Bayesian credibility methods can be combined with Interval Mapping to make confidence statements. A simulation study comparing all three methods of making confidence statement found the following results. When considered as a confidence interval, Bayes credible sets have the correct coverage probability. The 1.5-lod support interval for dense maps and a 1.0-lod support interval for sparse maps (markers 20 centimorgans apart) that were used, correspond approximately to 95% confidence intervals. There is not much difference between the Bayes credible set and the lod support interval in terms of size of the simulated confidence interval and the probability of including the true locus. However, as one expects, the Bayes credible method performs better than the lod-support when the prior is appropriate to the data and not so well when the prior is not appropriate. The likelihood method gives wider confidence intervals than both the lod-method and the Bayes credible sets with a good prior.

A problem mathematically similar to that of analyzing affected relative pairs through IBD information is the identification of quantitative trait loci in certain organisms. Lander and Botstein (1989) gave a model that implicitly assumes no epistasis (i.e. no interaction amongst loci) to identify genes that influence quantitative traits using a dense set of markers. Their analysis ignores dominance effects although they mention that the model could be extended to include such genetic effects. Patterson et al. (1991) applied the method to the tomato genome and found minimal evidence of epistasis but some dominance component to the phenotypic variance. This motivates taking a careful look at a more general model that also ignores epistasis but that includes both additive and dominance components. In Chapter 7, we describe
the model of Lander and Botstein (1989) for finding quantitative trait loci, and we propose a statistic to test for the presence of quantitative trait loci in the presence of both additive and dominance effects.

Chapter 8 contains a method to construct a joint confidence region for the quantitative locus and its additive and dominance effects. The confidence region is based on likelihood methods for change-points and has the correct coverage probability when the number of organisms available is large. However, a study simulating the tomato genome shows that as few as 100 tomatoes are enough for the procedure to include the true parameters with the accurate probability. Finally, we show how to use the likelihood confidence set, the Bayes credible sets and the lod support intervals defined in Chapter 6 to make confidence statements about the quantitative trait locus alone.
Chapter 2

Analysis of Genomic Mismatch Scanning Data

2.1 Introduction

In order to establish linkage through the affected relative pair method, the region of genetic identity-by-descent (IBD) between the two related individuals must be identified. If there exists a gene responsible for the disease, this gene will most likely be found in the region that is common to both affected relatives. A new method ideally suited to identifying such IBD regions is Genomic Mismatch Scanning (GMS, Nelson et al. 1993). GMS looks for mismatches at the DNA level and does not require conventional polymorphic markers. The output of GMS is a hybridization signal, consisting of a sequence of intensity readings along a chromosome or along the entire genome. High intensities indicate a segment common to both relatives. Accurate identification of regions of high intensity would give a complete map of IBD between the relatives. This can be accomplished by finding the change-points i.e., the locations where the intensity goes from low to high or from high to low.

To demonstrate the feasibility of GMS, Nelson et al. apply their method to yeast,
which offers certain advantages as a model organism. Their data can be described as follows. Two strains of yeast, Y55 and Y24, were mated and produced a tetrad of four spores (offspring). A tetrad (Suzuki et al., 1989, p. 125) is the four products from a single mating between the two yeast strains. Offspring of the same tetrad have the property that at any given location on their chromosomes, two offspring received their genetic material from the Y24 strain while the other two received it from the Y55 strain. The offspring will be identified as A, B, C, and D. The original yeast strains Y24 and Y55 and the four offspring are all haploid i.e., have only one copy of each chromosome instead of having pairs of chromosomes as is the case for humans. See Figure 2.1. The eight possible offspring-parent pairs will be identified by listing the parent followed by the offspring letter, such as Y55:A.

Figure 2.1: Mating between two strains of yeast, Y24 and Y55, producing four spores (offspring). The yeast stains and all the offspring are haploid.

GMS was performed on chromosome 5 for all eight parent-offspring pairs in order to identify the genetic material of each offspring inherited from each parent. The output from GMS for each pair is 31 intensity readings, each representing a small segment, called a clone, of chromosome 5. Chromosome 5 of the yeast is approximately 210 centimorgans (2.1 Morgans) long. The genetic distance (in Morgans) between two
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

points on a chromosome is defined to be expected number of crossovers between the
two points (1 Morgan = 100 centimorgans). So by this definition, we expect an average
of 2.1 crossovers to have occurred during meiosis for the yeast data i.e., we expect
about two change-points to occur along the 31 data points for each pair. Correctly
finding the location of the changes would tell us exactly which parts of chromosome
5 each offspring inherited from each parent.

Because we can observe all elements of a tetrad, this yeast experiment provides
information not available on human relative pairs. First, for any given offspring, say
A, the GMS output from Y24:A is complementary to Y55:A, since any segment of
chromosome 5 of A is IBD to either Y24 or Y55. Second, since all four offspring
come from the same tetrad or meiosis, any crossover that occurs in one offspring has
a dual change in one of the other offspring. So, any change-point in the hybridization
intensity for Y24:A say, has a change at exactly the same segment in either Y24:B,
Y24:C or Y24:D. This is only true for offspring of the same tetrad.

This complementarity information will not be used in the analysis of the data since
it will not be available in situations involving other organisms, especially humans. We
do, however, use the extra information to assist in evaluating the different statistical
procedures used to determine the IBD pattern of each pair.

We can express the mathematical problem in the following way. Let \( y_1, y_2, \ldots, y_n \)
be unobserved variables that take on values 0 or 1 with probability \( 1 - p \) and \( p \)
respectively. We say that the process is in state 1 at time \( i \) (or IBD on segment \( i \))
if \( y_i = 1 \). The \( y_i \)'s are not independent since changes (crossovers) occur rarely and
the process at time \( i \) is more likely to remain in the same state it was at time \( i - 1 \)
than to change states. The parameter \( p \) is the probability that the relatives are IBD
at any given point on their genome and depends on the relation between the affected
individuals only. For example, \( p = 1/2 \) for the yeast data and for human half sibling
and aunt-niece pairs, 1/4 for first cousins, etc.
We observe intensity readings $x_1, x_2, \ldots, x_n$ where the distribution function of $x_i$ depends on the value of $y_i$, i.e.,

$$P(x_i \leq x | y_i = 0) = F_0(x)$$
$$P(x_i \leq x | y_i = 1) = F_1(x)$$

We want to recover the $y$ process, although we only get to observe the $x_i$'s. Correctly identifying the points where the $y$ process switches states and $y_1$, the state at time 1, would suffice. Equivalently, we could identify the points where the $x$'s go from having distribution $F_0 (F_1)$ to $F_1 (F_0)$, which is a classical change-point problem.

There is an extensive literature on change-point problems. Tests for a single change in mean were described by many authors (Sen and Srivastava (1975), Worsley (1979), James, James and Siegmund (1987)). A test for two changes in mean was also presented in Siegmund (1986). Multiple change-point methods include the binary segmentation procedure (Vostrikova, 1981) and the use of Schwarz' Bayesian information criterion introduced by Yao (1988). The latter method is fairly computer intensive and was not applied to the yeast data. Venkatraman (1992) showed through a simulation study that the binary segmentation often fails to detect any change and tends to underestimate the number of changes, especially in the cases where the changes alternate in signs, which is the case for the yeast data. Both of these methods have been studied for detecting changes in mean only.

Two issues complicate the analyses of the yeast data and keep us from using directly the methods developed in the literature. First, the number of changes is unknown, although prior information on the distribution of the number of such changes (which are equivalent to crossovers) is available. Second, whereas most of these methods look for a change in mean only, in the GMS data the variance changes with the mean. In this paper, yeast data obtained from GMS are analyzed using (i) likelihood methods and (ii) an appropriately modified algorithm proposed by Churchill (1989)
that attempts to deal with both of the above difficulties.

Section 2.2 contains some likelihood methods for locating change points and presents their application to the yeast data. Section 2.3 implements the Churchill algorithm for our particular case, with illustrations on the yeast data.

2.2 Likelihood Ratio Analysis

In this section, we will define likelihood based tests to find the location of change points based on the observed \( z \)'s, the intensity readings. To simplify matters, we assume \( F_0 \) and \( F_1 \) to be normal distributions with equal variances. Recall that \( F_0 \) and \( F_1 \) are the distributions of the intensity readings corresponding to the non-IBD and IBD segments, respectively. The variance assumption can be relaxed with some modifications. QQ-plots of the data (not shown) indicate that the assumption of normality is reasonable. The equal variance model is probably not correct but still gives interesting results.

Ideally, we would like to compare the likelihood of no change to the likelihood of "some change," without being more precise as to the number or location of these changes. Unfortunately, maximizing the above likelihood ratio over the nuisance parameters (mean and variance of \( F_0 \) and \( F_1 \) and the location of the changes) is difficult. We therefore need to make the alternative hypothesis more specific and hope that it will give us some insight into more general alternative hypotheses.

We first consider the alternative with a single change occurring in the data. In mathematical terms, let \( x_i, (i = 1, \ldots, m) \) be a sequence of normal random variables with mean \( \mu^{(i)} \) and variance \( \sigma^2 \). The null hypothesis \( H_0 \) and the alternative \( H_1 \) are

\[
H_0 : \mu^{(1)} = \ldots = \mu^{(m)} \\
H_1 : \exists 1 \leq \nu < m \text{ such that } \mu_0 = \mu^{(1)} = \ldots = \mu^{(\nu)} \neq \mu^{(\nu+1)} = \ldots = \mu^{(m)} = \mu_1
\]
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

If $\nu$ and $\sigma$ are known, the log likelihood ratio statistic can be expressed as (James, James and Siegmund 1987)

$$\frac{(\frac{\nu}{m} S_m - S_\nu)^2}{\nu(1 - \frac{\nu}{m})\sigma^2},$$

(2.1)

where $S_n = \sum_{i=1}^{n} x_i$.

When $\nu$ is unknown, maximizing over all possible values of $\nu$ gives the log likelihood ratio statistic

$$\max_k \frac{(\frac{k}{m} S_m - S_k)^2}{k(1 - \frac{k}{m})\sigma^2}.$$ 

(2.2)

The value $k$ which maximizes the above expression is the Maximum Likelihood Estimate (MLE) of $\nu$. When $\sigma^2$ is unknown, we can substitute its MLE under $H_0$, $\hat{\sigma}^2 = m^{-1} \sum_{i=1}^{m} (x_i - \bar{x}_m)^2$. The significance of the new statistic can be assessed through the evaluation of a numerical integral, as described in James, James and Siegmund (1987). For example, when $m = 31$, values of the expression in (2.2) greater than 8.12 are statistically significant at the 5% level.

In place of the log likelihood ratio statistic, Pettitt (1980) suggested using

$$\max_k \left| \frac{k}{m} S_m - S_k \right|$$

(2.3)

This ad-hoc statistic has more power than (2.2) to detect changes in the middle of the data set and less power to detect change at the endpoints since the denominator $|k(1 - k/m)|$ has its maximum value at $k = m/2$. For unknown $\sigma$, we can use $\hat{\sigma}^{-1}(\frac{k}{m} S_m - S_k)$, where $\hat{\sigma}^2$ was defined previously. A value of 6.91 (for $m = 31$) of this modified statistic would be declared statistically significant at the 5% level (James, James and Siegmund 1987).

The expression (2.1) contains more information about possible change-points than is summarized by taking the maximum as in (2.2) and (2.3). A way of extracting this information for the purpose of studying the more general hypothesis of "some change" would be to plot $\hat{\sigma}^{-1}(\frac{k}{m} S_m - S_k)$ for all values of $k$ (with or without the
scaling factor of $k(1 - \frac{k}{m})$. Local maxima indicate changes from $F_0$ to $F_1$, if we assume that mean($F_0$) < mean($F_1$) and vice versa for local minima. Significant peaks and valleys in the plot suggest a classification of each point $x_i$ as coming from either $F_0$ or $F_1$. Note that the direction of the change would not be specified by a plot of $\tilde{\sigma}^{-2}(\frac{k}{m}S_m - S_k)^2$, where maxima represent change-points.

In Figure 2.2, we present plots of the intensity readings and of $\tilde{\sigma}^{-1}(\frac{k}{m}S_m - S_k)$ for all Y55-offspring pairs. The x-axis represents segments from chromosome 5 ordered from one end of the chromosome to the other. We used the plot without the scaling factor for reasons that will become apparent later. For pair Y55:A, further biological testing suggests that two crossovers occurred, one before clone 10 and the other one before clone 28. At $k = 9$, the value of $\tilde{\sigma}^{-1}(\frac{k}{m}S_m - S_k)$ is -7.33 (p-value = 0.031) and the value of the log likelihood ratio statistics, divided by $\tilde{\sigma}^2$ to account for the unknown variance, is 8.41 (p-value = 0.041). The graph also shows some indication of another change occurring at $k = 27$, although the peak is not sufficiently large to reach statistical significance.
Figure 2.2: Intensity readings and plot of $\hat{\sigma}^{-1}(\frac{k}{m}S_m - S_k)$ for all Y55 pairs, for chromosomes 5. The x-axis represents the 31 clones, ordered from one end of the chromosome to the other. Plots (a), (b), (c) and (d) represent the hybridization intensities for pairs Y55:A, Y55:B, Y55:C and Y55:D respectively. Plots (e), (f), (g) and (h) are the plots of $\hat{\sigma}^{-1}(\frac{k}{m}S_m - S_k)$. Further biological testing indicated that changes occurred during meiosis at clones 9 and 27 for pair Y55:A, clones 7 and 12 for pair Y55:B, clone 27 for pair Y55:C and clones 2 and 13 for pair Y55:D.

Since we expect on average two changes in the yeast data, we should also consider the two change alternative i.e.,

$H_0 : \mu^{(1)} = ... = \mu^{(m)}$

$H_1 : \exists 1 \leq \nu_1 < \nu_2 \leq m$ such that $\mu^{(1)} = ... = \mu^{(\nu_1)} = \mu_0$, 

$\mu^{(\nu_1+1)} = ... = \mu^{(\nu_2)} = \mu_1$, 

$\mu^{(\nu_2+1)} = ... = \mu^{(m)} = \mu_0$

The one change alternative is a special case. Then the log likelihood ratio statistic can be shown to be (Siegmund 1986)

$$\max_{i<j} \left( \frac{(S_j - S_i - \frac{i-i}{m}S_m)^2}{(j-i)(1-i/m)\sigma^2} \right)$$

(2.4)
Note that the maximum of the numerator only
\[ \max_{i<j} [S_j - S_i - \frac{j-i}{m} S_m] = \max_j [S_j - \frac{j}{m} S_m] - \min_i [S_i - \frac{i}{m} S_m] \]
is just the maximum minus the minimum of the statistic suggested by Pettitt. The plot of \( \tilde{\sigma}^{-1}(\frac{k}{m} S_m - S_k) \) can be used to visualize either the test for one change or the test for two changes in the data. The value of (2.4) for pair Y55:A is 17.56 which has an approximate p-value of 0.005 (See Figure 2.2). If we are using the correct test i.e., the test for two changes when in fact there are two changes, we apparently would have identified the complete map of IBD correctly.

From the above discussion it seems like the plot of \( \tilde{\sigma}^{-1}(\frac{k}{m} S_m - S_k) \) may be a very useful tool for locating multiple changes. However, there still remains the problem of the naive assumption of equal variance. If we relax this assumption, the likelihood ratio statistic for one change is
\[ \max_k \frac{\tilde{\sigma}_{0,m}^2}{\tilde{\sigma}_{0,k}^2 \tilde{\sigma}_{k,m}^2}, \tag{2.5} \]
where \( \tilde{\sigma}_{ij}^2 = (j-i)^{-1} \sum_{n=i+1}^{j} \left( x_n - \frac{S_j - S_i}{j-i} \right)^2. \)

Note that the quantity maximized in (2.5) is the statistic to test for equality of the mean and the variance of the first \( k \) and the last \( m-k \) observations, under the assumption of normality. A similar likelihood ratio statistic can be derived for the alternative of two changes occurring at \( i \) and \( j, \ i < j \). It would be equivalent to testing whether the middle observations \( x_{i+1} \) through \( x_j \) have the same mean and variance as the remaining observations. The statistic for one change (2.5) can be plotted against \( k \) (not shown) to detect a change in variance as well as in mean, but it does not offer some of the advantages of the statistic for equal variances. First, since only maxima indicate changes, the graph does not provide information as to the direction of the changes. Second, the same graph cannot be used to visualize
the statistic for one and two changes. The statistic for two changes requires a three-dimensional plot, since it is a function of both $i$ and $j$. A contour plot of the log likelihood ratio statistic for pair Y55:A is presented in Figure 2.3.

![Contour plot of log likelihood ratio statistic](image.jpg)

Figure 2.3: Log LR for two changes, unequal variances. The maximum of 59.08 occurs at $i = 9$ and $j = 27$, which agrees with the map suggested by further biological testing.

### 2.3 Churchill's Algorithm

A modification of an algorithm described by Churchill (1989) addresses some problems that the likelihood analysis did not. To begin, with Churchill's algorithm we do not need to decide ahead of time how many changes are present in the data. Also, the assumption that $F_0$ and $F_1$ have the same variance is unnecessary. However, our usage of the algorithm is based on the assumption that the state of the process (IBD
or not IBD) evolves slowly along the chromosome, according to a hidden Markov process. In the yeast data case, the \( y \)'s are the unobserved (hidden) "states" and the \( x \)'s are the observed intensity readings, whose distribution depends on the states of the process. This implies that the probability of a change between segments \( i - 1 \) and \( i \) only depends on the state of segment \( i - 1 \) and not on what happened previously in the data. This assumption of a hidden Markov process is equivalent to using the Haldane mapping function, which assumes no interference. Note that the transition matrix of the Markov process is known from the length of the chromosome and the number of segments or clones spanning the chromosome.

REMARK. The Markov assumption also includes the assumption of a stationary transition mechanism so it implies that the probability of a change between segments \( i \) and \( i + 1 \) is the same for all \( i \). This would be satisfied if we had equispaced clones, which is not exactly the case for the yeast data, but close enough for purposes of the analysis.

### 2.3.1 The Algorithm

The algorithm calculates the probability that the process is in state 1 (IBD) on each segment given the data, \( P(y_i = 1|x^m) \) where \( x^k = (x_1, ..., x_k) \). If these probabilities are close to 0 or 1, they can be used to identify the segments that are IBD. To calculate the probabilities, we need to specify \( F_0 \) and \( F_1 \) up to nuisance parameters. The algorithm can be implemented for a large class of distributions, both continuous and discrete. We will again define \( F_j \) to be the cumulative distribution of a normal variable, but this time with mean \( \mu_j \) and variance \( \sigma_j^2 \) for \( j = 0 \) or 1 with \( \mu_1 > \mu_0 \).

Three sets of equations form the basis of the algorithm. For our particular case, in the same terminology as Churchill (1989), they are
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

Predictive Equation

\[
P(y_i = 1 \mid x^{i-1}) = \sum_{\delta=0}^{\frac{1}{2}} P(y_i = 1, y_{i-1} = \delta \mid x^{i-1})
= \sum_{\delta=0}^{\frac{1}{2}} P(y_i = 1 \mid y_{i-1} = \delta) P(y_{i-1} = \delta \mid x^{i-1});
\tag{2.6}
\]

Filtering Equation

\[
P(y_i = 1 \mid x^i) = \frac{P(y_i = 1 \mid x_i, x^{i-1})}{P(x_i \mid x^{i-1})}
= \frac{P(x_i \mid y_i = 1) P(y_i = 1 \mid x^{i-1})}{P(x_i \mid x^{i-1})}
= \frac{\phi \left( \frac{x_i - \mu_i}{\sigma_i} \right) P(y_i = 1 \mid x^{i-1})}{\sigma_1 P(x_i \mid x^{i-1})},
\tag{2.7}
\]

where \(\phi\) is the standard normal density and

\[
P(x_i \mid x^{i-1}) = \sum_{\delta=0}^{\frac{1}{2}} P(x_i, y_i = \delta \mid x^{i-1})
= \sum_{\delta=0}^{\frac{1}{2}} P(x_i \mid y_i = \delta, x^{i-1}) P(y_i = \delta \mid x^{i-1})
= \sum_{\delta=0}^{\frac{1}{2}} \phi \left( \frac{x_i - \mu_i}{\sigma_i} \right) \frac{1}{\sigma_i} P(y_i = \delta \mid x^{i-1});
\]

Smoothing Equation

\[
P(y_i = 1 \mid x^m) = \sum_{\delta=0}^{\frac{1}{2}} P(y_i = 1, y_{i+1} = \delta \mid x^m)
= \sum_{\delta=0}^{\frac{1}{2}} P(y_i = 1 \mid y_{i+1} = \delta, x^m) P(y_{i+1} = \delta \mid x^m)
= \sum_{\delta=0}^{\frac{1}{2}} P(y_i = 1 \mid y_{i+1} = \delta, x^i) P(y_{i+1} = \delta \mid x^m)
= \sum_{\delta=0}^{\frac{1}{2}} \frac{P(y_i = 1, y_{i+1} = \delta \mid x^i)}{P(y_{i+1} = \delta \mid x^i)} \frac{1}{P(y_{i+1} = \delta \mid x^i)}
= P(y_i = 1 \mid x^i) \sum_{\delta=0}^{\frac{1}{2}} \frac{P(y_{i+1} = \delta \mid x^m) P(y_{i+1} = \delta \mid y_i = 1)}{P(y_{i+1} = \delta \mid x^i)}.
\tag{2.8}
\]
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

Note that \( P(y_i = 1|y_{i-1} = 0) \) is the probability that a recombination occurs between segments \( i - 1 \) and \( i \). For the yeast data, given the length of 210 cM of chromosome 5 and the number of clones (31), we calculated this probability to be approximately 6.33%.

From these equations, we can calculate \( P(y_i = 1|x^m) \) by two recursive steps. In the first step, starting with an initial value for \( P(y_0 = 1) \), we use the predictive and filtering equations to compute \( P(y_i|x^{i-1}) \) and \( P(y_i|x^i) \) for all values of \( i \). The second step consists of computing \( P(y_{m-1}|x^m) \) using the values \( P(y_{m}|x^m) \) and \( P(y_{m-1}|x^{m-1}) \) from the previous step and repeating recursively to obtain \( P(y_i|x^m) \) for all \( i \).

An appropriate initial value for \( P(y_0 = 1) \) would be \( p \), the unconditional probability that the relatives are IBD on any segment (\( p = 1/2 \) for yeast).

To implement the algorithm as described above, we need to specify the value of the nuisance parameter \( \theta = (\mu_0, \mu_1, \sigma_0, \sigma_1) \). The value of \( \theta \) is unknown and needs to be estimated by a method such as maximum likelihood. The complete likelihood of \( x^n \) and \( y^n \) is

\[
P(x^n, y^n) = \prod_{k=1}^{m} P(x_k|y_k, x^{k-1})P(y_k|y^{k-1})
= \prod_{k=1}^{m} P(x_k|y_k)P(y_k|y_{k-1}).
\]

The value of \( \theta \) that maximizes the likelihood of the complete data is easily seen to be

\[
\hat{\mu}_0 = \frac{\sum_{k=1}^{n} x_k(1 - y_k)}{\sum_{k=1}^{n}(1 - y_k)}, \quad \hat{\mu}_1 = \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} y_k}
\]

\[
\hat{\sigma}_0^2 = \frac{\sum_{k=1}^{n}(x_k - \hat{\mu}_0)^2(1 - y_k)}{\sum_{k=1}^{n}(1 - y_k)}, \quad \hat{\sigma}_1^2 = \frac{\sum_{k=1}^{n}(x_k - \hat{\mu}_0)^2 y_k}{\sum_{k=1}^{n} y_k}
\]

However, we do not observe the \( y \)'s. Baum, Petrie, Soules and Weiss (1970) proposed an iterative method to find the MLE in the present context. In the same
line of thought, Churchill suggests using the EM algorithm (Dempster, Laird and Rubin 1977) in the following way to get an estimate of \( \theta \).

Start with an initial value of \( \theta \), say \( \theta^0 \).

1) Run the algorithm with \( \theta = \theta^0 \).

2) Get a revised Maximum Likelihood Estimate (MLE) of \( \theta \) by treating the estimated values of \( \tilde{y}_i = P(y_i = 1 \mid x^n) \) as data.

Perform steps one and two until convergence of \( \theta \).

The above algorithm was programmed in C and applied to the yeast data. For the yeast data, the algorithm converges in six or seven iterations.

### 2.3.2 Yeast Data

![Graphs](image)

Figure 2.4: Plots of the probabilities of being IBD for each clone, for all Y55 pairs. The \( x \)-axis represents the ordered clones. The \( y \)-axis represents the probabilities calculated by the algorithm.
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

Figure 2.4 is a plot of $P(y_i = 1|\text{data})$ for all thirty-one clones for all Y55-offspring pairs. On the y-axis are the probabilities given the data. Probabilities close to 1 indicate a segment IBD and probabilities close to 0, a segment that is not IBD. Probabilities in-between represent regions of uncertainty. Too many such regions would prevent us from determining the complete map of IBD between the pair.

From Figure 2.4, we observe that the algorithm is very useful in locating the regions of IBD since most probabilities are close to 0 or 1. Use of the information provided by the tetrad and further biological analysis suggest that the true process had crossovers between segments 9 and 10 and 27 and 28. The algorithm classified accurately all segments with the exception of the first two segments. The method seems to locate the change-points accurately and points out the region along the chromosome where further biological testing might be necessary.

However, the method has some pitfalls. Two specific situations cause the algorithm to fail. The first case is when some aberrant values are present, in which case the outlier can be identified as being in one state and the rest of the observations in the other state. The second instance when the algorithm is less useful is when no change occurs in the data. For the yeast data, the probability of no crossovers on chromosome 5 is about 12% which is not negligible. We will present both situations in detail and offer some suggestions to improve the algorithm when faced with outliers or no change in the data set.

Presence of outliers

Figure 2.5 presents the data and the probabilities obtained from the algorithm for pair Y24:B. If we remove the first data point and reapply the algorithm, the IBD map that is suggested from the probabilities agree with the knowledge that we have on what happened during meiosis (Figure 2.5 (c)). However, if we keep the first data point in the analysis, we would be making a gross error (Figure 2.5 (b)). The outlier
seems to be defined as being in one state, with the rest of the data being defined as the other state.

Figure 2.5: (a) Intensity readings for pair Y24:B. (b) Probabilities calculated from the algorithm using all data points. (c) Probabilities calculated from the algorithm removing the first data point. (d) Probabilities calculated from the algorithm using a 15% trimmed mean (all data points).

One way to resolve this problem is by using robust methods for the estimation of \( \theta \), the nuisance parameter of \( F_0 \) and \( F_1 \) in the EM algorithm. Instead of using the maximum likelihood estimates, the median or a trimmed mean provides a suitable alternative. Figure 2.5 (d) presents the output of the algorithm where \( \mu_0 \) and \( \mu_1 \) have been estimated using a 15% trimmed mean. Trimming less that 15% was not sufficient to correct the situation. Using the median gave the same results as the 15% or more trimmed mean. Moreover, in the cases where no aberrant observations were present in the data, estimating the nuisance parameter using the trimmed mean gave the same IBD map as did the maximum likelihood estimates. Since this modified version of the algorithm performs better in the presence of outliers at a minimal cost,
we suggest using it for the remaining part of the analysis.

The robust estimation method offers protection from outliers coming from extremely large values from a high intensity section, or unusually small values from a low section, but does not offer any protection against large values of a low section or small values of a high section. (Even though those cases are also considered outliers, they are not unusual values for the whole data set.) One such example is given in Figure 2.6 for pair Y24:C, segment 13. The IBD map from the algorithm suggests that a change occurred between segments 12 and 13, as well as between segments 13 and 14. This is a very unlikely event because the presence of a crossover tends to inhibit the formation of other crossovers nearby. This phenomenon, called interference, is not taken into account by the Haldane map function, which entails that the crossovers occur independently. One way to correct for this type of outlier would be to ignore any changes that do not last for at least two clones (except at the endpoints). This is an ad-hoc suggestion, but it would compensate for the assumption of no interference, which we know is not satisfied.

Figure 2.6: (a) Intensity readings for pair Y24:C. (b) Probabilities calculated from the algorithm.
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

No Changes

When no changes are present in the data, we would like all the probabilities to be close to 1/2. This would indicate that the readings come either from $F_0$ or $F_1$, without "saying" which one it is since we have no basis for comparison to know whether the mean is high or low. Simulations of data without changes show that this happens in only approximately 41% of the cases. In the remaining 59% of the cases, the algorithm incorrectly finds two states in the data, but sometimes with less certainty. Too many segments with probabilities between 0.05 and 0.95 should be a warning sign that the data might all be in one state. Further biological tests might be applied to the data to try to determine what really happened during meiosis, but this could increase the workload considerably, especially if the method is to be used on shorter chromosomes where the probability of no crossover is large.

Another alternative would be to calculate the log likelihood ratio of the map obtained from the algorithm to the map with the hypothesis of no change. The likelihood of the map is

$$
\sigma_0^{-n_0/2} \sigma_1^{-n_1/2} (2\pi)^{-m/2} \exp\left\{ \sum_{i=1}^{m} \frac{\left[ z_i - \mu_0 1(\hat{\gamma}_i < 0.5) - \mu_1 1(\hat{\gamma}_i > 0.5) \right]^2}{\sigma_0^2 1(\hat{\gamma}_i < 0.5) + \sigma_1^2 1(\hat{\gamma}_i > 0.5)} \right\} \frac{\exp(-\lambda) \lambda^k}{k!}
$$

where $k$ is the number of changes suggested by the algorithm

- $\lambda$ is probability of crossover between two clones
- $n_0$ is the number of segments with $\hat{\gamma}_i < 0.5$
- $n_1$ is the number of segments with $\hat{\gamma}_i > 0.5$

The log likelihood ratio will tend to be small when no changes occurred in the data. We could compare the log likelihood ratio to its empirical distribution obtained from the simulation of 100,000 sequences of 31 intensity readings under the hypothesis of no crossover. Note that the simulated distribution does not depend on the parameters $\mu$ and $\sigma^2$. Table 2.1 presents the quantiles obtained from the simulation. This is
CHAPTER 2. ANALYSIS OF GENOMIC MISMATCH SCANNING DATA

Table 2.1: Quantiles for Statistic to test “no change” vs “some change”.

<table>
<thead>
<tr>
<th>Quantile</th>
<th>0.50</th>
<th>0.80</th>
<th>0.90</th>
<th>0.95</th>
<th>0.975</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value of Statistic</td>
<td>3.18</td>
<td>6.98</td>
<td>9.02</td>
<td>10.90</td>
<td>12.58</td>
<td>14.51</td>
</tr>
</tbody>
</table>

equivalent to turning the problem into one of testing similar to what is in Section 2. From the empirical distribution, one could determine the cutoff for the α-level desired.

This proposed test would significantly reduce the probability of the type-I error but not without cost. It would be interesting to investigate its power under various alternative hypotheses of interest. Further data would be required to define exactly what these alternative hypotheses are. For the moment, a low value of the log likelihood ratio and many segments of probabilities between 0.05 and 0.95 can be taken as a warning that there might be no change in the data.

2.4 Discussion

Two methods were presented in this chapter to analyze Genomic Mismatch Scanning data and recover a complete map of Identity-By-Descent between two relatives. The likelihood ratio method provides a guide for determining the IBD map. It has the advantage that the clones need not be equispaced and that we do not need to assume the absence of interference of crossovers. However, it requires one to more or less specify the number of changes that are present in the data ahead of time.

The Churchill algorithm addresses some of the issues left unsolved by the likelihood analysis. It works rather nicely on the yeast data and can be easily automated. It is a little more restrictive than the likelihood analysis in assuming the Haldane mapping function. Data sets with no changes also need special handling, although probabilities computed by the algorithm in the “uncertainty” zone can be taken as an indication
that there might be no change present in the data.

More Genomic Mismatch Scanning data, especially data from human pairs, are needed to evaluate further the methodology presented in this chapter. The next chapter presents how to use the continuous map of identity-by-descent between a pair of affected relatives to locate disease susceptibility genes.
Chapter 3

GMS and Genetic Linkage

This chapter is a description of the work of Feingold (1993) and Feingold, Brown and Siegmund (1993) for identifying disease susceptibility genes using pairs of affected relatives. The rational behind this method is that if there exists a genetic component to the disease, it will most likely be found in the region that is identical-by-descent between two affected relatives. The authors assume that a continuous map of identity-by-descent (IBD) between the relatives is available through the use of Genomic Mismatch Scanning (GMS, Nelson et al. (1993)) although their methods can be used with dense sets of polymorphic markers as well.

We first present how Feingold (1993) uses Markov-chain models to obtain p-values for testing for genetic linkage using GMS data on affected relative pairs. Then, Gaussian approximations to the Markov chain models introduced by Feingold, Brown and Siegmund (1993) are discussed. The Gaussian models require a further assumption of having a large number of affected relative pairs available but are easier to work with than the Markov chain versions. In their paper, the authors include power approximations and confidence interval for the test of genetic linkage using the Gaussian models. Finally, the need to extend the current analysis to polygenic diseases/traits will be discussed.
3.1 Markov models

We describe the analysis of grandparent-grandchild pairs first since it is the simplest case. We will discuss extensions to other relative pairs such as siblings later.

![IBD configuration between a grandparent-grandchild pair.](image)

We first consider a single grandparent-grandchild pair (see Figure 3.1) and concentrate on their first chromosome. Let $Y_t = 1$ if the pair is identical-by-descent at position $t$ on the chromosome and 0 otherwise. The times where the process goes from 0 to 1 or vice versa correspond to the crossovers that occurred during the parent meiosis. If we assume that the Haldane mapping function holds i.e., that the distance between crossovers is exponential, $Y_t$ is a continuous time Markov chain with two states, 0 and 1.
CHAPTER 3. GMS AND GENETIC LINKAGE

Under the hypothesis that there is no susceptibility gene, for fixed $t$, $Y_t$ is a Bernoulli($\frac{1}{2}$) random variable. If we take $N$ such grandparent-grandchild pairs and let $X_t = \sum_{i=1}^{N} Y_{it}$, $X_t$ is also a continuous time Markov chain, but this time with state space $\{0, \ldots, N\}$. For fixed $t$, $X_t$ is a binomial($N, \frac{1}{2}$) under the null hypothesis (no susceptibility genes).

The alternative hypothesis, $H_1$, specifies that there exists a gene at location $r$ that is responsible for the disease in a proportion $\alpha$ of grandparent-grandchild pairs. Under this alternative, the expected value of $X_r$ can be computed in the following way. In $\alpha$ of the pairs, $Y_r = 1$ and in the remaining $1 - \alpha$ of the pairs, $Y_r$ has probability $\frac{1}{2}$ of being 1. Therefore,

$$E(X_r) = N\alpha + N(1 - \alpha)\frac{1}{2} = \frac{N(1 + \alpha)}{2}.$$  

Under $H_1$, $X_r$ is binomial($N, \frac{1}{2} + \alpha$). For general $t$,

$$E(X_t) = NE(Y_t) = NE[E(Y_t|Y_r)] = N\left(1 + \alpha e^{-2\lambda |t-r|}\right).$$

Since the location of the disease susceptibility gene is unknown, an intuitive statistic to test $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ is $\max_t X_t$ or the maximum number of pairs that are identical-by-descent across locations on the genome. This maximum is over all chromosomes. How large must $\max_t X_t$ be in order to declare good evidence for genetic linkage? To answer this question, we need to compute $P_0(\max_t X_t > b)$, which is equivalent to a p-value for the test. Using Aldous’ Poisson Clumping Heuristic, Feingold (1993) approximated the above probability (for one chromosome of length $l$ cM) by

$$P(\max_t X_t > b) \approx 1 - \exp\left\{-l \left[\frac{\binom{N}{b}}{2^{N}} \frac{1}{(2b - N) \lambda}\right]\right\},$$

where $\lambda = 0.01$, the crossover rate by unit of genetic distance (in centimorgans) under the Haldane mapping function.
Note that the p-value of the test that accounts for all \( n \) chromosomes can be written as

\[
1 - \Pi_{i=1}^{n} \exp \left\{ -l_i \left[ \binom{N}{b} \frac{1}{2^N} (2b - N) \lambda \right] \right\} = 1 - \exp \left\{ \binom{N}{b} \frac{1}{2^N} (2b - N) \lambda \left[ \sum_{i=1}^{n} -l_i \right] \right\},
\]

where \( l_i \) is the length of chromosome \( i \) and \( n \) is the number of chromosomes (\( n = 23 \) for humans). This approximation is easy to compute and was shown to be quite accurate through simulations. Note that the overall p-value only depends on the length of the genome \( \sum_{i=1}^{n} l_i \) and not on the number of chromosomes.

Similar approximations are available for other types of relatives such as siblings, half siblings, cousins, etc. Getting power approximations for the above test or confidence intervals for the location of the genes or the proportion \( \alpha \) is a harder problem. To simplify the matter, Feingold, Brown and Siegmund (1993) explored a Gaussian approximation to the Markov chain models.

### 3.2 Gaussian approximations

Gaussian approximations have the extra limitation of needing large numbers of affected relative pairs to be valid, but they give some insight into more complicated problems. We will begin with the analysis of the same grandparent-grandchild pairs that was done with the Markov-chain models to show the connection between the two models.

#### 3.2.1 Grandparent-grandchild pairs

Let \( p \) be the probability that the relative pair is IBD at any given locus. Note that \( p = \frac{1}{2} \) for the grandparent-grandchild pairs. Let

\[
Z_t = \frac{X_t - Np}{\sqrt{N}}.
\]
CHAPTER 3. GMS AND GENETIC LINKAGE

As \( N \to \infty \), \( Z_t \) tends in distribution to an Ornstein-Uhlenbeck process with covariance function \( R(t) = \sigma^2 e^{-\beta t} \), where \( \beta = 2\lambda = 0.02, \sigma^2 = p(1 - p) = 1/4 \), and with mean function derived as follows. On chromosomes not containing the disease-susceptibility gene locus \( r \), \( Z_t \) has mean 0. On the chromosome with the locus \( r \),

\[
E(Z_t) = \frac{1}{\sqrt{N}} [E(X_t) - NP] = \sqrt{N \alpha p} e^{-\beta |t-r|} = \xi e^{-\beta |t-r|},
\]

(3.1)

where \( \xi = \sqrt{N \alpha p} \).

Testing for genetic linkage is equivalent to testing \( H_0 : \xi = 0 \) versus \( H_1 : \xi > 0 \). The log likelihood ratio statistic for the above test is

\[
\max_t \frac{Z_t}{\sigma}.
\]

(3.2)

This is equivalent to the test used for the Markov-chain model.

A large literature is available on Gaussian processes, much of which is applicable to the problem of finding the distribution of (3.2) under the null and the alternative hypotheses. The approximations to the p-value and the power suggested by Feingold, Brown and Siegmund (1993) are

\[
P_0 \{ \max_t \frac{Z_t}{\sigma} > b \} \approx 1 - \Phi(b) + \beta lb\phi(b),
\]

(3.3)

and

\[
P_{(r, \xi)} \{ \max_t \frac{Z_t}{\sigma} > b \} \approx 1 - \Phi(b - \frac{\xi}{\sigma}) + \phi(b - \frac{\xi}{\sigma}) \left[ 2 \left( \frac{\xi}{\sigma} \right)^{-1} - \left( b + \frac{\xi}{\sigma} \right)^{-1} \right],
\]

(3.4)

where \( l \) is the length of the chromosome in centimorgans (cM), \( \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \) and \( \Phi(x) = \int_{-\infty}^{x} \phi(x)dx \). For an overall significance of the test, (3.3) needs to be summed over all \( n \) chromosomes i.e.,

\[
p\text{-value} \approx \sum_{i=1}^{n} [1 - \Phi(b) + \beta l_i b\phi(b)].
\]
CHAPTER 3. GMS AND GENETIC LINKAGE

When the IBD information is not continuous but rather from a set of equispaced genetic markers, (3.3) becomes

$$P_0\{\max_i \frac{Z_i}{\sigma} > b\} \approx 1 - \Phi(b) + \beta b \phi(b) \nu(b\sqrt{2\beta\Delta}), \quad (3.5)$$

where $\Delta$ is the distance between the markers and $\nu(x) \approx e^{-0.583x}$. Note that this equation is equivalent to (3.3) when $\Delta = 0$. The function $\nu(x)$ is a discreteness correction factor to account for the fact that we are observing the likelihood ratio statistic at discrete points on the chromosomes instead of continuously as was the case for data from GMS. For an exact definition of $\nu(x)$, see Siegmund (1985) p.82.

The power in the case of equispaced markers is a little more complicated since it depends on the distance between the locus $r$ and its closest markers. For the case where the locus $r$ is exactly at a marker, (3.4) becomes

$$P_{(r,\xi)}\{\max_i \frac{Z_i}{\sigma} > b\} \approx 1 - \Phi(b - \frac{\xi}{\sigma}) + \phi(b - \frac{\xi}{\sigma}) \left[2\left(\frac{\xi}{\sigma}\right)^{-1} \nu - \left(b + \frac{\xi}{\sigma}\right)^{-1} \nu^2\right], \quad (3.6)$$

where $\nu = \nu(b\sqrt{2\beta\Delta})$. The case when the distance between $r$ and its closest marker is not 0 will be presented in the next chapter.

These approximations, for both continuous and discrete data, are valid for Gaussian processes with covariance functions of the general form

$$R(t) = \sigma^2(1 - \beta|t| + o(|t|)) \quad \text{as} \quad t \to 0,$$

which is satisfied by the Ornstein-Uhlenbeck process but not restricted to it. This property is very convenient for other types of relatives. For example, for avuncular pairs,

$$R(t) = \sigma^2\left(\frac{e^{-4\lambda t}}{2} + \frac{e^{-6\lambda t}}{2}\right),$$

and $Z_t$ is no longer an Ornstein-Uhlenbeck process. However, the approximations to the power and p-value can still be used with $\beta = 5\lambda$ since the covariance function
CHAPTER 3. GMS AND GENETIC LINKAGE

\[ R(t) = \sigma^2 (1 - 5\lambda|t| + o(|t|)) \quad \text{as} \quad t \to 0. \text{ (see Feingold, Brown and Siegmund (1993) for more details). Here is a list of covariance functions for different types of relatives:} \]

**Grandparent-Grandchild**

\[
R(t) = \sigma^2 \exp(-2\lambda|t|) \quad (\beta = 2\lambda),
\]

**Half-siblings**

\[
R(t) = \sigma^2 \exp(-4\lambda|t|) \quad (\beta = 4\lambda),
\]

**Avuncular**

\[
R(t) = \sigma^2 \left[ \frac{\exp(-4\lambda|t|) + \exp(-6\lambda|t|)}{2} \right] \quad (\beta = 5\lambda),
\]

**Cousins**

\[
R(t) = \sigma^2 \left[ \frac{\exp(-4\lambda|t|)}{2} + \frac{\exp(-6\lambda|t|)}{3} + \frac{\exp(-8\lambda|t|)}{6} \right] \quad (\beta = \frac{16}{3} \lambda).
\]

### 3.2.2 Siblings

We carefully omitted the case of siblings so far. The reason is that siblings can be identical-by-descent on one or both of their chromosomes, which complicates the analysis (see Figure 3.2).

Let \( X_{it} \) be the number of sibling pairs identical-by-descent on \( i \) chromosomes at location \( t, i = 0, 1 \) or 2. Define \( X_t = (X_{0t}, X_{1t}, X_{2t}) \) where \( X_{0t} + X_{1t} + X_{2t} = N \), the total number of pairs. Unfortunately, when a pair of siblings is IBD, the current method of Genomic Mismatch Scanning does not allow one to differentiate between IBD on one chromosome and IBD on both chromosomes. What this means is that only the sum \( X_{1t} + X_{2t} \) can be observed but not \( X_{1t} \) and \( X_{2t} \) individually. However, both the test when \( X_{1t} \) and \( X_{2t} \) are observed separately and the test when only the sum of the two is available have been studied.
Figure 3.2: IBD configuration between a sibling pair.

Let

\[
Z_{1t} = \frac{X_{1t} - \frac{N}{2}}{\sqrt{N}}, \quad Z_{2t} = \frac{X_{2t} - \frac{N}{4}}{\sqrt{N}}.
\]

In the ideal situation where \(X_{1t}\) and \(X_{2t}\) are observed, the likelihood ratio statistic to test \(H_0 : \alpha = 0\) versus \(H_1 : \alpha \neq 0\) is

\[
\max_i \frac{Z_{1t}/2 + Z_{2t}}{\sigma} \quad (3.7)
\]

where \(\sigma^2 = 1/8\). The process \(\frac{Z_{1t}/2 + Z_{2t}}{\sigma}\) is an Ornstein-Uhlenbeck process with covariance function \(R(t) = \sigma^2 e^{-4\lambda|t|}\). It has mean 0 under \(H_0\) and mean function \(\alpha \sqrt{N} e^{-4\lambda|t|}/4 = \xi e^{-4\lambda|t|}\) under \(H_1\). Therefore, the approximations given for the grandparent-grandchild pairs are applicable with \(\beta = 4\lambda\).
CHAPTER 3. GMS AND GENETIC LINKAGE

If we observe \( Z_t^* = Z_{1t} + Z_{2t} \), a plausible test statistic is

\[
\max_t \frac{Z_t^*}{\sqrt{\sigma_1^2 + \sigma_2^2}}
\]

(3.8)

where \( \sigma_1^2 = 1/16, \sigma_2^2 = 1/8 \).

Since this statistic is not actually the likelihood ratio statistic, approximations (3.3) and (3.4) need to be modified slightly. Feingold, Brown and Siegmund (1993) have shown that (3.8) is about 1/3 less efficient than (3.7) i.e., 1/3 less pairs are needed to obtain the same power when observing both \( X_{1t} \) and \( X_{2t} \) as opposed to the sum \( X_{1t} + X_{2t} \). Hopefully, the laboratory techniques will allow one to observe both \( X_{1t} \) and \( X_{2t} \) in the near future, making this comparison unnecessary.

3.2.3 Confidence Intervals

When it has been determined that there exists a genetic component to a disease, we would like to identify a region of a chromosome in which to concentrate the search for that gene. The location where the likelihood ratio statistic of the previous sections takes on its maximum provides a point estimate for the gene locus, but a confidence region is needed in order to know how close to the point estimate the real gene might lie. Methods borrowed from the change-point literature can be used to construct such confidence intervals. The value of the locus \( r \) where the mean of \( Z_t \) reaches its maximum is a change-point since there is a jump in the value of the derivative of the mean function (3.1) at \( t = r \).

By appropriately modifying the argument of Siegmund (1988), it was shown in Feingold, Brown and Siegmund (1993) that values of \( Z_v \) sufficiently close to \( Z^* = \max_t Z_t \) provide a confidence interval.
CHAPTER 3. GMS AND GENETIC LINKAGE

More formally, if we let all $R(t^* - v) = Z_v/Z^*$ and $Z^* = (\max_t Z_t)_{obs}$, then all values $v$ for which $Z_v$ satisfies

$$2\beta |R'(t^* - v)| \exp \left[-\frac{(Z_{v}^{2} - Z_{v}^{2})}{2\sigma^2}\right] \geq \gamma,$$

form a $1 - \gamma$ confidence interval for the gene locus. When the covariance function $R(t)$ is of the form $e^{-\beta||t||}$, $R'(t^* - v) = \beta R(t^* - v) = \beta Z_v/Z^*$. Otherwise, $R'(t^* - v)$ must be found numerically. Note that the confidence interval might not be connected. However, this is not a problem since further search for the exact location of the locus can be restricted to the disjoint regions of the chromosome that constitute the confidence interval.

3.2.4 Discussion

The methods discussed in this chapter are intended for a trait that are influenced by one major gene. If, however, the disease is polygenic (more than one gene locus are involved in increasing susceptibility) the methods are not optimal. The likelihood ratio will be larger at contributing gene loci but it may fail to exceed the threshold $b$ if there are a lot of loci each increasing susceptibility slightly and the number of affected relative pairs available is not sufficiently large. A statistic that searches simultaneously for multiple loci will offer some advantages in such a situation. The next chapter is devoted to the development of statistical tests for polygenic traits.
Chapter 4

Polygenic diseases

The previous chapter includes tests for locating a single disease susceptibility gene. If a few such genes exist and each makes only a small contribution to the susceptibility, then the test that is searching for only one gene may fail to detect any genetic component to the disease. In this chapter we develop methods to look simultaneously for more than one gene. There are two particular multilocus situations of interest. First, there could be genetic heterogeneity, meaning that the pairs of affected relatives can be divided in subgroups and different gene loci are responsible for the disease in different subgroups of the population. This implies that there is no interaction between loci. This situation will be modeled using Risch's additive model (Risch (1990a)). A second situation of interest is epistasis or gene interaction. In this instance, the presence of both genes is necessary to increase susceptibility to the disease. This can be modeled using Risch's multiplicative model.

The next section defines tests for polygenic diseases, with special attention given to the additive model. We will give approximations to the p-values of the tests in the case of traits with 2 or 3 loci involved. These approximations are provided for both the case where the IBD state between relatives is from GMS, and therefore continuous, and the case where the IBD state is from a set of equispaced markers.
Section 4.2 contains power approximations for the tests. Section 4.3 present results from a simulation study conducted to verify the accuracy of the approximations to the power. The derivations of the power approximations are included in Appendix A.

4.1 Tests and p-values

4.1.1 Tests

We assume a two-locus heterogeneous disease. More precisely, we assume that in a proportion \( \alpha_1 \) of the affected relative pairs the disease is influenced by a locus \( r_1 \) and in \( \alpha_2 \) of the pairs the disease is influenced by locus \( r_2 \). Obviously, \( \alpha_1 + \alpha_2 \leq 1 \). We will start with the simplest case of \( \alpha_1 = \alpha_2 \) and \( r_1 \) and \( r_2 \) unlinked or on different chromosomes and then remove the first assumption later.

Let \( Z_{it} \) be defined as in the previous chapter i.e.,

\[
Z_{it} = \frac{X_{it} - Np}{\sqrt{N}}
\]

where \( X_{it} \) is the number of relatives pairs that are IBD at location \( t \) on the \( i^{th} \) chromosome, \( N \) is the total number of relative pairs, and \( p \) is the probability that the relative pair is IBD at any given point on their genome.

To test \( H_0 : \xi_1 = \xi_2 = 0 \) versus \( H_1 : \xi_1 = \xi_2 \neq 0 \), where \( \xi_i = Np\alpha_i \), one sees that in the approximate Gaussian framework of Feingold, Brown and Siegmund (1993) the likelihood ratio statistic is

\[
\max_{s \neq t} \left( \frac{Z_{it} + Z_{js}}{\sqrt{2}} \right),
\]

where \( i \) and \( j \) are the chromosome numbers and \( s \) and \( t \) are locations along the chromosomes. This statistic is derived from the fact that (Feingold, Brown and Siegmund (1993))

\[
\frac{dP_{\xi}}{dP_0}(Z) = \exp\left(\frac{\xi Z_{r1}}{\sigma^2} - \frac{\xi^2}{2\sigma^2}\right) \exp\left(\frac{\xi Z_{r2}}{\sigma^2} - \frac{\xi^2}{2\sigma^2}\right).
\]
CHAPTER 4. POLYGENIC DISEASES

Maximizing the likelihood over the unknown parameters \( r_1, r_2 \) and \( \xi \) gives (4.1).

If we remove the assumption that \( \xi_1 = \xi_2 \) i.e., the hypotheses are \( H_0 : \xi_1 = \xi_2 = 0 \) and \( H_1 : \xi_1 \neq 0 \text{ or } \xi_2 \neq 0 \), we can show similarly that the likelihood statistic is

\[
\max_{i,j} \left( \frac{Z_{it}^2 + Z_{jt}^2}{\sigma^2} \right),
\]

where \( Z^+ = \max(0, Z) \).

The statistics (4.1) and (4.2) generalize easily to more than two loci. For the case of \( k \) equally contributing loci, the likelihood statistic is

\[
\max_{i_1 < \ldots < i_k} \left( \frac{Z_{i_1 t_1} + \ldots + Z_{i_k t_k}}{\sigma\sqrt{k}} \right),
\]

and when the \( k \) loci are not equally contributing, the likelihood statistic is

\[
\max_{i_1 < \ldots < i_k} \left( \frac{Z_{i_1 t_1}^2 + \ldots + Z_{i_k t_k}^2}{\sigma^2} \right).
\]

In order to define a rejection region for the tests, we will find approximations to the tail distribution of the likelihood statistics under \( H_0 \), which is equivalent to finding approximations for the p-value of the tests. We will first give approximations for the case of two contributing loci, then for the case of three loci.

4.1.2 Two-locus p-values

Let \( Z_i = \max_{t \in D} Z_{it}/(\sigma\sqrt{2}) \) or \( Z_i = \max_{t \in D} Z_{it}^2/\sigma^2 \), depending on whether we are using statistic (4.1) or (4.2). In the case when the data are from GMS, \( D = [0, l_i] \) where \( l_i \) is the length of the chromosome \( i \). When the data comes from a dense set of equispaced markers, \( D = \{ k\Delta : 0 \leq k\Delta \leq l_i \} \) and \( \Delta \) is the distance in centimorgans between markers. To get a simpler approximation to the p-value, we assume that each chromosome has the same length, hence the \( Z_i \)'s are iid under \( H_0 \). Let \( Z(j) \) be the \( j^{th} \) order statistic of \( Z_1, \ldots, Z_n \), \( n \) being the total number of chromosomes (23 in
humans). Then statistics (4.1) and (4.2) can be written as

\[ Z_{(n-1)} + Z_{(n)}, \]

which is the sum of the two largest order statistics. Define

\[ F(x) = P\left( \frac{Z_i}{\sigma} \leq x \right) \quad \text{and} \quad \bar{F}(x) = 1 - F(x). \]

To calculate the p-value for test (4.1), we define \( B = \{(x, y): x \leq y, x + y > b\} \) and use the equation

\[
P(Z_{(n-1)} + Z_{(n)} > b) = \int \int_{(x,y) \in B} P(Z_{(n-1)} \in dx, Z_{(n)} \in dy)
= \int_{0}^{F(b/2)} n(n-1)(1-y)^{n-2}ydy + \int_{-\infty}^{b/2} F^{n-2}(x) \bar{F}(b-x)f(x)dx.
\]

(4.5)

A similar equation for test (4.5) would have the added constraint that \( x \geq 0 \) but otherwise would be identical.

Unfortunately, we do not know the exact distribution \( F(x) \). However, Feingold, Brown and Siegmund (1993) provide some approximations to \( F(x) \) that were described in Chapter 3 and that we can use with equation (4.5) to determine the rejection region for the tests for 2 loci. An approximation to the density \( f(x) \) can be obtained by differentiating that of \( F(x) \). For statistic (4.1) which assumes equally contributing loci, we will use

\[
\bar{F}(x) = P(\max_{i} \frac{Z_{it}}{\sigma} > x\sqrt{2}) \approx 1 - \Phi(x\sqrt{2}) + \sqrt{2}\beta i x\phi(x\sqrt{2})\nu(2x\sqrt{\beta\Delta}).
\]

(4.6)

Since (4.6) holds best for large \( x \) and \( x\sqrt{\Delta} \) constant, we do not differentiate the function \( \nu(2x\sqrt{\beta\Delta}) \) in obtaining the approximation to the density \( f(x) \). Note that when the data is from Genomic Mismatch Scanning, \( \Delta = 0 \) and \( \nu(2x\sqrt{\beta\Delta}) = 1 \).

For statistic (4.2), without the assumption of \( \xi_1 = \xi_2 \), we can use

\[
\bar{F}(x) = P(\max_{i} \frac{Z_{it}^+}{\sigma^2} > x) \approx 1 - \Phi(\sqrt{x}) + \beta i \sqrt{x} \phi(\sqrt{x})\nu(\sqrt{2\beta\Delta x}).
\]

(4.7)
CHAPTER 4. POLYGENIC DISEASES

The approximations can be substituted in equation (4.5), and the calculations performed numerically. Solving for \( P_0(Z_{(n-1)} + Z_n > b) = 0.05 \), we obtain the thresholds in Tables 4.1 and 4.2. To make the results between statistics (4.1) and (4.2) comparable, we have found the thresholds for the square root of statistic (4.2). The tables also include a simulation study performed to evaluate the accuracy of the approximations.

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>4.31</td>
<td>4.32</td>
<td>4.48</td>
<td>4.49</td>
<td>4.62</td>
<td>4.63</td>
</tr>
<tr>
<td>H-S</td>
<td>4.36</td>
<td>4.35</td>
<td>4.58</td>
<td>4.58</td>
<td>4.76</td>
<td>4.74</td>
</tr>
<tr>
<td>Av.</td>
<td>4.37</td>
<td>4.36</td>
<td>4.60</td>
<td>4.60</td>
<td>4.80</td>
<td>4.77</td>
</tr>
<tr>
<td>C-I</td>
<td>4.37</td>
<td>4.39</td>
<td>4.61</td>
<td>4.60</td>
<td>4.81</td>
<td>4.79</td>
</tr>
</tbody>
</table>

Table 4.1: Threshold for 2-locus test (4.1) for type-1 error of 0.05. \( \Delta \) is the distance between markers in centimorgans. The continuous observation case is represented by \( \Delta = 0 \). App. stands for approximation. Sim. corresponds to the simulation threshold values. The type of relative pairs studied are: grandparent-grandchild (G-G), half-siblings (H-S), Avuncular (Av.) and first cousins (C-I).

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>4.34</td>
<td>4.36</td>
<td>4.51</td>
<td>4.56</td>
<td>4.65</td>
<td>4.85</td>
</tr>
<tr>
<td>H-S</td>
<td>4.39</td>
<td>4.38</td>
<td>4.61</td>
<td>4.61</td>
<td>4.79</td>
<td>4.79</td>
</tr>
<tr>
<td>Av.</td>
<td>4.40</td>
<td>4.41</td>
<td>4.63</td>
<td>4.63</td>
<td>4.82</td>
<td>4.80</td>
</tr>
<tr>
<td>C-I</td>
<td>4.40</td>
<td>4.42</td>
<td>4.63</td>
<td>4.64</td>
<td>4.83</td>
<td>4.82</td>
</tr>
</tbody>
</table>

Table 4.2: Threshold for 2-locus test (4.2) for type-1 error of 0.05.

In the simulation study we used the asymptotic distribution of \( Z_{it} \), which was shown by Feingold et al. (1993) to be an Ornstein-Uhlenbeck process for grandparent-grandchild and half sibling pairs, and the sum of Ornstein-Uhlenbeck processes for
avuncular and cousin pairs. Data for each chromosome was simulated using the fact that \( W(e^{2\beta t})/e^{\beta t} \) has the same distribution as an Ornstein-Uhlenbeck process with mean 0 and covariance function \( e^{-\beta |t|} \), where \( W(e^{2\beta t}) \) is standard Brownian Motion at time \( e^{2\beta t} \). The simulation of the Ornstein-Uhlenbeck process was done at discrete intervals of \( \Delta \) for 23 chromosomes of length 140 centimorgans each, to approximate the human genome. The simulation was repeated 10,000 times and for each iteration the value of the test statistic was computed. The 95th quantile was then calculated and is presented in Table 4.1 and 4.2. For the continuous case, the Ornstein-Uhlenbeck process was first simulated at \( \Delta = 1 \) cM. Then, interpolation between the discrete points was performed in the following way. Note that for small \( \Delta \) and \( t_2 - t_1 = \Delta \),

\[
P\left( \max_{t_1 < t < t_2} \frac{Z_{it}}{\sigma} > a, \frac{Z_{it_1}}{\sigma} = x_1, \frac{Z_{it_2}}{\sigma} = x_2 \right) \approx \exp\left[ -\frac{(a - x_1)(a - x_2)}{\beta \Delta} \right]. \quad (4.8)
\]

So for each interval of length \( \Delta \), a random uniform was drawn. We took the maximum reached by the continuous process in the interval \( (t_1, t_2) \) to be greater than \( a \) if the random uniform was smaller than (4.8). Otherwise, the maximum of the process was taken to be smaller than \( a \). The interpolation was repeated for all \( \frac{1}{\Delta} \) intervals. The results from the continuous case simulations and approximations correspond to \( \Delta = 0 \) in Tables 4.1 and 4.2.

### 4.1.3 Three-locus p-values

The distribution of the statistics for 3 loci is more complicated but is dealt with in a similar fashion. The hypotheses that we are testing can be stated as \( H_0 : \xi_1 = \xi_2 = \xi_3 = 0 \) versus \( H_1 : \xi_1 \neq 0 \) or \( \xi_2 \neq 0 \) or \( \xi_3 \neq 0 \). The likelihood statistic for testing for 3 contributing loci can be written as

\[
Z_{(n-2)} + Z_{(n-1)} + Z_{(n)}
\]

where \( Z_i = \max_{t \in D} Z_{it}/(\sigma \sqrt{3}) \) or \( Z_i = \max_{t \in D} Z_{it}^2/\sigma^2 \).

The equivalent of equation (4.5) for 3 loci can be written as
\[ P \left( Z_{(n-2)} + Z_{(n-1)} + Z_{(n)} > b \right) = \int\int_{(x,y,z) \in B} P \left( Z_{(n-2)} \in dx, Z_{(n-1)} \in dy, Z_{(n)} \in dz \right) \]

\[ = \int_{-\infty}^{b/3} \int_{x}^{b/2} \frac{n(n-1)(n-2)}{2} \tilde{F}(b-x-y) F^{(n-3)}(x) f(x) f(y) dy dx \]

\[ + \int_{-\infty}^{b/3} \frac{n(n-1)(n-2)}{2} \tilde{F}^2 \left( \frac{b-x}{2} \right) F^{(n-3)}(x) f(x) dx \]

\[ + \int_{0}^{F(b/3)} \frac{n(n-1)(n-2)}{2} (1-y)^{n-3} y^2 dy. \quad (4.9) \]

where \( B = \{(x,y,z) : x \leq y \leq z, x+y+z > b \} \).

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>20</th>
<th>20</th>
<th>10</th>
<th>10</th>
<th>5</th>
<th>5</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>4.86</td>
<td>4.90</td>
<td>5.09</td>
<td>5.10</td>
<td>5.26</td>
<td>5.29</td>
<td>5.52</td>
<td>5.52</td>
<td>5.76</td>
<td>5.72</td>
</tr>
<tr>
<td>H-S</td>
<td>4.94</td>
<td>4.98</td>
<td>5.23</td>
<td>5.24</td>
<td>5.46</td>
<td>5.46</td>
<td>5.82</td>
<td>5.81</td>
<td>6.14</td>
<td>6.08</td>
</tr>
<tr>
<td>Av.</td>
<td>4.96</td>
<td>4.99</td>
<td>5.26</td>
<td>5.28</td>
<td>5.51</td>
<td>5.49</td>
<td>5.90</td>
<td>5.90</td>
<td>6.25</td>
<td>6.19</td>
</tr>
<tr>
<td>C-I</td>
<td>4.96</td>
<td>4.99</td>
<td>5.26</td>
<td>5.29</td>
<td>5.52</td>
<td>5.52</td>
<td>5.92</td>
<td>5.90</td>
<td>6.29</td>
<td>6.24</td>
</tr>
</tbody>
</table>

Table 4.3: Threshold for 3-locus test (4.3) for type-1 error of 0.05.

<table>
<thead>
<tr>
<th>( \Delta )</th>
<th>20</th>
<th>20</th>
<th>10</th>
<th>10</th>
<th>5</th>
<th>5</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>4.92</td>
<td>4.96</td>
<td>5.14</td>
<td>5.16</td>
<td>5.32</td>
<td>5.33</td>
<td>5.58</td>
<td>5.57</td>
<td>5.80</td>
<td>5.76</td>
</tr>
<tr>
<td>H-S</td>
<td>4.99</td>
<td>5.05</td>
<td>5.27</td>
<td>5.29</td>
<td>5.50</td>
<td>5.51</td>
<td>5.86</td>
<td>5.85</td>
<td>6.18</td>
<td>6.12</td>
</tr>
<tr>
<td>Av.</td>
<td>5.00</td>
<td>5.04</td>
<td>5.30</td>
<td>5.32</td>
<td>5.55</td>
<td>5.53</td>
<td>5.94</td>
<td>5.93</td>
<td>6.29</td>
<td>6.23</td>
</tr>
<tr>
<td>C-I</td>
<td>5.00</td>
<td>5.05</td>
<td>5.30</td>
<td>5.34</td>
<td>5.56</td>
<td>5.56</td>
<td>5.96</td>
<td>5.95</td>
<td>6.32</td>
<td>6.27</td>
</tr>
</tbody>
</table>

Table 4.4: Threshold for 3-locus test (4.4) for type-1 error of 0.05.

The approximations to \( F(x) \) and \( f(x) \) can be used in conjunction with the equation (4.9) to calculate approximate thresholds that define rejection regions for the tests.
for 3 loci. Tables 4.3 and 4.4 present approximate thresholds and simulation results for statistic (4.3) and the square root of statistic (4.4). The approximations are quite accurate.

The degree of complexity of the approximation to the p-values increases with the number of loci in the model, but can be obtained as with the cases of two and three loci.

4.1.4 Siblings

We mentioned earlier that the case of siblings is a little more complex since they can be identical-by-descent on one or both of their chromosomes. If we can observed whether the siblings are identical-by-descent on one or two chromosomes, we can use the same thresholds as for the half-siblings (since $\beta = 0.04$) but with

$$Z_{it} = \frac{1}{2} \left( X_{it}^{(1)} \frac{N}{2} + \left( X_{it}^{(2)} - \frac{N}{4} \right) \right) \sqrt{\frac{\sigma}{N}}$$

where $X_{it}^{(1)}$ is the number of sibling pairs IBD on exactly one chromosome at location $t$ of the $i^{th}$ chromosome and $X_{it}^{(2)}$ is the number of sibling pairs IBD on both chromosomes at location $t$ of the $i^{th}$. The tests for two and three loci can be performed in a manner similar to that for the other types of relatives with this new definition of $Z_{it}$.

4.1.5 Which test should be used?

The case of equally contributing loci is the hardest to detect because if one of the loci makes a larger contribution to the susceptibility of disease, it will most likely be identified through a single locus search. For this reason, we prefer test (4.3) over test (4.4) for equally contributing loci.

Both tests (4.3) and (4.4) were derived using the Gaussian approximations to the Markov chain processes and Risch's additive model, which is designed to model genetic
CHAPTER 4. POLYGENIC DISEASES

heterogeneity. When a multiplicative model is indicated because of the presence of epistasis, one can show that statistic (4.3) is still appropriate, this time based on the Markov chain model of Chapter 3.

More precisely, assume a multiplicative model with two unlinked loci, \( r_1 \) and \( r_2 \), both with genetic effect \( \alpha \). Let \( X_{ijk} = 1(X_{ir_1} = j, X_{ir_2} = k) \), for \( i = 1, \ldots, N \), \( j, k = 0, 1 \) and \( N \) is the total number of relative pairs. For example, \( X_{i11} \) is the indicator that the \( i^{th} \) pair of relatives is IBD at both \( r_1 \) and \( r_2 \). Let \( Y_{jk} = \sum_{i=1}^{N} X_{ijk} \). We can write the log likelihood ratio of \( H_0 : \alpha = 0 \) vs \( H_1 : \alpha \neq 0 \) as

\[
Y_{11} \log [(1 + \alpha)^2] + (Y_{10} + Y_{01}) \log [1 - \alpha^2] + Y_{00} \log [(1 - \alpha)^2].
\]

Maximizing over all values of \( \alpha \), we get that the likelihood ratio statistic is

\[
Y_1 + Y_{11} - N,
\]

where \( Y_1 = \sum_{k=0}^{1} Y_{1k} \). If we maximize (4.10) over all possible values of \( r_1 \) and \( r_2 \) unlinked, we get statistic (4.1), that was derived for the additive model. Therefore, statistic (4.1) can be used for both heterogeneity and epistasis mode of inheritance.

When the number of contributing loci is not known, which test should be used? There is a way of performing the test for one, two and three loci sequentially while maintaining the overall significance level of 0.05 by increasing the threshold for each test slightly. However, to do so a limit on the number of tests that are going to be performed must be set ahead of time. For example, if we used a thresholds of 4.2, 5.6 and 6.7 for the tests of one, two and three loci with half-siblings pairs as opposed to the thresholds 4.09, 5.29, 6.14, we could perform the tests sequentially and still have an overall probability of 0.05 of erroneously detecting a genetic effect. Since the increases in the thresholds are small, we recommend using this approach to test for the presence of disease susceptibility loci. It would greatly improve the odds of detecting a disease having 2 or 3 equally contributing loci. However, when there is
CHAPTER 4. POLYGENIC DISEASES

only one locus with a strong genetic effect, the sequential procedure may reveal a false second locus. In this instance, repeating the experience on a different data set might help in determining if the contribution to the disease of the second locus is real.

4.2 Power approximations

In the previous section, we developed approximations to the distributions of the test statistics under the null hypothesis. The next step is to find approximations to the power. Power calculations would enable us to compute sample size requirements. The power calculations presented in this section are mainly for equally contributing loci, which is the most difficult situation in which to establish genetic linkage. Similar approximations can be obtained for the case where the loci have unequal genetic effects.

4.2.1 One locus

For the one locus model, the alternative was that there exists a gene locus \( r \), that increases susceptibility to the disease in a proportion \( \alpha \) of the pairs. The power approximation for the one locus statistic, presented in the previous chapter, is

\[
P_{\ell,r}(\max_{0 \leq \Delta \leq 1} \frac{Z_{1\Delta}}{\sigma} > b) \approx 1 - \Phi(b - \frac{\xi}{\sigma}) + \phi(b - \frac{\xi}{\sigma})[2(\frac{\xi}{\sigma})^{-1} \nu - (\frac{\xi}{\sigma} + b)^{-1} \nu^2],
\]

(4.11)

when the gene locus \( r \) corresponds to a marker (which includes the continuous case). The function \( \nu = \nu(b, \sqrt{2B\Delta}) \) is a discreteness correction factor defined in Chapter 3 and \( \nu = 1 \) for the continuous data case. In the above approximation it is assumed that \( b \) and \( \xi \) are large and this assumption is in force for the rest of the approximations in this section. Note that the parameter \( \beta \) enters into the equation only as a function of the level \( b \), since \( b \) depends on \( \beta \).
CHAPTER 4. POLYGENIC DISEASES

The case where $r$ is between markers is more complex. Let $t_1$ and $t_2$ be the two
flanking markers of $r$. Then

$$P_{\xi,\sigma}(\max_{0 \leq \Delta \leq \ell} \sigma^{-1} Z_\Delta > b) \approx 1 - \int_0^\infty h(x, t_1, r) \phi\left\{ \frac{g(x, t_2, t_1, r)}{d(\Delta)} \right\} dx$$

$$+ \nu (b \sqrt{2} \Delta) \int_0^\infty e^{-b(x, t_2, r)} \phi\left\{ \frac{g(x, t_2, t_1, r)}{d(\Delta)} \right\} dx$$

$$+ \nu (b \sqrt{2} \Delta) \int_0^\infty e^{-b(x, t_1, r)} \phi\left\{ \frac{g(x, t_2, t_1, r)}{d(\Delta)} \right\} dx$$

$$- \nu^2 (b \sqrt{2} \Delta) \int_0^\infty h(x, t_1, r) G(x, t_1, t_2) dx \quad (4.12)$$

where

$$d(\Delta) = \sqrt{1 - e^{-2b\Delta}},$$

$$h(x, t, r) = b - x - \frac{\xi}{\sigma} e^{-\beta|x-r|},$$

$$g(x, s, t, r) = b - \frac{\xi}{\sigma} e^{-\beta|x-s-r|} - e^{-\beta \Delta} \left[ b - x - \frac{\xi}{\sigma} e^{-\beta|x-r|} \right]$$

$$G(x, s, t, r) = \exp \left[ -bx - bg(x, t_2, t_1, r) + b^2 d(\Delta)/4 \right] \phi\left\{ \frac{g(x, t_2, t_1, r)}{d(\Delta)} \right\} - bd(\Delta).$$

The integrals must be evaluated numerically. Table 4.5 presents a simulation
study to verify the accuracy of (4.12). The approximation is very accurate. We can
use the power approximations to calculate sample size requirements. For example,
for grandparent-grandchild pairs, to reach a power of 80% with a continuous map of
IBD, $\xi/\sigma = \sqrt{Np} a/\sigma = 4.39$. For this type of relative pairs, $p = 1/2$ and $\sigma = 1/2$.
If it is believed that $a = 0.5$, then at least 78 pairs are required to reach the desired
power of 80%. The sample size will depend on the type of relative and the proportion
$\alpha$. Feingold et al. (1993) present a table on the relation between the parameter $\alpha$ and
the recurrence risk for different types of relatives (cf. Risch (1990a)), which can often
be estimated from epidemiologic data. The table can be used to determine sample
size requirements.
4.2.2 Two loci

A power approximation for test (4.1) for two equally contributing loci can be developed in the same way. The alternative hypothesis can be stated as follows. In a proportion \( \alpha_1 \) of the pairs locus \( r_1 \) is increasing susceptibility and in a proportion \( \alpha_2 \), locus \( r_2 \) is increasing susceptibility with \( \alpha_1 + \alpha_2 \leq 1 \). For \( \xi_1 = \xi_2 \) and both loci \( r_1 \) and \( r_2 \) corresponding to markers, the power for the 2 loci test can be shown to be

\[
P_{\xi,r_1,r_2} \left\{ \sigma^{-1} \left[ \max_{0 \leq i \leq \Delta \leq !} Z_1(i) + \max_{0 \leq i \leq \Delta \leq !} Z_2(j) \right] > b \sqrt{2} \right\} \approx 1 - \Phi \left( b - \sqrt{2} \frac{\xi}{\sigma} \right)
\]

\[
+ \phi \left( b - \sqrt{2} \frac{\xi}{\sigma} \right) \left[ \frac{2b^2 \nu^2}{(\frac{\xi}{\sigma})^2} + \frac{2b^2(1 - \nu - \nu^2)\nu}{(\frac{\xi}{\sigma})} + \frac{b^4}{(\frac{\xi}{\sigma} + \frac{\nu^2}{\sqrt{2}})^2} - \frac{(2 - 8\nu + \nu^2)\nu^2}{\sqrt{2}(\frac{\xi}{\sigma} + \frac{\nu^2}{\sqrt{2}})} \right],
\]

(4.13)

where \( \nu = \nu(b\sqrt{\beta\Delta}) \) and is equal to 1 for the continuous case. For statistic (4.2), which does not assume \( \xi_1 = \xi_2 \), the following power approximation can be used.

\[
P_{\xi,r_1,r_2} \left\{ \sigma^{-1} \left[ \max_{0 \leq i \leq \Delta \leq !} Z_1^2(i) + \max_{0 \leq i \leq \Delta \leq !} Z_2^2(j) \right] > b^2 \right\} \approx 1 - \Phi \left( b - \sqrt{2} \frac{\xi}{\sigma} \right) \frac{\sigma}{2\sqrt{2} \xi} \phi \left( b - \sqrt{2} \frac{\xi}{\sigma} \right)
\]

\[
+ \left[ \frac{b\sigma}{\sqrt{2} \xi} \right]^{1/2} \phi \left( b - \sqrt{2} \frac{\xi}{\sigma} \right) \left[ \frac{2b^2 \nu^2}{(\frac{\xi}{\sigma})^2} + \frac{2b^2(1 - \nu - \nu^2)\nu}{(\frac{\xi}{\sigma})} + \frac{b^4}{(\frac{\xi}{\sigma} + \frac{\nu^2}{\sqrt{2}})^2} - \frac{(2 - 8\nu + \nu^2)\nu^2}{\sqrt{2}(\frac{\xi}{\sigma} + \frac{\nu^2}{\sqrt{2}})} \right].
\]

(4.14)

When the loci are located between markers, providing approximations requires more work. If we first assume that \( r_2 \) is at a marker and that \( t_1 < r_1 < t_2 \) with \( t_2 - t_1 = \Delta \), we can obtain the following power approximation.

\[
P_{\xi,r_1,r_2} \left\{ \sigma^{-1} \left[ \max_{0 \leq i \leq \Delta \leq !} Z_1(s) + \max_{0 \leq j \Delta \leq !} Z_2(t) \right] > b \sqrt{2} \right\}
\]

\[
\approx 1 - I_1(t_1, t_2) - I_1(t_2, t_1) + I_2(t_1, t_2) + I_2(t_2, t_1),
\]

(4.15)
where

\[ I_1(s, t) = \int_{-\frac{b}{\sqrt{2}}}^{\infty} \Phi\left(\frac{b}{\sqrt{2}} + x - \frac{\xi}{\sigma}\right) h(x, s, r_1) H_1(x, s, t, r_1) \, dx, \]

\[ I_2(s, t) = \sum_{i=1}^{2} \int_{-\frac{b}{\sqrt{2}}}^{\infty} \phi\left(\frac{b}{\sqrt{2}} + x - \frac{\xi}{\sigma}\right) h(x, t, r_1) H_i(x, s, t, r_1) f_i(x) \, dx, \]

\[ h(x, t, r) = \phi\left(\frac{b}{\sqrt{2}} - x - \rho(t, r)\right), \]

\[ \rho(t, r) = \frac{\xi}{\sigma} e^{-\beta|t-r|}, \]

\[ g(x, s, t, r) = \frac{b}{\sqrt{2}} - \rho(s, r) - e^{-\beta \Delta} \left(\frac{b}{\sqrt{2}} - x - \rho(t, r)\right), \]

\[ H_1(x, s, t, r) = \Phi\left(\frac{g(x, s, t, r) - x}{d(\Delta)}\right), \]

\[ H_2(x, s, t, r) = \Phi\left(\frac{g(x, s, t, r) - x}{d(\Delta)} - \frac{b}{\sqrt{2}} d(\Delta)\right) \exp\left[-\frac{b}{\sqrt{2}} g(x, s, t, r) + \frac{b^2 d^2(\Delta)}{4}\right], \]

\[ f_1(\nu, x) = \frac{(2\nu - 1)\nu^2}{(\frac{b}{\sqrt{2}} - x + \frac{\xi}{\sigma})} + \frac{b\sqrt{2}\nu^2}{(x - \frac{\xi}{\sigma})^2} + \frac{(\nu^2 + 2\nu - 3)\nu}{(x - \frac{\xi}{\sigma})}, \]

and

\[ f_2(\nu, x) = \left[\frac{(3\nu^3 + 2\nu^2 - \nu)}{(x - \frac{\xi}{\sigma})} - \frac{(\nu^2 - 6\nu + 1)\nu^2}{(\frac{b}{\sqrt{2}} - x + \frac{\xi}{\sigma})^2} + \frac{b\sqrt{2}\nu^2}{(x - \frac{\xi}{\sigma})^2} + \frac{b\sqrt{2}\nu^4}{(\frac{b}{\sqrt{2}} - x + \frac{\xi}{\sigma})^2}\right] \exp\left[\frac{b}{\sqrt{2}} x\right]. \]

This messy expression was shown to be quite accurate through a simulation study.

We can take this a step further by assuming that both loci are between markers i.e.,

\[ t_1 < r_1 < t_2 \text{ and } s_1 < r_2 < s_2 \text{ with } t_2 - t_1 = s_2 - s_1 = \Delta, \]

Then
CHAPTER 4. POLYGENIC DISEASES

\[ P_{\xi_1, \xi_2} \left\{ \sigma^{-1} \left[ \max_{0 \leq s \Delta \leq 1} Z_1(s) + \max_{0 \leq j \Delta \leq 1} Z_2(t) \right] > b \sqrt{2} \right\} \approx 1 - I_1(s_1, s_2, t_1, t_2) - I_1(s_2, s_1, t_1, t_2) + I_2(s_1, s_2, t_1, t_2) + I_2(s_2, s_1, t_1, t_2) + I_2(s_1, s_2, t_2, t_1) + I_2(s_2, s_1, t_2, t_1), \] (4.16)

where

\[ I_1(s, t, u, v) = \int_{-\frac{b}{\sqrt{2}} - \infty}^{\infty} \int_{-\frac{b}{\sqrt{2}} - \infty}^{\infty} h(x, s, r_1) h(y, u, r_2) H_1(x, t, s, r_1) H_3(x, y, u, v, r_2) \, dy \, dx, \]

\[ I_2(s, t, u, v) = \sum_{j=1}^{2} \sum_{i=1}^{2} \int_{-\frac{b}{\sqrt{2}} - \infty}^{\infty} \int_{-\frac{b}{\sqrt{2}} - \infty}^{\infty} f_{ij}(x, y) h(x, s, r_1) h(y, u, r_2) H_i(x, s, t, r_1) H_j(x, y, u, v, r_2) \, dy \, dx, \]

\( h(x, t, r), \ \rho(t, r), \ g(x, s, t, r), \ \delta(\Delta), \ H_1(x, s, t, r), \ H_2(x, s, t, r) \) were previously defined and

\[ H_3(x, y, s, t, r) = \Phi \left( \frac{g(y, t, s, r) + x}{d(\Delta)} \right), \]

\[ f_{11}(x, y) = \left[ (2 - 2\nu + \nu^2) \nu + \frac{b}{\sqrt{2}} (x + y) \nu^2 \right] \exp \left[ -\frac{b}{\sqrt{2}} (x + y) \right], \]

\[ f_{12}(x, y) = \exp \left[ -\frac{b}{\sqrt{2}} x \right] \nu - \exp \left[ -\frac{b}{\sqrt{2}} (2x + y) \right] \nu^2 + \frac{b}{\sqrt{2}} (x + y) \exp \left[ -\frac{b}{\sqrt{2}} x \right] \nu^2, \]

\[ f_{21}(x, y) = \exp \left[ -\frac{b}{\sqrt{2}} y \right] (1 - \nu^2) \nu - \exp \left[ -\frac{b}{\sqrt{2}} (x + 2y) \right] \nu^2 + \frac{b}{\sqrt{2}} (x + y) \exp \left[ -\frac{b}{\sqrt{2}} y \right] \nu^2, \]

\[ f_{22}(x, y) = \frac{b}{\sqrt{2}} (x + y) \nu^2. \]

For the map of \( \Delta = 1 \) centimorgans, this approximation is equivalent to the simpler equation (4.13) in which it is assumed that the loci are located at markers. However,
for $\Delta$ as small as 5 cM, there are cases where the simpler equation will noticeably overestimate the power and hence underestimate the sample size needed for test (4.1) to reach a specific power value. For example, for half-sibling pairs with $\xi = 3.69$, equation (4.13) would predict a power of 75.1% while the more complicated equation (4.15) would predict a power of 67.5%, which is much closer to the simulated power of 68.6%.

4.2.3 Three loci

The complexity of the power approximations increase with the number of loci. For the three loci, we only present the continuous map case. The alternative is that there exit three gene loci, $r_1, r_2$ and $r_3$ each increasing susceptibility in a proportion $\alpha$ of the pairs. Then,

$$
P_{\xi,r_1,r_2,r_3}\left[\sigma^{-1}\left[\max_{0 \leq s \leq 1} Z_1(s) + \max_{0 \leq t \leq 1} Z_2(t) + \max_{0 \leq u \leq 1} Z_3(u) > b \sqrt{3}\right]\right] \approx 1 - \Phi\left(b - \sqrt{3} \frac{\xi}{\sigma}\right)$$

$$+ \Phi\left(b - \sqrt{3} \frac{\xi}{\sigma}\right)\left[\frac{8b^2}{(\frac{\sqrt{3} \xi}{\sigma})^3} - \frac{4b^2}{(b + \frac{\sqrt{3} \xi}{\sigma})^3} - \frac{16b}{(\frac{\sqrt{3} \xi}{\sigma})^2} - \frac{14b}{(b + \frac{\sqrt{3} \xi}{\sigma})^2} + \frac{32}{\frac{\sqrt{3} \xi}{\sigma}} - \frac{31}{b + \frac{\sqrt{3} \xi}{\sigma}}\right]. \quad (4.17)$$

4.3 Simulation study

We include a simulation study to evaluate the accuracy of the power approximations provided in section 4.2. The rejection regions were defined using the methods from section 4.1. Simulations for the data were done just as described for the null hypothesis distribution except that the mean function $\sigma^{-1}\xi e^{-\beta|t-r|}$ was added to the process on chromosomes containing a gene locus. Table 4.5 presents the results for the one locus approximation, for the case where the gene locus is mid-way between
markers. Table 4.6 contains the results for the two loci test (4.1) for the continuous IBD data or the discrete data with the loci located at markers. Table 4.7 is again for the test (4.1) with one of the loci located mid-way between markers. Table 4.8 is for the case where both loci are located between markers. The simulation results agree quite nicely with the approximations from the previous section.

<table>
<thead>
<tr>
<th></th>
<th>Δ</th>
<th>20</th>
<th>20</th>
<th>10</th>
<th>10</th>
<th>5</th>
<th>5</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>5.39</td>
<td>92.3</td>
<td>92.4</td>
<td>95.4</td>
<td>95.1</td>
<td>96.3</td>
<td>95.9</td>
<td>96.6</td>
<td>96.5</td>
<td>96.7</td>
<td>96.6</td>
</tr>
<tr>
<td></td>
<td>5.02</td>
<td>86.6</td>
<td>86.6</td>
<td>91.0</td>
<td>90.4</td>
<td>92.4</td>
<td>92.4</td>
<td>92.9</td>
<td>93.3</td>
<td>93.0</td>
<td>93.7</td>
</tr>
<tr>
<td>H-S</td>
<td>4.39</td>
<td>71.4</td>
<td>70.4</td>
<td>77.6</td>
<td>77.9</td>
<td>79.7</td>
<td>80.2</td>
<td>80.4</td>
<td>81.8</td>
<td>80.6</td>
<td>81.5</td>
</tr>
<tr>
<td></td>
<td>5.39</td>
<td>74.2</td>
<td>73.5</td>
<td>89.3</td>
<td>88.6</td>
<td>93.3</td>
<td>93.4</td>
<td>94.8</td>
<td>94.7</td>
<td>94.9</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>5.02</td>
<td>64.7</td>
<td>63.9</td>
<td>82.1</td>
<td>81.8</td>
<td>87.7</td>
<td>87.3</td>
<td>89.8</td>
<td>90.4</td>
<td>89.9</td>
<td>90.1</td>
</tr>
<tr>
<td>Av.</td>
<td>4.39</td>
<td>47.1</td>
<td>45.3</td>
<td>64.9</td>
<td>63.2</td>
<td>71.8</td>
<td>71.9</td>
<td>74.6</td>
<td>74.7</td>
<td>74.7</td>
<td>76.6</td>
</tr>
<tr>
<td></td>
<td>5.39</td>
<td>61.9</td>
<td>61.5</td>
<td>85.0</td>
<td>84.2</td>
<td>91.6</td>
<td>91.2</td>
<td>94.1</td>
<td>94.3</td>
<td>94.3</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>5.02</td>
<td>52.3</td>
<td>51.9</td>
<td>76.7</td>
<td>76.0</td>
<td>85.2</td>
<td>85.0</td>
<td>88.6</td>
<td>89.2</td>
<td>89.0</td>
<td>89.3</td>
</tr>
<tr>
<td>C-I</td>
<td>4.39</td>
<td>36.6</td>
<td>34.0</td>
<td>58.4</td>
<td>58.4</td>
<td>68.2</td>
<td>67.9</td>
<td>72.6</td>
<td>73.0</td>
<td>73.0</td>
<td>73.6</td>
</tr>
<tr>
<td></td>
<td>5.39</td>
<td>57.7</td>
<td>58.6</td>
<td>83.5</td>
<td>82.9</td>
<td>91.0</td>
<td>90.7</td>
<td>93.9</td>
<td>94.3</td>
<td>94.1</td>
<td>94.3</td>
</tr>
<tr>
<td></td>
<td>5.02</td>
<td>48.3</td>
<td>48.0</td>
<td>74.9</td>
<td>74.1</td>
<td>84.2</td>
<td>84.6</td>
<td>88.3</td>
<td>88.6</td>
<td>88.5</td>
<td>88.7</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>33.6</td>
<td>32.6</td>
<td>56.4</td>
<td>55.8</td>
<td>66.9</td>
<td>65.9</td>
<td>72.0</td>
<td>72.6</td>
<td>72.3</td>
<td>74.0</td>
</tr>
</tbody>
</table>

Table 4.5: Power for one locus test with type-1 error of 0.05. Δ is the distance between markers in centimorgans. App. stands for approximation. Sim. correspond to the simulation threshold values. The type of relative pairs studied are: grandparent-grandchild (G-G), half-siblings (H-S), Avuncular (Av.) and first cousins (C-I). The alternative is that there is one locus of effect ξ located mid-way between markers.
## Table 4.6: Power for 2-locus test (4.1) and type-1 error of 0.05. $\Delta$ is the distance between markers in centimorgans. The continuous observation case is represented by $\Delta = 0$. App. stands for approximation. Sim. correspond to the simulation threshold values. The type of relative pairs studied are: grandparent-grandchild (G-G), half-siblings (H-S), Avuncular (Av.) and first cousins (C-I). The alternative is that there are two equally contributing loci of effect $\xi$ located at markers.
### Table 4.7: Power for 2-locus test (4.1) and type-1 error of 0.05. One locus is located at a marker, the other midway between markers.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$\xi/\sigma$</th>
<th>20</th>
<th>20</th>
<th>10</th>
<th>10</th>
<th>5</th>
<th>5</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>5.02</td>
<td>99.5</td>
<td>99.5</td>
<td>99.6</td>
<td>99.6</td>
<td>99.6</td>
<td>99.7</td>
<td>99.6</td>
<td>99.6</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>95.9</td>
<td>96.1</td>
<td>96.5</td>
<td>96.0</td>
<td>96.6</td>
<td>96.6</td>
<td>96.5</td>
<td>96.8</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>79.9</td>
<td>79.2</td>
<td>81.3</td>
<td>81.8</td>
<td>81.4</td>
<td>81.5</td>
<td>81.3</td>
<td>82.1</td>
</tr>
<tr>
<td>H-S</td>
<td>5.02</td>
<td>97.8</td>
<td>97.9</td>
<td>98.8</td>
<td>98.9</td>
<td>99.1</td>
<td>99.1</td>
<td>99.1</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>89.4</td>
<td>89.6</td>
<td>92.6</td>
<td>93.1</td>
<td>93.5</td>
<td>93.7</td>
<td>93.5</td>
<td>94.3</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>65.3</td>
<td>65.0</td>
<td>70.5</td>
<td>70.5</td>
<td>72.0</td>
<td>72.3</td>
<td>71.8</td>
<td>74.0</td>
</tr>
<tr>
<td>Av.</td>
<td>5.02</td>
<td>96.4</td>
<td>96.8</td>
<td>98.3</td>
<td>98.4</td>
<td>98.7</td>
<td>98.9</td>
<td>98.9</td>
<td>99.0</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>85.2</td>
<td>85.9</td>
<td>90.5</td>
<td>90.6</td>
<td>91.9</td>
<td>91.9</td>
<td>92.2</td>
<td>93.1</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>58.6</td>
<td>58.2</td>
<td>66.0</td>
<td>66.2</td>
<td>68.3</td>
<td>68.4</td>
<td>68.5</td>
<td>70.6</td>
</tr>
<tr>
<td>C-I</td>
<td>5.02</td>
<td>95.8</td>
<td>96.1</td>
<td>98.1</td>
<td>97.9</td>
<td>98.6</td>
<td>98.7</td>
<td>98.8</td>
<td>98.8</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>83.8</td>
<td>83.9</td>
<td>89.6</td>
<td>89.7</td>
<td>91.4</td>
<td>91.4</td>
<td>91.8</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>56.6</td>
<td>57.6</td>
<td>64.4</td>
<td>64.7</td>
<td>67.1</td>
<td>67.5</td>
<td>67.4</td>
<td>68.5</td>
</tr>
</tbody>
</table>

### Table 4.8: Power for 2-locus test (4.1) and type-1 error of 0.05. Both loci are located midway between markers.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>$\xi/\sigma$</th>
<th>20</th>
<th>20</th>
<th>10</th>
<th>10</th>
<th>5</th>
<th>5</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-G</td>
<td>5.02</td>
<td>98.5</td>
<td>98.5</td>
<td>99.3</td>
<td>99.3</td>
<td>99.5</td>
<td>99.6</td>
<td>99.6</td>
<td>99.6</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>92.0</td>
<td>92.0</td>
<td>95.2</td>
<td>95.2</td>
<td>96.0</td>
<td>96.3</td>
<td>96.4</td>
<td>96.8</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>71.1</td>
<td>71.7</td>
<td>77.5</td>
<td>78.1</td>
<td>79.2</td>
<td>80.8</td>
<td>80.3</td>
<td>82.2</td>
</tr>
<tr>
<td>H-S</td>
<td>5.02</td>
<td>87.5</td>
<td>87.3</td>
<td>97.0</td>
<td>96.8</td>
<td>98.5</td>
<td>98.6</td>
<td>99.0</td>
<td>99.1</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>69.0</td>
<td>68.9</td>
<td>86.6</td>
<td>86.5</td>
<td>93.5</td>
<td>91.9</td>
<td>93.0</td>
<td>93.8</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>41.0</td>
<td>41.4</td>
<td>60.2</td>
<td>60.3</td>
<td>67.5</td>
<td>68.6</td>
<td>70.0</td>
<td>73.6</td>
</tr>
<tr>
<td>Av.</td>
<td>5.02</td>
<td>75.4</td>
<td>75.9</td>
<td>94.6</td>
<td>94.5</td>
<td>97.8</td>
<td>97.7</td>
<td>98.7</td>
<td>98.9</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>53.4</td>
<td>54.2</td>
<td>80.6</td>
<td>81.5</td>
<td>88.6</td>
<td>89.1</td>
<td>91.5</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>28.3</td>
<td>28.5</td>
<td>51.9</td>
<td>53.0</td>
<td>62.2</td>
<td>64.4</td>
<td>66.6</td>
<td>70.4</td>
</tr>
<tr>
<td>C-I</td>
<td>5.02</td>
<td>70.6</td>
<td>71.7</td>
<td>93.5</td>
<td>93.8</td>
<td>97.5</td>
<td>97.8</td>
<td>98.6</td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>48.3</td>
<td>50.1</td>
<td>78.2</td>
<td>78.7</td>
<td>87.6</td>
<td>88.4</td>
<td>91.0</td>
<td>92.5</td>
</tr>
<tr>
<td></td>
<td>3.69</td>
<td>24.8</td>
<td>25.3</td>
<td>49.0</td>
<td>48.9</td>
<td>60.5</td>
<td>61.9</td>
<td>65.4</td>
<td>69.4</td>
</tr>
</tbody>
</table>
Chapter 5

Interval Mapping

When the identity-by-descent (IBD) data come from a dense set of markers, Lander and Botstein (1986) suggest a way of interpolating between markers that they call "interval mapping." The purpose of interval mapping is to increase the likelihood of detecting a gene which is located between two markers. We begin by presenting the method of interval mapping. Then we describe how interval mapping can be used with the IBD data from affected relative pairs. We use interval mapping with both the original Markov Chain models (Feingold (1993)) and with their Gaussian approximations (Feingold et al. (1993)). We include the results from a simulation study that show that the gain in power from using interval mapping, in the context of IBD from affected relative pairs, is not substantial.

5.1 The Method

In classical linkage analysis, a marker is selected and tested to see if it is linked to a particular disease. This is done by examining how many times the marker is inherited with the disease. When the marker and the disease are not inherited together, they are said to have recombined. If the disease locus and the marker are unlinked, we expect
them to recombine by chance about half of the time and therefore the recombination fraction $\theta$, is equal to $1/2$. When they are linked, they will recombine less often, i.e., $\theta < 1/2$. To test whether a specific marker, say marker A, is linked to a disease gene G,

$$\frac{\text{lik}(\hat{\theta})}{\text{lik}(1/2)}$$

is used, where $\text{lik}(\theta)$ is the likelihood that the marker A and the disease gene G lie at a recombination fraction $\theta$ of each other, and $\hat{\theta}$ is the maximum likelihood estimate of $\theta$, based on the data. If this ratio is small and the marker is found to be unlinked to the disease, another marker is selected and the process of testing is repeated.

If the disease and the marker are "loosely" linked i.e., $\theta$ is large, this method may fail to detect linkage if not enough individuals are available. To improve the odds of making the correct decision, Lander and Botstein (1986) suggest testing linkage to an interval between two linked markers by comparing the hypothesis of the disease locus being located between the two markers to the hypothesis of no linkage. They propose using the likelihood ratio of these two hypotheses to test for linkage and they call this new method of testing interval mapping. In principal, we could use the likelihood of the gene being located anywhere between the two markers. In practice, we use the likelihood of the gene being located mid-way between the markers.

![Diagram](image)

Figure 5.1: Interval Mapping: Disease gene G is in the middle of markers A and B.

Lander and Botstein (1986) compare interval mapping to having a "virtual" marker that is tightly linked to all loci in the interval. Their reasoning is as follows. If there exits a disease locus D, located between two markers A and B (see
Figure 5.1), the disease locus will recombine from the interval if and only if two recombination events occur: a recombination between A and G and a recombination between G and B. If G is located midway between A and B, the probability of this happening is $\theta^2$, which is much lower than the probability of recombination ($\theta$) to one marker alone. Therefore, interval mapping tests for much tighter linkage than the classical one marker at a time method. However, with interval mapping, individuals in whom A and B have recombined do not provide any linkage information, since in this case G will be inherited with A or B with probability 1/2 (if G is indeed mid-way between A and B). This will decrease the sample size of informative individuals, but if A and B are close together, this reduction will not be large. Lander and Botstein claim that using interval mapping greatly reduces the number of individuals needed to detect linkage. In the next section, we will describe how interval mapping can be used with the IBD data to improve on the original test proposed in Chapter 3.

5.2 Interval Mapping for the Exact Process

We will first derive the likelihood ratio statistic for testing linkage at a specific marker (or location), then we will derive the likelihood ratio statistic for linkage to the midpoint of an interval i.e., the interval mapping interpolation.

We observe the process $X_i(t)$ for each relative pair $i, i = 1, \ldots, N$, at discrete intervals of length $\Delta$ centimorgans, where $X_i(t)$ is the indicator of the pair $i$ being IBD at location $t = k\Delta$ of the genome. Define $Y_1(t) = \sum_{i=1}^{N} X_i(t)$, the number of pairs IBD at location $t = k\Delta$, and $Y_0(t) = N - Y_1(t)$. We will derive the interval mapping interpolation only when $X_i(t)$ is an exact Markov Chain i.e., when the data are from grandparent-grandchild pairs or half-siblings pairs. It is the simpler case and gives a good illustration of how interval mapping works.

Under $H_0$, the hypothesis of no linkage, $X_i(t)$ has a Bernoulli($p$) distribution and
CHAPTER 5. INTERVAL MAPPING

$Y_1(t)$ has a binomial$(N, p)$ distribution, where $p$ is the probability of the relative pairs to be IBD at any given locus. For both grandparent-grandchild and half-sibling pairs, $p = 1/2$.

Under $H_1$, the hypothesis that there exists a gene locus at location $t = k\Delta$ that is responsible for the disease in a proportion $\alpha$ of the pairs, we have

$$P(X_i(t) = 1) = \alpha + (1 - \alpha)\frac{1}{2} = \frac{1}{2}(1 + \alpha), \quad P(X_i(t) = 0) = \frac{1}{2}(1 - \alpha).$$

Hence, under $H_1$, $Y_1(t)$ has a binomial$(N, \frac{1}{2}(1 + \alpha))$ distribution.

We can write the likelihood ratio statistic as

$$(1 + \alpha)^{Y_1(t)} (1 - \alpha)^{Y_0(t)}.$$

Maximizing over all possible values of $\alpha$ we get

$$\left[2LR(t)\right]^{1/2} = \left[2Y_1(t) \ln \left(\frac{2Y_1(t)}{N}\right) + 2Y_0(t) \ln \left(\frac{2Y_0(t)}{N}\right)\right]^{1/2}, \quad (5.1)$$

where $\ln$ is the natural log and $LR$ is the log likelihood ratio. Maximizing 5.1 over all values of $t = k\Delta$ (all markers) would give us the maximum likelihood ratio statistic for testing for the presence of a disease susceptibility gene under the assumption that the true locus is located at a marker. Note that this test is equivalent to Feingold's (1993) suggestion of using $\max_k Y_1(k\Delta)$. Therefore, we can use the approximations from Chapter 3, with slight modifications, to define a threshold, i.e., a value that the statistic $\max_k \left[2LR(k\Delta)\right]^{1/2}$ must exceed in order to be statistically significant.

The above test has maximum power when the gene locus is located very close to a marker. However, if the true gene locus is between markers, interval mapping might improve on the above test. Interval mapping compares the likelihood of the gene locus being located mid-way between markers to the likelihood of the gene locus being unlinked to the interval. We will derive these likelihoods below.

For $k\Delta < t < (k + 1)\Delta$, let $X_i(\delta_1, \delta_2)(t) = 1 \left(X_i(k\Delta) = \delta_1, X_i((k + 1)\Delta) = \delta_2\right)$ and $Y(\delta_1, \delta_2)(t) = \sum_i X_i(\delta_1, \delta_2)(t), i = 1, \ldots, N$ and $\delta_1, \delta_2 = 0, 1$. For example, $X_i(00)(t)$ is
the indicator that both markers flanking \( t \) are not IBD for pair \( i \). Note that \( X_{i(t_1,t_2)}(t) \) is constant in the intervals \((k\Delta, (k+1)\Delta)\).

Under the hypothesis of no linkage,

\[
P(X_{i(00)}(t) = 1) = P(X_{i(11)}(t) = 1) = \frac{1}{4}(1 + e^{-\beta\Delta}),
\]

\[
P(X_{i(01)}(t) = 1) = P(X_{i(10)}(t) = 1) = \frac{1}{4}(1 - e^{-\beta\Delta}).
\]

When the \( N \) pairs of relatives are independent, \((Y_{00}(t), Y_{01}(t), Y_{10}(t), Y_{11}(t))\) follows a multinomial distribution with the above proportions.

Under the hypothesis that there exists a disease locus located at \( t = (k + 1/2)\Delta \),

\[
P(X_{i(00)}(t) = 1) = \frac{1}{4}(1 + e^{-\beta\Delta}) - \frac{\alpha}{2} e^{-\beta\Delta/2},
\]

\[
P(X_{i(11)}(t) = 1) = \frac{1}{4}(1 + e^{-\beta\Delta}) + \frac{\alpha}{2} e^{-\beta\Delta/2},
\]

\[
P(X_{i(01)}(t) = 1) = P(X_{i(10)}(t) = 1) = \frac{1}{4}(1 - e^{-\beta\Delta}),
\]

and \((Y_{00}(t), Y_{01}(t), Y_{10}(t), Y_{11}(t))\) has a multinomial distribution with these new proportions.

The likelihood ratio of the above two hypotheses can be written as

\[
\left(1 - \frac{2\alpha e^{-\beta\Delta/2}}{1 + e^{-\beta\Delta}}\right) Y_{00}(t) \left(1 + \frac{2\alpha e^{-\beta\Delta/2}}{1 + e^{-\beta\Delta}}\right) Y_{11}(t).
\]

Maximizing over \( \alpha \), we get the following log likelihood ratio statistic:

\[
[2LR_{im}(t)]^{1/2} = \left[2Y_{11}(t) \ln \left(\frac{2Y_{11}(t)}{N}\right) + 2Y_{00}(t) \ln \left(\frac{2Y_{00}(t)}{N}\right)\right]^{1/2}.
\]  \((5.2)\)

When \( t \) is not the mid-point between two markers, the maximization of the likelihood ratio over \( \alpha \) has to be performed numerically. Since the case where the gene is exactly mid-way between markers is the worst case scenario, we are unlikely to be gaining much more power from maximizing over all possible values of \( t \). If the locus is located
CHAPTER 5. INTERVAL MAPPING

at a marker of mid-way between markers, we can combined statistics (5.1) and (5.2) to get a new likelihood ratio,

\[
\max_k \max \left[ \left( 2LR(k\Delta) \right)^{1/2}, \left( 2LR_{im}((k + 1/2)\Delta) \right)^{1/2} \right].
\] (5.3)

Since (5.3) \(\geq\) (5.1), the thresholds used for 5.3 will be greater than the ones for \(\max_k \left[ 2LR(k\Delta) \right]^{1/2}\).

We will compare the power using the interval mapping statistic (5.3) when the gene locus is mid-way between markers to (i) the test without interval mapping and (ii) the test with twice as many markers i.e., with true markers instead of interpolated values between markers. We did not include simulations to compare the power of the tests when the gene is at a marker since it is clear intuitively that statistic (5.1) without interval mapping is more powerful in that situation because of the lower threshold used. Since Lander and Botstein claimed that interval mapping was like having a "virtual" marker mid-way in the interval between markers, we would hope that the power from (5.3) and (ii) is comparable and that both are more powerful that situation (i). Tables 5.1, 5.2, and 5.3 present the simulation results from \(N = 100\) grandparent-grandchild pairs, and Tables 5.4, 5.5 and 5.6 are for the half-sibling case. We chose \(\xi/\sigma = 5.39, 5.02, 4.39, 3.69\) because these parameter values provide powers of 95%, 90%, 80% and 50% when using half-sibling pairs and continuous data.

To get the thresholds that would define tests with type-1 error of 5%, we simulated the process under \(H_0\). Each process \(X_i(t)\) was simulated as follows.

(i) Set \(X_{i0} = 0\) or 1 with probability 1/2.

(ii) Draw a random exponential random variable with rate \(\beta/2\) to determine the location where the process \(X_i(t)\) switches states.

(iii) Repeat step (ii) until the end of the chromosome is reached.

Even though it would have been more efficient to simulate the process for the sum of \(N\) relative pairs, it was more convenient to do it one pair at a time in order
to use the computer programs from Chapter 8. The value of $\beta$ used is 0.02 for the grandparent-grandchild pairs and 0.04 for the half-sibling pairs. All three steps were repeated 23 times (23 chromosomes of length 140 centimorgans) for each relative pair. The likelihood ratio statistic can be calculated from the $X_i(t)$ processes, and the 95% quantile provides the threshold.

To obtain the power, we simulated the process under $H_1$ and then calculated the proportion of time the threshold was exceeded by the test statistic.

Under $H_1$, when the gene locus is located at $r$, the first step is replaced by

(i) Set $X_{ir} = 0$ or 1 with probability $\frac{1}{2}(1 - \alpha)$ and $\frac{1}{2}(1 + \alpha)$ respectively.

Then step (ii) and (iii) are performed on each side of the locus $r$.

Since the process $X_i(t)$ is discrete, we cannot choose thresholds that would guarantee that the tests have a type-1 error of exactly 0.05. To make a fair comparison of the power from the different methods, we used a randomized tests to make all type-1 error approximately equal to 0.05. Because of the random nature of the tests, the threshold for the tests are not included. However, the reader is referred to Tables 5.3 to 5.3 for the Gaussian approximation case to see how the thresholds for the tests with and without interval mapping compare. The increase in the threshold values for statistic (5.3) to account for the interval mapping is quite small.

From the tables we can conclude that even though the power of the test with interval mapping is superior to the test without, it is far from being equivalent to having an extra marker at the midpoint of the interval. For example, for the half-sibling pairs, $\Delta = 20$ and $\xi/\sigma = 5.39$, the power from the test without interval mapping is 0.737, interval mapping increases it to 0.795 but a true marker at the midpoint of the intervals gives a power of 0.967.

Lander and Botstein predicted a huge increase in power from using interval mapping based on the following. They compared the expected value of the lod score at the mid-point ($LR_{im}((k + 1/2)\Delta)$) with the expected values of the lod scores at the
flanking markers \( E[LR(k\Delta)] \) and \( E[LR((k+1)\Delta)] \). However, a fairer approach would be to compare \( LR_{im}((k+1/2)\Delta) \) to the expected value of the maximum lod score at the two flanking markers i.e., \( E[\max(LR(k\Delta), LR((k+1)\Delta))] \), since the latter is the statistic being used in the absence of interval mapping.

Lander and Botstein based their conclusion that interval mapping greatly improves the power of a test on the substantial difference between \( LR_{im}((k+1/2)\Delta) \) and \( E[LR(k\Delta)] \) (or \( E[LR((k+1)\Delta)] \)). However, the more appropriate difference between \( LR_{im}((k+1/2)\Delta) \) and \( E[\max(LR(k\Delta), LR((k+1)\Delta))] \) is much smaller and explains the low increase in powers from our simulations.

**Grandparent-grandchild pairs**

<table>
<thead>
<tr>
<th></th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \xi/\sigma = 5.39 )</td>
<td>0.948</td>
<td>0.963</td>
<td>0.989</td>
</tr>
<tr>
<td>( \xi/\sigma = 5.02 )</td>
<td>0.895</td>
<td>0.918</td>
<td>0.966</td>
</tr>
<tr>
<td>( \xi/\sigma = 4.39 )</td>
<td>0.740</td>
<td>0.778</td>
<td>0.863</td>
</tr>
<tr>
<td>( \xi/\sigma = 3.69 )</td>
<td>0.500</td>
<td>0.534</td>
<td>0.637</td>
</tr>
</tbody>
</table>

Table 5.1: Power values for 100 grandparent-grandchild pairs for (i) the test without interval mapping, the test using interval mapping and (ii) the test with twice as many markers. Markers are \( \Delta = 20 \) centimorgans apart.

<table>
<thead>
<tr>
<th></th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \xi/\sigma = 5.39 )</td>
<td>0.977</td>
<td>0.977</td>
<td>0.986</td>
</tr>
<tr>
<td>( \xi/\sigma = 5.02 )</td>
<td>0.944</td>
<td>0.945</td>
<td>0.958</td>
</tr>
<tr>
<td>( \xi/\sigma = 4.39 )</td>
<td>0.815</td>
<td>0.817</td>
<td>0.848</td>
</tr>
<tr>
<td>( \xi/\sigma = 3.69 )</td>
<td>0.581</td>
<td>0.583</td>
<td>0.616</td>
</tr>
</tbody>
</table>

Table 5.2: Power values for 100 grandparent-grandchild pairs when \( \Delta = 10 \) centimorgans.
### Table 5.3: Power values for 100 grandparent-grandchild pairs when $\Delta = 5$ centimorgans.

<table>
<thead>
<tr>
<th>$\xi/\sigma$</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>0.979</td>
<td>0.980</td>
<td>0.986</td>
</tr>
<tr>
<td>5.02</td>
<td>0.947</td>
<td>0.952</td>
<td>0.963</td>
</tr>
<tr>
<td>4.39</td>
<td>0.830</td>
<td>0.837</td>
<td>0.855</td>
</tr>
<tr>
<td>3.69</td>
<td>0.598</td>
<td>0.604</td>
<td>0.621</td>
</tr>
</tbody>
</table>

### Half-sibling pairs

<table>
<thead>
<tr>
<th>$\xi/\sigma$</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>0.772</td>
<td>0.830</td>
<td>0.985</td>
</tr>
<tr>
<td>5.02</td>
<td>0.674</td>
<td>0.739</td>
<td>0.956</td>
</tr>
<tr>
<td>4.39</td>
<td>0.478</td>
<td>0.540</td>
<td>0.832</td>
</tr>
<tr>
<td>3.69</td>
<td>0.274</td>
<td>0.320</td>
<td>0.588</td>
</tr>
</tbody>
</table>

### Table 5.4: Power values for 100 half-sibling pairs when $\Delta = 20$ centimorgans.

<table>
<thead>
<tr>
<th>$\xi/\sigma$</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>0.924</td>
<td>0.940</td>
<td>0.977</td>
</tr>
<tr>
<td>5.02</td>
<td>0.854</td>
<td>0.875</td>
<td>0.943</td>
</tr>
<tr>
<td>4.39</td>
<td>0.673</td>
<td>0.704</td>
<td>0.817</td>
</tr>
<tr>
<td>3.69</td>
<td>0.426</td>
<td>0.455</td>
<td>0.563</td>
</tr>
</tbody>
</table>

### Table 5.5: Power values for 100 half-sibling pairs when $\Delta = 10$ centimorgans.
### Table 5.6: Power values for 100 half-sibling pairs when $\Delta = 5$ centimorgans.

<table>
<thead>
<tr>
<th>$\xi/\sigma$</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>0.958</td>
<td>0.963</td>
<td>0.976</td>
</tr>
<tr>
<td>5.02</td>
<td>0.905</td>
<td>0.912</td>
<td>0.938</td>
</tr>
<tr>
<td>4.39</td>
<td>0.750</td>
<td>0.762</td>
<td>0.801</td>
</tr>
<tr>
<td>3.69</td>
<td>0.500</td>
<td>0.513</td>
<td>0.549</td>
</tr>
</tbody>
</table>

#### 5.3 Interval Mapping for the Gaussian Approximations

When the number of relative pairs that are available is sufficiently large, it may be more convenient to use the Gaussian approximations to the Markov Chain models. They are simpler to work with and sometimes generalize easily to more difficult problems. We will derive the interval mapping interpolation for the scaled data

$$Z(t) = \frac{X(t) - N_p}{\sqrt{N}},$$

using its limiting distribution as derived in Chapter 3. We observe $N$ relative pairs, but this time we are not restricting ourselves to the grandparent-grandchild or the half-sibling pairs.

The log likelihood ratio (LR) statistic for the presence of a disease locus at $k\Delta$ is (for all relative pairs except siblings)

$$\frac{\xi}{\sigma^2}Z(k\Delta) - \frac{\xi^2}{2\sigma^2}.$$

If we replace $\xi$ by its maximum likelihood estimate, $\hat{\xi} = Z(k\Delta)$, we get

$$\left[2LR(k\Delta)\right]^{1/2} = \frac{Z(k\Delta)}{\sigma}. \quad (5.4)$$

To calculate the interval mapping likelihood ratio, we need to introduce new notation.
Let
\[
\mu = \begin{pmatrix} R(\frac{1}{2}) \\ R(\frac{1}{2}) \end{pmatrix}, \quad Z(\tilde{t}) = \begin{pmatrix} Z(k\Delta) \\ Z((k+1)\Delta) \end{pmatrix}, \quad \Sigma = \begin{bmatrix} 1 & R(\Delta) \\ R(\Delta) & 1 \end{bmatrix},
\]
where \( k\Delta < t < (k+1)\Delta \) and \( R(t) \) is the covariance function of \( \sigma^{-1}Z(t) \), and is assumed to be of the form \( 1 - \beta|t| + o(|t|) \) as \( t \to 0 \). When \( t \) is the midpoint of interval \( (k\Delta, (k+1)\Delta) \), the LR is
\[
\frac{\xi}{\sigma^2} \frac{\mu' \Sigma^{-1} Z(\tilde{t})}{\mu' \Sigma^{-1} \mu} - \frac{\xi^2}{2\sigma^2} \mu' \Sigma^{-1} \mu.
\]
Replacing \( \xi \) by its maximum likelihood estimate,
\[
\frac{\mu' \Sigma^{-1} Z(\tilde{t})}{\mu' \Sigma^{-1} \mu},
\]
we obtain
\[
\left[ 2LR_{im}(t) \right]^{1/2} = \frac{\mu' \Sigma^{-1} Z(\tilde{t})}{(\mu' \Sigma^{-1} \mu)^{1/2}} = \frac{Z_{t-\frac{\Delta}{2}} + Z_{t+\frac{\Delta}{2}}}{\sigma \sqrt{2[1 + R(\Delta)]}}, \tag{5.5}
\]
for \( t = (k + 1/2)\Delta \).

Just as we did in the previous section for the exact process, we can use
\[
\max_k \max \left[ (2LR(k\Delta))^{1/2}, (2LR_{im}((k + 1/2)\Delta))^{1/2} \right]. \tag{5.6}
\]
to test for the presence of a disease susceptibility gene.

The threshold defining the critical region for the new test will be larger than the threshold for \( \max_k [2LR(k\Delta)]^{1/2} \), but not by much (see Table 5.7 through 5.12). Moreover, when the covariance is of the form \( R(t) = e^{-\beta|t|} \) we can show that the thresholds for the process using interval mapping (5.6) will be smaller than the ones for the process that has twice as many markers. This can be more formally stated by the following lemma.

**Lemma 5.1** Let \( 0 \leq t_0 < t_1 < \ldots < t_n = l \) be a partition of the interval \((0, l)\) in \( n = \frac{2l}{\Delta} \) equispaced segments of length \( \frac{\Delta}{2} \).
Let
\[ \eta_1(t_i) = \begin{cases} \left[2LR(t_i)\right]^{1/2} & \text{for } i \text{ even} \\ \left[2LR_{im}(t_i)\right]^{1/2} & \text{for } i \text{ odd} \end{cases} \]

and
\[ \eta_2(t_i) = \left[2LR(t_i)\right]^{1/2} \quad \text{for all } i. \]

If \( \xi = 0 \),
\[ P\left\{ \max_{0 \leq i \leq n} \eta_1(t_i) > b_1 \right\} = P\left\{ \max_{0 \leq i \leq n} \eta_2(t_i) > b_2 \right\} \implies b_1 \leq b_2. \]

To prove the above lemma, we will use Slepian’s Inequality (Leadbetter, Lindgren and Rootzén, 1983 p.156).

**Lemma 5.2 (“Slepian’s Lemma”).**

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

**Proof of lemma 5.1:**

If we can show that the \( \eta_i \)’s satisfy the conditions of Slepian’s lemma, we have
\[ P\left\{ \max_{0 \leq i \leq n} \eta_1(t_i) \leq u \right\} \geq P\left\{ \max_{0 \leq i \leq n} \eta_2(t_i) \leq u \right\} \]
and the lemma is proved.
Clearly, \( \eta_1 \) and \( \eta_2 \) satisfy
\[
E(\eta_1(t)) = E(\eta_1(t)) = 0 \quad \text{and} \quad E(\eta_1(t)^2) = E(\eta_1(t)^2) = 1;
\]
all we need to show is that
\[
\rho_1(t_i, t_j) \geq \rho_2(t_i, t_j) \quad \text{for all } i \text{ and } j. \quad (5.7)
\]

We will look at the above inequality for the following three cases separately: (i) \( i \) and \( j \) even, (ii) \( i \) odd, \( j \) even and (iii) \( i \) and \( j \) odd.

(i) For \( i \) and \( j \) even, \( \rho_1(t_i, t_j) = \rho_2(t_i, t_j) \) and (5.7) is satisfied.

(ii) For \( i \) odd, \( j \) even,
\[
\rho_1(t_i, t_j) = e^{-\beta |j-i|} \sqrt{\frac{1 + e^{\beta \Delta}}{2}} \geq e^{-\beta |j-i|} = \rho_2(t_i, t_j).
\]

(iii) For \( i \) and \( j \) odd,
\[
\rho_1(t_i, t_j) = e^{-\beta |j-i|} \left[ \frac{1 + e^{\beta \Delta}}{2} \right] \geq e^{-\beta |j-i|} = \rho_2(t_i, t_j),
\]
which completes the proof of the lemma.

A simulation study was conducted to evaluate the advantage of interval mapping with the Gaussian approximations of the process. Results for the grandparent-grandchild pairs and sibling pairs follow. We compare the test with interval mapping to the test without and the test with twice as many markers. We described how the Gaussian processes were simulated in Chapter 4. The thresholds used to calculate the power were obtained by simulations rather than by using the approximations of Chapter 3.

The gain in power from using interval mapping is small. The biggest gain is for half-sibling pairs with \( \Delta = 10 \text{ cm} \). For example, for \( \xi/\sigma = 5.39 \), the test without interval mapping had simulated power 0.752 and the test using interval mapping had power 0.805. However, having extra markers in the middle of the intervals increases
the power to 0.968! Interval mapping does increase the odds of detecting a gene when it is located in the middle of two markers; however, it is not as powerful as having an actual marker at the gene locus.

Grandparent-grandchild pairs

<table>
<thead>
<tr>
<th>Type-I error = 0.05</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>thresholds</td>
<td>3.408</td>
<td>3.454</td>
<td>3.538</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.39$</td>
<td>0.926</td>
<td>0.936</td>
<td>0.976</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.02$</td>
<td>0.871</td>
<td>0.888</td>
<td>0.945</td>
</tr>
<tr>
<td>$\xi/\sigma = 4.39$</td>
<td>0.718</td>
<td>0.739</td>
<td>0.835</td>
</tr>
<tr>
<td>$\xi/\sigma = 3.69$</td>
<td>0.504</td>
<td>0.529</td>
<td>0.617</td>
</tr>
</tbody>
</table>

Table 5.7: Power values for grandparent-grandchild pairs for (i) the test without interval mapping, the test using interval mapping and (ii) the test with twice as many markers. Markers are $\Delta = 20$ centimorgans apart.

<table>
<thead>
<tr>
<th>Type-I error = 0.05</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>thresholds</td>
<td>3.538</td>
<td>3.563</td>
<td>3.632</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.39$</td>
<td>0.956</td>
<td>0.956</td>
<td>0.967</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.02$</td>
<td>0.916</td>
<td>0.919</td>
<td>0.943</td>
</tr>
<tr>
<td>$\xi/\sigma = 4.39$</td>
<td>0.783</td>
<td>0.789</td>
<td>0.824</td>
</tr>
<tr>
<td>$\xi/\sigma = 3.69$</td>
<td>0.561</td>
<td>0.571</td>
<td>0.604</td>
</tr>
</tbody>
</table>

Table 5.8: Power values for grandparent-grandchild pairs when $\Delta = 10$ centimorgans.
CHAPTER 5. INTERVAL MAPPING

<table>
<thead>
<tr>
<th>Type-I error = 0.05</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>thresholds</td>
<td>3.632</td>
<td>3.644</td>
<td>3.696</td>
</tr>
<tr>
<td>$\xi / \sigma = 5.39$</td>
<td>0.970</td>
<td>0.962</td>
<td>0.973</td>
</tr>
<tr>
<td>$\xi / \sigma = 5.02$</td>
<td>0.931</td>
<td>0.932</td>
<td>0.938</td>
</tr>
<tr>
<td>$\xi / \sigma = 4.39$</td>
<td>0.808</td>
<td>0.811</td>
<td>0.820</td>
</tr>
<tr>
<td>$\xi / \sigma = 3.69$</td>
<td>0.585</td>
<td>0.587</td>
<td>0.598</td>
</tr>
</tbody>
</table>

Table 5.9: Power values for grandparent-grandchild pairs when $\Delta = 5$ centimorgans.

Half-sibling pairs

<table>
<thead>
<tr>
<th>Type-I error = 0.05</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>thresholds</td>
<td>3.439</td>
<td>3.523</td>
<td>3.586</td>
</tr>
<tr>
<td>$\xi / \sigma = 5.39$</td>
<td>0.752</td>
<td>0.805</td>
<td>0.968</td>
</tr>
<tr>
<td>$\xi / \sigma = 5.02$</td>
<td>0.650</td>
<td>0.708</td>
<td>0.926</td>
</tr>
<tr>
<td>$\xi / \sigma = 4.39$</td>
<td>0.470</td>
<td>0.531</td>
<td>0.803</td>
</tr>
<tr>
<td>$\xi / \sigma = 3.69$</td>
<td>0.280</td>
<td>0.319</td>
<td>0.560</td>
</tr>
</tbody>
</table>

Table 5.10: Power values for half-sibling pairs when $\Delta = 20$ centimorgans.

<table>
<thead>
<tr>
<th>Type-I error = 0.05</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>thresholds</td>
<td>3.586</td>
<td>3.637</td>
<td>3.707</td>
</tr>
<tr>
<td>$\xi / \sigma = 5.39$</td>
<td>0.896</td>
<td>0.907</td>
<td>0.962</td>
</tr>
<tr>
<td>$\xi / \sigma = 5.02$</td>
<td>0.820</td>
<td>0.838</td>
<td>0.925</td>
</tr>
<tr>
<td>$\xi / \sigma = 4.39$</td>
<td>0.653</td>
<td>0.678</td>
<td>0.786</td>
</tr>
<tr>
<td>$\xi / \sigma = 3.69$</td>
<td>0.416</td>
<td>0.437</td>
<td>0.547</td>
</tr>
</tbody>
</table>

Table 5.11: Power values for half-sibling pairs when $\Delta = 10$ centimorgans.
CHAPTER 5. INTERVAL MAPPING

<table>
<thead>
<tr>
<th>Type-I error = 0.05</th>
<th>(i)</th>
<th>Interval Mapping</th>
<th>(ii)</th>
</tr>
</thead>
<tbody>
<tr>
<td>thresholds</td>
<td>3.707</td>
<td>3.735</td>
<td>3.813</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.39$</td>
<td>0.936</td>
<td>0.940</td>
<td>0.957</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.02$</td>
<td>0.887</td>
<td>0.893</td>
<td>0.908</td>
</tr>
<tr>
<td>$\xi/\sigma = 4.39$</td>
<td>0.728</td>
<td>0.737</td>
<td>0.777</td>
</tr>
<tr>
<td>$\xi/\sigma = 3.69$</td>
<td>0.486</td>
<td>0.494</td>
<td>0.522</td>
</tr>
</tbody>
</table>

Table 5.12: Power values for half-sibling pairs when $\Delta = 5$ centimorgans.

With the Gaussian process, it is easy to calculate the interval mapping interpolation for any value between two markers, not just the midpoint. This could not be done in closed form for the exact Markov-Chain models. For $t_1 < t < t_2$, $t_1$ and $t_2$ corresponding to markers and $t_2 - t_1 = \Delta$, the interpolation is

$$ [2LR_{im}(t)]^{1/2} = \frac{[R(t_1 - t) - R(\Delta)R(t_2 - t)]Z_{t_1} + [R(t_2 - t) - R(\Delta)R(t_1 - t)]Z_{t_2}}{\sqrt{[1 - R^2(\Delta)][R^2(t_1 - t) - 2R(t_1 - t)R(\Delta)R(t_2 - t) + R^2(t_2 - t)]}}. $$

This expression will be useful in the next chapter, for constructing a confidence interval for the gene locus when the continuous data are not available.
Chapter 6

Confidence Sets

Once it has been determined that there is a gene influencing the trait of interest on a particular chromosome, we would like to find its exact location. The maximum likelihood estimate of the gene locus gives an indication as to where the gene might lie, but we would like to know how far around that estimate to concentrate a search. A confidence interval for the gene locus would provide such a region in which to search for the gene.

When the data come from Genomic Mismatch Scanning, i.e., a continuous map of IBD between relatives is available, Feingold et al. (1993) suggest using a confidence interval for the gene locus based on likelihood methods (Chapter 3). We will not discuss further the case of continuous data, but rather develop confidence intervals for a dense set of markers.

In classical linkage analysis, confidence intervals are constructed in the form of lod support intervals. The lod support methods can be adapted to make confidence statements about the location of the gene using Identity-by-Descent information on affected relative pairs. Yet another suitable method of making confidence statements about the location of the gene is the use of Bayesian credible sets.

The next sections will present three methods of constructing confidence sets: (i)
likelihood methods, (ii) lod support intervals and (iii) Bayesian credible sets. The advantage of the lod support method and the Bayesian credible sets is that they can be combined with interval mapping to provided confidence sets when the map of markers is less dense ($\Delta \geq 5$).

6.1 Likelihood Methods

In Chapter 3, we defined a $(1 - \gamma)$ 100% confidence interval to include all points $v$ such that

$$2\beta |R^*(t^* - v)| \exp \left[ -\frac{(Z^2 - Z_v^2)}{2\sigma^2} \right] \geq \gamma, \quad (6.1)$$

where $R(t^* - v) = Z_v/Z^*$ and $Z^* = (\max_i Z_i)_{\text{obs}}$. When the covariance function $R(t)$ is of the form $e^{-\beta|t|}$, $R'(t^* - v) = \beta R(t^* - v) = \beta Z_v/Z^*$. Otherwise, $R'(t^* - v)$ has to be found numerically.

This confidence interval was derived along the lines of Siegmund (1988), which we will sketch below. It can be shown that $A_v$, the acceptance region for the likelihood ratio test of the hypothesis that the gene locus is at location $v$, is of the form

$$A_v = \{\max_i Z_i - Z_v \leq k\}.$$

The conditional probability of $A_v$ given $Z_v$ does not depend on $\xi$, so we can choose $k$ so that

$$P(A_v|Z_v) = 1 - \gamma.$$

Therefore, the set of all values $v$ that are accepted by the likelihood ratio test is a $(1 - \gamma)$ 100% confidence set. It is not necessary to solve explicitly for $k$ since

$$(\max_i Z_i)_{\text{obs}} - Z_v \leq k \iff P(\max_i Z_i > (\max_i Z_i)_{\text{obs}}|Z_v) \geq P(\max_i Z_i > k + Z_v|Z_v) = \gamma.$$

Feingold et al. (1993) used the approximation

$$P(\max_i Z_i > Z^*|Z_v) \approx 2\beta |R'(t^* - v)| \exp \left[ -\frac{(Z^2 - Z_v^2)}{2\sigma^2} \right]$$
to obtain the confidence interval defined by (6.1).

In the case of discrete data,

$$P(\max_t Z_t > Z^*|Z_v) \approx 2 \nu (Z^* \sqrt{2\beta \Delta})\beta|R^*(t^* - v)| \exp \left[ -\frac{(Z^2 - Z^2_v)}{2\sigma^2} \right],$$

and a confidence set can be constructed (for small \(\Delta\)) as the set of all markers \(v\) such that

$$2 \nu (Z^* \sqrt{2\beta \Delta})\beta|R^*(t^* - v)| \exp \left[ -\frac{(Z^2 - Z^2_v)}{2\sigma^2} \right] \geq \gamma. \quad (6.2)$$

For large values of \(\Delta\), this method is impractical since it does not give any indication as to what happened between markers. The next sections contain methods that are better suited to a less dense map of markers. In Section 6.4, we will compare this likelihood confidence set to the one obtain with the lod support method and the Bayesian credible set, in terms of the probability of coverage and the size of the sets.

### 6.2 Lod Support Intervals

In classical genetic linkage, a marker is selected and a test is performed to determine if the marker and the disease gene are linked. The test statistic most often used is the lod score statistic, which is the log base 10 of the likelihood ratio. A lod score above 3 is evidence of linkage; a lod score below -2 is taken to show no linkage. The values of the two bounds were determined to give a power of 0.99 and a significance level of 0.01 given a sequential test i.e., the lod score is recalculated every time new data are available. The significance level was chosen very low to take into account that the prior probability of the marker and the gene being linked is about 0.05 (Elston and Lange (1975)). The location where the lod score reaches its maximum is the maximum likelihood estimate of the gene locus. A one-lod support is formed by all the points \(v\) such that

$$\text{lod}(v) \geq \max_t \text{lod}(t) - 1. \quad (6.3)$$
CHAPTER 6. CONFIDENCE SETS

Note that this set is not necessarily a single interval, but, as makes intuitive sense, could be a union of intervals. A two-lod support is constructed similarly, with 1 replaced by 2 in (6.3).

\[
\text{max(lod) - 1}
\]

\[
\text{max(lod) - 2}
\]

Figure 6.1: One and Two-Lod Support Intervals.

Figure 6.1 presents how to construct one- or two-lod support intervals from the lod score function. This same method can be used to construct a confidence set with data from affected relatives. If we work with the Gaussian approximations, the lod score, using interval mapping, can be written as (Chapter 5)

\[
\text{lod}(t) = (\log_{10} e) \frac{\left[ R(t_1 - t) - R(\Delta) R(t_2 - t) \right] Z_{t_1} + \left[ R(t_2 - t) - R(\Delta) R(t_1 - t) \right] Z_{t_2}}{\left[ 1 - R^2(\Delta) \right] \left[ R^2(t_1 - t) - 2R(t_1 - t)R(\Delta)R(t_2 - t) + R^2(t_2 - t) \right]},
\]

(6.4)

where \( t_1 \) and \( t_2 \) are the flanking markers of locus \( t \) and \( R(t) \) is the covariance function of \( Z_t \). Equation (6.4) holds for all relative pairs but siblings. Not that when \( t_1 = t \),
lod(t) reduces to \((\log_{10} e)Z_t^2/2\). An \(x\)-lod support interval would include all loci \(v\) on the chromosome such that

\[
lod(v) \geq \max_t \text{lod}(t) - x.
\]

In classical linkage analysis, a one-lod support interval has an asymptotic coverage probability greater than 95% (Ott (1991) p.67). However, when we used the lod support interval with identity-by-descent data of affected relative pairs, a one-lod support interval usually has coverage less than 95%. A larger support interval need to be used in order to achieve a 95% confidence.

Table 6.1 presents the average size in centimorgans of 1, 1.5 and 2-lod support intervals for pairs of half siblings with a map of markers 1 centimorgan apart, based on a simulation sample of 1000. The coverage probability presented is for the case where the lod support interval is constructed whether or not the test to detect linkage is statistically significant. From Table 6.1 we see that the typical 1-lod support interval has a coverage of about 89%. If a 95% confidence set is desired, a 1.5-lod support interval might be more appropriate.

<table>
<thead>
<tr>
<th>(\xi/\sigma)</th>
<th>1.0-lod support</th>
<th>1.5-lod support</th>
<th>2.0-lod support</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>5.30 (0.927)</td>
<td>8.58 (0.976)</td>
<td>12.54 (0.995)</td>
</tr>
<tr>
<td>5.02</td>
<td>6.54 (0.899)</td>
<td>10.83 (0.966)</td>
<td>16.36 (0.989)</td>
</tr>
<tr>
<td>4.39</td>
<td>9.01 (0.903)</td>
<td>16.12 (0.971)</td>
<td>25.24 (0.992)</td>
</tr>
<tr>
<td>3.69</td>
<td>15.89 (0.869)</td>
<td>30.36 (0.952)</td>
<td>45.63 (0.986)</td>
</tr>
</tbody>
</table>

Table 6.1: Unconditional average size (probability of coverage) of lod support intervals for half-siblings pairs. Markers are one centimorgan apart.

If the lod support is only constructed when the lod score exceeds a predetermined threshold, the coverage probability is slightly lower and the intervals tighter. This can be explained intuitively by the fact that when the maximum lod score is large, the
lod support interval will tend to be narrower and might exclude the true locus more often. Table 6.2 contains the average size and coverage probability for the conditional lod support interval i.e., when the interval is constructed only if the test for linkage is statistically significant.

<table>
<thead>
<tr>
<th></th>
<th>1.0-lod support</th>
<th>1.5-lod support</th>
<th>2.0-lod support</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi/\sigma = 5.39$</td>
<td>4.83 (0.926)</td>
<td>7.60 (0.975)</td>
<td>10.62 (0.995)</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.02$</td>
<td>5.41 (0.898)</td>
<td>8.73 (0.965)</td>
<td>12.53 (0.986)</td>
</tr>
<tr>
<td>$\xi/\sigma = 4.39$</td>
<td>6.27 (0.870)</td>
<td>10.25 (0.966)</td>
<td>14.71 (0.987)</td>
</tr>
<tr>
<td>$\xi/\sigma = 3.69$</td>
<td>7.04 (0.844)</td>
<td>11.73 (0.944)</td>
<td>17.28 (0.981)</td>
</tr>
</tbody>
</table>

Table 6.2: Conditional average size (probability of coverage) of lod support intervals for half-siblings pairs. Markers are one centimorgan apart.

<table>
<thead>
<tr>
<th></th>
<th>1.0-lod support</th>
<th>1.5-lod support</th>
<th>2.0-lod support</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi/\sigma = 5.39$</td>
<td>22.61 (0.947)</td>
<td>30.34 (0.984)</td>
<td>39.07 (0.996)</td>
</tr>
<tr>
<td>$\xi/\sigma = 5.02$</td>
<td>24.78 (0.960)</td>
<td>34.89 (0.988)</td>
<td>46.50 (0.997)</td>
</tr>
<tr>
<td>$\xi/\sigma = 4.39$</td>
<td>33.02 (0.957)</td>
<td>47.62 (0.991)</td>
<td>63.42 (0.998)</td>
</tr>
<tr>
<td>$\xi/\sigma = 3.69$</td>
<td>44.82 (0.957)</td>
<td>66.29 (0.989)</td>
<td>85.75 (0.994)</td>
</tr>
</tbody>
</table>

Table 6.3: Unconditional average size (probability of coverage) of lod support intervals for half-siblings pairs. Markers are 20 centimorgans apart.

For less dense maps (larger $\Delta$), the size of the lod-support intervals are wider but the coverage probabilities increase. Table 6.3 contains the size and probability of coverage of the 1, 1.5 and 2-lod support intervals based on a map with markers 20 centimorgans apart. The locus was allocated at random to the interval between 60 and 80 cM. When the gene locus is at a marker, the size of the confidence set is smaller than the size shown in Table 6.3; when the locus is mid-marker, the confidence set is larger. However, the probability of coverage was not affected (in our simulations) by the location of the gene, except when it was located very close to the end of a
chromosome, in which case the probability of coverage was larger than the values from Table 6.3. Therefore, the estimate of coverage probability presented are conservative. Note that in the case of \( \Delta = 20 \) the 1-lod support interval is approximately a 95% confidence set. The conditional lod-support intervals were much tighter, but with a decrease of 1 to 3% in the coverage probabilities (Table not included).

6.3 Bayesian Credible Sets

Yet another way to provide a confidence set for the gene locus is through the use of Bayesian highest credible sets. Zhang (1991) showed that Bayes credible sets, in the context of change points, have good frequentist properties.

A \( 100(1 - \gamma)\% \) Bayesian credible set is a region that has a posterior probability \( 1 - \gamma \). If we observe the process \( \{ Z_t \} \), we can construct a Bayesian credible set \( B_\gamma \) by including all the loci \( r \) whose posterior density exceeds \( c_\gamma \). i.e.,

\[
B_\gamma = \{ r : \pi (r|\{Z_t\}) > c_\gamma \}
\]

where \( c_\gamma \) is chosen so that

\[
\int_{B_\gamma} \pi (r|\{Z_t\}) dr = 1 - \gamma.
\]

To implement the above procedure, we need to calculate the posterior probability \( \pi (r|\{Z_t\}) \). We will take a uniform prior on the location of the gene and a normal prior (with unit variance) on the effect size \( \xi' = \xi / \sigma \). These are choices of convenience, but no others commend themselves for general use. Hence, the prior density of \( (r, \xi') \) can be written as

\[
\pi (r, \xi) = \frac{1}{l} \phi(\xi' - \theta) d\xi' dr.
\]

The parameter \( l \) is the length of the chromosome where the gene locus is suspected to be or the length of the entire genome if no prior knowledge is available.
CHAPTER 6. CONFIDENCE SETS

We first calculate the posterior density \( \pi(r, \xi|\{Z_i\}) \) and then integrate over \( \xi \) to get the posterior density we want, namely \( \pi(r|\{Z_i\}) \).

Using Bayes' rule, we obtain

\[
\pi(r, \xi|\{Z_i\}) = \frac{\text{lik}(r, \xi|\{Z_i\}) \pi(r, \xi)}{\int_0^1 \int_{-\infty}^{\infty} \text{lik}(s, y) \pi(s, y) ds dy},
\]

where

\[
\text{lik}(s, y) = \frac{dP_s, \theta(\{Z_i\})}{dP_0(\{Z_i\})} = \exp[y Z_s - y^2/2].
\]

The denominator of (6.6) can be written as

\[
\int_0^1 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[y Z_s - \frac{y^2}{2}\right] \exp\left[-\frac{1}{2}(y - \theta)^2\right] dy ds
\]

\[= (1/\sqrt{2})^{-1} \int_0^1 \exp\left[\frac{1}{4} (Z_s^2 + 2Z_s \theta - \theta^2)\right] ds,
\]

so that

\[
\pi(r, \xi|\{Z_i\}) = \frac{\frac{1}{\sqrt{2}} \exp\left[\xi Z_r - \frac{\xi^2}{2} - \frac{1}{2}(\xi - \theta)^2\right]}{\int_0^1 \exp\left[\frac{1}{4} (Z_s^2 + 2Z_s \theta - \theta^2)\right] ds}.
\]

Integrating over \( \xi \) we get

\[
\pi(r|\{Z_i\}) = \frac{\exp\left[\frac{1}{4} (Z_r^2 + 2Z_r \theta)\right]}{\int_0^1 \exp\left[\frac{1}{4} (Z_s^2 + 2Z_s \theta)\right] ds}.
\]

When the data are continuous, we can combine (6.7) with (6.5) to construct a 1 - \( \gamma \) credible set. When the data are discrete, we can use the interval mapping interpolation (6.4) between markers to construct the credible set. What value of \( \theta \) should be used? In our simulations (next section), we used \( \theta = 4 \). An effect size of 4 is one barely detectable detectable; it might not be very informative to construct a confidence set for much smaller effect sizes. We constructed the Bayesian credible sets with other values of \( \theta \) (\( \theta = 3, 4, 5 \)) as well as for an uninformative prior on the effect size. If we use value of \( \theta \) that is smaller than the true effect \( \xi/\sigma \), the method is conservative and vice versa for large \( \theta \) values. A flat prior on \( \xi/\sigma \) resulted in much wider confidence intervals. In the next section we present a simulation study in which the three methods of constructing confidence sets are compared.
6.4 Simulations

We first compare the 95% confidence sets from the likelihood method, the lod support method and the Bayesian method with normal prior (N) and with uninformative prior (F) for very dense sets of markers ($\Delta = 1cM$). We use a 1.5-lod support interval since it is the closest to a 95% confidence set for small $\Delta$. We simulated 1000 grandparent-grandchild processes using the asymptotic Ornstein-Uhlenbeck distribution and constructed the various confidence sets. We calculated the size of the confidence sets and their coverage probability (Table 6.4). We did the same for half-sibling pairs (Table 6.5). For the range of $\xi/\sigma$ in the tables, the Bayesian method with normal prior performed the best i.e, gave the shortest confidence regions and the longest confidence sets were obtained from the Bayesian method with a flat prior. Surprisingly, the likelihood method gave larger intervals than the 1.5-lod support method, which is much simpler to implement. All methods had coverage probabilities very close to the desired level of 95%.

For less dense maps of markers, we compare the 1.0 to the Bayes credible set for grandparent-grandchild pairs for $\Delta = 20$ (Table 6.6). Again, 1000 processes were generated and the average size and probability of coverage was calculated for each method. Similar results for half-sibling pairs are presented in Table 6.7. Again, the Bayes method with a normal prior seems to give narrower confidence sets and the Bayes credible sets using the flat prior were the largest, as would be expected from other Bayesian problems.
### CHAPTER 6. CONFIDENCE SETS

<table>
<thead>
<tr>
<th>( \xi/\sigma )</th>
<th>Likelihood</th>
<th>1.5-lod support</th>
<th>Bayes (F)</th>
<th>Bayes (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>14.71 (0.969)</td>
<td>15.32 (0.978)</td>
<td>16.50 (0.965)</td>
<td>16.55 (0.979)</td>
</tr>
<tr>
<td>5.02</td>
<td>18.95 (0.942)</td>
<td>18.76 (0.952)</td>
<td>20.81 (0.958)</td>
<td>19.27 (0.968)</td>
</tr>
<tr>
<td>4.39</td>
<td>28.73 (0.950)</td>
<td>25.87 (0.957)</td>
<td>30.67 (0.963)</td>
<td>24.47 (0.970)</td>
</tr>
<tr>
<td>3.69</td>
<td>47.56 (0.946)</td>
<td>41.25 (0.954)</td>
<td>48.61 (0.964)</td>
<td>32.76 (0.943)</td>
</tr>
</tbody>
</table>

Table 6.4: Unconditional average size (probability of coverage) for grandparent-grandchild pairs. Markers are one centimorgans apart.

<table>
<thead>
<tr>
<th>( \xi/\sigma )</th>
<th>Likelihood</th>
<th>1.5-lod support</th>
<th>Bayes (F)</th>
<th>Bayes (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>7.93 (0.955)</td>
<td>8.58 (0.976)</td>
<td>9.49 (0.965)</td>
<td>8.48 (0.979)</td>
</tr>
<tr>
<td>5.02</td>
<td>10.53 (0.942)</td>
<td>10.83 (0.966)</td>
<td>13.27 (0.960)</td>
<td>10.47 (0.968)</td>
</tr>
<tr>
<td>4.39</td>
<td>17.33 (0.958)</td>
<td>16.12 (0.971)</td>
<td>22.84 (0.962)</td>
<td>13.80 (0.967)</td>
</tr>
<tr>
<td>3.69</td>
<td>33.43 (0.944)</td>
<td>30.36 (0.952)</td>
<td>42.51 (0.964)</td>
<td>20.53 (0.949)</td>
</tr>
</tbody>
</table>

Table 6.5: Unconditional average size (probability of coverage) for half-sibling pairs. Markers are one centimorgan apart.

<table>
<thead>
<tr>
<th>( \xi/\sigma )</th>
<th>1.0-lod support</th>
<th>Bayes (F)</th>
<th>Bayes (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>26.67 (0.943)</td>
<td>28.71 (0.9494)</td>
<td>27.62 (0.955)</td>
</tr>
<tr>
<td>5.02</td>
<td>29.07 (0.952)</td>
<td>33.09 (0.953)</td>
<td>30.19 (0.958)</td>
</tr>
<tr>
<td>4.39</td>
<td>37.26 (0.951)</td>
<td>43.48 (0.960)</td>
<td>36.54 (0.956)</td>
</tr>
<tr>
<td>3.69</td>
<td>49.15 (0.960)</td>
<td>59.83 (0.976)</td>
<td>44.54 (0.959)</td>
</tr>
</tbody>
</table>

Table 6.6: Unconditional average size (probability of coverage) for grandparent-grandchild pairs. Markers are 20 centimorgans apart.

<table>
<thead>
<tr>
<th>( \xi/\sigma )</th>
<th>1.0-lod support</th>
<th>Bayes (F)</th>
<th>Bayes (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.39</td>
<td>22.61 (0.947)</td>
<td>25.68 (0.949)</td>
<td>21.60 (0.950)</td>
</tr>
<tr>
<td>5.02</td>
<td>24.78 (0.960)</td>
<td>30.51 (0.954)</td>
<td>23.34 (0.955)</td>
</tr>
<tr>
<td>4.39</td>
<td>33.02 (0.957)</td>
<td>42.75 (0.967)</td>
<td>28.96 (0.958)</td>
</tr>
<tr>
<td>3.69</td>
<td>44.82 (0.957)</td>
<td>60.58 (0.973)</td>
<td>35.30 (0.950)</td>
</tr>
</tbody>
</table>

Table 6.7: Unconditional average size (probability of coverage) for grandparent-grandchild pairs. Markers are 20 centimorgans apart.
Chapter 7

Quantitative Trait Loci (QTLs)

The remaining chapters of this dissertation pertain to identifying genes that influence quantitative traits. By quantitative traits we mean characters that can be measured on a continuous scale such as blood pressure, mass of a fruit, etc. The methods are for the particular case of controlled breeding and do not apply as is to human data. However, a great deal can be learned about human genetics from studying animal models, like those involving mice and rats. The methods are also very helpful for agricultural purposes.

We begin by explaining the challenges faced when trying to locate quantitative trait loci. We then describe the work of Fisher (1919), who used analysis of variance (ANOVA) to model the variance in quantitative characters. Finally, we will present the work of Lander and Botstein (1989) who extended Fisher’s ideas to take full advantage of the increasing availability of genetic markers. The underlying mathematical problems for finding quantitative loci using Lander and Botstein’s model are strongly related to the identification of disease susceptibility genes in humans using identity-by-descent data from affected pairs of relatives.
7.1 Quantitative characters

Consider the simplest qualitative case of the color of a flower that is determined by a single gene that has two alleles, A and B. Flowers with genotype BB are white, AB are pink and AA are purple. There is a one-to-one correspondence between the genotype and the phenotype (color of the flower).

![Graph showing the distribution of values for BB, AB, and AA genotypes.](image)

Figure 7.1: Example of a quantitative trait influenced by a single gene

However, in the simplest quantitative traits, there is no such one-to-one correspondence. In Figure 7.1, the knowledge that the phenotype (e.g. length of stem of the flower) is 11 does not determine the genotype unambiguously; the genotype could be either BB or AB for this particular example. This is due to the fact that the variance in quantitative traits has two sources: the genetic variance and the environmental variance. In humans, it is very difficult to separate the environmental and the genetic effects. On the other hand, when breeding is controlled, we can provide the same environment for all organisms and look for differences in mean phenotype due to genetic factors only.

Of course, this simplest case of a single gene influencing the quantitative characters is rarely found in nature. Usually, many loci contribute to the genetic variance,
making it even harder to determined if the variation in the character has a genetic component to it.

Fisher (1918) used analysis of variance (ANOVA) to model the phenotype as a function of the genotype to determine if a particular gene influences a quantitative character of interest. The next section will briefly describe (in our own notation) his work.

### 7.2 Fisher's Method

For illustrative purposes we will present a model that includes only one gene. However, the model easily generalizes to more than one gene. The crucial assumption of Fisher's method is that the genetic and environmental components of the variance are uncorrelated i.e.,

$$V_P = V_G + V_E,$$

where \( V_P \) is the variance in phenotype, \( V_G \) is the genetic variance and \( V_E \) is the environmental variance.

We can write the model as follows. Let \( y_i \) be the phenotype (quantitative character) of interest in individual \( i \). Let \( x_i \) be the number of \( A \) alleles for a particular gene that we think might be influencing the phenotype. Then, we express \( y_i \) in terms of \( x_i \),

$$y_i = \mu + \alpha x_i + \delta 1(x_i=1) + e_i,$$

(7.1)

where the \( e_i \)'s are normally distributed with mean 0 and variance \( \sigma_e \), \( \mu \) is the overall mean, \( \alpha \) is the additive effect and \( \delta \) is the dominance effect. The \( e_i \)'s are independent of the \( x_i \)'s. We assume that the gene \( x \) has only two alleles, \( A \) and \( B \), so that \( x_i = 0, 1 \) or 2. If the gene has more than 2 alleles, the allele \( A \) is the one that is thought to influence the quantitative traits.
CHAPTER 7. QUANTITATIVE TRAIT LOCI (QTLS)

Note that $\delta = \alpha \neq 0$ represents the case of dominance (the genotype AB and AA have the same mean phenotype, different from the mean phenotype of BB) and $\delta = -\alpha$ corresponds to a recessive gene (genotype BB and AB have a different mean phenotype, smaller than the mean phenotype of BB).

Fisher used what later became known as ANOVA to test $H_0 : \alpha = \delta = 0$ versus $H_1 : \alpha \neq 0$ or $\delta \neq 0$.

Remark: When equation (7.1) does not include all the contributing loci, the $e'_s$ contain both the environmental effect and the genetic effects from the excluded loci. The estimate of the environmental variance will be inflated, making it harder to find statistically significant genetic effects. Therefore, it is better to include all contributing loci in the model. For simplicity, we assume that only one locus contributes to the trait variance and continue to write $\sigma^2_e$ for the environmental variance only.

When Fisher developed this method, the number of genes that could be tested was quite small. However, the number of genetic markers available today is larger and still growing. To take advantage of the dense sets of markers, Lander and Botstein (1989) developed a model based on Fisher’s ANOVA to test for the presence of quantitative trait loci at any location on the genome.

7.3 Lander and Botstein’s Model

Before we present the model to search the entire genome for quantitative trait locus, we first describe one of the key assumptions of the model: the data comes from a breeding experiment.

A breeding experiment consist 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 inbreeding or simply two different strains of the same organism with differing mean phenotype. Typically, the progeny are
produced by an intercross or a backcross. An intercross is generated by mating the two parental lines that produced the first generation of offspring (generation $F_1$). The $F_1$ generation is then mated together to produce the second generation ($F_2$). See Figure 7.3 for a schematic description of an intercross. We assume that the parental lines are completely different, so that we can call all the alleles from the strain with the highest mean phenotype “A”, and all the alleles from the lower strain “B”. Each $F_2$ offspring will have 0, 1 or 2 “A” alleles at any location on their genome. Informative individuals can be obtained through another type of design, the backcross, which is generated by mating an $F_1$ offspring to an individual from the parental line (see Figure 7.3). If we take the parental line with the lower mean for the trait, the offspring from the backcross will have 0 or 1 “A” alleles at any locus on their genome, as opposed to 0, 1 or 2 “A” alleles for the intercross.

![Diagram of intercross and backcross](image)

Figure 7.2: Intercross

The model proposed in Lander and Botstein (1989) is as follows. Let $x_i(d)$ be the number of “A” alleles at location $d$ on the genome of individual $i$ and $y_i$ be the value
CHAPTER 7. QUANTITATIVE TRAIT LOCI (QTLS)

Figure 7.3: Backcross

of the quantitative trait for individual $i$. If there exists only one gene located at $r$ that influences the traits, we can model the phenotype as

$$y_i = \mu + \alpha x_i(r) + \delta 1_{(x_i(r)=1)} + e_i,$$  \hspace{1cm} (7.2)

where $\mu$, $\alpha$, $\delta$ and $e_i$ were defined in the previous section. Note that this model could include more than one locus $r$ but we will deal with the simpler model first. The difference between (7.2) and Fisher's model is that $r$ is an unknown parameter. We are no longer testing whether a particular gene located at $r$ influences the trait, but whether there exists a gene anywhere along the genome that has some influence on the quantitative character.

For backcross data, since $x_i(r) = 0$ or 1, $\alpha$ and $\delta$ cannot be estimated uniquely and the model reduces to

$$y_i = \mu + \alpha x_i(r) + e_i,$$  \hspace{1cm} (7.3)

which is the model Lander and Botstein developed in their 1989 paper. We will come back to the full model (7.2) shortly. Note that we can estimate the additive effect $\alpha$ in the case of $r$ and $x_i(r)$ known, using least squares, as

$$\hat{\alpha}(r) = \frac{\sum_{i=1}^{N} [y_i - \bar{y}][x_i(r) - \bar{x}_i(r)]}{\sum_{i=1}^{N} [x_i(r) - \bar{x}_i(r)]^2}.$$  \hspace{1cm} (7.4)
The interesting question is to determine whether there exists a locus that influences the quantitative trait, i.e. \( \alpha \neq 0 \) for some \( r \). In the remaining part of this chapter, we will present the procedure suggested by Lander and Botstein (1989) to test for the presence of QTLs for the simple model (7.3) and extend their method to include testing for both additive and dominance effects.

### 7.3.1 Testing for the additive effects

The test suggested by Lander and Botstein for finding a quantitative trait locus looks for additive effects using the simpler model (7.3) and a backcross design.

Under the assumption that the \( e_i \)'s are normally distributed with mean 0 and known variance \( \sigma_e^2 \), the log likelihood ratio for testing \( H_0 : \alpha = 0 \) versus \( H_1 : \alpha \neq 0 \) for \( r \) and \( x_i(r) \) known is

\[
\left\{ \frac{\sqrt{N} \hat{\alpha}(r)}{2\sigma_e} \right\}^2.
\]

When \( r \) is unknown, as is the case for the Lander and Botstein model, the log likelihood ratio statistic becomes

\[
\max_d \left\{ \frac{\sqrt{N} \hat{\alpha}(d)}{2\sigma_e} \right\}^2,
\]

where the maximum is taken over all the loci \( d \) where \( x_i(d) \) is known. The variable \( x_i(d) \) is known when a marker is available at location \( d \). The usual assumption is that \( x_i(d) \) is known at equispaced distances of \( \Delta \) centimorgans. The value \( \hat{d} \) that maximizes (7.6) is the MLE of \( r \), and the MLE of the additive effect is \( \hat{\alpha}(\hat{d}) \).

**REMARK:** If \( \sigma_e \) is unknown, it can be estimated very accurately by producing offspring with identical genome where the variance in phenotype is solely due to environment. Such offspring can be produced via inbreeding for example.

In order to define a critical region for the test, we need to find the distribution of (7.6) under \( H_0 \).
CHAPTER 7. QUANTITATIVE TRAIT LOCI (QTLS)

Let

$$X(d) = \frac{\sqrt{N} \tilde{a}(d)}{2\sigma_c}.$$  

Under the assumption of equispaced markers, $X(d)$ is observed only when $d = k\Delta$ for some integer value $k$. As $n \to \infty$, $X(d)$ is a Gaussian process with mean 0 and covariance function $R(t) = e^{-2\lambda|t|}$, if we assume that the crossovers follow an exponential distribution with rate $\lambda$ (Haldane mapping function). This limiting distribution holds even when the $e_i$’s are not normally distributed. Therefore, $X(d)$ is an Ornstein-Uhlenbeck process, and the log likelihood ratio statistic, under $H_0$, has the same distribution as the extrema of the square of a Ornstein-Uhlenbeck process. This is exactly the same process we studied for the case of disease susceptibility genes using identity-by-descent data from affected pairs of relatives in Chapters 3 through 6. Lander and Botstein (1989) provided an approximation for the distribution of (7.6) when the data are continuous i.e., the maximum is taken over the entire genome. For the case where the $x_i(d)$ are only known at equispaced distance of $\Delta$ centimorgans, the authors gave the results from a simulation study to define the critical region of the test. We can compare their results to the ones we would obtain using the approximation to the tail distribution of (7.6) from Chapter 3,

$$P\left\{ \max_k X^2(k\Delta) > b^2 \right\} \approx 2 \left[ 1 - \Phi(b) + 2\lambda l b \phi(b) \nu \left( 2b \sqrt{\lambda \Delta} \right) \right],$$  

(7.7)

where $l$ is the length of the chromosome and $\nu(x) \approx \exp[-0.583x]$. The factor of 2 in the right hand side of equation (7.7) is to account for both events $\{X(k\Delta) > b\}$ and $\{X(k\Delta) < -b\}$.

To illustrate the closeness of the two methods, let’s take a genome with 10 chromosomes of length 100 centimorgans. According to Lander and Botstein, the maximum lod score i.e., ($\frac{1}{2} \log_{10}(e))X^2(d)$, must exceed 3.1 for the continuous data case and 2.5 for the map with $\Delta = 10cM$ and a test with type-1 error of 5%. If we use equation (7.7) we get thresholds for the lod score of 3.1 and 2.6 for the continuous and discrete
case respectively.

The model we have described so far includes only one gene, which does not agree with the general belief that quantitative traits are influenced by many loci. To comply with this reasonable assumption, once a gene influencing the trait has been identified, it is included in the model. A search for other loci that explain the remaining phenotypic variance can be performed similarly. Another way to deal with multilocus traits is to search simultaneously for more than one trait locus. For the simultaneous search, the likelihood ratio will have the same distribution as the statistics of Chapter 4 for polygenic human diseases. The methods of Chapter 4 can be applied directly.

One drawback of the Lander and Botstein’s approach is that it ignores the possible dominance effect in its testing scheme. Patterson et al. (1991) used the full model (7.2) to locate QTL in tomatoes in an intercross, where the dominance effects could be estimated. However, the statistical significance of the dominance effects were not estimated, for lack of methodology. We next address the issue of testing for the presence of both additive and dominance effects.

### 7.3.2  Testing for both additive and dominance effects

In the case of the full model,

\[ y_i = \mu + \alpha x_i(r) + \delta 1_{(x_i(r)=1)} + \epsilon_i, \]

with intercross data, the vectors \( x_i \) and \( 1_{(x_i(r)=1)} \) are asymptotically orthogonal so that we can estimate the additive effects and the dominance effects by (for known \( r \))

\[ \hat{\alpha}(r) = \frac{\sum_{i=1}^{N} [y_i - \bar{y}] [x_i(r) - \bar{x}(r)]}{\sum_{i=1}^{N} [x_i(r) - \bar{x}(r)]^2}, \quad \hat{\delta}(r) = \frac{\sum_{i=1}^{N} [y_i - \bar{y}] [1_{(x_i(r)=1)} - \bar{1}_{(x_i(r)=1)}]}{\sum_{i=1}^{N} [1_{(x_i(r)=1)} - \bar{1}_{(x_i(r)=1)}]^2}. \]

To test the more general hypothesis of \( H_0 : \alpha = \delta = 0 \) versus \( H_1 : \alpha \neq 0 \) or \( \delta \neq 0 \), the log likelihood ratio is (under the assumption that the gene is located at a marker)

\[ \max_k \left[ \left( \frac{\sqrt{N} \widehat{\alpha}(k\Delta)}{\sqrt{2}\sigma_x} \right)^2 + \left( \frac{\sqrt{N} \widehat{\delta}(k\Delta)}{2\sigma_x} \right)^2 \right]. \quad (7.8) \]
CHAPTER 7. QUANTITATIVE TRAIT LOCI (QTLS)

We will find an approximation to the distribution of (7.8) under $H_0$ in order to define a critical region for the test.

Let

$$X(d) = \frac{\sqrt{N}\hat{\alpha}(d)}{\sqrt{2}\sigma_e}, \quad \text{and} \quad Y(d) = \frac{\sqrt{N}\hat{\delta}(d)}{2\sigma_e}.$$

As $n \to \infty$, $X(d)$ and $Y(d)$ are Gaussian processes with mean 0 and covariance functions $e^{-2\lambda|d|}$ and $e^{-4\lambda|d|}$ respectively. Moreover, $X(d)$ and $Y(d)$ are asymptotically independent.

Using the above facts, we approximate the tail distribution of (7.8) by

$$P\left\{ \max_k \left[ X^2(k\Delta) + Y^2(k\Delta) \right] \geq b^2 \right\} \approx e^{-\frac{1}{2}b^2} + b^2 \nu \left( b\sqrt{\Delta(\beta_1 + \beta_2)} \right) e^{-\frac{1}{2}b^2},$$

(7.9)

where $\beta_1 = 2\lambda$ and $\beta_2 = 4\lambda$. For the case where the value of $x_i(d)$ is known continuously along the genome, so $\Delta = 0$, the factor $\nu \left( b\sqrt{\Delta(\beta_1 + \beta_2)} \right) = 1$.

The first term of the approximation is to account for the probability that the likelihood ratio statistic exceeds the threshold at the beginning of the chromosome i.e., $[X^2(0) + Y^2(0)] > b^2$. The second term was obtained by a suitable modification of Woodrooffe’s (1976) argument, which is presented in Appendix B.

Approximation (7.9) is for one chromosome only. The overall significance of the test will be approximately $n$ times equation (7.9), where $n$ is the number of chromosomes the organism under investigation has.

To verify the accuracy of the above approximation, we performed a simulation study. We calculated the threshold for the square root of statistic (7.8) to test for the presence of quantitative trait loci in tomatoes. The tomato has a genome of 12 chromosomes of approximate length 100 centimorgans each. We calculated the threshold for a test with type-1 error of 5%. For the simulations, we generated independent Ornstein-Uhlenbeck processes and calculated the 95% quantile of statistic (7.8). The results presented in Table 7.1 show that the approximation is very accurate.
Table 7.1: Thresholds for testing for the presence of additive or dominance effects

7.4 Discussion

Lander and Botstein’s 1989 paper presented a framework to search an entire genome to locate gene loci that influence quantitative traits. We extended their models so as to allow for testing both additive and dominance effects. Once we found evidence that there exists a gene influencing the quantitative trait of interest, we would want to narrow the region in which to look for the exact location of the locus. In the next chapter, we present a method for constructing a joint confidence region for the effects and the location of the gene locus. We also demonstrated how the methods of constructing confidence sets for the IBD data from Chapter 6 can be used in the context of QTLs.
Chapter 8

Confidence region for QTLs

Confidence sets are useful ways of narrowing the search for a QTL to a small portion of a chromosome. In the context of human identity-by-descent data on affected pairs of relatives, three methods of making confidence statements were used (Chapter 6): lod support intervals, likelihood intervals and Bayesian credible sets. In this chapter, we first show how to adapt all three methods to provide confidence sets for QTLs. We deal mainly with very dense maps of markers ($\Delta \leq 2cM$). We also include a comparison study of the 3 above methods for making confidence statements about the location of a quantitative trait locus.

Secondly, we present a likelihood method for constructing a joint confidence region for a quantitative trait locus, its additive effect ($\alpha$) and its dominance effect ($\delta$). We show through a simulation that the joint confidence region has the correct probability of coverage given that the number of organisms available is not too small ($N \geq 100$).
CHAPTER 8.  CONFIDENCE REGION FOR QTLS  

8.1 Confidence Sets for Locus Only

8.1.1 Lod Support Intervals

We briefly review how to construct a lod-support interval. A \( x \)-lod support interval includes all the loci \( v \) such that

\[
\text{lod}(v) \geq \max_t \text{lod}(t) - x,  \tag{8.1}
\]

where \( \text{lod}(v) \) is the logarithm base 10 of the likelihood ratio statistic at locus \( v \).

We work with the full model

\[
y_i = \mu + \alpha x_i(r) + \delta I(x_i(r) = 1) + e_i,  \tag{8.2}
\]

where \( y_i \) is the phenotype of individual \( i \) and \( x_i(d) \) is the number of A alleles individual \( i \) possesses at locus \( d \). The \( e_i \)'s are assumed to be normally distributed with mean 0 and known variance \( \sigma_e \).

From Chapter 7, we know that the asymptotic log likelihood ratio for the test \( H_0 : \alpha = \delta = 0 \) versus \( H_1 : \alpha \neq 0 \) or \( \delta \neq 0 \) for the presence of a quantitative trait locus at \( d \) is

\[
LR(d) = X^2(d) + Y^2(d),  \tag{8.3}
\]

where

\[
X(d) = \frac{\sqrt{N} \hat{\alpha}(d)}{\sqrt{2\sigma_e}}, \quad Y(d) = \frac{\sqrt{N} \hat{\delta}(d)}{2\sigma_e}.
\]

The functions \( \hat{\alpha}(d) \) and \( \hat{\delta}(d) \) are the least square estimates of the additive and dominance effects based on the model (8.2) with a locus at \( d \). The lod score can be written as

\[
\text{lod}(d) = \frac{1}{2} (\log_{10} e) LR(d).  \tag{8.4}
\]

We can use condition (8.1) and equation (8.4) to construct a lod-support interval. A simulation study showed that a 1.5-lod support interval is approximately equivalent to a 95\% confidence set for the gene locus.
8.1.2 Bayesian Credible Sets

Another method of making confidence statements for a disease susceptibility gene is through the use of Bayesian credible sets. They can also be used for quantitative trait loci.

In Chapter 6, we constructed a set \( B_\gamma \) of posterior probability \( 1 - \gamma \) by including all loci \( r \) whose posterior density exceeds \( c_\gamma \), i.e.,

\[
B_\gamma = \{ r : \pi(r|\{X_t,Y_t\}) > c_\gamma \}
\]  

(8.5)

where \( c_\gamma \) is chosen so that

\[
\int_{B_\gamma} \pi(r|\{X_t,Y_t\}) dr = 1 - \gamma.
\]

We will find an approximation to \( \pi(r|\{X_t,Y_t\}) \) and use it in (8.5) to construct a \((1 - \gamma)\%\) credible set for a quantitative trait locus.

First, we put a uniform prior on the location of the gene \( r \) and a bivariate normal prior on the additive and dominance effects, such that

\[
\pi(r, \xi_1, \xi_2) = \frac{1}{\Delta_2} \frac{1}{2\pi} e^{-\frac{1}{2}(\xi - \theta)^\top \Sigma^{-1} (\xi - \theta)},
\]

where

\[
\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad Z_t = \begin{pmatrix} X_t \\ Y_t \end{pmatrix}, \\
\Sigma_1 = \begin{pmatrix} \eta_1 & \eta_{12} \\ \eta_{12} & \eta_2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}.
\]

In addition, note that the likelihood ratio is of the form

\[
lik(r, \xi) = \frac{dP_{(r,\xi)}}{dP_0}(Z) = e^{r^\top Z - \frac{1}{2} \|\xi\|^2}.
\]
Bayes' rule with \( \eta_{12} = 0 \) entails that

\[
\pi(r, \xi|\{X_t, Y_t\}) = \frac{\text{lik}(r, \xi)\pi(r, \xi)}{\int_{0}^{\infty} \int_{-\infty}^{\infty} \text{lik}(s, y)\pi(s, y)dy_1dy_2ds}
\]

\[
= \frac{e^{\xi_1X_r-\xi_2Y_r-\xi_2/2} \frac{1}{\Delta} e^{-\frac{1}{2}(\xi-\theta)'\Sigma^{-1}(\xi-\theta)}}{\sum_{i=0}^{1/\Delta} \int_{0}^{\infty} \int_{-\infty}^{\infty} e^{\xi_1X_{\Delta_i}-\xi_2Y_{\Delta_i}-\xi_2/2} \frac{1}{\Delta} \frac{1}{2\pi} e^{-\frac{1}{2}(\xi-\theta)'\Sigma^{-1}(\xi-\theta)}dy_1dy_2}
\]

\[
= \frac{e^{\xi_1X_r-\xi_2/2+\xi_2Y_r-\xi_2/2} \frac{1}{\Delta} \frac{1}{2\pi} e^{-\frac{1}{2}(\xi-\theta)'\Sigma^{-1}(\xi-\theta)}}{\frac{1}{\sqrt{(1+1/\eta_1^2)(1+1/\eta_2^2)}} e^{\theta'\Sigma^{-1}\theta} \sum_{i=0}^{1/\Delta} \frac{1}{2(1+1/\eta_1^2)} \frac{e^{(X_{\Delta_i}+\theta_1)/\eta_1^2}}{e^{(1+1/\eta_1^2)}} \frac{1}{2(1+1/\eta_2^2)} \frac{e^{(Y_{\Delta_i}+\theta_2)/\eta_2^2}}{e^{(1+1/\eta_2^2)}}}
\]

Integrating \( \pi(r, \xi|\{Z_t\}) \) over \( \xi_1 \) and \( \xi_2 \) we get

\[
\pi(r|\{Z_t\}) = \frac{\left(\frac{X_r+\theta_1}{\eta_1^2}\right)^2 \left(\frac{Y_r+\theta_2}{\eta_2^2}\right)^2}{e^{\frac{X_r+\theta_1}{\eta_1^2}} e^{\frac{Y_r+\theta_2}{\eta_2^2}}}
\]

\[\text{(8.6)}\]

Choosing the value of \( \theta \) and \( \Sigma \) for the prior on the additive and dominance effects is more tricky than for the disease susceptibility genes since the effects can be negative (for the human data \( \xi \geq 0 \)). We chose a prior with mean 0 and standard deviation of 4 in our simulations (Section 8.1.4). The mean of 0 is to allow the parameters to be positive or negative and a standard deviation of 4 should be large enough to allow the parameters to vary freely.

For a prior centered at 0, the posterior probability (8.6) reduces to

\[
\pi(r|\{Z_t\}) = \frac{\frac{x_r^2}{e^{\frac{x_r^2}{2(1+1/\eta_1^2)}}} \frac{y_r^2}{e^{\frac{y_r^2}{2(1+1/\eta_2^2)}}}}{\sum_{i=0}^{1/\Delta} \frac{x_{\Delta_i}^2}{e^{\frac{x_{\Delta_i}^2}{2(1+1/\eta_1^2)}}} \frac{y_{\Delta_i}^2}{e^{\frac{y_{\Delta_i}^2}{2(1+1/\eta_2^2)}}}}
\]

\[\text{(8.7)}\]
Other priors for the additive and dominance effects have been investigated such as a mixture of normals and a uniform prior. We present the credible set obtained from the three priors in our comparison study (Section 8.1.4).

8.1.3 Likelihood Methods

Confidence statements for a quantitative trait locus can be obtained through the use of likelihood methods. Since the location of the locus is a change-point (Chapter 3), we can adapt the methods for constructing confidence sets for change-points presented by Siegmund (1988) to our particular problem. We will sketch how to derive the likelihood confidence set below.

It can be shown that $A_v$, the acceptance region for the likelihood ratio test of the hypothesis that the gene locus is at location $v$, is of the form

$$A_v = \{ \max_t ||Z_t||^2 - ||Z_v|| \leq k \}.$$  

The conditional probability of $A_v$ given $Z_v$ does not depend on $\xi$, so that

$$P(A_v | Z_v) = 1 - \gamma.$$  

Therefore, the set of all values $v$ that are accepted by the likelihood ratio test is a $(1 - \gamma) 100\%$ confidence set. It is not necessary to solve for $k$ since

$$(\max_t ||Z_t||^2)_{obs} - ||Z_v||^2 \leq k \iff P(\max_t ||Z_t||^2 > (\max_t ||Z_t||^2)_{obs} | Z_v) \geq P(\max_t ||Z_t||^2 > k + ||Z_v||^2 | Z_v) = \gamma.$$  

The likelihood confidence set will contain all loci $v$ that satisfy

$$P(\max_t ||Z_t||^2 > (\max_t ||Z_t||^2)_{obs} | Z_v) \geq \gamma. \quad (8.8)$$  

To construct the likelihood confidence set, we will use the approximation given by the following lemma.
Lemma 8.1 Let

\[ Z_t = \begin{pmatrix} Z_t^{(1)} \\ Z_t^{(2)} \end{pmatrix}, \]

where \( Z_t^{(1)} \) and \( Z_t^{(2)} \) are independent Gaussian processes with covariance functions

\[ R_i(t) = 1 - \beta_i|t| + o(|t|) \quad \text{as} \quad t \to 0. \]

Let \( 0 < ||x||^2 < b^2 \), and define \( t^* \) and \( \omega^* \) to be the positive solution of

\[ \begin{pmatrix} \frac{x_1}{R_1(t^*)} \\ \frac{x_2}{R_2(t^*)} \end{pmatrix} = \begin{pmatrix} b \cos \omega^* \\ b \sin \omega^* \end{pmatrix}. \]

Assume \( ||x|| \) and \( b \) to be large and that \( t^* \) is contained in \((0, t_1)\) and is bounded away from the upper endpoint \((t_1 > 0)\). Then

\[ P \left\{ \max_{0 \leq i \leq t} ||Z_i\Delta|| \geq b \mid Z_0 = x \right\} \approx \]

\[ \frac{e^{-\frac{1}{2}(t^2 - ||x||^2)} \left( \beta_1 \cos^2 \omega^* + \beta_2 \sin^2 \omega^* \right)}{\left| \dot{R}_1(t^*) \dot{R}_2(t^*) \cos^2 \omega^* + \dot{R}_1(t^*) \dot{R}_2(t^*) \sin^2 \omega^* \right|} \nu \left( b \left[ (\beta_1 + \beta_2)\Delta \right]^{\frac{3}{2}} \right), \]

where \( \dot{R}_i(t) = \frac{dR_i(t)}{dt} \).

A sketch of a proof of Lemma 8.1 is included in Appendix C.

For the particular application of testing for the presence of additive and dominance effects,

\[ R_1(t) = e^{-2\lambda|t|} \quad \text{and} \quad \dot{R}_2(t) = e^{-4\lambda|t|}, \]

which satisfy the conditions of Lemma 8.1; and the approximation reduces to

\[ P \left\{ \max_{0 \leq i \leq t} ||Z_i\Delta|| > b \mid Z_0 = x \right\} \approx \frac{(2b^2)^{3/2} e^{-\frac{1}{2}(t^2 - ||x||^2)}}{\left( x_1^2 + \sqrt{x_1^4 + 4b^2x_2^2} \right)^{3/2}} \nu \left( b \left[ 6\lambda\Delta \right]^{\frac{3}{2}} \right). \quad (8.9) \]
Using approximation (8.9) with condition (8.8), we conclude that an approximate $(1 - \gamma)%$ confidence set includes all loci $v$ such that
\[ 2 \frac{(2Z^2)^{3/2} e^{-\frac{1}{2}(Z^2 - ||z||^2)}}{(X_v^2 + \sqrt{X_v^4 + 4Z^2Y_v^2})^{3/2}} \nu \left( Z^* \left[ 6\lambda \Delta \right]^{\frac{1}{2}} \right) \geq \gamma, \]
where $Z^* = (\max \{ ||z|| \})_{\text{obs}}$. This method of constructing confidence sets for a quantitative trait locus is compared with the lod-support interval and the Bayesian credible sets in the following section.

### 8.1.4 Comparison Study

We constructed the likelihood confidence set, the 1.5-lod support interval and the Bayesian credible sets, with the three different priors mentioned in the previous section, for 350 tomato genomes simulated in the following way:

i) For each tomato, generate the crossover process for the chromosome containing the QTL using Haldane map function,

ii) Set the phenotype $(y_i)$ for each tomato to be
\[ y_i = \alpha x_i(r) + \delta I(x_i(r) = 1) + e_i, \]
where the $e_i$'s are normal random variables with mean 0 and variance 1.

The tomato genome has 12 chromosomes of approximate length 100 centimorgans. We performed the simulations for the additive model ($\delta = 0$) and the dominance model ($\delta = \alpha$), each at a power of 95% and 80%, giving four combinations of values of $\alpha$ and $\Delta$. The locus was picked to be in the middle of the chromosome. We generated 1000 sets of 350 tomatoes and calculated the average size and the probability of covering the true locus. The results are present in Table 8.1.4.

The lod-support interval and the Bayesian credible set with a prior consisting of a mixture of normals gave the tightest confidence sets while the likelihood method provided the largest ones. All three methods seem to give confidence sets with coverage
Table 8.1: Average size (probability of coverage) for 350 tomatoes. Add. stands for additive model, Dom. stands for dominance model. (F) is for the uniform prior, (N) for the normal prior with mean 0 and standard deviation 4 and (M) is for a mixture of 4 bivariate normal distribution centered at (4,4), (4,-4), (-4,4), (-4,-4) with standard deviation of 1.

probabilities close to 95%.

We repeated the simulations removing the assumption of known variance and the confidence sets were very similar in size and coverage probability. However, when we performed the simulations with fewer tomatoes (30, 50, 100 or 200), the sizes of the confidence sets were unaffected but the coverage probabilities were greatly reduced. The lod-support interval and the likelihood confidence sets were the least affected by a reduction of the sample sizes. For our range of values for the additive and the dominance effects, the lod-support interval was most robust to small sample sizes while providing tighter confidence sets.

8.2 Joint Confidence Region for genetic effects and locus

So far, we have provided methods of constructing confidence sets for the quantitative trait locus only. When the magnitude of the additive and dominance effects are of interest, a joint confidence region for the locus, its additive effect and its dominance effect might be more appropriate. We will use likelihood methods to construct an
approximate joint confidence region for the quantitative trait locus and its genetic effects.

8.2.1 Likelihood Methods

We will derive a joint confidence set along the lines of Siegmund (1988). We use the full model

\[ y_i = \mu + \alpha x_i(r) + \delta_{i(x_i(r)=1)} + e_i, \]

where the \( e_i \)'s are assumed normally distributed with mean 0 and known variance \( \sigma_e \).

Let \( \hat{\alpha}(t) \) and \( \hat{\delta}(t) \) be the estimate of the additive effect and of the dominance effect when there exists a trait locus at \( t \). As in the previous sections,

\[
Z_t = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} \sqrt{n \hat{\alpha}(t)} \\ \sqrt{n \hat{\delta}(t)} \end{pmatrix} \quad \text{and} \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \sqrt{n \hat{\alpha}} \\ \sqrt{n \hat{\delta}} \end{pmatrix}.
\]

The likelihood ratio for the test that there exist a quantitative trait locus at \( v \) with genetic effects \( \xi \) has acceptance region

\[
A_{v,\xi} = \left\{ \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 \leq c^2 \right\},
\]

where \( c = c(v, \xi) \) is chosen to satisfy \( P(A_{v,\xi}) = 1 - \gamma \) for all \( v \) and \( \xi \).

The set of all points \((v, \xi)\) that are accepted by the likelihood ratio test form a \((1 - \gamma)\)% joint confidence region. It is not necessary to solve for \( c \) since

\[
\left( \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 \right)_{obs} \leq c^2 \iff P\left\{ \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 \right\} \geq P\left\{ \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 > c^2 \right\} = \gamma.
\]

In order to construct the joint confidence region, all we need is an approximation to

\[
P\left\{ \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 > b^2 \right\}.
\]
CHAPTER 8. CONFIDENCE REGION FOR QTLS

We separate equation (8.10) into 2 parts.

\[
P\left\{ \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 > b^2 \right\} = \]
\[P\left\{ ||Z_v-\xi||^2 > b^2 \right\} + E\left[ P\left( \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 > b^2 \mid ||Z_v-\xi||, \omega \right) ; ||Z_v-\xi||^2 \leq b^2 \right],
\]

where \( \omega \) is the angle between \( Z_{v,1} \) and \( Z_{v,2} \).

The first term is the tail probability of a \( \chi^2 \) distribution with 2 degrees of freedom. To evaluate the second term, we can use approximation (8.9) for the inner probability and perform the expectation integral numerically. Hence we approximate (8.10) as

\[
P\left\{ \max_t ||Z_t||^2 - 2\xi'Z_v + ||\xi||^2 > b^2 \right\} \approx \]
\[e^{-\frac{1}{2}b^2} + \int \int \frac{(2\tilde{b})^{3/2}e^{-\frac{1}{2}(\tilde{b}^2-||v||^2)}}{B (x_1^4 + x_2^4 + 4\tilde{b}^2x_2^2)^{3/2}} \nu\left( \tilde{b} \left[ 6\lambda \Delta \right]^{1/2} \right) P\left( Z_{v,1} \in dx_1, Z_{v,2} \in dx_2 \right), \quad (8.11)
\]

where \( B = \{(v, \xi) : ||Z_v-\xi|| < \tilde{b}\} \) and \( \tilde{b} = b^2 + 2\xi'Z_v - ||\xi||^2 \). Note that for fixed \( v \), \( Z_{v,1} \) and \( Z_{v,2} \) are independent normal random variable with mean \( \xi_1 \) and \( \xi_2 \) and variance 1. A change of variable to polar coordinates makes this numerical integration easier to perform.

A point \((v, \xi)\) is included in the joint confidence region if equation (8.11) (with \( b = (\max_t ||Z_t||)_{obs} \)) is greater than or equal to \( \gamma \). The above procedure to construct a confidence region was implemented and tested through a simulation study under various alternative hypotheses. The next section has the results of the simulations.

8.2.2 Simulations

The genome and phenotype of \( n \) tomatoes (\( n = 30, 50, 100, 200 \) and 350) were generated as described in Section 8.1. The joint confidence region was constructed and the probability of the joint confidence region to include the true triplet \((r, \alpha, \delta)\) was calculated based on 1000 simulations. The simulations were performed for values of \( \alpha \sqrt{n/2} \) ranging from 0 to 5.5 and \( \delta \) values of 0, \( \pm \alpha/2 \) and \( \pm \alpha \). Table 8.2 presents the
coverage probability form 5 simulations for each sample size. The procedure seems to give a region with the correct probability of coverage with as few as 100 organisms. The coverage probability was not affected by the values of $\alpha$ and $\delta$ or the location of the gene, except when the gene was at the very end, in which case the probability of coverage was reduced. The confidence region was constructed assuming the environmental variance was known; however, similar results were obtained when this assumption was not in force.

<table>
<thead>
<tr>
<th>Number of tomatoes</th>
<th>Coverage Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>86.1 %</td>
</tr>
<tr>
<td>50</td>
<td>90.2 %</td>
</tr>
<tr>
<td>100</td>
<td>95.0 %</td>
</tr>
<tr>
<td>200</td>
<td>96.4 %</td>
</tr>
<tr>
<td>350</td>
<td>95.3 %</td>
</tr>
</tbody>
</table>

Table 8.2: Coverage probability of the likelihood joint confidence region.

The joint confidence region can be used to describe which values of $\alpha$ and $\delta$ are consistent with a particular value of the gene locus. If only the location of the gene is of interest, a confidence set of the gene only can be used by profiling the joint confidence region. However, the methods from Section 8.1 are much more efficient for providing confidence region for the gene only.
Appendix A

Power Approximations

A.1 One locus

The distributions of the test statistics under the alternative hypothesis are obtained by Feingold et al. (1993) by breaking the power into two parts. For the simplest case of a single locus and continuous map, equation (4.12),

$$P_{r,\xi} \left( \max_i \frac{Z_t}{\sigma} > b \right) = P_{r,\xi} \left( \frac{Z_r}{\sigma} > b \right) + \int_0^\infty P_{r,\xi} \left( \max_i \frac{Z_t}{\sigma} > b \mid \frac{Z_r}{\sigma} = b-x \right) P_{r,\xi} \left( \frac{Z_r}{\sigma} \in b-\infty \right).$$  

(A.1)

The key is to find exact expressions or approximations for all 3 probabilities of the right side of equation A.1.

Asymptotically, $\sigma^{-1}Z_r \sim N(\sigma^{-1}\xi, 1)$, so that

$$P_{r,\xi} \left( \frac{Z_r}{\sigma} > b \right) = 1 - \Phi \left( b - \frac{\xi}{\sigma} \right)$$

and

$$P_{r,\xi} \left( \frac{Z_r}{\sigma} \in b-\infty \right) = \phi \left( b - \frac{\xi}{\sigma} \right) dx.$$

For the third probability, note that given $Z_r$, $\max_{t<\tau} Z_t$ and $\max_{t>\tau} Z_t$ are independent when $Z_t$ is an Ornstein-Uhlenbeck process (not necessarily for the general
covariance function $1 - \beta|\tau| + o(|\tau|)$. So we can write

$$P_{r,\xi} \left( \max \frac{Z_t}{\sigma} > b \mid \frac{Z_r}{\sigma} = b - x \right) =$$

$$P_{r,\xi} \left( \max_{t<r} \frac{Z_t}{\sigma} > b \mid \frac{Z_r}{\sigma} = b - x \right) + P_{r,\xi} \left( \max_{t>r} \frac{Z_t}{\sigma} > b \mid \frac{Z_r}{\sigma} = b - x \right)$$

$$-P_{r,\xi} \left( \max_{t<r} \frac{Z_t}{\sigma} > b \mid \frac{Z_r}{\sigma} = b - x \right) P_{r,\xi} \left( \max_{t>r} \frac{Z_t}{\sigma} > b \mid \frac{Z_r}{\sigma} = b - x \right).$$

Feingold et al. (1993) provided the following approximation for the equispaced marker case or the continuous case ($\Delta = 0$).

$$P_{r,\xi} \left( \max_t \frac{Z_t}{\sigma} > b \mid \frac{Z_0}{\sigma} = b - x \right) \approx \nu \left( b \sqrt{2\beta \Delta} \right) e^{-bx}.$$

Substituting this in equation (A.1) and integrating gives the power approximation (4.12).

To derive (4.12), we used an approach similar to that of Feingold et al. (1993). when the locus is located between markers, i.e., $t_1 < r < t_2$ and $t_2 - t_1 = \Delta$; $Z_r$ is not observed, and the power is broken down differently as

$$P_{r,\xi} \left( \max_{0 \leq i \leq 1} \frac{Z_i \Delta}{\sigma} > b \right) = P_{r,\xi} \left( \max \left( \frac{Z_{t_1}}{\sigma}, \frac{Z_{t_2}}{\sigma} \right) > b \right)$$

$$+ \int_0^\infty \int_0^\infty P_{r,\xi} \left( \max_{0 \leq i \leq 1} \frac{Z_i \Delta}{\sigma} > b \mid \frac{Z_{t_1}}{\sigma} = b - x, \frac{Z_{t_2}}{\sigma} = b - y \right) P_{r,\xi} \left( \frac{Z_{t_1}}{\sigma} \in b - dx, \frac{Z_{t_2}}{\sigma} \in b - dy \right).$$

(A.2)

Again we need to find approximations to the 3 probabilities on the right side of (A.2).

Since $Z_{t_1}$ and $Z_{t_2}$ are normal random variables with mean $\xi e^{-\beta|t_1-r|}$ and $\xi e^{-\beta|t_2-r|}$ respectively, variance $\sigma^2$ and covariance $\sigma^2 e^{-\beta|t_2-t_1|} = \sigma^2 e^{-\beta \Delta}$, their joint density can be written as

$$P_{r,\xi} \left( \frac{Z_{t_1}}{\sigma} \in dx, \frac{Z_{t_2}}{\sigma} \in dy \right) =$$

$$\phi \left( b - x - \frac{\xi e^{-\beta|t_1-r|}}{\sigma} \right) \phi \left( \frac{b - y - \frac{\xi e^{-\beta|t_2-r|}}{\sigma} - e^{-\beta \Delta (b - x - e^{-\beta|t_1-r|})}}{\sqrt{1 - e^{-2\beta \Delta}}} \right) \frac{dx \, dy}{\sqrt{1 - e^{-2\beta \Delta}}},$$
so that

\[
P_{r,\xi}(\max(\frac{Z_{t_1}}{\sigma}, \frac{Z_{t_2}}{\sigma}) > b) = 1 - P_{r,\xi}(\max(\frac{Z_{t_1}}{\sigma}, \frac{Z_{t_2}}{\sigma}) < b) = 1 - P_{r,\xi}(\frac{Z_{t_1}}{\sigma} < b, \frac{Z_{t_2}}{\sigma} < b) = 1 - \int_0^\infty \int_0^\infty P_{r,\xi}(\frac{Z_{t_1}}{\sigma} \in b - dx, \frac{Z_{t_2}}{\sigma} \in b - dy) = 1 - \int_0^\infty h(x, t_1, r) \Phi\left(\frac{g(x, t_2, t_1, r)}{d(\Delta)}\right) dx,
\]

(A.3)

which is the first line of approximation (4.12). For the definitions of \( h, g, \) and \( d \), see equation (4.12).

For the other two probabilities, given \( Z_{t_1} \) and \( Z_{t_2} \), the variables \( \max_{0 \leq \Delta \leq t_1} Z_{i\Delta} \) and \( \max_{t_2 \leq \Delta \leq t_1} Z_{i\Delta} \) are independent and therefore,

\[
P_{r,\xi}\left(\max_{0 \leq \Delta \leq t_1} \frac{Z_{i\Delta}}{\sigma} > b \mid \frac{Z_{t_1}}{\sigma} = b - x, \frac{Z_{t_1}}{\sigma} = b - y\right) =

P_{r,\xi}\left(\max_{0 \leq \Delta \leq t_1} \frac{Z_{i\Delta}}{\sigma} > b \mid \frac{Z_{t_1}}{\sigma} = b - x\right) + P_{r,\xi}\left(\max_{t_2 \leq \Delta \leq t_1} \frac{Z_{i\Delta}}{\sigma} > b \mid \frac{Z_{t_2}}{\sigma} = b - y\right) - P_{r,\xi}\left(\max_{0 \leq \Delta \leq t_1} \frac{Z_{i\Delta}}{\sigma} > b \mid \frac{Z_{t_1}}{\sigma} = b - x\right) P_{r,\xi}\left(\max_{t_2 \leq \Delta \leq t_1} \frac{Z_{i\Delta}}{\sigma} > b \mid \frac{Z_{t_2}}{\sigma} = b - y\right)
\approx \nu\left(b\sqrt{2\beta \Delta}\right)e^{-\beta x} + \nu\left(b\sqrt{2\beta \Delta}\right)e^{-\beta y} + \nu^2\left(b\sqrt{2\beta \Delta}\right)e^{-\beta(x+y)}.
\]

Substituting in (A.2) and integrating we obtain approximation (4.12).

\section*{A.2 Two loci}

We will first outline how to obtain equation (4.13) for the power of the two loci test when the two loci are located at markers. Note that we can break the expression for the power into 2 parts as
\[ P_{\xi,r} \left( \max_t \frac{Z_{1t}}{\sigma} + \max_s \frac{Z_{1s}}{\sigma} > b\sqrt{2} \right) = P_{\xi,r_1,r_2} \left( \frac{Z_{1r_1}}{\sigma} + \frac{Z_{1r_2}}{\sigma} > b\sqrt{2} \right) \]

\[ + \int_0^\infty P \left( \max_t \frac{Z_{1t}}{\sigma} + \max_s \frac{Z_{1s}}{\sigma} > b\sqrt{2} \mid \frac{Z_{1r_1}}{\sigma} + \frac{Z_{1r_2}}{\sigma} = b\sqrt{2} - x \right) \frac{\rho_s \left( \frac{Z_{1r_1}}{\sigma} + \frac{Z_{1r_2}}{\sigma} \right) \in b\sqrt{2} - dx \right), \]

(A.4)

where \( r = (r_1, r_2) \) and \( r_1 \) and \( r_2 \) are on different chromosomes. Since \( \frac{Z_{1r_1}}{\sigma} + \frac{Z_{1r_2}}{\sigma} \) is normally distributed, we know its distribution function and density; but what requires more work is the conditional probability in the integral.

Using the approximation (for large \( y \))

\[ P \left( \max_t \frac{Z_{1t} - Z_{1r_1}}{\sigma} \leq y \mid \frac{Z_{1r_1}}{\sigma} + \frac{Z_{2r_2}}{\sigma} = b\sqrt{2} \right) \]

\[ \approx 1 - 2e^{-b\sqrt{2} + \sqrt{2}e^{-y} + \nu \left( b\sqrt{2} + \sqrt{2}e^{-y} \right)}, \quad y \geq 0 \]

we find that the distribution function of the convolution is (for small \( x \))

\[ P \left( \max_t \frac{Z_{1t} - Z_{1r_1}}{\sigma} + \max_s \frac{Z_{1s} - Z_{1r_2}}{\sigma} > x \mid \frac{Z_{1r_1}}{\sigma} + \frac{Z_{1r_2}}{\sigma} = b\sqrt{2} - x \right) \]

\[ \approx 2e^{-b\sqrt{2}x + \sqrt{2}e^{-4\nu} + 2\sqrt{2} - \nu \left( 1 - \nu \right) e^{x\sqrt{2}} - \frac{1}{\sqrt{2}} \nu \left( \nu - 1 \right) e^{-b\sqrt{2}x}}, \quad x \leq 0 \]

Substituting the approximation (for small \( x \))

\[ \phi \left( b - \frac{x}{\sqrt{2} - \sqrt{2}x} \right) dx \approx \exp \left[ x \left( b - \frac{\xi}{\sigma} \right) \right] \phi \left( b - \frac{\sqrt{2} \xi}{\sigma} \right) dx \]

in (A.4) before integrating gives the approximation (4.13) for the power of the two loci test with both loci at markers.

The last approximation that will be derived is equation (4.15) for the case where both loci are between markers, i.e., \( s_1 < r_1 < s_2, t_1 < r_2 < t_2 \) and \( s_2 - s_1 = t_2 - t_1 = \Delta \).

The power can be written as
\begin{align*}
P\left( \max_{0 \leq i \Delta \leq l} \frac{Z_{1i\Delta}}{\sigma} + \max_{0 \leq j \Delta \leq l} \frac{Z_{1j\Delta}}{\sigma} > b\sqrt{2} \right) &= \\
P\left( \max\left(\frac{Z_{1s1}}{\sigma}, \frac{Z_{1s2}}{\sigma}\right) + \max\left(\frac{Z_{1t1}}{\sigma}, \frac{Z_{1t2}}{\sigma}\right) > b\sqrt{2} \right) \\
+ \iiint_B P\left( \max_{0 \leq i \Delta \leq l} \frac{Z_{1i\Delta}}{\sigma} + \max_{0 \leq j \Delta \leq l} \frac{Z_{1j\Delta}}{\sigma} > b\sqrt{2} \right) \left| \begin{array}{c}
\frac{Z_{1s1}}{\sigma} = \frac{b}{\sqrt{2}} - x_1, \frac{Z_{1t1}}{\sigma} = \frac{b}{\sqrt{2}} - y_1 \\
\frac{Z_{1s1}}{\sigma} = \frac{b}{\sqrt{2}} - x_2, \frac{Z_{1t1}}{\sigma} = \frac{b}{\sqrt{2}} - y_2
\end{array} \right.
\right) \left( \frac{Z_{2t1}}{\sigma} \in \left[ \frac{b}{\sqrt{2}}, \frac{d_{y_1}}{\sqrt{2}} \right], \frac{Z_{2t1}}{\sigma} \in \left[ \frac{b}{\sqrt{2}}, \frac{d_{y_2}}{\sqrt{2}} \right] \right) \right)
\end{align*}

where \( B = \{(x_1, x_2, y_1, y_2) : \min(x_1, x_2) + \min(y_1, y_2) > 0\} \). The first part can be written as

\begin{align*}
P\left( \max\left(\frac{Z_{1s1}}{\sigma}, \frac{Z_{1s2}}{\sigma}\right) + \max\left(\frac{Z_{1t1}}{\sigma}, \frac{Z_{1t2}}{\sigma}\right) > b\sqrt{2} \right) &= 1 - P\left( \max\left(\frac{Z_{1s1}}{\sigma}, \frac{Z_{1s2}}{\sigma}\right) + \max\left(\frac{Z_{1t1}}{\sigma}, \frac{Z_{1t2}}{\sigma}\right) < b\sqrt{2} \right) \\
&= 1 - \int_0^\infty \int_0^{\min(x_1, x_2)} P\left( \max\left(\frac{Z_{2t1}}{\sigma}, \frac{Z_{2t2}}{\sigma}\right) < b\sqrt{2} - \min(x_1, x_2) \right) \left( \frac{Z_{1s1}}{\sigma} \in \left[ \frac{b}{\sqrt{2}}, \frac{d_{y_1}}{\sqrt{2}} \right], \frac{Z_{1s1}}{\sigma} \in \left[ \frac{b}{\sqrt{2}}, \frac{d_{y_2}}{\sqrt{2}} \right] \right) \right)
\end{align*}

If we further split the range of integration into \( x_1 \leq x_2 \) and \( x_1 > x_2 \), this corresponds to \( 1 - I_1(s_1, s_2, t_1, t_2) - I_1(s_2, s_1, t_1, t_2) \) in approximation (4.15).

The second part (quadruple integral) can be reduced to a double integral by using the approximation

\begin{align*}
P\left( \max_{0 \leq i \Delta \leq l} \frac{Z_{1i\Delta}}{\sigma} + \max_{0 \leq j \Delta \leq l} \frac{Z_{1j\Delta}}{\sigma} > b\sqrt{2} \right) &= \frac{Z_{1s1}}{\sigma} = \frac{b}{\sqrt{2}} - x_1, \frac{Z_{1t1}}{\sigma} = \frac{b}{\sqrt{2}} - y_1 \\
&= \nu e^{-\frac{b}{\sqrt{2}}(\min(x)+y_1)} + \nu e^{-\frac{b}{\sqrt{2}}(\min(x)+y_2)} - \nu^2 e^{-\frac{b}{\sqrt{2}}(2\min(x)+y_1+y_2)} \\
&+ \nu (1 - \nu)^2 e^{-\frac{b}{\sqrt{2}}(x_1+\min(y))} + \nu (1 - \nu)^2 e^{-\frac{b}{\sqrt{2}}(x_2 \min(y))} - \nu^2 e^{-\frac{b}{\sqrt{2}}(x_1+x_2+2\min(y))} \\
&+ \frac{b}{\sqrt{2}} \nu^2 (\min(x) + \min(y)) [e^{-\frac{b}{\sqrt{2}}(x_1+y_1)} + e^{-\frac{b}{\sqrt{2}}(x_1+y_2)} + e^{-\frac{b}{\sqrt{2}}(x_2+y_1)} + e^{-\frac{b}{\sqrt{2}}(x_2+y_2)}],
\end{align*}
where \( \min(x) = \min(x_1, x_2) \), \( \min(y) = \min(y_1, y_2) \) and \( \nu = \nu(b\sqrt{\beta \Delta}) \). Splitting the range of integration into the following four subsets

\[
\{x_1 \leq x_2, y_1 \leq y_2\}, \{x_1 \leq x_2, y_1 > y_2\}, \{x_1 > x_2, y_1 \leq y_2\}, \{x_1 > x_2, y_1 > y_2\},
\]

and integrating over the \( \max(x) \) and \( \max(y) \) we obtain

\[
I_2(s_1, s_2, t_1, t_2) + I_2(s_2, s_1, t_1, t_2) + I_2(s_1, s_2, t_2, t_1) + I_2(s_2, s_1, t_2, t_1),
\]

which is the second part of approximation (4.15).

The same method of breaking the power into two parts can be used to derive the power approximations for 3 loci.
Appendix B

Derivation of approximation (7.9)

We will show that for $X_{1,i\Delta}$ and $X_{2,i\Delta}$ independent Ornstein-Uhlenbeck processes with covariance function $R_1(t) = e^{-\beta_1 t}$ and $R_2(t) = e^{-\beta_2 t}$ respectively and large $b$,

$$P\left\{ \max_{0 \leq i \Delta \leq t} \|X_{i\Delta}\| \geq b \right\} \approx e^{-\frac{b^2}{2\Delta}} + b^2 l \left( \frac{\beta_1 + \beta_2}{2} \right) \nu \left( b \sqrt{\Delta (\beta_1 + \beta_2)} \right) e^{-\frac{b^2}{2\Delta}}. \quad (B.1)$$

The above is equivalent to approximation (7.9) for the tail distribution of the test statistic used for locating QTLs.

Derivation of approximation (7.9)

Define $D_i = \{ j \in Z : j \geq 1, (i+j)\Delta \leq l \}$, where $l$ and $\Delta$ are fixed. Let $\omega_{i\Delta}$ be the angle between $X_{1,i\Delta}$ and $X_{2,i\Delta}$. Then, we can write

$$P\left\{ \max_{0 \leq i \Delta \leq t} \|X_{i\Delta}\| \geq b \right\}$$

$$= \sum_{i=0}^{l/\Delta} P\left\{ \|X_{i\Delta}\| \geq b, \|X_{(i+j)\Delta}\| < b \ \forall j \in D_i \right\}$$

$$= \sum_{i=0}^{l/\Delta} \int_{0}^{\pi} P\left\{ \|X_{i\Delta}\| \geq b, \|X_{(i+j)\Delta}\| < b \ \forall j \in D_i, \ \omega_{i\Delta} \in dw \right\}$$

$$= \sum_{i=0}^{l/\Delta} \int_{0}^{\pi} \int_{b+y}^{\infty} P\left\{ \|X_{i\Delta}\| \in b+dy, \ \omega_{i\Delta} \in dw \right\} P\left\{ \|X_{(i+j)\Delta}\| < b \ \forall j \in D_i, \ \|X_{i\Delta}\| = b+y, \ \omega_{i\Delta} = w \right\}$$

(B.2)
APPENDIX B. DERIVATION OF APPROXIMATION (7.9)

We will find an approximation for both terms in this above expression and then perform the integration and summation step.

By the independence of $X_1$ and $X_2$ and the Gaussian nature of the processes, the first term is

$$P\left\{ ||X_i\Delta|| \in b + dy, \omega_i\Delta \in dw \right\} = \frac{1}{2\pi} e^{-\frac{1}{2} (b+y)^2 (b + y)} \ dy \ dx$$

$$\approx \frac{1}{\sqrt{2\pi}} b \phi(b + y) \ dy \ dx, \quad \text{for } y = O(1/b).$$

(B.3)

If we want to substitute (B.3) into (B.2), it is sufficient to get an approximation that is good for $y = O(1/b)$, since the second probability in the integral is non-negligible only for very small values of $y$.

The second term in the integral is a bit more difficult to evaluate. We can write

$$P\left\{ ||X_{(i+j)}\Delta|| < b \ \forall j \in D_i \mid ||X_i\Delta|| = b + y, \omega_i\Delta = w \right\}$$

$$= \ P\left\{ ||X_{(i+j)}\Delta|| (\cos^2 w + \sin^2 w) < b \ \forall j \in D_i \mid ||X_i\Delta|| = b + y, \omega_i\Delta = w \right\}$$

$$\approx P\left\{ X_{1,(i+j)}\Delta \cos w + X_{2,(i+j)}\Delta \sin w < b \ \forall j \in D_i \mid ||X_i\Delta|| = b + y, \omega_i\Delta = w \right\}$$

$$= P\left\{ \tilde{X}_{(i+j)}\Delta - \tilde{X}_{i}\Delta < b - (b + y) \ \forall j \in D_i \mid ||X_i\Delta|| = b + y, \omega_i\Delta = w \right\},$$

(B.4)

where $\tilde{X}_{k}\Delta = X_{1,k}\Delta \cos w + X_{2,k}\Delta \sin w.$

For small values of $j$ (fixed), $\tilde{X}_{(i+j)}\Delta - \tilde{X}_{i}\Delta$ has approximately the same distribution as does the sum of $j$ independent normal random variables, each with mean $[-b\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w)]$ and variance $[2\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w)]$. Therefore, the last line of equation (B.4) is approximately equal to

$$P_{-\mu,\sigma} \left\{ \max_{0 \leq (i+j)\Delta < l} S_j < -y \right\},$$

(B.5)

where $S_j$ is the sum of $j$ independent normal random variables with mean $-\mu$ and variance $\sigma^2$; $\mu = b\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w)$ and $\sigma^2 = 2\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w)$. 


Substituting (B.3) and (B.5) into (B.2), we obtain

\[ P \left\{ \max_{0 \leq t \leq T} ||X_t \Delta|| \geq b \right\} \]

\[ \approx \sum_{i=0}^{\lfloor \Delta \rfloor} \int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} b \phi(b + y) P_{\mu,\sigma} \left\{ \min_{j \geq 0} S_j > y \right\} dy \, dw \]

\[ \approx \sum_{i=0}^{\lfloor \Delta \rfloor} \int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} b \phi(b) \int_{0}^{\infty} e^{-by} P_{\mu,\sigma} \left\{ \min_{j \geq 0} S_j > y \right\} dy \, dw \]

\[ \approx \sum_{i=0}^{\lfloor \Delta \rfloor} \int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} b^2 \phi(b) \Delta (\beta_1 \cos^2 w + \beta_2 \sin^2 w) \nu \left( b \sqrt{2\Delta (\beta_1 \cos^2 w + \beta_2 \sin^2 w)} \right) \]

\[ = \frac{1}{\sqrt{2\pi}} b^2 \phi(b) \int_{-\pi}^{\pi} (\beta_1 \cos^2 w + \beta_2 \sin^2 w) \nu \left( b \sqrt{2\Delta (\beta_1 \cos^2 w + \beta_2 \sin^2 w)} \right) \]

\[ \approx b^2 e^{-\frac{1}{2} b^2 \left( \frac{\beta_1 + \beta_2}{2} \right)} \nu \left( b \sqrt{\Delta (\beta_1 + \beta_2)} \right), \]

(B.6)

where the function \( \nu(x) \) as defined on p. 82 of Siegmund (1985) and can be approximated by \( e^{-0.583x} \). The second line is obtained by ignoring the term \( e^{-y^2/2} \), which is negligible when \( y = O(1/b) \). The third line follows from Corollary 8.45 of Siegmund (1985), the fourth line from moving the summation sign inside the integral and the final line by integrating over \( w \).
Appendix C

Derivation of Lemma 8.1

In this appendix, we show how to obtain the approximation in Lemma 8.1. We will first restate the lemma and then show its derivation.

Lemma 8.1 Let

\[ Z_t = \begin{pmatrix} Z_t^{(1)} \\ Z_t^{(2)} \end{pmatrix}, \]

where \( Z_t^{(1)} \) and \( Z_t^{(2)} \) are independent Gaussian processes with covariance function

\[ R_t(t) = 1 - \beta_i |t| + o(|t|) \quad \text{as } t \to 0. \]

Let \( 0 < ||x||^2 < b^2 \) and define \( t^* \) and \( \omega^* \) to be the positive solution of

\[ \begin{pmatrix} \frac{z_1}{R_1(t^*)} \\ \frac{z_2}{R_2(t^*)} \end{pmatrix} = \begin{pmatrix} b \cos \omega^* \\ b \sin \omega^* \end{pmatrix}. \]

Assume \( ||x|| \) and \( b \) to be large and \( t^* \) is contained in \((0,t_1)\) and is bounded away from the upper endpoint \((t_1 > 0)\). Then

\[ P\{\max_{0 \leq i \leq t} ||Z_i\| \geq b \mid Z_0 = x\} \approx \]

\[ e^{-\frac{1}{2}(b^2 - ||x||^2)} \left( \beta_1 \cos^2 \omega^* + \beta_2 \sin^2 \omega^* \right) \nu(b [(\beta_1 + \beta_2)\Delta]^{\frac{1}{2}}), \]

where \( \dot{R}_i(t) = \frac{dR_i(t)}{dt} \).
Derivation of Lemma 8.1

Define $D_t = \{ j \in \mathbb{Z} : j \geq 1, (i + j)\Delta \leq l \}$, where $l$ and $\Delta$ are fixed. Let $\omega_{i\Delta}$ be the angle between $Z_{i\Delta}^{(1)}$ and $X_{i\Delta}^{(2)}$. Then, we can write

$$
P\left\{ \max_{0 \leq i \leq l} ||Z_{i\Delta}|| > b \mid Z_0 = x \right\}
= \sum_{i=0}^{l/\Delta} P\left\{ ||Z_{i\Delta}|| \geq b, ||Z_{(i+j)\Delta}|| < b \ \forall j \in D_t \mid Z_0 = x \right\}
= \sum_{i=0}^{l/\Delta} \int_{-\pi}^{\pi} P\left\{ ||Z_{i\Delta}|| \geq b, ||Z_{(i+j)\Delta}|| < b \ \forall j \in D_t, \omega_{i\Delta} \in dw \mid Z_0 = x \right\}
= \sum_{i=0}^{l/\Delta} \int_{\pi}^{\pi} \int_{0}^{\infty} P\left\{ ||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \mid Z_0 = x \right\}
* P\left\{ ||Z_{(i+j)\Delta}|| < b \ \forall j \in D_t \mid ||Z_{i\Delta}|| = b + y, \omega_{i\Delta} = w, Z_0 = x \right\}
= \phi \left( \frac{(b+y) \cos w - x_1 R_1(i\Delta)}{\sqrt{1-R_1^2(i\Delta)}} \right) \phi \left( \frac{(b+y) \sin w - x_2 R_2(i\Delta)}{\sqrt{1-R_2^2(i\Delta)}} \right) \frac{(b+y) \ dy \ dw}{\sqrt{1-R_1^2(i\Delta)}} \sqrt{1-R_2^2(i\Delta)}
$$

Using the fact that $Z_{i\Delta}^{(1)}$ and $Z_{i\Delta}^{(2)}$ are independent and normally distributed for fixed $t$, we have

$$
P\left\{ ||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \mid Z_0 = x \right\}
= \phi \left( \frac{(b+y) \cos w - x_1 R_1(i\Delta)}{\sqrt{1-R_1^2(i\Delta)}} \right) \phi \left( \frac{(b+y) \sin w - x_2 R_2(i\Delta)}{\sqrt{1-R_2^2(i\Delta)}} \right) \frac{(b+y) \ dy \ dw}{\sqrt{1-R_1^2(i\Delta)}} \sqrt{1-R_2^2(i\Delta)}
$$

If we expand the above around $t^*$ and $w^*$ where $t^*$ and $w^*$ are defined in the statement of the lemma, we get

$$
P\left\{ ||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \mid Z_0 = x \right\}
\approx \frac{b e^{-\frac{1}{2} b^2 - ||z||^2} e^{-\frac{1}{2} b^2 ((i\Delta-t^*)^2 a_1 + (w-w^*)^2 a_2 + 2(w-w^*)(i\Delta-t^*) a_3)}}{2\pi \sqrt{1-R_1^2(i\Delta)} \sqrt{1-R_2^2(i\Delta)}},
$$

where

$$
a_1 = \frac{\dot{R}_1(t^*) \cos^2 w^* + \dot{R}_2(t^*) \sin^2 w^*}{1 - R_1^2(t^*) + \frac{\dot{R}_2(t^*)}{1 - R_2^2(t^*)}},
$$
\[ a_2 = \frac{\sin^2 w^*}{1 - R_2^2(t^*)} + \frac{\cos^2 w^*}{1 - R_2^2(t^*)} - 1, \]
\[ a_3 = \cos w^* \sin w^* \left[ \frac{R_2^2(t^*) R_2^2(t^*)}{1 - R_2^2(t^*)} + \frac{\hat{R}_2^2(t^*) R_2^2(t^*)}{1 - R_2^2(t^*)} \right]. \]

The second probability in (C.1) can be treated in a manner similar to that of Appendix A i.e.,

\[ P \left\{ ||Z_{i+j}\Delta|| < b \; \forall j \in D_i \mid ||Z_i\Delta|| = b + y, \omega_i\Delta = w, Z_0 = x \right\} \]
\[ = P \left\{ ||Z_{i+j}\Delta|| (\cos^2 w + \sin^2 w) < b \; \forall j \in D_i \mid ||Z_i\Delta|| = b + y, \omega_i\Delta = w, Z_0 = x \right\} \]
\[ \approx P \left\{ Z_{i+j}\Delta \cos w + Z_{i+j}\Delta \sin w < b \; \forall j \in D_i \mid ||Z_i\Delta|| = b + y, \omega_i\Delta = w, Z_0 = x \right\} \]
\[ = P \left\{ \hat{Z}_{i+j}\Delta - \hat{Z}_{i}\Delta < -y \; \forall j \in D_i \mid ||Z_i\Delta|| = b + y, \omega_i\Delta = w, Z_0 = x \right\}, \]

(C.3)

where \( \hat{Z}_{k\Delta} = Z_{k\Delta}^{(1)} \cos w + Z_{k\Delta}^{(2)} \sin w. \)

If we Taylor expand the conditional mean and variance of \( \hat{Z}_{i+j}\Delta - \hat{Z}_{i}\Delta \) around \( t^* \) and \( w^* \), we find that for \( i\Delta \to t^* \),

\[ E \left[ \hat{Z}_{i+j}\Delta - \hat{Z}_{i}\Delta \mid ||Z_i\Delta|| = b + y, \omega_i\Delta = w, Z_0 = x \right] \approx -j\Delta \beta_1 \cos^2 w + \beta_2 \sin^2 w \]

and

\[ \text{Cov} \left[ \hat{Z}_{i+j}\Delta - \hat{Z}_{i}\Delta, \hat{Z}_{(i+j+k)\Delta} - \hat{Z}_{i}\Delta \mid ||Z_i\Delta|| = b + y, \omega_i\Delta = w, Z_0 = x \right] \]
\[ \approx 2j\Delta \beta_1 \cos^2 w + \beta_2 \sin^2 w, \]

so that \( \hat{Z}_{i+j}\Delta - \hat{Z}_{i}\Delta \) has approximately the same distribution as the sum of \( j \) independent normal random variables with mean \(-\Delta \beta_1 \cos^2 w + \beta_2 \sin^2 w\) and variance \( 2\Delta (\beta_1 \cos^2 w + \beta_2 \sin^2 w) \). Therefore, we can rewrite (C.3) as

\[ P_{-\mu, \sigma} \left\{ \max_{0 \leq i+j\Delta < l} S_j < -y \right\}. \]

(C.4)
Substituting (C.2) and (C.4) into (C.1) we obtain

\[ P\left\{ \max_{i} ||Z_i|| \geq b \mid Z_0 = x \right\} \]

\[ = \sum_{i=0}^{l+i} \int_{0}^{\pi} P\left\{ ||Z_{i\Delta}|| \leq b \mid \omega_{i\Delta} \in dw, Z_0 = x \right\} \]

\[ \ast P\left\{ ||Z_{(i+j)\Delta}|| \leq b \mid \omega_{i\Delta} = w, Z_0 = x \right\} \]

\[ \approx \sum_{i=0}^{l+i} \int_{0}^{\pi} \frac{be^{-\frac{1}{2}b^2||x||^2}e^{-\frac{1}{2}b^2[(i\Delta-t^*)^2a_1+(w-w^*)^2a_2+2(w-w^*)(i\Delta-t^*)^2a_3]}}{2\pi \sqrt{1-R_1^2(i\Delta)} \sqrt{1-R_2^2(i\Delta)}} b\Delta \bar{\beta} \nu(b\sqrt{2\Delta\bar{\beta}}) dw \]

\[ \approx \frac{e^{-\frac{1}{2}(b^2-||x||^2)}}{R_1(t^*) R_2(t^*) \cos^2 w^* + R_1(t^*) R_2(t^*) \sin^2 w^*} \nu(b[(\beta_1+\beta_2)\Delta]^{\frac{3}{2}}), \]

where \( \bar{\beta} = \beta_1 \cos^2 w^* + \beta_2 \sin^2 w^* \). The third line follows from Corollary 8.45 of Siegmund (1985). The last line is obtained by replacing \( i\Delta \) by \( t^* \) in the denominator and performing the summation first, then the integral. To perform the integral, the following identity was used:

\[ \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(ax^2+by^2+2cxy)} dx dy = \frac{1}{\sqrt{ab-c^2}}. \]
Bibliography


