Investigating a Variance-Components Approach for Linkage Analysis in Quantitative Traits

by

Lisa Kuramoto

B.Sc., University of British Columbia, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

September 2002

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct 10, 2002

DE-6 (2/88)

Abstract

Model-based linkage methods have had limited success in locating quantitative trait loci (QTLs) in complex traits since the underlying genetic mechanisms are not well known. As a result, robust or model-free approaches for detecting linkage have grown in popularity. We discuss a mixed effects model, which involves the estimation of genetic and non-genetic variance components, as well as recombination fractions. Using the Genometric Analysis Simulation Program (GASP), we first attempt to investigate the properties of this method on simple traits, which differ in terms of their variance components. To further understand its performance in a complex setting, we apply this method to simulated, familial data for an oligogenic disease with quantitative risk factors from the 10th Genetic Analysis Workshop (GAW10). We see that the ability of the variance-components approach to map QTLs depends on the amount of variability it contributes to the quantitative trait. As well, we find that the presence of the recombination fraction in the model results in consistent estimates of the variance components across the chromosome; however, it does not seem to improve the mapping ability of the model.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Dr. Harry Joe, for his expert guidance, generous support, and continuous inspiration, which greatly contributed to enhancing my statistical knowledge and research skills. In particular, Dr. Joe provided me with a tremendous amount of computer expertise which helped to address the computational challenges posed by this thesis. I would also like to thank my second reader, Dr. Nancy Heckman, for her thoughtful feedback and valuable suggestions which aided me to present the blend of statistics and genetics in this thesis much more clearly. Finally, I would like to thank all of the graduate students who contributed to my warm, extraordinary, character-building memories that I will cherish forever.

LISA KURAMOTO

The University of British Columbia
September 2002

Chapter 1

Introduction

With the growth of technology and wealth of genetic information, the role that statistics plays in quantitative genetics has flourished. Quantitative genetics involves studying genetic expression through continuous phenotypes (physical characteristics), which are known as quantitative traits. These traits may be influenced by multiple genes, and even environmental effects. In contrast to quantitative traits, phenotypes which fall into distinct classes and follow Mendelian laws of inheritance are known as Mendelian traits. For example, whether one has the ability to curl his or her tongue is a Mendelian trait. In general, Mendel's laws of inheritance describe how characteristics are passed down from parent to offspring. While statistical methods have helped us to thoroughly understand simple Mendelian traits, the challenge to reveal the genetic mechanisms underlying complex traits, especially those exhibited in humans, still remains. A complex trait is a "genetic condition whose mode of inheritance does not follow any of the known Mendelian laws" (Ott [22]); therefore, existing methods for the analysis of Mendelian traits have serious limitations in a more complex setting. As with all quantitative traits, an element of difficulty which arises when studying complex traits is that they may be controlled by more than one major gene, some polygenes, as well as interactions with the environment, just to name a few. Note that polygenes are genes that collectively have an effect on a trait. Some examples of complex diseases are schizophrenia, bipolar disease, and diabetes. Although we have some understanding of complex diseases, we would like to deepen this understanding by learning about the genetic mechanisms, if any, that are involved, or as Ott [22] states, in the context of schizophrenia: "the question is whether this diagnosis is genetically relevant or

whether the hypothesized underlying genes act on a wider or more narrowly defined phenotype." To answer this question, a variety of techniques has been proposed and studied, one of which is Amos' [3] robust variance-components approach to linkage analysis.

Variance-components methods attempt to partition the total variation of a quantitative trait into its genetic and non-genetic components. The origin of variance-components methods comes from the work of R. A. Fisher. In one of his classic papers, Fisher [10] proposed decomposing genotypic values into its components. Fisher [10] also realized that under the assumption that genetic and non-genetic factors are uncorrelated with each other, the analysis of variance, ANOVA, partitions the total phenotypic variance into the sum of its components. A major advantage of ANOVA is that it does not rely on the assumption of normality for estimation purposes. However, in the ANOVA setting balanced data are preferred, so it is difficult to jointly analyze families with different types of relationships, such as siblings and cousins. As variance-components approaches evolved, likelihood-based methods came to be favoured since they do not require balance and are able to handle arbitrary pedigrees; however, these methods usually assume that a family's traits follow a multivariate normal distribution, which may be a strong assumption. By modelling the familial correlation induced by any major genes underlying the phenotype of interest, modern variance-components methods have the advantage of not requiring the specification of a detailed genetic model involving allele frequencies or penetrances. Note that alleles are the various forms of a gene which may arise, and penetrance specifies the proportion of individuals with a certain genotype (genetic characteristic) who actually exhibit the characteristics of that genotype.

Variance-components techniques are currently used in the linkage analysis of quantitative traits. Linkage analysis is used to infer locations on chromosomes where major genes, which control a trait, reside relative to genetic markers. Genetic markers are genes with a known location along the chromosome so that we can observe the types of alleles that are present. Note that we do not know where major genes that control a quantitative trait reside, so we do not directly observe the types of alleles present at this location. The main idea underlying linkage analysis is that relatives, who have similar phenotypes, will have identical genes at the genetic marker only if the major gene controlling that phenotype is linked to the marker. Therefore, we are interested in markers which are tightly linked to a major gene. How tightly linked a marker

is to a major gene underlying the trait of interest is determined by the recombination fraction. Recombination is a phenomenon where genetic material is rearranged so that the materials inherited from parent to offspring at locations on a chromosome are not exactly identical. One of the earliest forms of linkage analysis involves simply counting the number of recombinants and non-recombinants. When there is no linkage recombination occurs about half of the time. Testing whether the estimated recombination fraction is less than $1/2$ gives an indication of linkage, since a recombination fraction close to 0 implies tight linkage. However, this primitive approach does not always suffice because of counting problems due to complications. For example, sometimes we do not know if a recombination occurred because the recombinant and non-recombinant genetic material looks the same. Current linkage analysis techniques involve maximum likelihood estimation and likelihood ratios. Likelihood techniques do not break down in the presence of incomplete information, so exact counts of recombinants and non-recombinants are not required.

Segregation analysis is performed before linkage analysis when using linkage methods that require one to specify a model for the genetic mechanism underlying the trait of interest. Segregation describes the separation and inheritance of alleles during reproduction. In particular, segregation analysis is used to determine the mode of inheritance and allele frequencies. Various methods have been devised to determine whether the mode of inheritance is dominant or recessive, for instance. This segregation phenomenon was first discovered by Mendel, the father of genetics, who formulated the law of segregation, which states that each of the two genes from a parent is equally likely to be passed on to an offspring (see Edlin [7], for example). Through his experiments with peas, Mendel observed the proportions of different discrete phenotypic characteristics, called segregation ratios, and was able to deduce modes of inheritance for simple traits. Segregation analysis is usually performed prior to using linkage methods that require the specification of a detailed genetic model; however, variance-components based linkage methods have alleviated the need for segregation analysis.

Association analysis is also jointly used with linkage analysis. Allelic association is the "excessive co-occurrence of certain combinations of alleles" which may be due to tight linkage (Sham [26]). If there exists tight linkage, then segments of chromosomes may be inherited from generation to generation unchanged. By detecting excessively frequent segments of chro-

mosomes in the population, association analysis also helps to map major genes. Risch and Merikangas [24] argue that major genes which do not have a large effect on the trait of interest are difficult to detect via linkage analysis, and advocate that association analysis is more powerful for fine mapping loci. While association analysis may be better at detecting major genes with small effects, it has the disadvantage of requiring one to identify a candidate gene before carrying out the analysis.

In the case of complex traits, it is nearly impossible to correctly specify the genetic model which governs the trait of interest; therefore, linkage methods which are not model dependent have grown in popularity. Modern linkage analysis methods alleviate the need for segregation analysis, which may prove to be a burden if incorrect genetic models are estimated. An approach to linkage analysis that does not require the specification of a detailed genetic model is said to be robust. In his paper, Amos [3] developed a robust variance-components approach to linkage analysis. Through his method, he uses a mixed effects model to decompose the total phenotypic variance into its genetic and non-genetic components without any major limiting assumptions on the underlying genetic mechanism. By modelling the correlation between relatives, Amos' approach also has the ability to handle arbitrary pedigrees.

The objective of this thesis is to investigate the ability of Amos' model to detect major genes which control a quantitative trait. We also determine the accuracy of the variance-components estimates from his model. These properties are compared with those of a sub-model, the so-called two-point model (Blangero et al. [5]), which puts a constraint on one of the parameters, namely the recombination fraction. In Chapter 2, we define some genetic terms and illustrate some concepts which are important to linkage analysis. Haseman and Elston [15] made a significant contribution to the linkage analysis of quantitative traits, so their method is presented in Chapter 3. In Chapter 4, we provide some theoretical details on the variance-components approach to linkage analysis and introduce Amos' model. Next, we evaluate the performance of the variance-components methods through some simulations, which involve a quantitative trait with one major gene, in Chapter 5. In Chapter 6, we use Amos' method to assess a simulated oligogenic disease with complex quantitative risk factors from the 10th Genetic Analysis Workshop. We are interested in analyzing this oligogenic disease since the quantitative traits which influence this disease are controlled by more than one major gene, as

well as interactions between them. Finally, in Chapter 7, we provide a summary and discussion of Amos' robust approach to linkage analysis in the context of complex traits.

Chapter 2

Genetics Background

This chapter is intended to introduce some terminology and concepts which pertain to genetics in the linkage analysis setting. We begin by reinforcing some basic genetics terminology, such as allele and locus. Next, we introduce some concepts which play key roles in understanding linkage analysis. Once the basic idea of linkage analysis is introduced, we elaborate on this topic to describe linkage analysis in the context of quantitative traits, as opposed to the well-studied Mendelian traits.

2.1 Terminology

To be able to understand the reasoning behind linkage analysis, it is important to know certain terms and be familiar with some genetic details of inheritance. It is well-known that chromosomes carry genetic information. A chromosome is composed of numerous genes which are the basic units of heredity. Humans have 23 pairs of chromosomes, one pair consists of the sex chromosomes, and the remaining 22 pairs are called autosomes. During reproduction, one chromosome from each pair is inherited from the mother and the other chromosome is inherited from the father. These pairs of chromosomes are called homologous pairs because, with the exception of the sex chromosomes, they have the shape shape. Since each cell contains 46 chromosomes, cell division must occur to ensure that the number of chromosomes from generation to generation remains constant. Due to a process called meiosis, chromosomes undergo a series of stages which result in the production of gametes. *Gametes* contain one chromosome

from each homologous pair. During meiosis, the chromosomes duplicate themselves and remain attached at a location called the centromere. Next, the homologous chromosomes pair up and may crossover. Division then occurs twice, once to divide the pairs of homologous chromosomes, and once to divide the chromosomes at their centromeres. So, meiosis results in four gametes with once copy of each chromosome, which may have different combinations of genes from the original chromosomes due to crossing-over. Note that an offspring receives one gamete from each parent to obtain a full set of chromosomes. We refer to the gamete from the mother and the father as a maternal and paternal gamete, respectively. When these two gametes are fused together it is referred to as a zygote.

Linkage analysis revolves around the characteristics of genes on a chromosome, so we introduce some related terminology. Firstly, the specific position of a gene along a chromosome is called the *locus*. Note that the locus for a certain gene is the same on two homologous chromosomes. Secondly, genes on a chromosome have various forms and these alternative forms are called *alleles*. At a given locus, every person receives one allele from each parent; therefore, each locus has two alleles. A genetic characteristic, such as allele type, is referred to as a *genotype*; whereas, a physical characteristic, such as height, is referred to as a *phenotype*. Thirdly, genetic *markers* are genes along the chromosome where we observe the types of alleles that are present. Of particular interest, in the context of quantitative traits, are *major genes* because they govern a portion of the variation in a trait. Markers provide valuable information when trying to detect major genes because of linkage of nearby markers to major genes in transmission from parent to offspring. Lastly, we will see that linkage analysis relies heavily on familial data or pedigrees, so it is necessary to distinguish between types of pedigrees. A *nuclear pedigree* consists of two generations of relatives, namely a father, a mother, and their offspring. On the other hand, a pedigree which has two sets of grandparents, one set of parents, and their offspring is called a *ceph pedigree*. Finally, we refer to any pedigree which has a more complex family structure than a ceph pedigree as an *extended pedigree*.

2.2 Basics of Linkage Analysis

Mendel's law of assortment states that the segregation of genes from one trait is independent from the segregation of genes from other traits (for example, see Edlin [7]). However, it is now known that this law mainly applies to loci on separate chromosomes. For loci on the same chromosome, the simultaneous transmission of genes may not be independent, so we say that these loci exhibit the phenomenon of linkage. In general, genes on the same chromosome tend to be inherited together, and this tendency decreases with the distance between loci. The goal of linkage analysis is to "infer relative positions of two or more loci by examining transmission from parent to offspring or allele sharing patterns of relatives" (Sham [26]). Linkage analysis has been built from essential genetics concepts, such as identical-by-descent and recombination. In this section, we will describe some important concepts which recur throughout the remaining chapters.

2.2.1 Identical-by-Descent

Firstly, two relatives are said to share alleles *identical-by-descent* (ibd) at a locus if they have the same form of the gene and that allele is descended from a common ancestor. Note that sharing alleles identical-by-descent differs from sharing alleles identical-by-state, which only requires two relatives to share the same form of a gene at a locus. At each locus, relative pairs may share 0, 1, or 2 alleles identical-by-descent. For example, consider the pedigree in Figure 2.1, which shows the genotype for each family member at a locus for a gene with two alleles, B and b. The nuclear family consists of a father and mother, who both have genotype Bb, as well as three offspring with genotypes BB, Bb, and Bb. Because each offspring inherits one gene from each parent, each parent-offspring pair shares exactly one allele ibd. The first and second siblings with genotypes BB and Bb share exactly one allele ibd. This is apparent since the first sibling inherited the unique B allele from both parents, and the second sibling must have inherited her only B allele from either the mother or the father. In the case of the second and third siblings, the number of alleles shared ibd is ambiguous. This sib-pair either shares 0 or 2 alleles ibd.

As opposed to examining actual counts, it is customary to study the proportion of alleles

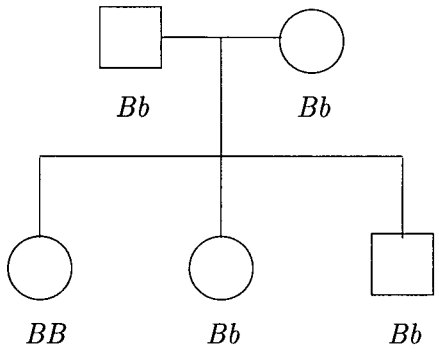

Figure 2.1: Nuclear Pedigree with Genotypes at a Locus with Two Alleles

shared ibd between relative pairs. If we have incomplete information, as we do in the case of the second and third siblings, then the conditional expectation of sharing 0, 1, or 2 alleles ibd between the jth relative pair given the locus information is denoted by π_j and may be estimated as $\pi_j = f_1/2 + f_2$, where f_i is the probability that the relative pair shares exactly i allele(s) ibd at a locus given the genotypes. In our example, the second and third siblings share 0 of their alleles ibd with probability 1/2, or share 2 of their alleles ibd with probability 1/2. So, they are estimated to share 1/2 of their alleles ibd. Various algorithms have been developed to rapidly compute the proportion of alleles ibd. Haseman and Elston [15] proposed an algorithm for computing the proportion of alleles shared ibd if some genotypes are unknown.

2.2.2 Coefficient of Relationship

Measures of relation play an important role in linkage analysis. In particular, the *coefficient of relationship* surfaces in the familial covariance structure. The coefficient of relationship is defined to be the expected proportion of genes that two relatives share ibd at a locus. To compute the coefficient of relationship, we must first determine the probability mass function for the number of genes shared ibd. Under the assumption of no inbreeding, a relative pair may share 0, 1, or 2 genes ibd at a locus. Let S be a discrete random variable for the number of genes shared ibd between two relatives at a locus and let f_i^* be the probability that two

relatives share i genes ibd. Then for any relative pair, the probability mass function for the number of genes shared ibd is shown in Table 2.1. From Table 2.1, we see that the coefficient of relationship, 2ϕ is

$$\begin{aligned} \mathrm{E}(S) &= 0 \times f_0^* + 1 \times f_1^* + 2 \times f_2^* \\ &= f_1^* + 2f_2^* \\ &\equiv 2\phi \end{aligned}$$

For illustrative purposes, we compute the coefficient of relationship for a sib-sib pair. Suppose the genotype of the mother and father are M_1M_2 and F_1F_2, respectively. Each offspring's genotype will consist of one maternal gene and one paternal gene; furthermore, each maternal gene is equally likely to be passed to the offspring, and likewise for the paternal genes. From these parents, offspring must have one of four possible genotypes: M_1F_1, M_1F_2, M_2F_1, M_2F_2. Therefore, there are 16 possible sib-sib pairs. Through a simple counting process, it can be seen that of the 16 possible sib-sib pairs, four pairs have 0 genes ibd, eight pairs have 1 gene ibd, and four pairs have 2 genes ibd; hence, $f_0^* = 1/4$, $f_1^* = 1/2$, and $f_2^* = 1/4$ for sib-sib pairs. So, we see that the coefficient of relationship is $1/2$. The probabilities of sharing genes ibd at a locus and the coefficient of relationship for sib-sib pairs and other relative pairs are given in Table 2.2. Note that the coefficient of relationship decreases by a factor of $1/2$ as the degree of relationship between two family members increases.

We also mention another measure of relation, which should not be confused with the coefficient of relationship, namely the coefficient of kinship. The coefficient of kinship is the probability that a randomly selected gene from one individual is ibd to a randomly selected gene from another individual. In non-inbred populations, the coefficient of relationship is twice the coefficient of kinship. This is the reasoning behind denoting the coefficient of kinship by ϕ and the coefficient of relationship by 2ϕ.

Table 2.1: Probability Mass Function for Number of Genes Shared Ibd at a Locus

$S = s$	0	1	2
$\Pr(S = s)$	f_0^*	f_1^*	f_2^*

Table 2.2: Probability of Sharing Genes Ibd and Coefficient of Relationships for Different Relative Pairs

Relationship	f_0^*	f_1^*	f_2^*	2ϕ
Sibs	1/4	1/2	1/4	1/2
Half-Sib	1/2	1/2	0	1/4
Grandparental	1/2	1/2	0	1/4
Avuncular	1/2	1/2	0	1/4
First Cousin	3/4	1/4	0	1/8

2.2.3 Recombination Fraction

Another fundamental concept in linkage analysis is recombination, since it is partially responsible for the genetic variation in phenotypes that we observe between generations. During meiosis, chromosomes may exchange genetic information due to physically crossing-over. Therefore, the produced gametes may have new gene combinations and are referred to as recombinants. To illustrate the phenomenon of recombination, consider the pedigree in Figure 2.2, which shows the genotype for each family member at two loci. The first locus has two alleles, B and b, and the second locus has two alleles, C and c. From Figure 2.2, we can see that the mother has genotype Bb at the first locus, and genotype cC at the second locus. Furthermore, the mother's paternal gamete is Bc and the maternal gamete is bC. Each offspring receives one gamete from each parent. If an offspring receives a Bc or bC gamete from the mother, then he or she has received a non-recombinant gamete. Alternatively, if an offspring receives a BC or bc gamete from the mother then he or she has received a recombinant gamete, and the two loci are said to have undergone recombination. In Figure 2.2, we see that the first and third offspring received non-recombinant gametes from their mother, but the second sibling received a recombinant gamete. We also note that we cannot easily determine whether each offspring received a non-recombinant or recombinant gamete from their father.

We quantify the phenomenon of recombination by the recombination fraction. The *recombination fraction* between two loci, θ, is the probability that a gamete is recombinant, or equivalently the probability that there are an odd number of crossovers occurring between the two loci. If two loci are on different chromosomes then we expect them to segregate independently, so the expected recombination fraction is 1/2. In the case where two loci are on the same chromosome, the physical distance between them is directly related to the chance

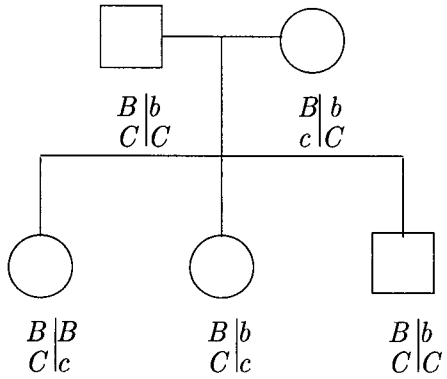

Figure 2.2: Nuclear Pedigree with Genotypes at Two Loci with Two Alleles Each

of recombination. So recombination is less likely to occur if the loci are close together. We say that a recombination fraction which is less than 1/2 implies that the two loci are linked. Furthermore, the smaller the recombination fraction is between two loci, the more tightly linked they are thought to be. Although the recombination fraction provides us with a sense of how close two loci reside on a chromosome, this does not directly translate into a measure of distance. Recombination fractions cannot be used as a measure of distance since they are not additive.

Genetic map distance, which is measured in centiMorgans (cM), is defined to be the expected number of crossovers occurring between two loci. A variety of mapping functions have been proposed to estimate the number of crossovers between two loci given the observed recombination fraction. Note that recombination between two loci occurs if there are an odd number of crossovers between them. Haldane [13] proposed a mapping function based on the Poisson distribution. He assumes that the number of crossovers occurs independently and randomly across the entire chromosome. Let W be a Poisson random variable, with mean λ, for the number of crossovers between two loci which are λ Morgans apart. Then the fraction

of recombinant gametes, θ, is estimated to be

$$\begin{aligned}
\theta &= \sum_{w=0}^{\infty} \Pr(W = 2w+1|\lambda) \\
&= \sum_{w=0}^{\infty} \frac{e^{-\lambda}\lambda^{2w+1}}{(2w+1)!} \\
&= e^{-\lambda} \sum_{w=0}^{\infty} \frac{\lambda^{2w+1}}{(2w+1)!} \\
&= e^{-\lambda}\Big(\frac{e^{\lambda} - e^{-\lambda}}{2}\Big) \\
&= \frac{1 - e^{-2\lambda}}{2}.
\end{aligned}$$

Therefore, the estimated map distance is

$$\lambda = \frac{-\log(1-2\theta)}{2}. \tag{2.1}$$

We can see from equation 2.1 that as linkage between two loci becomes tighter (ie. $\theta \to 0$), then the genetic distance between them decreases (ie. $\lambda \to 0$). As well, as linkage becomes looser (ie. $\theta \to \frac{1}{2}$), then the genetic distance between the loci increases (ie. $\lambda \to \infty$). Using this map function, the genetic map distance may be estimated. We note that this is just one of many proposed mapping functions.

2.2.4 LOD Scores

For likelihood-based methods, it is customary to use LOD, likelihood of odds, scores as evidence for linkage when mapping major genes (see Lynch and Walsh [17], for example). When using linkage methods to detect major genes, the usual null hypothesis is that there is no evidence of linkage between the marker locus and a major gene (ie. $\theta = 1/2$), and therefore, the alternative hypothesis states that there is evidence of linkage (ie. $\theta < 1/2$). Let $L(\theta|\mathbf{x})$ be the likelihood function, where θ is the parameter and $\mathbf{x}$ is the sampled data. Then the likelihood ratio statistic, $\Lambda(\mathbf{x})$, for testing for evidence of linkage is

$$\Lambda(\mathbf{x}) = \left[\frac{\max\limits_{\theta\in\Theta_0} \mathrm{L}(\theta|\mathbf{x})}{\max\limits_{\theta\in\Theta_A} \mathrm{L}(\theta|\mathbf{x})}\right], \tag{2.2}$$

where Θ_0 and Θ_A represent the parameter space under the null and alternative hypothesis, respectively. Note that $-2\log\Lambda(\mathbf{x})$ has a χ^2 distribution under the null model with the degrees of freedom equal to the difference in the number of parameters under the two hypotheses.

The LOD score is the base 10 logarithm of the likelihood ratio statistic:

$$\begin{aligned} \text{LOD}(\mathbf{x}) &= \log_{10}\Lambda(\mathbf{x}) \\ &= \log_{10}\left[\frac{\max\limits_{\theta\in\Theta_0} \text{L}(\theta|\mathbf{x})}{\max\limits_{\theta\in\Theta_A} \text{L}(\theta|\mathbf{x})}\right]. \end{aligned}$$

Notice that the LOD score is equivalent to scaling $-2\log\Lambda$ by a factor of $-1/(2\log 10)$. Throughout the remainder of this thesis, we define the LOD score to be the base 10 logarithm of the likelihood ratio statistic.

2.3 Linkage Analysis Summary

Linkage analysis is motivated by the phenomenon of recombination, since we know that genes which are inherited together tend to not undergo recombination. If we were to examine two loci which are close together, we would expect the number of recombinations between them to be close to 0. On the other hand, if we were to examine two loci which are far apart on a chromosome, then we would expect to observe recombination about half of the time. So, in its simplest form, linkage analysis tests for linkage between two loci by determining if the estimated recombination fraction differs significantly from 1/2.

In the context of quantitative traits, linkage analysis involves detecting major genes which contribute to the variability of the trait. These major genes are called *quantitative trait loci* (QTLs). When analyzing quantitative traits, we try to map any potential QTLs relative to marker loci across the chromosome. Linkage analysis is partly based on examining allele sharing patterns between relatives; therefore, the data consists of a sample of families or pedigrees. Within each pedigree the relationships between the family members are known. When mapping a QTL on a chromosome, we require data from a series of markers along the chromosome. For each family member the pair of alleles present at each marker locus is observed. Therefore, the proportion of alleles shared ibd at each marker locus may be computed between all relative pairs. Also, when dealing with quantitative traits, the quantitative trait of

interest and any covariates are measured for each family member. Intuitively, we would expect relatives who share more alleles ibd at the QTL to have more similar quantitative trait values.

In practice, we evaluate the evidence of linkage between a potential QTL and each marker separately. Markers which are tightly linked to a QTL should be able to explain the variability in the quantitative trait, whereas markers loosely linked to the QTL should show no association with the trait. Lander and Botstein [16] proposed a technique to summarize the results from these separate analyses. They introduced the idea of graphically displaying whether evidence for a QTL exists by plotting LOD scores versus marker position on the chromosome. A peak in the LOD scores suggests that a QTL may exist near the marker locus where the maximum is obtained. Furthermore, the LOD score curves visually help to not only detect QTLs, but also estimate their position along the chromosome.

One of the goals of linkage analysis is to map the location of any QTLs relative to the markers along a chromosome. A secondary goal of linkage analysis is to estimate the amount of variability in the quantitative trait due to genetic and non-genetic factors. Further details on linkage analyses involving quantitative traits will be presented in subsequent chapters.

Chapter 3

Haseman-Elston Approach to Linkage Analysis

In this chapter, we give an overview of Haseman and Elston's classic regression approach to linkage analysis, whose results also aided in the development of the variance-components approach. Through their work, Haseman and Elston [15] created the basic framework for utilizing marker information to map QTLs. The basic idea underlying their approach is that sib-pairs with a greater number of alleles identical-by-descent at a locus should have more similar quantitative trait values. This idea led Haseman and Elston to regress the squared sib-pair differences in the phenotype on the proportion of alleles identical-by-descent at a marker. So, a large negative slope estimate indicates that the marker is linked to a QTL.

We begin this chapter by describing the details of Haseman and Elston's method. In particular, we will mention results which arise in the variance-components context. Although Haseman and Elston conceived this idea in the early 1970's, their method continues to receive much attention. Therefore, we provide a summary of some of the modifications and extensions that have been made to this method. Finally, we conclude this chapter with a description of some of the advantages and disadvantages of the Haseman-Elston approach.

3.1 Haseman-Elston Model

The Haseman-Elston approach is appealing since it involves a simple linear regression model which may be easily fit using least squares methods. Before describing the regression model, we state the assumptions on the genetic mechanism underlying the quantitative trait.

Let Y_{1j} and Y_{2j} be the quantitative trait value for the 1st and 2nd sib of the jth sib pair. Then the proposed model for the quantitative trait is as follows:

$$Y_{ij} = \mu + g_{ij} + \epsilon_{ij}, \tag{3.1}$$

where μ is the overall mean, g_{ij} is the genetic effect, and ϵ_{ij} is the random deviation from the mean. We assume that g_{ij} and ϵ_{ij} are uncorrelated. In model 3.1, the random deviation term has mean 0 and variance σ_ϵ^2. The genetic effect has variance σ_g^2, which may be further decomposed into an additive component, σ_q^2, and a dominant component, σ_d^2. We assume that the dominant component is negligible for simplicity, so $\sigma_g^2 = \sigma_q^2$. The genetic effect is assumed to be from a single QTL with two alleles, B and b, having frequencies of p and r. Furthermore, the genetic effect may be quantified as follows:

$$g_{ij} = \begin{cases} a + d, & \text{for genotype } BB \\ d, & \text{for genotype } Bb \\ -a + d, & \text{for genotype } bb \end{cases} \tag{3.2}$$

Notice that a is one-half of the distance between the two homozygous genotypic means. With only an additive genetic component, the genetic effect of the heterozygous genotypic mean is halfway between the genetic effect of the two homozygous genotypic means. See Falconer [9] for further details on the dominant component. The simple, but elegant idea that Haseman and Elston proposed was to regress the squared sib-pair trait differences on the proportion of genes shared ibd. The proportion of genes shared ibd at the QTL would be ideal to regress on, however, in reality this information is not available because the locus position is initially unknown. Because we do not know the proportion of genes shared ibd between sib-pairs at the QTL, data at different markers are used, including markers that are thought to be near the QTL. Let π_j be the estimated proportion of genes shared ibd between the jth sib pair at a particular marker locus. A brief discussion on how to estimate the proportion of genes ibd at

a marker is given in section 2.2.1 in chapter 2. Then the regression calculation yields:

$$(Y_{1j} - Y_{2j})^2 = \hat{\alpha} + \hat{\beta}\pi_j, \tag{3.3}$$

where $\hat{\alpha}$ is the intercept and $\hat{\beta}$ is the regression coefficient. Note that $(Y_{1j} - Y_{2j})^2$ is expected to be smaller as the proportion of alleles ibd at a marker, which is tightly linked to a QTL, increases.

Conditional on the marker data, Haseman and Elston [15] showed that the expectation of $\hat{\beta}$ is $-2(1-2\theta)^2\sigma_q^2$; therefore, the regression coefficient depends on the genetic variance component and the recombination fraction. Although not of direct importance to the Haseman-Elston approach, the conditional expectation of $\hat{\alpha}$ is $2[1-2(1-\theta)\theta]\sigma_q^2 + \sigma_\epsilon^2$ (Haseman and Elston [15]), assuming no dominance. Note that the conditional expectations of the regression parameters are derived in the following section. Since $0 \leq \theta \leq 1/2$ with $\theta = 0$ implying tight linkage, a negative regression coefficient would be a strong indication of the presence of a QTL linked to the marker (ie. the marker is close to the QTL). As well, when $\theta = 0$, $-\hat{\beta}/2$ provides an estimate of the additive genetic component of variance. Estimating the regression coefficients from data at a series of markers allows one to detect the presence of a QTL. The conditional expected values for the regression parameters have been computed for other relative pairs (Amos and Elston [2]). As we will see in Chapter 4, these conditional expectations of the regression parameters play a key role in the development of the covariance structure in Amos' [3] variance components approach.

We note that the genetic effect shown in expression 3.2 may be extended to the case of a major locus with k alleles. Let $a_1, \ldots, a_k$ be the expected values for the alleles $B_1, \ldots, B_k$, respectively. Then under the assumption of no dominance, the expected mean for genotype B_iB_j is $a_i + a_j$. Expression 3.2 is a special case where the allelic effect of B and b are $(a+d)/2$ and $(-a+d)/2$, respectively. Furthermore, if there is no dominance, then the expected values of the regression parameters for a biallelic marker locus are still valid for a multiallele marker locus (Haseman [14]).

3.2 Expected Values of the Regression Parameters

The expected values of the regression coefficients in Haseman and Elston's [15] least squares approach to linkage analysis play an important role in the development of the covariance structure of Amos' [3] variance-components approach. In this section, we present some of the details for the derivation of the expected values of the regression coefficients.

Firstly, we introduce some additional notation. Let π_{qj} and π_{mj} be the proportion of alleles shared ibd between the jth sib-pair at the QTL and marker, respectively. Recall that we define π_j to be the estimated proportion of alleles shared ibd at the marker between the jth sib-pair. Also let the squared sib-pair difference for the jth pair be D_j, so $D_j = (Y_{1j} - Y_{2j})^2$. Secondly, we express the conditional expectation of the squared sib-pair differences on the estimated proportion of alleles shared ibd at the marker in terms of the true proportion of alleles shared ibd at the marker locus and QTL:

$$\begin{aligned} \mathrm{E}(D_j|\pi_j) &= \sum_{i=0}^{2} \mathrm{E}(D_j|\pi_{qj} = \tfrac{i}{2}) \Pr(\pi_{qj} = \tfrac{i}{2}|\pi_j) \\ &= \sum_{i=0}^{2}\sum_{k=0}^{2} \mathrm{E}(D_j|\pi_{qj} = \tfrac{i}{2}) \Pr(\pi_{qj} = \tfrac{i}{2}|\pi_{mj} = \tfrac{k}{2}) \Pr(\pi_{mj} = \tfrac{k}{2}|\pi_j). \end{aligned}$$

Thirdly, conditional on the proportion of alleles shared ibd at the QTL and assuming no dominance, the expectations of D_j are as follows:

$$\mathrm{E}(D_j|\pi_{qj} = 0) = \sigma_\epsilon^2 + 2\sigma_q^2, \tag{3.4}$$

$$\mathrm{E}(D_j|\pi_{qj} = \tfrac{1}{2}) = \sigma_\epsilon^2 + \sigma_q^2, \tag{3.5}$$

$$\mathrm{E}(D_j|\pi_{qj} = 1) = \sigma_\epsilon^2. \tag{3.6}$$

In the presence of dominance, the expressions for the conditional expectations may be found in Haseman and Elston's paper [15].

To compute the conditional expectation of the squared sib-pair differences, we need the joint distribution of π_{mj} and π_{qj}, as well as the joint distribution of π_j and π_{mj}. These joint distributions from Haseman and Elston [15] are reproduced in Tables 3.1 and 3.2. Further details are given in the Appendix. We let $\Psi = \theta^2 + (1-\theta)^2$.

Using equations 3.4, 3.5, 3.6, and Tables 3.1 and 3.2, we can compute the conditional expectation of the squared sib-pair differences given the estimated proportion of alleles shared

Table 3.1: Joint Distribution of π_{mj} and π_{qj}

π_{qj}	π_{mj} 0	$\frac{1}{2}$	1	Total
0	$\Psi^2/4$	$\Psi(1-\Psi)/2$	$(1-\Psi)^2/4$	$\frac{1}{4}$
$\frac{1}{2}$	$\Psi(1-\Psi)/2$	$(1-2\Psi+2\Psi^2)/2$	$\Psi(1-\Psi)/2$	$\frac{1}{2}$
1	$(1-\Psi)^2/4$	$\Psi(1-\Psi)/2$	$\Psi^2/4$	$\frac{1}{4}$
Total	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Table 3.2: Joint Distribution of π_{mj} and π_j

π_j	π_{mj} 0	$\frac{1}{2}$	1	Total
0	$\frac{1}{2}p^2r^2$	0	0	$\frac{1}{2}p^2r^2$
$\frac{1}{4}$	p^3r+pr^3	p^3r+pr^3	0	$2(p^3r+pr^3)$
$\frac{1}{2}$	$\frac{1}{4}(p^4+4p^2r^2+r^4)$	$\frac{1}{2}(p^4+6p^2r^2+r^4)$	$\frac{1}{4}(p^4+4p^2r^2+r^4)$	$(p^4+5p^2r^2+r^4)$
$\frac{3}{4}$	0	p^3r+pr^3	p^3r+pr^3	$2(p^3r+pr^3)$
1	0	0	$\frac{1}{2}p^2r^2$	$\frac{1}{2}p^2r^2$
Total	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

ibd at the marker locus. For example, if the estimated proportion of alleles ibd at the marker locus is 0, then

$$\begin{aligned}
\mathrm{E}(D_j|\pi_j=0) &= [\sigma_\epsilon^2+2\sigma_q^2][\Psi^2(1)+\Psi(1-\Psi)(0)+(1-\Psi^2)(0)] \\
&+ [\sigma_\epsilon^2+\sigma_q^2][2\Psi(1-\Psi)(1)+(1-2\Psi+2\Psi^2)(0)+2\Psi(1-\Psi)(0)] \\
&+ \sigma_\epsilon^2[(1-\Psi)^2(1)+\Psi(1-\Psi)(0)+\Psi^2(0)] \\
&= \sigma_\epsilon^2+2\Psi\sigma_q^2.
\end{aligned}$$

Similarly, it can be shown that

$$\begin{aligned}
\mathrm{E}(D_j|\pi_j=\tfrac{1}{4}) &= \sigma_\epsilon^2+(\tfrac{1}{2}+\Psi)\sigma_q^2 \\
\mathrm{E}(D_j|\pi_j=\tfrac{1}{2}) &= \sigma_\epsilon^2+\sigma_q^2 \\
\mathrm{E}(D_j|\pi_j=\tfrac{3}{4}) &= \sigma_\epsilon^2+(\tfrac{3}{2}-\Psi)\sigma_q^2 \\
\mathrm{E}(D_j|\pi_j=1) &= \sigma_\epsilon^2+2(1-\Psi)\sigma_q^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{E}(D_j|\pi_j) &= \sigma_\epsilon^2 + 2\Psi\sigma_q^2 + 2(1-2\Psi)\sigma_q^2\pi_j \\
&= \sigma_\epsilon^2 + 2[\theta^2 + (1-\theta)^2]\sigma_q^2 + 2[1 - 2(\theta^2 + (1-\theta)^2)]\sigma_q^2\pi_j \\
&= \sigma_\epsilon^2 + 2(1 - 2\theta + 2\theta^2)\sigma_q^2 - 2(4\theta^2 - 4\theta + 1)\sigma_q^2\pi_j \\
&= \sigma_\epsilon^2 + 2(1 - 2\theta + 2\theta^2)\sigma_q^2 - 2(1-2\theta)^2\pi_j \\
&= \sigma_\epsilon^2 + 2[1 - 2(1-\theta)\theta]\sigma_q^2 - 2(1-2\theta)^2\sigma_q^2\pi_j.
\end{aligned}$$

Finally, we can see that conditional on the marker data, the expected values of the regression parameters in Haseman and Elston's approach are

$$\mathrm{E}(\hat{\beta}|\pi_j) = -2(1-2\theta)^2\sigma_q^2 \tag{3.7}$$

$$\mathrm{E}(\hat{\alpha}|\pi_j) = \sigma_\epsilon^2 + 2[1 - 2(1-\theta)\theta]\sigma_q^2. \tag{3.8}$$

3.3 Extensions of the Haseman-Elston Approach

The Haseman-Elston approach is the foundation of linkage analysis for quantitative traits. This approach has been improved and enhanced over the years, and therefore is still an extremely popular tool for detecting QTLs. The power of this approach has been increased through insightful modifications to the dependent variable in the regression procedure. As well, attempts to incorporate relative pairs, in addition to sib-pairs, have been proposed.

In the late 1990s, questions regarding the use of the squared sib-pair differences as the dependent variable in the regression procedure arose. Wright [29] argued that only looking at squared sib-pair differences discards linkage information. He used a likelihood argument to allude to the consequences of only using sib-pair differences to test for linkage, and proposed also using the sib-pair sums. Noting that the sib-pair differences and sib-pair sums are independent, he argued that not all of the information from the sib-pair data is being used to its full potential. Drigalenko [6] expanded upon Wright's argument and proposed an extension of the Haseman-Elston approach which incorporated the sib-pair sum into the regression procedure. Using both the sib-pair differences and sib-pair sums in the regression procedure, Drigalenko showed that the variance for the estimated regression coefficient is smaller than when using only sib-pair differences or sib-pair sums. Furthermore, he showed that under certain assumptions, the

use of both sib-pair differences and sib-pair sums is equivalent to using the sib-pair products. Because the covariance gives the same information as the sib-pair differences and sums, he concluded that the success of the variance-components method must be due to the incorporation of the covariance structure. Finally, Elston et al [8] showed that the power to detect linkage will increase even further by regressing the mean-corrected product of the trait values on the proportion of alleles shared ibd at a marker locus.

The power of the Haseman-Elston approach has been improved by incorporating all relative pairs, in addition to sib-pairs, into a single regression model. Amos and Elston [2] developed an algorithm to compute the proportion of genes shared ibd between any relative pair. They also found the expectations of the regression parameters for relative pairs other than sibs. Olson and Wijsman [21] extended the Haseman-Elston approach by combining information from all types of relative pairs.

3.4 Haseman-Elston Discussion

In the literature, there are numerous discussions which compare and contrast Haseman-Elston's approach to the variance-components approach to linkage analysis. While both methods have their advantages and disadvantages, in this section we focus on some of the notable aspects of the Haseman-Elston approach.

One of the most appealing aspects of Haseman and Elston's approach is that it only involves least squares estimation for the regression parameters, which results in rapid computations. Therefore, if we have a vast amount of marker information to process and our main concern is to detect QTLs, this method is a prime candidate for an initial analysis. Another benefit of the Haseman-Elston approach is that it is robust from both a genetics viewpoint and a statistics viewpoint. In a genetics setting, this model is robust because we do not have to know genetic details such as the mode of inheritance or the frequency of the alleles at the QTL. In a statistics setting, this model is robust because the test statistic, for determining whether the regression coefficient is significantly less than zero, is not sensitive to departures from normality in the distribution of the quantitative trait.

Although the Haseman-Elston approach is useful for initially detecting QTLs, we can-

not obtain estimates of the genetic variance since it is confounded with the recombination fraction. If estimating the genetic variance in a quantitative trait is of primary interest, then this method may not be ideal. It is also not an easy task to apply this approach to extended pedigrees. Because extended pedigrees are known to provide more linkage information than nuclear pedigrees, Haseman-Elston's approach may not be efficient to use on some familial datasets. These concerns motivate the need for other approaches to linkage analysis, such as the variance-components method.

Chapter 4

Variance-Components Approach to Linkage Analysis

Variance-components based methods have flourished in the area of linkage analysis by giving rise to more robust techniques and alleviating the need for segregation analysis. In the linkage analysis context, we say that a method is "robust" if it does not require one to determine the mode of inheritance or how the alleles at the QTL segregate. Moreover, variance-components approaches require the estimation of fewer parameters than a penetrance-based approach. A variance-components approach attempts to decompose the variability of the phenotype into the variability due to the QTL, polygenes, and environment. In this chapter, we describe a series of models in increasing complexity and build up to the mixed effects models currently being exploited for linkage analysis.

4.1 Sporadic Model

Firstly, we describe the sporadic model (Blangero et al. [5]). The sporadic model attempts to explain the variation in the quantitative trait via only covariates, such as age. At this stage, the quantitative trait is not thought to be influenced by genetic factors. One of the goals at this step of the modelling procedure is to determine any covariates which may have an effect on the quantitative trait. Once any significant covariates have been found, the fitted sporadic model will serve as a basis of comparison for the detection of any genetic factors which may

further explain any remaining trends in the data.

Let Y_i be the quantitative trait value of the ith relative and x_{ik} be the value of the kth covariate for the ith relative. Then, the sporadic model is

$$Y_i = \mu + \sum_{k=1}^{K} \beta_k x_{ik} + \epsilon_i, \tag{4.1}$$

where μ is the overall mean, β_k is the effect of the kth covariate, and ϵ_i is the random deviation from the mean. We assume that $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. It is obvious that this model assumes that the total phenotypic variation, σ_T^2, is due to only a random deviation component, σ_ϵ^2.

For comparison with the subsequent models in this chapter, it is helpful to look more closely at the covariance structure stipulated by each model. The covariance structure of the sporadic model is

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_\epsilon^2, & \text{if } i = j \\ 0, & \text{if } i \neq j. \end{cases} \tag{4.2}$$

We note that the sporadic model does not account for any familial correlation. In fact, the sporadic model is a simple linear regression model.

4.2 Polygenic Model

To expand upon the sporadic model, the next step is to account for information on familial relationships, which results in the so-called polygenic model (Blangero et al. [5]). Because we are dealing with familial data, the quantitative trait values are correlated. Unlike the sporadic model, the polygenic model does not ignore the covariance structure within families. While for instance an exchangeable covariance structure could be used, the covariance structure is actually based on the type of relationship between any two individuals, and so is more complex. Furthermore, the covariance structure in the polygenic model accounts for the fact that siblings' trait values are likely to be more correlated than trait values between cousins, for example. The polygenic model decomposes the total phenotypic variance, σ_T^2, into two components: the polygenic component of variance, σ_G^2, and the random deviation component of variance, σ_ϵ^2. Recall that polygenes are a group of genes that collectively have a small effect on a quantitative trait. Using the expected proportion of genes ibd between relative pairs, this model essentially

attempts to divide the total phenotypic variance into a genetic component and non-genetic component.

Let Y_i be the quantitative trait value of the ith relative and x_{ik} be the value of the kth covariate for the ith relative. Then, the polygenic model is

$$Y_i = \mu + \sum_{k=1}^{K} \beta_k x_{ik} + G_i + \epsilon_i, \tag{4.3}$$

where μ is the overall mean, β_k is the effect of the kth covariate, G_i is the random polygenic effect, and ϵ_i is the random deviation. We assume that G_i has variance σ_G^2, and ϵ_i has variance σ_ϵ^2. As well, without loss of generality, we assume that $\mathrm{E}(G_i) = \mathrm{E}(\epsilon_i) = 0$, since the overall mean component may absorb any residual mean structure from these components. The polygenic and random deviation components, G_i and ϵ_i, are assumed to be independent for each individual. Furthermore, the vector of G_i in a family forms a random vector with dependent components, and the vector of ϵ_i in a family forms an random vector with independent components. As well, these two random vectors are independent of each other. The vectorized form of the model in 4.3 is

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{G} + \boldsymbol{\epsilon}, \tag{4.4}$$

where $\mathbf{Y}$ is a vector of quantitative traits, $\boldsymbol{\mu}$ is the mean vector, $\mathbf{X}$ is the matrix of covariates, $\boldsymbol{\beta}$ is a vector of covariate effects, $\mathbf{G}$ is the random vector of polygenic effects and $\boldsymbol{\epsilon}$ is the random vector of deviation effects. The covariance structure of the polygenic model is

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_G^2 + \sigma_\epsilon^2, & \text{if } i = j \\ 2\phi_{ij}\sigma_G^2, & \text{if } i \neq j. \end{cases} \tag{4.5}$$

Here, ϕ_{ij} is known as the coefficient of kinship for relatives i and j and $2\phi_{ij}$ is called the coefficient of relationship for relatives i and j (see section 2.2.2 in chapter 2). For example, $\phi_{ij} = 1$ if $i = j$, and $\phi_{ij} = 1/2$ if relative i and j are siblings. Notice that the total phenotypic variance is in terms of a genetic component and a non-genetic component. As well, the strength of correlation between relatives depends on the degree of their relationship.

Finally, we note that the sporadic model is nested within the polygenic model; therefore, by comparing the likelihoods of each model we can obtain an idea of how large of a role genetics plays in specifying the quantitative trait.

4.3 Two-Point Model

The sporadic model and polygenic model help to determine the degree to which genetics influence the quantitative trait; however, it is the two-point model (Blangero et al. [5]), which attempts to locate where the source of the variation underlying the phenotype lies via linkage analysis. In addition to familial relationships, the two-point model utilizes marker information in an attempt to map QTLs along a chromosome. The concept of recombination plays an important role in the development of this model. Recall that alleles at loci which are near one another tend to be passed from parent to offspring together. Therefore, one would think that markers which are closer to the QTL will be able to explain the variability in the quantitative trait better than those markers which are far from the QTL. The two-point model extracts information from the marker data by using the proportion of alleles shared ibd in its covariance structure. Furthermore, the two-point model decomposes the genetic variance into a component due to the QTL and a component due to the polygenes.

Let Y_i be the quantitative trait value of the ith relative and x_{ik} be the value of the kth covariate for the ith relative. Then, the two-point model is

$$Y_i = \mu + \sum_{k=1}^{K} \beta_k x_{ik} + q_i + G_i + \epsilon_i, \tag{4.6}$$

where μ is the overall mean, β_k is the effect of the kth covariate, q_i is the effect due to the QTL, G_i is the polygenic effect, and ϵ_i is the random deviation. We assume that q_i has variance σ_q^2, G_i has variance σ_G^2, and ϵ_i has variance σ_ϵ^2. Without loss of generality, we assume that $\mathrm{E}(q_i) = \mathrm{E}(G_i) = \mathrm{E}(\epsilon_i) = 0$. Finally, the QTL, polygenic, and random deviation components are assumed to be independent. Note that model 4.6 assumes that there is only one QTL underlying the trait of interest. The major gene effect is assumed to be from a single QTL with two alleles, B and b. Furthermore, the QTL effect may be quantified as follows:

$$q_i = \begin{cases} a + d, & \text{for genotype } BB \\ d, & \text{for genotype } Bb \\ -a + d, & \text{for genotype } bb \end{cases} \tag{4.7}$$

If we assume that the QTL effects are additive, we can easily extend this model to account for multiple QTLs.

Under the assumption of no dominance, the covariance structure of the two-point model is

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_q^2 + \sigma_G^2 + \sigma_\epsilon^2, & \text{if } i = j \\ \pi_{ij}\sigma_q^2 + 2\phi_{ij}\sigma_G^2, & \text{if } i \neq j. \end{cases} \tag{4.8}$$

Here, π_{ij} is the estimated proportion of genes shared ibd between relative i and relative j at the marker, and $2\phi_{ij}$ is the coefficient of relationship for relative i and relative j.

This mixed effects model partitions the total phenotypic variance into three variance components: the variance due to QTL, the variance due to polygenes, and the variance due to random deviation. We note that this two-point model implicitly assumes that there is no dominance, as well, the recombination fraction is assumed to be 0. The latter assumption will become more apparent after we define the Amos' model which relaxes the constraint on the recombination fraction. The former assumption can be seen from Amos [3] and his model which includes a dominance component. Because the two-point model assumes that the recombination fraction is 0 between the marker and the QTL, it is assuming that the marker is close to and tightly linked with the QTL. A comparison of the log-likelihoods from the two-point model and polygenic model at each marker will help us to map any underlying QTLs. Recall that a LOD score is the difference between the base 10 logarithm of the likelihoods. In practice, one would plot LOD scores versus the marker location. Ideally, such a plot would result in a peak in the LOD score curve over the region of the chromosome where the QTL is located.

4.4 Amos' Model

The robust variance-components model proposed by Amos [3] is a generalization of all of the models mentioned thus far. Based on familial relationships and the proportion of genes ibd at various markers, this model attempts to estimate not only the QTL variance component, polygenic variance component, and random deviation variance component, but also the recombination fraction. By allowing the recombination fraction to vary, Amos' model requires the estimation of an additional parameter but no longer assumes that the marker is tightly linked to the QTL, which is more realistic.

The genetic model for the QTL is the same as the two-point model:

$$q_i = \begin{cases} a+d, & \text{for genotype } BB \\ d, & \text{for genotype } Bb \\ -a+d, & \text{for genotype } bb \end{cases} \tag{4.9}$$

Here, the single QTL has two alleles, B and b.

The model for the quantitative trait is also similar to the two-point model with the exception of the covariance structure. Let Y_i be the quantitative trait value of the ith relative and x_{ik} be the value of the kth covariate for the ith relative. Then, Amos' model is

$$Y_i = \mu + \sum_{k=1}^{K} \beta_k x_{ik} + q_i + G_i + \epsilon_i, \tag{4.10}$$

where μ is the overall mean, β_k is the effect of the kth covariate, q_i is the effect due to the QTL, G_i is a random polygenic effect, and ϵ_i is the random deviation. We assume that q_i has variance σ_q^2, G_i has variance σ_G^2, and ϵ_i has variance σ_ϵ^2. Without loss of generality, we assume that $\mathrm{E}(q_i) = \mathrm{E}(G_i) = \mathrm{E}(\epsilon_i) = 0$. Finally the QTL, polygenic, and random deviation components are assumed to be independent for each individual.

Under the assumption of no dominance, the covariance structure of Amos' model is

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_q^2 + \sigma_G^2 + \sigma_\epsilon^2, & \text{if } i = j \\ f(\theta, \pi_{ij})\sigma_q^2 + 2\phi_{ij}\sigma_G^2, & \text{if } i \neq j, \end{cases} \tag{4.11}$$

where π_{ij} is the estimated proportion of genes ibd between relative i and relative j at the marker, and $2\phi_{ij}$ is the coefficient of relationship for relative i and relative j. Values of $f(\theta, \pi_{ij})$ are given in Amos [3], and reproduced in Table 4.1. Note when $\theta = 0$ which means that the marker is tightly linked to the QTL, the covariance structure of Amos' model reduces to the covariance structure of the two-point model. Also, when $\theta = 1/2$, meaning that the marker is unlinked to the QTL, the covariance structure of Amos' model reduces to the covariance structure of the polygenic model. In equation 4.9, we are assuming that dominance effects are negligible, and that the QTL effect is additive. A more general form of this mixed effects model, which includes dominance effects, is given by Amos [3].

Table 4.1: Fraction of Variance from Additive QTL Component

Relative Pair	Fraction of QTL Variance Component, $f(\theta, \pi_{ij})$
Sibs	$\frac{1}{2} + (1-2\theta)^2(\pi_{ij} - \frac{1}{2})$
Half-sibs	$\frac{1}{4} + (1-2\theta)^2(\pi_{ij} - \frac{1}{4})$
Avuncular	$\frac{1}{4} + (1-2\theta)^2(1-\theta)(\pi_{ij} - \frac{1}{4})$
Grandparental	$\frac{1}{4} + (1-2\theta)(\pi_{ij} - \frac{1}{4})$
First cousin	$\frac{1}{8} + (1-2\theta)^2(1 - \frac{4}{3}\theta + \frac{2}{3}\theta^2)(\pi_{ij} - \frac{1}{8})$

4.4.1 Covariance Structure

The utilization of marker information to map QTLs is one of the fundamental keys to linkage analysis. Haseman and Elston's [15] method of linkage analysis not only played a large role in developing the theory for a least squares approach, but also contributed to deriving the covariance structure used in the variance-components approach. In this section, we illustrate the contribution made by Haseman and Elston to Amos' covariance structure by deriving the component of variance due to the QTL effect, $f(\theta, \pi_{ij})\sigma_q^2$, for a sib-sib pair (as shown in Table 4.1).

Let Y_i and Y_j be the quantitative trait values for siblings i and j, whose estimated proportion of genes shared ibd at a marker locus is π_{ij}. We can compute the covariance between Y_i and Y_j by noting that

$$\begin{aligned} \mathrm{E}[(Y_i - Y_j)^2|\pi_{ij}] &= \mathrm{E}(Y_i^2) + \mathrm{E}(Y_j^2) - 2\mathrm{E}(Y_iY_j|\pi_{ij}) \\ &= \mathrm{E}(Y_i^2) - [\mathrm{E}(Y_i)]^2 + \mathrm{E}(Y_j^2) - [\mathrm{E}(Y_j)]^2 - 2\mathrm{E}(Y_iY_j|\pi_{ij}) + 2\mathrm{E}(Y_i)\mathrm{E}(Y_j) \\ &= 2\mathrm{Var}(Y_i) - 2\mathrm{Cov}(Y_i, Y_j|\pi_{ij}). \end{aligned}$$

Here, we assume that $\mathrm{E}(Y_i^2) = \mathrm{E}(Y_j^2)$. Therefore, the covariance is

$$\mathrm{Cov}(Y_i, Y_j|\pi_{ij}) = \mathrm{Var}(Y_i) - \tfrac{1}{2}\mathrm{E}[(Y_i - Y_j)^2|\pi_{ij}]. \tag{4.12}$$

Recall that Haseman and Elston [15] regressed the squared sib-pair differences on the proportion of genes shared ibd at a marker locus. Through their work, they found that

$$\mathrm{E}[(Y_i - Y_j)^2|\pi_{ij}] = 2[1 - 2(1-\theta)\theta]\sigma_q^2 + \sigma_\epsilon^2 - 2(1-2\theta)^2\sigma_q^2\pi_{ij}. \tag{4.13}$$

Using equation 4.12 and the result from equation 4.13, we can extract the coefficients of σ_q^2 to

find $f(\theta, \pi_{ij})$ for sib-pairs:

$$\begin{aligned} f(\theta, \pi_{ij}) &= 1 - [1 - 2(1-\theta)\theta - (1-2\theta)^2\pi_{ij}] \\ &= 2(1-\theta)\theta + (1-2\theta)^2\pi_{ij} \\ &= 2\theta - 2\theta^2 - \tfrac{1}{2} + \tfrac{1}{2} + (1-2\theta)^2\pi_{ij} \\ &= -\tfrac{1}{2}(1 - 4\theta + 4\theta^2) + \tfrac{1}{2} + (1-2\theta)^2\pi_{ij} \\ &= -\tfrac{1}{2}(1-2\theta)^2 + \tfrac{1}{2} + (1-2\theta)^2\pi_{ij} \\ &= \tfrac{1}{2} + (1-2\theta)^2(\pi_{ij} - \tfrac{1}{2}). \end{aligned}$$

We can see that as the distance between the marker and QTL becomes smaller (ie. as $\theta \to 0$), the fraction of variance due to the QTL tends towards the proportion of alleles shared ibd between the sibs. As well, as the distance between the marker and QTL increases (ie. as $\theta \to 1/2$), the fraction of variance due to the QTL tends towards the expected proportion of alleles shared ibd between the sibs. The remainder of Table 4.1 may be derived in a similar manner.

4.4.2 Likelihood Function

If we assume that the quantitative traits within a pedigree follow a multivariate normal distribution, then it is relatively easy to derive the likelihood function. We introduce some matrix notation so that the likelihood may be written in a compact form.

Let there be m pedigrees with n_l family members in the lth pedigree. Also, let $\mathbf{Y}_l$ be a $n_l \times 1$ vector of traits and $\mathbf{X}_l$ be a $n_l \times K$ matrix of K covariates for the lth pedigree. Then the vectorized form of Amos' model in 4.10 is

$$\mathbf{Y}_l = \boldsymbol{\mu}_l + \mathbf{X}_l\boldsymbol{\beta} + \mathbf{q}_l + \mathbf{G}_l + \boldsymbol{\epsilon}_l, \tag{4.14}$$

where $\boldsymbol{\mu}$ is a $n_l \times 1$ mean vector, $\boldsymbol{\beta}$ is a $K \times 1$ vector of covariate effects, $\mathbf{q}_l$ is a $n_l \times 1$ dependent random vector of QTL effects, $\mathbf{G}_l$ is a $n_l \times 1$ independent random vector of polygenic effects, and $\boldsymbol{\epsilon}_l$ is a $n_l \times 1$ independent random vector of the deviation from the mean effects. The vectors $\mathbf{q}_l$, $\mathbf{G}_l$, and $\boldsymbol{\epsilon}_l$ are assumed to be independent. Suppose $\boldsymbol{\Pi}_l$ is a $n_l \times n_l$ matrix of the estimated proportion of genes shared ibd at a marker for the lth pedigree, $2\boldsymbol{\Phi}_l$ is a $n_l \times n_l$ matrix of the

coefficients of relationship for the lth pedigree. Then, the covariance matrix, $\mathbf{\Omega}_l$, is

$$\mathbf{\Omega}_l = \mathbf{\Pi}_l\sigma_q^2 + 2\mathbf{\Phi}_l\sigma_G^2 + \mathbf{I}_l\sigma_\epsilon^2, \tag{4.15}$$

where $\mathbf{I}_l$ is an identity matrix of dimension n_l. Hence, $\mathbf{q}_l \sim \text{MVN}(\mathbf{0}, \mathbf{\Pi}_l\sigma_q^2)$, $\mathbf{G}_l \sim \text{MVN}(\mathbf{0}, 2\mathbf{\Phi}_l\sigma_G^2)$, and $\boldsymbol{\epsilon}_l \sim \text{MVN}(\mathbf{0}, \mathbf{I}_l\sigma_\epsilon^2)$.

The log-likelihood is

$$l(\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma_q^2, \sigma_G^2, \sigma_\epsilon^2 | \mathbf{y}, \mathbf{x}) = \sum_{l=1}^{m} \Big\{ -\tfrac{n_l}{2}\log(2\pi) - \tfrac{1}{2}\log|\mathbf{\Omega}_l| - \tfrac{1}{2}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta})^{\mathrm{T}}\mathbf{\Omega}_l^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta}) \Big\} \tag{4.16}$$

Note that by assuming the quantitative traits in a family follow a multivariate normal distribution, the log-likelihood functions for the sporadic, polygenic, and two-point models may be constructed in a similar manner.

We can maximize the log-likelihood in equation 4.16 to obtain the maximum likelihood estimates of the parameters. The covariate effects may be estimated simultaneously; although, in practice they are sometimes estimated prior to fitting a variance-components model. Note that from matrix theory (Searle [25]), if M is a square matrix and z is a scalar variable, then

$$\frac{\partial \log|M|}{\partial z} = tr\Big(M^{-1}\frac{\partial M}{\partial z}\Big) \tag{4.17}$$

and

$$\frac{\partial M^{-1}}{\partial z} = -M^{-1}\frac{\partial M}{\partial z}M^{-1} \tag{4.18}$$

where tr is the trace of a square matrix, which is the sum of the diagonal elements. Using these properties, the partial derivatives of the log-likelihood in equation 4.16 with respect to each parameter are as follows:

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = \sum_{l=1}^{m} \mathbf{1}^{\mathrm{T}}\mathbf{\Omega}_l^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta}) \tag{4.19}$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{l=1}^{m} \mathbf{x}_l^{\mathrm{T}}\mathbf{\Omega}_l^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta}) \tag{4.20}$$

$$\frac{\partial l}{\partial \sigma_q^2} = \sum_{l=1}^{m} \Big[-\tfrac{1}{2}\text{tr}(\mathbf{\Omega}_l^{-1}\mathbf{\Pi}_l) + \tfrac{1}{2}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta})^{\mathrm{T}}\mathbf{\Omega}_l^{-1}\mathbf{\Pi}_l\mathbf{\Omega}_l^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta}) \Big] \tag{4.21}$$

$$\frac{\partial l}{\partial \sigma_G^2} = \sum_{l=1}^{m} \Big[-\tfrac{1}{2}\text{tr}(\mathbf{\Omega}_l^{-1}2\mathbf{\Phi}_l) + \tfrac{1}{2}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta})^{\mathrm{T}}\mathbf{\Omega}_l^{-1}2\mathbf{\Phi}_l\mathbf{\Omega}_l^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta}) \Big] \tag{4.22}$$

$$\frac{\partial l}{\partial \sigma_\epsilon^2} = \sum_{l=1}^{m}\left[-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Omega}_l^{-1}) + \tfrac{1}{2}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Omega}_l^{-1}\boldsymbol{\Omega}_l^{-1}(\mathbf{y}_l - \boldsymbol{\mu}_l - \mathbf{x}_l\boldsymbol{\beta})\right] \tag{4.23}$$

Setting these partial derivatives equal to 0, we may solve these equations for the maximum likelihood estimates. Note that the estimates for the fixed effects depend on the estimates of the random effects. Since there is no closed form for the maximum likelihood estimators, numerical methods must be used to maximize the likelihood. In Chapters 5 and 6, we use a quasi-Newton approach to compute the maximum likelihood estimates.

4.5 Summary

In this chapter, we have described several nested mixed effect models used in variance components analysis. To make this nesting structure more apparent, we now re-express these models in terms of heritabilities and constraints. Recall that σ_T^2, σ_q^2, σ_G^2, and σ_ϵ^2 are the total phenotypic, QTL, polygenic, and random deviation components of variance. We define $h_q^2 = \sigma_q^2/\sigma_T^2$, $h_G^2 = \sigma_G^2/\sigma_T^2$, $h_\epsilon^2 = \sigma_\epsilon^2/\sigma_T^2$ to be the QTL heritability, polygenic heritability, and random deviation heritability, respectively. In matrix form, Amos' model is

$$\begin{cases} \mathbf{Y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{q} + \mathbf{G} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Omega}) \\ \boldsymbol{\Omega} = \sigma_T^2[f(\theta, \boldsymbol{\Pi})h_q^2 + 2\boldsymbol{\Phi}h_G^2 + \mathbf{I}h_\epsilon^2]\ . \end{cases} \tag{4.24}$$

Under certain constraints, this model reduces to the sporadic, polygenic, and two-point model. The nesting structure of these models is shown in Table 4.2. From this nesting structure the relationship between the models is apparent. For example, we can see by comparing the likelihoods of the polygenic and two-point models, we can test whether the QTL heritability differs from 0.

Table 4.2: Nesting Structure of Variance-Components Models

Model	Constraint
Sporadic	$h_\epsilon^2 = 1,\ h_q^2 = h_G^2 = 0;$
Polygenic	$h_\epsilon^2 + h_G^2 = 1,\ h_q^2 = 0;$
Two-Point	$h_\epsilon^2 + h_q^2 + h_G^2 = 1;$ $\quad \theta = 0$
Amos	$h_\epsilon^2 + h_q^2 + h_G^2 = 1$

Chapter 5

Simulations

In order to understand the performance and properties of the variance-components approach to linkage analysis, we conduct some simulations. We use GASP to simulate a quantitative trait which is influenced by one major locus and some polygenes. In our simulations, the heritabilities due to the major locus and polygenes are varied, while the random deviation effect is held constant. Through these simulations we would like to understand how well the variance-components method is able to map major loci, in addition to evaluate the accuracy of the estimated variance components.

In this chapter, we begin by describing the program used to simulate the data and the programs used to fit the models of interest. Next, we assess the pattern of LOD scores across the simulated chromosomes in their ability to map the major locus. Then, we compare and contrast the estimated variance components from the two-point model and Amos' model. Finally, following a discussion of the estimated recombination fractions from Amos' model, we give a brief summary of our findings.

5.1 Software

5.1.1 GASP

The Genometric Analysis Simulation Program (GASP) [11] is a freely available program which is supported on the Unix operating system and provided by the National Human Genome Research Institute at http://research.nhgri.nih.gov/gasp/. The main purpose of GASP is to

provide researchers with a means of creating familial data from a user-specified genetic model so that they may study the properties of statistical methods, such as power and error rates. GASP has the ability to generate pedigree data from nuclear families to extended families. As well, it can simulate quantitative traits which involve major loci effects, polygenic effects, random deviation effects, and covariate effects.

GASP (Wilson et al. [28]) provides the basis of a genetic model which is partly dictated by the user. In GASP, the quantitative trait is a linear combination of the QTL, polygenic, and random deviation components with weights equal to the proportion of the variance due to each component. For a QTL with two alleles, say B and b, the genotypic specific relative means for an individual with genotypes BB, Bb, bb are -1, 0, 1, respectively; therefore, GASP assumes that there is no dominance. The polygenic component and random deviation component are each assumed to have a standard normal distribution. GASP allows the user to explicitly state the proportion of the variance in the quantitative trait due to the QTL(s), polygenes, and random deviation. GASP also allows the user to specify the distance between successive markers, successive QTLs, or an adjacent marker and QTL via a recombination fraction. As well, the user must specify the allele frequencies for all QTLs and markers. For further details on GASP, please refer to the GASP website [11].

We use GASP to simulate a chromosome with twenty-five equally spaced markers and one major gene. This major gene is a quantitative trait locus with two equally frequent alleles. Each marker locus also has alleles occurring at equal frequencies. In our simulations, the percentage of the variation attributed to the QTL ranges from 0% to 80%. As well, the polygenic component varies in accordance with the QTL variance component from 10% to 90%. The random deviation variance component is held fixed at 10% throughout all of the simulations. Note that the only sources of variation in the quantitative trait are the QTL, the polygenic component, and the random deviation component; therefore, the percentage of variation due to all three components should total 100%. The twenty-five markers and single QTL were simulated to be equally spaced. Also, the QTL was simulated to lie between markers 14 and 15. In one case, we specified the recombination fraction to be 0.02 between each pair of adjacent loci. For instance, the recombination fraction between markers 1 and 2 was 0.02. As well, the recombination fraction between marker 14 and the QTL was 0.02, and the recombination fraction between

the QTL and marker 15 was 0.02. Since we are also interested in the effect of the density of the markers, we also conducted a second set of simulations where the recombination fraction was set to be 0.04 between each pair of adjacent loci. In this chapter, we denote the simulated recombination fraction between two successive loci to be θ_s. To be consistent with the notation in Chapter 4, we denote the parameter for the recombination fraction between a marker and a QTL to be θ. One hundred ceph families of size 10 were simulated in each replicate. Recall that a ceph family consists of three generations, and always includes two sets of grandparents and one set of parents. Finally, for each genetic model that we specified, we generated 100 replicates.

5.1.2 SOLAR

SOLAR (Blangero et al. [5]), which stands for Sequential Oligogenic Linkage Analysis Routines, is a well developed package for the variance components approach to linkage analysis. Amongst its many capabilities in version 1.7.3, this package is able to rapidly compute the proportion of genes shared ibd between relative pairs in a pedigree of arbitrary size and complexity. The sporadic, polygenic, and two-point models for linkage analysis involving a quantitative trait may be fit using SOLAR. As well, it is able to screen for covariates which may influence the quantitative trait of interest, and carry out a multipoint linkage analysis. Mainly written by Blangero, Lange, Almasy, Williams, Dyer, and Peterson, SOLAR is supported by the Southwest Foundation for Biomedical Research in San Antonio Texas. It may be downloaded for use on operating systems such as Unix and Linux from the public domain at http://www.sfbr.org/sfbr/public/software/solar.

We use SOLAR's capabilities to compute the proportion of genes ibd at each marker location along a chromosome via its "ibd" routine. The coefficients of relationship are also extracted from an intermediate file created by SOLAR. Finally, the "mibd" routine is used to classify the types of relationships within a pedigree. Further details on these routines may be found in the documentation located at the SOLAR website [5].

5.1.3 C Program

Although SOLAR is capable of fitting the sporadic model, polygenic model, and two-point model, it is not yet able to fit Amos' model. SOLAR's source code is not available in the public domain, so we wrote a C program to fit all models of interest. We parameterized the models in terms of heritabilities; therefore, the estimated variance components will be expressed as a fraction of the total phenotypic variance in this chapter and the following chapter.

When estimating Amos' model and the two-point model there are two stages. The first stage involves estimating the proportion of alleles shared ibd for all relative pairs at each marker locus using SOLAR's "ibd" routine. In the second stage, the estimated proportion of alleles are treated as known for the maximum likelihood estimation of the regression, heritability, and recombination parameters. Our program is able to simultaneously estimate the fixed effects and random effects. For each model, the negative log-likelihood is minimized using a quasi-Newton routine [19]. The running time of the program increases with the complexity of the pedigree.

5.2 Detection of Quantitative Trait Loci via LOD Scores

One of the major goals of linkage analysis is to map major loci relative to the markers on a chromosome. Once the region of the chromosome where any major loci may lie is estimated, association analysis techniques may be used to fine map these loci. By assessing the extent to which marker information helps to explain the variability in the quantitative trait, we are able to see how strongly linked these markers are to any QTLs in this vicinity. Currently, patterns of LOD scores across chromosomes are used to detect regions where major loci may reside. Comparing the likelihood of the polygenic model to the likelihood of the two-point model results in LOD scores which give an indication of whether a QTL is present. Therefore, when examining LOD scores across a chromosome, we would expect to see a rise in LOD scores as we approach regions with a QTL. In addition, a comparison of the likelihood of the two-point model to the likelihood of Amos' model will give us an idea of the role that the recombination fraction plays in explaining the quantitative trait. In this section, we will focus on the simulations where the recombination between successive loci is 0.02 (ie. $\theta_s = 0.02$).

5.2.1 Two-Point Model

To determine how well the variance-components approach to linkage analysis detects regions where QTLs reside, we look at the LOD scores between the polygenic model and two-point model. We begin by plotting the average of 100 LOD scores at each of 25 marker locations. Figure 5.1 shows how the pattern of LOD scores across a chromosome changes as the heritability due to the QTL increases from 0.0 to 0.8.

From the plots of the polygenic model and two-point model LOD scores versus the marker location, we see a distinct peak in LOD scores around markers 14 and 15. This peak becomes visible when the QTL component is around 40 to 60%, and is more apparent as the proportion of variance due to the QTL increases. When the QTL component ranges from 0% to 20% the LOD scores reach a maximum of about 1.1 and it is difficult to tell if a QTL is actually present on the chromosome. In contrast, when the QTL component is 80%, the LOD scores reach a maximum of around 12.9, and one can more readily map the QTL region.

The fact that the LOD scores rise with h_q^2 coincides with our intuition, since one would expect that if the majority of the variability in a quantitative trait is due to the QTL, then it should be easier to detect. A high LOD score implies that the two-point model explains the data better than the more simplistic polygenic model. Moreover, the data imply that the variability in the trait can be partly explained by the information given from the markers since they must be linked to a QTL. From the analyses of our simple quantitative trait, we can foresee the difficulties which may arise in detecting QTLs underlying a complex trait. Since many major loci contribute to the variability in a complex trait, we can appreciate that it is difficult to map all of these loci, especially if the majority of them account for less than 20% of the total variation each. In addition, the proportion of the genes ibd at each marker gives a stronger indication of the presence of a QTL as the distance between the QTL and marker decreases or equivalently as the recombination fraction tends to zero. From Figure 5.1, we can also see that, on average, the peakedness of the LOD score curves increases as the heritability due to the QTL increases. We believe that the non-additive and non-linear nature of the recombination fraction is being reflected through the steep ascents in the LOD score curves. Recall that although the recombination is constant between adjacent markers, it is not a measure of genetic distance. In fact, as we discussed in Chapter 2, the genetic distance between the QTL and farther markers

grows exponentially according to Haldane [13]. So we see that both the magnitude of the LOD scores and the peakedness of the LOD score curve give an indication of not only the presence of a QTL, but also its influence on the quantitative trait.

5.2.2 Amos' Model

The magnitude of a recombination fraction gives us an indication of how close a QTL is to a marker. Unlike the two-point model, Amos' model attempts to estimate the recombination fraction. We investigate whether Amos' model is able to map QTLs better than the two-point model. Recall that the two-point model assumes that the recombination fraction between the QTL and any marker is zero.

We construct LOD score curves from the average of 100 replicates of LOD scores which compare the two-point model to Amos' model. Figure 5.2 shows the LOD score curves comparing the two-point model to Amos' model under a constant recombination fraction of 0.02 between markers, but varying proportions of heritability due to the QTL. From Figure 5.2, we see that the LOD scores are quite small since they are less than 0.5, regardless of the magnitude of h_q^2; therefore, the recombination fraction does not seem to significantly contribute to the model. For mapping purposes, it appears that assuming all markers are tightly linked to any QTLs suffices to effectively map them.

We also note that for large values of h_q^2, the LOD scores decline as the distance between the QTL and marker decreases. Since the two-point model assumes that $\theta = 0$, we expect that Amos' model would not improve upon the two-point model significantly at markers close to the QTL, hence the small LOD scores. Recall that the recombination fraction between a marker and the QTL tends to zero as the distance between them decreases. Furthermore, the rise in LOD scores at markers far from the QTL indicates that the two-point model can be improved by allowing the recombination fraction to differ from zero.

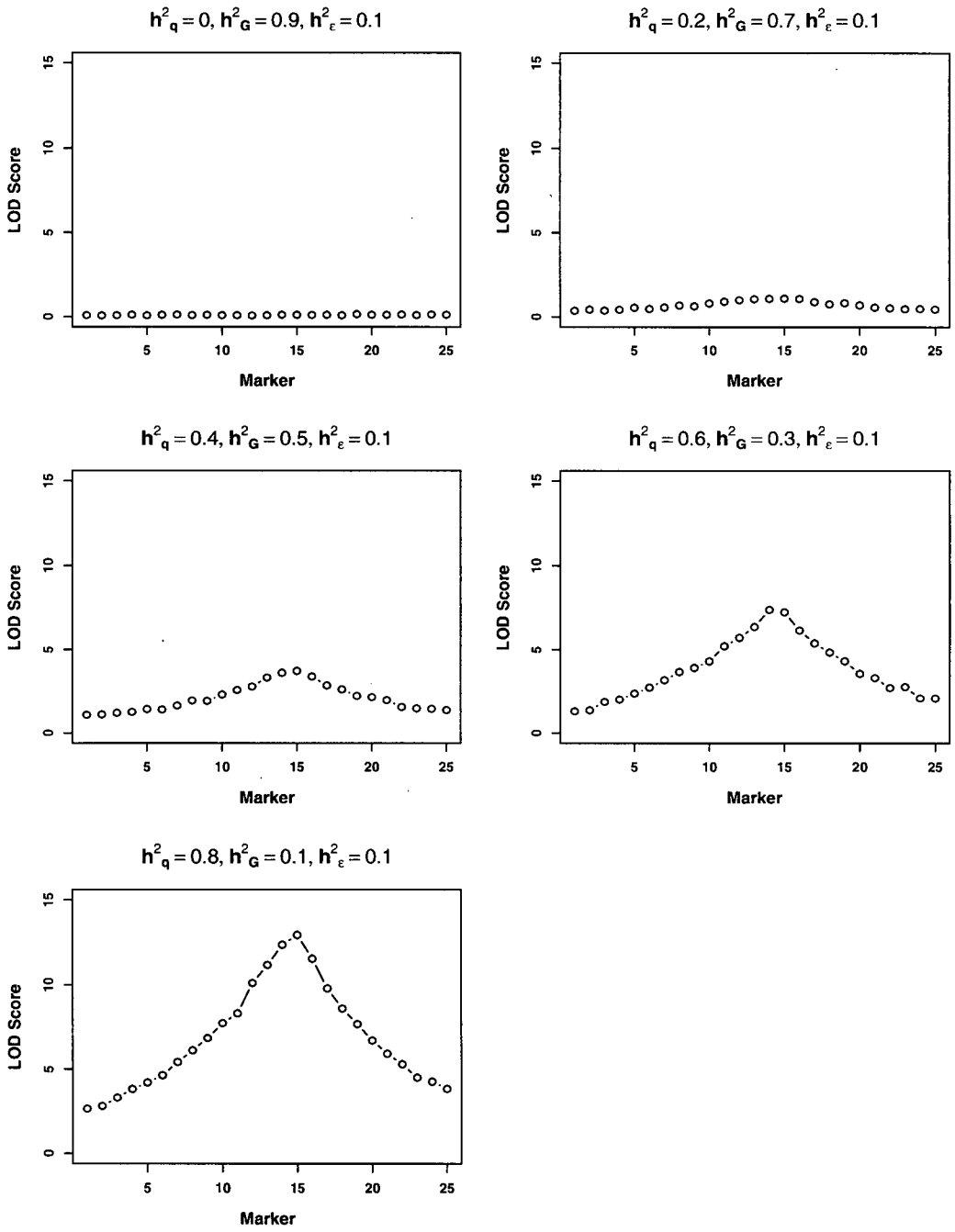

Figure 5.1: Polygenic versus Two-Point LOD Score Curves, $\theta_s = 0.02$

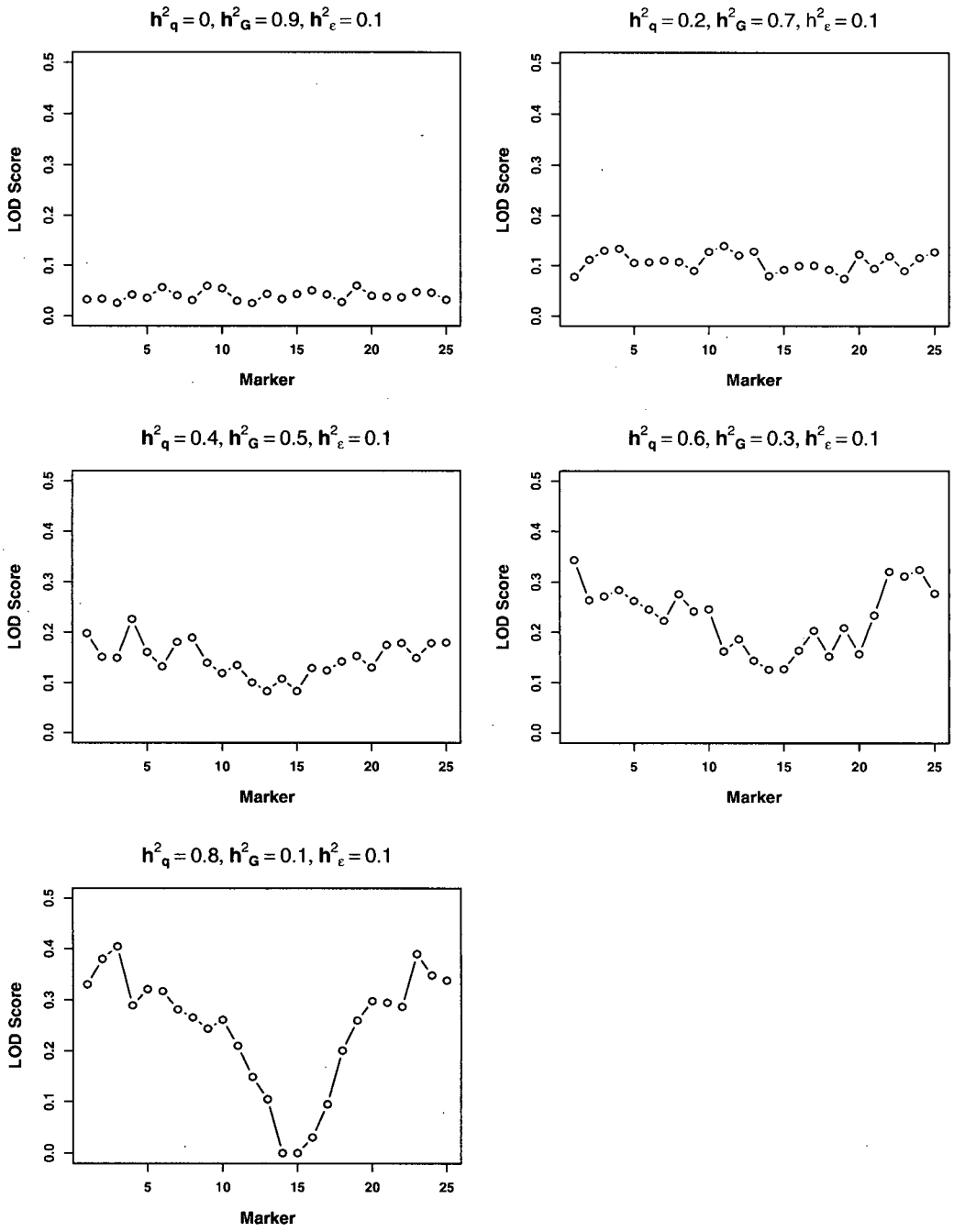

Figure 5.2: Two-Point versus Amos' LOD Score Curves, $\theta_s = 0.02$

5.3 Estimation of Variance Components

A secondary goal of linkage analysis is to estimate the variance components. Both the two-point model and Amos' model allow us to estimate the QTL variance component, polygenic variance component, and random deviation variance component. As opposed to directly estimating the variance components, we estimate the variance components as fractions of the total phenotypic variability, also known as heritabilities. We denote the QTL and polygenic heritabilities to be h_q^2 and h_G^2, respectively. Note that heritabilities are bounded between 0 and 1. In this section, we compare the estimated heritabilities from the two-point model and Amos' model. As in the previous section, we focus on the simulations where the QTL and polygenic variance components vary, and the recombination fraction between successive loci is fixed at 0.02.

5.3.1 Two-Point Model

Firstly, we examine the estimated heritabilities from the two-point model. We are particularly interested in the heritability due to the QTL and the heritability due to the polygenes. At each marker we averaged the 100 estimated QTL heritabilities. The average QTL heritabilities were then plotted versus marker location, as shown in Figure 5.3. From Figure 5.3, it is apparent that, on average, in the presence of a QTL, h_q^2, is always underestimated. The degree of underestimation becomes more severe as we move towards the extremities of the chromosome, and as the true QTL heritability increases. We suspect that the restricting assumption on the recombination fraction is the culprit for the poor estimation at the extremities of the chromosome. By assuming that the recombination fraction between the QTL and any marker is 0, we are inevitably fitting incorrect models at marker locations which are far from the QTL; hence, it is not unexpected to obtain poor estimates of the QTL heritability at these locations.

Although the estimated QTL heritability tends to be underestimated on average, it seems that we can obtain a fairly accurate estimate of h_q^2. Figure 5.3 shows that the maximum value of the estimated h_q^2 curve lies extremely close to the true value. Table 5.1 shows the average maximum value of the QTL heritability components, $\bar{h}^2_{q\_\max}$, over 100 replicates, along with 95% empirical probability intervals. From Table 5.1, we can see that, on average, the QTL heritability is overestimated, but the estimates improve as h_q^2 increases. As well, we can see

that the width of the empirical probability intervals decrease from about 0.30 to 0.20 as the QTL component increases.

Next we look at the estimation of the other genetic component, namely the polygenic heritability. The average estimated polygenic heritabilities are plotted against marker location in Figure 5.4. From these curves, we see that the true contribution of the polygenes tends to be overestimated. Although h_q^2 tends to be underestimated and and h_G^2 tends to be overestimated the accuracy of the estimates seem to be affected by the same factors. The problem of overestimation grows worse as the distance between the QTL and marker increases. As well, as h_G^2 decreases, the bias increases. Such poor estimates may be due to the fact that model misspecification becomes worse as the distance between the QTL and markers increases.

Because the problem of model misspecification weakens at markers close to the QTL, the estimates of the polygenic heritabilities at these markers seem to be quite accurate. From Figure 5.4, we see that the minimum value of the polygenic heritability curve gives a reasonable estimate of h_G^2. Table 5.2 shows the average minimum value of the estimated polygenic heritability, $\bar{h}^2_{G\_\min}$, over 100 replicates, along with 95% empirical probability intervals. From Table 5.2, we see that, on average, the polygenic heritability is underestimated. The widths of the 95% empirical probability intervals for the polygenic heritabilities are also larger than those for the QTL heritabilities. It seems that the polygenic component is more difficult to estimate than the QTL component.

5.3.2 Amos' Model

Next, we study the performance of Amos' model in estimating the heritabilities. Amos' model relaxes the assumption on the recombination fraction which is used in the two-point model. By not constraining the recombination fraction to be 0, model misspecification at the ends of the chromosome is no longer inevitable, but we add the burden of estimating another parameter.

The average of 100 estimated QTL heritabilities from Amos' model across the chromosome is shown in Figure 5.5. In contrast to the estimated QTL heritability curves from the two-point model, Amos' method appears to give a constant estimate of the QTL heritability regardless of the marker location. By not constraining the recombination fraction parameter, the model space becomes larger and Amos' model becomes less prone to model misspecification.

From Figure 5.5, we see that when h_q^2 is less than or equal to 0.60, it tends to be consistently overestimated on average; however, when h_q^2 is 0.80, it tends to be generally underestimated. Although the direction of bias is not apparent, the magnitude of the bias seems to increase as the QTL heritability decreases.

Recall that for the two-point method, it seemed obvious to use the maximum value of the QTL heritability curve as an estimate for h_q^2. In Amos' method, it is not apparent how to estimate h_q^2. One possible estimate may be to average the estimates of h_q^2 across the chromosome For comparison purposes, Table 5.3 shows the average estimated mean value of the QTL heritability estimates, $\bar{h}^2_{q\_\text{avg}}$, across the chromosome over 100 replicates. In general, the two-point model seems to provide more accurate estimates of the QTL heritability since the widths of the probability intervals are usually smaller.

We also look at how accurately Amos' model estimates the heritability due to polygenes. Figure 5.6 shows the average of the 100 estimates of h_G^2 from Amos' model at each marker. Like the two-point model, Amos' model seems to give biased estimates of h_G^2 and the direction of bias is opposite to the direction of bias for the estimates of h_q^2. Again the severity of the bias depends upon the magnitude of h_G^2. Although the estimates are biased, the bias appears to be relatively constant across the entire chromosome.

The question of how to estimate the polygenic component from Amos' model arises, as it did for the estimation of the QTL component. We again look at the average estimated mean value of the polygenic heritability, $\bar{h}^2_{G\_\text{avg}}$, over the entire chromosome (see Table 5.4). Note that Amos' model again results in wider probability intervals for the polygenic heritability. Using the minimum value of the estimated polygenic heritability across the chromosome from the two-point model appears to be a more reliable estimate than the average value of the estimated polygenic heritabilities across the entire chromosome from Amos' model.

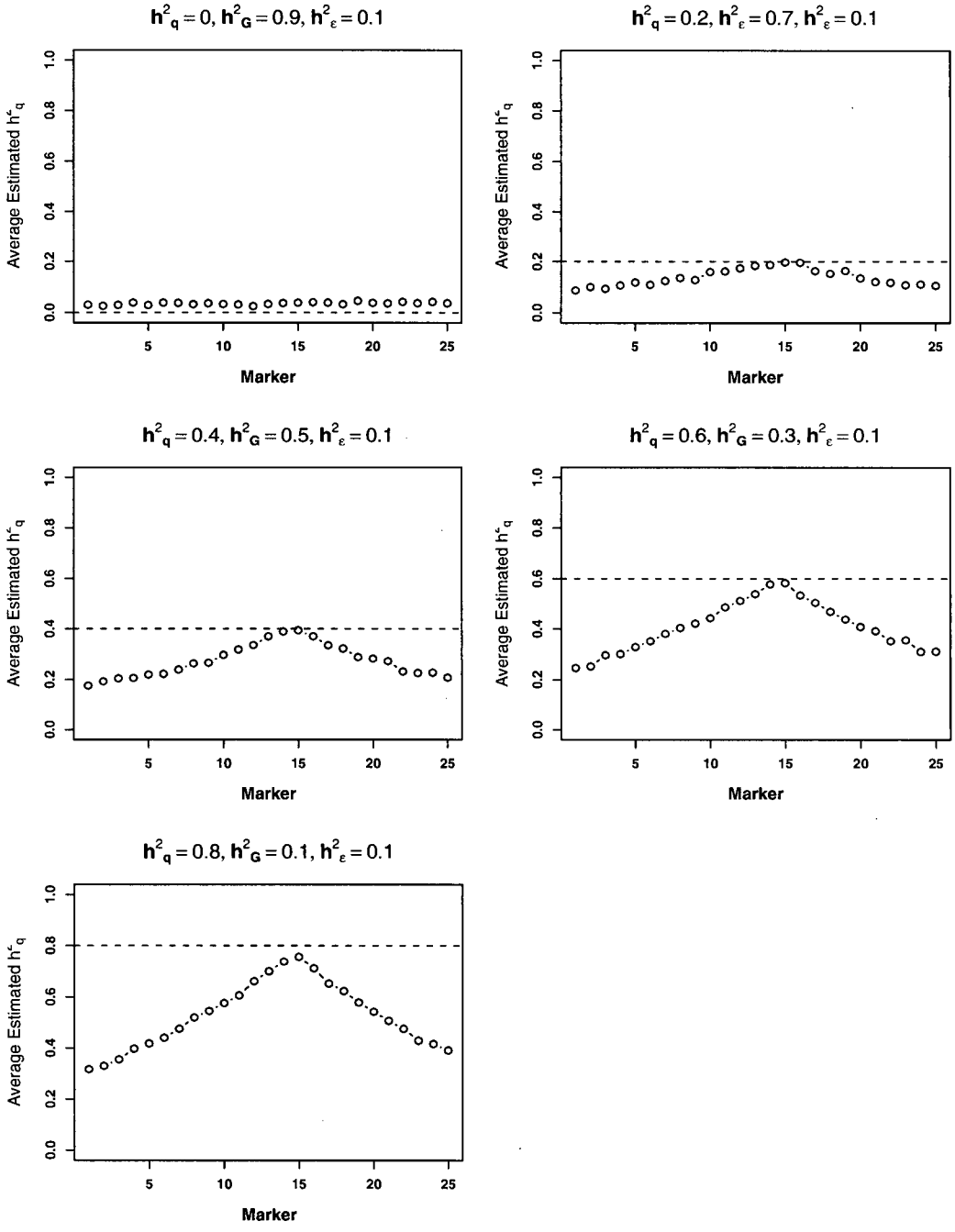

Figure 5.3: Average Estimated QTL Heritability from Two-Point Model, $\theta_s = 0.02$

Table 5.1: Average Estimated Maximum Value of QTL Heritability from Two-Point Model

h_q^2	$\bar{h}^2_{q\_\max}$	95% Emp. Prob. Interval
0.0	0.168	(0.062,0.304)
0.2	0.328	(0.168,0.501)
0.4	0.512	(0.351,0.646)
0.6	0.672	(0.537,0.835)
0.8	0.817	(0.706,0.900)

Table 5.2: Average Estimated Minimum Value of Polygenic Heritability from Two-Point Model

h_G^2	$\bar{h}^2_{G\_\min}$	95% Emp. Prob. Interval
0.9	0.733	(0.596, 0.865)
0.7	0.565	(0.380, 0.743)
0.5	0.385	(0.217, 0.593)
0.3	0.223	(0.006, 0.397)
0.1	0.076	(0.000, 0.203)

Table 5.3: Average Estimated Mean Value of QTL Heritability from Amos' Model

h_q^2	$\bar{h}^2_{q\_\text{avg}}$	95% Emp. Prob. Interval
0.0	0.183	(0.006,0.453)
0.2	0.381	(0.157,0.661)
0.4	0.514	(0.328,0.755)
0.6	0.664	(0.507,0.841)
0.8	0.756	(0.629,0.857)

Table 5.4: Average Estimated Mean Value of Polygenic Heritability from Amos' Model

h_G^2	$\bar{h}^2_{G\_\text{avg}}$	95% Emp. Prob. Interval
0.9	0.716	(0.464,0.913)
0.7	0.513	(0.250,0.734)
0.5	0.386	(0.124,0.588)
0.3	0.233	(0.066,0.404)
0.1	0.139	(0.032,0.272)

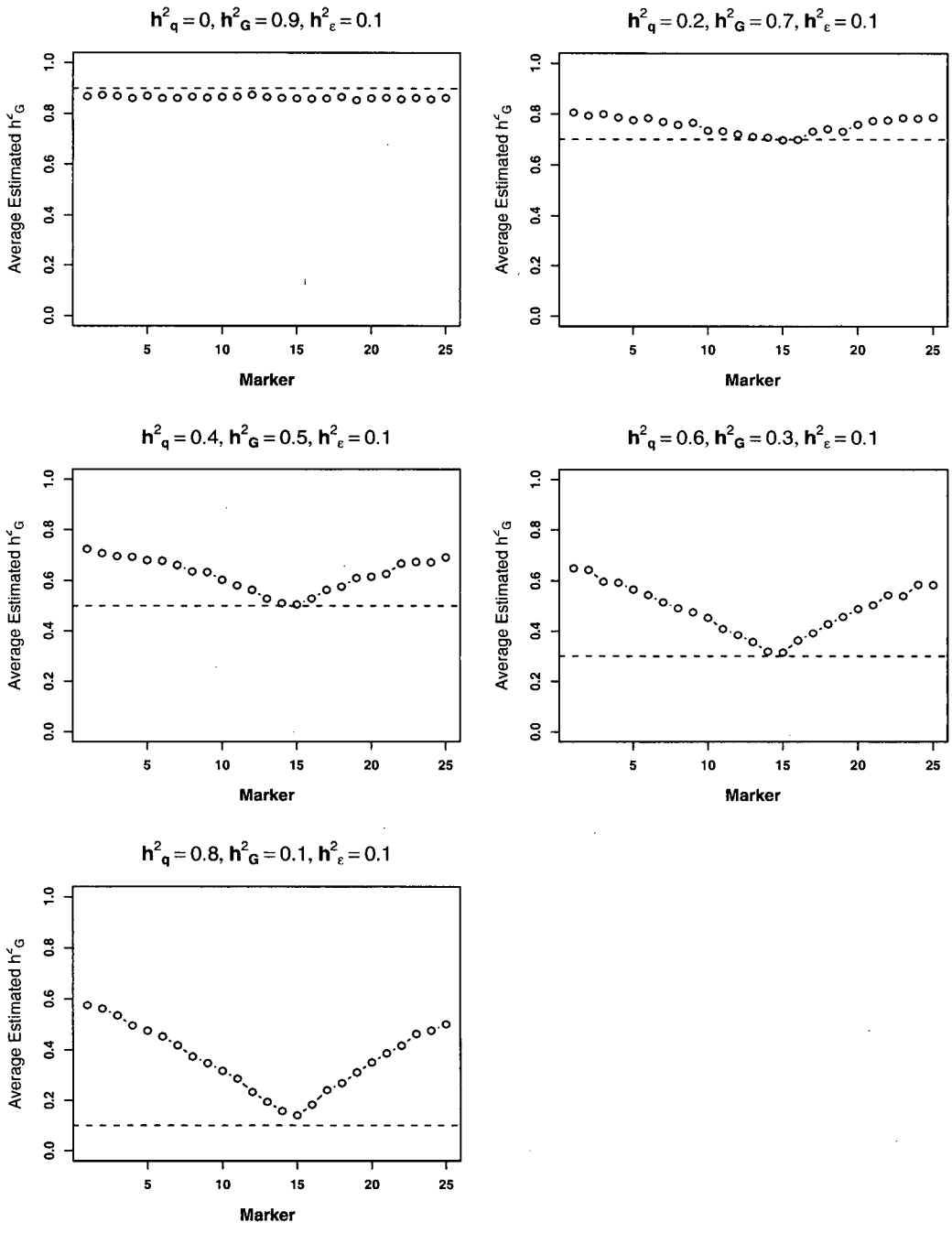

Figure 5.4: Average Estimated Polygenic Heritability from Two-Point Model, $\theta_s = 0.02$

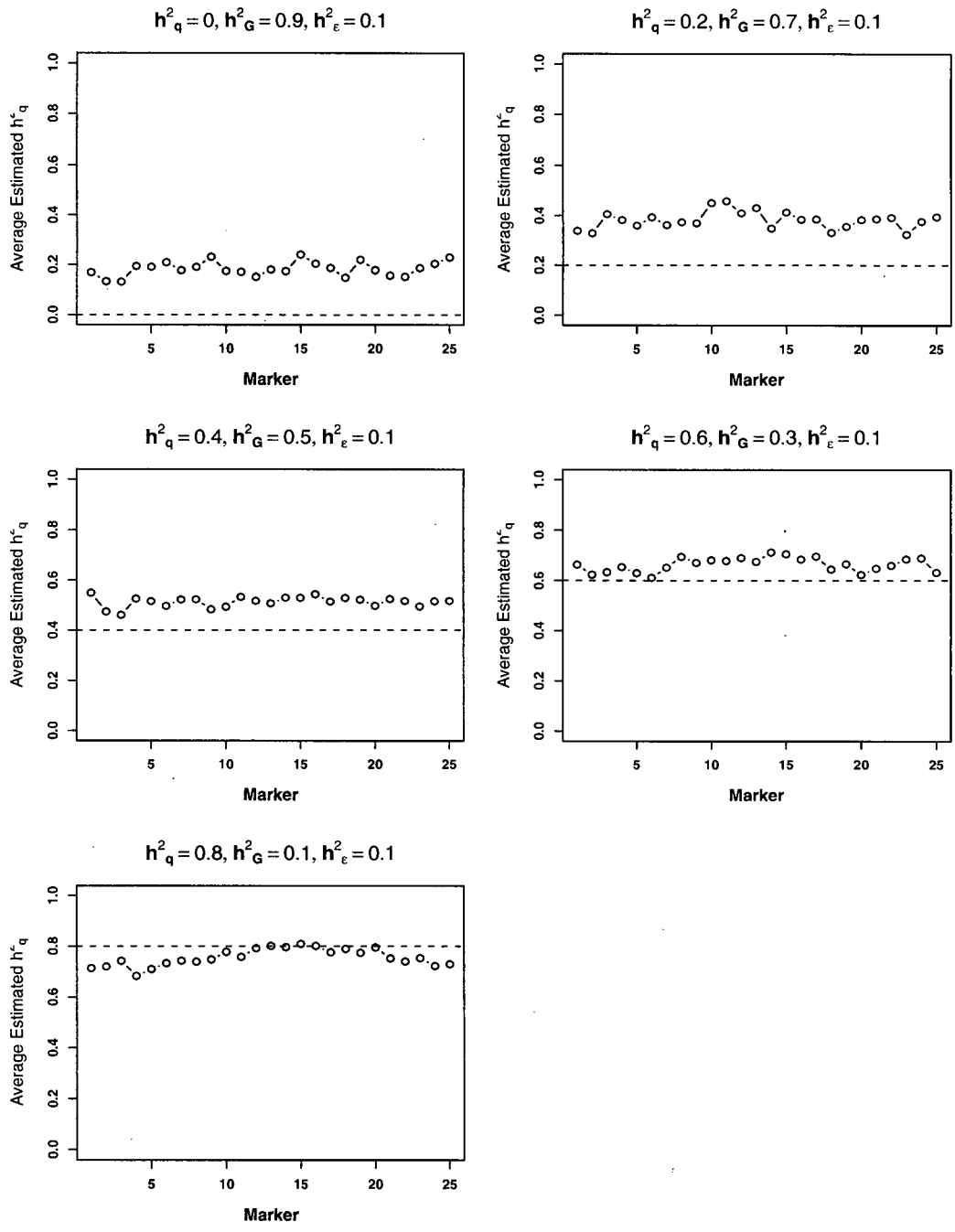

Figure 5.5: Average Estimated QTL Heritability from Amos' Model, $\theta_s = 0.02$

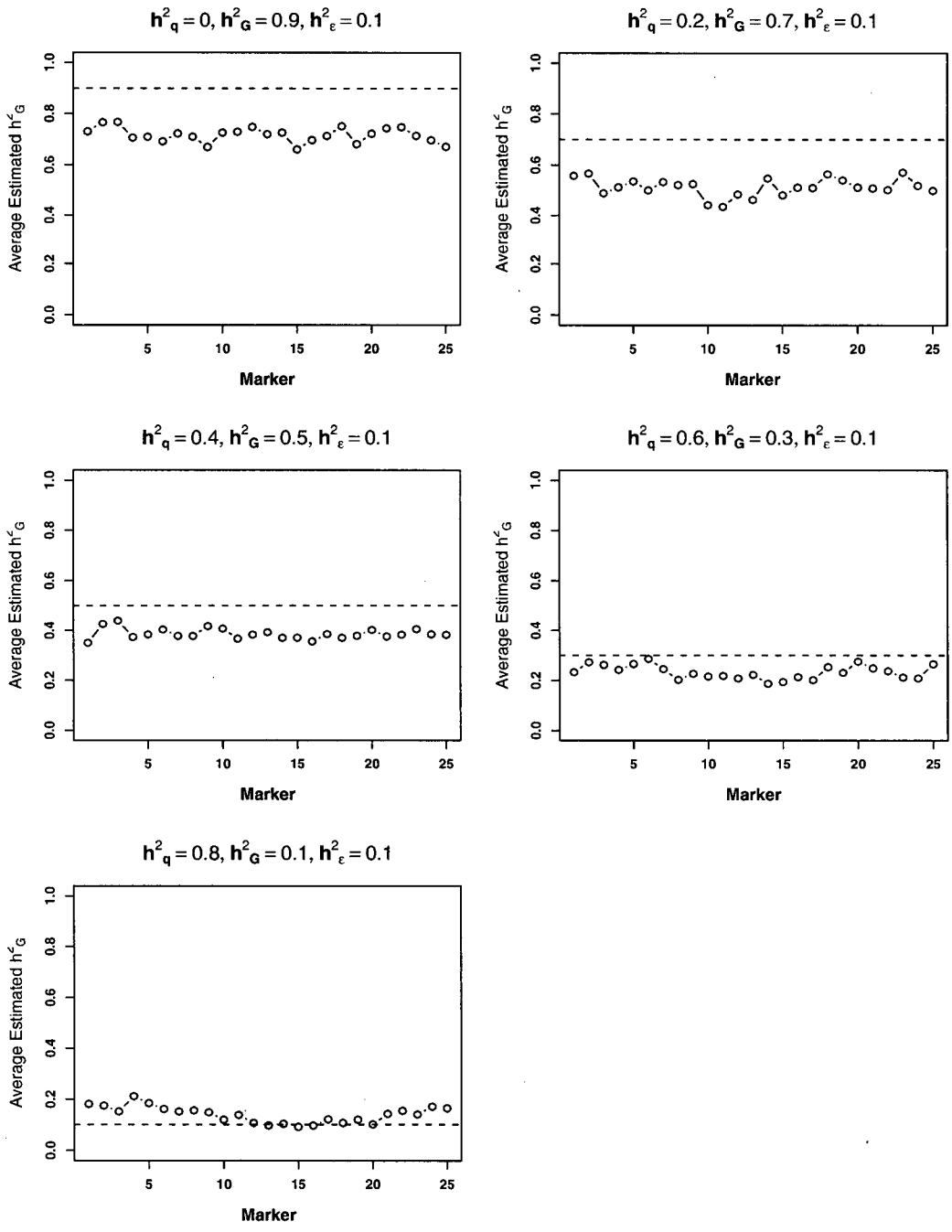

Figure 5.6: Average Estimated Polygenic Heritability from Amos' Model, $\theta_s = 0.02$

5.4 Estimation and Effect of the Recombination Fraction

Recombination fractions may not be used as a distance measure, but they may still give us an idea of the distance between a QTL and a marker; whereas, a LOD score curve can only tell us what regions of the chromosome are likely to contain the QTL. We investigate how well Amos' model is able to estimate the recombination fraction. As well, in this section, we discuss the effect that recombination has on predicting regions housing QTLs via LOD score curves and estimating variance components.

5.4.1 Amos' Model

Firstly, we assess the estimates of the recombination fractions from Amos' model. Figure 5.7 shows the average of 100 estimated recombination fractions at each marker on the chromosome. We can see that as the QTL heritability component increases a trough in the curve slowly emerges. This trend in the estimated recombination fractions is expected. Although the true recombination fraction is 0.02 between all adjacent markers and loci, the recombination fraction between the QTL and successively farther markers increase non-linearly. Recall that the estimated variance components from the two-point model were non-constant across the chromosome, whereas the variance components estimates from Amos' model were constant. It appears that Amos' model is able to unconfound the trends seen in the estimates of the two-point model by the inclusion of a recombination fraction parameter.

Because the mapping function from the recombination fraction to the genetic map distance is unknown, it is difficult to assess the accuracy of the estimated recombination fractions. From our simulations we know that the recombination fraction between the QTL and marker 14 and the recombination fraction between the QTL and marker 15 is 0.02; however, we cannot be certain of the true recombination fraction between the QTL and any other markers.

5.4.2 Varying Recombination Fraction

We also investigated the effect that varying the recombination fraction would have on the ability to detect QTLs. Equivalently, this would be examining how the distance between markers affects the ability to map QTLs. Recall that a recombination fraction close to 0 implies tight linkage

and a recombination fraction close to 1/2 implies loose linkage.

In our simulations, we allowed the true value of the recombination fraction between successive loci to vary between 0.02 and 0.04. By increasing the true value of the recombination fraction, we are increasing the genetic distance between markers. Firstly, we assess how looser linkage affects the LOD scores. Figure 5.8 shows the average LOD scores of 100 replicates from comparing the likelihoods of the polygenic model to the two-point model. By comparing Figures 5.1 and 5.8, we see that the LOD scores are higher when $\theta_s = 0.02$, or when the markers are more tightly linked to the QTL. This observation is expected since the more tightly linked the markers are to the QTL, the greater the likelihood that the marker information will be able to explain the variability in the phenotype. As well, we see that the LOD score curves have more narrow peaks above the QTL location when $\theta_s = 0.04$. This observation also concurs with our intuition. When $\theta_s = 0.04$, the genetic distance between the QTL and more distant markers grows more rapidly than when $\theta_s = 0.02$; therefore, we would expect a sharper decline in the LOD scores when the recombination fraction increases. The LOD scores which compare the two-point model to Amos' model share similar properties to those LOD scores discussed above. From Figure 5.9, the dips toward 0 over the QTL location are more narrow when $\theta_s = 0.04$. As mentioned before, because the two-point model assumes that $\theta = 0$, model misspecification becomes more severe as the true recombination fraction increases.

Secondly, we determine how the spacing between markers affects the estimates of the variance components. Figure 5.10 and Figure 5.11 show the average estimated QTL heritabilities and polygenic heritabilities from the two-point model. In comparison with Figures 5.3 and 5.4, the observation of a narrower peak or dip again emerges. We also note that at markers which are farther from the QTL, the estimates of the heritabilities are worse when $\theta_s = 0.04$. The average estimated QTL and polygenic heritabilities from Amos' model are shown in Figures 5.12 and 5.13, respectively. Increasing the recombination fraction by a factor of 2 causes the estimates to no longer be constant across the entire chromosome, which is what we observed when $\theta_s = 0.02$ in Figures 5.5 and 5.6. Instead, the estimated heritabilities become significantly worse as we move farther from the QTL. Again, the estimated QTL heritabilities and estimated polygenic heritabilities seem to be more severely biased for small values of h_q^2 or equivalently large values of h_G^2.

Finally, we plot the average estimated recombination fractions in Figure 5.14. As expected, the estimated recombination fractions decrease more rapidly as the distance between the QTL and markers decreases.

5.5 Summary

From our simulation study, we evaluated the properties of the two-point model and Amos' model in terms of their ability to detect regions where QTLs reside and their ability to accurately estimate the variance components. We saw that although the two-point model is prone to model misspecification at markers which are far from the QTL, it is still able to detect QTL regions. Furthermore, Amos' model does not appear to significantly improve upon its sub-model's ability to detect QTLs.

When estimating variance components, the two-point model inevitably gives poor estimates at marker locations far from the QTL; however, if the QTL heritability is large, then there is a distinct rise or fall in the estimates across the chromosome and the optimal value seems to give a reasonable estimate of the heritability component. On the other hand, Amos' approach is not prone to model misspecification, but it seems to be quite challenging to simultaneously estimate the variance components, as well as the recombination fraction. Because these parameters seem to be confounded, the estimated heritabilities, although fairly consistent across the chromosome, seem to be quite poor.

Lastly, by varying the distance between markers, we see that the two-point model's estimates of the variance components improves as we increase the concentration of markers along the chromosome. However, if detection of the QTL region is the primary goal, the spacing between successive markers does not need to be extremely dense.

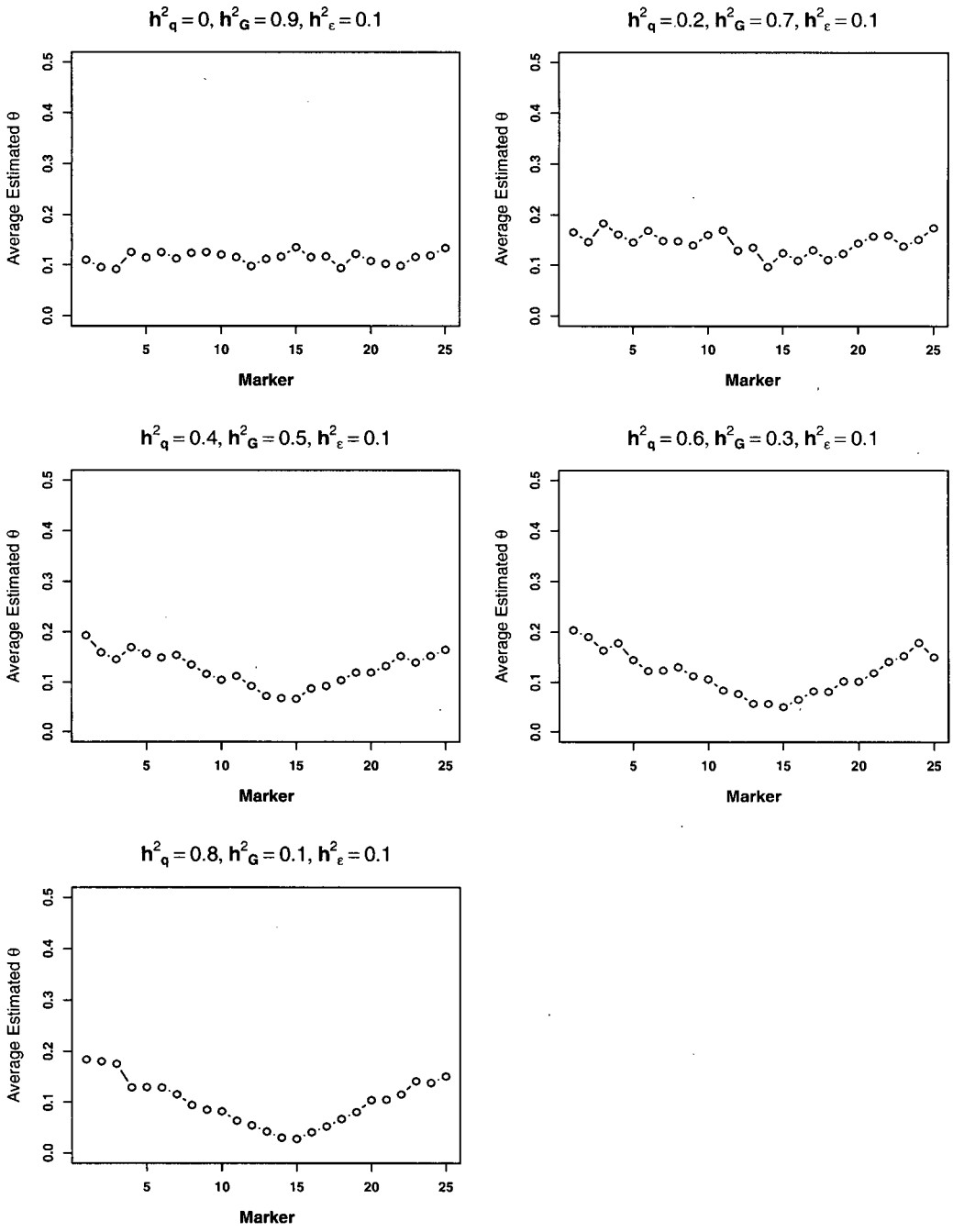

Figure 5.7: Average Estimated Recombination Fractions from Amos' Model, $\theta_s = 0.02$

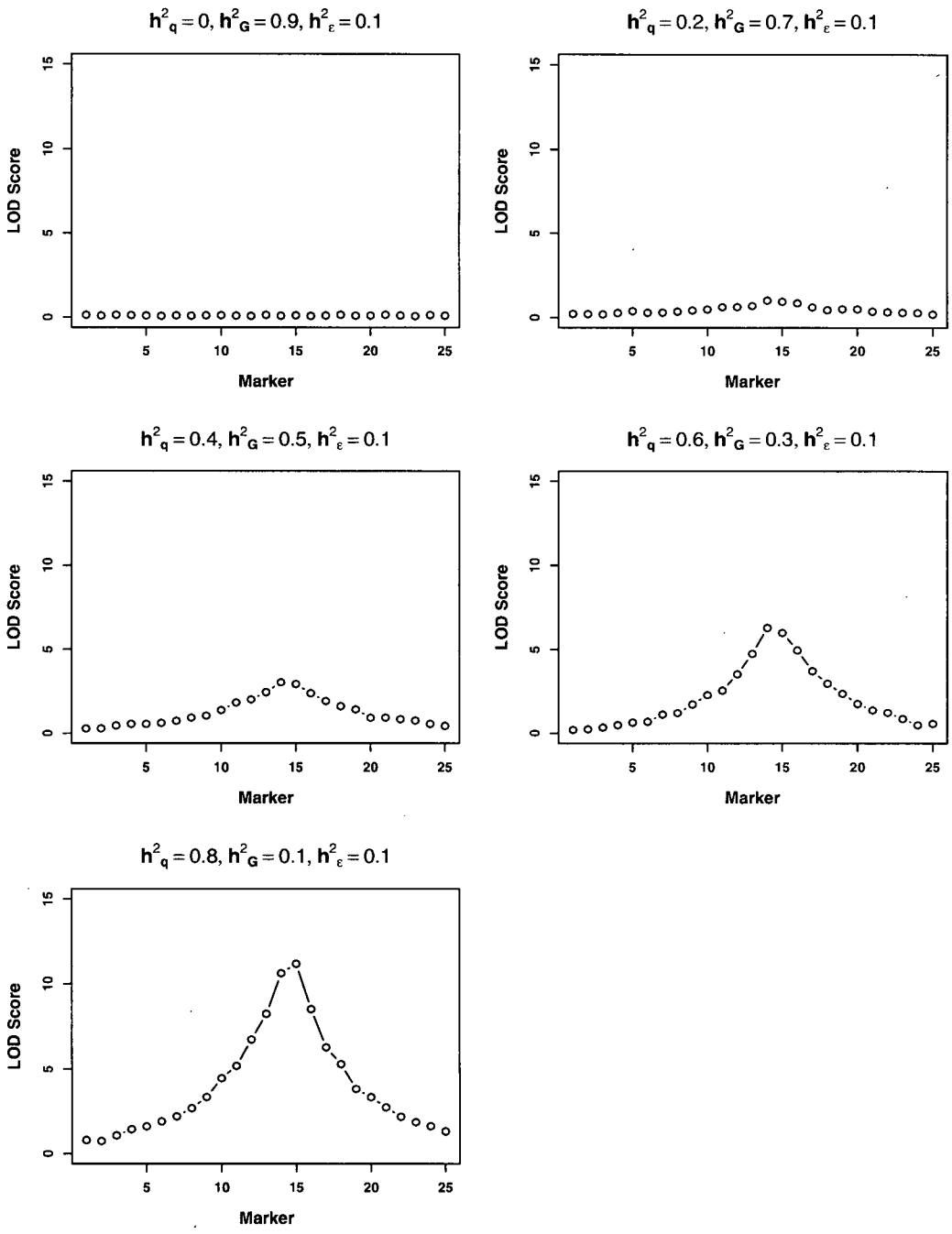

Figure 5.8: Polygenic versus Two-Point LOD Score Curves, $\theta_s = 0.04$

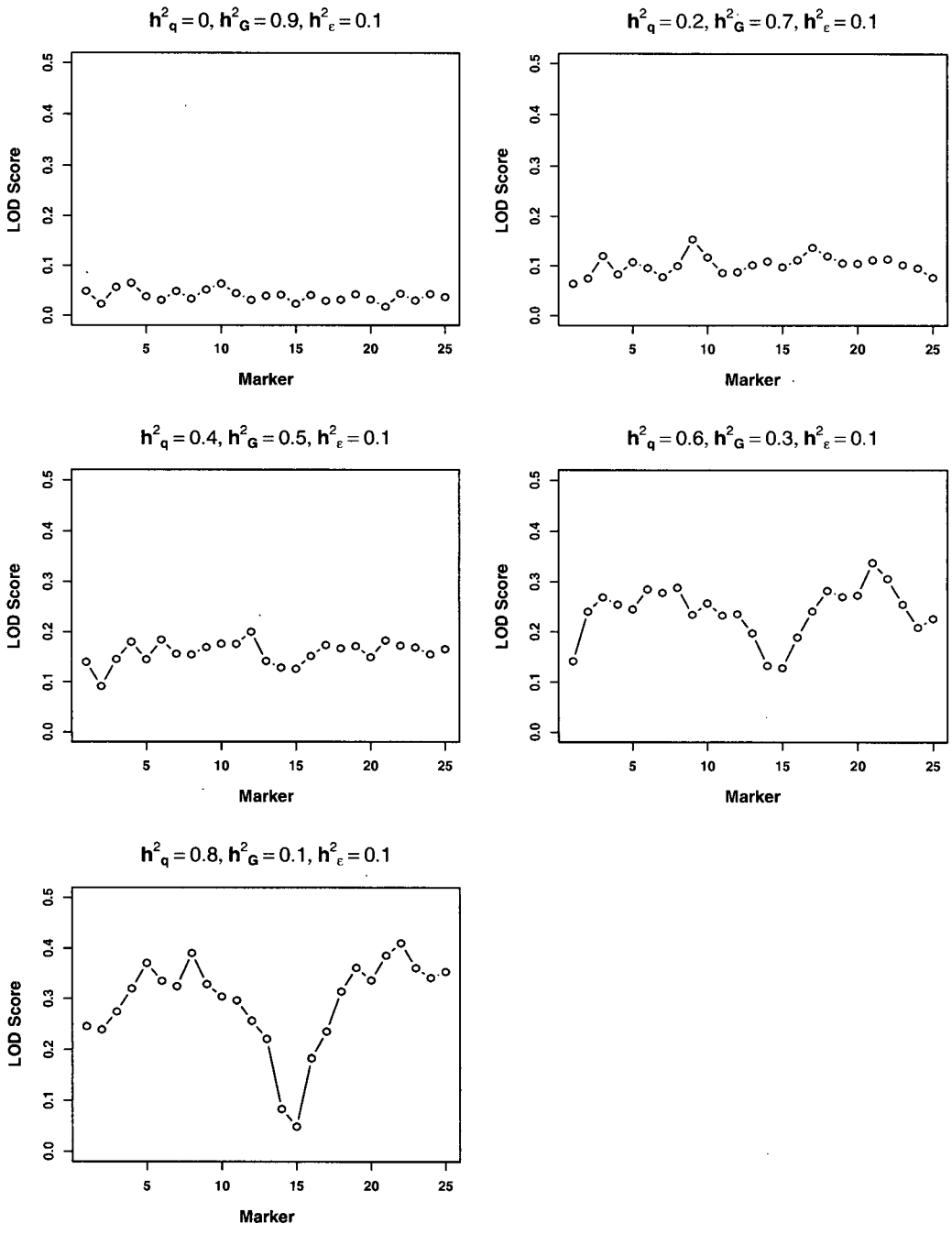

Figure 5.9: Two-Point versus Amos' LOD Score Curves, $\theta_s = 0.04$

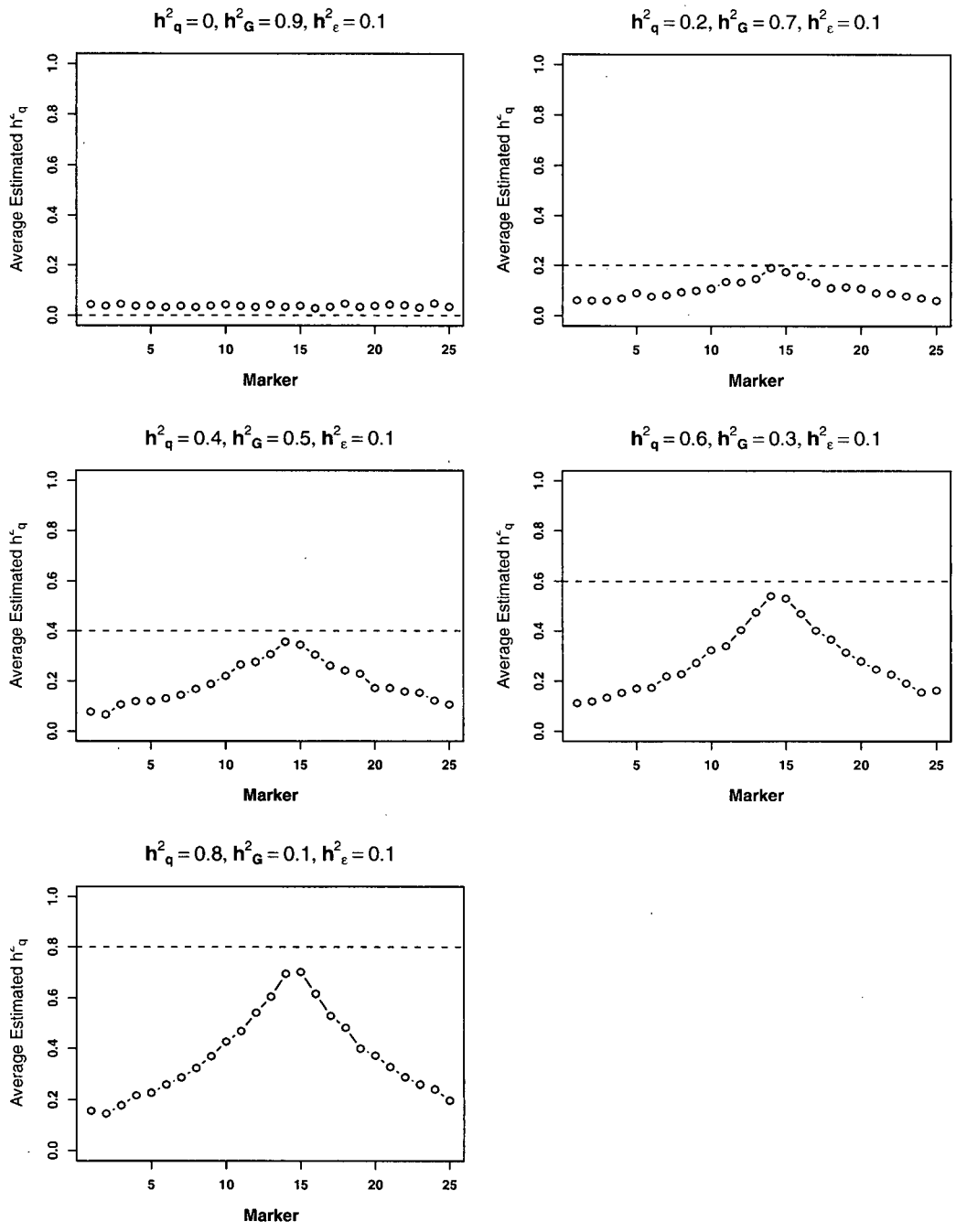

Figure 5.10: Average Estimated QTL Heritability from Two-Point Model, $\theta_s = 0.04$

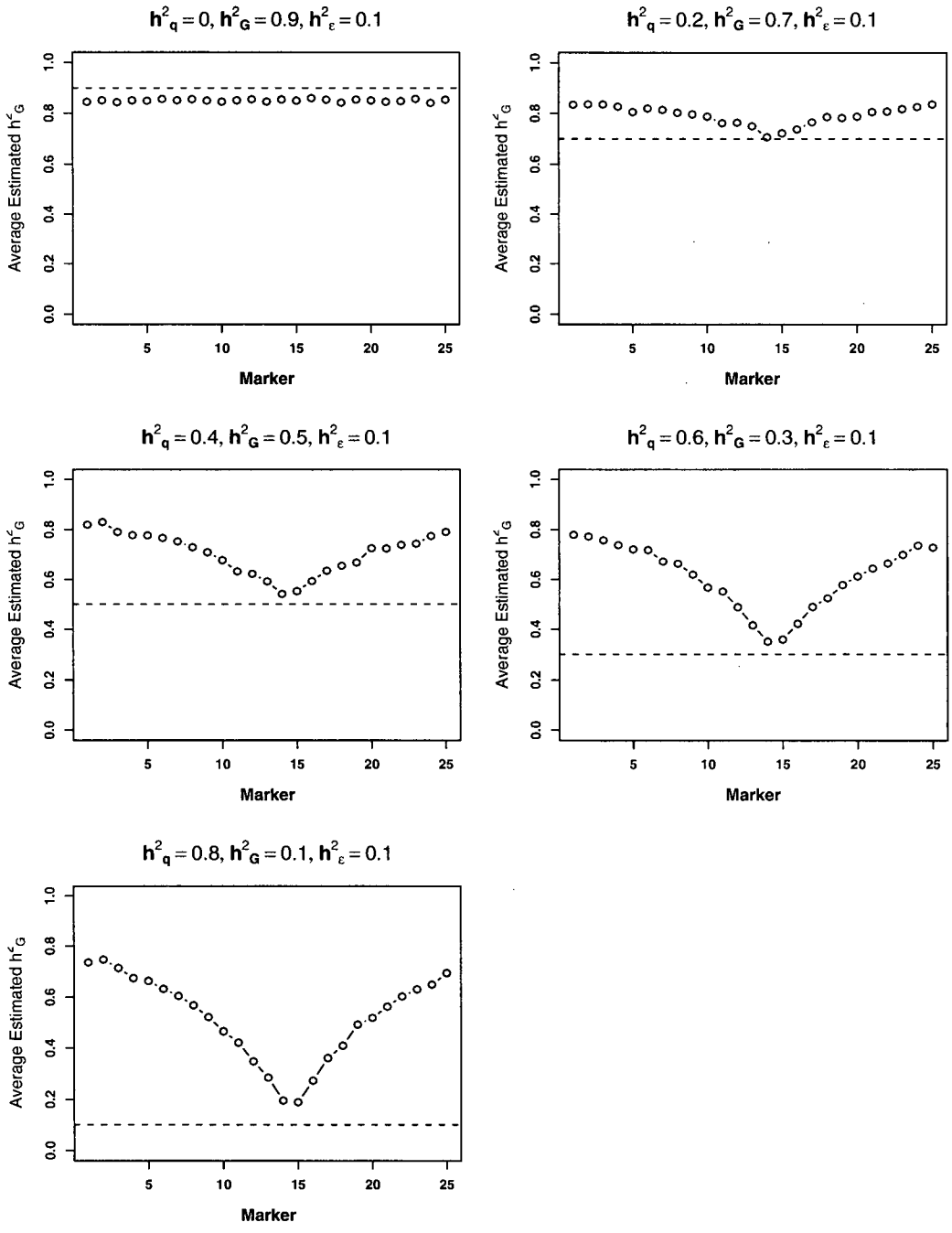

Figure 5.11: Average Estimated Polygenic Heritability from Two-Point Model $\theta_s = 0.04$

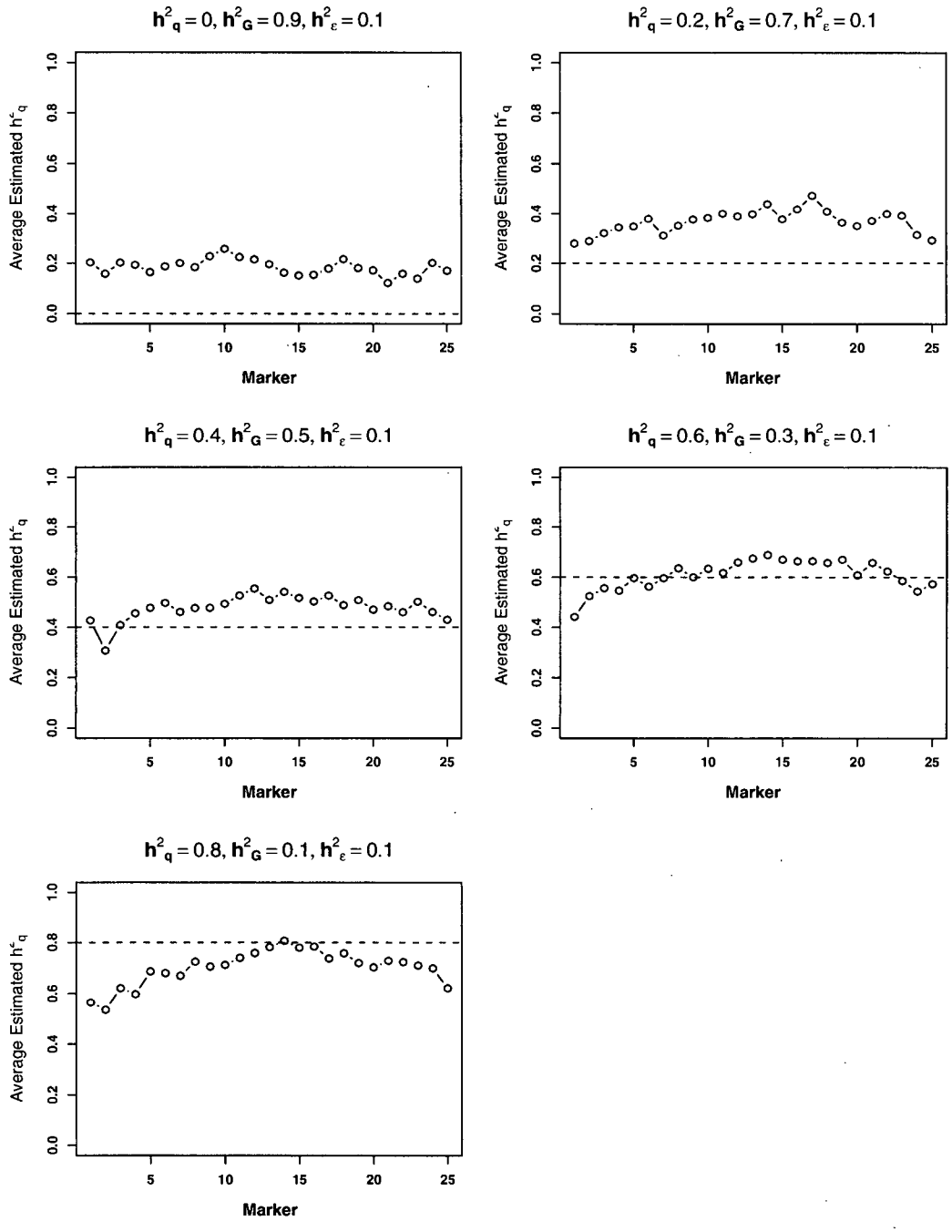

Figure 5.12: Average Estimated QTL Heritability from Amos' Model, $\theta_s = 0.04$

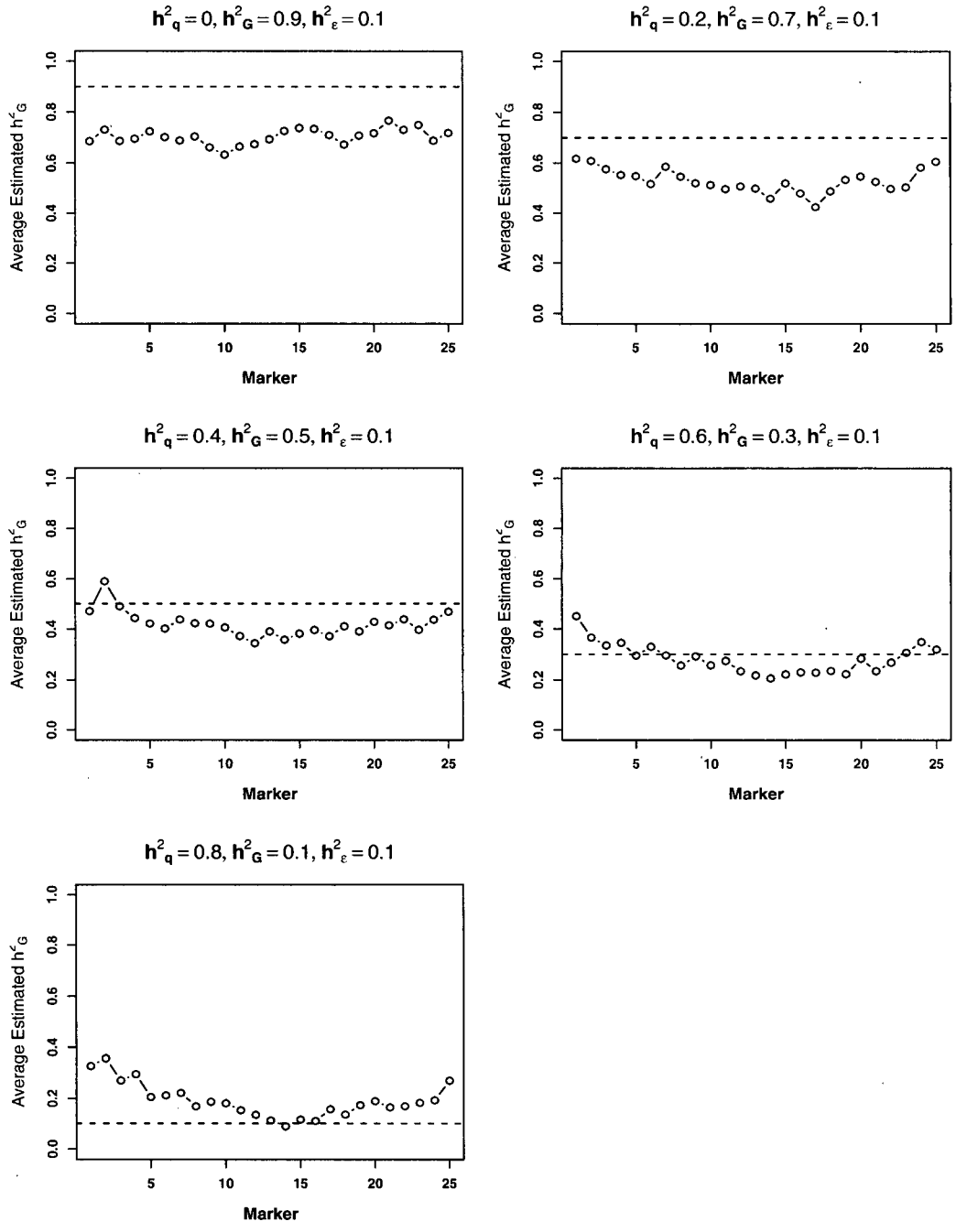

Figure 5.13: Average Estimated Polygenic Heritability from Amos' Model, $\theta_s = 0.04$

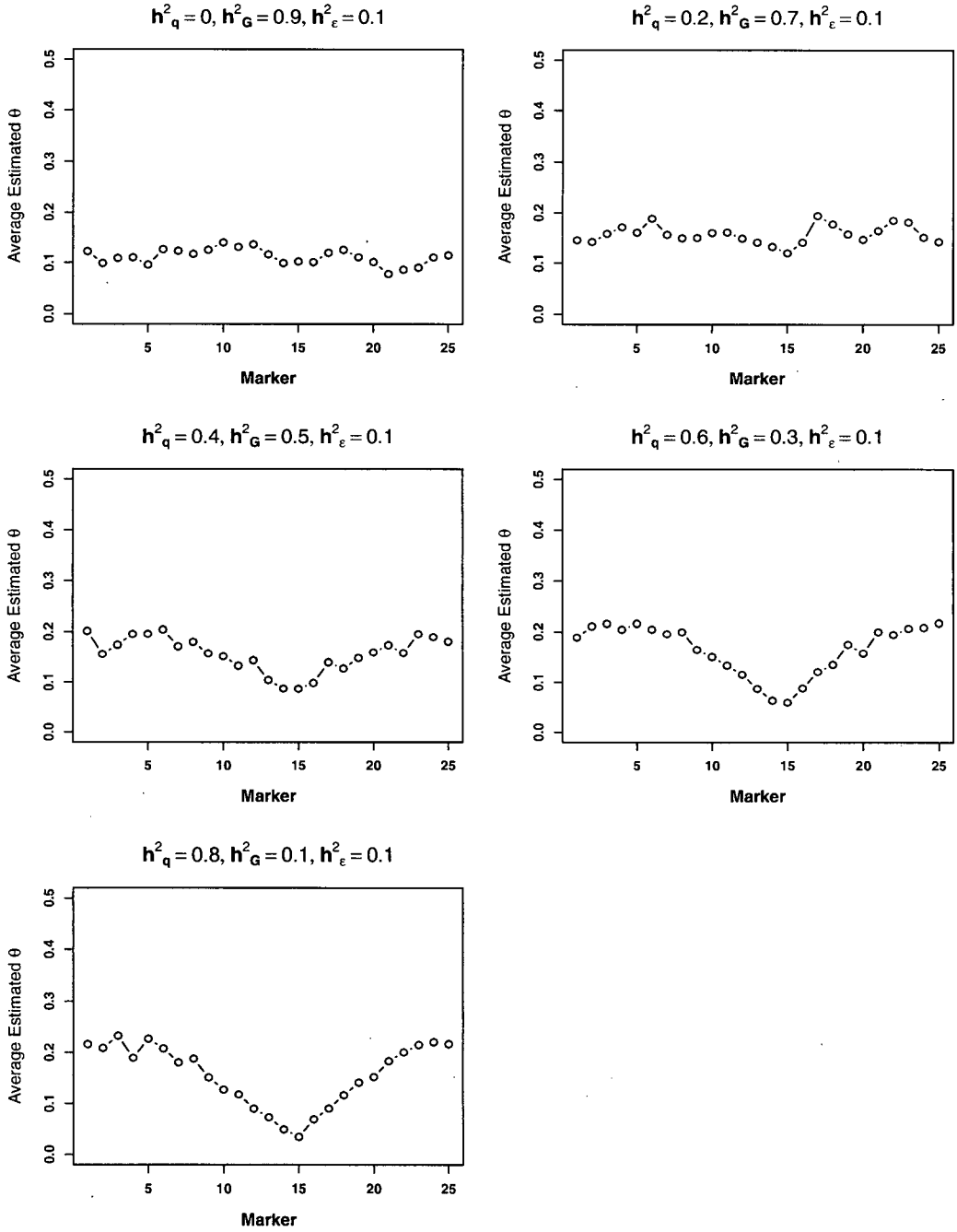

Figure 5.14: Average Estimated Recombination Fractions from Amos' Model, $\theta_s = 0.04$

Chapter 6

Application to a Complex Trait

6.1 Genetic Analysis Workshops

The Southwest Foundation for Biomedical Research encourages collaboration amongst researchers who are using statistical genetics, through biennial Genetic Analysis Workshops (GAWs) [12]. The focus of each workshop revolves around a current analytical problem in genetic epidemiology. These workshops provide researchers with the opportunity to discuss, compare, and assess the performance of competing statistical methods. As a basis for comparison, the participants are given a common problem consisting of real or simulated data. Through the participants' independent analyses, it is not only interesting to see the various approaches to tackling a problem, but also enlightening to realize the diverse results that can be obtained from a common dataset. Further information on the Genetic Analysis Workshops may be found at http://www.sfbr.org/external/gaw/. We utilize the simulated data from the tenth Genetic Analysis Workshop (GAW10) to assess the performance of the variance-components method of linkage analysis on a complex trait.

6.1.1 Tenth Genetic Analysis Workshop Data

GAW10 focussed on comparing statistical methods in terms of power in detecting major genes and accuracy of parameter estimates when studying an oligogenic disease with quantitative risk factors, or complex traits. Recall that a complex trait is a phenotype which is determined by multiple genes and environmental factors, and does not follow Mendelian laws of inheritance.

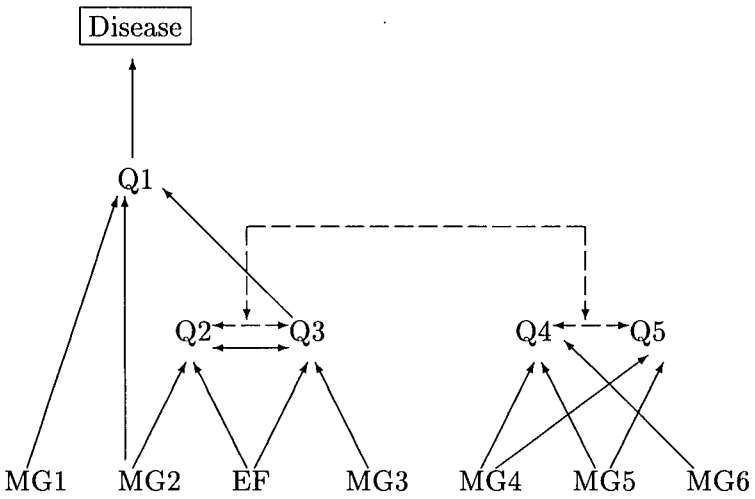

Figure 6.1: Schematic Diagram of Simulated Oligogenic Disease in GAW10

The mechanisms underlying such traits are challenging to uncover since the variability of a complex trait arises from many factors which may interact in an intricate manner. The simulated oligogenic disease in GAW10 is directly or indirectly influenced by various interactions amongst six major genes (MG1, MG2, MG3, MG4, MG5, and MG6), five quantitative risk factors (Q1, Q2, Q3, Q4, and Q5), as well as environmental (EF) and polygenic (PG) factors. An individual is declared to be affected if the first quantitative trait, Q1, exceeds a threshold of 40. The model which generated this common oligogenic disease is reproduced from MacCluer et al [18] in Figure 6.1. In Figure 6.1, single-headed solid arrows indicate the influence of the major genes or quantitative traits on one of the five quantitative traits. Also, double-headed solid arrows represent residual genetic correlation, and double-headed dashed arrows indicate residual environmental correlations.

In addition to the general structure of the disease, we are also know other specific genetic details regarding gene location and the magnitude of each variance component. Each major gene has two alleles with differing frequencies. As well, some traits were influenced by gender, age, and environmental effects. The percentage of variability in each trait attributable to each major gene and the random components range from less than 1% to 63%. The simulated variance components table from GAW10 is reproduced in Table 6.1. $Q1_M$ and $Q1_F$ denote the first quantitative trait for males and females, respectively. Similarly, $Q3_M$ and $Q3_F$ represent

Table 6.1: GAW10 Simulated Variance Components (in %)

Source	$Q1_M$	$Q1_F$	Q2	$Q3_M$	$Q3_F$	Q4	Q5
MG1	21.67	20.65	0.00	0.00	0.00	0.00	0.00
MG2	0.47	0.45	18.48	0.00	0.00	0.00	0.00
MG1HMG2	12.99	12.38	0.00	0.00	0.00	0.00	0.00
MG3	1.45	6.06	0.00	5.82	21.35	0.00	0.00
MG4	0.00	0.00	0.00	0.00	0.00	28.00	14.00
MG5	0.00	0.00	0.00	0.00	0.00	16.00	23.00
MG6	0.00	0.00	0.00	0.00	0.00	11.00	0.00
PG	6.26	5.97	14.78	25.15	21.00	0.00	0.00
Age	7.98	7.61	0.00	0.00	0.00	0.00	0.00
EF	5.56	5.30	13.04	22.33	18.65	0.00	0.00
Random	43.62	41.58	53.70	46.70	39.00	45.00	63.00
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Table 6.2: GAW10 Major Gene Locations

Major Gene	Chromosome	Location (distance in centiMorgans)
MG1	5	D5G14 $\leftarrow$ 0.8 $\rightarrow$ MG1 $\leftarrow$ 1.6 $\rightarrow$ D5G15
MG2	8	D8G26 $\leftarrow$ 0.3 $\rightarrow$ MG2 $\leftarrow$ 0.6 $\rightarrow$ MG4 $\leftarrow$ 0.6 $\rightarrow$ D8G27
MG3	4	D4G14 $\leftarrow$ 0.8 $\rightarrow$ MG3 $\leftarrow$ 1.9 $\rightarrow$ D4G15
MG4	8	D8G26 $\leftarrow$ 0.3 $\rightarrow$ MG2 $\leftarrow$ 0.6 $\rightarrow$ MG4 $\leftarrow$ 0.6 $\rightarrow$ D8G27
MG5	9	D9G8 $\leftarrow$ 1.0 $\rightarrow$ MG5 $\leftarrow$ 0.3 $\rightarrow$ D9G9
MG6	10	D10G7 $\leftarrow$ 0.5 $\rightarrow$ MG6 $\leftarrow$ 2.6 $\rightarrow$ D10G8

the third quantitative trait for males and females, respectively. Note that there is epistasis or masking, between MG1 and MG2. The six major genes were located on five of ten chromosomes with 24 to 50 unequally spaced markers. The location of each major gene as given by MacCluer et al [18] is shown in Table 6.2. Adhering to the GAW10 notation of marker locations, we note that D5G14 refers to marker 14 on chromosome 5. A more detailed description of the simulation procedure may be found in MacCluer et al [18].

The participants of GAW10 were given two types of data. The first type consisted of 200 replicates of 239 nuclear families with a total of 1164 individuals. Each nuclear family was simulated to have at least two offspring. The second type of data consisted of 200 replicates of 23 extended pedigrees with 1497 individuals. Each extended family included two parents and all first, second, and third degree relatives.

In our linkage analyses, we concentrated on estimating some of the variance components

for quantitative trait Q1. From Table 6.1, we see that Q1 is influenced by three covariates: age, sex, environment. As well, approximately 21% of the variation in Q1 is directly due to the MG1 and about 0.46% is directly from MG2. The influence of sex can be seen through the effect of MG3 on Q1. MG3 indirectly contributes about 1.5% or 6.1% of the variance through Q3 depending on sex. The polygenic variance component is simulated to be approximately 6.12%. When fitting the sporadic, polygenic, two-point, and Amos' models, age, sex, and environmental covariates are fit as x_{ik}'s in the model since each covariate was found to be significant in the sporadic model. Furthermore, we simultaneously estimated the variance components and covariate effects. Because we know that extended families are more powerful in being able to detect linkage than nuclear families, we also focussed on analyzing the data from the extended pedigrees. A summary of the GAW10 participants' analyses is given by Wijsman and Amos [27].

Our objective is to determine how well the variance-components approach to linkage analysis is able to detect MG1 and MG3, which are located on chromosome 5 and 4 respectively, and estimate the variance attributable to them. Note that chromosome 5 has 25 markers and chromosome 4 has 29 markers. We apply the variance-components methods on the first 60 of the extended pedigree replicates simulated for GAW10.

6.2 Detection of Quantitative Trait Loci via LOD Scores

To estimate the regions of the chromosomes with major loci affecting the complex trait Q1, we determine if any markers are linked to the QTL by looking at LOD score curves. Figure 6.2 shows the average of 60 replicates of LOD scores across chromosomes 5 and 4. The two left panels display LOD scores for chromosome 5 and the two right panels show LOD scores from chromosome 4. Firstly, we comment on the top two panels, which correspond to the LOD scores based on comparing the likelihood of the polygenic model to the likelihood of the two-point model. Along chromosome 5, there is a distinct peak in the LOD scores around markers 14 and 15; however, along chromosome 4, there is no indication of the presence of a major gene. Because MG3 accounts for less than 10% of the total variation in Q1, the variance component approach fails to locate this locus. On the other hand, MG1 accounts for approximately 20%

of the total variation in Q1, so the LOD scores reach a maximum of about 4.2 at marker 15.

Secondly, we comment on the bottom two panels of Figure 6.2, which show the LOD score curves for chromosome 5 and 4 which compare the likelihood of the two-point model to the likelihood of Amos' model. The inclusion of a parameter for the recombination fraction does not seem to significantly improve the model since all of the LOD scores fall below 0.2. As we saw in section 5.2.2, the LOD scores seem to be approximately 0 near any region of a major gene which contributes to at least 20% of the total phenotypic variation. Although it is not distinct, we see that there are two dips in the plot of the LOD scores from chromosome 5, which gives a subtle indication of the presence of a QTL. Because the markers which are close to the QTL have recombination fractions close to 0, Amos' model does not improve upon the two-point model as greatly as it does for markers which are far from the QTL.

6.3 Estimation of Variance Components

Estimating the variance components of a complex trait can be quite challenging, especially if the components are small. In this section, we compare the estimates of the variance-components from the two-point model and Amos' model. We denote the QTL heritability and polygenic heritability to be h_q^2 and h_G^2, respectively. Figure 6.3 shows the average estimated QTL and polygenic heritabilities from the marker data on chromosome 5. Note that the two reference lines in the three left panels correspond to the heritability due to MG1 and the heritability due to MG1 and MG1HMG2, as given in Table 6.1. These reference lines correspond to the heritability due to MG1 alone and the total heritability due to MG1. The reference line in the three right panels correspond to the polygenic heritability. Also, we averaged the female and male heritability components since there is a gender effect. In general, we see that the QTL heritability component is underestimated and the polygenic heritability component is overestimated. The top two panels in Figure 6.3 show the estimated QTL heritabilities and estimated polygenic heritabilities from the two-point model. The bottom four panels show the estimated QTL and polygenic heritabilities from Amos' model. Recall that Amos' model expresses part of the covariance as a fraction of the additive genetic component, and that this fraction is a function of not only the proportion of genes ibd at the marker, but also the

recombination fraction. Furthermore, the form of this fraction, $f(\theta, \pi_{ij})$, is dependent on the relationship between family members. Because we are dealing with extended pedigrees, the types of relative pairs in the GAW10 data exceed those in Table 4.1. Table 6.3 presents the frequencies of the types of relative pairs in one replicate of the GAW10 data. From this table, we can see that the number of higher degree relative pairs is not insignificant. To handle the higher degree relative pairs, we make one of two assumptions. One possibility is to assume that they provide no linkage information, so $\theta = 1/2$. This conservative assumption causes $f(\theta, \pi_{ij})$ to have no dependence on the recombination fraction, which reduces to the expected proportion of genes shared ibd. Under this assumption, the two-point model is no longer nested within Amos' model. Alternatively, we also fit Amos' model under the assumption that high degree relatives exhibit tight linkage, so $\theta = 0$. In this case, $f(\theta, \pi_{ij}) = \pi_{ij}$. In Figure 6.3, the middle two panels show the estimates of the heritability components when tight linkage information is assumed from higher degree relatives, and so the bottom two panels assume loose linkage. When loose linkage is assumed, the QTL heritability estimates seem to be slightly larger and the polygenic heritability estimates seem to be slightly smaller than the estimates under the assumption of tight linkage.

In comparison with chromosome 5, Figure 6.4 shows the average estimated heritability components from chromosome 4. MG3 is located on chromosome 4 between markers 14 and 15, and only accounts for less than 10% of the total phenotypic variance. Because MG3 has a distinctly different effect on Q1, which depends on gender, we plot reference lines for the heritability component for males and females separately (as shown in Table 6.1). Whether we use the two-point model or Amos' model, the estimated heritability components are relatively similar. The average estimated QTL heritability seems to be less than 0.07 and the average estimated polygenic heritability seems to be about 0.40. It is interesting to note that the polygenic component is severely overestimated. However, roughly flat curves indicate that there is no major gene on this chromosome with a large effect. Because we do not simultaneously account for multiple QTLs, the variation due to other major loci seems to be absorbed by the polygenic component. Therefore, it is apparent that if we would like to accurately estimate the polygenic component in a complex trait, it is necessary to account for multiple QTLs via multipoint linkage analysis (Almasy et al. [1]). If the QTL components are assumed to be

Table 6.3: Frequencies of Relative Pairs in GAW10 Data

Relative Pair Type	Frequency
Unrelated	34954
Self	1497
Parent-offspring	2076
Siblings	1096
Grandparent-grandchild	2038
Avuncular	2723
Half-siblings	30
Great grandparent-grandchild	1394
Grand avuncular	1496
Half avuncular	68
1st cousins	2969
Great great grandparent-grandchild	266
Great grand avuncular	241
Half grand avuncular	13
1st cousins, 1 rem	3086
Half 1st cousins	27
1st cousins, 2 rem	423
2nd cousins	169

additive, then Amos' model can be easily extended.

6.4 Estimation of Recombination Fraction

Although we do not know the mapping function from genetic distances to recombination fractions, we can visually assess the ability of Amos' model to estimate recombination fractions. Figure 6.5 shows the average estimated recombination fractions along chromosomes 5 and 4. The top two panels correspond to chromosome 5 and the bottom two panels correspond to chromosome 4. Because we are dealing with extended families, for higher degree relative pairs, we either assumed that they exhibited tight linkage or loose linkage. In Figure 6.5, the left panels are the estimated recombination fractions when tight linkage was assumed for high degree relatives, and the right panels are the estimated recombination fractions when loose linkage was assumed. The two different assumptions made on the higher degree relatives does not seem to have a substantial impact on the estimation of the recombination fraction. It is apparent that marker information from both chromosome 5 and 4 are exhibiting signs of linkage to a QTL

since all estimated recombination fractions lie well below 1/2. Because chromosome 5 houses the major gene which accounts for about 20% of the variation in Q1, we see that the estimated recombinations become smaller at those marker locations closer to MG1. The estimated recombination fractions from chromosome 4 gives us a sense that a major gene is present, but since there is no distinct minimum value so we cannot tell where MG3 resides.

6.5 Summary

In this chapter, we evaluated the performance of Amos' robust approach to linkage analysis to detect major loci and estimate the variance components of a complex trait, which was simulated for GAW10. We see that QTLs which account for less than 10% of the total phenotypic variation are not detected; however, QTLs which account for at least one-fifth of the total variation are readily detected via LOD score curves. As well, QTL variance-components tend to be underestimated and polygenic variance-components tend to be overestimated on average. Furthermore, the polygenic variance component seems to absorb the variance components due to QTLs which are not accounted for in the model. So, multipoint linkage analysis (Almasy et al. [1]) should be used to obtain a reasonable estimate of the polygenic component.

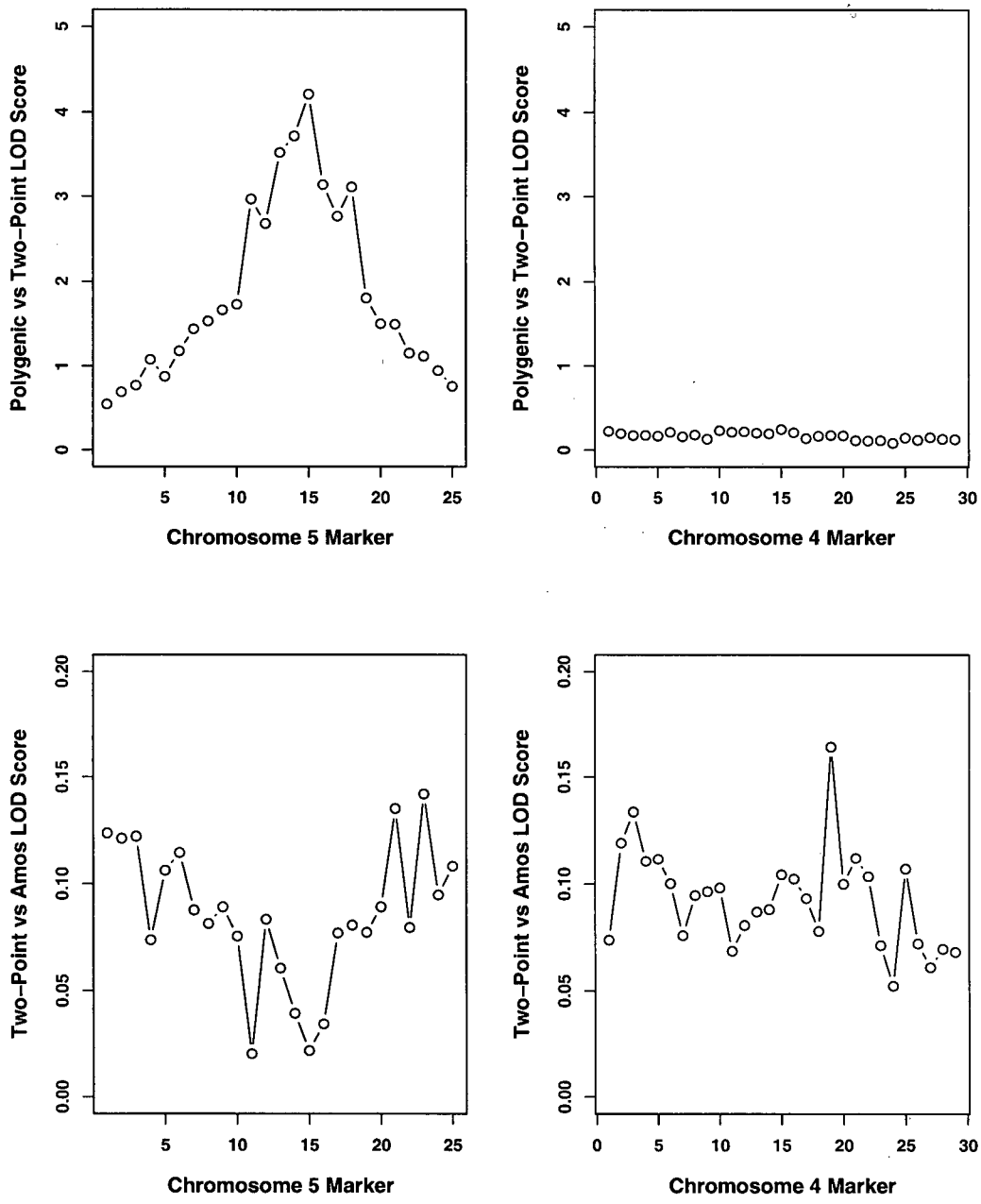

Figure 6.2: LOD Score Curves for Complex Trait Q1

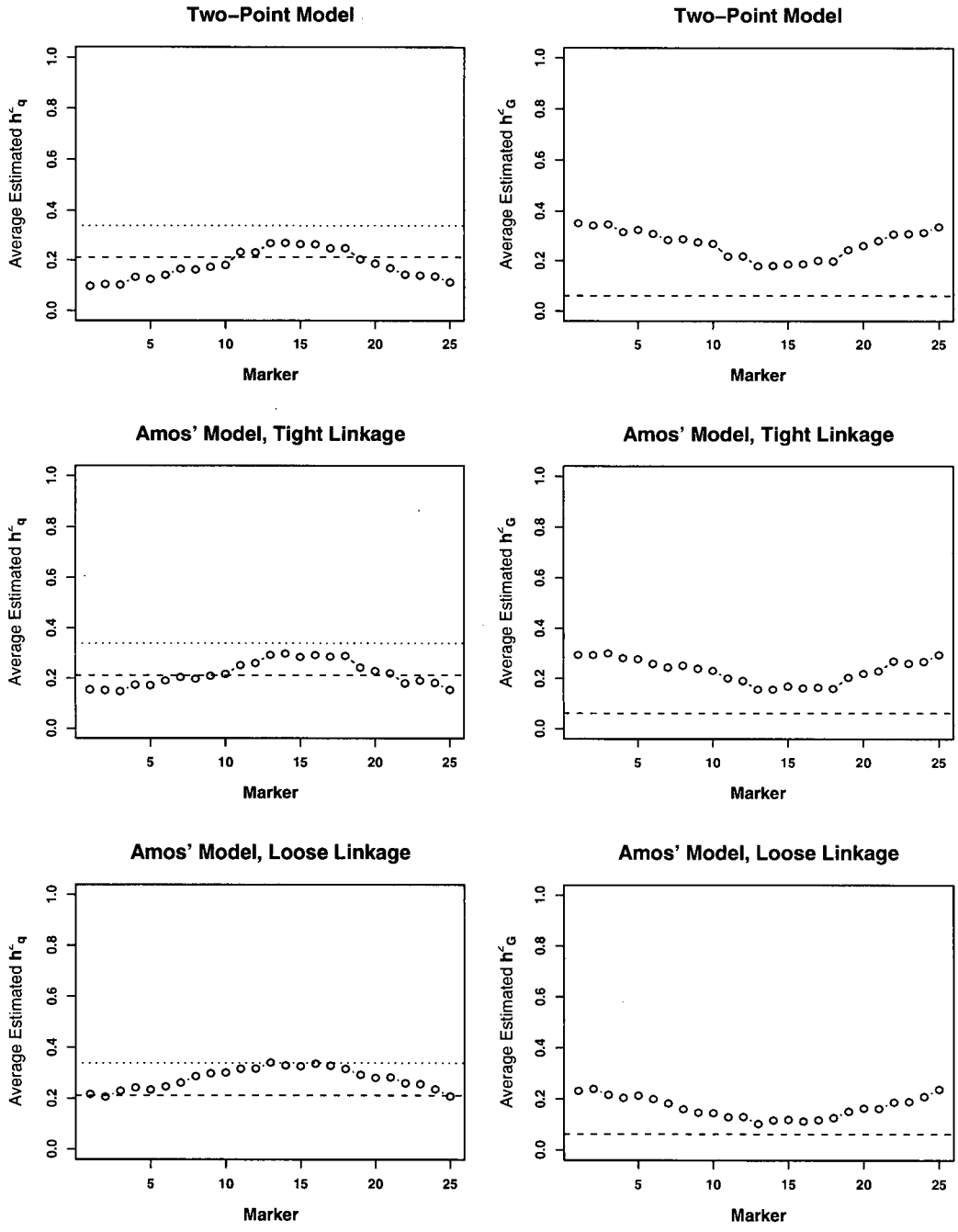

Figure 6.3: Average Estimated Heritabilities for Q1 along Chromosome 5

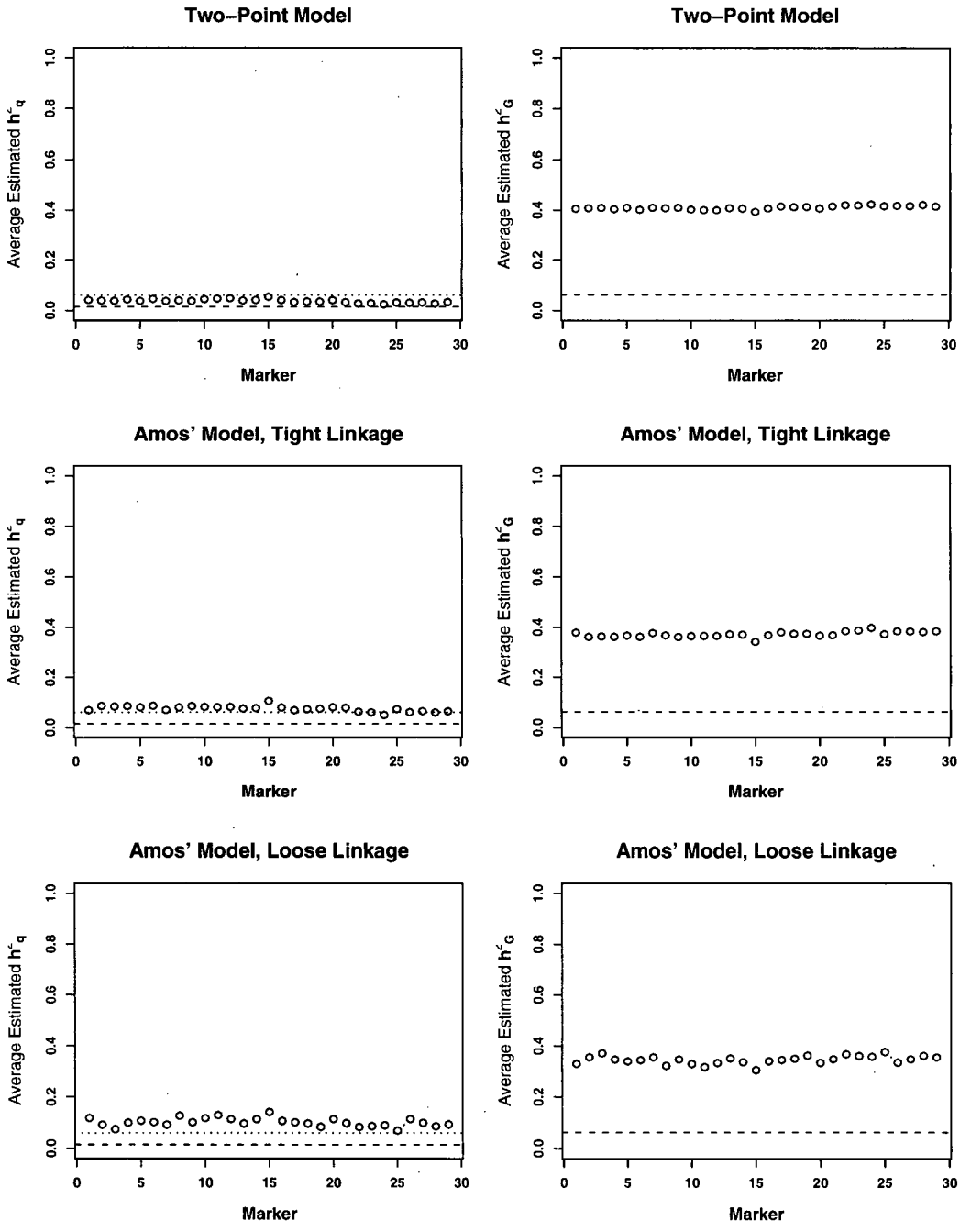

Figure 6.4: Average Estimated Heritabilities for Q1 along Chromosome 4

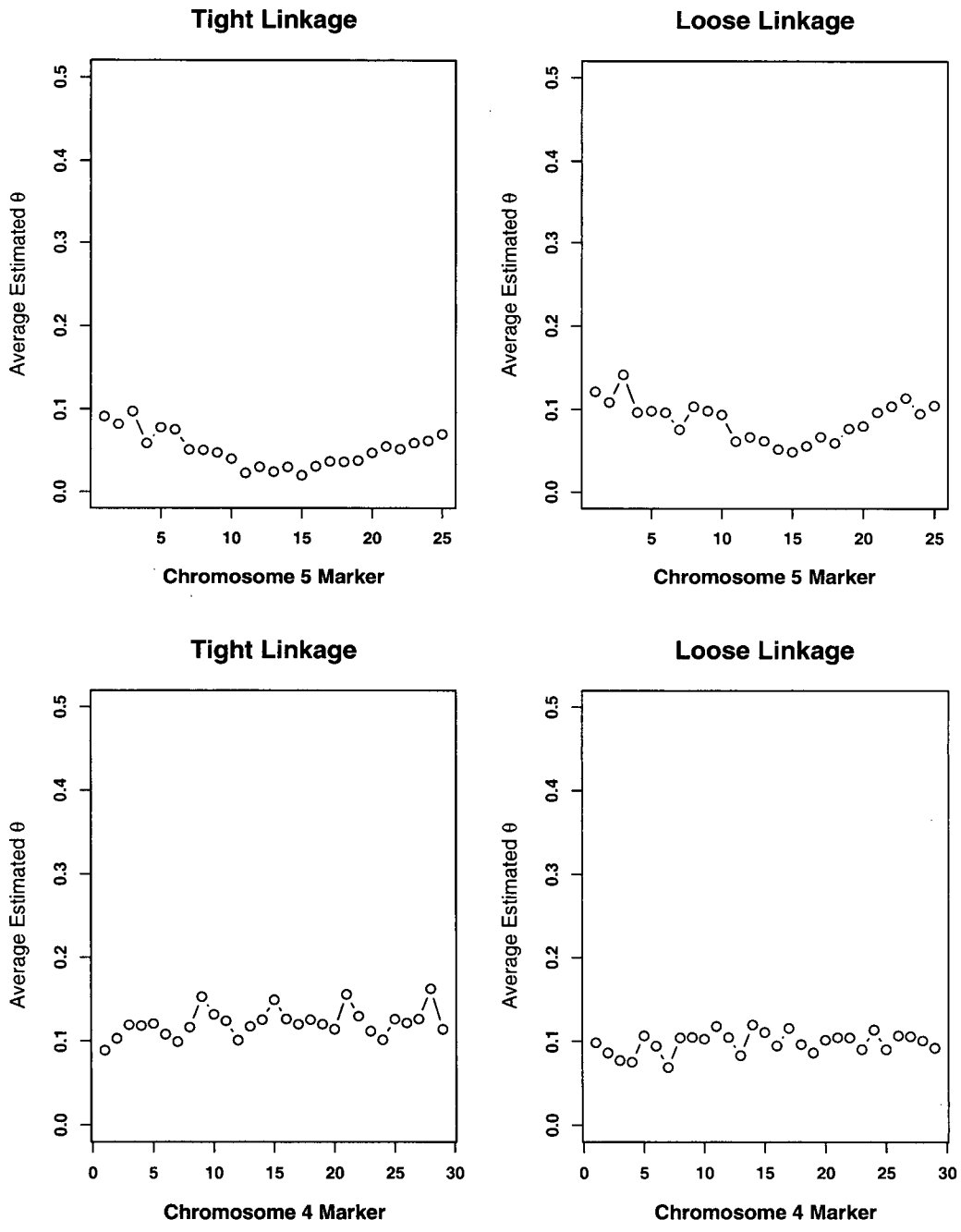

Figure 6.5: Average Estimated Recombination Fractions along Chromosomes 5 and 4

Chapter 7

Discussion

We investigated the performance of Amos' [3] variance-components model in its ability to detect major genes and partition the total variation of a complex trait into its genetic and non-genetic components. The consequences of neglecting the varying distances between the markers along a chromosome and a potential QTL by assuming tight linkage were also investigated by assessing a sub-model of Amos' model, which we referred to as the two-point model. In this chapter, we highlight some of the observations from our investigation, and discuss some potential future work.

Firstly, in Chapters 5 and 6, we saw that the estimated variance-components from the two-point model tended to be biased. On average, the QTL heritability tended to be underestimated and the polygenic heritability tended to be overestimated. We also saw that the magnitude of the error in the estimates increased as the distance between the marker and QTL increased. Because the two-point model incorrectly assumes tight linkage at markers which are far from the QTL, we suspect that the degree of model misspecification is associated with the severity of the error in the estimated heritabilities. Although Amos' full-model accounts for the varying distances between markers and a QTL, the estimated heritabilities still appeared to be biased, and in most cases overestimated. In mixed effects models, it is known that maximum likelihood estimators result in biased estimates of variance components. Therefore, we may want to use other forms of estimation, like those based on estimating equations (Amos et al. [4], Prentice et al. [23], Nelder et al. [20]), to determine whether the accuracy of the estimated heritabilities improves.

Secondly, it is apparent that there is confounding between the recombination fraction and the variance components. In Chapter 5, we saw that if we assume tight linkage, the estimated heritabilities from the two-point model at different markers exhibit a distinct trend across the chromosome. However, when we simultaneously estimate the recombination fraction and the heritability components of the quantitative trait, we see that the trend in the estimated heritabilities is removed. It seems that less accurate recombination fraction estimates correspond to less accurate estimated heritabilities. When dealing with extended pedigrees, we either assumed that higher degree relative pairs exhibited tight linkage or loose linkage. It would be interesting to see how the estimated recombination fractions and heritabilities are affected if we eliminate these assumptions. This would require determining the exact form of the fraction of additive variance due to the QTL for higher relative pairs as a function of the recombination fraction and proportion of alleles shared ibd.

Thirdly, in Chapter 6 we saw that modelling only one QTL effect for a complex trait does not provide accurate estimates for all variance components. Because a complex trait may have more than one QTL, by only accounting for one major gene, the variance attributable to the remaining major genes becomes absorbed into the polygenic variance component. To decrease the amount of over-estimation in the polygenic heritability, we may want to investigate whether multipoint linkage analysis (Almasy et al. [1], and Amos et al. [4]) improves the accuracy of the estimated variance components.

Finally, the ability of the variance-components approach to detect QTLs strongly depends on the proportion of variation it contributes to the trait of interest. Through our simulations in Chapter 5, we saw that the two-point model can readily map QTLs which contribute more than 40% of the total phenotypic variation. When analyzing a complex trait in Chapter 6, we were able to detect a QTL accounting for approximately 20% of the variation in one of the quantitative risk factors. However, the two-point model was unable to locate a QTL which contributed less than 10% of the variation. In addition, Amos' full model did not appear to improve upon the two-point model's detection ability. We point out that whether the trait was simple or complex, this model-free variance-components approach was able to detect QTLs contributing at least 20% of the total variation without specifying genetic details, such as the mode of inheritance, allele frequencies, and penetrances. Because the success of the variance-

components approach to linkage analysis does not seem to depend on all of the intricate genetic details underlying a quantitative trait, it does indeed appear to be robust and alleviate the need for segregation analysis. On the other hand, because the variance-components approach fails to detect QTLs with a modest effect on a quantitative trait, linkage analysis should still be used together with association analysis techniques (see Sham [26], for example).

Bibliography

[1] Almasy, L., and Blangero J. (1998). Multipoint Quantitative Trait Linkage Analysis in General Pedigrees. *American Journal of Human Genetics,* **62**, 1198-1211.

[2] Amos, C. I., and Elston, R. C. (1989). Robust Methods for the Detection of Genetic Linkage for Quantitative Data from Pedigrees. *Genetic Epidemiology,* **6**, 349-360.

[3] Amos, C. I. (1994). Robust Variance-Components Approach for Assessing Genetic Linkage in Pedigrees. *American Journal of Human Genetics,* **54**, 535-543.

[4] Amos, C.I., and De Andrade, M. (2001). Genetic Linkage Methods for Quantitative Traits. *Statistical Methods in Medical Research,* **10**, 3-25.

[5] Blangero, J., Lange, K., Almasy L., Williams, J., Dyer, T., Peterson, C. (December 2001). *SOLAR: Sequential Oligogenic Linkage Analysis Routines.* Southwest Foundation for Biomedical Research. Retrieved September 2002, from http://www.sfbr.org/sfbr/public/software/solar.

[6] Drigalenko, E. (1998). How Sib Pairs Reveal Linkage. *American Journal of Human Genetics,* **63**, 1242-1245.

[7] Edlin, G. (1990). *Human Genetics.* Jones and Bartlett, Boston.

[8] Elston, R. C., Buxbaum, S., Jacobs, K. B., and Olson, J. M. (2000). Haseman and Elston Revisited. *Genetic Epidemiology,* **19**, 1-17.

[9] Falconer, D. S. (1989). *Introduction to Quantitative Genetics, 3E.* Longman Sci. and Tech., Harlow, UK.

[10] Fisher, R. A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. Royal Soc. Edinburgh,* **52**, 399-433.

[11] *Genometric Analysis Simulation Program (G.A.S.P.).* (June 2002). National Human Genome Research Institute. Retrieved September 2002, from http://research.nhgri.nih.gov/gasp/.

[12] *Genetic Analysis Workshop.* (October 2002). Southwest Foundation for Biomedical Research. Retrieved September 2002, from http://www.sfbr.org/external/gaw/.

[13] Haldane, J. B. S. (1919). The Combination of Linkage Values, and the Calculation of Distance between the Loci of Linked Factors. *Journal of Genetics,* **8**, 299-309.

[14] Haseman, J. K. (1970). *The Genetic Analysis of Quantitative Traits using Twin and Sib Data.* Ph.D. thesis. University of North Carolina.

[15] Haseman, J. K., and Elston, R. C. (1972). The Investigation of Linkage Between a Quantitative Trait and a Marker Locus. *Behavior Genetics,* **2**, 3-19.

[16] Lander, R. S., and Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits using RFLP Linkage Maps. *Genetics,* **121**, 185-199.

[17] Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Massachusetts.

[18] MacCluer, J. W., Blangero, J., Dyer, T. D., and Speer, M. C. (1997). GAW10: Simulated Family Data for a Common Oligogenic Disease with Quantitative Risk Factors. *Genetic Epidemiology,* **14**, 737-742.

[19] Nash, J.C. (1979). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* Adam Hilger Ltd., Bristol.

[20] Nelder, J.A., and Pregibon, D. (1987). An Extended Quasi-likelihood Function. *Biometrika,* **74**, 221-232.

[21] Olson, J.M., and Wijsman, E.M. (1993). Linkage between Quantitative Trait and Marker Loci: Methods Using All Relative Pairs. *Genetic Epidemiology,* **10**, 87-102.

[22] Ott, J. (1999). *Analysis of Human Genetic Linkage, 3E.* Johns Hopkins University Press, Baltimore.

[23] Prentice, R.L., and Zhao, L.P. (1991). Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. *Biometrics,* **47**, 825-839.

[24] Risch, N., and Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science,* **273**, 1516-1517.

[25] Searle, S. R. (1982). *Matrix Algebra useful for Statistics.* John Wiley and Sons, New York.

[26] Sham, P. (1998). *Statistics in Human Genetics.* Arnold, London.

[27] Wijsman, E. M. and Amos, C. I. (1997). Genetic Analysis of Simulated Oligogenic Traits in Nuclear and Extended Pedigrees: Summary of GAW10 Contributions. *Genetic Epidemiology,* **14**, 719-735.

[28] Wilson, A. F., Bailey-Wilson, J. E., Pugh, E. W., Sorant, A. J. M. The Genometric Analysis Simulation Program (G.A.S.P.): a Software Tool for Testing and Investigating Methods in Statistical Genetics. *American Journal of Human Genetics,* **59**, A193.

[29] Wright, F. A. (1997). The Phenotypic Difference Discards Sib-Pair QTL Linkage Information. *American Journal of Human Genetics,* **60**, 740-742.

Appendix A

A.1 Joint Distribution of π_{mj} and π_{qj}

Haseman and Elston [15] derived the joint distribution of the proportion of alleles shared ibd at a marker locus, π_{mj}, and the proportion of alleles shared ibd at a QTL, π_{qj}, for the jth sib-pair. Their derivation of this distribution is illustrated below.

Let the genotype at the marker locus for the mother and father be B_1B_2 and B_3B_4, respectively. Also, let the genotype at the QTL for the mother and father be C_1C_2 and C_3C_4, respectively. We let the alleles with odd subscripts be parents' maternal alleles and alleles with even subscripts be parents' paternal alleles. Therefore, mating between these two parents may be expressed as follows:

$$\frac{B_1C_1}{B_2C_2} \times \frac{B_3C_3}{B_4C_4}.$$

Let θ be the recombination fraction between the marker locus and the QTL. Then the possible gametes, along with their frequencies, from each parent are shown in Table A.1. A recombination occurs with probability θ, and therefore no recombination occurs with probability $1-\theta$. As well, the frequencies of the two possible non-recombinant gametes are equal for each parent, and likewise for the frequencies of the two possible recombinant gametes. Note that if the two loci are not linked (ie. $\theta = 1/2$), then the frequencies of the gametes from each parent are $1/4$. In this case, all gametes are equally likely to occur since recombination between two unlinked loci occurs about one half of the time. Alternatively, if two loci are tightly linked (ie. $\theta = 0$), then recombination does not occur between the two loci. Furthermore, recombinant gametes are not formed, and so they have a frequency of 0. Because an offspring independently

Table A.1: Possible Gametes from Two Loci and their Frequencies

Parent	Gamete	Frequency
Mother	B_1C_1	$(1-\theta)/2$
	B_2C_2	$(1-\theta)/2$
	B_1C_2	$\theta/2$
	B_2C_1	$\theta/2$
Father	B_3C_3	$(1-\theta)/2$
	B_4C_4	$(1-\theta)/2$
	B_3C_4	$\theta/2$
	B_4C_3	$\theta/2$

receives one gamete from each parent, there are 16 possible combinations of gametes, which are called zygotes.

To compute the probability that a sib-pair shares all of the alleles ibd at the marker and QTL, we simply sum the probabilities over all such instances. For example, the proportion of alleles shared ibd at a marker and a QTL will be 1 if both siblings have genotypes $B_1C_1B_3C_3$. The probability that this occurs is $[(1-\theta)/2]^2[(1-\theta)/2]^2 = (1-\theta)^4/16$. The summation over all 16 cases where the sib have identical genotypes gives the probability that a sib-pair shares all of their alleles ibd at both the marker and QTL:

$$\begin{aligned}
\Pr(\pi_{mj}=1,\pi_{qj}=1) &= 4[\theta^4/16]+8[\theta^2(1-\theta)^2/16]+4[(1-\theta)^4/16] \\
&= [\theta^4+2\theta^2(1-\theta)^2+(1-\theta)^4]/4 \\
&= [\theta^2+(1-\theta)^2]^2/4 \\
&= \Psi^2/4
\end{aligned}$$

where $\Psi^2 = \theta^2+(1-\theta)^2$.

The remaining joint probabilities in Table 3.1 may be computed in a similar manner. We may also exploit symmetry and knowledge of the marginal distributions to derive these probabilities. For further details please see Haseman and Elston [15].

A.2 Joint Distribution of π_{mj} and π_j

In their paper, Haseman and Elston [15] derive the joint distribution of the proportion of alleles shared ibd at a marker, π_{mj}, and the estimated proportion of alleles shared ibd at a marker,

Table A.2: Mating Types and Sib-Pair Types at a Locus with Two Alleles

Mating Type	Sib-Pair Type	Probability	f_0	f_1	f_2	π_j
$BB-BB$	$BB-BB$	p^4	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
$BB-bb$	$Bb-Bb$	$2p^2r^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
$BB-Bb$	$BB-BB$	p^3r	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$
	$BB-Bb$	$2p^3r$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$
	$Bb-Bb$	p^3r	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$
$Bb-Bb$	$BB-BB$	$p^2r^2/4$	0	0	1	1
	$bb-bb$	$p^2r^2/4$	0	0	1	1
	$BB-bb$	$p^2r^2/2$	1	0	0	0
	$BB-Bb$	p^2r^2	0	1	0	$\frac{1}{2}$
	$bb-Bb$	p^2r^2	0	1	0	$\frac{1}{2}$
	$Bb-Bb$	p^2r^2	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$
$bb-bB$	$bb-bb$	r^3p	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$
	$bb-Bb$	$2r^3p$	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{4}$
	$Bb-Bb$	r^3p	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{4}$
$bb-bb$	$bb-bb$	r^4	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$

π_j, for the jth sib-pair. In particular, they consider the case where there are two alleles at the marker locus. As well, it is assumed that we have complete parental information and no dominance.

We illustrate some of the details in the derivation of this joint distribution. Let the marker locus have two alleles, B and b, with allele frequencies p and r, respectively. Then the possible mating types and sib-pairs are shown in Table A.2. Table A.2 also shows the probability of the occurrence of each mating type and sib-pair type. Conditional on the parental and sibling marker data, the probability of sharing i alleles ibd at the marker is given by f_i. Therefore, $\pi_j = f_1/2 + f_2$ is the estimated proportion of alleles shared ibd at the marker. Note that the results in this table may be generalized to a multiallele marker. Please see Haseman and Elston [15] for further details.

Using Table A.2, we can derive the joint distribution of π_j and π_{mj}. For example, $\pi_j = 0$ if and only if the mating type $Bb-Bb$ results in a sib-pair of $BB-bb$. Note that it is not possible for this sib-pair to share any alleles ibd. Therefore, $\Pr(\pi_j = 0, \pi_{mj} = 1/2) = \Pr(\pi_j = 0, \pi_{mj} = 1) = 0$ and $\Pr(\pi_j = 0, \pi_{mj} = 0) = p^2r^2/2$. Similarly, $\pi_j = 1/4$ if the mating type $BB - Bb$ (or $bb-bB$) results in a sib-pair of $BB-Bb$ (or $bb-bB$). Note that these sib-pairs cannot share two alleles ibd, so $\Pr(\pi_j = 1/4, \pi_{mj} = 1) = 0$. However, they can share 0 or 1 alleles ibd and each

case is equally likely to occur, so $\Pr(\pi_j = 1/4, \pi_{mj} = 0) = \Pr(\pi_j = 1/4, \pi_{mj} = 1/2) = p^3r + pr^3$. The remainder of Table 3.2 may be derived in a similar manner.