Analysis of quantitative traits: segregation and conditional linkage

By

J Concepción Loredo-Osti

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
at
Dalhousie University
Halifax, Nova Scotia
August 1999

0-612-49278-8

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Analysis of quantitative traits: Segregation and conditional linkage"

by J. Concepcion Loredo-Osti

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 13, 1999

External Examiner [redacted]

Research Supervisor

Examining Committee

Dalhousie University

Date: **August 1999**

Author: **J Concepción Loredo-Osti**

Title: **Analysis of quantitative traits: segregation and conditional linkage**

Department: **Mathematics and Statistics**

Degree: **Ph.D.** Convocation: **October** Year: **1999**

Signature of Author

To Carmen, Pallas Athena, and Asdrúbal Ígor.

Contents

Acknowledgement

I would like to express my gratitude to Prof. Bruce R. Smith. His advice and guidance were fundamental in the development of this work.

My studies in Halifax would not have been possible without the support of Dalhousie University, the Canadian International Development Agency, the University of Prince Edward Island (especially Dr. Alfonso López), and La Universidad Autónoma de Tamaulipas y El Pograma para el Mejoramiento del Profesorado.

Analysis of quantitative traits: segregation and conditional linkage

J Concepción Loredo-Osti
Doctor of Philosophy
Department of Mathematics and Statistics
Dalhousie University
1999

Abstract

This thesis addresses the problem of assessing segregation of a quantitative trait under a mixed model of inheritance in non experimental populations, when the trait is linked to some monitored marker. There are two main approaches to the linkage problem in non experimental populations, the difference among them being the kind of information about the model of inheritance which is used in making inferences. Sib-pair methods are based only on identity by descent, i.e., they use partial information about the model, and their inference is conditioned on the identity by descent status of the markers. The alternative approach uses the full model information and, traditionally, the inference for these methods is done through the unconditional analysis of cosegregation, usually requiring Monte Carlo methods to evaluate the likelihood. This thesis presents an extension to one of the techniques (the Morton-MacLean integral) to compute the 'exact' likelihood under mixed models of inheritance when the data is at the bottom of the pedigree, regardless the complexity of such pedigree. With the same arguments used to derive this extension, a Monte Carlo sampling scheme in blocks of sibships is proposed. The idea of making linkage analysis conditional on the marker inheritance vector is also explored, together with some issues related to the computation of the conditional likelihood.

Chapter 1

Introduction

The identification of chromosome regions containing specific genes responsible for significant variation of a quantitative trait is one of the current problems of interest in genetic analysis. One effective way of addressing this problem is by linkage methods, which have been successfully used in the study of qualitative traits. For quantitative traits, a degree of success has been attained using linkage analysis with experimental crosses. Obviously, to approach the problem by using inbred lines in not always feasible, often, but not always, due to the nature of the species. Furthermore, it is often of interest to study the segregation of quantitative traits within non-inbred populations.

When experimental crosses are not possible or are not available, the problem is far more difficult. Several methods for the detection of linkage to quantitative traits in humans have been proposed; the Haseman-Elston sib-pair regression method, its variants and extensions [21, 32], and different flavors of variance components analysis and regression [37, 72] with maximum likelihood estimation [4, 25, 76, 88, 90]. Most of these methods are based on the presumption of unrelated nuclear families or zero loop pedigrees, a presumption that may be not appropriate for many 'natural' populations. To overcome this inconvenience some modifications have been proposed. However, the derived methods are much less efficient than the original methods or those methods in use with experimental crosses, and in many cases, the modifications require a large set of implicit assumptions which may damage the inference.

Exclusion is in some sense the dual problem to linkage. The goal of exclusion analysis is the identification of regions of the genome where the putative gene or genes are unlikely to be located. While exclusion analysis has long been established for qualitative traits [67], it is only recently under development for quantitative traits. Traditionally, exclusion analysis for quantitative traits has been just an appendix to linkage analysis, and there are few works to study the adequacy of linkage methods to assess exclusion.

The likelihood approach to linkage analysis began in the 1930's, beginning with the classical works of Haldane [30] and Fisher [19, 20]. Developed in the context of large size samples, the original likelihood methods were abandoned in the fifties mainly because of concerns about applicability to small samples and difficulties with computing the scores for matings of known phase. However, the basic statistical framework remains. Of course, many contributions have been made since then. Likelihood estimation of the recombinant fraction in man was introduced in two papers by Haldane and Smith [31] and Smith [79]. Haldane and Smith were the first to use the LOD (logarithm of odds) in the analysis of linkage, an approach that was further developed by Morton [60], who brought the whole machinery of sequential analysis to linkage. These works envisioned how inferences could be drawn.

Sib-pair analysis also began in the 1930's, pioneered by Penrose [71, 72]. His essential contribution was to provide a single generation data method, efficient in terms of information per unit of resource, to declare linkage. This approach did not become popular because, among some other inconveniences, it neglects information in the parents and other relatives, and does not give an estimate of recombinant fraction with acceptable precision. However, sib-pair methods have been revived after the work of

Haseman and Elston [32] and proved to be successful in complex inheritance problems [61].

At the beginning of the 1970's, Elston and Stewart [17] developed a recursive algorithm for the computation of probabilities on zero-loop pedigrees and discussed its application to the problem of inferring genetic models. Years later, the method was generalized, at least in theory, to pedigrees of arbitrary size and complexity [9, 52]. However, the implementation of that generalization when the model of inheritance considered is not monogenic is still hard to achieve, even for pedigrees of moderate size. That is a strong limitation of the Elston-Stewart algorithm. Morton and MacLean [62] proposed an alternative way to evaluate the segregation probabilities with mixed models of inheritance for nuclear families which has some computational advantages over the Elston and Stewart formulation. Also, as an alternative to the Elston and Stewart algorithm, Lander and Green [49] introduced inheritance vectors and hidden Markov model theory as a device for inference in linkage analysis of qualitative traits. This approach is still a very active area of research [38, 47, 48]. The last decade has seen an explosion in the usage of computer intensive techniques, particularly Monte Carlo methods, in segregation and linkage analysis of quantitative traits for complex pedigrees [82, 83].

The outline of the remainder of this thesis is as follows. In section 1.1 the distribution of a quantitative phenotypic trait is derived under the assumption that the phenotype consists of contributions from a discrete quantitative trait locus (QTL), a continuous polygenic component, and an environmental component. The ultimate goal is to assess the presence of the QTL, and its proximity to one or more marker genes which are monitored together with the phenotype. Models of inheritance of the discrete and continuous components are discussed in section 1.2, leading to the formulation of the phenotypic density as a mixture distribution. In monogenic models, the Elston-Stewart algorithm allows for a simplification of the mixture structure, providing a form suitable for direct computation.

For mixed models of inheritance the Morton–McLean algorithm is discussed. This result is generalized in chapter four.

As mentioned, the principle goal of this thesis is the development of methods to assess the proximity of a QTL to a marker gene. Linkage, as represented by co–segregation of the QTL and marker, is discussed in section 1.3 together with the recombinant fraction ρ, which parameterizes the linkage distance between marker and QTL. With this parameterization the problem of assessing proximity of QTL to marker becomes a problem in parametric inference, and the rest of the thesis is focused on the development of methods for estimating ρ.

Traditional methods for linkage analysis and estimation of recombinant frequency have utilized controlled crosses of pure genetic lines, which is the subject of chapter 2. The object of such crosses has often been point and interval estimation of recombinant fraction, and is usually based on the LOD (log–of–odds) score. Potential problems with LOD based interval estimates have been discussed by several authors, the difficulties often being based on an incorrect specification of the asymptotic distribution of the LOD score. A conservative interval estimate of ρ is proposed here. Some commonly used linear model methods for controlled crosses are discussed, and a number of problems with the underlying distributional assumptions are pointed out. A statistically valid permutation based approach to testing is proposed. As the remainder of the thesis is not concerned with experimental crosses, the methods proposed in chapter two will not be considered in further detail. The most important point identified in the chapter is that in all of the popular methods for linkage analysis with experimental crosses, the inference is conditioned on marker information.

In non experimental populations, linkage between a quantitative trait and marker becomes a more challenging problem and is the subject of much current research. Chapter 3 surveys linkage methods based on the concept of gene identity by descent. While sib–pair methods are in wide use by the genetics community, their inefficiency is unquestioned, having been mentioned since Penrose's time [61]. In this chapter the Haseman-Elston regression is presented in brief, and a convincing formal argument to show one cause of its inefficiency is set down. Several other inconveniences associated with Haseman-Elston regression are noted, including the fact that the method is a detection only procedure. The variance components approach based on identity by descent is reviewed in section 3.3. Arguments are given to point out that a full likelihood analysis in the sib-pair framework is equally as complex as a full likelihood analysis in a model-based framework.

Modern methods for likelihood analysis of data collected on a pedigree are discussed in chapter 4. One principle tool is the Elston-Stewart algorithm, which decomposes complex likelihoods into tractable pieces. A generalization of this algorithm, known as peeling, is the basis of many current methods for likelihood analysis, and is discussed in the context of a particular example. Peeling algorithms are highly successful with models which incorporate only an oligogenic, or only a polygenic, component. However, for polygenic models, methods from linear algebra are highly efficient as well, and are preferred in many cases. A recursive matrix algorithm for the Cholesky factorization of the inverse covariance matrix of the polygenic effect is derived in section 4.2. The algorithm turns out to be equivalent to previously published methods for calculating the inverse covariance matrix.

In section 4.4 an extension to the Morton-MacLean algorithm for likelihood evaluation with mixed models of inheritance is developed for the case of a nuclear family at the bottom of an arbitrarily complex pedigree. This is the principle contribution of this thesis. The extension allows for likelihood estimation with arbitrary known pedigree, complete observation of putative trait and marker data on all offspring at the bottom of the pedigree, i.e. all individuals without progeny, full or partial observations on quantitative trait and marker data of their parents, and full or partial observations of markers in the remainder of the pedigree. This generalization of the Morton-MacLean algorithm provides a deterministic evaluation of the likelihood which is exact, apart from quadrature approximations to integrals. Specifications are also given for several stochastic methods of likelihood approximation, including gene dropping, the Gibbs sampler, and the Hastings-Metropolis method.

Much of the recent effort in evaluating joint likelihoods on pedigrees with observed marker and quantitative trait data is based on the analysis of their joint segregation. In these cases the computational requirements grow enormously when the markers are highly polymorphic. On the other hand, highly polymorphic markers will typically be essential for accurate assessment of recombinant fraction. An alternative approach, which has been successfully utilized in the study of qualitative traits, is to use inheritance vectors or segregation indicators (which identify the parental origin of marker alleles) in the place of the markers themselves. The methods are reviewed in chapter 5, and provide an enormous computational simplification for methods of likelihood evaluation in the present context.

In chapter 6 the extended Morton-MacLean algorithm is applied to the estimation of recombinant fraction in several small simulated datasets with relatively complex underlying pedigrees. Gene drop methods and the Gibbs sampler are also applied to these data in an attempt to assess their performance in relation to the method of choice – the extended Morton-MacLean algorithm. While the estimation of recombinant fraction is a problem of major import to the identification of quantitative trait loci, most of the examples of such estimates in the statistical genetics literature are restricted to evaluation of likelihood ratios in the vicinity of the object of interest itself, the unknown value of ρ. Therefore, one contribution of this thesis is the aggregate collection of algorithms and methods which underly the likelihood evaluation and approximations for these simulations.

The thesis concludes with a discussion of results and suggestions for further work.

1.1 The problem

The analysis of quantitative trait data with a major gene involved is usually done under the assumption that the value of the quantitative trait for one individual, the ith say, can be written as

$$Y_i = \mu + \zeta_i + \eta_i + e_i \tag{1.1}$$

where μ is an unknown constant, ζ_i is the effect of a single major gene for which segregation with a marker is being monitored, η_i is the residual polygenic effect, and e_i, the environmental effects. These three random variables are assumed to have null expectation. If gametic phase equilibrium holds, the variance of Y_i is

$$\mathrm{Var}(Y_i) = \sigma_\zeta^2 + \sigma_\eta^2 + \sigma_e^2 \tag{1.2}$$

To know the covariance between a pair of individuals, it is necessary to introduce the concept of identity by descent.

Two alleles are said to be identical by descent, IBD, if they are copies coming from the same ancestor allele. Two individuals can have zero, one, or both alleles IBD at some particular locus. For example, let us consider the genotypic configurations from the mating type $Q_1Q_2 \times Q_3Q_4$. There are four possible offspring genotypes, each with an equal frequency. The genotype array for these is

$$\frac{1}{4}Q_1Q_3 \;:\; \frac{1}{4}Q_1Q_4 \;:\; \frac{1}{4}Q_2Q_3 \;:\; \frac{1}{4}Q_2Q_4$$

Denote by π_{ij} the proportion of alleles IBD shared by the individuals i and j. Whenever we observe the pair of sibs (Q_1Q_3, Q_1Q_3), we draw the conclusion that both individuals have received exactly the same alleles from their parents, so π_{ij} is 1; technically, the pair behaves as twins for this locus. Likewise, π_{ij} is $\frac{1}{2}$ for the pair (Q_1Q_3, Q_1Q_4) since they have one allele, Q_1, coming form the same parental allele and π_{ij} is zero for (Q_1Q_3, Q_2Q_4), with similar arguments applying to the remaining seven pairs. So, conditioned on π_{ij} at the putative quantitative locus, the covariance between the individual i and j is

$$\mathrm{Cov}(Y_i, Y_j \mid \pi_{ij}) = \sigma_a^2 \pi_{ij} + \sigma_d^2 \mathrm{I}_{[\pi_{ij}=1]} + 2\sigma_\eta^2 \psi_{ij} \tag{1.3}$$

where σ_a^2 and σ_d^2, the additive and dominance components of genetic variance are based in the decomposition $\sigma_\zeta^2 = \sigma_a^2 + \sigma_d^2$, and ψ_{ij} is Malécot's coefficient of *parenté* [39, 59] between i and j, defined as the probability that a gene selected randomly from i and a gene selected randomly from the same autosomal locus of j are identical by descent. In cases where a shared environment is modeled, it may also be necessary to split σ_e^2 as $\sigma_w^2 + \sigma^2$, where σ_w^2, the variance attributable to shared environment is added to the first term of (1.3) for appropriately related pairs. By definition of ψ_{ij}, it follows that

$$\mathrm{E}(\pi_{ij}) = 2\psi_{ij} \tag{1.4}$$

Knowledge of the pedigree is enough to determine ψ_{ij}, but not for π_{ij}, because in addition to the pedigree component, π_{ij} involves Mendelian sampling. A single observation of π_{ij} contains little information about ψ_{ij} and vice versa. One knows that $\psi_{ij} = 0$ implies $\pi_{ij} = 0$ and $\pi_{ij} \neq 0$ implies $\psi_{ij} \neq 0$, but no too much more can be said.

An interesting fact related to (1.3) is that

$$\mathrm{E}(Y_i \mid \pi_{ij}) = \mathrm{E}(Y_i) \qquad \text{and} \qquad \mathrm{Var}(Y_i \mid \pi_{ij}) = \mathrm{Var}(Y_i) \tag{1.5}$$

In fact, for any pair of individuals Y_i and Y_j with a proportion π_{ij} of shared alleles IBD at the putative trait, the marginal distributions of Y_i given π_{ij} and Y_j given π_{ij} are the same as the unconditional ones. To see this, it is enough to consider only the joint distribution of the genotypes at the putative trait for both individuals. Table 1.1 contains this joint conditional

q_i	q_j	π_{ij} 0	 $\frac{1}{2}$	 1
QQ	QQ	p^4	p^3	p^2
QQ	Qq	$2p^3(1-p)$	$p^2(1-p)$	0
QQ	qq	$p^2(1-p)^2$	0	0
Qq	QQ	$2p^2(1-p)$	$p^2(1-p)$	0
Qq	Qq	$4p^2(1-p)^2$	$p(1-p)$	$2p(1-p)$
Qq	qq	$2p(1-p)^3$	$p(1-p)^2$	0
qq	QQ	$p^2(1-p)^2$	0	0
qq	Qq	$2p(1-p)^3$	$p(1-p)^2$	0
qq	qq	$(1-p)^4$	$(1-p)^3$	$(1-p)^2$

Notice: p is the relative frequency of the allele Q.

Table 1.1: Joint distribution of the genotypes of two individuals at the putative locus given the IBD sharing proportion.

distribution for the model with two alleles at the quantitative trait locus.

If Y_i were normally distributed, the mean and variance would be enough to characterize the statistical problem. Sadly, it is unlikely that this will be the case, simply because ζ_i, the effect of a major gene, has a discrete distribution, usually with only a few points of support. For example, if there are n_q alleles for the quantitative trait in question, there are not more than $\frac{1}{2}n_q(n_q+1)$ different values for ζ_i. A standard assumption is that

$$\eta_i + e_i \sim \mathcal{N}\left(0, \sigma_\eta^2 + \sigma_e^2\right)$$

and consequently the distribution of Y_i is the mixture

$$f_{Y_i}(y) = \sum_{j=1}^{n_q} p_{ij}\, \varphi_{\sigma_\eta^2+\sigma_e^2}(y - \mu - \zeta_{q_j}) \tag{1.6}$$

where

$$p_{ij} = \Pr(\zeta_i = \zeta_{q_j})$$

and ζ_{q_j} is the genetic value associated with the q_j genotype.

We do not observe the genotypes q_j; so, the mixing proportions p_{ij} are unknown. Different ways of dealing with this problem have been proposed; most of them derive from the pioneering works of Elston and Stewart [17] and Morton and MacLean [62]. One special case arises when a quantitative trait has been monitored jointly with some marker gene. The central idea is that if the marker is close enough to the position of a major QTL in the chromosome, they should segregate together and one can make inferences about the p_{ij}s using the marker information. The linkage-exclusion problem is to determine whether or not a major QTL is associated with the marker. Moreover, the goal is to estimate both the size of the effect of the major QTL, and its distance from the marker.

1.2 Models of inheritance

There are a minimum of four ingredients of a genetic model for segregation analysis [15]. The first is the description of the type or types of genetic effects including the genetic basis of variation in the population; the second ingredient is the statistical model which relates those genetic effects to the trait values; the third describes the mechanism which explains how the variation is passed through generations; the last describes the way in which a group of individuals is sampled from the population for study.

The model of inheritance from one generation to the next can be summarized mathematically by the genotypic distribution of the offspring conditioned on the two parental genotypes. There are three fundamental models of inheritance for quantitative traits: oligogenic, polygenic, and a mixture of oligogenic and polygenic models. An oligogenic model assumes that the genetic component in a quantitative trait is explained in terms of a few loci, which means that the genetic component is a discrete random variable; monogenic models are a special case of oligogenic ones where the genetic variation is attributed to just one gene. No matter how many loci are involved in the model, the main characteristic is that the oligogenic component of the quantitative trait can be considered as a discrete random variable. The polygenic model, on the other hand, assumes that the genetic component for some measurable quantitative trait is the sum of effects of a vast number of genes whose individual contributions are very small. The polygenic assumption has the effect of shifting likelihood computation problems from the domain of combinatorics to the domain the linear algebra [26].

In general, the density of the observations y can be written as

$$f(y) = \sum_{g} f(y \mid g)\ \Pr(g) \tag{1.7}$$

$$= \int_{a} f(y \mid a)\ dF(a) \tag{1.8}$$

$$= \sum_{g} \int_{a} \Pr(g)\, f(y \mid g,\ a)\ dF(a \mid g) \tag{1.9}$$

where the latent variables **g** and **a** have some associated genetic meaning, for example, major gene genotypes and polygenic effect. Other expressions may be appropriate for some particular analyses. Which expression is the best for computations depends very much on the complexity of pedigree and the inheritance model assumed.

Modern segregation analysis starts with the works of Hilden [36] and Elston and Stewart [17]. For linkage analysis the second paper is particularly relevant since it provided the first systematic method for computing likelihoods using entire pedigrees. In the context of quantitative traits, the Elston-Stewart algorithm has had overwhelming success for oligogenic models, especially monogenic, on general pedigrees. Also, it shows acceptable performance for polygenic models but it is still computationally prohibitive for mixed models of inheritance [8, 69]. Some time later, Morton and MacLean [62] presented an algorithm which is a better choice for mixed models of inheritance on nuclear families. The difference between these algorithms is the decomposition of the joint probabilities used [7]. Let us have a closer look at these different representations on a nuclear family with parents m and f, n_k children, and phenotypes

$$\begin{aligned} \mathbf{y} &= (y_f,\ y_m,\ y_1,\ \dots,\ y_{n_k})' \\ &= (y_f,\ y_m,\ \mathbf{y}^{*\prime})' \end{aligned}$$

where $\mathbf{y}^* = (y_1, \dots, y_{n_k})'$.

Monogenic inheritance. Assume the model $\mathbf{y} = \boldsymbol{\zeta} + \mathbf{e}$ with $\mathbf{y} \mid \boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\zeta}, \sigma^2 \mathbf{I})$. The Elston-Stewart algorithm is based in the decomposition

$$\mathrm{f}(\mathbf{y}) = \sum_{\zeta_f} \sum_{\zeta_m} \cdots \sum_{\zeta_{n_k}} \Pr(\zeta_f, \zeta_m, \dots, \zeta_{n_k}) \prod_{j=f}^{n_k} \varphi_{\sigma^2}(y_j - \zeta_j) \tag{1.10}$$

$$= \sum_{\zeta_f} \sum_{\zeta_m} \cdots \sum_{\zeta_{n_k}} \Pr(\zeta_f, \zeta_m) \prod_{j=1}^{n_k} \Pr(\zeta_j \mid \zeta_f, \zeta_m)$$

$$\varphi_{\sigma^2}(y_f - \zeta_f)\ \varphi_{\sigma^2}(y_m - \zeta_m) \prod_{j=1}^{n_k} \varphi_{\sigma^2}(y_j - \zeta_j) \tag{1.11}$$

$$= \sum_{\zeta_f} \Pr(\zeta_f)\ \varphi_{\sigma^2}(y_f - \zeta_f) \sum_{\zeta_m} \Pr(\zeta_m \mid \zeta_f)\ \varphi_{\sigma^2}(y_m - \zeta_m)$$

$$\prod_{j=1}^{n_k} \sum_{\zeta_j} \Pr(\zeta_j \mid \zeta_f, \zeta_m)\ \varphi_{\sigma^2}(y_j - \zeta_j) \tag{1.12}$$

The calculations in (1.11) increase exponentially with the family size, while with (1.12) the increase is approximately linear.

The step from (1.11) to (1.12) is a special case of the following result. Suppose $f(\cdot,\cdot)$ is a bivariate density, then

$$\begin{aligned}\sum_{g_n}\cdots\sum_{g_1}\prod_{j=1}^{n} f_j(g_j, x_j) &= \sum_{g_n}\cdots\sum_{g_2}\prod_{j=2}^{n} f_j(g_j, x_j)\sum_{g_1} f_1(g_1, x_1)\\ &= f_{X_1}(x_1)\sum_{g_n}\cdots\sum_{g_3}\prod_{j=3}^{n} f_j(g_j, x_j)\sum_{g_2} f_2(g_2, x_2) \qquad (1.13)\\ &\;\vdots\\ &= \prod_{j=1}^{n} f_{X_j}(x_j)\end{aligned}$$

Note that the application of (1.13) to (1.11) requires the phenotypes of two sibs to be independent once their genotypes are given, a restriction that excludes the possibility of non-genetic covariation as, for example, shared environment.

Polygenic (additive) inheritance. Without loss of generality, assume a model $\mathbf{y} = \boldsymbol{\eta} + \mathbf{e}$ with $\boldsymbol{\eta}$ accounting for only additive genetic effects; $\boldsymbol{\eta}$ and $\mathbf{e}$ independent; $\mathbf{e} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $\boldsymbol{\eta} \sim \mathcal{N}\left(0, \sigma_\eta^2\mathbf{A}\right)$ where $\mathbf{A} \in \mathbb{R}^{n_k+2}$ and

$$\mathbf{A} = \tfrac{1}{2}\begin{pmatrix} 2 & 0 & \mathbb{1}' \\ 0 & 2 & \mathbb{1}' \\ \mathbb{1} & \mathbb{1} & (\mathbf{I} + \mathbb{1}\mathbb{1}') \end{pmatrix}$$

i.e., a family with unrelated, non-inbred parents ($\mathbb{1}$ is a vector of 1's). The decomposition on which the Elston-Stewart algorithm is based for this case, obtained similarly as for (1.12), is

$$\mathrm{f}(\mathbf{y}) = \int_{\eta_f} \varphi_{\sigma^2_\eta}(\eta_f)\ \varphi_{\sigma^2}(y_f - \eta_f) \int_{\eta_m} \varphi_{\sigma^2_\eta}(\eta_m)\ \varphi_{\sigma^2}(y_m - \eta_m) \tag{1.14}$$

$$\prod_{j=1}^{n_k} \int_{\eta_j} \varphi_{\frac{1}{2}\sigma^2_\eta}(\eta_j - \tfrac{1}{2}(\eta_f + \eta_m))\ \varphi_{\sigma^2}(y_j - \eta_j)\ \mathrm{d}\eta_j\, \mathrm{d}\eta_m \mathrm{d}\eta_f$$

As before, the calculations are linear in the family size, although in this case the integral can be explicitly evaluated.

Mixed (monogenic and polygenic-additive) inheritance. With the addition of a polygenic component to the oligogenic model, i.e, the model $\mathbf{y} = \boldsymbol{\zeta} + \boldsymbol{\eta} + \mathbf{e}$ with the same assumptions on η and $\mathbf{e}$ as in the polygenic model, the Elston-Stewart formulation for writing down the density in this case is

$$\mathrm{f}(\mathbf{y}) = \sum_{\zeta_f} \Pr(\zeta_f) \int_{\eta_f} \varphi_{\sigma^2_\eta}(\eta_f)\ \varphi_{\sigma^2}(y_f - \zeta_f - \eta_f) \tag{1.15}$$

$$\sum_{\zeta_m} \Pr(\zeta_m) \int_{\eta_m} \varphi_{\sigma^2_\eta}(\eta_m)\ \varphi_{\sigma^2}(y_m - \zeta_m - \eta_m)$$

$$\prod_{j=1}^{n_k} \sum_{\zeta_j} \Pr(\zeta_j \,|\, \zeta_f, \zeta_m) \int_{\eta_j} \varphi_{\frac{1}{2}\sigma^2_\eta}(\eta_j - \tfrac{1}{2}(\eta_f + \eta_m))$$

$$\varphi_{\sigma^2}(y_j - \zeta_j - \eta_j)\ \mathrm{d}\eta_j\, \mathrm{d}\eta_m \mathrm{d}\eta_f$$

The expression (1.15) is similar to (1.11) but cannot be written in the form (1.12) due to the integral over η_m, η_f. Because of this, the Elston-Stewart algorithm is impractical even for relatively small pedigrees [7, 69].

In order to deal with this problem, a different decomposition of the density is needed. As starting point, one can think of something like

$$\begin{aligned} \mathrm{f}(\mathbf{y}) &= \mathrm{f}(y_f, y_m)\ \mathrm{f}(\mathbf{y}^* \mid y_f, y_m) \\ &= \mathrm{f}(y_f, y_m) \int_a \mathrm{f}(\mathbf{y}^* \mid y_f, y_m, a)\ \mathrm{dF}(a \mid y_f, y_m) \end{aligned} \tag{1.16}$$

The idea is to pick an a such that the computations can be carried out in an efficient way.

Define $\zeta^* = (\zeta_1, \ldots, \zeta_{n_k})'$, then, in the case of unrelated and non-inbred parents,

$$\mathrm{E}(\mathbf{y}^* \mid \zeta, y_f, y_m) = \zeta^* + \tfrac{1}{2}h^2(y_f - \zeta_f + y_m - \zeta_m)\mathbb{1} \tag{1.17}$$

$$\mathrm{Var}(\mathbf{y}^* \mid \zeta, y_f, y_m) = \tfrac{1}{2}\sigma_\eta^2\left(1 - h^2\right)\mathbb{1}\mathbb{1}' + \left(\tfrac{1}{2}\sigma_\eta^2 + \sigma^2\right)\mathbf{I} \tag{1.18}$$

where $h^2 = \frac{\sigma_\eta^2}{\sigma_\eta^2+\sigma^2}$. Because the joint conditional distribution of $\mathbf{y}$ given ζ is normal, the distribution of $\mathbf{y}^*$ given ζ, y_f, and y_m is also normal and, in the presence of (1.18), it is invariant under permutation as well, i.e., conditioned on ζ, y_f, and y_m, $\mathbf{y}^*$ is a set of exchangeable random variables [10]. These properties allow us to choose a so that

$$\mathrm{f}(\mathbf{y}^* \mid \zeta, y_f, y_m, a) = \prod_{j=1}^{n_k} \mathrm{f}(y_j \mid \zeta, y_f, y_m, a) \tag{1.19}$$

and this expression together with

$$\Pr(\zeta^* \mid \zeta_f, \zeta_m) = \prod_{j=1}^{n_k} \Pr(\zeta_j \mid \zeta_f, \zeta_m) \tag{1.20}$$

can be worked through (1.13) to obtain a more manageable expression than (1.15). While F in (1.16) is a normal distribution, different choices for its mean are possible, subject to the constraints imposed by (1.16) and (1.17). A reasonable choice is to pick a distribution for which subsequent computations are simplified.

It is known that a random vector $\mathbf{x}$ of normal variables with covariance $c_0\mathbf{I}+c_1\mathbb{1}\mathbb{1}'$ can be written as $\mathbf{x}=\mathbf{x}_0+X_1\mathbb{1}$, where $\mathbf{x}_0$ is a vector of independent normal variables, all of them having variance c_0, and X_1 a normal random variable, independent of $\mathbf{x}_0$, for which variance and covariances with the elements of $\mathbf{x}_0$ are all equal to c_1. With this in mind, a quick inspection of (1.18) reveals that the variance of a must be $\frac{1}{2}\sigma_\eta^2(1-\hbar^2)$. In fact, the computations are simplified if a is taken as

$$a \sim \mathcal{N}\left(\tfrac{\hbar^2}{2}(y_f-\zeta_f+y_m-\zeta_m),\ \tfrac{1}{2}\sigma_\eta^2\left(1-\hbar^2\right)\right)$$

This choice also provides some free genetic interpretation for a: it is the parental contribution to the polygenic component, the breeding value, of the children's trait given the phenotypes and oligogenotypes of their parents.

After the application of (1.13), (1.16) becomes

$$\begin{aligned}
\mathrm{f}(\mathbf{y}) = & \sum_{\zeta_f} \Pr(\zeta_f)\ \varphi_{\sigma_\eta^2+\sigma^2}(y_f-\zeta_f) \\
& \sum_{\zeta_m} \Pr(\zeta_m)\ \varphi_{\sigma_\eta^2+\sigma^2}(y_m-\zeta_m) \\
& \int_a \varphi_{\frac{1}{2}\sigma_\eta^2(1-\hbar^2)}\left(a-\tfrac{\hbar^2}{2}(y_f-\zeta_f+y_m-\zeta_m)\right) \\
& \prod_{j=1}^{n_k} \sum_{\zeta_j} \Pr(\zeta_j \mid \zeta_f, \zeta_m)\ \ \varphi_{\frac{1}{2}\sigma_\eta^2+\sigma^2}(y_j-\zeta_j-a)\ \mathrm{d}a
\end{aligned} \tag{1.21}$$

Morton and MacLean [62] derived an expression similar to (1.21), but based on a stronger set of assumptions and genetic arguments than those used here. Because of that, this equation is referred as the Morton-MacLean algorithm. In most non-trivial situations there is no analytical solution to the integral in (1.21). It must be integrated numerically, but is of the form suitable for accurate evaluation using quadrature. In spite of its required numerical integration, the Morton–MacLean algorithm is still economically acceptable because it is linear in the number of family members.

Actually, Morton and MacLean's model also included an environmental component common to all the members of the same sibship. This is equivalent to specifying the covariance between $(y_f, y_m)'$ and $\mathbf{y}^*$ to remain as before, while requiring that $\sigma_w^2 \mathbb{1}\mathbb{1}'$ be added to $\mathrm{Var}(\mathbf{y}^*)$ in order to account for the variance of the environment common to sibs, i.e., it can be written

$$\mathrm{Var}(\mathbf{y}^*) = \tfrac{1}{2}\sigma_\eta^2(\mathbf{I} + \mathbb{1}\mathbb{1}') + \sigma_w^2 \mathbb{1}\mathbb{1}' + \sigma^2 \mathbf{I} \tag{1.22}$$

Despite this consideration, the elements of $\mathbf{y}^*$ remain equicorrelated. The model is again derivable from the exchangeability argument. Taking (1.22) into account, $\sigma_w^2 \mathbb{1}\mathbb{1}'$ must be added to (1.18) to include the common environment. With this addition, the variance of a has to be $\frac{1}{2}\sigma_\eta^2(1 - \hbar^2) + \sigma_w^2$ in order to meet (1.19), i.e., the natural choice for a must be

$$a \sim \mathcal{N}\left(\tfrac{\hbar^2}{2}(y_f - \zeta_f + y_m - \zeta_m), \tfrac{1}{2}\sigma_\eta^2\left(1 - \hbar^2\right) + \sigma_w^2\right)$$

Therefore, $\varphi_{\frac{1}{2}\sigma_\eta^2(1-\hbar^2)}(\cdot)$ must be replaced by $\varphi_{\frac{1}{2}\sigma_\eta^2(1-\hbar^2)+\sigma_w^2}(\cdot)$ in (1.21), to get the same expression as in Morton and MacLean's original paper. Using this same principle, as part of this work, an extension to the Morton–MacLean algorithm which allows the parents to be inbred and related will be presented in chapter 4.

Incidentally, this has touched on another crucial point, that being the common environment for the sibship. In a polygenic model of inheritance this effect can be accounted for without additional computational complication, while in the oligogenic model the inclusion of this effect drastically changes the computational strategy because the conditional independence needed to move from (1.11) to (1.12) does not hold.

1.3 Linkage

Linkage is the phenomenon by which two genes on the same chromosome stick, or segregate, together at meiosis. Two genes are unlinked if they are 'sampled' independently through a pedigree path. Loci on non-homologous chromosomes segregate independently during the meiosis process; so, genes on these loci are unlinked. In contrast, a pair of genes on the same chromosome tend to remain together during the formation of gametes. The physically closer the loci lie, the higher the chance that the genes remain coupled and segregate as a unit; the more distant, the more independent they become. In general, linkage is one of the most important forces in retarding the rate of decay of gametic phase disequilibrium from generation to generation [14].

Crossing-over is the phenomenon that disrupts linkage. At meiosis, each member of a pair of homologous chromosomes replicates to form two sister chromosomes called chromatids. These chromatids align perfectly to form a bundle of four chromatids, after which crossing-over may occur at points known as chiasmata. At each chiasma, one sister chromatid from each pair may be randomly chosen and cut at a cross-over point. The cell rejoins the partial paternal and maternal chromatids, exchanging the genetic material beyond the cut point so that two hybrid chromatids are formed. In absence of chromatidal interference, two chromatids are chosen independently, and after crossing-over and two binary divisions, the recombinant chromatids go to the four gametes.

If one thinks of both chiasma formation and crossing over as processes occurring on a fixed interval of the real line, the number and positions of the chiasmata along the chromosome bundle can be modeled as a point process [12]. Without chromatidal interference, each crossover process is created from the chiasma process by random thinning of chiasmata. This random thinning determines whether or not the gamete participates in the underlying cross-over at each chiasma point; because of the symmetry of the process, both choices have the same probability [50].

If the haplotype of an individual contains two alleles coming from the same grandparent, recombination at the parental level has not occurred. An offspring is termed recombinant or non-recombinant depending on whether or not the offspring indicates that a recombination has occurred in one of the parents [67]. For two unlinked genes on the same chromosome, the same proportion of recombinants and non-recombinants is expected. The recombinant fraction ρ between two loci at positions a and b is connected to $N_{[a,b]}$, the number of chiasmata occurring on the interval $[a, b]$, by Mather's formula

$$\rho_{[a,b]} = \tfrac{1}{2}\Pr(N_{[a,b]} > 0) \tag{1.23}$$

To prove Mather's formula define r_n as

$$r_n = \Pr(\text{the gamete is recombinant}|N_{[a,b]} = n)$$

so that

$$\rho_{[a,b]} = \sum_{n=0}^{\infty} r_n \Pr(N_{[a,b]} = n)$$

Clearly $r_0 = 0$ and for $n > 0$

$$r_n = \tfrac{1}{2}r_{n-1} + \tfrac{1}{2}(1 - r_{n-1})$$

since the gamete is recombinant if it is after $n-1$ cross-overs and does not participate in the nth cross-over, or if it is not recombinant after $n-1$ cross-overs and participates in the nth cross-over. The solution to the previous equation is $r_n = \frac{1}{2}$ for all $n > 0$, which completes the proof.

The genetic map distance $\gamma_{[a,b]}$ is defined in terms of the expected number of chiasmata on $[a, b]$ per gamete as

$$\gamma_{[a,b]} = \tfrac{1}{2}\,\mathrm{E}(N_{[a,b]}) \tag{1.24}$$

The term map function is used indistinctly to denote ρ, the recombinant fraction, expressed as a function of γ, the map distance, in Morgans and the opposite, to represent γ in terms of ρ. Under Haldane's model of independence between numbers of chiasmata falling on disjoint intervals, the chiasma process is an homogeneous Poisson process; so, its map function is

$$\rho = \tfrac{1}{2}\,(1 - \mathrm{e}^{-2\gamma}) \tag{1.25}$$

with inverse

$$\gamma = -\tfrac{1}{2}\log(1 - 2\rho) \qquad 0 \le \rho \le \tfrac{1}{2} \tag{1.26}$$

Since most of our understanding of the chiasma process is merely phenomenological, many map functions have been proposed to account for chromatidal interference, finite numbers of chiasmata, etc. For example, Karlin's model assumes at most N crossovers, independently distributed on the interval $[a, b]$, and following a binomial distribution; its map function is

$$\rho = \tfrac{1}{2}\left(1 - (1 - 2\gamma/\mathrm{N})^{1/\mathrm{N}}\right) \tag{1.27}$$

with inverse

$$\gamma = \tfrac{1}{2}\mathrm{N}\left(1 - (1 - 2\rho)^{1/\mathrm{N}}\right) \tag{1.28}$$

Kosambi's model [46], intended to deal with chromatidal interference, postulates that the chiasma process is determined by a stationary renewal model with map function

$$\rho = \tfrac{1}{2}\tanh(2\gamma) \qquad \text{and} \qquad \gamma = \tfrac{1}{2}\tanh^{-1}(2\rho) \tag{1.29}$$

It seems clear now why in addition to the 'size' of the effect of the trait, the inference about linkage between a quantitative trait and a marker depends on the distance between loci, expressed in genetic map distance units or as recombinant fraction.

Naïvely, constructing a genetic linkage map is just a matter of counting recombinants and non recombinants. While this may be approximately correct whenever geneticists have the ability of arranging crosses to avoid or resolve potential ambiguities, with non-experimental populations it may be impossible simply to 'count recombinants' in a cross, due to the lack of information to identify unambiguously where recombinant events have occurred [49]. That is the goal of linkage analysis.

Chapter 2

Linkage in controlled crosses

This thesis is concerned with methods of analysis for non-experimental populations. Nevertheless, a quick review of some of the standard methods for linkage analysis appropriate to controlled crosses of inbred lines is an appropriate way to commence a study of linkage analysis in non experimental populations. This chapter points out some of the problems arising in linkage analysis with controlled crosses. A key point identified in the chapter is that that the currently most popular approaches to linkage analysis for controlled crosses all condition on marker information.

Traditionally, the studies of linkage between a quantitative trait and marker in controlled crosses have been carried out through the observation of recombinants within families. The principle is very simple: if there exists association between marker type and trait value, it is likely that the major trait locus is close to the marker locus.

The simplest approach uses the markers as classification variables for analysis of variance, regression, t-tests, or some such analysis. Another approach is to construct the likelihood using the joint segregation of the quantitative trait gene and the marker [56, 57]. To appreciate the basic ideas, consider a two-allele model with co-dominant marker, recombinant fraction ρ, and parental genotypes MQ/MQ and mq/mq as in the figure 2.1. All of the individuals from the F_1 generation have the same genotype, and they

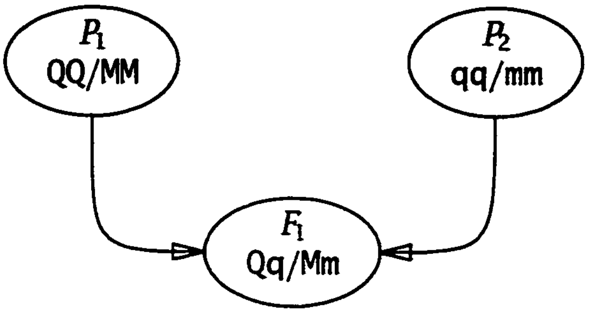

Figure 2.1: An example to illustrate linkage analysis with controlled crosses and known phase

produce gametes with frequencies

$$\frac{1-\rho}{2}\,\mathsf{MQ} \;:\; \frac{1-\rho}{2}\,\mathsf{mq} \;:\; \frac{\rho}{2}\,\mathsf{Mq} \;:\; \frac{\rho}{2}\,\mathsf{mQ}$$

If F_1 is back-crossed with the parental genotype P_1, only the genotypes QQ and Qq are obtained on the quantitative trait locus under scrutiny. Similarly, the back-cross on P_2 produces only qq and Qq on the major quantitative trait locus. The joint segregation for B_1, the back-cross on the parent P_1, can be obtained from table 2.1. For B_2, the back-cross on the other parental genotype, the same table can be used after replacing the homozygotes QQ and MM by qq and mm, respectively.

$\Pr_\rho(q \mid m)$		q		
		QQ	Qq	$f_M(m)$
m	MM	$(1-\rho)$	ρ	$\frac{1}{2}$
	Mm	ρ	$(1-\rho)$	$\frac{1}{2}$

Table 2.1: B_1 segregation proportions for a quantitative trait locus conditional on a linked codominant marker.

The joint segregation for F_2, the product of randomly crossing the F_1 individuals among themselves, can be obtained from the table 2.2

Assuming normality, if the quantitative trait expected values given the

$\Pr_\rho(q \mid m)$		q QQ	 Qq	 qq	$f_M(m)$
	MM	$(1-\rho)^2$	$2\rho(1-\rho)$	ρ^2	$\frac{1}{4}$
m	Mm	$\rho(1-\rho)$	$(1-\rho)^2+\rho^2$	$\rho(1-\rho)$	$\frac{1}{2}$
	mm	ρ^2	$2\rho(1-\rho)$	$(1-\rho)^2$	$\frac{1}{4}$

Table 2.2: F_2 segregation proportions for a quantitative trait locus conditional on a linked codominant marker.

putative quantitative locus genotypes QQ, Qq, and qq are μ_{QQ}, μ_{Qq}, and μ_{qq}, respectively, and if the variance for all of the marker-QTL classes is the same, the likelihood for the ith observation, Y_i, conditioned on the marker information is

$$\ell_i = \sum_j \Pr_\rho(q_j \mid m_i)\ \varphi_{\sigma^2}(Y_i - \mu_{q_j}) \tag{2.1}$$

and the full likelihood, ℓ, is the product of the individual likelihoods.

The hypothesis of no linkage can be tested with the likelihood ratio statistic

$$\lambda = \frac{\max\limits_{\rho=0.5} \ell}{\max \ell} \tag{2.2}$$

If the regularity conditions were met, the likelihood ratio test statistic under the hypothesis of no linkage would have asymptotically a $\chi^2_{(1)}$ distribution. However, the fact that in the reduced model some parameters are fixed in the boundary of the parameter space (for example, if some the mixing proportions in (2.1) are not strictly positive, some μ_q's are meaningless) and the requirement of absence of chromatidal interference [43], i.e., the assumption that $0 \le \rho \le \frac{1}{2}$, tell us that regularity conditions are not met. Despite these considerations it is not uncommon to assume that $-2\log\lambda \xrightarrow{\mathcal{L}} \chi^2_{(1)}$ under the hypothesis of no linkage. Another practice is to replace the likelihood ratio statistic (2.2) by a scaled version, the LOD score and use some bound for the distribution of this statistic. The lod-score is defined as

$$\operatorname{lod} r = \log_{10} \frac{\max\limits_{\rho=r} \ell}{\max\limits_{\rho=0.5} \ell} \tag{2.3}$$

and its maximum

$$\mathsf{LOD} = \max_{r} \operatorname{lod} r \tag{2.4}$$

$$= -\log_{10} \lambda \tag{2.5}$$

Large positive values of lod provide evidence of linkage, and negative values support its exclusion. Originally proposed in the context of sequential testing, the lod is nowadays applied almost invariably in a non-sequential framework. Currently, the lod-score criterion in human genetics studies is to declare linkage if $\mathsf{LOD} > 3$ and exclude the possibility of linkage between the marker and some quantitative trait locus whenever the recombinant fraction ρ lies in $\{\rho : \operatorname{lod} \rho < -2\}$. The basis for these criteria is given by Markov's inequality. Under the hypothesis of non linkage $\mathrm{E}\,(\lambda^{-1}) = 1$ which implies

$$\Pr_{\rho}(\mathsf{LOD} > \log_{10} c \mid \rho = \tfrac{1}{2}) < c^{-1} \tag{2.6}$$

regardless the distribution of LOD. Then, declaring linkage when $\mathsf{LOD} > 3$ is equivalent to saying that the p-value for the hypothesis of non-linkage is smaller that 0.001.

There exists an alternative definition for the lod-score as

$$\operatorname{lod}^* r = \log_{10} \frac{\max\limits_{\rho=r} \ell}{\max\limits_{\text{no QTL}} \ell} \tag{2.7}$$

where 'no QTL' means 'there is not a quantitative trait locus' or equivalently that 'the trait expectations, given the quantitative trait genotypes, are equal'. Clearly, if there is no segregating quantitative trait, nothing can be linked to it; so, the parameter ρ is meaningless. In this sense, 'no QTL' implies 'no linkage'. This variant of the lod-score is more powerful for declaring linkage when the rule LOD* > 3 is used regardless of any other consideration, as follows from the fact that

$$\max_{\rho=0.5} \ell \geq \max_{\text{no QTL}} \ell \tag{2.8}$$

It is interesting to note that the relationship between the rule LOD > 3 and the likelihood test with λ as the test statistic is quite different from the relationship between the rule LOD* > 3 and its corresponding likelihood ratio test, where $\text{LOD}^* = \max_r \text{lod}^* r$. Remember that LOD is just a scaled version of λ which implies the existence of an unique p-value for the likelihood ratio test such that both methods lead to the same conclusion, since under the mentioned regularity conditions λ always has an asymptotic $\chi^2_{(1)}$ distribution. The asymptotic distribution of LOD* under the same regularity conditions changes with the number of nuisance parameters in the likelihood. For example, for B_1 it is $\chi^2_{(2)}$ while for F_2 it is $\chi^2_{(3)}$. So, to generate the same conclusion, the p-value of the likelihood ratio test has to be modified as the asymptotic distribution of its test statistic changes.

The computation of the so-called support intervals is another important scenario where both variants of the lod-score can be applied. Many ways to assess the accuracy of the estimates of ρ have been suggested. One of these is through confidence regions, confidence intervals preferentially. From a statistical viewpoint, a sensible approach is to invert the acceptance region of the likelihood ratio test for the hypothesis $\rho = r$ versus the two sided alternative $\rho \neq r$. For the hypothesis in question, the likelihood ratio statistic can be written as

$$\lambda(r) = \frac{\max_{\rho=r} \ell}{\max \ell} \tag{2.9}$$

The figure (2.2) shows an artificial example of how the likelihood intervals may look. With present computational resources, inverting the acceptance region determined by the distribution of (2.9) does not represent a difficult numerical task. However, in spite of their known good statistical properties, likelihood intervals are rarely used; the so-called LOD-one-down intervals being much more popular.

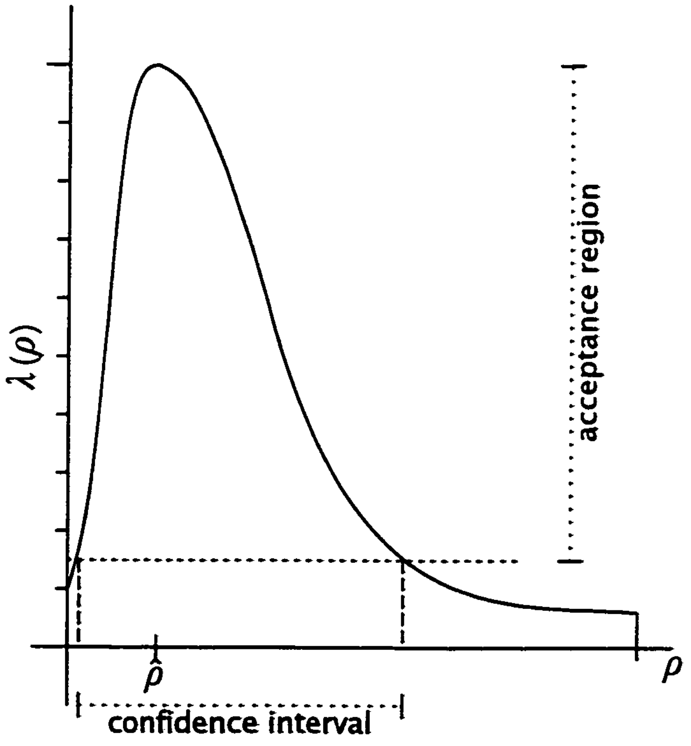

Figure 2.2: Confidence interval obtained by inversion of the acceptance region of the likelihood ratio test

Figure (2.3) illustrates how the LOD-one-down support intervals are obtained. After plotting lod ρ and drawing a line at LOD $-$ 1, the support interval is obtained by projecting over the ρ-axis the two points where the graph of $\operatorname{lod}\rho$ is intersected by that line. Narrower support intervals may be attained by using lod^* instead of lod.

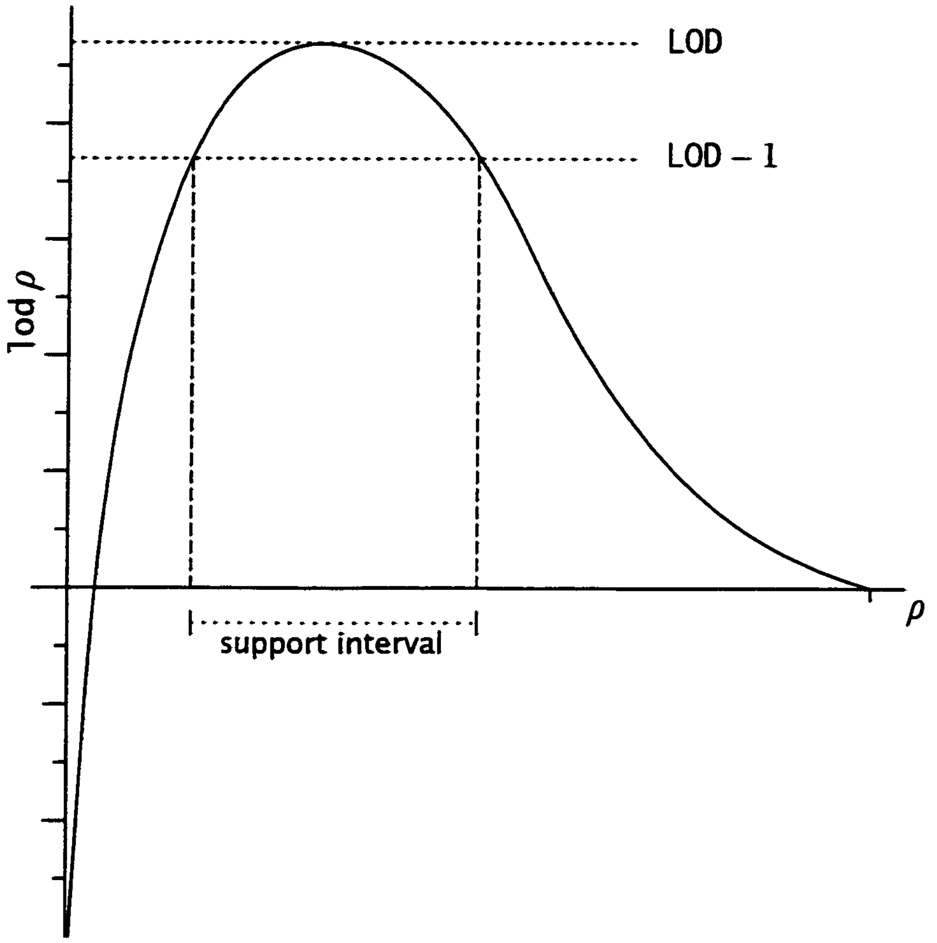

Figure 2.3: Typical LOD-one-down support interval

The LOD-one-down support intervals may be numerically convenient, but are by no means equivalent to likelihood based intervals. Moreover, the determination of the coverage of the LOD-one-down intervals may be more difficult than inverting the likelihood ratio acceptance region because

$$\operatorname{lod} r = \log_{10}\lambda(r) - \log_{10}\lambda \tag{2.10}$$

i.e., lod r is a linear combination of two non-independent random variables with different marginal distributions, and only one of them can be asymptotically proportional to the distribution $\chi^2_{(1)}$ as the hypotheses underlying $\lambda(r)$ and λ cannot be simultaneously true. While the likelihood intervals are coherent and parsimonious, there is no guarantee that LOD-one-down support intervals are going to yield logically consistent regions no matter which hypothesis is accepted as true or how high the LOD has to be before the intervals are computed. Examples of those logical inconsistencies are mentioned by Ott [67].

Nevertheless, given support intervals, it is feasible to compute at least a bound of the probability coverage, i.e., the probability that the support interval covers the true recombinant fraction, through Bonferroni's inequality. To do so, just observe that the intersection between any pair of upper and lower bounds for ρ is a confidence interval for the parameter. In the likelihood framework, the confidence bounds come from inverting the acceptance regions of two one-side tests. Bonferroni's inequality guarantees that the intersection of those bounds yields a confidence region whose confidence coefficient is not less than $1 - p_1 - p_2$, where p_1 and p_2 are the p-values for the one-sided tests. Since the extremes of a support interval can be taken as the values of upper and lower confidence bounds, there are two one-side likelihood ratio tests related to those bounds with known p-values that can be used to obtain a reasonable approximation for the coverage probability.

The likelihood approach is not the only one, and is certainly not the most popular among experimental cross scientists. In typical studies of linkage on controlled crosses, individuals are separated into marker classes and those classes are compared. Given some value m for the marker in any individual, the expected value of the putative trait is

$$\begin{aligned}\mu_m &= \mathrm{E}(Y \mid m)\\ &= \sum_j \mathrm{E}(Y \mid q_j, m)\ \Pr_p(q_j \mid m)\\ &= \sum_j \mathrm{E}(Y \mid q_j)\ \Pr_p(q_j \mid m)\\ &= \sum_j \mu_{q_j}\ \Pr_p(q_j \mid m) \qquad (2.11)\end{aligned}$$

since given q_j, the distribution of the quantitative trait does not depend on the marker. That is the expectation for the marker class m. Also, if the variance given q_j is the same for all j, say σ^2, then

$$\begin{aligned}\sigma_m^2 &= \mathrm{E}\left((Y-\mu_m)^2 \mid m\right)\\ &= \sum_j \left(\mathrm{Var}(Y \mid q_j) + (\mu_{q_j} - \mu_m)^2\right)\ \Pr_p(q_j \mid m)\\ &= \sigma^2 + \sum_j (\mu_{q_j} - \mu_m)^2\ \Pr_p(q_j \mid m) \qquad (2.12)\end{aligned}$$

For the B_1 design, the means for the marker classes are

$$\mu_{\mathrm{MM}} = (1-\rho)\,\mu_{\mathrm{QQ}} + \rho\,\mu_{\mathrm{Qq}}$$

$$\mu_{\mathrm{Mm}} = \rho\,\mu_{\mathrm{QQ}} + (1-\rho)\,\mu_{\mathrm{Qq}}$$

and the variances are

$$\begin{aligned}\sigma_{\mathrm{MM}}^2 &= \sigma^2 + \rho(1-\rho)\,(\mu_{\mathrm{QQ}} - \mu_{\mathrm{Qq}})^2\\ &= \sigma_{\mathrm{Mm}}^2\end{aligned}$$

The difference between marker class means is

$$\mu_{\mathrm{MM}} - \mu_{\mathrm{Mm}} = (1 - 2p)\,(\mu_{\mathrm{QQ}} - \mu_{\mathrm{Qq}}) \tag{2.13}$$

which differs from zero only when $p < \frac{1}{2}$ and $\mu_{\mathrm{QQ}} \neq \mu_{\mathrm{Qq}}$, a fact that can be used to test for linkage through the t-statistic

$$t = \frac{\bar{y}_{\mathrm{MM}} - \bar{y}_{\mathrm{Mm}}}{s\sqrt{\left(\frac{1}{n_{\mathrm{MM}}} + \frac{1}{n_{\mathrm{Mm}}}\right)}} \tag{2.14}$$

where the $\bar{y}$'s are marker class averages, and s^2 is the pooled variance within marker classes [56, 57]. The likelihood ratio statistic for the B_1 design is

$$\lambda = \frac{\max\,\ell\,(\mu_{\mathrm{MM}}, \mu_{\mathrm{Mm}}, \sigma^2, p = \frac{1}{2})}{\max\,\ell\,(\mu_{\mathrm{MM}}, \mu_{\mathrm{Mm}}, \sigma^2, p)} \tag{2.15}$$

Of course (2.14) is not equivalent to (2.15), since they are testing different hypotheses. In this respect the t-statistic is closer to

$$\lambda^* = \frac{\max\,\ell\,(\mu_{\mathrm{MM}} = \mu_{\mathrm{Mm}}, \sigma^2)}{\max\,\ell\,(\mu_{\mathrm{MM}}, \mu_{\mathrm{Mm}}, \sigma^2, p)} \tag{2.16}$$

with $-2\log\lambda^* \xrightarrow{\mathcal{L}} \chi^2_{(2)}$; however, as before, the two tests are not equivalent in the sense that they do not have to lead to the same conclusion. In fact, λ^* allows only the declaration of existence of genetic control through a major gene by showing that it is linked to the marker, since as for the t-test case, size of the trait effect and its position are confused.

For the F_2 design the marker means are

$$\begin{aligned}
\mu_{\mathrm{MM}} &= (1-\rho)^2\mu_{\mathrm{QQ}} + 2\rho(1-\rho)\mu_{\mathrm{Qq}} + \rho^2\mu_{\mathrm{qq}} \\
\mu_{\mathrm{Mm}} &= \rho(1-\rho)\mu_{\mathrm{QQ}} + \left(\rho^2 + (1-\rho)^2\right)\mu_{\mathrm{Qq}} + (1-\rho)^2\mu_{\mathrm{qq}} \\
\mu_{\mathrm{mm}} &= \rho^2\mu_{\mathrm{QQ}} + 2\rho(1-\rho)\mu_{\mathrm{Qq}} + (1-\rho)^2\mu_{\mathrm{qq}}
\end{aligned}$$

and the variances are

$$\begin{aligned}
\sigma^2_{\mathrm{MM}} &= \sigma^2 + 2\rho(1-\rho)\left((\mu_{\mathrm{QQ}} - \mu_{\mathrm{Qq}}) - \rho\left(\mu_{\mathrm{QQ}} - 2\mu_{\mathrm{Qq}} + \mu_{\mathrm{qq}}\right)\right)^2 \\
&\quad + \rho^2(1-\rho)^2(\mu_{\mathrm{QQ}} - 2\mu_{\mathrm{Qq}} + \mu_{\mathrm{qq}})^2 \\
\sigma^2_{\mathrm{Mm}} &= \sigma^2 + \rho(1-\rho)\left((\mu_{\mathrm{QQ}} - \mu_{\mathrm{Qq}})^2 + (\mu_{\mathrm{qq}} - \mu_{\mathrm{Qq}})^2\right) \\
&\quad + \rho^2(1-\rho)^2(\mu_{\mathrm{QQ}} - 2\mu_{\mathrm{Qq}} + \mu_{\mathrm{qq}})^2 \\
\sigma^2_{\mathrm{mm}} &= \sigma^2 + 2\rho(1-\rho)\left((\mu_{\mathrm{qq}} - \mu_{\mathrm{Qq}}) - \rho\left(\mu_{\mathrm{QQ}} - 2\mu_{\mathrm{Qq}} + \mu_{\mathrm{qq}}\right)\right)^2 \\
&\quad + \rho^2(1-\rho)^2(\mu_{\mathrm{QQ}} - 2\mu_{\mathrm{Qq}} + \mu_{\mathrm{qq}})^2
\end{aligned}$$

This differs substantially from the B_1 case, as these variances are equal only under a non-dominance model, i.e., only if

$$\mu_{\mathrm{Qq}} = \frac{\mu_{\mathrm{QQ}} + \mu_{\mathrm{qq}}}{2}$$

Two contrasts of interest are

$$\begin{aligned}
\mu_{\mathrm{MM}} - \mu_{\mathrm{mm}} &= (\mu_{\mathrm{QQ}} - \mu_{\mathrm{qq}})(1-2\rho) \\
\mu_{\mathrm{Mm}} - \tfrac{1}{2}(\mu_{\mathrm{MM}} + \mu_{\mathrm{mm}}) &= \left(\mu_{\mathrm{Qq}} - \tfrac{1}{2}(\mu_{\mathrm{QQ}} + \mu_{\mathrm{qq}})\right)(1-2\rho)
\end{aligned}$$

and under the non-dominance model

$$\frac{\mu_{\mathrm{MM}} - \mu_{\mathrm{mm}}}{\mu_{\mathrm{MM}} - \mu_{\mathrm{Mm}}} = 2(1-2\rho)$$

In classical studies of linkage with F_2, hypothesis testing is carried out with analysis of variance techniques, since in the absence of linkage the three marker classes have the same mean, regardless of the penetrance model. However, that analysis may not be appropriate when the model includes a dominance term because of the heterogeneity of variances in the marker classes. As in the B_1 case, for F_2, the likelihood ratio approach that uses (2.2) is not equivalent to the analysis of variance.

No matter what the testing procedure, the underlying idea is that no linkage implies that the marker does not contain information about the trait. This suggests an alternative testing strategy, namely permutation. If there is no linkage between marker and the quantitative trait, the test statistic should not be affected if the marker genotypes are held fixed and the trait values shuffled. A set of test statistic values obtained by resampling after random permutation of the trait values yields an empirical distribution for the statistic. The hypothesis of no linkage is rejected whenever the observed value of the test statistic lies among the most extreme values generated by permutation [13]. In a similar fashion, confidence regions for the recombinant fraction can be obtained by parametric bootstrap [87].

Because in experimental populations the crosses are designed to exhibit and contrast the genetic values of the quantitative trait, they can produce reasonably accurate estimates. In general, regression techniques, t-tests, and maximum likelihood inference give more or less equivalent answers for declaring linkage between marker and the major gene loci. As regards position, i.e., the recombinant fraction, the likelihood approach tends to be superior.

The point must be stressed that all of the inferences for experimental crosses described here are conditioned on the marker information. The markers are not important by themselves; they only matter as far as the content of the information about the segregation of the putative quantitative trait locus.

Chapter 3

Linkage methods based on identity by descent

For quantitative traits, the processes that determine the model of inheritance may be unknown. But how can one assess linkage when the genetic model is unknown?. This chapter describes several 'modern' methods which have attempted to answer this question. The proposals are known generically as *robust methods*, and while their genetic foundation resides in the concept of identity by descent, statistically they are quite simple and ordinary: regression and variance components analysis [4, 25, 32, 70]. With the IBD based methods there exists a compromise between simplicity and efficiency. In the recent past, some likelihood approaches to the analysis based on IBD have been proposed. However, the specification of the likelihood function is highly dependent on the genetic model and the desired simplicity disappears. Moreover, the computations under any sensible likelihood approach are as complex as the likelihood approach which uses the whole model.

The pristine work for the problem of detecting linkage between a quantitative trait and a marker was based on sib pair data, and is due to Penrose [72]. Haseman and Elston developed a similar idea and proposed the modern version of sib pair analysis, the Haseman–Elston regression [32], which is the topic of section 3.1. Interval and multipoint mapping, discussed in section 3.2, provide extensions of Haseman-Elston regression, and the 'robust' variance components approach is summarized in section 3.3.

3.1 The Haseman-Elston regression

Haseman and Elston addressed the linkage analysis problem by taking n pairs of relatives of the same kind and regressing the squared difference of their genetic values on the proportion of alleles IBD. Applied to the model (1.1)-(1.3), the rationale goes as follows: assume a diallelic model for the quantitative trait on a zero-loop pedigree and suppose the individuals in the pair $(2i-1, 2i)$ are relatives with *parenté* coefficient ψ, and that this pair shares π_i alleles IBD at the quantitative trait locus. Define $Z_i = (Y_{2i} - Y_{2i-1})$ and notice that the model implies

$$\mathrm{E}(Y_{2i} \,|\, \pi_i) = \mathrm{E}(Y_{2i})$$

$$\mathrm{Var}(Y_{2i} \,|\, \pi_i) = \mathrm{Var}(Y_{2i})$$

Moreover, because $\mathrm{E}(Y_{2i}) = \mathrm{E}(Y_{2i-1})$ and $\mathrm{Var}(Y_{2i}) = \mathrm{Var}(Y_{2i-1})$ by hypothesis, it follows that

$$\begin{aligned}
\mathrm{E}(Z_i^2 \,|\, \pi_i) &= 2\ (\mathrm{Var}(Y_{2i}) - \mathrm{Cov}(Y_{2i}, Y_{2i-1} \,|\, \pi_i)) \\
&= 2\left(\sigma_a^2(1-\pi_i) + \sigma_d^2 \mathrm{I}_{[\pi_i<1]} + \sigma_\eta^2(1-2\psi) + \sigma^2\right) \\
&= \sigma_0^2 - 2\sigma_a^2\pi_i - 2\sigma_d^2 \mathrm{I}_{[\pi_i=1]} \\
&= \sigma_0^2 - 2\sigma_\zeta^2\pi_i + \sigma_d^2 \mathrm{I}_{[\pi_i=\frac{1}{2}]} \qquad (3.1)
\end{aligned}$$

where $\sigma_0^2 = 2\left(\sigma_\zeta^2 + (1-2\psi)\sigma_\eta^2 + \sigma^2\right)$. Since the variances are non-negative, it is evident that the difference in genetic value between two relatives is expected to decrease as they share more alleles identical by descent at the quantitative trait locus.

Under a non-dominance model, σ_d^2 is null and (3.1) becomes

$$\begin{aligned} \mathrm{E}(Z_i^2 \mid \pi_i) &= 2\left(\sigma_\zeta^2(1-\pi_i) + \sigma_\eta^2(1-2\psi) + \sigma^2\right) \\ &= \sigma_0^2 - 2\sigma_\zeta^2 \pi_i \qquad (3.2) \end{aligned}$$

and so only when $\sigma_\zeta^2 > 0$ does the regression of Z_i^2 on π_i have negative slope. Thus if the identities by descent at the trait locus were known for each individual and a simple linear regression model of Z_i^2 on π_i fitted, the usual least squares estimator of the slope would be an unbiased estimator of $-2\sigma_\zeta^2$ whenever (3.2) holds. With non additivity, it is fortunate that if the simple linear regression is carried out for sib pairs, $\hat{\beta}$, the least squares estimator of the slope under the model (3.1) has expectation given by

$$\begin{aligned} \mathrm{E}\left(\hat{\beta} \mid \boldsymbol{\pi}\right) &= \frac{\sum_{i=1}^{n} (\pi_i - \bar{\pi})\ \mathrm{E}(Z_i^2 \mid \pi_i)}{\sum_{i=1}^{n} (\pi_i - \bar{\pi})^2} \\ &= \frac{-2\sigma_\zeta^2 \sum_{i=1}^{n} \pi_i(\pi_i - \bar{\pi}) + \sigma_d^2 \sum_{i=1}^{n} \mathrm{I}_{[\pi_i = \frac{1}{2}]}(\pi_i - \bar{\pi})}{\sum_{i=1}^{n} (\pi_i - \bar{\pi})^2} \\ &= -2\sigma_\zeta^2 + \sigma_d^2 \frac{\sum_{i=1}^{n} \mathrm{I}_{[\pi_i = \frac{1}{2}]}(\pi_i - \bar{\pi})}{\sum_{i=1}^{n} (\pi_i - \bar{\pi})^2} \end{aligned}$$

and because

$$\sum_{i=1}^{n} \mathrm{I}_{[\pi_i = \frac{1}{2}]}(\pi_i - \bar{\pi}) = \frac{n_1}{2n}(n_0 - n_2)$$

with n_j being the number of individuals with j alleles IBD, it follows that in sib pair analysis, $\hat{\beta}$ will be asymptotically unbiased even when dominance is present, since by symmetry n_0 and n_2 tend to equality as the sample size increases.

Of course, linkage analysis would be fatuous if π_i were known; the fact that it cannot be observed makes the problem more interesting. Conditioning on π_i^m, the proportion of IBD alleles at some marker locus, instead of on π_i, the expectation (3.1) can be rephrased as

$$\mathrm{E}(Z_i^2 \mid \pi_i^m) = \sum_{\pi_i} \mathrm{E}(Z_i^2 \mid \pi_i, \pi_i^m) \; \Pr(\pi_i \mid \pi_i^m) \tag{3.3}$$

As long as there are no pleiotropic effects between trait and marker loci, i.e., one trait does not interfere in the expression of the other, it follows that

$$\mathrm{E}(Z_i^2 \mid \pi_i^m) = \sum_{\pi_i} \mathrm{E}(Z_i^2 \mid \pi_i) \; \Pr(\pi_i \mid \pi_i^m) \tag{3.4}$$

When joint Hardy-Weinberg equilibrium at quantitative trait and marker loci holds, the term $\Pr(\pi_i | \pi_i^m)$ accounts for the type of genetic relationship between the pair of individuals under consideration and the recombinant fraction between trait and marker loci, but it does not depend upon what is observed at the marker locus [2]. $\Pr(\pi_i | \pi_i^m)$ for some different types of relationship have been published [2, 32]. Table 3.1 contains this conditional distribution for sib-ships.

Using table (3.1) and the equation (3.1) for siblings (3.4) yields

$$\begin{aligned} \mathrm{E}(Z_i^2 \mid \pi_i^m) &= \sigma_0^2 + 2\left(\varrho\sigma_\zeta^2 + \varrho(1-\varrho)\sigma_d^2\right) \\ &\qquad + 2(1-2\varrho)\sigma_\zeta^2\pi_i^m + (1-2\varrho)^2\sigma_d^2 \mathrm{I}_{[\pi_i^m = \frac{1}{2}]} \\ &= \beta_0 + \beta_1\pi_i^m + \beta_2 \mathrm{I}_{[\pi_i^m = \frac{1}{2}]} \end{aligned} \tag{3.5}$$

$\Pr(\pi_i \mid \pi_i^m)$		π_i^m		
		0	$\frac{1}{2}$	1
	0	ϱ^2	$\varrho(1-\varrho)$	$(1-\varrho)^2$
π_i	$\frac{1}{2}$	$2\varrho(1-\varrho)$	$1-2\varrho(1-\varrho)$	$2\varrho(1-\varrho)$
	1	$(1-\varrho)^2$	$\varrho(1-\varrho)$	ϱ^2
		$\varrho = \rho^2 + (1-\rho)^2$		

Table 3.1: Probability of the IBD sharing proportions at the trait locus conditioned on the IBD sharing proportions at the marker in siblings.

The regression coefficients in (3.5) comprise the information about the variance components and recombinant fraction available in Z_i and π_i^m. Because $(1-2\varrho) = -(1-2\rho)^2$, negative values of β_1 provide evidence of linkage. Analogous linear relationships have been established for other types of relationship, and in those cases also, β_1 is negative only if there is linkage [2]. For example,

$$\beta_1 = \begin{cases} -2(1-2\rho)^2\sigma_a^2 & \text{half-sibs} \\ -2(1-2\rho)\sigma_a^2 & \text{grandparent-grandchild} \\ -2(1-2\rho)^2(1-\rho)\sigma_a^2 & \text{avuncular} \\ -2(1-2\rho)^2(1-\frac{4}{3}\rho+\frac{2}{3}\rho^2)\sigma_a^2 & \text{first-cousins} \end{cases}$$

What Haseman and Elston proposed was to not condition on the marker IBD sharing proportions, but instead on $\mathcal{J}_{mi}$, the marker information. For example, if the marker is fully informative then the marker information is the marker. If one parent is homozygous and the other heterozygous for the marker, then the marker provides only partial information on linkage, while in the case that both parents are homozygous, the markers are completely uninformative. Conditioning on the marker information, it follows that

$$\begin{aligned}\mathrm{E}(Z_i^2 \mid \mathcal{J}_{\mathrm{m}i}) &= \sum_{\pi_i^{\mathrm{m}}} \mathrm{E}(Z_i^2 \mid \pi_i^{\mathrm{m}}, \mathcal{J}_{\mathrm{m}i})\ \Pr(\pi_i^{\mathrm{m}} \mid \mathcal{J}_{\mathrm{m}i}) \\ &= \sum_{\pi_i^{\mathrm{m}}} \sum_{\pi_i} \mathrm{E}(Z_i^2 \mid \pi_i) \Pr(\pi_i \mid \pi_i^{\mathrm{m}}) \Pr(\pi_i^{\mathrm{m}} \mid \mathcal{J}_{\mathrm{m}i}) \qquad (3.6)\end{aligned}$$

an expression that can be reparameterized as

$$\mathrm{E}(Z_i^2 \mid \mathcal{J}_{\mathrm{m}i}) = \beta_0 + \beta_1 \pi_i^{\mathcal{J}_{\mathrm{m}}} + \beta_2 \Pr(\pi_i^{\mathrm{m}} = \tfrac{1}{2} \mid \mathcal{J}_{\mathrm{m}i}) \qquad (3.7)$$

where

$$\pi_i^{\mathcal{J}_{\mathrm{m}}} = \Pr(\pi_i^{\mathrm{m}} = 1 \mid \mathcal{J}_{\mathrm{m}i}) + \tfrac{1}{2} \Pr(\pi_i^{\mathrm{m}} = \tfrac{1}{2} \mid \mathcal{J}_{\mathrm{m}i}) \qquad (3.8)$$

With full marker information for the pair and the individuals connecting them in the pedigree, $\Pr(\pi_i^{\mathrm{m}} \mid \mathcal{J}_{\mathrm{m}i})$ can be determined from the data. However, when the marker provides only partial information for linkage, $\Pr(\pi_i^{\mathrm{m}} \mid \mathcal{J}_{\mathrm{m}i})$ becomes a quantity which depends also on the marker allele proportions in the population. Haseman and Elston's approach was to throw away β_2, to replace $\pi_i^{\mathcal{J}_{\mathrm{m}}}$ by its expectation $\hat{\pi}_i^{\mathcal{J}_{\mathrm{m}}}$, to carry out a simple linear regression of Z_i^2 on $\hat{\pi}_i^{\mathcal{J}_{\mathrm{m}}}$, and to declare linkage between marker and quantitative trait whenever β_1 was found to be significantly less than zero. That rule is still in use.

There is an alternate reparameterization to (3.7) which is useful in many of the extensions to the Haseman–Elston regression, which requires

$$\begin{aligned}\mathrm{E}(Z_i^2 \mid \mathcal{J}_{\mathrm{m}i}) &= \sum_{\pi_i} \mathrm{E}(Z_i^2 \mid \pi_i) \Pr(\pi_i \mid \mathcal{J}_{\mathrm{m}i}) \\ &= \sigma_0^2 - 2\sigma_\zeta^2 \pi_\tau^{\mathcal{J}_{\mathrm{m}i}} + \sigma_d^2 \Pr(\pi_i = \tfrac{1}{2} \mid \mathcal{J}_{\mathrm{m}i}) \qquad (3.9)\end{aligned}$$

where

$$\Pr(\pi_i \mid \mathcal{I}_{\mathfrak{m}i}) = \sum_{\pi_i^{\mathfrak{m}}} \Pr(\pi_i \mid \pi_i^{\mathfrak{m}}) \Pr(\pi_i^{\mathfrak{m}} \mid \mathcal{I}_{\mathfrak{m}i}) \tag{3.10}$$

and

$$\pi_{\mathfrak{t}}^{\mathcal{I}_{\mathfrak{m}i}} = \Pr(\pi_i = 1 \mid \mathcal{I}_{\mathfrak{m}i}) + \tfrac{1}{2}\Pr(\pi_i = \tfrac{1}{2} \mid \mathcal{I}_{\mathfrak{m}i}) \tag{3.11}$$

While (3.7) seems more convenient from a computational viewpoint, (3.9) can be easily adapted to situations with more than one marker. In fact, if σ_d^2 is dropped from consideration, a parameterization equivalent to (3.9) has been used extensively in interval mapping [21] and multipoint interval mapping [21, 23, 47].

The Haseman–Elston regression has many advantages: it allows multiple allelism in the marker, it does not require an accurate knowledge of the genetic mechanism underlying the trait, even when it was derived under a two allele model; it is unbiased, provided that there is not dominance at the trait locus or the bias is small for large samples; and many more [2, 6, 32]. But, the biggest asset of the Haseman–Elston regression is its simplicity. Modifications to the original formulation have taken advantage of informative marker data by using reweighted least squares or carrying out the regression through the EM algorithm, but these modifications suffer a substantial loss of simplicity [3, 47, 89].

There are several caveats with the Haseman–Elston regression. It is a 'detection only' procedure since the linkage parameter and the variance component for the quantitative trait are confounded. It uses only pairs of individuals with the same type of relationship, although this inconvenience can be overcome using iteratively reweighted least squares [66, 65]. Other potential problems are related to heteroscedasticity of the errors (but not under the null hypothesis) and the lack of independence among different pairs as far as individuals belonging to the same family and/or the pairs have an ancestor in common. A more recent criticism is that the Haseman–Elston regression wastes information. In the early 80's it was accepted that non-normality was not an issue with samples of moderate size and the usual t-test was said to be satisfactory [6]. Recently, this drawback has been challenged as the adoption of maximum likelihood methods has become matter of course. The fact that Z_i^2 does not have a normal distribution implies that linear regression does not yield maximum likelihood estimates. After pointing this out, Kruglyak and Lander suggested use of the likelihood

$$\ell = \prod_i \sum_j p_{ij} \varphi_{\sigma_j^2}(z_j) \tag{3.12}$$

where $\sigma_1^2 = \sigma_0^2 - \sigma_a^2$, $\sigma_2^2 = \sigma_0^2 - \sigma_\zeta^2$, and $p_{ij} = \Pr(\pi_i = \frac{j}{2})$, $j = 0, 1, 2$ with estimation of the three variances through the EM algorithm [47]. While this approach may be superior to the traditional Haseman–Elston regression, the offered solution seems to be the wrong answer to the right question. First of all, underlying (3.12) there is a strong assumption that conditional on $\pi_i = \frac{j}{2}$, $Z_i \sim \mathcal{N}(0, \sigma_j^2)$, which obviously does not hold. Recall that the distribution of Y_i is a mixture of normals, so the distribution of Z_i has to be a mixture as well. For example, the density of Z_i conditioned on π_i for

	$p_j(\pi_i)$	π_i		
		0	$\frac{1}{2}$	1
	0	$p^4 + 4pq + q^4$	$p^2 + q^2$	1
μ_j	$\zeta_{QQ} - \zeta_{Qq}$	$4p^3q$	$2p^2q$	0
	$\zeta_{Qq} - \zeta_{qq}$	$4pq^3$	$2pq^2$	0
	$\zeta_{QQ} - \zeta_{qq}$	$2p^2q^2$	0	0

Table 3.2: Coefficients for evaluating the density of Z_i given π_i for sib-pairs (p is the proportion of alleles Q in the population, $q = 1 - p$).

sib pairs is

$$\mathrm{f}(z_i \mid \pi_i) = \sum_j \mathrm{p}_j(\pi_i)\, \varphi_{\sigma_\eta^2+2\sigma^2}(|z_i| - \mu_j) \tag{3.13}$$

where the coefficients μ_j and $\mathrm{p}_j(\pi_i)$ come from table 3.2. Then, conditioned on $\mathcal{J}_{mi}$, the density of Z_i is

$$\begin{aligned} \mathrm{f}(z_i \mid \mathcal{J}_{mi}) &= \sum_{\pi_i^m} \sum_{\pi_i} \mathrm{f}(z_i \mid \pi_i) \Pr(\pi_i \mid \pi_i^m) \Pr(\pi_i^m \mid \mathcal{J}_{mi}) \\ &= \sum_{\pi_i} \mathrm{f}(z_i \mid \pi_i) \Pr(\pi_i \mid \mathcal{J}_{mi}) \end{aligned} \tag{3.14}$$

Secondly, as Fulker and Cherny [22] indicated, the use of sib-pair differences does not lead to an optimal maximum likelihood approach for quantitative sib-pair data, which are in bivariate form. In fact, the Haseman-Elston regression wastes information as shown in the following argument. Define $Z_1 = Y_1 - Y_2$ and $Z_2 = Y_1 + Y_2$, then

$$\mathrm{f}(Y_1, Y_2 \mid \pi) = \mathrm{f}(Z_1 \mid \pi)\, \mathrm{f}(Z_2 \mid \pi) \tag{3.15}$$

i.e., using only Z_1 throws away the information contained in Z_2. Moreover, it has been shown that

$$\mathrm{E}(Z_1^2 \mid \pi) = \beta_0 + \beta_1 \pi + \beta_2 \mathrm{I}_{[\pi=\frac{1}{2}]} \tag{3.16}$$

and it is straightforward to show that

$$\begin{aligned} \mathrm{E}(Z_2^2 \mid \pi) &= \mathrm{Var}(Z_2 \mid \pi) + 4\mu^2 \\ &= 4\mu^2 + 2\,\mathrm{Var}(Y_1) + 2\sigma_\zeta^2 \pi - \sigma_d^2 \mathrm{I}_{[\pi=\frac{1}{2}]} \end{aligned}$$

which implies that

$$-\mathrm{E}(Z_2^2 \mid \pi) = \beta_0' + \beta_1 \pi + \beta_2 \mathrm{I}_{[\pi=\frac{1}{2}]} \tag{3.17}$$

where β_1 and β_2 are as in (3.16). That means the regression of the squared sum pair on π is parallel to the regression of the squared difference, which means the sum contains as much information about linkage as the difference does. So, the simple combination of both regressions must do a better job. Fulker and Cherny advocate using the bivariate form by testing the difference between observed and expected covariance matrices based on the Whishart distribution; unfortunately, this approach also requires normality for Y_i.

3.2 Interval and multipoint mapping

With two alleles at the marker locus the number of informative combinations is small. When only two-allele markers are available, an increase in the number of markers may be a method of increasing the IBD information at the trait locus. On the other hand, no matter how informative a marker can be, for the traditional Haseman-Elston regression, the size of the effect and position are confounded. With two or more markers both size and position can be determined. Fulker and Cardon [21] developed an extension to the Haseman-Elston regression, known as interval mapping, which employs the information on two flanking markers separated by a known map distance to locate the quantitative trait gene and estimate the size of its effects. Later on, this approach was further extended to multipoint mapping, which is nothing but the inclusion of more markers [23].

The idea behind interval and multipoint mapping is to take (3.2) and to approximate π_t as a linear function of the IBD sharing proportions in the markers. To do so, it was proposed to use the best linear mean square predictor. To derive this consider that, given the relationship, the expectation and variance of IBD sharing proportions are the same for any locus, and the correlation between the IBD proportions for any pair of loci depends only on the recombinant fraction, which is known provided the map distance is known. Table 3.3 contains these constants for some selected pair types. The proposed approximation for π can be written as

$$\pi = \alpha_0 + \boldsymbol{\alpha}'\boldsymbol{\pi}^{\mathrm{m}} \tag{3.18}$$

where

$$\boldsymbol{\alpha} = \mathrm{Cov}(\pi, \boldsymbol{\pi}^{\mathrm{m}})\,\mathrm{Var}(\boldsymbol{\pi}^{\mathrm{m}})^{-1} \qquad \alpha_0 = \mathrm{E}(\pi)\,(1 - \boldsymbol{\alpha}'\mathbb{1})$$

$\boldsymbol{\pi}^{\mathrm{m}}$ is the vector of markers linked to the quantitative trait. $\mathrm{E}(\pi)$ and the expectation of the IBD proportions for any other pair of loci on the same pair of individuals is 2ψ, as was mentioned before. The matrix $\mathrm{Var}(\boldsymbol{\pi}^{\mathrm{m}})$ can be evaluated because the map distances between markers are known; in contrast, the row vector $\mathrm{Cov}(\pi, \boldsymbol{\pi}^{\mathrm{m}}) = \mathrm{Var}(\pi)\,\mathrm{Corr}(\pi, \boldsymbol{\pi}^{\mathrm{m}})$ depends on the unknown recombinant fractions between each marker with the quantitative trait. To overcome this inconvenience, the chromosome is divided into a number of intervals and sequentially, a point ξ inside each interval is chosen, and the assumption made that the quantitative trait lies at that point (usually, the middle of each interval). Once this is done, the map distance between each ξ and the position of all of the markers can be transformed into recombinant fractions to be plugged into $\mathrm{Cov}_{\xi}(\pi, \boldsymbol{\pi}^{\mathrm{m}})$ so that π_{ξ} can be computed. Afterwards, as in Haseman-Elston regression, a simple linear regression of Z^2 on each π_{ξ} is carried out. The estimate of position for the quantitative trait is that ξ for which the residual sum of squares is minimum.

3.3 The robust variance components approach

The approach to the problem of a major quantitative trait which seems to be becoming most popular is the IBD based variance components linkage method. Based on the work of Golgar [25] and Schork [76], Amos [4] made the first formal proposal of such methods. The problem that Golgar and Schork addressed was to estimate the genetic variance due to the additive

Pair type	Distribution of 2π	$\text{Corr}(\pi, \pi^*)$
full-sibs	$\mathcal{B}in(2, \frac{1}{2})$	$(1-2\rho^*)^2$
half-sibs	$\mathcal{B}er(\frac{1}{2})$	$(1-2\rho^*)^2$
grandparent-grandchild	$\mathcal{B}er(\frac{1}{2})$	$(1-2\rho^*)$
avuncular	$\mathcal{B}er(\frac{1}{2})$	$(1-2\rho^*)^2(1-\rho^*)$
first-cousins	$\mathcal{B}er(\frac{1}{4})$	$(1-2\rho^*)^2(1-\frac{4}{3}\rho^*+\frac{2}{3}\rho^{*2})$
half-avuncular	$\mathcal{B}er(\frac{1}{4})$	$(1-2\rho^*)(1-2\rho^*+\frac{4}{3}\rho^{*2})$

[1] ρ^* represents the recombinant fraction between the loci whose IBD proportions are π and π^*.

Table 3.3: Distribution of IBD counts and correlation between any two IBD proportions for some pair types.

effects of one or more loci located in a segment of the chromosome defined by two flanking markers. Amos [4] took the same problem addressed by Haseman and Elston [32] but from a mixed effects variance components approach and it can be easily extended to interval and multipoint mapping by the same strategy used to extend the Haseman-Elston regression [1, 90].

The model (1.1) in matrix notation is

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta} + \mathbf{e} \tag{3.19}$$

with $\boldsymbol{\mu}$ holding the family means and any other fixed effects. If one splits the major gene effect, $\boldsymbol{\zeta}$, into additive and dominance effects and assumes that the polygenic effect is only additive, then we have

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{a} + \mathbf{d} + \boldsymbol{\eta} + \mathbf{e} \tag{3.20}$$

with

$$\begin{aligned} \mathrm{E}(\mathbf{y} \mid \mathbf{\Pi}) &= \boldsymbol{\mu} \\ &= \mathrm{E}(\mathbf{y}) \end{aligned} \tag{3.21}$$

and

$$\begin{aligned} \mathrm{Var}(\mathbf{y} \mid \mathbf{\Pi}) &= \sigma_a^2 \, \mathbf{\Pi} + \sigma_d^2 \, \mathbf{\Delta} + 2\sigma_\eta^2 \, \mathbf{\Psi} + \sigma_e^2 \, \mathbf{R} \\ &= \sigma_\zeta^2 \, \mathbf{\Pi} - \tfrac{1}{2}\sigma_d^2 \, \mathbf{\Delta}^* + 2\sigma_\eta^2 \, \mathbf{\Psi} + \sigma_e^2 \, \mathbf{R} \\ &= \mathbf{\Sigma}_\pi \end{aligned} \tag{3.22}$$

where $\pi_{ii} = 1$ and π_{ij} is the IBD sharing proportion between the ith and jth individuals, $\delta_{ij} = \mathrm{I}_{[\pi_{ij}=1]}$, $\delta^*_{ij} = \mathrm{I}_{[\pi_{ij}=\frac{1}{2}]}$, $\mathbf{\Psi}$ the Malécot's *parenté* matrix, and $\mathbf{R}$ may be an identity matrix or something else to model shared environment among families.

Dropping σ_d^2, the previous expression becomes

$$\mathrm{Var}(\mathbf{y} \mid \mathbf{\Pi}) = \sigma_\zeta^2 \, \mathbf{\Pi} + 2\sigma_\eta^2 \, \mathbf{\Psi} + \sigma_e^2 \, \mathbf{R} \tag{3.23}$$

Then, following similar steps as in the Haseman-Elston regression, it can be shown that

$$\begin{aligned} \mathrm{Var}(\mathbf{y} \mid \mathbf{\Pi}^{\mathrm{m}}) &= \sigma_\zeta^2 \, (\mathbf{\Pi}^{\mathrm{m}} - 2\mathbf{\Psi}) \star \text{ß}^{(\varrho)} + 2(\sigma_\zeta^2 + \sigma_\eta^2) \, \mathbf{\Psi} + \sigma_e^2 \, \mathbf{R} \\ &= \mathbf{\Sigma}_{\pi^{\mathrm{m}}} \end{aligned} \tag{3.24}$$

where '$\star$' denotes Hadamard product and the elements of the matrix $\text{ß}^{(\varrho)}$ are given by $\text{ß}_{ij}^{(\varrho)} = \mathrm{Corr}_\rho\left(\pi_{ij}, \pi_{ij}^{\mathrm{m}}\right)$, i.e. they depend on the recombinant fraction between trait and marker loci and the type of relationship between the ith and jth individuals. For some selected types of kinship, the correlations between two IBD proportions are shown in table 3.3; a more complete table can be found in Almasy and Blangero [1].

Taking (3.21) and (3.24) as a base, it is common to propose for $\mathbf{y}$ the following log-likelihood

$$-\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log\left|\boldsymbol{\Sigma}_{\pi^m}\right| - \tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_{\pi^m}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{3.25}$$

where n is the dimension of $\mathbf{y}$, i.e., the number of individuals in the pedigree. This ignores two key facts: $\mathbf{y}$ is not normal and $\boldsymbol{\Sigma}_{\pi^m}$ is a conditional variance. Xu [88] pointed out the latter fact and proposed to use the likelihood conditioned on the marker information, given by

$$\sum_{\boldsymbol{\Pi}} \Pr(\boldsymbol{\Pi} \mid \mathcal{I}_m)\ \varphi_{\boldsymbol{\Sigma}_\pi}(\mathbf{y}-\boldsymbol{\mu}) \tag{3.26}$$

This assumes that conditioned on the IBD status at the quantitative trait, the distribution of $\mathbf{y}$ is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\pi)$, i.e., it implies that $\boldsymbol{\zeta}$ has a normal distribution, which is obviously inaccurate. While this formulation may be superior to (3.25), it is still unjustified. In fact, conditioned on $\boldsymbol{\Pi}$, the density of $\mathbf{y}$ looks like

$$\mathrm{f}(\mathbf{y} \mid \boldsymbol{\Pi}) = \sum_{\boldsymbol{\zeta}^\pi} \Pr(\boldsymbol{\zeta}^\pi \mid \boldsymbol{\Pi})\ \varphi_{2\sigma_\eta^2\boldsymbol{\Psi}+\sigma_e^2\mathbf{R}}(\mathbf{y}-\boldsymbol{\mu}-\boldsymbol{\zeta}^\pi) \tag{3.27}$$

where $\boldsymbol{\zeta}^\pi$ is a vector $\boldsymbol{\zeta}$ compatible with $\boldsymbol{\Pi}$, by example, $\pi_{ij} = 1 \iff \zeta_i^\pi = \zeta_j^\pi$. So, the likelihood conditioned on the marker information must look like

$$\sum_{\boldsymbol{\Pi}} \Pr(\boldsymbol{\Pi} \mid \mathcal{I}_m)\ \mathrm{f}(\mathbf{y} \mid \boldsymbol{\Pi}) \tag{3.28}$$

There are some considerations to be pointed out:

- To assume normality of $\boldsymbol{\zeta}$ may be convenient but is not realistic in most practical situations.
- Even if the number of alleles coding for each ζ_i is large, only a few genotypes will be heavily represented in the population, a situation in

which it would be unwise to invoke the Central Limit Theorem for the distribution of the quantitative trait values.

- It is well known that a 'non-degenerate' finite mixture of normals cannot be normal [85].

In light of these facts, it is questionable whether sensible answers regarding linkage can be obtained under the assumption of normality.

Chapter 4

Likelihood evaluation

In this chapter several methods for evaluating likelihoods for quantitative trait data are discussed. For simple models which include only a monogenic or only a polygenic component, there are a collection of so-called peeling algorithms which allow for decomposition of the likelihood into tractable components. Two such algorithms are discussed in sections 4.1 and 4.2. The mixed model described in section 4.3, which includes both major gene and polygenic effects, cannot be peeled, and requires other techniques for likelihood evaluation. A generalization of the Morton-MacLean algorithm is derived in section 4.4 which allows for exact computation, up to numerical quadrature, of likelihoods in cases where phenotypic data is available for the nuclear families at the bottom of the pedigree, and possibly for the parents of those families. An arbitrary pedigree structure for the parents' ancestors is allowed. In many situations there will be phenotypic observations throughout the pedigree. In such cases, a suite of methods have been developed for Monte Carlo approximation of the likelihood. Several such methods are discussed in section 4.5, including gene dropping, the Gibbs sampler, and the Hastings-Metropolis algorithm. The performance of these methods on some simulated data sets will be discussed in chapter 6.

4.1 Peeling monogenic models.

Earlier we made mention of the Elston–Stewart algorithm in the context of the likelihood for nuclear families. In fact, this algorithm is much more general. Originally intended for simple pedigrees without loops, the algorithm has been extended considerably by Ott [68], Lange and Elston [52], and Cannings *et al.* [9] among many others. The main idea behind the algorithm is that instead of dealing with all of the genotypes at the same time, the computations are broken down into many pieces, each involving a small number of individuals. In a pedigree, the phenotypes and/or genotypes of related individuals may be dependent upon each other. However, a decomposition is possible because the phenotypes of some individuals will be independent conditioned on the genotypes of some others. For example, in a pedigree without loops, the relationship of an individual to its ancestors comes through its parents; hence, knowing the genotype of the parents, any knowledge of the genotype of the other ancestors does not yield further information. Another example is the genotype of a group of full sibs. As a consequence of Mendel's first law, which for diploid organisms establishes that the genotype at a given locus is determined by random sampling of parental gametes, one has that given the genotype of the parents, the genotypes of their children are independent, in the absence of a family effect.

The spirit of the algorithm is best explained by working through an example. Consider the pedigree given in figure 4.1 which is a typical example of a pedigree without loops. Let us evaluate the likelihood under a monogenic model $\mathbf{y} = \boldsymbol{\zeta} + \mathbf{e}$.

In agreement with (1.7), the likelihood can be written as

$$\mathrm{f}(\mathbf{y}) = \sum_{\zeta_1}\sum_{\zeta_2}\cdots\sum_{\zeta_{13}}\prod_{j=1}^{13}\Pr(\zeta_j)\ \mathrm{f}(y_j \mid \zeta_j) \tag{4.1}$$

While it is straightforward to evaluate the product in the previous equation,

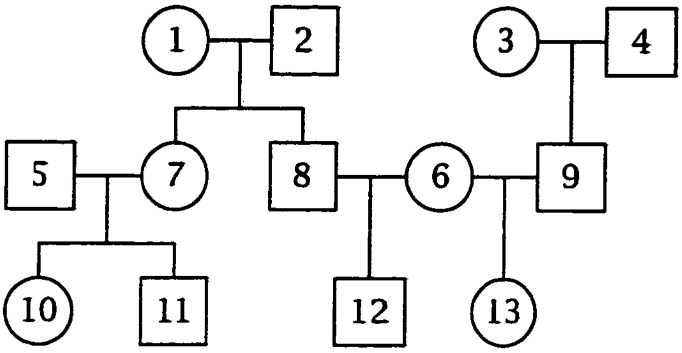

Figure 4.1: Pedigree without loops

the sum over all ζs may not be needed. On pedigrees of limited size, it may be possible to enumerate all of the configurations for ζ and to then compute the sums; however, in general this approach will be impractical for most cases. Elston and Stewart [17] proposed a recursive algorithm that points out the sequence in which sums and products are to be performed in order to efficiently evaluate (4.1). This sequential process is known as *peeling*. In our example, such a sequence can be described as follows.

i). For each possible pair (ζ_5, ζ_7), evaluate

$$f(y_{10} \mid \zeta_5, \zeta_7) = \sum_{\zeta_{10}} \Pr(\zeta_{10} \mid \zeta_5, \zeta_7)\, f(y_{10} \mid \zeta_{10})$$

and

$$f(y_{11} \mid \zeta_5, \zeta_7) = \sum_{\zeta_{11}} \Pr(\zeta_{11} \mid \zeta_5, \zeta_7)\, f(y_{11} \mid \zeta_{11})$$

then, compute

$$f(y_{10}, y_{11} \mid \zeta_5, \zeta_7) = f(y_{10} \mid \zeta_5, \zeta_7)\, f(y_{11} \mid \zeta_5, \zeta_7)$$

ii). For each possible ζ_7, evaluate

$$\begin{aligned} f(\mathbf{y}_7^* \mid \zeta_7) &= f(y_5, y_7, y_{10}, y_{11} \mid \zeta_7) \\ &= f(y_7 \mid \zeta_7) \sum_{\zeta_5} \Pr(\zeta_5)\, f(y_5 \mid \zeta_5)\, f(y_{10}, y_{11} \mid \zeta_5, \zeta_7) \end{aligned}$$

then, compute

$$f(\mathbf{y}_7^* \mid \zeta_1, \zeta_2) = \sum_{\zeta_7} \Pr(\zeta_7 \mid \zeta_1, \zeta_2)\, f(\mathbf{y}_7^* \mid \zeta_7)$$

for all possible pairs (ζ_1, ζ_2).

iii). For each possible ζ_9 evaluate

$$\begin{aligned} f(\mathbf{y}_9^*, \zeta_9) &= f(y_3, y_4, y_9, \zeta_9) \\ &= f(y_9 \mid \zeta_9) \sum_{\zeta_3} \Pr(\zeta_3)\, f(y_3 \mid \zeta_3) \\ &\quad \sum_{\zeta_4} \Pr(\zeta_4)\, f(y_4 \mid \zeta_4) \Pr(\zeta_9 \mid \zeta_3, \zeta_4) \end{aligned}$$

iv). After computing $f(y_{13} \mid \zeta_6, \zeta_9)$, for each possible ζ_6, evaluate

$$\begin{aligned} f(\mathbf{y}_6^*, \zeta_6) &= f(y_6, \mathbf{y}_9^*, y_{13}, \zeta_6) \\ &= \Pr(\zeta_6)\, f(y_6 \mid \zeta_6) \sum_{\zeta_9} f(\mathbf{y}_9^*, \zeta_9)\, f(y_{13} \mid \zeta_6, \zeta_9) \end{aligned}$$

v). Compute $f(y_{12} \mid \zeta_6, \zeta_8)$ and after that, for each possible ζ_8, evaluate

$$\begin{aligned} \Pr(\mathbf{y}_8^* \mid \zeta_8) &= f(\mathbf{y}_6^*, y_8, y_{12} \mid \zeta_8) \\ &= f(y_8 \mid \zeta_8) \sum_{\zeta_6} f(\mathbf{y}_6^*, \zeta_6)\, f(y_{12} \mid \zeta_6, \zeta_8) \end{aligned}$$

then, for all possible pairs (ζ_1, ζ_2), compute

$$f(y_8^* | \zeta_1, \zeta_2) = \sum_{\zeta_8} \Pr(\zeta_8 | \zeta_1, \zeta_2)\, f(y_8^* | \zeta_8)$$

vi). Compute

$$f(y_7^*, y_8^* | \zeta_1, \zeta_2) = f(y_7^* | \zeta_1, \zeta_2)\, f(y_8^* | \zeta_1, \zeta_2)$$

and, finally,

$$\Pr(y) = \sum_{\zeta_1} \Pr(\zeta_1)\, f(y_1 \mid \zeta_1) \sum_{\zeta_2} \Pr(\zeta_2)\, f(y_2 \mid \zeta_2)\, \Pr(y_7^*, y_8^* | \zeta_1, \zeta_2)$$

The key feature of the algorithm is to look at the edges of the pedigree for points where the computations should start, in order to limit the number of genotypes that must be considered simultaneously. A pedigree without loops can be seen as a sequence of nuclear families with neighbors connected by a single individual, the pivot. Conditioning on the pivot, the computation for non-pivot members of a peripheral family can be carried out, and the summation over those non-pivot members leads to a function of only pivot genotypes, and thereby gets rid of the non-pivot members. The process is repeated successively for each family, incorporating previous contributions through the pivots, until the pedigree is exhausted. At the end, the likelihood for the pedigree is obtained. The specification of the order of pivots in which the computations are carried out is known as the *peeling sequence*, and in many situations the efficiency of the algorithm may be affected by the choice of such sequence [45].

The Elston-Stewart algorithm is not the only one to compute 'exact' likelihoods for oligogenic models. Heuch and Li [35] proposed another recursive algorithm. There are also iterative algorithms intended to deal with large pedigrees without loops and others than can be implemented recursive or iteratively [18, 51, 64, 73, 86].

4.2 Peeling polygenic models

Elston and Stewart [17] have shown that the peeling algorithm for computing likelihoods in the simple additive polygenic model with independent errors, i.e., no underlying environmental effect associated with sibship, is essentially the same algorithm as the one for oligogenotypes, with integrals instead of sums. Additionally, in the polygenic case the integrals can be explicitly evaluated since

$$\int \mathrm{e}^{-\frac{1}{2}\mathbf{w}'\mathbf{B}\mathbf{w}}\,\mathrm{d}w_k = \sqrt{\frac{2\pi}{b_{kk}}}\,\mathrm{e}^{-\frac{1}{2}\mathbf{w}^{*\prime}(\mathbf{B}^{*}-b_{kk}^{-1}\mathbf{b}\mathbf{b}')\mathbf{w}^{*}} \tag{4.2}$$

where

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}^{*} \\ w_k \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}^{*} & \mathbf{b} \\ \mathbf{b}' & b_{kk} \end{pmatrix}$$

This property has been used extensively to compute the likelihood for polygenic models [27, 84].

While peeling polygenic models may be a ‘natural’ thing from the genetics viewpoint, the assumption of normality for trait values $\mathbf{y}$ allows computations to be shifted from the domain of combinatorics to linear algebra, where the classical apparatus of multivariate analysis can be invoked, thereby making the computations, up to numerical considerations, straightforward. For example, the additive polygenic model can be parameterized as

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\eta} + \mathbf{e} \tag{4.3}$$

where $\boldsymbol{\mu}$ is a vector of parameters, and $\boldsymbol{\eta}$ and $\mathbf{e}$ are genetic and environmental effects, respectively, with

$$\mathrm{E}\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\eta} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ 0 \\ 0 \end{pmatrix} \tag{4.4}$$

and

$$\mathrm{Var}\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\eta} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \sigma_\eta^2\mathbf{A} + \sigma_e^2\mathbf{R} & \sigma_\eta^2\mathbf{A} & \sigma_e^2\mathbf{R} \\ \sigma_\eta^2\mathbf{A} & \sigma_\eta^2\mathbf{A} & 0 \\ \sigma_e^2\mathbf{R} & 0 & \sigma_e^2\mathbf{R} \end{pmatrix} \tag{4.5}$$

where $\mathbf{A} = 2\boldsymbol{\psi}$. The log-likelihood for this model is

$$\ell = -\tfrac{n}{2}\log(2\pi) - \log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.6}$$

where

$$\mathbf{V} = \sigma_\eta^2\mathbf{A} + \sigma_e^2\mathbf{R}$$

and

$$\mathbf{V}^{-1} = \tfrac{1}{\sigma_e^2}\mathbf{R}^{-1} - \tfrac{1}{\sigma_e^2}\mathbf{R}^{-1}\left(\mathbf{R}^{-1} + \tfrac{\sigma_e^2}{\sigma_\eta^2}\mathbf{A}^{-1}\right)^{-1}\mathbf{R}^{-1}$$

It can be seen that computing $\mathbf{A}$ and its inverse becomes an important part of the computations. Let ψ_{ij} be the coefficient of *parenté* between the ith and jth individuals, and $(\mathrm{q}_{i1}, \mathrm{q}_{i2})$ and $(\mathrm{q}_{j1}, \mathrm{q}_{j2})$ be their respective gametes. Then, by definition,

$$\psi_{ij} = \tfrac{1}{4}\left(\Pr\left(\mathrm{q}_{i1} \overset{bd}{\equiv} \mathrm{q}_{j1}\right) + \Pr\left(\mathrm{q}_{i1} \overset{bd}{\equiv} \mathrm{q}_{j2}\right) + \Pr\left(\mathrm{q}_{i2} \overset{bd}{\equiv} \mathrm{q}_{j1}\right) + \Pr\left(\mathrm{q}_{i2} \overset{bd}{\equiv} \mathrm{q}_{j2}\right)\right)$$

where ‘ $\overset{bd}{\equiv}$ ’ means ‘identical by descent to’. It is straightforward to show that

$$\psi_{ij} = \tfrac{1}{4}\left(\psi_{i_f j_f} + \psi_{i_f j_m} + \psi_{i_m j_f} + \psi_{i_m j_m}\right) \tag{4.7}$$

$$= \tfrac{1}{2}\left(\psi_{i j_f} + \psi_{i j_m}\right) \tag{4.8}$$

where $\psi_{i_f j_f}$ is the coefficient of *parenté* between the fathers of the ith and jth individuals, and so on. Then, the elements of A can be computed recursively [59]. However, computing the inverse of A may not be as direct. Without loss of generality, it can be assumed that each individual in the pedigree is a founder, or both of its parents are in the pedigree. Let $\mathcal{F}$ be the set of indices of founder individuals and suppose also that founders precede non-founders. Define $n_{\mathcal{F}} = |\mathcal{F}|$, the number of founders. Now, rewrite the recursive computations for A as

$$\mathbf{A}_{i+1} = \begin{cases} \mathbf{I}_{i+1} & i < n_{\mathcal{F}} \\ \begin{pmatrix} \mathbf{A}_i & \mathbf{A}_i\mathbf{k}_i \\ \mathbf{k}_i'\mathbf{A}_i & 1 + F_{i+1} \end{pmatrix} & i \geq n_{\mathcal{F}} \end{cases} \tag{4.9}$$

where $\mathbf{k}_i$ is a vector with $\frac{1}{2}$ in the positions f and m, the indices for the parents of the $(i+1)$-th individual, and zeroes elsewhere; F_{i+1} is the inbreeding coefficient of the $(i+1)$-th individual, i.e, $F_{i+1} = \frac{1}{2}a_{fm}$.

From the previous representation and the fact that

$$\mathbf{k}_i'\mathbf{A}_i\mathbf{k}_i = \frac{1}{2} + F_{i+1} + \frac{1}{4}(F_f + F_m) \tag{4.10}$$

it follows that

$$|\mathbf{A}_{i+1}| = \begin{cases} 1 & i < n_{\mathcal{F}} \\ |\mathbf{A}_i|\left(\frac{1}{2} - \frac{1}{4}(F_f + F_m)\right) & i \geq n_{\mathcal{F}} \end{cases} \tag{4.11}$$

Now, since $\mathbf{A}$ is a p.d. matrix, it can be written as $\mathbf{A} = \mathbf{LL}'$ where $\mathbf{L}$ is a lower triangular matrix which can also be computed recursively as

$$\mathbf{L}_{i+1} = \begin{cases} \mathbf{I}_{i+1} & i < n_{\mathcal{F}} \\ \begin{pmatrix} \mathbf{L}_i & \mathbf{0} \\ \mathbf{k}_i'\mathbf{L}_i & \sqrt{\frac{1}{2} - \frac{1}{4}(F_f + F_m)} \end{pmatrix} & i \geq n_{\mathcal{F}} \end{cases} \tag{4.12}$$

and

$$\mathbf{L}_{i+1}^{-1} = \begin{cases} \mathbf{I}_{i+1} & i < n_{\mathcal{F}} \\ \begin{pmatrix} \mathbf{L}_i^{-1} & \mathbf{0} \\ \frac{-\mathbf{k}_i'}{\sqrt{\frac{1}{2}-\frac{1}{4}(F_f+F_m)}} & \frac{1}{\sqrt{\frac{1}{2}-\frac{1}{4}(F_f+F_m)}} \end{pmatrix} & i \geq n_{\mathcal{F}} \end{cases} \tag{4.13}$$

so that,

$$\mathbf{A}_{i+1}^{-1} = \begin{cases} \mathbf{I}_{i+1} & i < n_{\mathcal{F}} \\ \begin{pmatrix} \mathbf{A}_i^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\frac{1}{2}-\frac{1}{4}(F_f+F_m)} \begin{pmatrix} \mathbf{kk}' & -\mathbf{k}_i \\ -\mathbf{k}_i' & 1 \end{pmatrix} & i \geq n_{\mathcal{F}} \end{cases} \tag{4.14}$$

The expressions (4.12)–(4.14) are essentially Henderson's rules to compute the inverse of $\mathbf{A}$ reworked through partitioned matrix algebra [33, 34, 74].

4.3 The mixed model

When both oligogenic and polygenic contributions to the quantitative trait are considered, the distribution of the trait for each individual, as we saw before, is a discrete mixture of normals, and since the individuals are not independent of one another the whole model is a mixture of multivariate normals, with as many components as the number of feasible configurations for ζ, the vector of genotypes in the pedigree. The simplest mixed model, a model with only a major gene effect, additive polygenic effect, and independent environmental effects, can be written as

$$\mathbf{y} = \boldsymbol{\zeta} + \boldsymbol{\eta} + \mathbf{e} \tag{4.15}$$

where the conditional distribution of $\mathbf{y}$ given $\boldsymbol{\zeta}$ is $\mathcal{N}\left(\boldsymbol{\zeta}, \sigma_\eta^2\mathbf{A} + \sigma_e^2\mathbf{R}\right)$. That makes the density of $\mathbf{y}$

$$f(\mathbf{y}) = \sum_{\boldsymbol{\zeta}} \int_{\boldsymbol{\eta}} \Pr(\boldsymbol{\zeta})\, f(\mathbf{y} \mid \boldsymbol{\eta}, \boldsymbol{\zeta})\, dF(\boldsymbol{\eta}) \tag{4.16}$$

Because (4.16) is neither normal nor discrete, the mixed model cannot be peeled, and other methods are required for likelihood evaluation.

4.4 The Generalized Morton-MacLean Algorithm

We have seen that with nuclear family data, (4.16) can be efficiently evaluated by the Morton-MacLean algorithm [62, 63]. As was mentioned in chapter one, the Morton-MacLean algorithm can be extended to the case when one has a pedigree with quantitative trait records only for a nuclear family at the bottom of the pedigree, no matter how complex this pedigree may look. The key fact which supports the extension follows from the next argument. Let $\mathbf{y}_1$, and $\mathbf{y}_2$ be the records of one sibship and the records of any set of non-descendant relatives of the individuals with record in $\mathbf{y}_1$. Then the covariance matrix for $\mathbf{y}_1$ given $\mathbf{y}_2$ and $\boldsymbol{\zeta}$ can be written as

$$\boldsymbol{\Sigma}_{1\cdot 2} = \kappa_1 \mathbf{1}\mathbf{1}' + \kappa_2 \mathbf{I} \tag{4.17}$$

and because the conditional distribution of $\mathbf{y}_1$ given $\mathbf{y}_2$ and ζ is normal, the same argument used for (1.16)–(1.21) applies, i.e., there exists an a such that

$$\mathrm{f}(\mathbf{y}_1 \mid \mathbf{y}_2, \zeta) = \int_a \varphi_{\kappa_1}(a-\mu_a) \prod_{j\in C_{fm}} \sum_{\zeta_j} \Pr(\zeta_j \mid \zeta_f, \zeta_m)\, \varphi_{\kappa_2}(y_j - \zeta_j - a)\ \mathrm{d}a \tag{4.18}$$

for some appropriately chosen μ_a. Here f and m are the indices of the parents of the children with records in $\mathbf{y}_1$ and C_{fm} is the set of indices of those children.

To prove (4.17), note that the conditional variance of $\mathbf{y}_1$ and $\mathbf{y}_2$ given ζ can be written as

$$\begin{pmatrix} \Sigma_{11} & \mathbb{1}\mathbf{c}' \\ \mathbf{c}\mathbb{1}' & \Sigma_{22} \end{pmatrix} \tag{4.19}$$

for some vector $\mathbf{c}$, because any pair of individuals in a sibship have exactly the same 'degree' of relationship with any other individual in the pedigree. Therefore, the covariance between $\mathbf{y}_1$ and the kth individual in $\mathbf{y}_2$ has to be $c_k\mathbb{1}$. This implies that

$$\mathrm{Var}(\mathbf{y}_1 \mid \mathbf{y}_2, \zeta) = \Sigma_{11} - \mathbb{1}\mathbf{c}'\Sigma_{22}^{-1}\mathbf{c}\mathbb{1}' \tag{4.20}$$

which can be expressed in the form indicated by (4.17) because

$$\Sigma_{11} = \sigma_\eta^2 \mathbf{A}_{11} + \sigma_w^2 \mathbb{1}\mathbb{1}' + \sigma^2 \mathbf{I}$$

and

$$\mathbf{A}_{11} = \left(\tfrac{1}{2} + \tfrac{1}{4}(F_f + F_m) + F\right)\mathbb{1}\mathbb{1}' + \left(\tfrac{1}{2} - \tfrac{1}{4}(F_f + F_m)\right)\mathbf{I} \tag{4.21}$$

The generalization of the Morton-MacLean integral (4.18) may be useful in segregation studies, particularly when one is interested in a set of unrelated nuclear families with most of their data at the bottom of the pedigree. Also, because the evaluation f(y) is 'exact', this generalization provide a good reference point for checking the accuracy of some other computational techniques.

Because it is trivial to compute the joint probability for the parental ζs and because the conditional densities are normal, the evaluation of (4.18) is particularly easy when $\mathbf{y}_2$ contains nothing or the data for one or both parents. In these specific cases κ_1 and κ_2 in (4.17) can be determined as follows:

i). when both parents have data records

$$E(\mathbf{y}_1 \mid \zeta, y_f, y_m) = \zeta_2 + \mu_o \mathbb{1} \qquad (4.22)$$

where

$$\mu_o = \mathfrak{k}_0 \left(\mathfrak{k}_f \left(y_f - \zeta_f\right) + \mathfrak{k}_m \left(y_m - \zeta_m\right)\right)$$

$$\mathfrak{k}_f = \left(\hbar^{-2} + F_m\right)\left(\tfrac{1}{2} + \tfrac{1}{2}F_f + F\right) - 2F\left(\tfrac{1}{2} + \tfrac{1}{2}F_m + F\right)$$

$$\mathfrak{k}_m = \left(\hbar^{-2} + F_f\right)\left(\tfrac{1}{2} + \tfrac{1}{2}F_m + F\right) - 2F\left(\tfrac{1}{2} + \tfrac{1}{2}F_f + F\right)$$

$$\mathfrak{k}_0^{-1} = \left(\hbar^{-2} + F_f\right)\left(\hbar^{-2} + F_m\right) - 4F$$

and, as before

$$\hbar^2 = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_e^2}$$

Also,

$$\mathrm{Var}(\mathbf{y}_1 \mid \zeta, y_f, y_m) = \left(c_1\sigma_\eta^2 + \sigma_w^2\right)\mathbb{1}\mathbb{1}' + \kappa_2\mathbf{I} \qquad (4.23)$$

where

$$c_1 = \tfrac{1}{2} + \tfrac{1}{4}\left(F_f + F_m\right) + F - \mathfrak{t}_0\left(\mathfrak{t}_f\left(\tfrac{1}{2} + \tfrac{1}{2}F_f + F\right) + \mathfrak{t}_m\left(\tfrac{1}{2} + \tfrac{1}{2}F_m + F\right)\right)$$

$$\kappa_2 = \left(\tfrac{1}{2} - \tfrac{1}{4}\left(F_f + F_m\right)\right)\sigma_\eta^2 + \sigma^2$$

then $\kappa_1 = \left(c_1\sigma_\eta^2 + \sigma_w^2\right)$.

ii). when phenotypic data for one of the parents is missing, say y_f, then

$$\mathrm{E}(\mathbf{y}_1 \mid \zeta, y_m) = \zeta_2 + \mu_o'\mathbb{1} \qquad (4.24)$$

where

$$\mu_o' = \frac{\hbar^2}{2}\left(\frac{1 + F_m}{1 + \hbar^2 F_m}\right)$$

$$\mathrm{Var}(\mathbf{y}_1 \mid \zeta, y_m) = \left(c_1'\sigma_\eta^2 + \sigma_w^2\right)\mathbb{1}\mathbb{1}' + \kappa_2\mathbf{I} \qquad (4.25)$$

and

$$c_1' = \tfrac{1}{2} + \tfrac{1}{4}\left(F_f + F_m\right) + F - \frac{\hbar^2\left(1 + F_m\right)^2}{4\left(1 + \hbar^2 F_m\right)}$$

with $\kappa_1 = \left(c_1'\sigma_\eta^2 + \sigma_w^2\right)$.

iii). and, when both parents have missing data records, trivially

$$\mathrm{E}(\mathbf{y}_1 \mid \zeta) = \zeta_2 + \mu_o''\mathbb{1} \qquad (4.26)$$

with $\mu_o'' = 0$ and

$$\mathrm{Var}(\mathbf{y}_1 \mid \zeta) = \left(c_1''\sigma_\eta^2 + \sigma_w^2\right)\mathbb{1}\mathbb{1}' + \kappa_2\mathbf{I} \qquad (4.27)$$

where

$$c_1'' = \tfrac{1}{2} + \tfrac{1}{4}(F_f + F_m) + F$$

then $\kappa_1 = \left(c_1''\sigma_\eta^2 + \sigma_w^2\right)$.

Also, to simplify the computations, μ_a, the mean of a in (4.18), can be chosen as μ_o, μ_o', or μ_o'' depending on whether both parents have records, one is missing a record, or both have missing records, respectively. Note that if the model includes a constant mean, only μ_o, μ_o', and μ_o'' are affected.

Then, for a nuclear family at the bottom of the pedigree, f(y) can be evaluated as:

i). when both parents have data,

$$\mathrm{f}(\mathbf{y}) = \sum_{\zeta_f} \Pr(\zeta_f)\, \varphi_{\sigma_\eta^2+\sigma^2}(y_f - \zeta_f) \tag{4.28}$$

$$\sum_{\zeta_m} \Pr(\zeta_m \mid \zeta_f)\, \varphi_{\kappa_3}(y_m - \zeta_m - \kappa_4(y_f - \zeta_f))\ \ \mathrm{f}(\mathbf{y}_1 \mid y_f, y_m, \zeta_f, \zeta_m)$$

where

$$\kappa_3 = (1 + F_m - 2\kappa_4 F)\sigma_\eta^2 + \sigma^2$$

and

$$\kappa_4 = \frac{2F\sigma_\eta^2}{(1 + F_f)\sigma_\eta^2 + \sigma^2}$$

ii). when the data for one parent, say y_f, is missing,

$$\mathrm{f}(\mathbf{y}) = \sum_{\zeta_f} \Pr(\zeta_f) \sum_{\zeta_m} \Pr(\zeta_m \mid \zeta_f)\, \varphi_{\kappa_3'}(y_m - \zeta_m)\ \ \mathrm{f}(\mathbf{y}_1 \mid y_m, \zeta_f, \zeta_m) \tag{4.29}$$

where

$$\kappa_3' = (1 + F_m)\sigma_\eta^2 + \sigma^2$$

iii). and in the case in which data for both parents is missing,

$$f(y) = \sum_{\zeta_f} \Pr(\zeta_f) \sum_{\zeta_m} \Pr(\zeta_m \mid \zeta_f)\ f(y_1 \mid \zeta_f, \zeta_m)$$

This specific version of the Morton-MacLean algorithm allows for the incorporation of all of the information contained in the pedigree for the nuclear family. This provides a very efficient way of evaluating the likelihood of the family under study regardless of how complex the pedigree may look, and regardless of the size of the family.

4.5 Monte Carlo methods

In this section a number of Monte Carlo methods for approximating mixed model likelihoods are discussed. As mentioned, (4.16) is unpeelable. However, Thompson and Guo [84] have shown that for a given $\boldsymbol{\eta}$,

$$f(\mathbf{y} \mid \boldsymbol{\eta}) = \sum_{\zeta} f(\mathbf{y} \mid \boldsymbol{\eta}, \zeta)\ \Pr(\zeta) \qquad (4.30)$$

may be computed by Mendelian peeling, and that for a given ζ,

$$f(\mathbf{y} \mid \zeta) = \int_{\boldsymbol{\eta}} f(\mathbf{y} \mid \boldsymbol{\eta}, \zeta)\ dF(\boldsymbol{\eta}) \qquad (4.31)$$

can be computed by polygenic peeling, or as we have seen, by methods of linear algebra and multivariate analysis. These expressions allow us to rephrase (4.16) as

$$f(y) = \sum_{\zeta} f(y \mid \zeta) \Pr(\zeta) \tag{4.32}$$

$$= \int_{\eta} f(y \mid \eta)\, dF(\eta) \tag{4.33}$$

which can be rewritten as

$$f(y) = \mathop{E}_{\zeta}(f(y \mid \zeta)) \tag{4.34}$$

$$= \mathop{E}_{\eta}(f(y \mid \eta)) \tag{4.35}$$

While exact computation of (4.32) or (4.33) may be computationally prohibitive, the forms of (4.34) and (4.35) suggest that Monte Carlo methods may provide a successful solution via the sampling of ζ or η from an appropriate distribution.

4.5.1 Gene drop

The simplest Monte Carlo evaluation of (4.32) is to sample from $\Pr(\zeta)$ to obtain $\{\zeta^{(1)}, \ldots, \zeta^{(N)}\}$, N realizations of ζ, and, as (4.34) suggests, to estimate the value of the sum by

$$\mathop{\hat{E}}_{\zeta}(f(y \mid \zeta)) = \frac{1}{N}\sum_{i=1}^{N} f(y \mid \zeta^{(i)}) \tag{4.36}$$

To simulate a realization of ζ is straightforward. Just note that $\Pr(\zeta)$ can be factored as

$$\Pr(\zeta) = \prod_{j \in \mathcal{F}} \Pr(\zeta_j) \prod_{j \notin \mathcal{F}} \Pr(\zeta_j \mid \zeta_{f_j}, \zeta_{m_j}) \tag{4.37}$$

where, as before, $\mathcal{F}$ is the set of indices for founders and f_j and m_j are the indices of the parents of the jth individual. Therefore, it is only necessary to sample the genes to form the founder genotypes, and then drop them down sequentially through the pedigree following the rules of Mendelian segregation [58]. This sampling procedure is a special case of the *forward sampling* studied in probabilistic expert systems and Bayesian networks [42, 54]. Under the usual assumptions for a two-allele model, sampling the founder genes is just sampling from the Bernoulli distribution. With qualitative traits, genotype configurations obtained by dropping may be inconsistent with the observed data, and should be rejected. Fortunately, this is not an issue in the analysis of quantitative traits, since all of the configurations obtained by dropping are in principle consistent with the data. In addition to simplicity, one does not need to worry about issues of irreducibility and chain-mixing; those properties follow automatically with the gene drop. Perhaps due to bad experiences with simulating qualitative traits in the presence of data, it has been pointed out that most of the realizations of ζ will provide only an infinitesimal contribution to (4.36), and those few genotypic configurations that do provide substantial a contribution have minuscule probability of being realized, even in a large sample [69, 84]. However, since for the simulation of quantitative traits there is no rejection of simulated configurations due to genotype inconsistency with the observed phenotypes, this method may perform reasonably well provided that the simulated sample is large enough. In fact, it has been pointed out that the cases where gene dropping happens to be successful do not involve rejection [77, 78, 80].

Another possible sampling scheme may be backwards gene dropping, i.e., beginning with the individuals at the bottom of the pedigree, simulate up to the founders. This approach is not as simple as dropping genes down the pedigree and, for complex pedigrees with cross generational mating of relatives, may be very difficult to carry out.

At any rate, it is desirable to have procedures which use available information, and a naïve way to accomplish forward sampling in the presence of data is as follows.

For $i \in \mathcal{F}$, sample ζ_i from a distribution proportional to

$$\varphi_{\sigma_\eta^2+\sigma^2}(y_i - \zeta_i - \mu)\,\Pr(\zeta_i)$$

whenever y_i is observed, or from $\Pr(\zeta_i)$ if it is missing.

Sampling genotypes for non-founders is not quite as straightforward. One proposal is to sample from a distribution proportional to

$$\varphi_{\kappa_1+\kappa_2}(y_i - \zeta_i - \mu_a)\,\Pr(\zeta_i \mid \zeta_{f_i}, \zeta_{m_i})$$

when y_i is observed, or from $\Pr(\zeta_i \mid \zeta_{f_i}, \zeta_{m_i})$ if it is missing. Here κ_1, κ_2, and μ_a are determined taking into account all of the available phenotypic and genotypic information about the ancestors of y_i.

This forward sampling is a version of *posterior gene-dropping* [28] which does not use the conditional covariance structure for sibships, and which samples the children as if they were independent. It is tempting to instead sample simultaneously all full sibs from the appropriate distribution, i.e., to sample from a distribution proportional to

$$\int_a \varphi_{\kappa_1}(a - \mu_a) \prod_{j \in C_{fm}} \Pr(\zeta_j \mid \zeta_f, \zeta_m)\ \varphi_{\kappa_2}(y_j - \zeta_j - a)\ \mathrm{d}a$$

This approach is highly dependent upon the technique used to sample from the associated multinomial distribution, and in large problems the repeated evaluation of the integral may be computationally expensive. A more plausible scheme which avoids the above integral is to condition sequentially on the children, shuffling the indices of the sibs before sampling to prevent outcome dependency on the order in which the computations are carried out. We refer this approach as *forward gene dropping*. Details of its implementation are given in chapter 6.

Posterior gene-dropping has been proposed before in the context of the mixed model, usually for obtaining initial values for the Gibbs sampler [28, 29]. In those applications it is implicitly assumed that $\text{Var}(\eta) = \sigma_\eta^2 I$, i.e., the totality of the *parenté* information is thrown away, and it is perhaps due to this that the method has received little further attention. However, posterior gene dropping can be useful in conjunction with some of the ideas presented in the following subsection.

4.5.2 Importance sampling

To look for an alternative approximation, let $\Pr_*(\zeta)$ be another density for ζ with the same support as $\Pr(\zeta)$. Then note that f(y) can be expressed as

$$\begin{aligned} f(y) &= \sum_{\zeta} f(y \mid \zeta) \frac{\Pr(\zeta)}{\Pr_*(\zeta)} \Pr_*(\zeta) \\ &= \underset{\zeta}{E_*}\left(f(y \mid \zeta) \frac{\Pr(\zeta)}{\Pr_*(\zeta)} \right) \qquad (4.38) \end{aligned}$$

which suggests simulating from $\Pr_*(\cdot)$ and reweighting the realizations. The generating density $\Pr_*(\cdot)$ is called the importance density and sampling from it is called importance sampling. An account of the frequentist properties of importance sampling is given by Gamerman [24].

The effectiveness of importance sampling depends upon the choice of $\Pr_*(\cdot)$. Ideally, three additional properties for $\Pr_*(\cdot)$ are required:

i). realizations of ζ are easy to simulate from $\Pr_*(\cdot)$,

ii). $f(y \mid \zeta) \Pr(\zeta)$ has the "same shape" as $\Pr_*(\zeta)$, and

iii). at the realized values of ζ, $f(y \mid \zeta) \frac{\Pr(\zeta)}{\Pr_*(\zeta)}$ can be evaluated without a great deal of computational effort.

These desiderata are far from being easy to meet [75]. In fact, without the "same shape" requirement the realizations bear no relation to the observed data whatsoever. Therefore, as in simple gene drop, the majority of the realizations would make minuscule contributions to

$$\underset{\zeta}{\hat{\mathrm{E}}_*}\left(\mathrm{f}(\mathrm{y} \mid \zeta)\,\frac{\Pr(\zeta)}{\Pr_*(\zeta)}\right) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{f}(\mathrm{y} \mid \zeta^{(i)})\,\frac{\Pr(\zeta^{(i)})}{\Pr_*(\zeta^{(i)})} \tag{4.39}$$

if $\zeta^{(1)}, \ldots, \zeta^{(N)}$ were sampled from $\Pr_*(\cdot)$.

Importance sampling would work nicely if

$$\Pr_*(\zeta) \propto \mathrm{f}(\mathrm{y} \mid \zeta)\,\Pr(\zeta) \tag{4.40}$$

and since

$$\mathrm{f}(\mathrm{y} \mid \zeta)\,\Pr(\zeta) \propto \mathrm{f}(\zeta \mid \mathrm{y}) \tag{4.41}$$

this means that any sampling distribution that mimics the integrand must be close to the conditional distribution of ζ, the latent variables, given the data. Direct Monte Carlo sampling from exactly this distribution is pointless; if the proportionality constant were explicitly known, it would also be the value of the integral [83, 84]. Nevertheless, (4.41) suggests another importance sampling scheme,

$$\mathrm{f}(\mathrm{y}) = \underset{\zeta}{\mathrm{E}_*}\left(\mathrm{f}(\mathrm{y} \mid \zeta)\,\frac{\Pr(\zeta)}{\Pr_*(\zeta \mid \mathrm{y})} \,\middle|\, \mathrm{y}\right) \tag{4.42}$$

to be estimated as

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{f}(\mathrm{y} \mid \zeta^{(i)})\,\frac{\Pr(\zeta^{(i)})}{\Pr_*(\zeta^{(i)} \mid \mathrm{y})} \tag{4.43}$$

where now $\zeta^{(1)}, \ldots, \zeta^{(N)}$ come from $\Pr_*(\zeta \mid \mathrm{y})$. While this fixes the "same shape" requirement, it remains questionable as to how easy it is to sample from that distribution, and posterior gene dropping may be an attractive option. As we will see below, another alternative is to build a Markov chain having the desired stationary distribution by using the Metropolis-Hastings algorithm or the Gibbs sampler.

It has been pointed out that the previous formulation may be better suited to the evaluation of likelihood ratios [81, 84]. Suppose that $f(y)$ is parameterized by $\boldsymbol{\theta}$ and, instead of considering the likelihood $\ell(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(y)$, suppose that one is interested in the likelihood ratio $\lambda(\boldsymbol{\theta}, \boldsymbol{\theta}_o)$

$$\begin{aligned} \lambda(\boldsymbol{\theta}, \boldsymbol{\theta}_o) &= \frac{\ell(\boldsymbol{\theta})}{\ell(\boldsymbol{\theta}_o)} \\ &= \frac{1}{f_{\boldsymbol{\theta}_o}(y)} \underset{\zeta}{E_{\boldsymbol{\theta}_o}} \left(f_{\boldsymbol{\theta}}(y \mid \zeta) \frac{\Pr_{\boldsymbol{\theta}}(\zeta)}{f_{\boldsymbol{\theta}_o}(\zeta \mid y)} \,\middle|\, y \right) \end{aligned} \tag{4.44}$$

to be evaluated as

$$\begin{aligned} \hat{\lambda}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) &= \frac{1}{N} \sum_{i=1}^{N} \frac{f_{\boldsymbol{\theta}}(y \mid \zeta^{(i)}) \Pr_{\boldsymbol{\theta}}(\zeta^{(i)})}{f_{\boldsymbol{\theta}_o}(y)\, f_{\boldsymbol{\theta}_o}(\zeta^{(i)} \mid y)} \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{f_{\boldsymbol{\theta}}(y \mid \zeta^{(i)}) \Pr_{\boldsymbol{\theta}}(\zeta^{(i)})}{f_{\boldsymbol{\theta}_o}(y \mid \zeta^{(i)}) \Pr_{\boldsymbol{\theta}_o}(\zeta^{(i)})} \\ &= \frac{1}{N} \sum_{i=1}^{N} \frac{f_{\boldsymbol{\theta}}(y, \zeta^{(i)})}{f_{\boldsymbol{\theta}_o}(y, \zeta^{(i)})} \end{aligned} \tag{4.45}$$

where the $\zeta^{(1)}, \ldots, \zeta^{(N)}$ come from $f_{\boldsymbol{\theta}_o}(\zeta \mid y)$. Now, for the special case in which the difference between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_o$ is only in the segregation of the major gene, then $f_{\boldsymbol{\theta}}(y \mid \zeta^{(i)}) = f_{\boldsymbol{\theta}_o}(y \mid \zeta^{(i)})$ and (4.45) becomes

$$\hat{\lambda}(\boldsymbol{\theta}, \boldsymbol{\theta}_o) = \frac{1}{N} \sum_{i=1}^{N} \frac{\Pr_{\boldsymbol{\theta}}(\zeta^{(i)})}{\Pr_{\boldsymbol{\theta}_o}(\zeta^{(i)})} \tag{4.46}$$

That represents a substantial simplification since the explicit evaluation of the polygenic component is avoided.

Importance sampling also allows us to explore other possibilities: to evaluate the the density of y or the likelihood ratio thinking about f(y) as a multi-normal mixture of discrete distributions, i.e.,

$$\begin{aligned} f(y) &= \int_\eta \frac{f(y, \eta)}{f_*(\eta \mid y)}\, dF_*(\eta \mid y) \\ &= \underset{\eta}{E_*}\left(\frac{f(y, \eta)}{f_*(\eta \mid y)} \,\middle|\, y \right) \end{aligned} \qquad (4.47)$$

and

$$\begin{aligned} \lambda(\theta, \theta_o) &= \int_\eta \frac{f_\theta(y, \eta)}{f_{\theta_o}(y, \eta)}\, dF_{\theta_o}(\eta \mid y) \\ &= \underset{\eta}{E_{\theta_o}}\left(\frac{f_\theta(y, \eta)}{f_{\theta_o}(y, \eta)} \,\middle|\, y \right) \end{aligned} \qquad (4.48)$$

or, if one wants,

$$\begin{aligned} f(y) &= \sum_\zeta \int_\eta \frac{f(y, \zeta, \eta)}{f_*(\zeta, \eta \mid y)}\, dF_*(\zeta, \eta \mid y) \\ &= \underset{\zeta\eta}{E_*}\left(\frac{f(y, \zeta, \eta)}{f_*(\zeta, \eta \mid y)} \,\middle|\, y \right) \end{aligned} \qquad (4.49)$$

and

$$\begin{aligned} \lambda(\theta, \theta_o) &= \sum_\zeta \int_\eta \frac{f_\theta(y, \zeta, \eta)}{f_{\theta_o}(y, \zeta, \eta)}\, dF_{\theta_o}(\zeta, \eta \mid y) \\ &= \underset{\zeta\eta}{E_{\theta_o}}\left(\frac{f_\theta(y, \zeta, \eta)}{f_{\theta_o}(y, \zeta, \eta)} \,\middle|\, y \right) \end{aligned} \qquad (4.50)$$

It should be stressed that these approaches are likely to work well when θ_o is in a neighborhood of θ, but otherwise the behavior of the method may be erratic.

4.5.3 Metropolis-Hastings algorithm and the Gibbs sampler

The Metropolis-Hastings algorithm is a general tool to generate samples from some state space $\mathcal{S}$ on which a probability measure P is defined. The idea is to define a Markov chain having stationary distribution with density f(x) for $\mathbf{x} \in \mathcal{S}$, the density that one is interested in sampling from. The algorithm proceeds by simulating a candidate or proposal value z from a transition distribution $q(\cdot, \mathbf{x})$. At the next step, $\mathbf{x}_{t+1}$ is randomly assigned to be either z with probability $r(\mathbf{z}, \mathbf{x}_t)$, or $\mathbf{x}_t$ with probability $1 - r(\mathbf{z}, \mathbf{x}_t)$, where

$$r(\mathbf{z}, \mathbf{x}_t) = \min\left(\frac{\mathrm{f}(\mathbf{z})\, q(\mathbf{x}_t, \mathbf{z})}{\mathrm{f}(\mathbf{x}_t)\, q(\mathbf{z}, \mathbf{x}_t)}, 1\right)$$

is the acceptance probability. This chain has transition kernel

$$Q(\mathbf{x} \longrightarrow \mathbf{z}) = \begin{cases} r(\mathbf{z}, \mathbf{x})q(\mathbf{z}, \mathbf{x}) & \text{if } \mathbf{z} \neq \mathbf{x} \\ 1 - \sum_{\mathbf{z}' \neq \mathbf{x}} r(\mathbf{z}', \mathbf{x})q(\mathbf{z}', \mathbf{x}) & \text{if } \mathbf{z} = \mathbf{x} \end{cases}$$

Whenever $Q(\mathbf{x} \longrightarrow \mathbf{z})$ is aperiodic and irreducible, the ergodic theorem guarantees that

$$\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i) \xrightarrow{as} \mathrm{E}_{\mathrm{P}}(g(\mathbf{x})) \quad \text{as } N \longrightarrow \infty$$

and by choosing different transition densities $q(\cdot, \cdot)$, one obtains different Monte Carlo Markov chain algorithms, including the Gibbs sampler.

The Gibbs sampler is a particular version of the Metropolis-Hastings algorithm in which the elements of a vector of latent variables are updated one at each stage, without a rejection step. The most common implementation of the Gibbs sampler for quantitative trait data in human populations is due to Guo and Thompson [28]. Since their implementation of the

Gibbs sampler requires conditional independence, both $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ must be simulated, and if $\mathbf{R}$ is not diagonal (it typically includes a full-sibs shared environment component) the model must be reparameterized by splitting $\mathbf{e}$ as $\mathbf{e} = \mathbf{W}\boldsymbol{\varepsilon} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, each element of $\boldsymbol{\varepsilon}$, $\varepsilon_j \sim \mathcal{N}\left(0, \sigma^2_{c_j}\right)$, and $\mathbf{W}$ is a matrix relating sibship environmental effect to the individuals, so that $\boldsymbol{\varepsilon}$ needs to be simulated as well. Under the traditional scheme the sampling may be carried out sequentially as

- for $\boldsymbol{\zeta}$ sample each ζ_i from

$$\begin{aligned} \mathrm{f}(\zeta_i \mid \boldsymbol{\zeta}_{-i}, \boldsymbol{\eta}, \boldsymbol{\varepsilon}, \mathbf{y}) &= \Pr(\zeta_i \mid \boldsymbol{\zeta}_{v_i}, \eta_i, \varepsilon_{f_i m_i}, y_i) \\ &\propto \Pr(\zeta_i \mid \zeta_{f_i}, \zeta_{m_i})\, \mathrm{f}(y_i \mid \zeta_i, \eta_i, \varepsilon_{f_i m_i}) \prod_{\substack{k \in C_{ij} \\ j \in \mathcal{P}_i}} \Pr(\zeta_k \mid \zeta_i, \zeta_j) \end{aligned} \tag{4.51}$$

- for $\boldsymbol{\eta}$ sample each η_i from

$$\begin{aligned} \mathrm{f}(\eta_i \mid \boldsymbol{\zeta}, \boldsymbol{\eta}_{-i}, \boldsymbol{\varepsilon}, \mathbf{y}) &= \mathrm{f}(\eta_i \mid \zeta_i, \boldsymbol{\eta}_{v_i}, \varepsilon_{f_i m_i}, y_i) \\ &\propto \mathrm{f}(\eta_i \mid \eta_{f_i}, \eta_{m_i})\, \mathrm{f}(y_i \mid \zeta_i, \eta_i, \varepsilon_{f_i m_i}) \prod_{\substack{k \in C_{ij} \\ j \in \mathcal{P}_i}} \mathrm{f}(\eta_k \mid \eta_i, \eta_j) \end{aligned} \tag{4.52}$$

- for $\boldsymbol{\varepsilon}$ sample each ε_{fm} from

$$\mathrm{f}(\varepsilon_{fm} \mid \boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\varepsilon}_{-fm}, \mathbf{y}) = \mathrm{f}(\varepsilon_{fm} \mid \zeta_j, \eta_j, y_j, j \in C_{fm}) \tag{4.53}$$

where the subscript $-i$ indicates all individuals but i, C_{ij} is the index set for the children of marriage formed by the ith and jth individuals, and $\mathcal{P}_i$ is the index set for the partners of the ith individual. The subscript on the common environmental effect ε_{fm} specifies the two parents for the sibship, and v_i defines the neighborhood of the ith individual, which consists of the indices of its parents, progeny, and mates. The associated densities are

$$\mathrm{f}(y \mid \zeta, \eta, \varepsilon_{fm}) = \begin{cases} 1 & \text{if } y \text{ missing} \\ \varphi_{\sigma^2}(y - \mu - \zeta - \eta - \varepsilon_{fm}) & \text{otherwise} \end{cases}$$

$$\mathrm{f}(\eta \mid \eta_f, \eta_m) = \varphi_{\frac{1}{2}\sigma_\eta^2}(\eta - \tfrac{1}{2}(\eta_f + \eta_m))$$

$$\mathrm{f}(\varepsilon_{fm} \mid \zeta_j, \eta_j, y_j, j \in C_{fm}) = \varphi_{(1 - n_{fm}\mathfrak{h}^2)\sigma_w^2}\left(\mathfrak{h}^2 \sum_{j \in C_{fm}} (y_j - \mu - \zeta_j - \eta_j)\right)$$

where

$$\mathfrak{h}^2 = \frac{\sigma_w^2}{n_{fm}\sigma_w^2 + \sigma^2} \quad \text{and} \quad n_{fm} = |C_{fm}|$$

It can be shown that this approach uses an expression similar to (4.16) with an additional integration over a latent vector accounting for $\boldsymbol{\varepsilon}$, i.e., it is sampling from $\mathrm{f}(\boldsymbol{\zeta}, \boldsymbol{\eta}, \boldsymbol{\varepsilon} \mid \mathbf{y})$. It has been mentioned for the previous procedure that even if the Gibbs Markov chain is irreducible, it may show very poor mixing rates for the genotype states, resulting in difficulties in achieving convergence and in reaching states with low transition probabilities. To improve mixing rates an alternative may be to apply Gibbs sampling to sub-vectors, thereby treating those as a block, rather than by using a complete breakdown of the latent vector into its scalar components [41]. This suggests the following proposal for a modified Gibbs sampling scheme to be applied when only the children at the bottom of the pedigree have records.

The main idea is to use the random field structure of the pedigree by extending the scheme for forward sampling in the presence of data presented earlier to some sort of Gibbs sampler in blocks. More precisely, the proposal is to sample from $\Pr(\boldsymbol{\zeta} \mid \mathbf{y})$, taking blocks of full sibs simultaneously, i.e.,

- sample $\zeta_j;\ j \in \mathcal{F}$ with probability proportional to

$$\Pr(\zeta_j) \prod_{\substack{k\in C_{ij} \\ i\in\mathcal{P}_j}} \Pr(\zeta_k \mid \zeta_i, \zeta_j) \tag{4.54}$$

  and since this is computationally inexpensive, the set $\{\zeta_j;\ j \in \mathcal{F}\}$ can be sampled simultaneously as a block;

- for sampling $\{\zeta_j;\ j \in C_{fm}\}$ a plausible alternative may be to use component by component updating of the sibship-block $\{\zeta_j;\ j \in C_{fm}\}$, where each element is sampled with probability proportional to

$$\Pr\left(\zeta_j \mid \zeta_f, \zeta_m\right)\ \mathrm{f}(y_j \mid \zeta_j, \zeta_{-j}, \mathbf{y}_{-j}) \tag{4.55}$$

  where $\mathrm{f}(y_j \mid \zeta_j, \zeta_{-j}, \mathbf{y}_{-j})$ is an univariate normal distribution, $\mathcal{N}(\mu^c_j, \sigma^2_c)$ say, with

$$\mu^c_j = \zeta_j + \mu_a + \mathrm{g} \sum_{\substack{k\in C_{fm} \\ k\neq j}} (y_k - \zeta_k - \mu_a) \tag{4.56}$$

  and variance

$$\sigma^2_c = \kappa_2 + \kappa_1\Big(1 - (n_{fm} - 1)\,\mathrm{g}\Big) \tag{4.57}$$

  where

$$\mathrm{g} = \frac{\kappa_1}{(n_{fm} - 1)\,\kappa_1 + \kappa_2} \tag{4.58}$$

  The constants κ_1, κ_2, and μ_a are computed conditioning on all the available phenotypic and genotypic information outside of the sibship.

Chapter 5

Linkage analysis

This chapter is devoted to the discussion of segregation analysis in the context of linkage. The chapter begins with an introduction to the traditional approach to linkage through the analysis of the joint segregation of the putative trait and marker loci. It is noted that the methods of chapter four apply without modification to the joint distribution of trait and marker. Furthermore, under the usual parameterization, the vector of marker genotypes is ancillary with respect to quantitative trait and linkage parameters and, in the general case, will be at least partially ancillary. Certain advantages of conditioning are mentioned, without delving too deeply into the general debate on the merits of conditional versus unconditional inference. The chapter ends with the proposal of conditioning on the marker inheritance vector rather than the markers themselves, an approach that allows for highly polymorphic markers without further complication.

In its simplest form, linkage analysis consists of counting recombinants and non-recombinants, estimating the recombinant fraction, and testing whether or not this fraction is significantly less than one half [16]. In a more integral form, the aim of linkage analysis is to locate the genes contributing to a trait by analyzing the cosegregation of the trait with the genetic marker or markers within a pedigree. In this sense, the mathematical model for linkage is just an extension to the segregation analysis for a one locus model. However, the model can be represented in different ways. Let us first see the standard form. Consider a two loci scenario with a gene **q** and a marker **m**. A *haplotype* is the joint state given by $(\mathrm{q}^{(i)}, \mathrm{m}^{(i)})$, the pair of alleles that a person gets from the parent i. Hence, each individual has two haplotypes that together form the *ordered genotype*. The *joint genotype* or composite genotype, as it is also known, is the joint state of the two haplotypes disregarding order and parental origin. Define $\mathbf{g} = (\mathbf{q}, \mathbf{m})$, then we already know that

$$\mathrm{Pr}_\rho(\mathbf{g}) = \prod_{i \in \mathcal{F}} \mathrm{Pr}_\rho(g_i) \prod_{j \notin \mathcal{F}} \mathrm{Pr}_\rho(g_j \mid g_{f_j}, g_{m_j}) \tag{5.1}$$

The marginals $\mathrm{Pr}_\rho(g_i)$ are usually determined assuming Hardy-Weinberg equilibrium, and each $\mathrm{Pr}_\rho(g_j \mid g_{f_j}, g_{m_j})$ is entirely determined by the inheritance model. For example, if the individual j has haplotypes

$$\left(\left(\mathrm{q}_j^{(f_j)}, \mathrm{m}_j^{(f_j)}\right), \left(\mathrm{q}_j^{(m_j)}, \mathrm{m}_j^{(m_j)}\right)\right)$$

then $\left(\mathrm{q}_i^{(j)}, \mathrm{m}_i^{(j)}\right)$, the haplotype of the ith child of the jth individual would be

$$\left(\mathrm{q}_i^{(j)}, \mathrm{m}_i^{(j)}\right) = \begin{cases} \left(\mathrm{q}_j^{(f_j)}, \mathrm{m}_j^{(f_j)}\right) & \text{with probability} \quad \frac{1}{2}(1-\rho) \\ \left(\mathrm{q}_j^{(f_j)}, \mathrm{m}_j^{(m_j)}\right) & \text{with probability} \quad \frac{1}{2}\rho \\ \left(\mathrm{q}_j^{(m_j)}, \mathrm{m}_j^{(f_j)}\right) & \text{with probability} \quad \frac{1}{2}\rho \\ \left(\mathrm{q}_j^{(m_j)}, \mathrm{m}_j^{(m_j)}\right) & \text{with probability} \quad \frac{1}{2}(1-\rho) \end{cases}$$

where ρ is the recombinant fraction. Note that $j = f_i$ or $j = m_i$. It seems clear that the determination of $\mathrm{Pr}_\rho(g_i \mid g_{f_i}, g_{m_i})$ involves not only the parental genotypes but also information about the phase, i.e., which of the grandparents the alleles are coming from, no matter if the genotypes are treated as ordered or unordered. It is useful to note that ordered genotypes preserve knowledge of parental origin of each allele at each participating locus of the genotype and that ordering is slightly more detailed than ordering by phase. Unordered genotypes preserve knowledge of phase, but not origin [26].

It follows that when one deals with ordered genotypes, $\Pr_{\rho}(g_i \mid g_{f_i}, g_{m_i})$ can be computed as

$$\Pr_{\rho}(g_i \mid g_{f_i}, g_{m_i}) = \Pr_{\rho}\left(\mathrm{q}_i^{(f_i)}, \mathrm{m}_i^{(f_i)} \mid g_{f_i}\right) \Pr_{\rho}\left(\mathrm{q}_i^{(m_i)}, \mathrm{m}_i^{(m_i)} \mid g_{m_i}\right) \tag{5.2}$$

The traditional approach to the problem is to analyze the joint segregation of trait and marker(s). Therefore the methods of the previous section can be applied without modification, just by replacing $\mathbf{y}$ by $(\mathbf{y}, \mathbf{m})$ where $\mathbf{y}$ contains the trait phenotypes and $\mathbf{m}$ the marker observations. Adopting this approach, Knot and Haley [44] presented the method for computing the exact likelihood for nuclear families under the oligogenic model with familial environmental effects, i.e., the method for evaluating $f(\mathbf{y}, \mathbf{m})$ when

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\zeta} + \mathbf{e}$$

with $\mathbf{e}$ accounting for familial effects and residual variation.

When the data to be analyzed come from complex pedigrees with more than unrelated nuclear families, the Monte Carlo segregation equations are the natural choice to be employed for both oligogenic models with familial effects, and mixed models of inheritance. In general, the joint segregation equation can be written as

$$\begin{aligned} f(\mathbf{y}, \mathbf{m}) &= \sum_{\zeta} f(\mathbf{y}, \mathbf{m}, \zeta) \\ &= \sum_{\zeta} f(\mathbf{y}, \mathbf{m} \mid \zeta) \Pr(\zeta) \end{aligned} \tag{5.3}$$

When no pleiotropy is assumed, the previous expression becomes

$$f(y, m) = \sum_{\zeta} f(y \mid \zeta) \Pr(\zeta \mid m) \Pr(m) \tag{5.4}$$

$$= \sum_{\zeta} f(y \mid \zeta) \Pr(m \mid \zeta) \Pr(\zeta) \tag{5.5}$$

Also

$$f(y, m) = \frac{f(y, m, \zeta)}{\Pr(\zeta \mid y, m)} \tag{5.6}$$

which means that $\lambda(\theta, \theta_o)$ can be rewritten as

$$\begin{aligned}\lambda(\theta, \theta_o) &= \frac{f(y, m)}{f_{\theta_o}(y, m)} \\ &= \sum_{\zeta} \frac{f(y, m, \zeta)}{f_{\theta_o}(y, m)} \\ &= \sum_{\zeta} \frac{f(y, m, \zeta)}{f_{\theta_o}(y, m, \zeta)} \Pr_{\theta_o}(\zeta \mid y, m) \\ &= \sum_{\zeta} \frac{f_{\theta}(y \mid \zeta) \Pr_{\theta}(m \mid \zeta) \Pr_{\theta}(\zeta)}{f_{\theta_o}(y \mid \zeta) \Pr_{\theta_o}(m \mid \zeta) \Pr_{\theta_o}(\zeta)} \Pr_{\theta_o}(\zeta \mid y, m) \qquad (5.7)\end{aligned}$$

This or analogous expressions are given in different papers [28, 40, 82]. It is interesting to note that if θ contains only the set of parameters associated with the quantitative trait plus the linkage parameter, then

$$\Pr_{\theta}(m) = \Pr(m) \tag{5.8}$$

i.e., m is ancillary for θ, in which case (5.7) can be reduced to

$$\lambda(\theta, \theta_o) = \sum_{\zeta} \frac{f_{\theta}(y \mid \zeta) \Pr_{\theta}(\zeta \mid m)}{f_{\theta_o}(y \mid \zeta) \Pr_{\theta_o}(\zeta \mid m)} \Pr_{\theta_o}(\zeta \mid y, m) \tag{5.9}$$

In general, Pr(**m**) does not depend on the putative gene location or the putative trait effects, as long as there are no pleiotropic effects involved. Lange and Sobel [53] pointed this out in the context of qualitative trait analysis, and they concluded that inferences about linkage can be equally well based on either the conditional or unconditional likelihoods. Similar arguments can be applied to linkage inferences with quantitative trait data, since in the worst scenario, where Pr(**m**) is unknown, **m** would be at least partially ancillary for linkage and quantitative trait parameters. Even so, there is not a clear answer to the question of which of these approaches to the likelihood should be used. It has been mentioned that, in the long run, conditional inference may be less powerful than unconditional. However, conditioning, like sufficiency or invariance, leads to a reduction of the data, and as a result, it often leads to great simplification. One may argue that conditioning on an ancillary statistic is appropriate because it makes the inference more relevant to the situation in hand. It is accepted that likelihood without supplementing it with available ancillary information may result in anomalous inference [5, 11, 55]. We have seen that in the analysis of experimental crosses and in the IBD methods, conditional inference has been used without further consideration and it is widely accepted. The remainder of this chapter is devoted to the exposition of some ideas about how to compute the conditional likelihood in the context of the mixed model.

In some cases the choice of appropriate latent variables may also simplify computations. For example, nothing changes in our representation of the conditional distribution $f(\mathbf{y} \mid \zeta)$ if one considers ordered genotypes instead of unordered for the major gene, and to compute $\Pr(\zeta \mid \mathbf{m})$ is simpler for ordered genotypes. Another possibility is to change the representation of the linkage model and to describe the model in terms of haplotypes and *inheritance vectors* [49, 53] or in terms of haplotypes and vectors of *segregation indicators* [45, 81]. Both representations are completely equivalent. Introduced by Lander and Green [49], the idea of using inheritance vectors has recently been the focus of attention for linkage analysis of qualitative traits [38, 47, 48]. For a given haplotype the segregation indicators vector for non-founder individuals is a bitwise vector indicating the origin of each allele for each locus, conventionally, the indicator of grandmother origin, which means that, for each individual haplotype, the dimension of the segregation indicators vector is twice the number of loci. The inheritance vector is also an indicator of the origin of each allele; the difference is that there is a vector for each locus. Then, an inheritance vector has dimension two times the number of individuals being considered. Let ξ_q and ξ_m be the inheritance pattern vectors in the (q, m) system. The simplification comes by noticing that given the marker inheritance vectors, nothing in the remaining marker information adds anything about linkage, i.e,

$$\Pr_\rho(\zeta, \xi_q \mid \mathrm{m}, \xi_m) = \Pr_\rho(\zeta, \xi_q \mid \xi_m) \tag{5.10}$$

This is quite nice, because it implies that the marker can be very polymorphic without adding too much computational effort or complexity. In unconditional cosegregation analysis the computation grows exponentially with the degree of polymorphism. Moreover, since the computation of $\Pr(\zeta \mid \xi_q)$ does not involve any knowledge about the recombinant fraction, it can be verified that

$$\begin{aligned} \Pr_\rho(\zeta, \xi_q \mid \xi_m) &= \Pr(\zeta \mid \xi_q) \Pr_\rho(\xi_q \mid \xi_m) \\ &= \Pr(\zeta \mid \xi_q)\ \rho^{|\xi_q - \xi_m|}(1-\rho)^{1-|\xi_q-\xi_m|} \end{aligned} \tag{5.11}$$

no matter if the values for the putative trait genotypes, ζ, are taken as ordered or unordered. $|\xi_q - \xi_m|$ is the Hamming distance between ξ_q and ξ_m, i.e., the number of bits where the vectors differ.

Chapter 6

A Simulation Study

A generalization of the Morton–MacLean algorithm was described in chapter four which is appropriate for any pedigree in which the putative trait data are available only at the bottom. Because it is exact, apart from quadrature approximation, this is the method of choice for likelihood evaluation in such cases. Several Monte Carlo approaches to likelihood approximation were also discussed. In this chapter a simulation study is carried out to assess the performance of these various methods. It was noted in chapter five that the likelihood framework applies equally well to the conditional distribution of the putative trait data given markers. Furthermore, it was indicated that conditioning on either inheritance vectors or segregation indicator vectors is equivalent to conditioning on the marker data themselves. Therefore, the simulations, which consider a single marker, assume without loss of generality that the inheritance vectors are given.

Two particular cases were chosen for the simulations, each involving the mating of inbred parents. The first example has two nuclear families based in the pedigree shown in figure 4.1. The families were the product of the marriages 10 × 12 and 11 × 13 with 5 and 3 children, respectively. It was assumed that only the children had data records. The parameters used to generate the y's for this data set were:

$$\begin{aligned} \sigma^2_\eta &= 0.14 & \zeta' &= (8.7\ 7.4\ 2.3)' \\ \sigma^2_w &= 0.2 & \rho &= 0.2 \\ \sigma^2 &= 0.6 & p &= 0.3 \end{aligned}$$

where p and $1-p$ are the QTL allele frequencies; σ^2_η, σ^2_w, σ^2 are the polygenic, common environmental and error variances; ρ is the recombinant fraction; and ζ contains the QTL effects.

The data set for this case is shown in table 6.1.

parents	ξ_{m_i}	y_i
10 × 12	(0, 1)	7.62947
	(1, 0)	9.32674
	(1, 1)	10.0158
	(1, 0)	8.36626
	(0, 1)	3.21165
11 × 13	(0, 1)	9.44775
	(0, 0)	9.11278
	(1, 1)	9.62598

Table 6.1: Simulated data for case 1.

For the second case, an highly inbred family was considered, with the objective of challenging the forward gene dropping and Gibbs schemes. The pedigree for the parents is shown in table 6.2. and the data consists of records for seven children who are the product of the marriage 9 × 10. For this case two data sets were considered. The first was generated with

i	f_i	m_i
1		
2		
3		
4		
5	1	2
6	1	3
7	1	4
8	1	5
9	7	6
10	8	6

Table 6.2: Parental pedigree for case 2.

the following parameters:

$$\begin{aligned} \sigma^2_\eta &= 0.14 & \zeta' &= (9.7\ 7.4\ 2.3)' \\ \sigma^2_w &= 0.2 & \rho &= 0.2 \\ \sigma^2 &= 0.6 & p &= 0.53 \end{aligned}$$

The data set resulting from this choice of parameters is shown in the table 6.3.

The second set of parameters used was:

$$\begin{aligned} \sigma^2_\eta &= 0.03 & \zeta' &= (2.51\ 3.14\ 2.68)' \\ \sigma^2_w &= 0.01 & \rho &= 0.23 \\ \sigma^2 &= 0.32 & p &= 0.37 \end{aligned}$$

and the data set arising from this choice of parameters is shown in table 6.4.

i	$\breve{\zeta}_{m_i}$	y_i
11	(1, 1)	7.70787
12	(0, 1)	7.72817
13	(0, 1)	7.75753
14	(1, 1)	2.64921
15	(1, 0)	3.50281
16	(0, 0)	2.72604
17	(1, 1)	2.67313

Table 6.3: Simulated data for case 2, set 1.

Apart from a quadrature approximation the likelihood for the nuclear family case can be evaluated exactly by the extended Morton-MacLean algorithm, which was therefore chosen as an 'exact' basis for comparison. The case with two related families is evaluated by an hybrid Morton-MacLean integral with summation over the family with less individuals.

For each data set, the likelihood $\ell(\rho)$ was evaluated, holding everything else constant. In addition to the Morton-MacLean extended algorithm (the reference point), two approximations based in the conditional marginal densities for the nuclear families were also tried. The first one, applicable only to the data set with two sibships, was to ignore the polygenic covariance between the two sibships, taking the product of conditional marginal densities for the nuclear families as an approximation to the conditional density of the children given the parents set. The second one was to use the Morton-MacLean algorithm assuming independence of the parents just at the time of computing their joint probability, i.e, the joint probability of the parental ζs was computed from the allelic proportions without taking into account the pedigree information and then following the same steps as in the previous approximation. The Monte Carlo methods used to estimate the likelihood were plain gene dropping, forward sampling (posterior gene dropping) with sequential conditioning on each block of full sibs, and

i	ξ_{m_i}	y_i
11	(1, 1)	3.32901
12	(0, 1)	3.30548
13	(0, 1)	3.3047
14	(1, 1)	2.82521
15	(1, 0)	3.39917
16	(0, 0)	2.82104
17	(1, 1)	2.78196

Table 6.4: Simulated data for case 2, set 2.

Gibbs sampling with component-wise update for each sibship.

Finally, a simulation for ten nuclear families with moderate degree of inbreeding in one parent was tried. This example involved 67 children with record. In this case the likelihood $\ell(p)$ was obtained by the product of the familial likelihoods.

6.1 Computations

Computationally, the strategy followed was to eliminate as many latent variables as possible before starting the Monte Carlo evaluation. Therefore, the first step was to compute the joint distribution of the major genes of the parents. With an additive polygenic background, the *parenté* matrix and the joint distribution of the polygenes condenses all of the information contained in the pedigree. Note that if there are no marker records for the ancestors of the parents then the joint distribution in question can be computed from the *parenté* matrix. After this point, each technique for approximating the likelihood was tried. Since in our examples, parents and ancestors have no records, the expressions were simplified considerably.

For the Morton-MacLean method, each sibship conditional density was numerically evaluated through

$$\int_a \varphi_{\kappa_1}(a-\mu_a) \prod_{j \in C_{fm}} \sum_{\zeta_j} \Pr(\zeta_j \mid \zeta_f, \zeta_m\, \xi_{m_j})\ \varphi_{\kappa_2}(y_j - \zeta_j - a)\ \mathrm{d}a \tag{6.1}$$

The constants κ_1, κ_2, and μ_a are the ones defined for a nuclear family. After standardization, 16-point Legendre quadrature over five non-overlapping intervals was used in this work.

In the case of two related families, the first approximation to the exact Morton-MacLean method was obtained by multiplying the product of the two conditional familial densities by the joint probability of the parental ζs, and adding. The second approximation to the Morton-MacLean method throws away all of the pedigree information for the parents, and computes $\Pr(\zeta_f, \zeta_m)$ using allele proportions.

For the case with one nuclear family and observations y and ξ, the computations where carried out in the following way.

Morton-MacLean method. The procedure can be summarized in the following steps:

1. Compute κ_1, κ_2, and μ_a.
2. Compute $\Pr(\zeta_f, \zeta_m)$ exactly, and using the two approximations described above.
3. Evaluate by quadrature methods $\mathrm{f}(\mathbf{y} \mid \zeta_f, \zeta_m, \xi)$.
4. Evaluate $\mathrm{f}(\mathbf{y} \mid \xi) = \sum_{\zeta_f, \zeta_m} \Pr(\zeta_f, \zeta_m)\, \mathrm{f}(\mathbf{y} \mid \zeta_f, \zeta_m, \xi)$.

Fore the Monte Carlo methods, $\mathrm{f}(\mathbf{y} \mid \xi_m)$ was estimated as follows.

Plain gene dropping. The procedure consists of the following steps:

0. Compute the covariance matrix of y given ζ and also compute $\Pr(\zeta_f, \zeta_m)$.

1. Take a sample from $\Pr(\zeta_f, \zeta_m)$.

2. Sample each child's ζ_j from

$$\Pr\left(\zeta_j \mid \zeta_{f_j}, \zeta_{m_j}\ \xi_{m_j}\right)$$

3. Compute $f(\mathbf{y} \mid \zeta^{(i)})$.

4. Repeat steps 1 to 3 N times and then compute

$$\hat{f}(\mathbf{y} \mid \xi_m) = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{y} \mid \zeta^{(i)})$$

Forward sampling. The procedure consists of the following steps:

0. Compute the covariance matrix of $\mathbf{y}$ given ζ, $\Pr(\zeta_f, \zeta_m)$ and, using the parameter values in θ_o, the constants κ_1, κ_2, and μ_a.

1. Take a sample from $\Pr(\zeta_f, \zeta_m)$.

2. Take a random permutation of the children, and update $\zeta_{j_1}, \zeta_{j_2}, \ldots,$ by sequentially sampling them from a distribution proportional to

$$\Pr\left(\zeta_{j_k} \mid \zeta_f, \zeta_m, \xi_{m_{j_k}}\right)\ \varphi_{\sigma_k^2}(y_{j_k} - \mu_k)$$

where

$$\mu_k = \zeta_{j_k} + \mu_a + g_k \sum_{k'<k} (y_{j_{k'}} - \zeta_{j_{k'}} - \mu_a)$$

and

$$\sigma_k^2 = \kappa_2 + \kappa_1(1 - (k-1)\,g_k)$$

with

$$g_k = \frac{\kappa_1}{(k-1)\kappa_1 + \kappa_2}$$

3. Compute $f(y \mid \zeta^{(i)}) \dfrac{\Pr_{\theta}(\zeta^{(i)})}{\Pr_{\theta_0}(\zeta^{(i)} \mid y)}$.

4. Repeat steps 1 to 3 N times and then compute

$$\hat{f}(y \mid \xi_m) = \frac{1}{N} \sum_{i=1}^{N} f(y \mid \zeta^{(i)}) \frac{\Pr_{\theta}(\zeta^{(i)})}{\Pr_{\theta_0}(\zeta^{(i)} \mid y)}$$

Gibbs sampling. The steps for this procedure can be summarized as:

0. As in forward sampling.

1. Obtain a valid initial configuration by gene dropping.

2. Take a sample for (ζ_f, ζ_m) from a distribution proportional to

$$\Pr(\zeta_f, \zeta_m) \prod_{k=1}^{n} \Pr(\zeta_k \mid \zeta_f, \zeta_m, \xi_{m_k})$$

3. Take a random permutation of the children, and update $\zeta_{j_1}, \zeta_{j_2}, \ldots,$ by sampling each component from a distribution proportional to

$$\Pr\left(\zeta_{j_k} \mid \zeta_f, \zeta_m, \xi_{m_{j_k}}\right) \; \varphi_{\sigma_*^2}(y_{j_k} - \mu_k)$$

where

$$\mu_k = \zeta_{j_k} + \mu_a + g \sum_{k' \neq k} (y_{j_{k'}} - \zeta_{j_{k'}} - \mu_a)$$

and

$$\sigma_*^2 = \kappa_2 + \kappa_1 (1 - (n-1)\, g)$$

with

$$g = \frac{\kappa_1}{(n-1)\kappa_1 + \kappa_2}$$

4. Compute $f(y \mid \zeta^{(i)}) \frac{\Pr_\theta(\zeta^{(i)})}{\Pr_{\theta_0}(\zeta^{(i)} \mid y)}$.

5. Repeat steps 2 to 4 N times and then compute

$$\hat{f}(y \mid \xi_m) = \frac{1}{N} \sum_{i=1}^{N} f(y \mid \zeta^{(i)}) \frac{\Pr_\theta(\zeta^{(i)})}{\Pr_{\theta_0}(\zeta^{(i)} \mid y)}$$

6.2 Results

The results for the data sets considered are shown in appendix A, and are summarized in figures 6.1-6.5. Except for the ten families example, which was based in 30,000 iterations, each of the Monte Carlo estimates is based on 100,000 iterations.

Briefly, one can say that, in the particular case of the first data set, ignoring the polygenic covariance between the two sibships does not seem to substantially affect the computations. However, this may not be true in general. On the other hand, completely ignoring the pedigree may result in an over or under estimation of the likelihood. Even if the shape of the likelihood function is more or less the same as the reference likelihood based on the extended Morton-MacLean algorithm and the maxima of the two functions are attained in more or less the same neighborhood, this practice may lead to a considerable loss of power in the long run.

For all the cases considered, in the neighborhood of 0 the Gibbs scheme was disastrous. However, this method provide excelent estimates of the likelihood for non null ρ. On the other side, forward sampling yields 'good' estimates in many cases, but in some others behaves quite suspicious. Both methods perform better when the other model parameters are close to their true values (figure 6.1) and both methods err by more or less the same magnitude when the other parameters are far away from their true values (figure 6.2). This is a well known characteristic of importance sampling methods, and both schemes are used in this context. On the other hand, since $\rho = 0$ for practical proposes implies that the marker marks itself, the same sorts of reducibility issues will arise with the Gibbs scheme as are found in qualitative trait analysis [77, 78]. That is, when ρ is 0, the Gibbs sampling method may work in principle, but will take an exceedingly long time to traverse the space of latent variables. Also, when the sibships are highly correlated and the likelihood is quite flat, the Gibbs sampler does not yield good approximations to the likelihood in a reasonable number of iterations (figure 6.4).

In general, plain gene dropping does not work well unless ρ is null. Exceptions to this may occur when the parents are highly inbred and related, as is illustrated in figures 6.3 and 6.4.

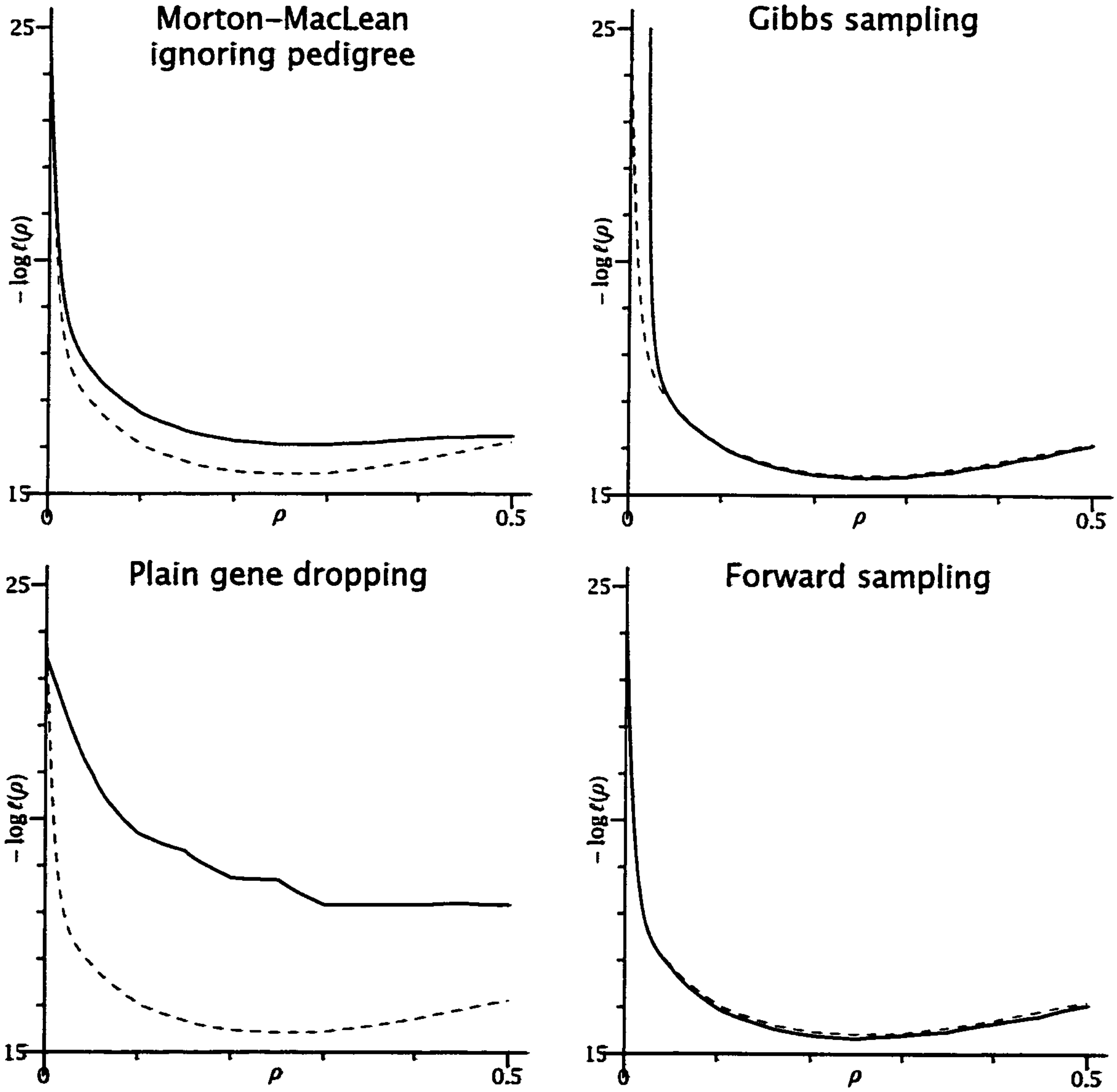

Figure 6.1: Likelihood for data set from table 6.1. (Dashed is $-\log \ell(\rho)$ computed by extended Morton–MacLean algorithm.)

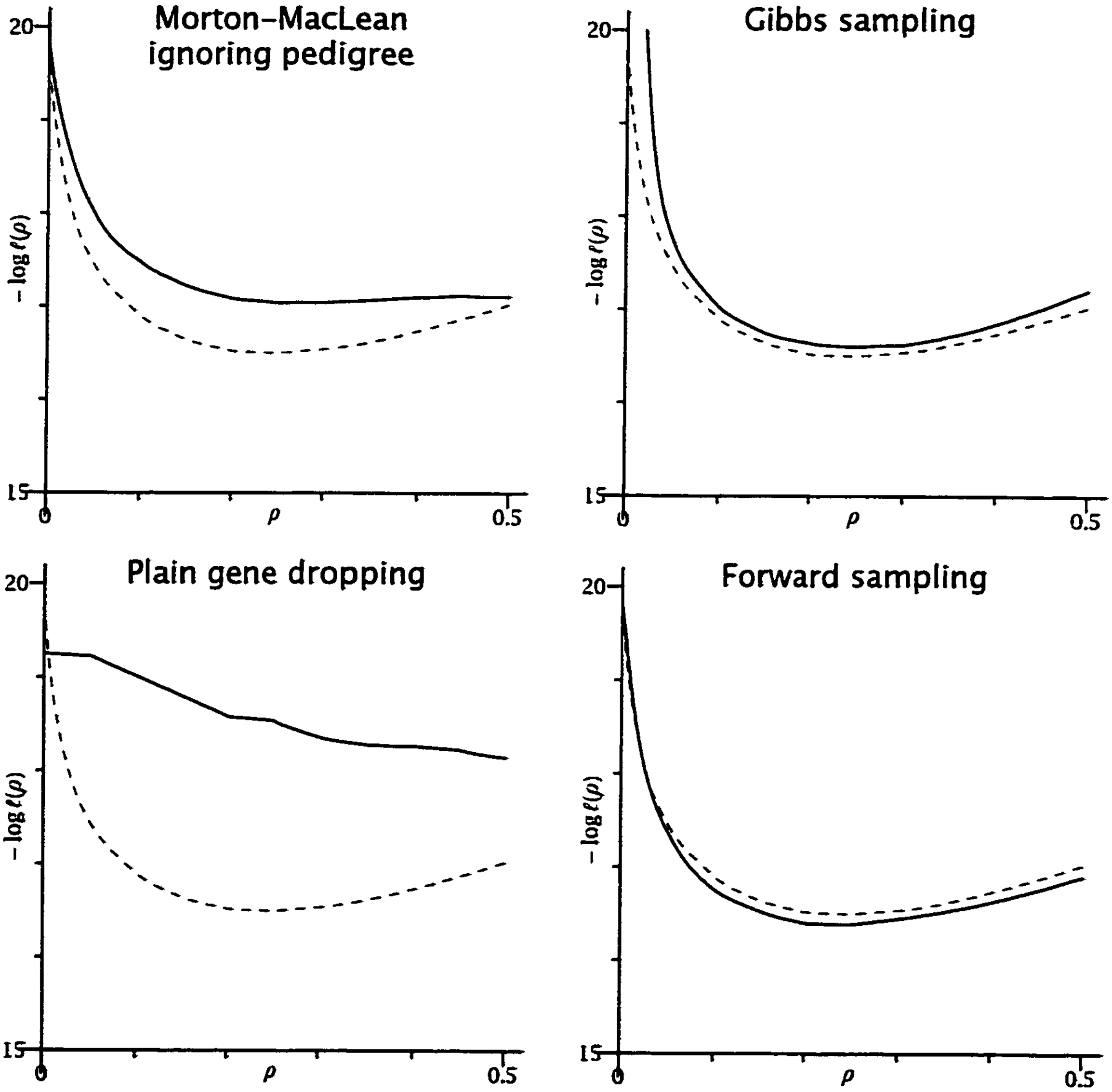

Figure 6.2: Likelihood for data set from table 6.1 with the variance of polygenic effects ten times the 'true' value. (Dashed is $-\log \ell(\rho)$ computed by Morton-MacLean algorithm.)

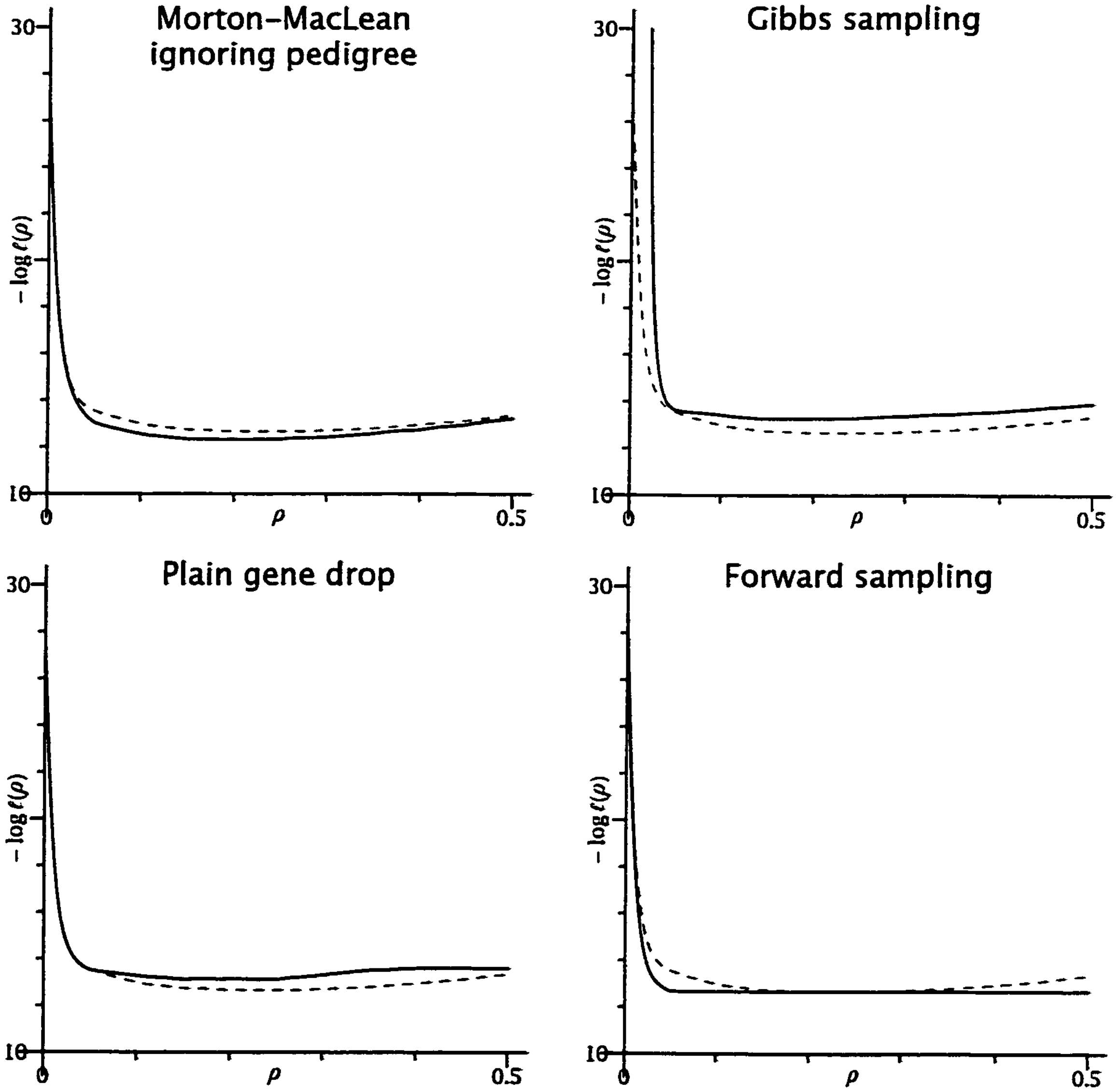

Figure 6.3: Likelihood for data set from table 6.3. (Dashed is $-\log \ell(\rho)$ computed by extended Morton-MacLean algorithm.)

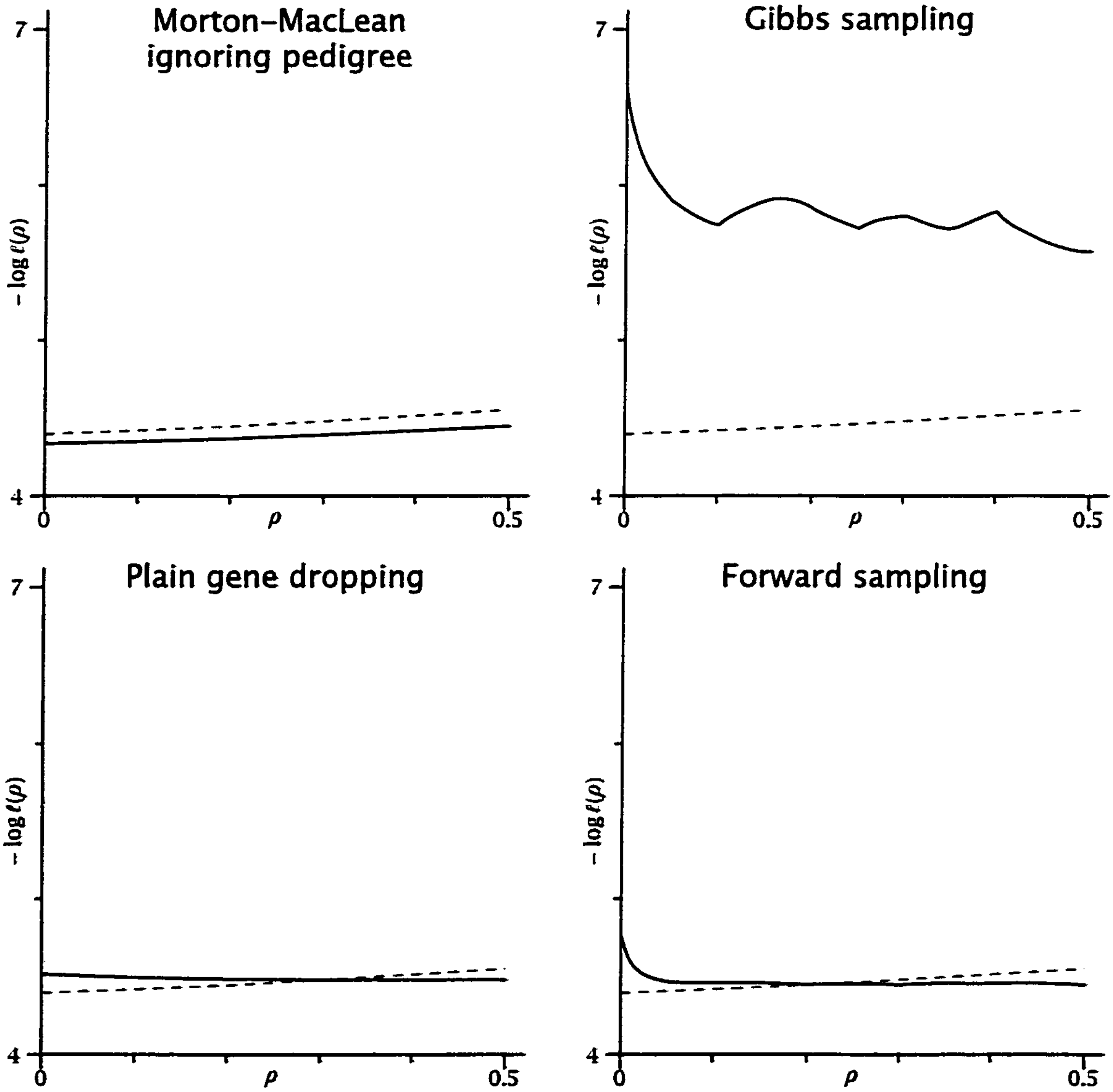

Figure 6.4: Likelihood for data set from table 6.4. (Dashed is $-\log \ell(\rho)$ computed by extended Morton-MacLean algorithm.)

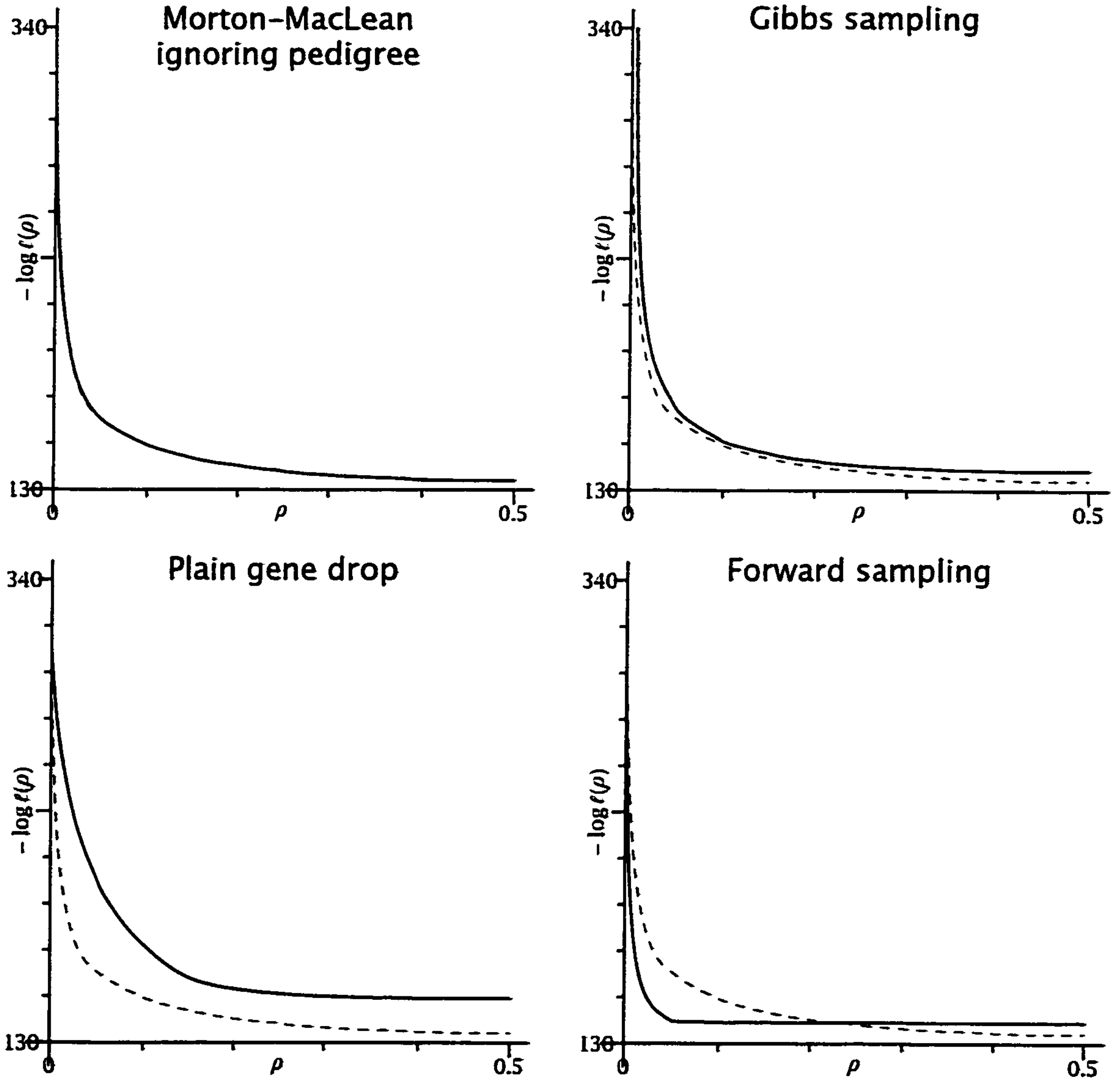

Figure 6.5: Comparison with a 'large' data set (ten nuclear families). (Dashed is the 'true' curve computed by Morton-MacLean integral.)

Chapter 7

Conclusion

The analysis of quantitative traits with an underlying mixed model of inheritance presents formidable problems in pedigree analysis. While for oligogenic inheritance models, algorithms and computer programs have proliferated in the last two decades, for mixed models the search for 'good' methods is still very active. The problem is that peeling methods, very successful for oligogenic models, require independence of the childrens' observations once the parental genotypes for the major gene are given. That is something that cannot be obtained in the presence of polygenic covariation or any other kind of covariation that breaks the required conditional independence. Of course, computations may be carried out by brute force. However, this will only be successful with a very few individuals. In 1972, Morton and MacLean introduced an integral to compute the likelihood under the mixed model of inheritance for nuclear families with unrelated parents. When their result is derived using exchangeability arguments, it becomes clear that the algorithm can be extended to allow any kind of relationship between the parents. This extension allows the computation of 'exact' likelihoods for families with phenotypic data at the bottom of the pedigree, taking into account all of the pedigree information. The usefulness of this extension may be fully appreciated in context of linkage analysis of quantitative traits with mixed models of inheritance and low penetrance, since in this situation the likelihood is expected to be flat enough to produce, with small or moderate sample sizes, a really bad scenario for inference. There are a number of other problems than can be studied with this tool, as for example assessing the effect that the misspecification of the pedigree may have on the inference and on power calculations.

When phenotypic data is available along a complex pedigree, however, one has to look for alternative methods of likelihood evaluation, or even to look for alternative methods of inference. Originally thought of as quick preliminary techniques, the methods based on identity by descent have become more and more popular because of their simplicity and because of the belief that those methods do not require model specification. However, as soon as one attempts to write down the likelihood for such models, it becomes clear that this function depends on the model and, even in the simplest cases, the computational problem is as substantial as that which was intended to be avoided.

Conditional inference is an alternative way to simplify the problem. In the particular case of linkage analysis we have seen that under a standard set up, the markers are ancillary for all of the parameters associated with the quantitative trait and the linkage parameter. When this argument is used together with the appropriate choice of latent variables, the problem may become simpler without loss of efficiency. Inheritance vectors have been proposed in the context of multipoint mapping of human diseases. In this thesis, we used them as a device to infer about linkage in quantitative traits and, from a small simulation study, the approach seems to work reasonably well.

As regards the numerical evaluation of the likelihood, two new Monte Carlo procedures were considered in this work: forward sampling in blocks of sibships and a Gibbs sampler blocking in a similar way. With these proposed procedures, only realizations of the major gene are simulated and the evaluation of the multivariate normal density is done through standard linear algebra methods instead of Monte Carlo integration. The idea behind these procedures is to use the structure of the pedigree to drop down genes in the presence of data (forward sampling) or to simulate realizations. These procedures turned out to yield more or less equivalent estimates of the likelihood in most of the situations tried. A possible explanation to this is that both procedures use the same amount of information about the target distribution. However, work needs to be done to compare the relative efficiency and accuracy of both methods.

So far, only the case of one quantitative trait and one marker was considered. Extension to interval and multipoint mapping problems may be useful, as well as to multivariate trait analysis.

Appendix A

ρ	Morton-MacLean	Morton-MacLean[1]	Morton-MacLean[2]	Gibbs sampling	Plain dropping	Forward sampling
0	24.4669	24.4678	24.5571	71.8094	23.4617	24.5123
0.05	16.8731	16.8722	17.5565	16.8502	20.9780	16.8219
0.1	16.0567	16.0559	16.7347	16.0219	19.6891	15.9757
0.15	15.6623	15.6615	16.3326	15.6281	19.3039	15.5826
0.2	15.4710	15.4701	16.1298	15.4321	18.7360	15.3843
0.25	15.4081	15.4070	16.0484	15.3662	18.6819	15.3295
0.3	15.4384	15.4371	16.0474	15.3955	18.1576	15.3814
0.35	15.5413	15.5398	16.0961	15.4960	18.1421	15.4726
0.4	15.7013	15.6996	16.1641	15.6566	18.1578	15.6416
0.45	15.9014	15.8995	16.2187	15.8506	18.1898	15.8029
0.5	16.1185	16.1165	16.2316	16.0747	18.1423	16.0471

[1] Extended integral only to compute the marginal for each nuclear family.

[2] Traditional algorithm taking parents as if they were unrelated.

Table A.1: $-\log \ell(\rho)$ for the data set 6.1 by different methods. (100,000 batches).

ρ	Morton- MacLean	Morton- MacLean [1]	Morton- MacLean [2]	Gibbs sampling	Plain dropping	Forward sampling
0	19.7937	19.8085	19.8421	50.4390	19.2526	19.8255
0.05	17.4469	17.4596	18.0093	17.7228	19.2254	17.3541
0.1	16.8927	16.9047	17.4759	17.0233	18.9814	16.7559
0.15	16.6335	16.6448	17.2135	16.7424	18.7866	16.5217
0.2	16.5172	16.5278	17.0835	16.6324	18.5707	16.3936
0.25	16.4919	16.5017	17.0334	16.5946	18.5399	16.3838
0.3	16.5318	16.5409	17.0335	16.6178	18.3541	16.4384
0.35	16.6206	16.6288	17.0605	16.7420	18.2746	16.4859
0.4	16.7438	16.7511	17.0932	16.8690	18.2655	16.6039
0.45	16.8857	16.8922	17.1121	17.0372	18.2273	16.7391
0.5	17.0274	17.0331	17.1038	17.2048	18.1401	16.8954

[1] Extended integral only to compute the marginal for each nuclear family.

[2] Traditional algorithm taking parents as if they were unrelated.

Table A.2: $-\log \ell(\rho)$ for the data set 6.1 computed with $\sigma_\eta^2 = 1.4$ and everything else as in table A.1. (100,000 batches).

	Morton-	Morton-	Gibbs	Plain	Forward
ρ	MacLean	MacLean[1]	sampling	dropping	sampling
0	27.4927	26.9226	46.5035	27.5135	28.8571
0.05	13.5222	13.0123	13.6349	13.5175	12.6520
0.1	12.9881	12.5343	13.4087	13.2534	12.6536
0.15	12.7565	12.3551	13.2566	13.1198	12.6479
0.2	12.6588	12.3065	13.2586	13.1277	12.6561
0.25	12.6433	12.3372	13.3055	13.1120	12.6708
0.3	12.6884	12.4255	13.3735	13.2990	12.6523
0.35	12.7837	12.5612	13.4660	13.4832	12.6632
0.4	12.9248	12.7398	13.5785	13.5651	12.6538
0.45	13.1102	12.9596	13.7276	13.6024	12.6595
0.5	13.3415	13.2215	13.9192	13.5774	12.6536

[1] Traditional method taking parents as if they were unrelated.

Table A.3: $-\log \ell(\rho)$ for the data set 6.3 by different methods. (100,000 batches)

ρ	Morton-MacLean	Morton-MacLean [1]	Gibbs sampling	Plain dropping	Forward sampling
0	4.39671	4.33775	6.63411	4.51477	4.76733
0.05	4.40668	4.34087	5.89215	4.50530	4.46277
0.1	4.41831	4.34671	5.73610	4.49402	4.45938
0.15	4.43140	4.35496	5.89733	4.48341	4.45744
0.2	4.44574	4.36524	5.84469	4.48262	4.44952
0.25	4.46111	4.37716	5.72067	4.47632	4.45350
0.3	4.47728	4.39036	5.79742	4.47654	4.44975
0.35	4.49406	4.40450	5.71413	4.47586	4.45606
0.4	4.51124	4.41930	5.82731	4.47769	4.45584
0.45	4.52866	4.43448	5.63339	4.47695	4.45651
0.5	4.54616	4.44985	5.57161	4.47696	4.44938

[1] Traditional method taking parents as if they were unrelated.

Table A.4: $-\log \ell(\rho)$ for the data set 6.4 by different methods. (100,000 batches)

ρ	Morton-MacLean	Morton-MacLean [1]	Gibbs sampling	Plain dropping	Forward sampling
0	326.825	325.853	447.087	309.177	337.700
0.05	162.067	161.929	166.161	203.195	139.696
0.1	150.316	150.154	152.070	179.894	139.582
0.15	144.387	144.197	146.387	159.075	139.537
0.2	140.823	140.599	143.146	155.971	139.498
0.25	138.487	138.225	141.307	153.310	139.502
0.3	136.872	136.576	140.092	152.492	139.502
0.35	135.729	135.412	139.464	149.910	139.485
0.4	134.937	134.613	138.847	149.587	139.460
0.45	134.439	134.122	138.501	150.155	139.413
0.5	134.212	133.911	138.576	150.068	139.603

[1] Traditional method taking parents as if they were unrelated.

Table A.5: $-\log \ell(\rho)$ for the simulation of ten nuclear families by different methods. (30,000 batches)

Bibliography

1. L. Almasy and J. Blangero. Multipoint quatitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62:1198-1221, 1998.

2. C. I. Almos and R. C. Elston. Robust methods for detection of genetic linkage for quatitative data from pedigrees. *Genetic Epidemiology*, 6:349-360, 1989. Errata 6:727.

3. C. I. Almos, R. C. Elston, A. F. Wilson, and J. E. Bailey-Wilson. A more powerful robust sib-pair test of linkage for quatitative traits. *Genetic Epidemiology*, 6:435-449, 1989.

4. C. I. Amos. Robust variance-components approach for assesing genetic linkage in pedigrees. *American Journal of Human Genetics*, 54:535-543, 1994.

5. D. Basu. Recovery of ancillary information. *Sankhyā*, 26:2-16, 1962.

6. W. C. Blackwelder and R. C. Elston. Power and robustness of sib-pair linkage tests and extensions to large sibships. *Communications in Statistics. Theory and Methods*, 15:449-484, 1982.

7. G. E. Bonney. Compound regresive models for family data. *Human Heredity*, 42:28-41, 1992.

8. G. E. Bonney and R. C. Elston. Integrals of products and multinomial mixtures. *Applied Mathematics and Computations*, 16:93-104, 1985.

9. C. Cannings, E. A. Thompson, and M. H. Skolnick. Probability functions on complex pedigrees. *Advances in Applied Probability*, 10:26-61, 1978.

10. Y. S. Chow and H. Teicher. *Probability Theory. Independence, Interchangeability, Matingales.* Springer-Verlag, New York, 1997.

11. D. R. Cox. The choice between ancillary statistics. *Journal of the Royal Statistical Society, Series B*, 33:251-255, 1971.

12. D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Process.* Springer-Verlag, New York, 1988.

13. R. W. Doerge and G. A. Churchill. Permutation tests for multiple loci affecting a quatitative character. *Genetics*, 142:285-294, 1996.

14. A. W. F. Edwards. *Foundations of Mathematical Genetics.* Cambridge University Press, New York, 1977.

15. R. C. Elston. Segregation analysis. In J. H. Mielke and M. H. Crawford, editors, *Current Developmdents in Anthropological Genetics*, volume 1, pages 327-354. Plenum Press, New York, 1980.

16. R. C. Elston. Methods of linkage analysis-and the assumptions undelying the. *American Journal of Human Genetics*, 63:931-934, 1998.

17. R. C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Human Heredity*, 21:523-542, 1971.

18. R. L. Fernando, C. Stricker, and R. C. Elston. An efficient algorithm to compute the posterior genotypic distribution for every memeber of a pedigree without loops. *Theoretical and Applied Genetics*, 87:89-93, 1993.

19. R. A. Fisher. The amount of information supplied by records of families as a function of the linkage in the population sampled. *Annals of Eugenics*, 6:66-70, 1934.

20. R. A. Fisher. The detection of linkage with 'dominant' abnormalities. *Annals of Eugenics*, 6:187-201, 1935.

21. D. W. Fulker and L. R. Cardon. A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics*, 54:1092-1103, 1994.

22. D. W. Fulker and S. S. Cherny. An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics*, 26(26):527-532, 1996.

23. D. W. Fulker, S. S. Cherny, and L. R. Cardon. Multipoint interval maping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics*, 56:1224-1233, 1995.

24. D. Gamerman. *Markov Chain Monte Carlo. Stochastic Simulation for Bayesian Inference.* Chapman & Hall, New York, 1997.

25. D. E. Goldgar. Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics*, 47:957-967, 1990.

26. T. M. Goradia, K. Lange, P. L. Miller, and P. M. Nadkarni. A general model for the genetic analysis of pedigree data. *Human Heredity*, 42:42-62, 1992.

27. S. W. Guo and E. A. Thompson. Monte Carlo estimation of variance component models for large complex pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology*, 8:171-189, 1991.

28. S. W. Guo and E. A. Thompson. A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics*, 51:1111-1126, 1992.

29. S. W. Guo and E. A. Thompson. Monte Carlo estimation of mixed models for large complex pedigrees. *Biomerics*, 50:417-432, 1994.

30. J. B. S. Haldane. Methods for the detection of autosomal linkage in man. *Annals of Eugenics*, 6:26-65, 1934.

31. J. B. S. Haldane and C. A. B. Smith. A new estimate of the linkage between the genes for haemophilia and colour-blindeness in man. *Annals of Eugenics*, 14:10-31, 1947.

32. J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2:3-19, 1972.

33. C. R. Henderson. A simple method for computing the inverse of a large numerator relationship matrix used in prediction of breeding values. *Biomerics*, 32:69-83, 1976.

34. H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23:53-60, 1981.

35. I. Heuch and F. H. F. Li. PEDIG -A computer program for calculation of genotype probabilities using phenotype information. *Clinical Genetics*, 3:501-504, 1972.

36. J. Hilden. GENEX -An algebraic approach to pedigree probability calculus. *Clinical Genetics*, 1:319-348, 1970.

37. A. P. Hill. Quantitative linkage: a statistical procedure for its detection and estimation. *Annals of Human Genetics*, 38:443-449, 1975.

38. R. M. Indury and R. C. Elston. A faster and more general hidden Markov model algorithm for multipoint likelihood calculations. *Human Heredity*, 47:197-202, 1997.

39. A. Jacquard. *The Genetic Structure of Populations.* Springer-Verlag, New York, 1974.

40. R. C. Jansen. A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics*, 142:305-311, 1996.

41. L. L. G. Janss, R. Thompson, and J. A. M. van Arendonk. Application of Gibbs sampling for inference in a mixed major gene-poligenic inheritance model in animal populations. *Theoretical and Applied Genetics*, 91:1137-1147, 1995.

42. C. S. Jensen, A. Kong, and U. Kjærulff. Blocking Gibbs sampling in very large probabilistic expert systems. Technical report, Institute for Electronic Systems, Department of Mathematics and Computer Science, Denmark, October 1993.

43. S. Karlin and U. Liberman. Measuring interference in the chiasma renewal formation process. *Advances in Applied Probability*, 15:471-487, 1983.

44. S. A. Knott and C. S. Haley. Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics*, 132:1211-1222, 1992.

45. A. Kong. Efficient methods for computing linkage likelihoods of recesive diseases in inbred pedigrees. *Genetic Epidemiology*, 8:81-103, 1991.

46. D. D. Kosambi. The estimation of map distance from recombination values. *Annals of Eugenics*, 12:172-175, 1944.

47. L. Kruglyak and E. S. Lander. Complete multipoint sib-pair analysis of qualitative and quatitative traits. *American Journal of Human Genetics*, 57:439-459, 1995.

48. L. Kruglyak and E. S. Lander. Faster multipoint linkage analysis using Fourier transforms. *Journal of Computational Biology*, 5:1-7, 1998.

49. E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proccedings of the National Academy of Sciences USA*, 84:2363-2367, 1987.

50. K. Lange. *Statistics for Biology and Health.* Springer-Verlag, New York, 1997.

51. K. Lange and M. Boehnke. Extensions to pedigree analysis. v. optimal calculation of Mendelian likelihoods. *Human Heredity*, 33:291-301, 1983.

52. K. Lange and R. C. Elston. Extension to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Human Heredity*, 25:95-105, 1975.

53. K. Lange and E. Sobel. A random walk method for computing genetic location scores. *American Journal of Human Genetics*, 49:1320-1334, 1991.

54. S. L. Lauritzen, A. P. David, B. N. Larsen, and H. G. Leimer. Independence properties of directed Markov fields. *Networks*, 20:491-505, 1990.

55. E. L. Lehmann. *Testing Statistical Hipotheses.* Springer-Verlag, New York, 2nd edition, 1997.

56. B. H. Liu. *Statistical Genomics.* CRC Press, Boca Raton, FL, 1998.

57. M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Inc., Sunderland, MA, 1998.

58. J. W. MacCuler, J. L. Vandeberg, B. Read, and O. A. Ryder. Pedigree analysis by computer simulation. *Zoo Biology*, 5:147-160, 1986.

59. G. Malécot. *Les Mathématiques de l'Herédité.* Mason, Paris, 1948.

60. N. E. Morton. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7:277-318, 1955.

61. N. E. Morton. LODs past and present. *Genetics*, 140:7-12, 1995.

62. N. E. Morton and C. J. MacLean. Analysis of family resembranse. III. Complex segregation of quatitative traits. *American Journal of Human Genetics*, 26:489-503, 1974.

63. N. E. Morton, D. C. Rao, and J. M. Lalouel. *Methods in Genetic Epidemiology*, volume 4 of *Contribution to Epidemiology and Biostatisitcs*. Karger, Basel, Switzerland, 1983.

64. J. R. O'Connell and D. E. Weeks. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics*, 11:402-408, 1995.

65. J. M. Olson. Some empirical properties of an all-relative-pairs linkage test. *Genetic Epidemiology*, 10:87-102, 1994.

66. J. M. Olson and E. M. Wijsman. Linkage between quatitative trait and marker loci: methods using all relative pairs. *Genetic Epidemiology*, 10:87-102, 93.

67. J. Ott. *Analysis of Human Genetic Linkage*. John Hopkings Univ. Press, Baltimore, MA, 1971.

68. J. Ott. Estimation of recombinant fraction in human pedifrees: efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics*, 26:588-597, 1974.

69. J. Ott. Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics*, 31:161-175, 1979.

70. G. P. Page, C. I. Amos, and E. Boerwinkle. The quantitative LOD score: test statistic and sample size for exclusion and linkage of quantitative traits in human sibships. *American Journal of Human Genetics*, 62:962-986, 1998.

71. L. S. Penrose. The detection of linkage in data which consist of pairs of brothers ans sisters of unespecified parentage. *Annals of Eugenics*, 6:133-138, 1935.

72. L. S. Penrose. Genetic analysis in graded human characters. *Annals of Eugenics*, 8:233-237, 1937.

73. L. M. Ploughman and M. Boehnke. Estimating the power of a proposed linkage study fo a complex genetic trait. *American Journal of Human Genetics*, 44:543-551, 1989.

74. R. L. Quass. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biomerics*, 58:949-953, 1976.

75. S. M. Ross. *A Course in Simulation.* Macmillan Publishing Co., New York, 1990.

76. N. J. Schork. Extended multipoint identity-by-descent analysis of human quantitative data traits: efficiency, power, and modeling considerations. *American Journal of Human Genetics*, 53:1306-1319, 1993.

77. N. Sheehan. Sampling genotypes on complex pedigrees with phenotypic constrains: the origin of the B allele among the polar esquimos. *IMA Journal of Mathematics Applied in Medicine and Biology*, 9:1-18, 1992.

78. N. Sheehan and A. Thomas. On irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biomerics*, 49:163-175, 1993.

79. C. A. B. Smith. The detection of linkage in human genetics. *Journal of the Royal Statistical Society, Series B*, 15:153-184, 1953.

80. A. Thomas. A comparison of an exact and a simulation method for calculating gene extintion probabilities in pedigrees. *Zoo Biology*, 9:259-274, 1990.

81. E. A. Thompson. Monte Carlo likelihood in the genetic mapping. *Statistical Science*, 9:355-366, 1994.

82. E. A. Thompson. Monte Carlo likelihood in the genetic mapping of complex traits. *Philosophical Transactions of the Royal Society of London. Series B*, 334:345-351, 1994.

83. E. A. Thompson. Likelihood and linkage: from Fisher to the future. *Annals of Statistics*, 24:449-465, 1996.

84. E. A. Thompson and S. W. Guo. Evaluation of likelihood ratios for complex genetic models. *IMA Journal of Mathematics Applied in Medicine and Biology*, 8:149-169, 1991.

85. D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York, 1985.

86. J. A. M. van Arendok, C. Smith, and B. W. Kennedy. Method to estimate genotypic probabilities at individual loci in farm livestock. *Theoretical and Applied Genetics*, 78:735-740, 1989.

87. P. M. Visscher, R. Thompson, and C. S. Haley. Confidence intervals in QTL mapping by bootstrap. *Genetics*, 143:1013-1020, 1996.

88. S. Xu. Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics*, 144:1951-1960, 1996.

89. S. Xu. Iteratively reweighed least squares mapping of quatitative trait loci. *Behavior Genetics*, 28:341-355, 1998.

90. S. Xu and W. R. Atchley. A random model approach to interval mapping of quantitative trait loci. *Genetics*, 141:1189-1197, 1995.