Thesis for the Degree of Licentiate of Philosophy

Two-Locus Affected Sib-Pair Identity By Descent Probabilities

Constraints, Parameterisation and Estimation

Olof Bengtsson

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Göteborg, Sweden 2001

Two-Locus Affected Sib-Pair Identity By Descent Probabilities
Constraints, Parameterisation and Estimation
Olof Bengtsson

ISSN 0347-2809/NO 2001:8
Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Chalmers University of Technology
Göteborg, Sweden 2001

Abstract

The identity by descent (IBD) probabilities are the parameters of primary interest in affected sib-pair (ASP) linkage analysis for disease gene mapping. Under realistic genetic models the ASP IBD probabilities at one locus fall into a region defined by two linear inequalities, the possible triangle. The power of ASP analysis may be increased by constrained estimation.

In this thesis known constraints on the joint ASP IBD probabilities of two unlinked disease loci are expanded to a set of inequalities that is the natural two-locus analogue of the triangle. Constraints on sex-specific IBD probabilities, distinguishing between maternal and paternal alleles, are proved under two genetic models. One model, which allows general mating type frequencies, has quite strict monotonicity assumptions on the penetrances. The other model includes the populations genetic assumptions Hardy-Weinberg equilibrium, random mating and linkage equilibrium, but no penetrance assumptions.

IBD probabilities satisfying the constraints may be represented in terms of a set of multinomial probabilities. It is demonstrated that IBD probabilities at loci linked to the disease loci conform to the constraints and are conveniently expressed using the derived parameterisation. Furthermore, the parameterisation facilitates maximum-likelihood estimation of the IBD probabilities subject to the constraints, by means of a double EM algorithm.

Keywords: Affected sib-pair, EM algorithm, gene mapping, IBD probabilities, imprinting, linkage analysis, triangle constraints, two-locus.

Acknowledgements

First of all, I want to thank my supervisor Professor Olle Nerman for his ideas and comments on my work, his encouragement and always friendly attitude.

I would also like to express my gratitude to Professor Terry Speed at The Walter and Eliza Hall Institute of Medical Research, Melbourne / University of California, Berkeley. He suggested the subject and the main ideas of this thesis one and a half years ago, and since then he has given me guidance, encouragement and inspiration, mostly in the form of rapid and helpful answers to my e-mails.

I am indebted to Dr Sandrine Dudoit at Stanford University, who has provided some useful unpublished notes and valuable comments on an early version of the manuscript.

I want to thank Dr Dragi Anevski for careful reading of the manuscript and lots of useful comments.

My years as a graduate student have included valuable practical experience of linkage analysis during a total of four months, financed by Astra Hässle AB (today AstraZeneca R&D Mölndal), spent at the Wellcome Trust Centre for Human Genetics in Oxford, where Dr Martin Farrall was my host and excellent teacher.

Finally I am grateful to the head of non-clinical statistics at AstraZeneca Mölndal, Tomas Morsing, who prompted me to start working in the fascinating field of statistical genetics and has always been supportive of my half-time arrangement.

This licentiate thesis has been financially supported by the National Network in Applied Mathematics, NTM and by the Swedish Natural Science research Council.

Olof Bengtsson
Göteborg, February 2001

Contents

Chapter 1

Introduction

The identity by descent (IBD) probabilities are the parameters of interest in affected sib-pair linkage analysis for localisation of disease genes. This thesis is devoted to constraints, representation and estimation of joint IBD probabilities at two unlinked genetic loci. An explanation of these concepts requires some genetic and statistical genetic terminology. However, before we introduce genetic concepts and linkage analysis in section 1.2 and 1.3 respectively, we give a brief and nontechnical background in section 1.1. An outline of the rest of the thesis is given in section 1.4.

1.1 Background

A major challenge of current biomedical research is to explain the genetic and biochemical mechanisms of common diseases, such as cardiovascular disease, diabetes and cancer. One approach in the early stages of this process is to roughly localise disease susceptibility genes in chromosome segments by linkage analysis, in a screen of the entire human genome. This requires data in the form of observations of genetic markers, with known positions, in families with affected individuals. A successful linkage analysis identifies regions in the genome that are not inherited independently of disease affection status, indicating physical proximity of the marker to a potential disease gene.

The diseases mentioned above are instances of complex diseases, that is they are not controlled by a single gene, but by the poorly understood interaction of several genes and environmental factors. Hence, nonparametric statistical methods, which do not require detailed specification of the genetic model, are appealing for complex traits. The by far most common approach for nonparametric linkage analysis is affected sib-pair analysis, aiming at identifying loci (positions) where affected siblings exhibit increased similarity. The analysis at a certain locus depends on the proportions of affected sib-pairs having inherited DNA originating from the same parental chromosome from none, one or both parents. These are the identity by descent (IBD) proportions, which are the subject of this thesis.

All IBD proportions are not compatible with realistic genetic models. It has been proved (see e.g. Holmans [41], Dudoit and Speed [21]) that the IBD proportions under general genetic models with one disease gene satisfy inequalities, known as the possible triangle or Holman's triangle. The power to detect disease genes can be improved by taking the triangle constraints into account in the statistical test procedure. Similar constraints have been derived by Cordell et al. [10] and Dudoit and Speed [21] for the joint IBD proportions at two loci, under a model with two disease genes located at separate chromosomes.

Derivation of a parameterisation of two-locus IBD proportions satisfying these constraints leads us to conjecturing additional constraints. The entire set of constraints, which is a natural two-locus analogue of the triangle, is subsequently proved under general genetic models. The representation facilitates maximum likelihood estimation of the two-locus IBD proportions subject to the constraints. We regard our representation promising for the development of new well-founded statistical tests of joint effects of two loci.

1.2 Genetic Concepts

As is well-known, all genetic information necessary for the development, maintenance and reproduction of an organism is encoded into DNA and stored in every cell. The genetic instructions are implemented when DNA is translated into proteins that are required for the structure, function, and regulation of cells, tissues, and organs. The pieces of DNA that code for a protein are called genes.

In humans, all the genetic material, the *genome*, is linearly arranged in 23 pairs of *homologous* chromosomes. Of these, 22 pairs are the autosomes and the remaining two chromosomes are the sex chromosomes, a pair of X chromosomes in females and one X and one Y in males. One of each pair of autosomes was received from the mother and the other from the father, the *maternal* and *paternal* chromosomes respectively.

The major part of the human genome is in fact occupied by non-coding DNA with no known function. Also the non-coding regions contain differences among individuals in a population. This variation is used in linkage analysis to keep track of inheritance. A *genetic marker* is a segment of DNA showing population variation, but having the same, known position in all individuals. It may or may not be part of a gene. The type of marker ordinarily used in modern linkage analysis, the microsatellite (see Vance and Othmane [75]), is composed by repeats of short DNA sequences, where differences in the number of repeats provide the population variation. Among the several thousands of detected microsatellites, standard sets of a few hundred markers, evenly distributed over the genome, are typically used in a genome screen. Blood samples from the individuals of the ascertained families may be analysed to determine the number of repeats in high-throughput, highly automated assays.

A *locus* (pl. *loci*) is defined as a specific position on a chromosome. The term is often used in a wider sense, not restricted to mean only an address in the genome, but also the DNA of gene or marker at that position. In this text, 'disease gene' and 'disease locus' will be used interchangeably. The different variants of the DNA sequence at a locus are called *alleles*. The pair of alleles at a locus constitutes a *genotype*, which is said to be *unordered* if we disregard whether an allele was inherited from the mother or the father. Observations on marker genotypes are unordered, but sometimes ordered genotypes can be inferred (see Fig. 1.1 (1)). Combinations of paternal and

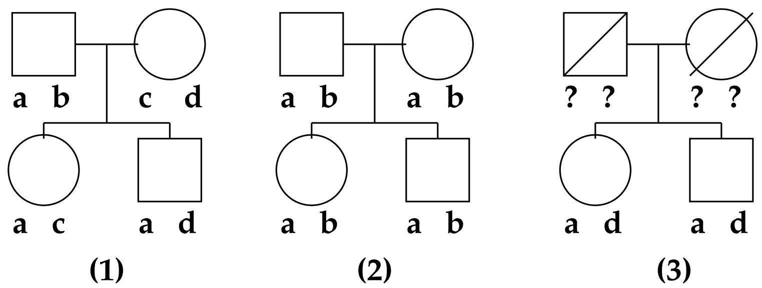

Figure 1.1: Pedigrees with observations of unordered genotypes for a marker in a sib-pair with parents. **(1)**. Ordered genotypes can be inferred in the sib-pair, which shares one allele IBD. **(2)** Zero or two alleles IBD. **(3)** Parental genotypes not observed (parents deceased). All IBD states possible.

maternal unordered genotypes form *mating types*[1].

Two alleles are said to be *identical by descent* (IBD) if they originate in the same ancestral allele, which in our case of sib-pairs (in families with no inbreeding) means that they are copies of the same parental allele. Hence a sib-pair can share 0, 1 or 2 alleles IBD. Three examples with different degree of marker information are given in figure (1.1).

One or more observable physical characteristics of an individual is called

[1]NB: With the mating type notion we keep track of the sex of the parents, but disregard the grandparental origin.

a *phenotype*. In this thesis we consider the qualitative phenotype 'affection status'.

1.2.1 Recombination, Linkage and Genetic Maps

Mendel's first law says that genetic traits are controlled by pairs of discrete entities, today called genes, and that a random gene is passed, or *segregated*, to the offspring. Mendel's second law states that genes controlling different characteristics are independently inherited. Today we know that it is chromosomes that segregate, and that genes on the same chromosome are not independently inherited. However, additional shuffling of the genetic material is accomplished by a random phenomenon called *crossover*, which enables linkage analysis. In meiosis, the cell division leading to the formation of *gametes* (egg or sperm cells), the chromosomes are duplicated and homologous chromosomes line up next to each other (Fig 1.2 (2)). The homologous chromosomes adhere to each other at random points, where crossover, that is exchange of DNA between maternal and paternal chromosomes, can occur (Fig 1.2 (3)). At fertilisation one of the four resulting gametes (Fig 1.2 (4)) is transmitted to the offspring.

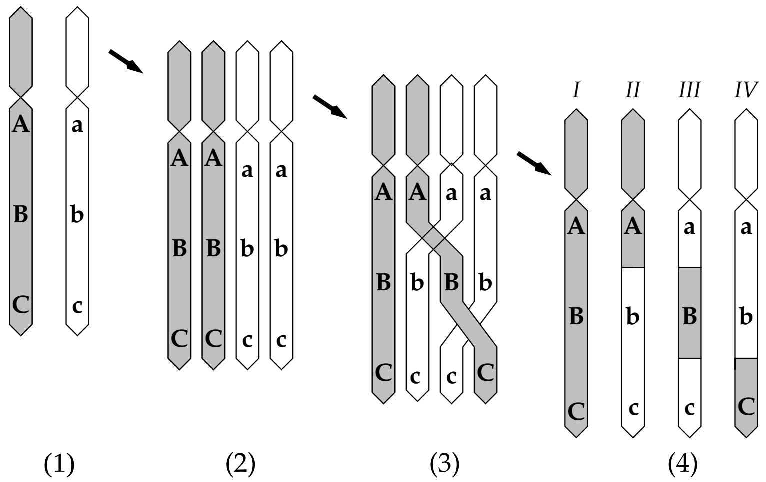

Figure 1.2: Schematic picture of crossover during meiosis. Two crossovers occur between a pair of chromosomes with three indicated loci, with alleles A;a, B;b and C;c, respectively.

When the alleles at two loci on the same chromosome were inherited from different parental chromosomes, a *recombination* has occurred. This requires an odd number of crossovers between the loci, e.g. gamete *III* in Fig 1.2 is scored as nonrecombinant if we only observe the first and third locus. Because of the independent segregation of chromosomes, the probability of recombination or the recombination fraction, θ, is $\frac{1}{2}$ for two loci on separate chromosomes. For two loci on the same chromosome, $\theta < \frac{1}{2}$ and we say that they are *linked*. Linkage analysis is based on the observation that the closer two loci are, the smaller is θ, which thus provides a measure of distance between loci.

As a consequence of the incidence of multiple crossovers between a pair of loci, the recombination fractions in adjacent intervals are not additive. Therefore, distances on genetic maps, showing the relative positions of markers and genes on chromosomes, are expressed in the expected number of crossovers, with unit *centiMorgan* (cM). One Morgan (100 cM) is the distance along which one crossover is expected per gamete per meiosis.

Translation between recombination fraction and genetic distance is complicated by *interference* (see e.g. Ott [62], Broman and Weber [8]), which is the dependence of the probability of crossover in one region on crossovers in other, nearby regions. Various *map functions* (Ott [62], Speed [68]), relate θ to the genetic distance d (cM), based on different stochastic models of crossover. Most widely used in human genetics is the Haldane map function,

$$\theta = \frac{1}{2}(1 - e^{-2d/100}), \tag{1.1}$$

which is derived assuming no interference and that locations of crossovers are determined by a Poisson process.

1.2.2 Genetic Models

We may consider the genetic model for the disease as constructed from three main components. First, a population genetic model describes the distribution of disease genotypes in the population. Second, one component describes how alleles at the trait loci are segregated from parents to offspring. Finally, we need a model for the conditional probability of disease phenotype given genotypes at the disease loci.

Population Genetic Models

Common population genetic assumptions are *Hardy-Weinberg equilibrium* (HWE), *random mating* (RM) and *linkage equilibrium* (LE). Under HWE the frequency of the genotype at a locus is the product of the frequencies of the alleles constituting the genotype. Under the additional assumption of RM, the frequency of a parental mating type is the product of the frequencies of the parents individual genotypes. LE between loci means independence between alleles at different loci. Thus, multilocus genotype frequencies under LE are simply the product of single-locus genotype frequencies. It follows that, under HWE, RM and LE, multi-locus mating type frequencies are easily calculated as products of allele frequencies. However, these are strong assumptions, that are hard to verify and it is unclear whether they are true for complex diseases.

Linkage disequilibrium (LD), that is population association between alleles, can be caused by close linkage, but also by for instance population admixture and stratification (see e.g. Sham [67] for explanation and further causes of LD). Since LD refers to association between specific alleles at two loci, which may or may not be linked, the more appropriate term *allelic association* is sometimes used.

We will use two models for the population genotype frequencies, one general model and one equilibrium model, where HWE, RM and LE are assumed.

Segregation Model

We assume that there is no segregation distortion, i.e. there is no preferential transmission of certain alleles and thus that Mendel's first law is in effect.

We derive constraints on joint IBD probabilities for unlinked loci. Hence, recombination fractions are not part of the genetic model of the disease. However, we also consider IBD probabilities at markers linked to the disease loci. These calculations are carried through with sex-averaged recombination fraction, but may be modified to include separate male and female recombination fractions, to adjust for the fact that crossing over is more frequent in females (Ott [62], Broman et al. [7]).

Link Between Genotype and Phenotype

In the final part of the genetic model we first make the crucial assumption of conditional independence of sibling phenotypes, given the genotypes at the disease loci. This assumption, often used but not mentioned, does not allow for covariance between sibling phenotypes caused by influence of environmental factors. Hence, it is not true for complex diseases, by definition. However, we need it to derive general constraints on the IBD proportions. It is unclear if this assumption may be slightly relaxed to allow for some forms of environmental covariance (see the discussion in section 7.3).

The conditional probability of a phenotype, given genotype, is called a *penetrance*. For a disease, genetically influenced by a single locus with m

alleles, $D_1, \dots, D_m$, define the penetrance

$$f_{ij} = \mathcal{P}(\text{Affected}|D_i D_j), \tag{1.2}$$

$i, j = 1, \dots, m$, where D_i is the allele inherited from the father and D_j is inherited from the mother. As examples, we consider diallelic models, i.e. $m = 2$, with a 'disease allele' d and a 'normal allele' n. The most commonly used penetrance functions are:

- *strict-recessive* model: $0 = f_{nn} = f_{nd} < f_{dd}$
- *strict-dominant* model: $0 = f_{nn} < f_{nd} = f_{dd}$.

In the first case the disease allele is recessive to the normal allele and two disease alleles are required for disease predisposition, whereas in the second case the disease allele is dominant and hides or modifies the effect of the normal allele. Often the prefix 'strict' is used more restrictively to denote the fully penetrant case, that is when $f_{dd} = 1$. These models may be slightly modified to (Day and Simons [15], Dudoit and Speed [21]):

- *quasi-recessive* model: $0 < f_{nn} = f_{nd} \leq f_{dd}$
- *quasi-dominant* model: $0 < f_{nn} \leq f_{nd} = f_{dd}$.

These are examples of functions that can model *phenocopies* or *sporadic cases*, that is individuals with a non-risk genotype that develop the disease, or a disorder which is phenotypically indistinguishable from the disease. Possible explanations for this phenomenon are that the disease is caused solely by environmental factors or by a gene at a different location (*locus heterogeneity*). Further penetrance functions suggested in the literature are:

- *additive* model: $f_{nd} = \frac{f_{nn}+f_{dd}}{2}$ (Motro and Thomson [57])
- *intermediate* model: $0 = f_{nn} \leq f_{nd} \leq f_{dd}$ (Spielman et al. [69])
- *overdominant* model: $0 = f_{nn} = f_{dd} < f_{nd}$ (Lynch and Walsh [54]).

The general penetrance notation (1.2) allows for *parental imprinting* or *parent-of-origin effect* (Lalande [49], Bartolomei and Tilghman [2]), which means that the effect of the same allele is different, depending on whether it was inherited from the father or from the mother.

In section 2.3 we discuss two-locus penetrance functions and assumptions regarding these, required to prove IBD constraints under the general population genetic model.

1.3 Linkage Analysis

Genetic linkage analysis aims at localising (disease) loci in chromosomes in relation to a map of markers. It is based on the recombination between marker loci and disease loci. Since the disease genotypes are unobserved, marker segregation in relation to the phenotype is studied. For a certain marker locus we want to test the null hypothesis of no linkage,

$$H_0 : \theta = \frac{1}{2}, \tag{1.3}$$

against the alternative hypothesis

$$H_A : \theta < \frac{1}{2},$$

where θ is the recombination fraction between the marker and the disease locus. In traditional parametric linkage analysis (Ott [62]), the likelihood, $L(\theta)$, of observed family data at a marker locus is calculated and maximised with respect to θ, assuming some disease model, including disease allele frequencies and penetrances. Usually the test is based on the logarithm (with base 10) of the likelihood ratio $L(\theta)/L(\frac{1}{2})$, which is known under the name of *LOD score* (logarithm of odds).

Linkage Analysis of Complex Diseases

Parametric linkage analysis has been successful in mapping genes for Mendelian diseases, which usually have low population prevalence and result from highly penetrant changes in a single major gene, for example cystic fibrosis and Huntington disease (Thomson and Esposito [73]). Now focus is on common, complex diseases, caused by interaction of environmental factors and multiple genes. Furthermore, in different populations there may be different disease predisposing alleles at the same locus (allelic heterogeneity), or at different loci (locus heterogeneity).

Consider as an example coronary artery disease and myocardial infarction (CAD/MI), where risk factors such as smoking, unhealthy diet and reduced physical activity are known to have a major impact. Several genes have been identified (Winkelmann and Hager [77]) as having an effect on some mechanism, e.g. the lipid metabolism, that is believed to be involved in CAD/MI. The risk of getting a heart attack is associated with quantitative variables, such as concentration of LDL-bound cholesterol, controlled by a network of biochemical pathways, in turn controlled by a network of related genes.

Since the underlying genetic mechanism is unknown, nonparametric methods based on IBD sharing, such as the affected sib-pair design, are in most cases preferred over the LOD score approach. However, some authors (Durner et al. [22], Xu et al. [78]) advocate the use of parametric models also for complex diseases.

Mapping projects are launched under the assumption that some major genes will have sufficiently large effects to be detected by linkage methods, but it is expected that most genes influencing common diseases have minor effects. Up to now, the success has been limited (Thomson and Esposito [73], Collins [9]) and attempts of independent replication of evidence for linkage have often failed (Lernmark and Ott [52], Moldin [56], Roberts et al. [66]). Hence, improvement and development of statistical methods for linkage data are still essential, even considering the increased use of alternative approaches for gene mapping (Horvath and Baur [42]).

1.3.1 Affected Sib-Pair Analysis

Affected sib-pairs (ASP), with or without parents, are the most common sampling units for a variety of statistical methods that are based on IBD sharing in sets of relatives (Hauser and Boehnke [38], Kruglyak et al. [47], Cordell et al. [11], Teng and Siegmund [72], Sham [67]). ASP analysis was first described by Penrose [63] as early as 1935, but its popularity has increased dramatically during the last 10-15 years, with regard to application as well as development of statistical methodology. It is based on the simple idea that if both members of an ASP have inherited the same disease allele, then IBD sharing at linked nearby markers will on the average also be higher than expected by chance.

Compared to large families with multiple affected individuals, it is relatively easy to find ASP's available for sampling. It has been shown (see e.g. Risch [65]) that, under some genetic models, the statistical power of affected relative pairs is greater compared to discordant pairs. Furthermore, analysis of affected individuals only[2] deals with the problem of incomplete penetrance, which may bias analysis that is based on both affected and unaffected individuals. This is particularly important for diseases with late age of onset and when affection status depends on underlying quantitative variation. Individuals may enter the state of being affected when some threshold is exceeded for one or more of the underlying variables, but in fact individuals may not be classified as truly 'unaffected'. The reverse problem of phenocopies, and that of genetic heterogeneity, of course remain.

ASP analysis is 'nonparametric' or 'model free' in the sense that we do not need to specify a genetic model for the trait, in terms of disease allele frequencies and penetrances (see e.g.the discussion of Farrall [26] and

[2] Parents, or additional unaffected siblings, are used only to determine the IBD distribution and their phenotypes are disregarded.

Greenberg et al. [33]). Like most nonparametric statistical methods, it still rests on a number of assumptions, as will be discussed in the following chapters.

In most practical applications IBD states can not be inferred with certainty from marker data, as was illustrated in Fig. 1.1 (2) and (3). This situation of incomplete IBD data is particularly evident for late-onset diseases where parental genotypes are not available. Computer programs (e.g. MapMaker/Sibs, cf. Kruglyak and Lander [48]) are routinely used to extract inheritance information at loci of interest, utilising the information from all linked markers. This requires the assumption of no interference, essentially the Poisson cross-over model resulting in the Haldane mapping function (1.1), and LE between markers, and between markers and the disease loci. Furthermore, if parental genotypes are not available, estimates of marker allele frequencies are needed. The marker information is captured in the *inheritance distribution*, that is a probability distribution over the possible IBD configurations (Lander and Green [50]), which is the basis for subsequent statistical inference.

Test Statistics

Let z_i, $i = 0, 1, 2$ be the probability that an ASP shares i alleles IBD at a marker locus. The expected IBD proportions at a locus that segregates independently of any disease locus are $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, see section 2.2, and the null hypothesis of no linkage (1.3) may be restated

$$H_0 : (z_0, z_1, z_2) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}). \tag{1.4}$$

Several authors have suggested test statistics of the form $n_2 + cn_1$, where n_1 and n_2 are the number of sib-pairs sharing one and two alleles IBD respectively, and c is a constant (e.g. $c = 0.5$: Blackwelder and Elston [3], Dudoit and Speed [20], $c = 0.355$: Whittemore and Tu [76], $c = 0.25$: Feingold and Siegmund [29]). These statistics, that may be extended to handle incomplete IBD data, were all suggested from different theoretical starting points. Their relative power advantages obviously depend on the genetic model. The 'mean IBD' statistic, $n_2 + 0.5n_1$, can for instance be derived as a score statistic in θ (Dudoit and Speed [20]) which is locally most powerful for θ close to zero (Cox and Hinkley [12]).

The preceding paragraph was merely intended to provide a few examples out of the sometimes overwhelming number of suggested ASP statistics. This may be further illustrated by the paper by Davis and Weeks [14], who compare the power of 23 statistics implemented in 10 software packages.

However, the most widely used statistic is perhaps the MLS (maximum LOD score) statistic (Risch [65], Kruglyak and Lander [48]), which is the logarithm (with base 10) of the likelihood ratio

$$\frac{L(\hat{z}_0, \hat{z}_1, \hat{z}_2)}{L(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})}. \tag{1.5}$$

Here, $(\hat{z}_0, \hat{z}_1, \hat{z}_2)$ are the maximum likelihood estimates of the the IBD probabilities, unconstrained or constrained to the possible triangle (to be described in section 1.3.2). The MLS statistic (multiplied by $2\ln(10)$) is under the null hypothesis (1.4) asymptotically chi-square distributed with 2 degrees of freedom in the unconstrained case and has a mixture of chi-square distributions with 0, 1 and 2 degrees of freedom when maximisation is restricted to the possible triangle (Holmans [41]).

To account for the multiple tests performed during a genome scan, Lander and Kruglyak [51] gave guidelines, providing the recommended pointwise significance level corresponding to a 5% genome-wide significance level, for various linkage statistics. This is the probability under the intersection of the null hypotheses of exceeding the corresponding threshold somewhere in the genome. The thresholds were based on the approximation of the null distribution of, for instance, the MLS statistic by an Ornstein-Uhlenbeck process (see e.g. Grimmet and Stirzaker [35]), assuming no interference, complete IBD information and complete resolution along the genome. By application of the theory of extreme values for this process, the genome-wide significance level corresponding to a certain threshold may be obtained given the the number of chromosome pairs (i.e. 23 in humans) and the total genome length (approximately 33 Morgan in humans). For more details, see Lander and Kruglyak [51], Feingold et al. [28] or Sham [67].

Discussion of existing approaches to statistical tests involving two loci is deferred until section 7.2.

Errors in Marker Data

Whatever particular ASP statistic we choose, our success in mapping genes will be highly dependent on the quality of the data, especially since we probably are searching for rather modest genetic effects. One important aspect of a linkage study is to reduce errors in the data (see e.g. Broman [5]).

There are two types of errors, pedigree errors and typing errors. For sib-pair data, pedigree errors include half-sibs or even unrelated individuals among the supposed sibling pairs. This may be due to nonpaternity, unreported adoption, errors in data entry or sample mix-ups. Several implementations (e.g. Boehnke and Cox [4], Göring and Ott [32]) of statistical methods for detection of false sib-pairs using marker data, not requiring parental genotypes, are available. With data from a full genome scan, these methods have high accuracy.

Genotyping errors occur when the observed genotypes are not the true underlying genotypes due to laboratory errors, incorrect data interpretation or entry. When hundreds of thousands, or even millions of marker genotypes are determined for a genome scan, these errors are inevitable. When parental genotypes are available, genotyping errors will be detected through Mendelian inconsistencies, such as children possessing alleles not present in the parents. O'Connel and Weeks [59] have made available a computer program that is guaranteed to find all Mendelian inconsistencies and identifies the individuals most likely to be in error.

For late-onset diseases, parental genotypes are rarely available and most genotyping errors will pass by undetected, which may have a potentially large effect on linkage power. However, recently Douglas et al. [18] described a method for detection of errors in sib-pairs with no parents. The probability of genotyping errors or mutations is calculated for each combination of sib-pair and marker, conditional on all marker data and an assumed genotype-error rate. To have a reasonable power, the method probably needs more tightly spaced markers than usually afforded in a genome scan (5-10 cM apart). Furthermore, with a higher marker density of 1-5 cM, e.g. in a follow-up study of interesting regions, follows increased power, but also a high false-negative rate and risk for bias in linkage statistics calculated from the cleaned data (Ewen et al. [23], Bahlo et al. [1]). This bias reflects the fact that errors are easier to detect for $IBD = 2$ than for $IBD = 0$.

1.3.2 The Possible Triangle

Realistic genetic models impose restrictions on the conditional IBD probabilities (z_0, z_1, z_2). Holmans [41] and Faraway [24] drew attention to the fact that the power of affected sib-pair analysis can be increased by restricting estimation of the IBD probabilities to the region known as the possible triangle or Holmans triangle. Under the assumptions of RM and HWE they showed that

$$z_1 \leq \frac{1}{2} \tag{1.6}$$

and

$$z_0 \leq \frac{1}{2} z_1. \tag{1.7}$$

Dudoit and Speed [21] studied conditional IBD probabilities for affected, unaffected and discordant sib-pairs without the assumption of RM and HWE. They derived the ASP triangle constraints (1.6) and (1.7) under general mating type frequencies with the following monotonicity assumption on the penetrances: For all $i, j, k, l = 1, \ldots, m$

$$(f_{ik} - f_{jk})(f_{il} - f_{jl}) \geq 0 \tag{1.8}$$

and

$$(f_{ik} - f_{il})(f_{jk} - f_{jl}) \geq 0. \tag{1.9}$$

This assumption essentially requires the existence of an ordering of the alleles such that an allele affects disease susceptibility in the same direction (or not at all) irrespective of the second allele in the genotype. The assumptions are satisfied by all models described in section 1.2.2, except the overdominant model. Dudoit and Speed [21] also suggested that the IBD

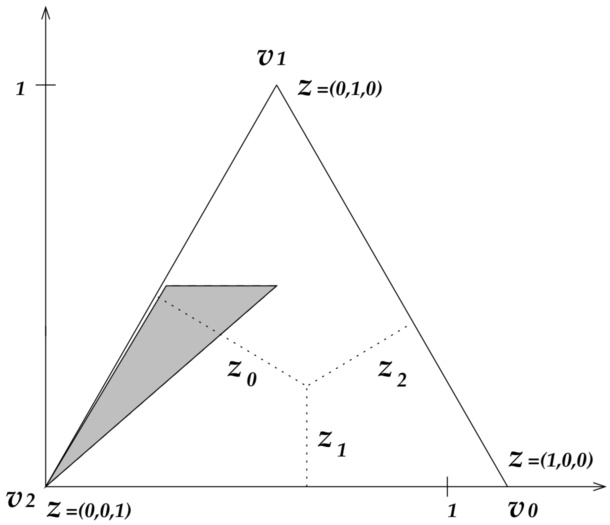

Figure 1.3: Barycentric representation of ASP IBD probabilities. The possible triangle is the shaded area.

probabilities may be visualised using a barycentric representation. If we let

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ z_2 \end{bmatrix},$$

then the points (x, y) will fall into an equilateral triangle (Fig. 1.3, where the shaded area is the possible triangle). Vertex v_i in the triangle corresponds to $z_i = 1$ and the vertical distance from a point to the side opposite to vertex v_i is equal to z_i.

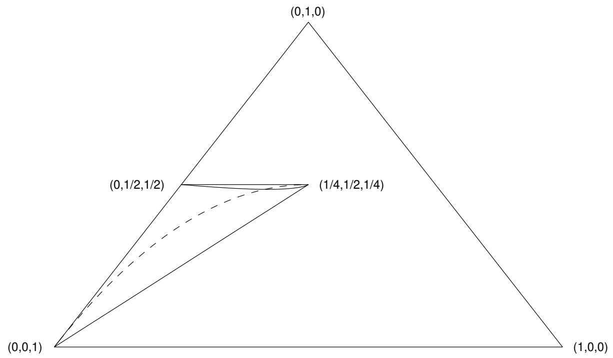

Figure 1.4: Barycentric plot of ASP IBD probabilities, where the smaller triangle is the possible triangle. The dashed (solid) curve is possible values of the IBD probabilities for a strict-recessive (strict-dominant) model, with disease allele frequency between 0 and 1, under RM and HWE.

Coverage

Consider two examples where we have calculated ASP IBD probabilities (as will be described in section 2.2.1) for the bi-allelic strict-recessive and strict-dominant models, described in section 1.2.2. The parental mating type frequencies were calculated from a disease allele frequency p, assuming RM and HWE. In the strict-recessive case the IBD probabilities are given by the simple expression

$$(z_0, z_1, z_2) = \left(\frac{p^2}{(1+p)^2}, \frac{2p}{(1+p)^2}, \frac{1}{(1+p)^2} \right). \tag{1.10}$$

Also for the dominant model the IBD probabilities are independent of f_{dd}, i.e. the probability of getting the disease for an individual with the predisposing genotype. For p between 0 and 1 the IBD probabilities of the strict-recessive model follow the dashed curve in Fig 1.4. The IBD proportions for the strict-dominant model lie on the solid curve, close to the upper side of the possible triangle ($z_1 = \frac{1}{2}$).

Dudoit and Speed [21] showed that the possible triangle is covered by a strict-recessive model with varying general mating type frequencies, that is all possible IBD probabilities within the triangle can be obtained under this model. With the curve of the RM/HWE case in Fig. 1.4 fresh in mind, it is clear that these population genetic assumptions potentially have a strong influence on the IBD probabilities. Furthermore, Dudoit and Speed [21] proved that the triangle is covered also under RM and HWE, when varying penetrance models are considered.

1.4 Outline

In chapter 2 we introduce additional concepts and notation, while giving more details of the genetic model discussed in section 1.2.2. In particular we define *sex-specific* and *nonsex-specific* IBD status. With the notion sex-specific IBD status we consider the parental origin of alleles shared IBD. The main part of the genetic model with no population genetic assumptions, i.e. with general mating type frequencies, is the penetrance monotonicity assumptions stated in section 2.3.2, where we also investigate if they are satisfied by various two-locus penetrance functions. The penetrance assumptions are required for the proof of two-locus constraints, analogous to the triangle constraints (1.6) and (1.7), under the general population genetic model.

In chapter 3 we present a first set of constraints on nonsex-specific IBD probabilities of two unlinked disease loci, proved by Cordell et al. [10] under HWE, RM and LE, and a corresponding set of sex-specific IBD probabilities proved by Dudoit and Speed [21] under the general population genetic model. Next, it is described how single locus IBD probabilities in the possible triangle may be expressed in terms of three parameters that are unconstrained, except being positive and summing to one. This parameterisation is extended to the two-locus case as the square, using Kronecker product, of the matrix defining the single locus representation. However, this requires new constraints on the IBD probabilities.

With the sex-specific constraints derived by Dudoit and Speed [21] as starting point, we conjecture a set of sixteen nonredundant linear inequalities, which are derived under the general population genetic model and the penetrance monotonicity assumptions. From this set follow nine nonsex-specific inequalities, including those derived by Cordell et al. [10] and four additional inequalities. In fact the representation of two-locus IBD probabilities used by Cordell et al. [10] (section 4.2.3), implies all nine constraints, but all of them were not explicitly stated. Furthermore, we prove that the parameter space is covered under a jointly recessive two-locus penetrance model with varying general parental mating type frequencies. Hence, the derived set of constraints is the tightest possible under the given model. Finally, we discuss how the parameters of the representation may be interpreted.

In chapter 4 the two-locus constraints are proved under RM, HWE and LE, without restrictions on the penetrances. First the proof is obtained by expression of the IBD probabilities in terms of genetic variance components, which was the approach used by Cordell et al. [10]. This gives an interpretation of the parameters of the general representation from chapter 3, under RM, HWE and LE. Second, the constraints are proved by a more direct approach.

In chapter 5 the expected IBD probabilities at two unlinked loci, for instance marker loci, linked to one disease locus each, are expressed in terms of the recombination fractions and the IBD probabilities at the disease loci. It is shown that the IBD probabilities at the markers also satisfy the constraints.

In chapter 6 we describe how the two-locus IBD probabilities may be estimated from marker data by the maximum likelihood method, using a double EM algorithm facilitated by our representation. The various assumptions needed for this procedure in practical situations with incomplete IBD data are discussed.

Naturally, the ultimate goal of these efforts is the invention and evaluation of new statistical tests for detection of disease genes and interaction between genes. However, this is outside the scope of this thesis, but in the final chapter we indicate the use of the parameterisation for statistical testing, with reference to existing single- and two-locus methods. Finally, various remaining questions and open problems are listed.

Chapter 2

The Genetic Model

In this chapter we present additional terminology and notation, while describing the genetic model discussed in section 1.2.2, now in more detail. The first section deals with the population genetic model and in the second section inheritance vectors are introduced in connection to the segregation model. The main part is section 2.3 where we discuss two-locus penetrances, define assumptions on these and investigate to what extent the assumptions are satisfied by various penetrance functions.

2.1 Parental Mating Type Frequencies

Let $\mathcal{D}^1$ and $\mathcal{D}^2$ denote the two disease loci, where locus $\mathcal{D}^l$ has m_l alleles, $D_1^l, \ldots, D_{m_l}^l$, $l = 1, 2$. However, when discussing models with two alleles per loci we will diverge from this notation and represent the two alleles at a locus by n and d to make the description more visually clear.

Let $pg^l = (pg_1^l, pg_2^l, pg_3^l, pg_4^l)$ be the *ordered parental genotype* at disease locus D^l, where pg_i^l is the allele on parental chromosome labeled i, with $i = 1, 2$ ($i = 3, 4$) corresponding to the paternal (maternal) chromosomes. For a locus with m alleles there are $m^2 \times m^2$ ordered genotypes. These can be arranged into *parental mating types*, by grouping genotypes that are alike when alleles 1 and 2 and/or 3 and 4 are permuted. Thus, as remarked earlier, mating types distinguish the sex of the parents, but not the sex of the grandparents. As an example, for a diallelic locus there are nine mating types (Table 2.1). In general the number of mating types is equal to $(m(m+1)/2)^2$ for a locus with m alleles. Let mt^l be the parental mating type at locus $\mathcal{D}^l$.

Now consider two diallelic loci, with alleles n^1, d^1 and n^2, d^2 respectively. If, for example, the parental mating types at both loci, mt^1 and mt^2, corre-

Mating type, mt	Genotypes, pg
1	$nn \times nn$
2	$nn \times nd, nn \times dn$
3	$nd \times nn, dn \times nn$
4	$nn \times dd$
5	$dd \times nn$
6	$nd \times nd, nd \times dn$ $dn \times nd, dn \times dn$
7	$nd \times dd, dn \times dd$
8	$dd \times nd, dd \times dn$
9	$dd \times dd$

Table 2.1: Parental mating types and ordered parental genotypes for a diallelic locus.

spond to mating type 2 in table 2.1, then there are four possible two-locus parental genotypes:

$$n^1n^2/n^1n^2 \times n^1n^2/d^1d^2, \tag{2.1}$$
$$n^1n^2/n^1n^2 \times n^1d^2/d^1n^2, \tag{2.2}$$
$$n^1n^2/n^1n^2 \times d^1n^2/n^1d^2, \tag{2.3}$$
$$n^1n^2/n^1n^2 \times d^1d^2/n^1n^2. \tag{2.4}$$

With this notation the *phase* of the genotypes is indicated, that is in the right (i.e. maternal) part of genotype (2.1), n^1n^2 is inherited from one grandparent and d^1d^2 from the other. Consequently, if the loci are linked, n^1 and n^2 reside on one chromosome and d^1 and d^2 on the other. Unless we have LE, the frequency of genotypes (2.1) and (2.4) may be different from the frequency of genotypes (2.2) and (2.3), even if the loci are unlinked. However, working with unlinked disease loci, we do not have to take phase into account, since the loci segregate independently. In a two-locus ordered parental genotype, $pg = (pg^1, pg^2)$, we let the labeling of the chromosomes at the two loci be independent. Hence, the notation provides no information about phase. From this follows that we may consider all ordered parental genotypes within a two-locus mating type as having the same frequency. Thus, for $pg^1 \in mt^1$ and $pg^2 \in mt^2$

$$\mathcal{P}(pg^1, pg^2) = \frac{\mathcal{P}(mt^1, mt^2)}{\#\{pg \in mt^1\} \times \#\{pg \in mt^2\}}, \tag{2.5}$$

where $\#\{pg \in mt^l\}$ is the number of ordered parental genotypes in mating type mt^l.

Note that we have made no assumptions about the genotype frequencies in the population. Studies of the properties of IBD probabilities have up to now, with few exceptions, e.g. Dudoit and Speed [21, 19], been done under the population genetic assumptions HWE, RM and LE, which are questionable for complex diseases.

2.2 Inheritance Vectors and Calculation of IBD Probabilities

The inheritance pattern of any pedigree is completely determined by an *inheritance vector* (Lander and Green [50]). The vector consists of two components for each *nonfounder*, that is the individuals whose parents are in the pedigree, in our case the siblings. The components specify which parental chromosome segments have been transmitted at the locus of interest. The sib-pair inheritance vector at $\mathcal{D}^l$ is

$$x^l = (x^l_{p1}, x^l_{m1}, x^l_{p2}, x^l_{m2}), \tag{2.6}$$

where $x_{pi} \in \{1,2\}$ ($x_{mi} \in \{3,4\}$) is the label of the chromosome from which the ith sibling has inherited his paternal (maternal) allele at $\mathcal{D}^l$.

In Table 2.2 the 16 possible single locus inheritance vectors are grouped according to the number of alleles the siblings share IBD. Let $\mathit{IBD} \in \{0,1,2\}$ be the number of alleles IBD at a locus and let $\mathit{SIBD} \in \{(i,j), i,j = 0,1\}$, denote the event that the siblings share i paternal and j maternal alleles IBD. The *sex-specific* IBD sharing, *SIBD*, is of interest not only for models of parental imprinting, but will also prove useful in derivation of inequalities for nonsex-specific IBD probabilities. Although we are focusing on two-locus ASP IBD probabilities, the single locus sex-specific and nonsex-specific IBD probabilities need to be defined as they will be used in the description. Define the unconditional nonsex-specific IBD probabilities,

$$\alpha_i = \mathcal{P}(\mathit{IBD} = i),$$

SIBD	*IBD*	Inheritance Vectors
(0,0)	0	$(1,3,2,4)$,$(1,4,2,3)$,$(2,3,1,4)$,$(2,4,1,3)$
(1,0)	1	$(1,3,1,4)$,$(1,4,1,3)$,$(2,3,2,4)$,$(2,4,2,3)$
(0,1)	1	$(1,3,2,3)$,$(1,4,2,4)$,$(2,3,1,3)$,$(2,4,1,4)$
(1,1)	2	$(1,3,1,3)$,$(1,4,1,4)$,$(2,3,2,3)$,$(2,4,2,4)$

Table 2.2: Inheritance vectors at one locus, sex-specific IBD sharing, *SIBD*, and nonsex-specific IBD sharing, *IBD*.

for $i = 0, 1, 2$. According to Mendel's first law, all inheritance vectors have the same probability. Thus the unconditional probability of any inheritance vector is $\frac{1}{16}$. It follows that the unconditional nonsex-specific IBD probabilities are given by $(\alpha_0, \alpha_1, \alpha_2) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. (The sex-specific unconditional IBD probabilities are equal to $\frac{1}{4}$ for all IBD states.)

We have in section 1.3 already used the nonsex-specific single locus ASP IBD probabilities,

$$z_i = \mathcal{P}(\mathit{IBD} = i|\text{ASP})$$

for $i = 0, 1, 2$, where ASP denotes the event that both siblings are affected with the disease in question. The sex-specific single locus ASP IBD probabilities are

$$\pi_{ij} = \mathcal{P}(\mathit{SIBD} = (i, j)|\text{ASP}),$$

$i, j = 0, 1$. We will have to remember that these probabilities depend on the genetic model of the disease, which is not indicated by the notation.

Above, we have implicitly used the single locus version of assumption 1 below, stating independence between inheritance vectors and parental genotypes at the disease loci. This means that there is no preferential transmission of certain alleles and that Mendel's first law is in effect.

Assumption 1 *There is no segregation distortion, i.e. for parental two-locus genotypes pg and inheritance vectors $x = (x^1, x^2)$ at two unlinked disease loci,*

$$\mathcal{P}(x|pg) = \left(\frac{1}{16}\right)^2.$$

□

Next we extend our notation for IBD events to two disease loci, $\mathcal{D}^1$ and $\mathcal{D}^2$. Let $\mathit{IBD} \in \{(i, j), i, j = 0, 1, 2\}$ be the nonsex-specific IBD sharing, where i (j) is the number of alleles IBD at $\mathcal{D}^1$ ($\mathcal{D}^2$). Similarly, let $\mathit{SIBD} \in \{(i_1, j_1, i_2, j_2),\ i_1, j_1, i_2, j_2 = 0, 1\}$ be the event that a sib-pair shares i_l paternal and j_l maternal alleles at locus $\mathcal{D}^l$, $l = 1, 2$.

The unconditional nonsex-specific two-locus IBD probabilities, $\alpha_{ij}, i, j = 0, 1, 2$, are for two unlinked loci given by $\alpha_{ij} = \alpha_i \alpha_j$ and the unconditional sex-specific IBD probabilities are equal to $\frac{1}{16}$ for all IBD states. For linked loci this simple relationship is obviously not true and the probabilities depend on the recombination fraction between the loci.

Finally, define the main parameters of this thesis, the two-locus ASP IBD probabilities,

$$z_{ij} = \mathcal{P}(\mathit{IBD} = (i, j)|\text{ASP})$$

for $i, j = 0, 1, 2$, and

$$\pi_{i_1 j_1 i_2 j_2} = \mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|\mathrm{ASP})$$

for $i_1, j_1, i_2, j_2 = 0, 1$. Note that $\pi_{0000} = z_{00}$, $\pi_{0001} + \pi_{0010} = z_{01}$, $\pi_{0011} = z_{02}$, etc.

The IBD probabilities may be obtained by summation of conditional probabilities of inheritance vectors, $\mathcal{P}(x^1, x^2|\mathrm{ASP})$, calculated as described in the following section, for an appropriate choice of inheritance vectors, x^1 and x^2. As an example there are four (Table 2.2) single-locus inheritance vectors corresponding to allele sharing zero. It follows that there are 16 combinations of inheritance vectors corresponding to IBD zero at two loci, and e.g. z_{00} is a sum of 16 conditional probabilities. Calculations are simplified by the fact that inheritance vectors belonging to the same IBD configuration, i.e. corresponding to the same sex-specific IBD status, have the same conditional probability given phenotype (Dudoit [19]) and we can work with one representative inheritance vector for each IBD configuration.

2.2.1 Conditional Distribution of Affected Sib-Pair Inheritance Vectors

In this section we describe how to calculate the conditional distribution of the two-locus inheritance vectors, $x = (x^1, x^2)$, given the phenotype of the sib-pair, as described by Dudoit and Speed [21]. These probabilities are used in the calculation of the conditional affected sib-pair IBD probabilities, z_{ij}. Define the phenotype vector $\phi = (\phi_1, \phi_2)$, where component ϕ_i indicates whether the ith sib is affected (ϕ_i=1) or unaffected (ϕ_i=0) by the disease in question.

From Bayes theorem and assumption 1 follows

$$\begin{aligned}
\mathcal{P}(x|\phi) &= \frac{\mathcal{P}(\phi|x)\mathcal{P}(x)}{\sum_x \mathcal{P}(\phi|x)\mathcal{P}(x)} \\
&= \frac{\sum_{pg} \mathcal{P}(\phi|x, pg)\mathcal{P}(x|pg)\mathcal{P}(pg)}{\sum_x \sum_{pg} \mathcal{P}(\phi|x, pg)\mathcal{P}(x|pg)\mathcal{P}(pg)} \\
&= \frac{\sum_{pg} \mathcal{P}(\phi|x, pg)\mathcal{P}(pg)}{\sum_x \sum_{pg} \mathcal{P}(\phi|x, pg)\mathcal{P}(pg)}. \qquad (2.7)
\end{aligned}$$

The two-locus ordered genotype, $sg_i = (sg_i^1, sg_i^2)$, of sibling i, $i = 1, 2$, is completely determined by the inheritance vectors and the parental ordered genotypes. Thus, assuming independence of sibling phenotypes, given disease genotypes (cf. assumption 2, section 2.3) $\mathcal{P}(\phi|x, pg)$ is equal to $\mathcal{P}(\phi_1|sg_1)\mathcal{P}(\phi_2|sg_2)$, where

$$\mathcal{P}(\phi_i|sg_i) = \mathcal{P}(\phi_i|pg^l_{x^l_{pi}}, pg^l_{x^l_{mi}}, l = 1, 2) \text{ for } i = 1, 2.$$

Now, in more detail, equation (2.7) can be written

$$\mathcal{P}(x^1, x^2|\phi) = \frac{\sum_{pg^1}\sum_{pg^2}\mathcal{P}(\phi_1|sg_1)\mathcal{P}(\phi_2|sg_2)\mathcal{P}(pg^1, pg^2)}{\sum_{x^1}\sum_{x^2}\sum_{pg^1}\sum_{pg^2}\mathcal{P}(\phi_1|sg_1)\mathcal{P}(\phi_2|sg_2)p(pg^1, pg^2)}. \quad (2.8)$$

The parental genotype frequencies, $\mathcal{P}(pg^1, pg^2)$, are computed from the population mating type frequencies using equation (2.5) or from allele frequencies under the equilibrium model, and $\mathcal{P}(\phi_i|sg_i)$ is a two-locus penetrance.

2.3 Penetrances

Now we have reached the last, but crucial part of the genetic model, the conditional probabilities of sibling phenotypes given their genotypes. First we will make the strong, but commonly used assumption, about independence of sib phenotypes given genotypes.

Assumption 2 *The phenotype of an individual is conditionally independent of the phenotypes and genotypes of his/her sibs, given the genotypes at the disease loci, i.e.*

$$\mathcal{P}(\phi_1, \phi_2|sg_1, sg_2) = \mathcal{P}(\phi_1|sg_1)\mathcal{P}(\phi_2|sg_2).$$

□

This assumption will obviously rarely be completely true, as it does not allow for environmental covariance. However it is needed for the proof of the triangle constraints, inequalities (1.6) and (1.7), and similar constraints, in the two-locus case. It is unclear if assumption 2 may be slightly relaxed to allow for some types of environmental covariance between sibling phenotypes. Note that in a two-locus model it is assumed that the total genetic contribution to disease susceptibility comes from the two loci in the model. If there are additional loci affecting the disease, then we must condition also by the genotypes at these loci in order to have the conditional independence between sib phenotypes, disregarding environmental covariance. Hence, if there are additional disease loci not accounted for in our two-locus model, then there will, in addition to environmental covariance, be residual genetic covariance between sib phenotypes.

Under assumption 2, the conditional probabilities of the sib-pair phenotypes, given genotypes can be expressed as a product of two-locus penetrances. Let

$$f_{i_1 j_1 i_2 j_2} = \mathcal{P}(\text{Affected } |D^1_{i_1} D^1_{j_1} D^2_{i_2} D^2_{j_2}),$$

where $D^l_{i_l}$ ($D^l_{j_l}$) is the paternally (maternally) inherited allele at locus $\mathcal{D}^l$, for $i_l, j_l = 1, \ldots, m_l$, $l = 1, 2$.

2.3.1 Examples of Two-Locus Penetrance Functions

With the single locus models described in section 1.2.2 as starting point we may form a considerable number of plausible two-locus models. As a first classification, multilocus models may be divided into heterogeneity models, where genotypes at each locus independently cause disease susceptibility, and interaction models, where certain combinations of genotypes at both loci are needed to develop the disease. One example of interaction is when genotypes at one locus hide the effect of genotypes at the other locus. This type of interaction is an instance of *epistasis*, which may be thought of as the counterpart of dominance relations between alleles at one locus. Epistasis is often used to denote interaction between two loci in a general sense.

Strickberger [70] gave 12 examples of two-locus models for categorical characters in plants and animals. One of these examples pertain to comb shape in poultry. At locus 1, allele A for rose comb is dominant over the allele a, for non-rose comb. At locus 2, the allele for pea shaped comb, B is dominant over the allele for non-pea shape, b. Combination of two dominant genotypes or the homozygous recessive genotypes results in novel phenotypes, walnut or single comb as showed in table (2.3). This and the

	locus 1		
locus 2	AA	Aa	aa
BB	*walnut*	*walnut*	*pea*
Bb	*walnut*	*walnut*	*pea*
bb	*rose*	*rose*	*single*

Table 2.3: Example of two-locus inheritance model (Strickberger [70]) where two diallelic loci jointly affect the shape of combs in poultry; pea shape, rose, single and walnut.

other examples provided by Strickberger, may be converted to a penetrance model for human disease by assigning a penetrance value for each phenotype class, or a model for a quantitative character by assigning a distribution to each class.

Ott [62] lists a few references for examples of two-locus inheritance in man and provides some examples of two-locus penetrance functions. In one of these models, table 2.4 (a), locus 1 has a mode of inheritance which depends on the genotypes at locus 2, which can be viewed as a modifier locus. Locus 1 acts strictly dominant given genotype dd at locus 2, or strictly recessive given genotype nn or nd at locus 2.

These are just a few examples out of the vast number of conceivable two-locus models. In the setting not allowing for parental imprinting there are for instance $2^9 - 2 = 510$ di-allelic penetrance models with two phe-

	locus 1		
locus 2	nn	nd	dd
nn	0	0	1
nd	0	0	1
dd	0	1	1

(a)

	locus 1		
locus 2	nn	nd	dd
nn	0	0	0
nd	0	1	1
dd	0	1	1

(b)

Table 2.4: Example of two-locus penetrance functions. **(a)** Model with modifier locus. **(b)** Jointly dominant model.

notype categories. Examples of this type of model are the modifier model (Table 2.4 (a)), mentioned above, and the fully penetrant jointly dominant model with no phenocopies (Table 2.4 (b)). The number of models is reduced by considering models equivalent if they are obtained from each other by permutations (Hartl and Maruyama [36]). This class of two-locus models has been studied by Li and Reich [53], who considered three types of permutations; (1) exchange of first and second locus, (2) exchange of two alleles at one or both loci and (3) exchange of affection status. They showed that if models are considered equivalent when alike under permutations (1) and/or (2) there are 100 nonredundant models. If in addition permutation (3) is applied there are 50 nonredunant models in this class.

In design of a linkage study or investigation of a statistical method it is convenient to work with broader classes of models. Risch [64, 65] discussed three classes of simple models which are commonly used in power considerations, method evaluations and examinations of multi-locus inheritance. These are mathematical constructions, but they may potentially approximate many true genetic models. Below the models are defined for two loci, but extensions to more loci are straight forward.

- *multiplicative model (AND model):* $f_{i_1 j_1 i_2 j_2} = x_{i_1 j_1} y_{i_2 j_2}$,
  where $x_{i_1 j_1}$ and $y_{i_2 j_2}$ are penetrance factors for locus 1 and locus 2 respectively. This type of model represents epistasis between two loci (Hodge [39]). One example is the jointly dominant model 2.4 (b). This model class has the property that multilocus IBD probabilities factorise into products of the marginal IBD probabilities (Cordell et al. [10]).

- *additive model:* $f_{i_1 j_1 i_2 j_2} = x_{i_1 j_1} + y_{i_2 j_2}$,
  for $i_l, j_l = 1, \ldots, m_l$, $l = 1, 2$. This model can be used as an approximation to the heterogeneity model below.

- *heterogeneity model (OR model):* $f_{i_1 j_1 i_2 j_2} = 1 - (1 - x_{i_1 j_1})(1 - y_{i_2 j_2})$

$= x_{i_1 j_1} + y_{i_2 j_2} - x_{i_1 j_1} y_{i_2 j_2}$.
This is a model of the situation when two loci can cause the disease independently. The risk genotype at either locus leads to disease susceptibility.

A disease model may obviously be heterogeneous even if the penetrances can not be defined in terms of penetrance factors and penetrance terms. Therefore we will follow the example of Li and Reich [53] and call the heterogeneity model above the '*OR* model', since it defines a logical OR structure when penetrance factors are only 0 or 1. Similarly the multiplicative model defines a logical AND structure.

Further examples of two-locus models are given in the following section, in the discussion of the penetrance assumptions and whether they are satisfied by various two-locus models.

2.3.2 Penetrance Assumptions

The following assumption of monotonicity of penetrances, are needed to prove constraints on two-locus IBD probabilities under the general population genetic model, with no assumptions regarding parental mating type frequencies. The assumptions contain one monotonicity property for each subset, of size one, two or three, of the four index positions. The assumptions are mathematical requirements that do not necessarily have a genetic interpretation. However, it will be seen that they are satisfied by several types of two-locus models.

Assumption 3 *Let a, $\tilde{a}$, b, $\tilde{b}$ and c, $\tilde{c}$ be arbitrary alleles at locus $\mathcal{D}^1$ or $\mathcal{D}^2$. For all a, $\tilde{a}$, b, $\tilde{b}$, c, $\tilde{c}$ and $i_l, j_l, k_l, l_l = 1, \ldots, m_l$, $l = 1, 2$, the effects of alleles and the interaction effects of alleles have the following monotonicity properties.*

- *The effect of a single allele:*
$$\begin{aligned} &(f_{i_1 abc} - f_{j_1 abc})(f_{i_1 \tilde{a}\tilde{b}\tilde{c}} - f_{j_1 \tilde{a}\tilde{b}\tilde{c}}) \geq 0, \\ &(f_{a i_1 bc} - f_{a j_1 bc})(f_{\tilde{a} i_1 \tilde{b}\tilde{c}} - f_{\tilde{a} j_1 \tilde{b}\tilde{c}}) \geq 0 \end{aligned} \tag{2.9}$$
and the corresponding two inequalities with fixed arbitrary alleles at $\mathcal{D}^1$ and the combinations of indices defined in inequality (2.9) at $\mathcal{D}^2$.

- *The interaction between two alleles at the same locus:*
$$\begin{aligned} &(f_{i_1 k_1 ab} + f_{j_1 l_1 ab} - f_{i_1 l_1 ab} - f_{j_1 k_1 ab})\times \\ &(f_{i_1 k_1 \tilde{a}\tilde{b}} + f_{j_1 l_1 \tilde{a}\tilde{b}} - f_{i_1 l_1 \tilde{a}\tilde{b}} - f_{j_1 k_1 \tilde{a}\tilde{b}}) \geq 0 \end{aligned} \tag{2.10}$$
and the corresponding inequality with fixed arbitrary alleles at $\mathcal{D}^1$ and the combination of indices defined in inequality (2.10) at $\mathcal{D}^2$.

- *The interaction between two alleles at different loci:*

$$\begin{aligned}&(f_{i_1ai_2b}+f_{j_1aj_2b}-f_{i_1aj_2b}-f_{j_1ai_2b})\times\\&(f_{i_1\tilde{a}i_2\tilde{b}}+f_{j_1\tilde{a}j_2\tilde{b}}-f_{i_1\tilde{a}j_2\tilde{b}}-f_{j_1\tilde{a}i_2\tilde{b}})\geq 0\end{aligned}\tag{2.11}$$

  and the corresponding three inequalities with one fixed arbitrary allele at each loci and the combination of indices at the remaining positions as defined in inequality (2.11).

- *The interaction between one allele and the pair of alleles at the other locus:*

$$\begin{aligned}&(f_{i_1ai_2j_2}+f_{i_1ak_2l_2}-f_{i_1ai_2l_2}-f_{i_1ak_2j_2}\\&-f_{j_1ai_2j_2}-f_{j_1ak_2l_2}+f_{j_1ai_2l_2}+f_{j_1ak_2j_2})\times\\&(f_{i_1\tilde{a}i_2j_2}+f_{i_1\tilde{a}k_2l_2}-f_{i_1\tilde{a}i_2l_2}-f_{i_1\tilde{a}k_2j_2}\\&-f_{j_1\tilde{a}i_2j_2}-f_{j_1\tilde{a}k_2l_2}+f_{j_1\tilde{a}i_2l_2}+f_{j_1\tilde{a}k_2j_2})\geq 0\end{aligned}\tag{2.12}$$

  and the corresponding three inequalities with one fixed arbitrary allele and the combination of indices at the remaining three positions as defined in inequality (2.12).

□

The monotonicity assumption (2.9) says that an allele affects disease susceptibility in the the same direction, irrespective of the other allele in the genotype and the genotype at the other locus. Assumption (2.10) [(2.11)] requires that the interaction effects of two alleles at the same [different] locus have the same sign irrespective of the combination of two alleles at the remaining positions. Finally, Assumption (2.12) is a similar monotonicity assumption about the interaction between one allele at a locus and the genotype at the other locus.

Running over all possible positions of the fix indices, inequalities (2.9) - (2.12) give 14 monotonicity assumptions. It may be useful to comment on the analogy to main and interaction effects in a 2^4 factorial experiment, with the alleles corresponding to factors. There is one restriction on each of the four 'main' effects (2.9), six restrictions on 'two-factor' interactions (2.10) and (2.11) and four restrictions on the 'three-factor' interactions (2.12). Assumption 3 says that these effects have the same sign, irrespective of the level of the other factors, that is the alleles not included in the effect.

Monotonicity of Penetrances under the Multiplicative, Additive and *OR* Models

Assumption 3 holds for the multiplicative, additive and *OR* models defined in section 2.3.1, if the penetrance factors satisfy the following assumptions, analogous to the single locus penetrance assumptions, (1.8) and (1.9). For $i_l, j_l, k_l, l_l = 1, \ldots, m_l$, $l = 1, 2$

$$\begin{aligned}
&(x_{i_1k_1} - x_{j_1k_1})(x_{i_1l_1} - x_{j_1l_1}) \geq 0,\\
&(x_{i_1k_1} - x_{i_1l_1})(x_{j_1k_1} - x_{j_1l_1}) \geq 0,\\
&(y_{i_2k_2} - y_{j_2k_2})(y_{i_2l_2} - y_{j_2l_2}) \geq 0,\\
&(y_{i_2k_2} - y_{i_2l_2})(y_{j_2k_2} - y_{j_2l_2}) \geq 0.
\end{aligned} \tag{2.13}$$

This assumption, requiring an ordering of the penetrance factors, implies that the marginal penetrances satisfy the single locus monotonicity assumptions (1.8) and (1.9). To see that inequality (2.11) is true under this assumption for the multiplicative model consider the following representative example.

$$\begin{gathered}
(f_{i_1k_1i_2k_2} + f_{j_1k_1j_2k_2} - f_{i_1k_1j_2k_2} - f_{j_1k_1i_2j_2})\times\\
(f_{i_1l_1i_2l_2} + f_{j_1l_1j_2l_2} - f_{i_1l_1j_2l_2} - f_{j_1l_1i_2l_2}) =\\
(x_{i_1k_1} - x_{j_1k_1})(x_{i_1l_1} - x_{j_1l_1})(y_{i_2k_2} - y_{j_2k_2})(y_{i_2l_2} - y_{j_2l_2}),
\end{gathered}$$

which is non negative under assumption (2.13). Under the additive model the left side of inequality (2.11) is equal to zero, and hence the assumption is true for all additive models. For the *OR* model the penetrance terms vanish and the remaining products are non-negative, by the argument used for the multiplicative model.

Next, look at an example of inequality (2.12), expressed in penetrance factors of a multiplicative model,

$$\begin{gathered}
(x_{i_1k_1}y_{i_2j_2} + x_{i_1k_1}y_{k_2l_2} - x_{i_1k_1}y_{i_2l_2} - x_{i_1k_1}y_{k_2j_2}\\
-x_{j_1k_1}y_{i_2j_2} - x_{j_1k_1}y_{k_2l_2} + x_{j_1k_1}y_{i_2l_2} + x_{j_1k_1}y_{k_2j_2}) \times\\
(x_{i_1l_1}y_{i_2j_2} + x_{i_1l_1}y_{k_2l_2} - x_{i_1l_1}y_{i_2l_2} - x_{i_1l_1}y_{k_2j_2} -\\
x_{j_1l_1}y_{i_2j_2} - x_{j_1l_1}y_{k_2l_2} + x_{j_1l_1}y_{i_2l_2} + x_{j_1l_1}y_{k_2j_2}) =\\
(x_{i_1k_1} - x_{j_1k_1})(x_{i_1l_1} - x_{j_1l_1})(y_{i_2j_2} + y_{k_2l_2} - y_{i_2l_2} + y_{k_2j_2})^2.
\end{gathered}$$

The product of the first two factors is non-negative under assumption (2.13) and it follows that (2.12) is true for the multiplicative model. From the arguments used above for the *OR* and additive models, it follows that (2.12) is true also for these models. The proofs of (2.9) and (2.10) are similar.

Monotonicity of Penetrances Under Models with Two Phenotype Classes

Since the penetrance products in assumption 3 contain as many positive as negative terms, it follows that the assumption is satisfied also by any model that can be transformed to a multiplicative, additive or *OR* model by addition or subtraction of a constant to all penetrances. Consider for example a fully penetrant jointly dominant model with a phenocopy rate of 0.1, that is $f_{nnij} = f_{ijnn} = 0.1$ for $i, j = n, d$ and $f_{ijkl} = 1$ otherwise. This model does not fit any of the three classes, but if we subtract 0.1 from all f_{ijkl}, then these penetrances will fit a multiplicative model.

It follows that to investigate the penetrance assumptions for dichotomous models, that is models with two phenotype classes, it suffices to study the class of models with no phenocopies and a single penetrance, f, for all risk genotypes. All dichotomous models can be transformed to a model of this type by subtraction of the phenocopy rate.

Among the 510 bi-allelic dichotomous models with no imprinting (denote this model class M22 for future reference), there are 42 models satisfying assumption 3. If we consider models as equivalent under permutations as discussed in section 2.3.1, then there are six types of models satisfying the assumptions, displayed in table 2.5. In M22 only multiplicative and *OR* models satisfy the assumptions. In table 2.6 we give examples of models in M22 that do not satisfy assumption 3. Models with a structure corresponding to single locus overdominance, table 2.6 (a) - (c), do not satisfy inequality (2.9). One example is a jointly dominant model where the risk is lower for the double homozygous genotype than for the other risk genotypes, table 2.6 (a). Inequality (2.10) is not fulfilled by the modifier model 2.6 (d) nor by the threshold models where individuals with at least two or three disease alleles are at risk (table 2.6 (e) and (f)). In M22 all models that satisfy assumptions (2.9) and (2.10) also satisfy (2.11) and (2.12).

Monotonicity of Penetrances Under Models with More Than Two Phenotype Classes

If we consider models with more than two phenotype classes, then we can find models that satisfy (2.9) and (2.10) but not (2.11) and (2.12) (see table 2.7 (a) and (b)).

The model in table 2.7 (c) is interesting because it satisfies assumption 3 but it does not have a multiplicative, additive or *OR* structure, and can not be transformed to any of these forms by addition of a constant. Hence these penetrance structures are not necessary to fulfil the monotonicity assumptions for models with more than two phenotype classes. Note that the model in table 2.7 (c) is similar to the modifier model in table 2.6 (d), which

.

0	0	0
0	0	0
0	0	f

(a)

0	0	0
0	0	0
0	f	f

(b)

0	0	0
0	f	f
0	f	f

(c)

0	f	f
f	f	f
f	f	f

(d)

0	0	f
f	f	f
f	f	f

(e)

0	0	f
0	0	f
f	f	f

(f)

Table 2.5: Bi-allelic, dichotomous two-locus penetrance models satisfying assumption 3 (cf. table 2.4). **(a)** Jointly Recessive model (mult.). **(b)** Dominant-Recessive model (mult.). **(c)** Jointly Dominant model (mult.). **(d)** Dominant+Dominant Heterogeneity model (OR). **(e)** Recessive+Dominant Heterogeneity model (OR). **(f)** Recessive+Recessive Heterogeneity Model(OR).

.

0	0	0
0	f	f
0	f	0

(a)

0	0	0
0	0	f
0	f	0

(b)

0	0	0
0	f	0
0	0	f

(c)

0	0	0
0	0	f
f	f	f

(d)

0	0	f
0	f	f
f	f	f

(e)

0	0	0
0	0	f
0	f	f

(f)

Table 2.6: Bi-allelic, dichotomous two-locus penetrance models that do not satisfy assumption 3 (cf. table 2.4). **(a)-(c)** Overdominance models. **(d)** Modifier model. **(e)** Treshold model (two or more disease alleles). **(f)** Treshold model (three or more disease alleles).

does not satisfy assumption 3. The model in table 2.7 (d) is additive and hence satisfies assumption 3, whereas 2.7 (e) does not satisfy assumption 3 even though it looks similar at first appearance.

The last two examples illustrate the fact that while we for the M22 model class easily can deduce from the structure of the model if the monotonicity assumptions hold or not, it is for models with more phenotype classes a question of subtle mathematical differences between the models. Assump-

0	0	0
0	0.5	1
0	1	1

(a)

0	0	0
0	0	0.5
0	0.5	1

(b)

0	0	0
0	0	0.5
0.5	0.5	1

(c)

0	0.3	0.5
0.2	0.5	0.7
0.4	0.7	0.9

(d)

0	0.3	0.5
0.2	0.5	0.7
0.4	0.9	1

(e)

Table 2.7: Examples of models with more than two phenotype classes (cf. table 2.4). **(a)** Model satisfying (2.9) and (2.10), but not (2.11). **(b)** Model satisfying assumptions (2.9) - (2.11) , but not (2.12)). **(c)** Modifier model with reduced penetrance for risk genotypes that are not double homozygots. All assumptions satisfied. The model does not have a mult., add. or OR structure. **(d)** Additive model. **(e)** Model similar to (d), but(2.11) and (2.12) are not satisfied.

tions (2.10) - (2.12) seem to be merely mathematical requirements about symmetries in the 3×3-penetrance table without any natural interpretation. Inequality (2.9) on the other hand has a genetic interpretation in that it does not allow overdominance models.

Monotonicity of Penetrances Under Imprinting Models

Finally we will consider imprinting models. In table 2.8 we have tabulated two imprinting models that may or may not be biologically relevant. In model (a) we have a model which resembles a jointly dominant model (table 2.5 (c)) if the disease allele in the heterozygous genotype at locus 1, d_h say, is of maternal origin. For paternally inherited d_h the model mimics a recessive-dominant model (table 2.5 (b)). These two non-imprinting models satisfy the penetrance assumptions and so does the imprinting model in table 2.8 (a).

Table 2.8 (b) displays a model which also is jointly dominant if d_h comes from the mother. If d_h is of paternal origin the model resembles a threshold model requiring at least three disease alleles (table 2.6 (f)). The threshold model does not satisfy the penetrance assumptions and hence this is the case also for the imprinting model in table 2.8 (b). These two examples illustrate that imprinting per se does not make a model incompatible with the penetrance assumptions. It suggests that if the non-imprinting models, that are combined into the imprinting model for different parental origin of the disease alleles, satisfy assumption 3, then this is the case also for the imprinting model.

	locus 1			
locus 2	nn	nd	dn	dd
nn	0	0	0	0
nd	0	1	0	1
dn	0	1	0	1
dd	0	1	0	1

(a)

	locus 1			
locus 2	nn	nd	dn	dd
nn	0	0	0	0
nd	0	1	0	1
dn	0	1	0	1
dd	0	1	1	1

(b)

Table 2.8: Imprinting models. **(a)** Locus 1: Maternal disease allele acts dominantly, paternal disease allele is recessive. Locus 2: Disease allele acts dominantly. **(b)** If the disease allele in the heterozygous genotype at locus 1 is of maternal origin, then one disease allele at locus 2 is enough to develop the disease. Otherwise a total of three disease alleles is reuqired.

Chapter 3

Two-locus IBD Probabilities under General Mating Type Frequencies

In this chapter we derive and prove a set of constraints on two-locus IBD probabilities under the general population genetic model described in section 2.1 and the penetrance assumptions of section 2.3.2. The starting point is the linear constraints derived by Cordell et al. [10] and Dudoit and Speed [21], which are described in section 3.1. Next, in section 3.2 we describe a parameterisation of the IBD probabilities that facilitates maximum likelihood estimation. The parameterisation implies additional constraints on the IBD probabilities, which are proved in section 3.3. Furthermore, we show that the resulting parameter space is covered, i.e. all IBD probabilities within the constrained region can be obtained under the used model, and we prove the necessity of the penetrance assumptions. Finally, we discuss interpretation of the parameters of our representation.

3.1 A First Set of Constraints

Cordell et al. [10] derived a set of constraints on two-locus ASP nonsex-specific IBD probabilities, z_{ij}, $i, j = 0, 1, 2$, implicitly assuming HWE, RM and LE. The proof, based on the expression of IBD probabilities in terms of genetic variance components, will be discussed in section 4.2.3. The same inequalities also follow from a result by Dudoit and Speed [21], concerning constraints on multilocus sex-specific IBD probabilities with sharing specified at one locus and fixed arbitrary sharing at the other loci. We will state their proposition here, but in a two-locus setting. The constraints were proved under general mating type frequencies and penetrance monotonicity assumptions (2.9) and (2.10).

Proposition 1 *Under a two-locus model with arbitrary mating type frequencies and assumptions 1, 2 and monotonicity properties (2.9) and (2.10) in assumption 3, the sex-specific two-locus IBD probabilities at the disease loci satisfy the following constraints. For all IBD states $i_2, j_2 = 0, 1$ at the second locus*

$$\pi_{00i_2j_2} \leq \pi_{10i_2j_2}, \tag{3.1}$$

$$\pi_{00i_2j_2} \leq \pi_{01i_2j_2} \tag{3.2}$$

and

$$\pi_{01i_2j_2} + \pi_{10i_2j_2} \leq \pi_{00i_2j_2} + \pi_{11i_2j_2}. \tag{3.3}$$

Analogous constraints hold with fixed arbitrary sharing at the first locus and specified IBD states at the second locus.

□

In section 3.3.3 we prove additional constraints with the technique used by Dudoit and Speed [21], so the proof of proposition 1 is omitted here.

The inequalities for nonsex-specific IBD probabilities are obtained by letting $i = i_1 + j_1$, $j = i_2 + j_2$ and using the sex-specific inequalities in proposition 1. Addition of inequalities (3.1) and (3.2) leads to

$$z_{i0} \leq \tfrac{1}{2} z_{i1}, \tag{3.4}$$

for $i = 0, 1, 2$. From (3.3) follows directly that

$$z_{i1} \leq z_{i0} + z_{i2}. \tag{3.5}$$

Similarly we get

$$z_{0j} \leq \tfrac{1}{2} z_{1j}, \tag{3.6}$$

and

$$z_{1j} \leq z_{0j} + z_{2j} \tag{3.7}$$

for $j = 0, 1, 2$. The twelve inequalities (3.4)-(3.7) are the constraints derived by Cordell et al. [10].

Note that these constraints imply that the marginal IBD probabilities, $z_{i+} = \sum_{j=0}^{2} z_{ij}$ and $z_{+j} = \sum_{i=0}^{2} z_{ij}, i, j = 0, 1, 2$, satisfy the single locus triangle inequalities (1.6) and (1.7).

3.2 Parameterisation of ASP IBD Probabilities

Naturally, we want to obtain maximum likelihood estimates of the two-locus IBD probabilities, subject to the constraints, from observed marker data.

In the single locus situation the IBD probabilities, z_i, $i = 0, 1, 2$ are routinely estimated by maximum likelihood, restricted to the possible triangle defined by inequalities (1.6) and (1.7). In the implementation of Kruglyak and Lander [48], unconstrained maximisation is first performed. If the estimates fall outside the triangle, then the likelihood is remaximised along the sides of the triangle. In the two locus-case, with the mutually dependent constraints (3.4)-(3.7), a similar algorithm for maximisation would be more complicated. However, this can be accomplished by a parameterisation which enforces the constraints and replaces all constraints by a single such constraint, along the lines suggested by Zhao et al. [79] for another application.

In section 3.2.1 the parameterisation is derived for single locus IBD probabilities restricted to the possible triangle, to indicate what we want to accomplish in the two-locus case. Extension of the parameterisation to the two-locus case implies additional constraints on the two-locus IBD probabilities, which are proved in section 3.3. The estimation procedure is described in chapter 6.

3.2.1 Parameterisation of Single-Locus IBD Probabilities

We want to express (z_0, z_1, z_2), restricted to the possible triangle, in a set of parameters, (q_0, q_1, q_2), which are positive and sum to one, that is, standard trinomial probabilities. Inequalities (1.6) and (1.7) can be written

$$\begin{aligned} z_1 - 2z_0 &\geq 0 \\ z_0 - z_1 + z_2 &\geq 0. \end{aligned} \tag{3.8}$$

This suggests writing

$$\begin{aligned} q_0 &= k_0 z_0, \\ q_1 &= k_1(z_1 - 2z_0), \\ q_2 &= k_2(z_0 - z_1 + z_2) \end{aligned} \tag{3.9}$$

for some constants, $k_i > 0$, $i = 0, 1, 2$. Inverting these relations we get

$$\begin{aligned} z_0 &= \tfrac{1}{k_0} q_0, \\ z_1 &= \tfrac{2}{k_0} q_0 + \tfrac{1}{k_1} q_1, \\ z_2 &= \tfrac{1}{k_0} q_0 + \tfrac{1}{k_1} q_1 + \tfrac{1}{k_2} q_2, \end{aligned}$$

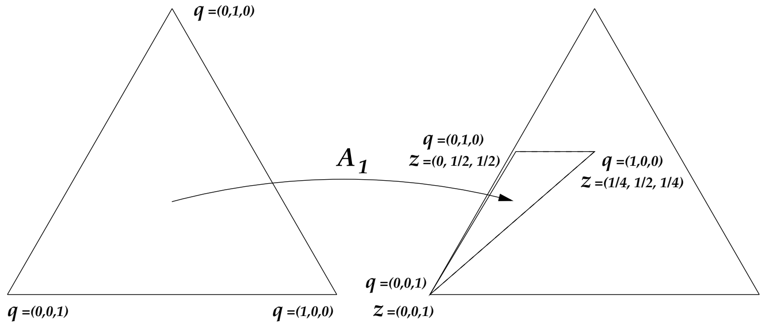

Figure 3.1: Barycentric representation of the simplex S_3 and the possible triangle as its image under A_1.

with $\sum z_i = \frac{4}{k_0}q_0 + \frac{2}{k_1}q_1 + \frac{1}{k_2}q_2 = 1$. Since $\sum q_i$=1 and $0 \leq q_i \leq 1$, we can choose $k_0 = 4$, $k_1 = 2$ and $k_2 = 1$. This results in the new parameterisation

$$\boldsymbol{z} = \boldsymbol{A_1 q}, \tag{3.10}$$

where $\boldsymbol{z} = (z_0, z_1, z_2)'$, $\boldsymbol{q} = (q_0, q_1, q_2)'$, $q_i > 0$, $\sum q_i = 1$ and

$$\boldsymbol{A_1} = \frac{1}{4}\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 1 & 2 & 4 \end{bmatrix}. \tag{3.11}$$

From the defining relations (3.9) and the fact that q_i are positive for all i, it is evident that the triangle constraints hold. It is also clear that any set of IBD probabilities within the triangle can be obtained from (3.10).

This parameterisation represents the possible triangle as the image of the simplex

$$S_3 = \{(q_0, q_1, q_2) : q_i > 0, \sum q_i = 1\}$$

under the invertible stochastic matrix $\boldsymbol{A_1}$ (Fig. 3.1). The null hypothesis of no linkage, $\boldsymbol{z} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ is the image of $\boldsymbol{q} = (1, 0, 0)$.

Now consider the sex-specific single locus IBD probabilities. Suppose $\boldsymbol{q_s}$ is a column vector consisting of four positive parameters summing to one. Further, let $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ and

$$\boldsymbol{A_{s1}} = \frac{1}{4}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 2 & 2 & 4 \end{bmatrix}. \tag{3.12}$$

A representation of the sex-specific single locus IBD probabilities, subject to the single locus version of inequalities (3.1)-(3.3) (simply disregard the indices, i_2, j_2, of the second locus), is obtained by

$$\boldsymbol{\pi} = \boldsymbol{A_{s1} q_s}. \tag{3.13}$$

The form of $\boldsymbol{z}$ and $\boldsymbol{\pi}$, expressed in terms of an unobserved multinomial distribution, makes the EM algorithm available for estimation.

3.2.2 Parameterisation of Two-Locus IBD Probabilities

We want to find a parameterisation, corresponding to the one-locus representation (3.10), where the two-locus nonsex-specificIBD probabilities z_{ij}, $i, j = 0, 1, 2$, restricted by inequalities (3.4) - (3.7), are expressed in terms of nine parameters that are unconstrained, except being positive and summing to one. Suppose that this can be accomplished by first letting

$$(z_{i0}, z_{i1}, z_{i2})' = \boldsymbol{A_1}(\tilde{q}_{i0}, \tilde{q}_{i1}, \tilde{q}_{i2})' \tag{3.14}$$

and then

$$(\tilde{q}_{i0}, \tilde{q}_{i1}, \tilde{q}_{i2})' = \boldsymbol{A_1}(q_{i0}, q_{i1}, q_{i2})', \tag{3.15}$$

for $i, j = 0, 1, 2$. This results in the parameterisation

$$\begin{aligned}
z_{00} &= \tfrac{1}{16} q_{00}, \\
z_{01} &= \tfrac{1}{8} q_{00} + \tfrac{1}{8} q_{01}, \\
z_{02} &= \tfrac{1}{16} q_{00} + \tfrac{1}{8} q_{01} + \tfrac{1}{4} q_{02}, \\
z_{10} &= \tfrac{1}{8} q_{00} + \tfrac{1}{8} q_{10}, \\
z_{11} &= \tfrac{1}{4} q_{00} + \tfrac{1}{4} q_{01} + \tfrac{1}{4} q_{10} + \tfrac{1}{4} q_{11}, \\
z_{12} &= \tfrac{1}{8} q_{00} + \tfrac{1}{4} q_{01} + \tfrac{1}{2} q_{02} + \tfrac{1}{8} q_{10} + \tfrac{1}{4} q_{11} + \tfrac{1}{2} q_{12}, \\
z_{20} &= \tfrac{1}{16} q_{00} + \tfrac{1}{8} q_{10} + \tfrac{1}{4} q_{20}, \\
z_{21} &= \tfrac{1}{8} q_{00} + \tfrac{1}{8} q_{01} + \tfrac{1}{4} q_{10} + \tfrac{1}{4} q_{11} + \tfrac{1}{2} q_{20} + \tfrac{1}{2} q_{21}, \\
z_{22} &= \tfrac{1}{16} q_{00} + \tfrac{1}{8} q_{01} + \tfrac{1}{4} q_{02} + \tfrac{1}{8} q_{10} + \tfrac{1}{4} q_{11} + \tfrac{1}{2} q_{12} + \tfrac{1}{4} q_{20} + \tfrac{1}{2} q_{21} + q_{22},
\end{aligned} \tag{3.16}$$

for $q_{ij} \geq 0$ and $\sum_{i=0}^{2} \sum_{j=0}^{2} q_{ij} = 1$. With this representation the constraints (3.4)-(3.7) are satisfied. However, it also imposes additional constraints, for instance

$$\begin{aligned}
4z_{00} - 2z_{01} - 2z_{10} + z_{11} &\geq 0, \\
-2z_{00} + 2z_{01} - 2z_{02} + z_{10} - z_{11} + z_{12} &\geq 0, \\
-2z_{00} + z_{01} + 2z_{10} - z_{11} - 2z_{20} + z_{21} &\geq 0, \\
z_{00} - z_{01} + z_{02} - z_{10} + z_{11} - z_{12} + z_{20} - z_{21} + z_{22} &\geq 0, \quad (3.17)
\end{aligned}$$

which do not follow from (3.4) - (3.7). These new constraints follow from constraints on sex-specific IBD probabilities that will be derived in section 3.3.3, under the general population genetic model, and in chapter 4 under the equilibrium model. In fact, it will be seen that they follow from the approach used by Cordell et al. [10] in their derivation of (3.4)-(3.7). For the moment suppose that the inequalities (3.17) are true.

The repeated use of the single-locus representation in equations (3.14) and (3.15) implies that the two-locus representation can be expressed as a Kronecker product of two single locus matrices, $\boldsymbol{A_1}$, defined in equation (3.11). Let

$$\boldsymbol{z} = (z_{00}, z_{01}, \ldots, z_{22})' \quad (3.18)$$

and let $\boldsymbol{q}$ be a column vector with q_{ij}, $i, j = 0, 1, 2$ in the corresponding order. Then

$$\boldsymbol{z} = \boldsymbol{A_1}^{\otimes 2}\boldsymbol{q} = \boldsymbol{A_2}\boldsymbol{q}, \quad (3.19)$$

where $\otimes$ denotes left Kronecker product.

A similar representation of sex-specific two-locus IBD probabilities is obtained by

$$\boldsymbol{\pi} = \boldsymbol{A_{s1}}^{\otimes 2}\boldsymbol{q} = \boldsymbol{A_{s2}}\boldsymbol{q}, \quad (3.20)$$

where

$$\boldsymbol{\pi} = (\pi_{0000}, \pi_{0001}, \pi_{0010}, \pi_{0011}, \pi_{0100}, \ldots \pi_{1111})', \quad (3.21)$$

$\boldsymbol{q}$ is a column vector of parameters that have the corresponding indices, are positive and sum to one, and $\boldsymbol{A_{s1}}$, defining the single-locus parameterisation, is given in (3.12). Representation (3.20) imposes additional restrictions on $\boldsymbol{\pi}$ that do not follow from proposition 1. These are derived in section 3.3.3, under assumptions 1-3.

3.3 A Complete Set of Constraints

We will now define and prove a set of linearly interdependent inequalities, including the new inequalities (3.17). The proof is accomplished by first proving a set of 16 inequalities for sex-specific IBD probabilities, which then are added to obtain a set of 9 nonsex-specific inequalities.

The derived set of inequalities is complete in the sense that all inequalities above can be obtained and the resulting parameter space is covered under the genetic model with general mating type frequencies, which is proved in section 3.3.4.

3.3.1 Constraints on Sex-Specific Two-Locus IBD Probabilities

Consider figure (3.2) representing the 16 sex-specific two-locus IBD proba-

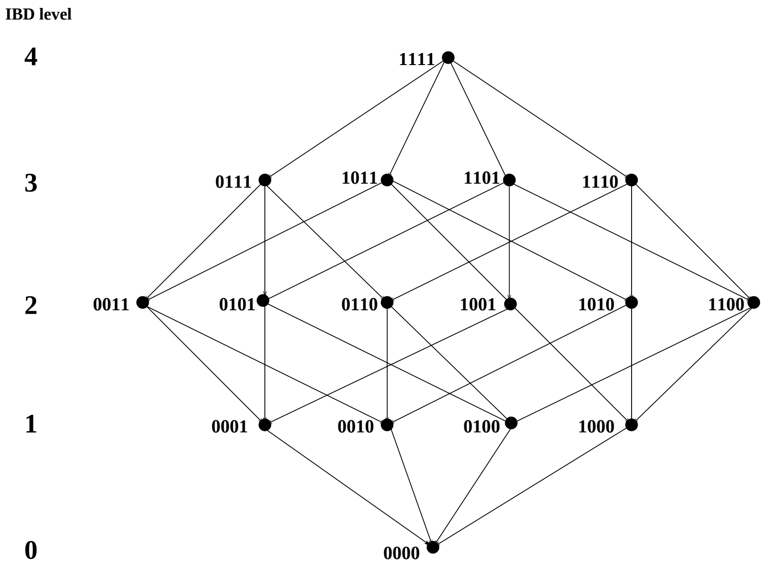

Figure 3.2: Graph representing the 16 sex-specific two-locus IBD probabilities, $\pi_{i_1 j_1 i_2 j_2}$, $i_1, j_1, i_2, j_2 = 0, 1$. The five levels, 0-4, correspond to the total number of alleles IBD.

bilities, with five levels corresponding to the total number of alleles shared IBD. Note that a node at level k is connected to a node at level $k-1$ if one and only one index has changed from 1 to 0. Remember that according to

Proposition 1 we have the inequalities

$$\begin{aligned}
\pi_{0001} - \pi_{0000} &\geq 0 \\
\pi_{0010} - \pi_{0000} &\geq 0 \\
\pi_{0100} - \pi_{0000} &\geq 0 \\
\pi_{1000} - \pi_{0000} &\geq 0 \\
\pi_{0011} - \pi_{0001} - \pi_{0010} + \pi_{0000} &\geq 0 \\
\pi_{1100} - \pi_{0100} - \pi_{1000} + \pi_{0000} &\geq 0.
\end{aligned}$$

Note that the inequalities involving two terms are constructed by taking one probability at level 1 and subtracting the level 0 term. The inequalities with four terms are alternating sums of one term from level 2, minus the two connected terms at level 1, plus the level 0 term. This pattern suggests that we may get a complete set of constraints by considering all alternating sums starting with a term at level k, k=0,...,4, minus all connected terms at level $k-1$, plus all connected term at level $k-2$ and so on. In this way we get the following 16 inequalities for the sex-specific IBD probabilities.

$$\pi_{0000} \geq 0 \quad (3.22)$$
$$\pi_{1000} - \pi_{0000} \geq 0 \quad (3.23)$$
$$\pi_{0100} - \pi_{0000} \geq 0 \quad (3.24)$$
$$\pi_{0010} - \pi_{0000} \geq 0 \quad (3.25)$$
$$\pi_{0001} - \pi_{0000} \geq 0 \quad (3.26)$$
$$\pi_{1100} - \pi_{0100} - \pi_{1000} + \pi_{0000} \geq 0 \quad (3.27)$$
$$\pi_{0011} - \pi_{0001} - \pi_{0010} + \pi_{0000} \geq 0 \quad (3.28)$$
$$\pi_{1010} - \pi_{0010} - \pi_{1000} + \pi_{0000} \geq 0 \quad (3.29)$$
$$\pi_{1001} - \pi_{0001} - \pi_{1000} + \pi_{0000} \geq 0 \quad (3.30)$$
$$\pi_{0110} - \pi_{0010} - \pi_{0100} + \pi_{0000} \geq 0 \quad (3.31)$$
$$\pi_{0101} - \pi_{0001} - \pi_{0100} + \pi_{0000} \geq 0 \quad (3.32)$$
$$\pi_{1110} - \pi_{0110} - \pi_{1010} - \pi_{1100} + \pi_{0010} + \pi_{0100} + \pi_{1000} - \pi_{0000} \geq 0 \quad (3.33)$$
$$\pi_{1101} - \pi_{1100} - \pi_{1001} - \pi_{0101} + \pi_{1000} + \pi_{0100} + \pi_{0001} - \pi_{0000} \geq 0 \quad (3.34)$$
$$\pi_{1011} - \pi_{1010} - \pi_{1001} - \pi_{0011} + \pi_{1000} + \pi_{0010} + \pi_{0001} - \pi_{0000} \geq 0 \quad (3.35)$$
$$\pi_{0111} - \pi_{0011} - \pi_{0110} - \pi_{0101} + \pi_{0001} + \pi_{0010} + \pi_{0100} - \pi_{0000} \geq 0 \quad (3.36)$$
$$\begin{aligned}
&\pi_{1111} - \pi_{1110} - \pi_{1101} - \pi_{1011} - \pi_{0111} + \pi_{0011} + \pi_{0101} + \pi_{1001} \\
&+\pi_{0110} + \pi_{1010} + \pi_{1100} - \pi_{0001} - \pi_{0010} - \pi_{0100} - \pi_{1000} + \pi_{0000} \geq 0 \quad (3.37)
\end{aligned}$$

Consider the inverse of the 16×16 matrix, $\boldsymbol{A_{s2}}$, that define parameterisation (3.20). It may be seen that each row of $\boldsymbol{A_{s2}}^{-1}$ corresponds to one of the inequalities (3.22) - (3.37), or equivalently to one node and the connected nodes at lower levels in the graph in figure (3.2). (See the nonsex-specific case in section 3.3.2, where the corresponding matrix is displayed.) Hence, the set of inequalities may be expressed as

$$\boldsymbol{A_{s2}}^{-1}\boldsymbol{\pi} \geq 0.$$

This set of constraints contains no redundant inequalities, as they are linearly independent. They imply a large number of additional inequalities, for instance 32 inequalities involving two probabilities. In particular all 24 inequalities in proposition 1 follow from (3.22)-(3.37).

In order to understand what these constraints represent, it may be helpful to again use the analogy with main and interaction effects in a 2^4 factorial experiment. This was mentioned in relation to the penetrance monotonicity assumptions in section 2.3.2. We see that each of the inequalities (3.23)-(3.36) corresponds to one of the 14 monotonicity properties in assumption 3. Inequalities (3.23)-(3.26) correspond to assumption (2.9), Inequalities (3.27) and (3.28) correspond to assumption (2.10) etc. The proof of each inequality (section 3.3.3) relies on the corresponding monotonicity assumption.

3.3.2 Constraints on Nonsex-Specific Two-Locus IBD Probabilities

In Fig. (3.3) the graph in Figure (3.2) of the sixteen sex-specific IBD probabilities has been collapsed into a graph of the nine nonsex-specific IBD probabilities. Each inequality is obtained by summation of the sex-specific inequalities whose first term is located at the same node in the graph. For instance, the inequality starting with z_{12}, is obtained by adding (3.35) and (3.36), which after conversion to non-sex-specific notation results in

$$-2z_{00} + 2z_{01} - 2z_{02} + z_{10} - z_{11} + z_{12} \geq 0.$$

Each inequality is represented by a row in the inverse of the matrix $\boldsymbol{A_2}$, defined in equation (3.19),

$$\boldsymbol{A_2}^{-1} = \begin{bmatrix} 16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -16 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & -4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ -16 & 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 \\ 16 & -8 & 0 & -8 & 4 & 0 & 0 & 0 & 0 \\ -4 & 4 & -4 & 2 & -2 & 2 & 0 & 0 & 0 \\ 4 & 0 & 0 & -4 & 0 & 0 & 4 & 0 & 0 \\ -4 & 2 & 0 & 4 & -2 & 0 & -4 & 2 & 0 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}.$$

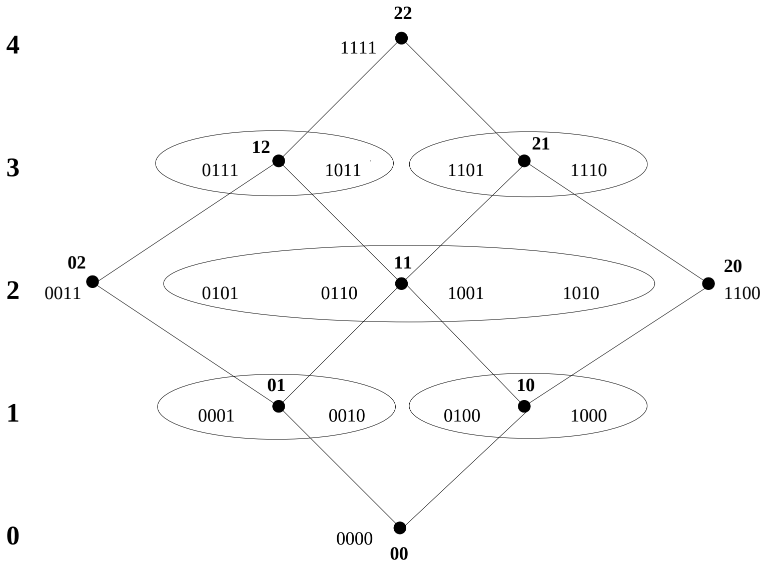

Figure 3.3: Graph representing the nine nonsex-specific IBD probabilities and the corresponding sex-specific probabilities.

Thus, we get the following nine inequalities,

$$
\begin{aligned}
z_{00} &\geq 0 && (3.38)\\
z_{01} - 2z_{00} &\geq 0 && (3.39)\\
z_{10} - 2z_{00} &\geq 0 && (3.40)\\
z_{00} - z_{01} + z_{02} &\geq 0 && (3.41)\\
z_{00} - z_{10} + z_{20} &\geq 0 && (3.42)\\
4z_{00} - 2z_{10} - 2z_{01} + z_{11} &\geq 0 && (3.43)\\
-2z_{00} + 2z_{01} - 2z_{02} + z_{10} - z_{11} + z_{12} &\geq 0 && (3.44)\\
-2z_{00} + z_{01} + 2z_{10} - z_{11} - 2z_{20} + z_{21} &\geq 0 && (3.45)\\
z_{00} - z_{01} + z_{02} - z_{10} + z_{11} - z_{12} + z_{20} - z_{21} + z_{22} &\geq 0 && (3.46)
\end{aligned}
$$

which may be expressed

$$\boldsymbol{A_2}^{-1}\boldsymbol{z} \geq 0. \qquad (3.47)$$

With nine IBD probabilities we can at most have nine linearly independent inequalities. The inequalities given by (3.47) contain no redundant constraints as they are linearly independent. They include the new constraints (3.43) - (3.46) and generate all twelve constraints (3.4)-(3.7) derived

1		

-2	1	

1	-1	1

-2	**2**	**-2**
1	**-1**	**1**

-2		
1		

1		
-1		
1		

-2	**1**	
2	**-1**	
-2	**1**	

4	**-2**	
-2	**1**	

1	**-1**	**1**
-1	**1**	**-1**
1	**-1**	**1**

Table 3.1: The nine two-locus constraints: Each table corresponds to a nonnegative linear combination of z_{ij}, $i, j = 0, 1, 2$, representing an inequality . Constraints additional to those previously derived by Cordell at.al [10] and proved under general mating type frequencies by Dudoit and Speed [21] are printed in boldface.

by Cordell et al. [10]. We are able to construct examples of IBD probabilities satisfying the first set of constraints (3.4)-(3.7) but not the additional constraints (3.43)-(3.46).

To remember and understand the structure of the constraints, table 3.1 may be helpful. The new constraints are indicated in boldface. Note that while the old constraints are analogous to the single locus constraints and have fixed sharing at one locus, the new constraints involve varying sharing at both loci, and imply restrictions on interaction in the two-locus space.

3.3.3 Proof of Constraints on Sex-Specific Probabilities

First we will, under the assumption of no segregation distortion (assumption 1) alone, prove a lemma which is the key to the proof of inequalities (3.22)-(3.37). Lemma 1 and accompanying proof are a reproduction of a result proved by Dudoit and Speed [21] for a function of inheritance vectors at one locus under a multilocus model, here with slight changes to suit our two-locus model.

Lemma 1 *Let $c(x^1, x^2)$ be any function of the inheritance vectors at two disease loci. If, under assumption 1,*

$$\sum_{x^1}\sum_{x^2} c(x^1, x^2) \sum_{\{pg^1 \in mt^1\}} \sum_{\{pg^2 \in mt^2\}} \mathcal{P}(ASP|x^1, x^2, pg^1, pg^2) \geq 0 \qquad (3.48)$$

for all mt^1, mt^2, then

$$\sum_{x^1}\sum_{x^2} c(x^1, x^2)\mathcal{P}(x^1, x^2|ASP) \geq 0. \qquad (3.49)$$

Proof. With $x = (x^1, x^2)$, condition (3.48) implies for all mt^1, mt^2

$$\frac{\mathcal{P}(mt^1, mt^2)}{\#\{pg^1 \in mt^1\}\#\{pg^2 \in mt^2\}} \sum_x c(x) \sum_{\{pg^l \in mt^l, l=1,2\}} \mathcal{P}(\text{ASP}|x, pg^1, pg^2) \geq 0.$$

Using equation (2.5) for the probability of ordered parental genotypes and then taking the sum over all mt^1, mt^2 it follows that

$$\begin{aligned} 0 &\leq \sum_x c(x) \sum_{pg^1}\sum_{pg^2} \mathcal{P}(\text{ASP}|x, pg^1, pg^2)\mathcal{P}(pg^1, pg^2) \\ &= \sum_x c(x) \sum_{pg} \mathcal{P}(\text{ASP}|x, pg)\mathcal{P}(pg) \end{aligned} \qquad (3.50)$$

with $pg = (pg^1, pg^2)$. This is equivalent to

$$\begin{aligned} 0 &\leq \sum_x c(x) \sum_{pg} \mathcal{P}(x|\text{ASP}, pg)\mathcal{P}(\text{ASP}, pg)/\mathcal{P}(x|pg) \\ &= 16^2 \mathcal{P}(\text{ASP}) \sum_x c(x) \sum_{pg} \mathcal{P}(x|\text{ASP}, pg)\mathcal{P}(pg|\text{ASP}) \end{aligned}$$

where the last inequality follows since $\mathcal{P}(x|pg) = (\frac{1}{16})^2$ under assumption 1. Thus (3.49) holds.

$\square$

Now we are equipped to prove the constraints (3.22)-(3.37) assuming no segregation distortion (assumption 1), conditional independence of sibling phenotypes (assumption 2) and the monotonicity assumptions on the penetrances (assumption 3).

Proposition 2 *Under assumptions 1-3 and general mating type frequencies, the sex-specific conditional affected sib-pair IBD probabilities at two unlinked disease loci satisfy inequalities (3.22)-(3.37).*

Proof. Inequality (3.22) is trivial. Inequalities (3.23)-(3.28) involve probabilities with varying sharing at one locus and fixed sharing at the other locus, and thus they follow from proposition 1. How the proof was accomplished will be evident from the proof of the remaining inequalities. Now recall definition (2.6) of inheritance vectors and define

$$C_{i_1 j_1 i_2 j_2} = \{(x^1, x^2) : \delta(x^l_{p1}, x^l_{p2}) = j_l, \delta(x^l_{m1}, x^l_{m2}) = i_l,\ l = 1, 2\} \quad (3.51)$$

where $\delta(x, y)$ is equal to 1 if $x = y$, and 0 otherwise. Thus $C_{i_1 j_1 i_2 j_2}$ is a set of inheritance vectors corresponding to certain sex-specific IBD sharing. The sex-specific IBD probabilities are computed

$$\pi_{i_1 j_1 i_2 j_2} = \sum_{\{(x^1,x^2) \in C_{i_1 j_1 i_2 j_2}\}} \mathcal{P}(x^1, x^2|\text{ASP})$$

and hence in order to prove inequality (3.29), $\pi_{1010} - \pi_{0010} - \pi_{1000} + \pi_{0000} \geq 0$, we need to show

$$\sum_{C_{1010}} \mathcal{P}(x^1, x^2|\text{ASP}) - \sum_{C_{0010}} \mathcal{P}(x^1, x^2|\text{ASP})$$

$$- \sum_{C_{1000}} \mathcal{P}(x^1, x^2|\text{ASP}) + \sum_{C_{0000}} \mathcal{P}(x^1, x^2|\text{ASP}) \geq 0.$$

According to lemma 1 it suffices the show that for all $mt = \{mt^1, mt^2\}$

$$\sum_{\{pg \in mt\}} \left\{ \sum_{C_{1010}} \mathcal{P}(\text{ASP}|x^1, x^2, pg) - \sum_{C_{0010}} \mathcal{P}(\text{ASP}|x^1, x^2, pg) \right.$$

$$\left. - \sum_{C_{1000}} \mathcal{P}(ASP|x^1, x^2, pg) + \sum_{C_{0000}} \mathcal{P}(\text{ASP}|x^1, x^2, pg) \right\} \geq 0. \quad (3.52)$$

The calculations can be further simplified by the following observation. There are 16 combinations of inheritance vectors (x^1, x^2) in each $C_{i_1 j_1 i_2 j_2}$, but they can be obtained from each other by permuting 1 and 2 and/or 3 and 4 in the labeling of paternal chromosomes, at one or both loci. It follows that for (x^1, x^2) and $(\tilde{x^1}, \tilde{x^2}) \in C_{i_1 j_1 i_2 j_2}$

$$\sum_{\{pg \in mt\}} \mathcal{P}(\text{ASP}|\tilde{x^1}, \tilde{x^2}, pg) = \sum_{\{pg \in mt\}} \mathcal{P}(\text{ASP}|x^1, x^2, pg),$$

parental genotypes: locus 1	locus 2: $D^2_{i_2}D^2_{j_2}$, $D^2_{k_2}D^2_{l_2}$	$D^2_{i_2}D^2_{j_2}$, $D^2_{l_2}D^2_{k_2}$	$D^2_{j_2}D^2_{i_2}$, $D^2_{k_2}D^2_{l_2}$	$D^2_{j_2}D^2_{i_2}$, $D^2_{l_2}D^2_{k_2}$
$D^1_{i_1}D^1_{j_1} \times D^1_{k_1}D^1_{l_1}$	$f_{i_1k_1i_2k_2} \cdot f_{i_1l_1i_2l_2}$	$f_{i_1k_1i_2l_2} \cdot f_{i_1l_1i_2k_2}$	$f_{i_1k_1j_2k_2} \cdot f_{i_1l_1j_2l_2}$	$f_{i_1k_1j_2l_2} \cdot f_{i_1l_1j_2k_2}$
$D^1_{i_1}D^1_{j_1} \times D^1_{l_1}D^1_{k_1}$	$f_{i_1l_1i_2k_2} \cdot f_{i_1k_1i_2l_2}$	$f_{i_1l_1i_2l_2} \cdot f_{i_1k_1i_2l_2}$	$f_{i_1l_1j_2k_2} \cdot f_{i_1k_1j_2l_2}$	$f_{i_1l_1j_2l_2} \cdot f_{i_1k_1j_2k_2}$
$D^1_{j_1}D^1_{i_1} \times D^1_{k_1}D^1_{l_1}$	$f_{j_1k_1i_2k_2} \cdot f_{j_1l_1i_2l_2}$	$f_{j_1k_1i_2l_2} \cdot f_{j_1l_1i_2k_2}$	$f_{j_1k_1j_2k_2} \cdot f_{j_1l_1j_2l_2}$	$f_{j_1k_1j_2l_2} \cdot f_{j_1l_1j_2k_2}$
$D^1_{j_1}D^1_{i_1} \times D^1_{l_1}D^1_{k_1}$	$f_{j_1l_1i_2k_2} \cdot f_{j_1k_1i_2l_2}$	$f_{j_1l_1i_2l_2} \cdot f_{j_1k_1i_2k_2}$	$f_{j_1l_1j_2k_2} \cdot f_{j_1k_1j_2l_2}$	$f_{j_1l_1j_2l_2} \cdot f_{j_1k_1j_2k_2}$

Table 3.2: Example of $\mathcal{P}(\mathrm{ASP}|x^1, x^2, pg)$. The conditional probabilities of ASP given inheritance vectors, $x^1 = x^2 = (1,3,1,4)$, and ordered parental genotypes (the paternal above the maternal at locus 2), $pg = (pg^1, pg^2)$ for a representative mating $mt = (mt^1, mt^2)$.

and hence we only have to consider one representative combination of inheritance vectors for each $C_{i_1j_1i_2j_2}$. Hence, instead of (3.52) it suffices to show that for all mt

$$\begin{aligned}\textstyle\sum_{\{pg\in mt\}} \{&\mathcal{P}(\mathrm{ASP}|(1,3,1,4),(1,3,1,4),pg)\\ &-\mathcal{P}(\mathrm{ASP}|(1,3,2,4),(1,3,1,4),pg)\\ &-\mathcal{P}(\mathrm{ASP}|(1,3,1,4),(1,3,2,4),pg)\\ &+\mathcal{P}(\mathrm{ASP}|(1,3,2,4),(1,3,2,4),pg)\} \geq 0.\end{aligned} \tag{3.53}$$

Under assumption 2 the conditional probabilities of ASP given inheritance vectors and parental genotype factorise into products of two penetrances. As an example table (3.2) shows the conditional probabilities $\mathcal{P}(\mathrm{ASP}|x^1, x^2, pg)$ for $x^1 = x^2 = (1,3,1,4)$ and a general representative mating type. The left side of inequality (3.53) results in 64 such terms, which after some rearrangement factorise into

$$
\begin{aligned}
&2(f_{i_1k_1i_2k_2} - f_{i_1k_1j_2k_2} - f_{j_1k_1i_2k_2} + f_{j_1k_1j_2k_2})\times \\
&\quad(f_{i_1l_1i_2l_2} - f_{i_1l_1j_2l_2} - f_{j_1l_1i_2l_2} + f_{j_1l_1j_2l_2}) + \\
&2(f_{i_1k_1i_2l_2} - f_{i_1k_1j_2l_2} - f_{j_1k_1i_2l_2} + f_{j_1k_1j_2l_2}) \times \\
&\quad(f_{i_1l_1i_2k_2} - f_{i_1l_1j_2k_2} - f_{j_1l_1i_2k_2} + f_{j_1l_1j_2k_2}),
\end{aligned}
$$

where each product of four by four terms is nonnegative according to monotonicity property (2.11), and hence inequality (3.29) is proved.

The constraints (3.30)-(3.32) similarly lead to four products, each with two sets of four terms of penetrances. The desired inequalities then follow from (2.11).

To prove inequality (3.33) it suffices to show (cf. lemma 1), for a representative mating type mt,

$$
\sum_{C_{i_1j_1i_2j_2}} \sum_{\{pg\in mt\}} S(C_{i_1j_1i_2j_2})\mathcal{P}(\text{ASP}|x^1, x^2, pg) \geq 0, \qquad (3.54)
$$

where

$$
S(C_{i_1j_1i_2j_2}) = \begin{cases} -1 & \text{if } C_{i_1j_1i_2j_2} \in \{C_{0110}, C_{1010}, C_{1100}, C_{0000}\} \\ 1 & \text{if } C_{i_1j_1i_2j_2} \in \{C_{1110}, C_{0010}, C_{0100}, C_{1000}\} \\ 0 & \text{otherwise.} \end{cases}
$$

Again we only have to consider one representative combination of inheritance vectors for each $C_{i_1j_1i_2j_2}$. The left side of (3.54) results in 128 products of two penetrances each, which factorise into

$$
\begin{aligned}
&2(f_{i_1k_1i_2k_2} - f_{j_1k_1i_2k_2} - f_{i_1l_1i_2k_2} - f_{i_1k_1j_2k_2} \\
&+ f_{j_1l_1i_2k_2} + f_{j_1k_1j_2k_2} + f_{i_1l_1j_2k_2} - f_{j_1l_1j_2k_2}) \\
&\times(f_{i_1k_1i_2l_2} - f_{j_1k_1i_2l_2} - f_{i_1l_1i_2l_2} - f_{i_1k_1j_2l_2} \\
&+ f_{j_1l_1i_2l_2} + f_{j_1k_1j_2l_2} + f_{i_1l_1j_2l_2} - f_{j_1l_1j_2l_2}).
\end{aligned}
$$

This is nonnegative under assumption (2.12). Similarly inequalities (3.34)-(3.36) lead to two products, each with two sets of eight terms, which are nonnegative under assumption (2.12).

A proof of inequality (3.37) along the same lines leads to 256 penetrance products which factorise into a square of 16 terms. Hence no penetrance monotonicity assumptions are required to prove this constraint.

□

1	$nn \times nn$
2	$nn \times nd, nn \times dn$
3	$nd \times nn, dn \times nn$
4	$nn \times dd$
5	$dd \times nn$
6	$nd \times nd, nd \times dn$ $dn \times nd, dn \times dn$
7	$nd \times dd, dn \times dd$
8	$dd \times nd, dd \times dn$
9	$dd \times dd$

(a)

$p_1 = p_{66}$
$p_2 = p_{67} = p_{68}$
$p_3 = p_{69}$
$p_4 = p_{76} = p_{86}$
$p_5 = p_{77} = p_{78} = p_{87} = p_{88}$
$p_6 = p_{79} = p_{89}$
$p_7 = p_{96}$
$p_8 = p_{97} = p_{98}$
$p_9 = p_{99}$

(b)

Table 3.3: **(a)** The nine mating types at one bi-allelic disease locus. **(b)** The 16 two-locus mating type probabilities that contribute to ASP IBD probabilities for a jointly recessive model. Grouping of the probabilities into 9 classes.

Note that if we take phase into account in our two-locus notation for genotypes we can not use equation (2.5) and we can then not use the simplification that follows from lemma 1. Proposition 1 still follows, but we have to work with the complete summation in the right side of inequality (3.50) to prove it.

3.3.4 Coverage of Parameter Space

We will now show that the ASP IBD parameter space is covered by a jointly recessive model, table 2.5 (a), under general mating type frequencies, i.e. that all IBD probabilities satisfying the constraints can be obtained under this model. The nine single locus mating types, table 3.3 (a), can be combined into 81 two-locus parental mating types, (mt^1, mt^2). Let $p_{ij} = \mathcal{P}(mt^1 = i, mt^2 = j)$ for $i, j = 1, \ldots, 9$. If the father for example has genotype nd at locus 1 and dd at locus 2, while the mother has genotype nd at both loci, the parental mating type is $(6, 8)$. For the jointly recessive model only mating types where both parents have at least one disease allele at both loci can contribute with positive terms in calculation of conditional probabilities of inheritance vectors. Hence we only have to consider p_{ij} for $i, j \geq 6$. The calculations can be further simplified if we do not distinguish between single locus mating types 7 and 8, which is natural when the penetrance function does not account for parental imprinting. This way we end up with nine two-locus mating types relevant for the model we are considering, whose probabilities we rename according to table 3.3 (b). When conditional probabilities of inheritance vectors given ASP are calculated, using equation (2.8), and summed, we get

$$
\boldsymbol{z} = \frac{1}{D} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 16 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 16 & 32 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 16 & 16 \\ 0 & 0 & 0 & 0 & 0 & 16 & 0 & 0 & 32 \\ 0 & 0 & 0 & 0 & 16 & 32 & 0 & 32 & 64 \\ 0 & 0 & 0 & 4 & 16 & 16 & 8 & 32 & 32 \\ 0 & 0 & 4 & 0 & 0 & 16 & 0 & 0 & 16 \\ 0 & 4 & 8 & 0 & 16 & 32 & 0 & 16 & 32 \\ 1 & 4 & 4 & 4 & 16 & 16 & 4 & 16 & 16 \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_9 \end{bmatrix}, \tag{3.55}
$$

where D is a constant. This relation can be inverted to

$$
\boldsymbol{p} = \frac{c}{16} \begin{bmatrix} 16(z_{00}-z_{01}+z_{02}-z_{10}+z_{11}-z_{12}+z_{20}-z_{21}+z_{22}) \\ 4(-2z_{00}+z_{01}+2z_{10}-z_{11}-2z_{20}+z_{21}) \\ 4(z_{00}-z_{10}+z_{20}) \\ 4(-2z_{00}+2z_{01}-2z_{02}+z_{10}-z_{11}+z_{12}) \\ 4z_{00}-2z_{10}-2z_{10}+z_{11} \\ z_{10}-2z_{00} \\ 4(z_{00}-z_{01}+z_{02}) \\ z_{01}-2z_{00} \\ z_{00} \end{bmatrix},
$$

where $\boldsymbol{p} = (p_1, \dots, p_9)$ and c is constant such that $\sum_{i=1}^{9} p_i \leq 1$. Note that the linear combinations of IBD probabilities defining $\boldsymbol{p}$, are multiples of exactly those occurring in inequalities (3.38) - (3.46). Hence the mating type frequencies will be positive for all z_{ij} that satisfy the two-locus constraints. With mating type frequencies constructed this way, equation (3.55) gives us any $z_{00}, \dots, z_{22}$ satisfying the constraints, which was to be proved.

3.3.5 Necessity of the Penetrance Assumptions

Here we give counter examples showing the necessity of the penetrance assumptions for the two-locus constraints to hold under general mating type frequencies.

In the proof of inequalities (3.23) - (3.26), equivalent to (3.39) and (3.40) in the non-sex-specific setting, assumption (2.9) is utilised. The overdominance model in table 2.6 (a) is one example of a model not satisfying this assumption. If ASP IBD probabilities are calculated for general mating type frequencies, p_{ij}, $i,j = 1, \dots, 9$ under this model we get

$$
\begin{aligned}
&z_{01} - 2z_{00} \\
&\quad = \tfrac{1}{d}\{4(2p_{42} + 2p_{43} + 2p_{52} + 2p_{53} + p_{62} + p_{63} + 2p_{72} + 2p_{73} - 2p_{76} \\
&\qquad +2p_{82} + 2p_{83} - 2p_{86} + 2p_{92} + 2p_{93} - 4p_{96} + 2p_{97} + 2p_{98}\}
\end{aligned}
$$

where d is a positive normalising constant. We see that this difference contains a few negative terms and hence the mating type frequencies can be chosen such that $z_{01} - 2z_{00} < 0$.

In proving (3.27) and (3.28), equivalent to (3.41) and (3.42), we used assumption (2.10). To see that this was necessary, IBD probabilities were calculated under the modifier model, 2.6 (d), which satisfies assumption (2.9) but not (2.10). Then, e.g. $z_{00} - z_{01} + z_{02}$ contains negative terms, and hence the mating type frequencies can be chosen such that $z_{00} - z_{01} + z_{02} < 0$.

Model (a) in table 2.7, satisfies (2.9) and (2.10), but not (2.11), which was used in the proof of (3.29)-(3.32), leading to (3.43). IBD probabilities calculated under this model makes it possible to chose mating type frequencies such that $4z_{00} - 2z_{10} - 2z_{10} + z_{11} < 0$. Finally, model (b) in table 2.7 satisfy (2.9)-(2.11), but not (2.12) which was used in the proof of (3.33) - (3.36), or (3.44) and (3.45). Consequently, we can chose mating type frequencies such that for instance $-2z_{00}+2z_{01}-2z_{02}+z_{10}-z_{11}+z_{12} < 0$.

3.4 Interpretation of Parameters

In this section we derive an interpretation of the parameters in our representation. We will see that the parameters are conditional probabilities of certain partitions of the parental mating types and sibling genotypes, given that a sibling pair is affected. The interpretation is artificial as the partitioning of the genotype information depends on the genetic model.

Parts of this section are sketchy and based on examples, and are merely intended to give an idea of what the parameters represents, or rather, indicate the lack of a completely satisfactory interpretation.

In chapter 4 an interpretation will be given in terms of genetic variance components under the equilibrium model.

3.4.1 Single-Locus Sex-Specific Case

Let *SIBD* $= (i, \cdot)$ [or *SIBD* $= (\cdot, j)$] denote the marginal event that a sib-pair shares i parental [j maternal] alleles IBD at a locus, $i, j = 0, 1$. Define the following notation for the elements of the matrix of parameterisation (3.12)

$$\boldsymbol{A_s} = \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} = \begin{bmatrix} a_{00,\_\_} & a_{00,\_1} & a_{00,1\_} & a_{00,11} \\ a_{01,\_\_} & a_{01,\_1} & a_{01,1\_} & a_{01,11} \\ a_{10,\_\_} & a_{10,\_1} & a_{10,1\_} & a_{10,11} \\ a_{11,\_\_} & a_{11,\_1} & a_{11,1\_} & a_{11,11} \end{bmatrix}.$$

Note that

$$
\begin{aligned}
a_{ij,\_\_} &= \mathcal{P}(\mathit{SIBD} = (i,j)) \\
a_{ij,\_1} &= \mathcal{P}(\mathit{SIBD} = (i,j)|\mathit{SIBD} = (\cdot,1)) \\
a_{ij,1\_} &= \mathcal{P}(\mathit{SIBD} = (i,j)|\mathit{SIBD} = (1,\cdot)) \\
a_{ij,\_\_} &= \mathcal{P}(\mathit{SIBD} = (i,j)|\mathit{SIBD} = (1,1))
\end{aligned}
\tag{3.56}
$$

Thus, to obtain representation (3.12) we need to condition on information that can be partitioned into

$$
G = \{G_{\_\_}, G_{\_1}, G_{1\_}, G_{11}\} \tag{3.57}
$$

in such a way that

$$
\begin{aligned}
\pi_{ij} &= \mathcal{P}(\mathit{SIBD} = (i,j)|\mathrm{ASP}) \\
&= \sum_G \mathcal{P}(\mathit{SIBD} = (i,j)|G, \mathrm{ASP})\mathcal{P}(G|\mathrm{ASP})
\end{aligned}
\tag{3.58}
$$

and

$$
\begin{aligned}
&\mathcal{P}(\mathit{SIBD} = (i,j)|G_{\_\_}, \mathrm{ASP}) = \tfrac{1}{4} \\
&\mathcal{P}(\mathit{SIBD} = (i,j)|G_{\_1}, \mathrm{ASP}) = \begin{cases} 0 & \text{if } j = 0 \\ \frac{1}{2} & \text{if } j = 1 \end{cases} \\
&\mathcal{P}(\mathit{SIBD} = (i,j)|G_{1\_}, \mathrm{ASP}) = \begin{cases} 0 & \text{if } i = 0 \\ \frac{1}{2} & \text{if } i = 1 \end{cases} \\
&\mathcal{P}(\mathit{SIBD} = (i,j)|G_{11}, ASP) = \begin{cases} 1 & \text{if } i = j = 1 \\ 0 & \text{otherwise} \end{cases} .
\end{aligned}
\tag{3.59}
$$

The parameters q_{ij}, $i,j = 0,1$ of representation (3.13) are now given by

$$
(q_{00}, q_{01}, q_{10}, q_{11}) =
$$

$$
(\mathcal{P}(G_{\_\_}|\mathrm{ASP}), \mathcal{P}(G_{\_1}|\mathrm{ASP}), \mathcal{P}(G_{1\_}|\mathrm{ASP}), \mathcal{P}(G_{11}|\mathrm{ASP})). \tag{3.60}
$$

This interpretation is probably of little practical use, since what type of information, G, we condition on and how it should be partitioned depends on the unknown genetic model. However, for a strict recessive model, we will see that G corresponds to a quite natural partition of the parental mating types.

Strict Recessive model

For a strict recessive penetrance model ($f_{nn} = f_{nd} = 0$, $f_{dd} = f$) we may condition on each of the four possible parental mating types, $mt_6, \ldots, mt_9$ in table 3.4. Given mt_6 an affected sib-pair must share two alleles IBD, given mt_7 we know that they share the paternal allele and so on (see column three in table 3.4). It is intuitively clear that parental mating types and the fact that the sibs are affected give no further information about IBD in addition to what can be said with certainty (column 3). Consequently equation (3.59) is true if we let

$$G = \{mt_9, mt_8, mt_7, mt_6\},$$

with the order of the partitions as defined by (3.57).

Given q_{ij}, $i, j = 0, 1$, as defined in (3.60), IBD probabilities for the strict recessive model may now be calculated using equation (3.58) or equivalently representation (3.13).

We may check that (3.59) is satisfied with the following calculations.

$$\begin{aligned} & \mathcal{P}(\mathit{SIBD} = (i,j)|mt, \mathrm{ASP}) \\ = \ & \frac{\mathcal{P}(\mathrm{ASP}|mt, \mathit{SIBD} = (i,j))\mathcal{P}(mt, \mathit{SIBD} = (i,j))}{\mathcal{P}(\mathrm{ASP}|mt)\mathcal{P}(mt)}. \end{aligned} \tag{3.61}$$

Under assumption 1 (no segregation distortion) equation (3.61) simplifies to

$$\mathcal{P}(\mathit{SIBD} = (i,j)|mt, \mathrm{ASP}) = \frac{\mathcal{P}(\mathrm{ASP}|mt, \mathit{SIBD} = (i,j))\frac{1}{4}}{\mathcal{P}(\mathrm{ASP}|mt)}$$

As an example

$$\begin{aligned} \mathcal{P}(\mathrm{ASP}|mt_7, \mathit{SIBD} &= (1,1)) = \frac{f^2}{2}, \\ \mathcal{P}(\mathrm{ASP}|mt_7) &= \frac{f^2}{4} \end{aligned} \tag{3.62}$$

and hence

$$\mathcal{P}(\mathit{SIBD} = (1,1)|mt_7, \mathrm{ASP}) = \frac{1}{2}.$$

The parameter values for the recessive model are calculated by

$$\mathcal{P}(mt|\mathrm{ASP}) = \frac{\mathcal{P}(\mathrm{ASP}|mt)\mathcal{P}(mt)}{\mathcal{P}(\mathrm{ASP})}. \tag{3.63}$$

mt	mt	*SIBD* $\mid$ ASP_{rec}, mt	possible sg $\mid$ mt	*SIBD* $\mid$ mt, sg
1	$nn \times nn$	–	(nn, nn)	$(\cdot, \cdot)$
2	$nn \times nd$	–	(nn, nn)	$(\cdot, 1)$
	$nn \times dn$		$(nn, nd), (nd, nn)$	$(\cdot, 0)$
			(nd, nd)	$(\cdot, 1)$
3	$nd \times nn$	–	(nn, nn)	$(1, \cdot)$
	$dn \times nn$		$(nn, nd), (nd, nn)$	$(0, -)$
			(nd, nd)	$(1, \cdot)$
4	$nn \times dd$	–	(nd, nd)	$(\cdot, \cdot)$
5	$dd \times nn$	–	(nd, nd)	$(\cdot, \cdot)$
6	$nd \times nd$	$(1, 1)$	(nn, nn)	$(1, 1)$
	$nd \times dn$		$(nn, nd), (nd, nn)$	$(\cdot, \cdot)$
	$dn \times nd$		(nn, dd)	$(0, 0)$
	$dn \times dn$		(nd, nd)	$(\cdot, \cdot)$
			$(nd, dd), (dd, nd)$	$(\cdot, \cdot)$
			(dd, dd)	$(1, 1)$
7	$nd \times dd$	$(1, \cdot)$	(nd, nd)	$(1, \cdot)$
	$dn \times dd$		$(nd, dd), (dd, nd)$	$(0, \cdot)$
			(dd, dd)	$(1, \cdot)$
8	$dd \times nd$	$(\cdot, 1)$	(nd, nd)	$(\cdot, 1)$
	$dd \times dn$		$(nd, dd), (dd, nd)$	$(\cdot, 0)$
			(dd, dd)	$(\cdot, 1)$
9	$dd \times dd$	$(\cdot, \cdot)$	(dd, dd)	$(\cdot, \cdot)$

Table 3.4: Parental mating types at a bi-alleleic disease locus. column 2: The ordered parental genotypes in the mating type. column 3: Known sex-specific IBD sharing for a sib-pair affected by a strict recessive disease, given the parental mating type. Here '–' means that the mating type is not possible for parents of an affected sib-pair. column 4: The possible (unordered) sibling genotypes, given parental mating type. column 5: Known sex-specific IBD sharing, given parental mating types and sibling genotypes.

As an example

$$\mathcal{P}(\text{ASP}|mt_6) = \frac{1}{16}f^2.$$

Under RM and HWE, $\mathcal{P}(mt_6) = 4p^2(1-p)^2$ and $\mathcal{P}(ASP) = \frac{p^2(p+1)^2}{4}f^2$, which result in

$$q_{11} = \mathcal{P}(mt_6|\text{ASP}) = \frac{(1-p)^2}{(1+p)^2}.$$

With similar calculations we get

$$(q_{00}, q_{01}, q_{10}, q_{11}) = \left(\frac{4p^2}{(p+1)^2}, \frac{2p(1-p)}{(p+1)^2}, \frac{2p(1-p)}{(p+1)^2}, \frac{(1-p)^2}{(p+1)^2}\right). \quad (3.64)$$

Nonsex-Specific Case

The matrix defining the nonsex-specific representation (3.10) is

$$\boldsymbol{A_1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix}. \quad (3.65)$$

Let a_{ij}, $i,j = 0,1,2$ be the elements of this matrix. To obtain representation (3.10) we need to condition on information partitioned into[1] $G = \{G_0, G_1, G_2\}$, where $G_0 = G_{\_\_}$, $G_1 = G_{1\_} \cup G_{\_1}$ and $G_2 = G_{11}$. Now, representation (3.10) can be written

$$\mathcal{P}(\mathit{IBD} = i|\text{ASP}) = \sum_j \mathcal{P}(\mathit{IBD} = i|G_j, \text{ASP})\mathcal{P}(G_j|\text{ASP}),$$

where $\mathcal{P}(\mathit{IBD} = i|G_j, \text{ASP}) = a_{ij}$. The probabilities a_{ij}, $i,j = 0,1,2$ are obtained from the sex-specific representation.

For instance,

$$a_{11} = \mathcal{P}(\mathit{IBD} = 1|G_1, \text{ASP}) =$$

$$\frac{1}{2}(\mathcal{P}(\mathit{SIBD} = (1,0)|G_{1\_}) + \mathcal{P}(\mathit{SIBD} = (1,0)|G_{\_1}) +$$

$$\mathcal{P}(\mathit{SIBD} = (0,1)|G_{1\_}) + \mathcal{P}(\mathit{SIBD} = (0,1)|G_{\_1})) = \frac{1}{2}.$$

Here we used $\mathcal{P}(G_{1\_}|G_1) = \frac{1}{2}$, which is true for our partitioning in the case of a recessive disease, with $G_1 = mt_7 \cup mt_8$.

[1] In what follows it will be convenient to use G_0 instead of $G_{\_}$, which would have been more consistent. We will have to remember that G_0 denotes genotypes giving no certain IBD information and does *not* mean genotypes implying IBD 0.

3.4.2 Two-locus case

The sex-specific two-locus representation (3.20) is defined by the 16×16 matrix $\boldsymbol{A_{s2}} = \boldsymbol{A_s}^{\otimes 2}$. The elements of $\boldsymbol{A_{s2}}$ can be seen to be probabilities of two-locus sex-specific IBD sharing, either given no extra information or given paternal and/or maternal sharing at one or both loci.

Similar to the one-locus interpretation (3.58), we may write

$$\mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|\mathrm{ASP}) =$$

$$\sum_G \mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|G, \mathrm{ASP})\mathcal{P}(G|\mathrm{ASP}). \qquad (3.66)$$

In this case we condition on information partitioned into 16 classes,

$$G = \{G_{\_\_\_\_}, G_{\_\_\_1}, G_{\_\_1\_}, G_{\_\_11}, G_{\_1\_\_}, \ldots, G_{111\_}, G_{1111}\},$$

and the conditional probabilities of these partitions, given ASP, are the parameters $q_{i_1 j_1 i_2 j_2}$, for $i_1, j_1, i_2, j_2 = 0, 1$. $G_{\_\_\_1}$, for instance, represents genotype information that specifies that the maternal allele at locus 2 is IBD, but does not give certain information about IBD at locus 1 and the paternal allele at locus 2. From the construction of $\boldsymbol{A_{s2}}$ as the Kronecker product of the single locus matrix follows that we must be able to express $(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|G,\ \mathrm{ASP})$ as products of the elements of the single-locus matrix (3.56),

$$\mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|G_{\_\_\_\_}, \mathrm{ASP}) = a_{i_1 j_1, \_\_} a_{i_2, j_2, \_\_}$$

$$\mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|G_{\_\_\_1}, \mathrm{ASP}) = a_{i_1 j_1, \_\_} a_{i_2, j_2, \_1}$$

$$\vdots$$

$$\mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|G_{111\_}, \mathrm{ASP}) = a_{i_1 j_1, 11} a_{i_2, j_2, 1\_}$$

$$\mathcal{P}(\mathit{SIBD} = (i_1, j_1, i_2, j_2)|G_{1111}, \mathrm{ASP}) = a_{i_1 j_1, 11} a_{i_2, j_2, 11}.$$

Hence, the partitioning of G must be such that we get conditional independence between IBD status at the two loci, given this information and ASP. For a jointly recessive two-locus model (Table 2.5 (a)) representation (3.66), or (3.20), is obtained by letting

$$G = \{(mt_9^1, mt_9^2), (mt_9^1, mt_8^2), (mt_9^1, mt_7^2), (mt_9^1, mt_6^2),$$

$$(mt_8^1, mt_9^2), \ldots, (mt_6^1, mt_7^2), (mt_6^1, mt_6^2)\}.$$

Two-locus nonsex-specific case

In the nonsex-specific case we let $G_{00} = G_{\_\_\_\_}$, $G_{01} = G_{\_\_\_1} \cup G_{\_\_1\_}$, $G_{02} = G_{\_\_11}$, etc to get

$$\mathcal{P}(IBD = (i,j)|\mathrm{ASP}) = \sum_{k=0}^{2}\sum_{l=0}^{2} \mathcal{P}(IBD = (i,j)|G_{kl}, \mathrm{ASP})\mathcal{P}(G_{kl}|\mathrm{ASP}) =$$

$$\sum_{k=0}^{2}\sum_{l=0}^{2} a_{ik}a_{jl}\mathcal{P}(G_{kl}|\mathrm{ASP}),$$

where a_{ij} are the elements from the single locus matrix defined in equation (3.65). Now we can interpret the parameters as

$$\mathcal{P}(G_{kl}|\mathrm{ASP}) = q_{kl}$$

for $k, l = 0, 1, 2$.

3.4.3 Non-Recessive Models

While the partitioning of the parental genotypes to obtain our representation is quite natural for the recessive model, it is not as obvious how to proceed for other models. Consider for example a single-locus dominant model ($f_{nn} = 0$, $f_{nd} = f_{dd} = f$). First, since only one disease allele is required for an individual to develop the disease, it exists no parental mating type such the fact that the sibs are affected implies that they have to share two alleles IBD. It follows that in order to obtain G_{11} we will have to condition also on sibling genotypes. Consider the fifth column in table 3.4 where we have listed what is known about sex-specific IBD status, given parental mating type and (unordered) sibling genotypes. For some genotype combinations we know with certainty that a maternal/paternal allele is not IBD. This information class does not fit into our representation (3.58), and accordingly G cannot be a partitioning of all combinations of parental mating types and sibling genotypes. G must be composed from some partitions determined solely by mating types and some which are determined by both mating types and sibling genotypes.

It is unclear how to obtain the partitions in general. However, merely as an example to get an idea of what such partitions might look like, we will display a partition of mating types and sibling genotypes that give us representation (3.58) for a strict recessive model.

$$
\begin{aligned}
G_{--} &= \{mt_1, mt_4, mt_5, mt_7, mt_8, mt_9\} \cup \{mt_6, sgt \neq (dd, dd)\} \\
G_{-1} &= \{mt_2\} \\
G_{1-} &= \{mt_3\} \\
G_{11} &= \{mt_6, sgt = (dd, dd)\}.
\end{aligned}
$$

This partitioning was validated by comparison of the resulting parameter values, q_{ij}, $i, j = 0, 1$, to those obtained from the variance component approach described in section 4.2.2.

Next, consider a model with phenocopies, that is when an individual may have the disease independent of his genotype. We define phenocopy status as an indicator, independent of genotypes, for the event that a sibling is affected from other causes than the disease locus under study. Preliminary calculations suggest that a partitioning leading to our representation may be obtained by letting the partitions be combinations of mating types, genotypes and the four possible combinations of phenocopy states for the two siblings.

Finally consider an over dominant model, which may be instructive as a new insight as to why the penetrance monotonicity properties (assumption 3) are needed.

Under a dominant model we know for instance that

$$\mathcal{P}(\mathit{SIBD} = (i, j) | mt_7, \mathrm{ASP}) =$$

$$\mathcal{P}(IBD = (i, j) | pg \in \{nd \times dd, dn \times dd\}, \mathrm{ASP}) = \frac{1}{4},$$

for all $i, j = 0, 1$. Parental mating types and ASP give no more information about IBD in addition to what can be said with certainty. For an overdominant model this is however not the case. Let for instance $f_{nn} = 0$, $f_{nd} = f$ and $f_{dd} = f/2$. Then we have

$$\mathcal{P}(\mathit{SIBD} = (1, 1) | pg \in \{nd \times dd, dn \times dd\}, \mathrm{ASP}) =$$

$$\frac{\mathcal{P}(\mathrm{ASP} | \mathit{SIBD} = (1, 1), pg \in \{nd \times dd, dn \times dd\}) \mathcal{P}(\mathit{SIBD} = (1, 1))}{\mathcal{P}(\mathrm{ASP} | pg \in \{nd \times dd, dn \times dd\})}$$

$$= \frac{\frac{1}{2}(f^2 + \frac{1}{4}f^2)\frac{1}{4}}{\frac{1}{4}(\frac{1}{2}f^2 + \frac{1}{2}f^2 + f^2 + \frac{1}{4}f^2)} = \frac{5}{18}$$

and (3.59) is not satisfied. This suggests that asymmetries of this type may lead to IBD probabilities that do not satisfy our constraints.

Chapter 4

IBD Probabilities under the Equilibrium Model

The constraints on IBD probabilities proved in chapter 3 are true also without the penetrance monotonicity assumptions, provided that the population genetic assumptions, HWE, RM and LE are satisfied. Now, the the following proposition will be proved.

Proposition 3 *Under assumptions 1 and 2, Hardy-Weinberg equilibrium, random mating and linkage equilibrium, the sex-specific conditional affected sib-pair IBD probabilities at two unlinked disease loci satisfy inequalities (3.22)-(3.37).*

Proposition 3 will be proved in two ways. First by expression of the IBD probabilities in terms of genetic variance components, second by a direct method. The reason for including both approaches is that while the direct proof is mathematically appealing, the variance component approach is instructive. It was used by Holmans [41], when deriving the possible triangle and by Cordell et al. [10] in derivation of the first set of two-locus constraints (proposition 1). Furthermore, it gives us an interpretation, under the equilibrium model, of the parameters of the general representation.

The general structure and the key needed to bring the direct proof to its conclusion were observed and provided by Terry Speed.

4.1 Genetic Variance Components

Models of quantitative phenotypes are often composed by a genotypic value and a residual environmental deviation. The genotypic value is the sum of the effects from all loci influencing the phenotype. It may be further decomposed into additive, dominance and epistatic effects. For a more solid introduction to these concepts and the genetic variance components described in this section, see for instance the comprehensive treatment of quantitative genetics by Lynch and Walsh [54].

For a phenotype whose genetic component is determined by a single locus, let y_{ij} be the genotypic value (deviation from the overall population mean) for a paternal allele D_i and a maternal allele D_j, $i, j = 1, \dots, m$. We decompose y_{ij} into additive effects, μ_i and μ_j , and a residual dominance deviation, δ_{ij}, as

$$y_{ij} = \mu_i + \mu_j + \delta_{ij}, \tag{4.1}$$

for, $i, j = 1, \dots, m$. The additive effects are chosen to minimise the residual sum of squares, weighted by the population genotype frequencies. This weighted sum of squares is the dominance variance, σ_D^2, i.e. the variance of δ_{ij}. The additive effect μ_i (μ_j), of a paternal (maternal) allele, can be shown to be the difference between the mean phenotype of individuals who have inherited the allele from the father (mother), and the population mean phenotype. Under RM and HWE the total genetic variance, σ_G^2, may be decomposed into

$$\sigma_G^2 = \sigma_{A^p}^2 + \sigma_{A^m}^2 + \sigma_D^2,$$

where $\sigma_{A^p}^2$ ($\sigma_{A^m}^2$) is the population variance of the additive paternal (maternal) effects. The sex-specific additive components are usually summed to a single additive variance component, σ_A^2.

We will use variance component approach for qualitative traits by exchanging the continuous responses for penetrances. The variance components are then functions of penetrances and allele frequencies. As an example, for a bi-allelic locus where the population frequency of the disease allele, d, is p, the variance components have the following simple forms (Suarez et al. [71])

$$\begin{aligned} \sigma_A^2 &= 2p(1-p)\left[(1-p)(f_{nn} - f_{nd}) + p(f_{nd} - f_{dd})\right]^2, \\ \sigma_D^2 &= p^2(1-p)^2\left[f_{nn} - 2f_{nd} + f_{dd}\right]^2. \end{aligned}$$

When the linear model (4.1) is extended to two unlinked loci, epistatic terms are added for the interaction effects between one allele at each locus (add$\times$add), interaction between genotypes at one locus and one allele at the other (dom$\times$add) and a residual (dom$\times$dom) term. For alleles $D_{i_1}^1, D_{j_1}^1$ and $D_{i_2}^2, D_{j_2}^2$ at two unlinked loci, we decompose the genotypic value $y_{i_1 j_1, i_2 j_2}$

into fifteen terms,

$$\begin{aligned}
y_{i_1j_1,i_2j_2} &= \mu_{i_1}+\mu_{j_1}+\mu_{i_2}+\mu_{j_2} && \text{(add)}\\
&+ \delta_{i_1j_1}+\delta_{i_2j_2} && \text{(dom)}\\
&+ (\mu\mu)_{i_1,i_2}+(\mu\mu)_{i_1,j_2}+(\mu\mu)_{j_1,i_2}+(\mu\mu)_{j_1,j_2} && \text{(add} \times \text{add)}\\
&+ (\mu\delta)_{i_1,i_2j_2}+(\mu\delta)_{j_1,i_2j_2}+(\delta\mu)_{i_1j_1,i_2}(\delta\mu)_{i_1j_1,j_2} && \text{(add} \times \text{dom)}\\
&+ (\delta\delta)_{i_1j_1,i_2j_2}. && \text{(dom} \times \text{dom)}
\end{aligned} \tag{4.2}$$

As in the single locus case, an additive effect is the difference between the mean phenotype of individuals possessing the allele and the population mean, for instance the additive effect of allele i_1 at locus 1, inherited from the father, is

$$\mu_{i_1} = \bar{y}_{i_1\ldots} - \bar{y}_{\ldots.}.$$

A dominance effect is the difference between the conditional mean, given two alleles at a locus, and the additive effects of the alleles, for instance

$$\delta_{i_1j_1} = \bar{y}_{i_1j_1..} - \mu_{i_1} - \mu_{j_1}.$$

The epistatic terms are defined in a similar fashion as the differences between the conditional means, given the alleles and/or genotypes indicated by the indices, and the lower order terms.

Provided that we have HWE, RM and LE, the total genetic variance in the population is the sum of the variances of the individual effects. We will use the decomposition into all fifteen components when proving the two-locus constraints on IBD probabilities, but usually the terms are grouped according to equation (4.2), resulting in five variance components,

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2. \tag{4.3}$$

The variance decomposition for one locus dates back to Fisher in 1918 [30]. Kempthorne [44] derived the multilocus extension in the fifties. More recently Tiwari and Elston [74] described a less strenuous, general method to calculate the components for any multilocus model in terms of allele frequencies and phenotype means or penetrances.

Numerical values of variance components may be obtained from the sums of squares from an analysis of variance in any statistical software, such as SAS. This is done in a weighted analysis with all interactions between four factors, where the factors are the paternal and maternal alleles at the loci and the number of levels is the number of alleles. The weights used are the products of the frequencies for the allele combinations in question, which are the two-locus genotype frequencies under HWE, RM and LE.

It should be emphasised that variance components are population parameters. The relative magnitude of the variance components may be considerably different for the same penetrance function, but with different allele frequencies. Relatively small epistatic variance components are not incompatible with a strong physiological interaction between two genes (Lynch and Walsh [54]).

4.2 IBD Probabilities Expressed in Variance Components

To express IBD probabilities in terms of genetic variance components we will employ the approach used by Cordell et al. [10]. The IBD probabilities are in an intermediate step expressed in terms of population disease prevelance and risks for individuals with a certain IBD sharing pattern with an affected individual.

Let ϕ_1 and ϕ_2 denote the affection status of two random relatives of type R, where $\phi_i = 1$ if individual i is affected and $\phi_i = 0$ otherwise. Then the population prevalence is

$$K = E[\phi_i]$$

and the recurrence risk for a relative of type R of an affected individual is

$$K_R = E[\phi_2 | \phi_1 = 1].$$

As noted by James [43], it follows from the definition of covariance that the recurrence risk may be written

$$K_R = K + \frac{1}{K}\mathrm{cov}_R(\phi_1, \phi_2). \tag{4.4}$$

Recall that we are working under the assumption of no environmental covariance (assumption 2). Hence, $\mathrm{cov}_R(\phi_1, \phi_2)$ is pure genetic covariance.

4.2.1 Genetic Covariance Between Relatives

As will be seen below, $\mathrm{cov}_R(\phi_1, \phi_2)$, for various true or artificial relationships, are essential components of ASP IBD probabilities.

Kempthorne [44, 45] derived a general formula for genetic covariance in terms of variance components for any type of relatives and any number of unlinked loci under RM and LE. We will use this formula and its derivation.

The mean value of any effect in equation (4.2) is by definition zero. Hence the covariance of the genotypic values for two individuals is the expectation of the crossproduct of the terms in equation (4.2). Under the equilibrium assumptions all effects are uncorrelated both between and within subjects. It follows that the expectation of any product is equal to zero unless the effects are of the same type and the indicated alleles are IBD, in which case the expectation is a variance.

Let c_R be the probability that an allele drawn at random from an individual is IBD to any one of the two alleles in a relative of type R, and let τ_R be the probability that the relatives share two alleles IBD. These probabilities are listed for MZ twins, parent-offspring and siblings in table 4.1.

relationship	c_R	τ_R
MZ twins	1	1
parent-offspring	$\frac{1}{2}$	0
siblings	$\frac{1}{2}$	$\frac{1}{4}$

Table 4.1: Expected proportion of alleles IBD, c_R, and probability for two alleles IBD, τ_R, for three relationship types.

Additive effects involve single alleles and are IBD with probability c_R. Dominance effects are IBD if both involved alleles are IBD, which occur with probability τ_R. It follows that the genetic covariance, for a trait influenced by two loci may be expressed

$$\mathrm{cov}_R(\phi_1, \phi_2) = c_R\sigma_A^2 + \tau_R\sigma_D^2 + c_R^2\sigma_{AA}^2 + c_R\tau_R\sigma_{AD}^2 + \tau_R^2\sigma_{DD}^2. \qquad (4.5)$$

For a more detailed derivation of equation (4.5) we refer to the original work by Kempthorne [44] or the clear description by Lynch and Walsh [54].

4.2.2 Single Locus IBD Probabilities in Terms of Variance Components

The single locus sex-specific IBD probabilities may be written

$$\begin{aligned}\pi_{ij} &= \frac{\mathcal{P}(\mathrm{ASP}|\mathit{SIBD} = (i,j))\mathcal{P}(\mathit{SIBD} = (i,j))}{\mathcal{P}(\mathrm{ASP})} \\ &= \frac{\mathcal{P}(\phi_1 = 1|\phi_2 = 1, \mathit{SIBD} = (i,j))\mathcal{P}(\phi_2 = 1)\frac{1}{4}}{\mathcal{P}(\phi_1 = 1|\phi_2 = 1)\mathcal{P}(\phi_2 = 1)} \\ &= \frac{1}{4}\frac{K_{ij}^\star}{K_S},\end{aligned}$$

for $i,j = 0,1$. Here 'S' denotes siblings and $K_{ij}^\star$ is the prevalence for a individual sharing i paternal and j maternal alleles IBD with an affected proband. This type of prevalence parameters, introduced by Cordell et al. [10] in the nonsex-specific two-locus case, corresponds to imaginary relationships for $i \neq j$, while $K_{00} = K$ and $K_{11} = K_M$, where 'M' denotes monozygotic twins. Note that for nonsex-specific IBD sharing $K_1 = K_O$, where 'O' denotes parent-offspring relationship.

Using equation (4.4) we get

$$\pi_{ij} = \frac{1}{4KK_S}(K^2 + \mathrm{cov}_{ij}), \tag{4.6}$$

for $i, j = 0, 1$, where cov_{ij} is the genetic covariance between a pair of relatives with the indicated IBD sharing.

Naturally, cov_{00} is equal to zero and the other covariances may be expressed in terms of variance components, using the arguments from the derivation of equation (4.5). First consider π_{01}. The only effect that is IBD in two individuals with this IBD sharing, is the additive effect of the maternal allele, consequently $\mathrm{cov}_{01} = \sigma^2_{A^m}$. Similarly, $\mathrm{cov}_{10} = \sigma^2_{A^p}$ and $\mathrm{cov}_{11} = \sigma^2_{A^p} + \sigma^2_{A^m} + \sigma^2_D$. From these expressions and equation (4.6) follow that $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{00})'$ may be expressed

$$\begin{aligned}
\boldsymbol{\pi} &= \frac{1}{4KK_S}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} K^2 \\ \sigma^2_{A^m} \\ \sigma^2_{A^p} \\ \sigma^2_D \end{bmatrix} \\
&= \frac{1}{4KK_S}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 2 & 2 & 4 \end{bmatrix}\begin{bmatrix} K^2 \\ \frac{1}{2}\sigma^2_{A^m} \\ \frac{1}{2}\sigma^2_{A^p} \\ \frac{1}{4}\sigma^2_D \end{bmatrix},
\end{aligned} \tag{4.7}$$

where

$$K_S = K + \frac{1}{K}(\frac{1}{2}\sigma^2_{A^m} + \frac{1}{2}\sigma^2_{A^p} + \frac{1}{4}\sigma^2_D).$$

Note that the 4×4 matrix in (4.7), multiplied by $\frac{1}{4}$, is the matrix $\boldsymbol{A_{s1}}$, defining the general sex-specific single locus representation (3.13). From this follows that the parameters of the general representation, q_{ij}, $i, j = 0, 1$, under RM and HWE, have the following interpretation in terms of prevalences and genetic variance components

$$(q_{00}, q_{10}, q_{01}, q_{11}) = \frac{1}{KK_S}(K^2, \frac{1}{2}\sigma^2_{A^m}, \frac{1}{2}\sigma^2_{A^p}, \frac{1}{4}\sigma^2_D).$$

Using equation (4.7) and the fact that $z_1 = \pi_{01} + \pi_{10}$ and $\sigma_A^2 = \sigma_{A^m}^2 + \sigma_{A^p}^2$, the representation of nonsex-specific single locus IBD probabilities in terms of variance components is obtained by

$$\boldsymbol{z} = \boldsymbol{A_1 v},$$

where $\boldsymbol{v} = \frac{1}{KK_S}\left(K^2, \frac{\sigma_A^2}{2}, \frac{\sigma_D^2}{4}\right)'$ and $\boldsymbol{A_1}$ is the matrix (3.11) defining the single locus parameterisation in the general case. Thus, $\boldsymbol{v}$ provides an interpretation of the general parameters (q_0, q_1, q_2).

Using this parameterisation and the fact that variance components are positive, it is easy to see that the triangle constraints (1.6) and (1.7) are true. This is how Holmans [41] originally derived the triangle.

Note that for a trait with a negligible genetic component $K_S \approx K$. Then $q_0 \approx 1$ and $q_1, q_2 \approx 0$, corresponding to IBD sharing close to $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

Consider a strict recessive model with $f_{nn} = f_{nd} = 0$ and $f_{dd} = f$, where the disease allele has frequency p. For this model we get simple expressions for the required quantities, $\sigma_A^2 = 2p^3(1-p)f^2$,
$\sigma_D^2 = p^2(1-p)^2 f^2$, $K = p^2 f$ and $K_S = \frac{f}{4}(p+1)^2$. This leads to (cf. equation (3.64))

$$(q_0, q_1, q_2) = \left(\frac{4p^2}{(p+1)^2}, \frac{4p(1-p)}{(p+1)^2}, \frac{(1-p)^2}{(p+1)^2}\right). \tag{4.8}$$

When the disease allele frequency, p, approaches 0, q_2 approaches 1, corresponding to allele sharing $(0, 0, 1)$. On the other hand, if the trait allele is very common, then q_0 is close to 1, corresponding to the null IBD proportions.

As a second example consider an additive disease model

$$f_{nd} = \frac{f_{nn} + f_{dd}}{2}. \tag{4.9}$$

For this model the dominance variance is equal to 0 and q_2 is accordingly equal to zero. It follows that $z_1 = \frac{1}{2}(q_0 + q_1) = \frac{1}{2}$ for all f_{11}, f_{22} and p.

4.2.3 The Two-Locus Case

Let $K^\star_{i_1 j_1 i_2 j_2}$, $i_1, j_1, i_2, j_2 = 0, 1$, be the prevalence for an individual, which shares alleles IBD with an affected individual as specified by the indices.

Then we have

$$\begin{aligned}\pi_{i_1j_1i_2j_2} &= \frac{\mathcal{P}(\text{ASP}|\mathit{SIBD}=(i_1,j_1,i_2,j_2))\mathcal{P}(\mathit{SIBD}=(i_1,j_1,i_2,j_2))}{\mathcal{P}(\text{ASP})} \\ &= \frac{\mathcal{P}(\phi_1=1|\phi_2=1,\mathit{SIBD}=(i_1,j_1,i_2,j_2))\mathcal{P}(\phi_2=1)\frac{1}{16}}{\mathcal{P}(\phi_1=1|\phi_2=1)\mathcal{P}(\phi_2=1)} \\ &= \frac{1}{16}\frac{K^\star_{i_1j_1i_2j_2}}{K_S}.\end{aligned}$$

Now, using equation (4.4), we want to express $K^\star_{i_1j_1i_2j_2}$ in terms of the fifteen sex- and locus-specific variance components, corresponding to the fifteen effects in the linear model (4.2). Consider first cov_{0001}. The only effect that has an IBD counterpart in individuals with this IBD sharing, is the additive effect of the maternal allele at locus 2. Thus $\text{cov}_{0001}=\sigma^2_{A_2^m}$. Next consider cov_{0011}. In this case both alleles at locus 2 are IBD, and accordingly both the additive effects and the dominance effect at locus 2 are IBD with probability one and contribute to the covariance, while the remaining variance components vanish. Hence

$$\text{cov}_{0011}=\sigma^2_{A_2^p}+\sigma^2_{A_2^m}+\sigma^2_{D_2}.$$

In this way we also get, for instance,

$$cov_{0101}=\sigma^2_{A_1^m}+\sigma^2_{A_2^m}+\sigma^2_{A_1^mA_2^m}$$

and

$$\text{cov}_{1101}=\sigma^2_{A_1^m}+\sigma^2_{A_1^p}+\sigma^2_{D_2}+\sigma^2_{A_2^m}+\sigma^2_{A_1^pA_2^m}+\sigma^2_{A_1^mA_2^m}+\sigma^2_{D_1A_2^m}.$$

Here $\sigma^2_{A_1^mA_2^m}$ is the variance components for the additive$\times$additive effect of the maternally derived alleles at locus 1 and locus 2 respectively, $\sigma^2_{D_1A_2^m}$ is the dominance$\times$additive component for the interaction of genotypes at locus 1 and the maternal allele at locus 2 etc.

We refrain from displaying all sixteen covariances and the individual IBD probabilities and turn directly to the matrix representation. By expressing all covariances, $cov_{i_1j_1i_2j_2}$, in variance components as indicated above and then applying equation (4.4), the sex-specific IBD probabilities may be expressed as

$$\boldsymbol{\pi}=\boldsymbol{A_{2s}v}, \tag{4.10}$$

where $\boldsymbol{A_{2s}}$ is the matrix from the general representation defined in equation (3.20) and

$$\begin{aligned}\boldsymbol{v} = \;& \tfrac{1}{K_SK}(K^2,\sigma^2_{A_2^m},\sigma^2_{A_2^p},\sigma^2_{D_2},\sigma^2_{A_1^m},\sigma^2_{A_1^mA_2^m},\sigma^2_{A_1^mA_2^p},\sigma^2_{A_1^mD_2},\\ &\sigma^2_{A_1^p},\sigma^2_{A_1^pA_2^m},\sigma^2_{A_1^pA_2^p},\sigma^2_{A_1^pD_2},\sigma^2_{D_1},\sigma^2_{D_1A_2^m},\sigma^2_{D_1A_2^p},\sigma^2_{D_1D_2})'.\end{aligned}$$

Like in the single locus case, $\boldsymbol{v}$ now gives an interpretation of the general parameters under RM, HWE and LE.

The recurrence risk for a sibling of an affected individual, K_S, is expressed using equation (4.4). In the expansion of the covariance for siblings in terms of the fifteen components, the coefficient for the additive components is equal to $\frac{1}{2}$, since an allele of specified parental origin is IBD to the corresponding allele in the sibling with probability $\frac{1}{2}$. Similarly, since two alleles are IBD with probability $\frac{1}{4}$, the coefficient is $\frac{1}{4}$ for the two dominance variance components and the four additive$\times$additive components, $\frac{1}{8}$ for the four dominance$\times$additive components and $\frac{1}{16}$ for the dominance$\times$dominance component.

Proposition 3 follows easily from representation (4.10). For example, inequalities (3.23) and (3.29) result in

$$\pi_{1000} - \pi_{0000} = \frac{1}{16\lambda_S K^2}\sigma^2_{A_1^p} \geq 0$$

and

$$\pi_{1010} - \pi_{0010} - \pi_{1000} + \pi_{0000} = \frac{1}{16\lambda_S K^2}\sigma^2_{A_1^p A_2^p} \geq 0.$$

Similarly the left sides of inequalities (3.23) - (3.26) are all proportional to one sex- and locus-specific additive variance component. Similarly, inequalities (3.27) - (3.32) are proportional to one dominance or additive$\times$additive component and (3.33) - (3.36) each corresponds to one dominance$\times$additive component. Finally the left side of inequality (3.37) is proportional to the dominance$\times$dominance component.

The representation of nonsex-specific two-locus IBD probabilities, derived by Cordell et al. [10] follows by addition of the sex-specific probabilities expressed by representation (4.10) and conversion to nonsex-specific IBD probabilities. The sex-specific variance components are converted to nonsex-specific, for instance $\sigma^2_{A_1} = \sigma^2_{A_1^m} + \sigma^2_{A_1^p}$ and $\sigma^2_{A_1 D_2} = \sigma^2_{A_1^m D_2} + \sigma^2_{A_1^p D_2}$. This results in

$$\boldsymbol{z} = \boldsymbol{A_2 v}, \tag{4.11}$$

where

$$\boldsymbol{v} = \frac{1}{KK_S}\left(K^2, \frac{\sigma^2_{A_2}}{2}, \frac{\sigma^2_{D_2}}{4}, \frac{\sigma^2_{A_1}}{2}, \frac{\sigma^2_{A_1 A_2}}{4}, \frac{\sigma^2_{A_1 D_2}}{8}, \frac{\sigma^2_{D_1}}{4}, \frac{\sigma^2_{D_1 A_2}}{8}, \frac{\sigma^2_{D_1 D_2}}{16}\right)'$$

and $\boldsymbol{A_2}$ is the matrix in equation (3.19) defining the parametrisation in the general case. The nonsex-specific inequalities (3.38) - (3.46) follow from representation (4.11).

4.3 A Direct Proof

Inequalities (3.22) - (3.37) are of the type

$$\sum_x c(x)\mathcal{P}(x|\text{ASP}) \geq 0,$$

where $x = (x^1, x^2)$ are two-locus inheritance vectors. From the calculations in the proof of lemma 1 (from equation (3.50) and forward), it follows under assumption 1 that this is equivalent to

$$\sum_x c(x) \sum_{pg} \mathcal{P}(\text{ASP}|x, pg)\mathcal{P}(pg) \geq 0.$$

As motivated in the proof of proposition 2, we only have to select one representative combination of inheritance vectors for each IBD state. Hence, for example, in order to prove inequality (3.25),

$$\pi_{0010} - \pi_{0000} \geq 0,$$

it suffices to show that

$$\sum_{pg} \{\mathcal{P}(\text{ASP}|(1,3,2,4),(1,3,1,4),pg) -$$

$$\mathcal{P}(\text{ASP}|(1,3,2,4),(1,3,2,4),pg)\}\mathcal{P}(pg) \geq 0.$$

Under assumption 1 and 2, RM, HWE and LE this is equivalent to proving that

$$\begin{aligned}
&\sum_{i_1,j_1,k_1,l_1} \sum_{i_2,j_2,k_2,l_2} \{f_{i_1k_1i_2k_2} f_{j_1l_1i_2l_2} \\
&\qquad\qquad - f_{i_1k_1i_2k_2} f_{j_1l_1j_2l_2}\} \, p_{i_1}p_{j_1}p_{k_1}p_{l_1}p_{i_2}p_{j_2}p_{k_2}p_{l_2} \\
&= \sum_{i_2} p_{i_2} \sum_{i_1,k_1,k_2} f_{i_1k_1i_2k_2} p_{i_1}p_{k_1}p_{k_2} \sum_{j_1,l_1,l_2} f_{j_1l_1i_2l_2} p_{i_1}p_{l_1}p_{l_2} \\
&- \sum_{i_1,k_1,i_2,k_2} f_{i_1k_1i_2k_2} p_{i_1}p_{k_1}p_{i_2}p_{k_2} \sum_{j_1,l_1,j_2,l_2} f_{j_1l_1j_2l_2} p_{j_1}p_{l_1}p_{j_2}p_{l_2} \qquad (4.12)
\end{aligned}$$

for a representative mating type (cf. table 3.2). Now, as an aid to see the structure of these sums, introduce four i.i.d. discrete random variables $X_1, \ldots, X_4$. Let $\mathcal{P}(X_i = j) = p_j$, for $i = 1, \ldots, 4$, $j = 1, \ldots, m_l$. Suppose f is a function of $X_1, \ldots, X_4$. Then (4.12) may be written

$$E[(E[f|X_3])^2] - (E[f])^2 \geq 0. \qquad (4.13)$$

This inequality follows from Jensen's inequality, since $E[f] = E[E[f|X_3]]$ and x^2 is a convex function.

The remaining inequalities may also be expressed in terms of conditional expectations, given subsets of $X_1, \ldots, X_4$. For instance, inequality (3.29),

$$\pi_{1010} - \pi_{0010} - \pi_{1000} + \pi_{0000} \geq 0$$

is

$$E[(E[f|X_1, X_3])^2] - E[(E[f|X_1])^2] - E[(E[f|X_3])^2] + (E[f])^2 \geq 0.$$

However, proofs of the inequalities involving more than two terms are not as straight forward as inequality (4.13). To further slim notation[1] and identify the structure of the inequalities, define for any subset $u \subseteq \{1, 2, 3, 4\}$

$$E^u = E[\ \cdot\ |X_i : i \in u].$$

Then we have

$$E^u E^v = E^v E^u = E^{u \cap v} \text{ for } u, v \subseteq \{1, 2, 3, 4\}, \tag{4.14}$$

which, for instance, may be seen by manipulation of the conditional expectations, expressed in terms of sums of the type in equation (4.12). Further, let $f^u = E^u f$. Then the proof of proposition 3 will follow from the following property of conditional expectations.

$$E[f^u f^v] = E(f^{u \cap v})^2, \tag{4.15}$$

with empty intersection meaning (unconditional) expectation. To prove (4.15) we make repeated use of equality (4.14) and basic properties of conditional expectations.

$$\begin{aligned} E[f^u f^v] &= E[E^u f^u f^v] = E[f^u E^u f^v] \\ &= E[f^u f^{u \cap v}] = E[E^{u \cap v} f^u f^{u \cap v}] \\ &= E[f^{u \cap v} E^{u \cap v} f^u] = E[(f^{u \cap v})^2] \end{aligned}$$

and thus (4.15) holds. Now we are equipped to prove proposition 3.

Proof of proposition 3

Inequalities (3.23) - (3.26) are of the type

$$E(f^{\{1\}})^2 - (Ef)^2 \geq 0.$$

[1]The notation and details of the proof in the reminder of this section was kindly provided by Terry Speed.

As mentioned above this inequality follows from Jensen's inequality, since $Ef = Ef^{\{1\}}$ and x^2 is convex. Inequalities (3.27) - (3.32) are of the type

$$E(f^{\{1,2\}})^2 - E(f^{\{1\}})^2 - E(f^{\{2\}})^2 + (Ef)^2 \geq 0. \tag{4.16}$$

The conditional Jensen's Inequality says that

$$E^{\{2\}}\left(f^{\{1,2\}} - f^{\{1\}}\right)^2 \geq \left(E^{\{2\}}(f^{\{1,2\}} - f^{\{1\}})\right)^2. \tag{4.17}$$

The expectation of the left side of (4.17) is

$$\begin{aligned} &E[(f^{\{1,2\}})^2 - 2f^{\{1,2\}}f^{\{1\}} + (f^{\{1\}})^2] \\ =\ &E(f^{\{1,2\}})^2 - E(f^{\{1\}})^2 \end{aligned}$$

since $E[f^{\{1,2\}}f^{\{1\}}] = E(f^{\{1\}})^2$ according to (4.15). The expectation of the right side of inequality (4.17) is

$$E(f^{\{2\}} - Ef)^2 = E(f^{\{2\}})^2 - (Ef)^2,$$

since $E[f^{\{2\}}Ef] = EfEf^{\{2\}} = (Ef)^2$. Hence inequality (4.16) follows as the expectation of inequality (4.17). Similarly, inequalities (3.33) - (3.36) are of the type

$$\begin{aligned} E(f^{\{1,2,3\}})^2 - E(f^{\{1,2\}})^2 - E(f^{\{1,3\}})^2 - E(f^{\{2,3\}})^2 & \\ +E(f^{\{1\}})^2 + E(f^{\{2\}})^2 + E(f^{\{3\}})^2 - (Ef)^2 &\geq 0. \end{aligned} \tag{4.18}$$

Again, it follows from Jensen's inequality, that

$$\begin{aligned} &E^{\{1,2\}}\left(f^{\{1,2,3\}} - f^{\{1,3\}} - f^{\{2,3\}} + f^{\{3\}}\right)^2 \\ \geq\ &\left(E^{\{1,2\}}(f^{\{1,2,3\}} - f^{\{1,3\}} - f^{\{2,3\}} + f^{\{3\}})\right)^2. \end{aligned} \tag{4.19}$$

If we take the expectation of inequality (4.19) and simplify, using (4.14) and (4.15), then inequality (4.18 follows).

Finally,

$$\begin{aligned} &E^{\{1,2,3\}}\Big(f^{\{1,2,3,4\}} - f^{\{1,2,4\}} - f^{\{1,3,4\}} - f^{\{2,3,4\}} \\ &+ f^{\{1,4\}} + f^{\{2,4\}} + f^{\{3,4\}} - f^{\{4\}}\Big)^2 \\ \geq\ &\Big(E^{\{1,2,3\}}(f^{\{1,2,3,4\}} - f^{\{1,2,4\}} - f^{\{1,3,4\}} - f^{\{2,3,4\}} \\ &+ f^{\{1,4\}} + f^{\{2,4\}} + f^{\{3,4\}} - f^{\{4\}})\Big)^2, \end{aligned} \tag{4.20}$$

by Jensen's. The last inequality, (3.37), follows by simplification of the expectation of (4.20).

Chapter 5

IBD Probabilities at Markers

The IBD proportion at two candidate loci[1], e.g. markers linked to the disease loci, are in this chapter expressed in terms of recombination fractions and the IBD proportions at the disease loci. Given the representations derived in chapter 3, the marker IBD probabilities may be expressed directly in terms of recombination fractions and the parameters of the representation. Using the obtained expressions, it is shown that IBD proportions at the candidate loci are contained in the two-locus analogue of the possible triangle.

5.1 The Single-Locus Case

Under the model with one disease locus, consider a marker, $\mathcal{M}(\theta)$, where θ is the recombination fraction between the marker and the disease locus. Furthermore, let $\mathit{SIBD}_\theta = (i,j)$, $i,j = 0,1$, be the sex-specific IBD status at the marker and, as before, let *SIBD* refer to the disease locus. The sex-specific ASP IBD probabilities at the marker are, for $i,j = 0,1$,

$$\begin{aligned}\pi_{ij}(\theta) &= \mathcal{P}(\mathit{SIBD}_\theta = (i,j)|\mathrm{ASP}) \\ &= \sum_{k,l=0,1} \mathcal{P}(\mathit{SIBD}_\theta = (i,j)|\mathit{SIBD} = (k,l), \mathrm{ASP}) \cdot \\ &\qquad\qquad \mathcal{P}(\mathit{SIBD} = (k,l)|\mathrm{ASP}). \qquad (5.1)\end{aligned}$$

[1]The term 'candidate locus' is often used to mean a chromosome region, whose protein product suggests that it may be a disease gene. However, here it means any locus under investigation, for instance, during a genome screen.

The IBD status at the marker locus is conditionally independent of disease phenotypes given the IBD status at the disease locus. It follows that $\pi_{ij}(\theta)$ may be written in terms of θ and the IBD proportions at the disease locus (see e.g. Haseman and Elston [37]), as follows.

Let $\psi = \theta^2 + (1-\theta)^2$. This is the probability that recombination has occurred between the marker and the disease locus, in both or none of the chromosomes transmitted to the siblings from a given parent. In other words, ψ is the probability that the IBD status of alleles transmitted from this parent is the same at the disease locus and at the marker. Hence, the conditional sex-specific marker IBD proportions, given the IBD state at the disease locus, are the elements of the matrix (cf. Dudoit and Speed [20])

$$\boldsymbol{\Psi_s} = \begin{pmatrix} \psi^2 & \psi\overline{\psi} & \psi\overline{\psi} & \overline{\psi}^2 \\ \psi\overline{\psi} & \psi^2 & \overline{\psi}^2 & \psi\overline{\psi} \\ \psi\overline{\psi} & \overline{\psi}^2 & \psi^2 & \psi\overline{\psi} \\ \overline{\psi}^2 & \psi\overline{\psi} & \psi\overline{\psi} & \psi^2 \end{pmatrix}, \tag{5.2}$$

where rows (columns) correspond to the marker (disease locus). Now equation (5.1) can be written

$$\boldsymbol{\pi}(\theta) = \boldsymbol{\Psi_s}\boldsymbol{\pi}, \tag{5.3}$$

where $\boldsymbol{\pi}(\theta) = (\pi_{00}(\theta), \pi_{01}(\theta), \pi_{10}(\theta), \pi_{11}(\theta))'$ and $\boldsymbol{\pi}$ is the IBD proportions at the disease locus . Equation (5.3) is true under any single locus model. For models where the IBD probabilities are expressed using parameterisation (3.13), we can write

$$\boldsymbol{\pi}(\theta) = \boldsymbol{\Psi_s}\boldsymbol{A_{s1}}\boldsymbol{q}, \tag{5.4}$$

which simplifies to

$$\boldsymbol{\pi}(\theta) = \boldsymbol{A_{s1}}(\psi)\boldsymbol{q}$$

where

$$\boldsymbol{A_{s1}}(\psi) = \begin{pmatrix} \frac{1}{4} & \frac{1}{2}\overline{\psi} & \frac{1}{2}\overline{\psi} & \overline{\psi}^2 \\ \frac{1}{4} & \frac{1}{2}\psi & \frac{1}{2}\overline{\psi} & \psi\overline{\psi} \\ \frac{1}{4} & \frac{1}{2}\overline{\psi} & \frac{1}{2}\psi & \psi\overline{\psi} \\ \frac{1}{4} & \frac{1}{2}\psi & \frac{1}{2}\psi & \psi^2 \end{pmatrix} \tag{5.5}$$

and $\boldsymbol{q} = (q_{00}, q_{01}, q_{10}, q_{11})'$. Matrix (5.5) may be obtained directly from the interpretation of the parameters, discussed in section 3.4.1. Recall that q_{00}

corresponds to no information about IBD at the disease locus. Accordingly, all IBD states at the marker also have equal probability, $\frac{1}{4}$, which give us column one of matrix (5.5). Column two is obtained from the fact that the parameter q_{01} corresponds to information implying that the maternal allele at the disease locus is IBD, but giving no information about IBD status of the paternal allele. Hence, the two paternal IBD states have equal probability, $\frac{1}{2}$. The maternal IBD states, 0 and 1, have probability $\overline{\psi}$ and ψ respectively. With similar arguments we get the columns corresponding to q_{10} (column three) and q_{11} (column four).

In the nonsex-specific case the marker IBD proportions $\boldsymbol{z} = (z_0(\theta), z_1(\theta), z_2(\theta))'$, can be written (see e.g. Haseman and Elston [37])

$$\boldsymbol{z}(\theta) = \boldsymbol{\Psi}\boldsymbol{z},$$

where $\boldsymbol{\Psi}$ is the matrix

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi^2 & \psi\overline{\psi} & \overline{\psi}^2 \\ 2\psi\overline{\psi} & \psi^2 + \overline{\psi}^2 & 2\psi\overline{\psi} \\ \overline{\psi}^2 & \psi\overline{\psi} & \psi^2 \end{pmatrix}. \tag{5.6}$$

For models leading to IBD probabilities within the possible triangle, we have

$$\boldsymbol{z}(\theta) = \boldsymbol{\Psi}\boldsymbol{A_1}\boldsymbol{q},$$

where $\boldsymbol{A_1}$ is the the matrix (3.11). This may be simplified to

$$\boldsymbol{z}(\theta) = \boldsymbol{A_1}(\psi)\boldsymbol{q} \tag{5.7}$$

where

$$\boldsymbol{A_1}(\psi) = \begin{pmatrix} \frac{1}{4} & \frac{1}{2}\overline{\psi} & \overline{\psi}^2 \\ \frac{1}{2} & \frac{1}{2} & 2\psi\overline{\psi} \\ \frac{1}{4} & \frac{1}{2}\psi & \psi^2 \end{pmatrix}. \tag{5.8}$$

Similar to the sex-specific case above, matrix (5.8) may be obtained directly using the interpretation of the parameters, q_0, q_1 and q_2 (see section 3.4.1).

It has been proved, see Holmans [41] or Dudoit and Speed [21], that the IBD probabilities at markers are contained in the possible triangle. This may be seen from representation (5.7) as follows. The triangle is defined by linear inequalities, (1.6) and (1.7), hence it is convex. Write representation (5.7) in the form

$$\begin{bmatrix} z_0(\theta) \\ z_1(\theta) \\ z_2(\theta) \end{bmatrix} = q_0 \begin{bmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{bmatrix} + q_1 \begin{bmatrix} \frac{1}{2}\overline{\psi} \\ \frac{1}{2} \\ \frac{1}{4}\psi \end{bmatrix} + q_2 \begin{bmatrix} \overline{\psi}^2 \\ 2\psi\overline{\psi} \\ \psi^2 \end{bmatrix}. \tag{5.9}$$

It may be shown that the column vectors on the right hand side satisfy inequalities (1.6) and (1.7) for $0 \leq \theta \leq \frac{1}{2}$, and hence they are contained in the triangle. Now, since $\sum q_i = 1$, $0 \leq q_i \leq 1$ and the triangle is convex, it follows that $z_i(\theta)$, $i = 0, 1, 2$ are contained in the triangle.

5.2 The Two-Locus Case

Consider two unlinked markers $\mathcal{M}(\theta_1)$ and $\mathcal{M}(\theta_2)$, where $\mathcal{M}(\theta_l)$ is linked to disease locus l, $l = 1, 2$, with recombination fraction θ_l and unlinked to the other disease locus. Let, $\psi_l = \theta_l^2 + (1 - \theta_l)^2$ and define for, locus l, $\boldsymbol{\Psi_s^l}$, $\boldsymbol{A_{s1}}(\psi_l)$, $\boldsymbol{\Psi^l}$ and $\boldsymbol{A_1}(\psi_l)$ as in equations (5.2), (5.5), (5.6) and (5.8) respectively, with ψ replaced by ψ_l. Also let $\boldsymbol{\pi}(\theta_1, \theta_2)$ and $\boldsymbol{z}(\theta_1, \theta_2)$ be the vectors of the sex-specific and non-sex-specific marker IBD probabilities.

The conditional probabilities of IBD states at the marker given the IBD configuration at the disease loci are products of four probabilities. Each probability, $\overline{\psi}_l$ or ψ_l for locus l, $l = 1, 2$, corresponds to one sex-specific IBD state being changed or unchanged between the disease locus and the marker. It may be seen that the matrix of conditional probabilities can be expressed as the Kronecker product of the single-locus matrices of conditional marker IBD probabilities, that is $\boldsymbol{\Psi_s^1} \otimes \boldsymbol{\Psi_s^2}$.

It follows that, for models were the sex-specific IBD probabilities are expressed with representation (3.20), we have

$$\begin{aligned} \boldsymbol{\pi}(\theta_1, \theta_2) &= \boldsymbol{\Psi_s^1} \otimes \boldsymbol{\Psi_s^2} \boldsymbol{\pi} \\ &= \boldsymbol{\Psi_s^1} \otimes \boldsymbol{\Psi_s^2} \boldsymbol{A_{s1}}^{\otimes 2} \boldsymbol{q} \\ &= \boldsymbol{A_{s1}}(\psi_1) \otimes \boldsymbol{A_{s1}}(\psi_2) \boldsymbol{q}, \end{aligned}$$

where $\boldsymbol{\pi}(\theta_1, \theta_2)$, $\boldsymbol{\pi}$ and $\boldsymbol{q}$ are 16×1 vectors. Similarly, in the nonsex-specific case we get

$$\begin{aligned} \boldsymbol{z}(\theta_1, \theta_2) &= \boldsymbol{\Psi^1} \otimes \boldsymbol{\Psi^2} \boldsymbol{z} \\ &= \boldsymbol{\Psi^1} \otimes \boldsymbol{\Psi^2} \boldsymbol{A_1}^{\otimes 2} \boldsymbol{q} \\ &= \boldsymbol{A_1}(\psi_1) \otimes \boldsymbol{A_1}(\psi_2) \boldsymbol{q}. \end{aligned}$$

The two-locus analogue of the possible triangle, defined by representation (3.19), is convex since it is defined by intersections of linear inequalities, (3.38) - (3.46). Furthermore, the sixteen column vectors in the matrix

$\boldsymbol{A_1}(\psi_1) \otimes \boldsymbol{A_1}(\psi_2)$, all comply with the constraints. We have not pursued trying to prove this analytically, but it has been confirmed by plotting. Each of the nine linear combinations of the column elements, corresponding to inequalities (3.38) - (3.46), was plotted for $0 \leq \theta_1, \theta_2 \leq \frac{1}{2}$ and was seen to be nonnegative. Hence, since the conditional IBD probabilities at the markers are linear combinations of the column vectors, where the coefficients, q_{ij}, $i, j = 0, 1, 2$, are nonnegative and sum to one, it follows from the convexity that the marker IBD proportions are contained in the triangle analogue.

Chapter 6

Maximum Likelihood Estimation of IBD Probabilities

In this chapter we describe maximum likelihood estimation of the nonsex-specific two-locus IBD probabilities subject to the constraints, using a double EM algorithm as described by Zhao et al. [79] for a different application. Sampling assumptions are not stated in the text, but the likelihood calculations require the usual assumptions of independence between samples. Here this implies independence of genotypes and phenotypes between sib-pairs and hence excludes related families. Other types of assumptions are discussed when they are needed.

First, we describe the likelihood of complete and incomplete IBD data and the inheritance distribution required in the latter case. Section 6.2 gives a general introduction to the EM algorithm, followed by a description of the double EM-algorithm.

Let $IBD \in \{(i,j), i,j = 0,1,2\}$ be the nonsex-specific IBD status at two unlinked candidate loci. It may be instructive to note that, to be consistent with the notation of section 5.1, $IBD = IBD_{\theta_1,\theta_2}$, where θ_1 and θ_2 are (unknown) recombination fractions between the candidate loci and the two disease loci.

We want to estimate the joint ASP IBD proportions, $z_{ij} = z_{ij}(\theta_1, \theta_2)$ for $i,j = 0,1,2$.

6.1 Likelihood of Marker Data

Let $M_1, \ldots, M_n$ be the marker data of n affected sib-pairs, that is M_k is the sibling genotypes at all typed markers and parental marker genotypes where available, for sib-pair k, $k = 1, \ldots, n$. Our first objective is to calculate the likelihood of the marker data,

$$\begin{aligned} \mathcal{P}(M_1, \ldots, M_n|\text{ASP}) &= \prod_{k=1}^{n} \mathcal{P}(M_k|\text{ASP}) \\ &= \prod_{k=1}^{n} \sum_{i,j=0}^{2} \mathcal{P}(M_k|\mathit{IBD} = (i,j), \text{ASP}) z_{ij}. \quad (6.1) \end{aligned}$$

Consider the situation when parental genotypes are observed and sufficiently informative to determine the IBD states of the sib-pairs with certainty. Let $ibd_k = (i,j)$, $i,j = 0,1,2$, be IBD status of sib-pair k, $k = 1, \ldots, n$. Now, since $\mathcal{P}(M_k|\mathit{IBD} = (i,j), \text{ASP})$ is zero for all IBD states except the observed, it follows that (6.1) is proportional to

$$\prod_{k=1}^{n} z_{ibd_k} = \prod_{i=0,j=0}^{2} z_{ij}^{N_{ij}} =: L_C(\boldsymbol{z}), \quad (6.2)$$

where N_{ij} is the number of sib-pairs sharing i alleles IBD at locus 1 and j alleles at locus 2. L_C denotes the 'complete data'-likelihood, a notion that will be useful in the context of the EM algorithm.

In practice parental genotypes are often not available and, even if they are, data may not be sufficiently informative for reduction to IBD counts. In this situation the likelihood is calculated from conditional IBD probabilities for each sib-pair, given marker data,

$$\mathcal{P}(\mathit{IBD} = (i,j)|M_k), \quad (6.3)$$

for $k = 1, \ldots n$, to be discussed in section 6.1.1. However, this requires a strong assumption. In the likelihood (6.1), $\mathcal{P}(M_k|\mathit{IBD} = (i,j), \text{ASP})$ is now greater than zero for more than one IBD state and furthermore this probability depends on the genetic model. In order to be able to use the conditional probabilities (6.3), we will need the following assumption:

$$\mathcal{P}(M_k|\mathit{IBD} = (i,j), ASP) = \mathcal{P}(M_k|\mathit{IBD} = (i,j)). \quad (6.4)$$

This is true when the candidate loci are at the two disease loci, that is when $z_{ij} = z_{ij}(0,0)$, provided that there is LE between the markers and the disease loci. In addition to this, two perhaps less questionable assumptions are needed. The first is that of no segregation distortion at markers and disease locus (cf. assumption 1), the second is conditional independence of phenotype, given disease genotypes. That equation (6.4) is true under these assumptions may be proved[1] by summation of the left side of equality (6.4) over founder genotypes at the disease loci and inheritance vectors at markers and disease locus. Then expansion, by the law of total probability, followed by simplification using the above assumptions, result in the right side of equality (6.4).

The important fact here is that equality (6.4) is *not* true in many practical applications, because then we do not know the positions of the disease loci. It is clear that the sib-pair affection status is not independent of marker genotypes when there is LD between marker and disease loci. Furthermore, even under LE, the phenotype provides information about IBD, and hence about genotypes, at markers between the disease locus and the candidate locus.

Nevertheless, property (6.4) is required in practical applications. We will use it in the description of the double EM algorithm, having in mind that we are not working with the proper likelihood (6.1).

6.1.1 The Inheritance Distribution

Computer programs (e.g. MapMaker/Sibs, cf. Kruglyak and Lander [48]) are routinely used to extract inheritance information from incomplete marker data using a Hidden Markov Model (HMM).

Under the assumption of no cross-over interference, the inheritance vectors constitute a Markov Chain along the chromosome, where recombination forces transition between the inheritance vectors. The probability distribution of the inheritance vectors at any locus in the genome, i.e. also in between markers, can be computed conditional on data for all markers linked to the locus. Probabilities of inheritance vectors are then summed to give

$$\omega_{ki}^{l} = \mathcal{P}(IBD_l = i | M_k), \tag{6.5}$$

for $i = 0, 1, 2$, sib-pair k, $k = 1, \ldots, n$, at locus l .

If parental data are not available, then all possible parental genotypes have to be taken into account and population mating type frequencies are required. For a single marker it may be feasible to include general marker mating type frequencies as unknown parameters of the likelihood, in simultaneous estimation of these and the IBD probabilities. Then the calculation of ω_{ki}^{l} would be conditional not only on M_k, but on the marker data of all sib-pairs.

[1] Proved in unpublished notes, kindly provided by Sandrine Dudoit.

However, we want to use all linked markers for estimation of IBD. This may not be feasible without assuming HWE, RM and LE between markers, considering the vast number of possible multi-locus genotypes under the general model.

In any case, in the implementation of our estimation algorithm, we make use of a HMM algorithm implemented in for instance MapMaker/Sibs. Population allele frequencies are given as input for use in computation of the inheritance distribution of sib-pairs where no parental data are available. This implies assumption of LE, RM and HWE. The frequencies may be estimated from independent samples, or perhaps most commonly, from $M_1, \ldots, M_n$, either by selection of random individuals from each sib-pair, or using all data, taking the relationships into account (see Broman [6]).

In summary, the computation of ω_{ki}^l requires (i) no crossover interference, (ii) LE between markers and (iii) the assumption that single-marker genotype frequencies may be constructed assuming HWE and RM.

In what follows it is assumed that the marker data of a sib-pair, M_k, are reduced to ω_{ki}^1 and ω_{kj}^2, $i, j = 0, 1, 2$ for two candidate loci. Now, since ω_{ki}^l is calculated without conditioning on phenotype and DNA at unlinked loci segregate independently, we can simply multiply the probabilities for the two loci to get

$$\omega_{kij} = \mathcal{P}(\mathit{IBD} = (i,j)|M_k) = \omega_{ki}^1 \omega_{kj}^2 \tag{6.6}$$

for $i, j = 0, 1, 2$ and $k = 1, \ldots, n$ and ω_{ki}^l defined by (6.5). However, if parental genotypes are not observed assumption (ii) is needed again for equation (6.6) to be true.

6.2 A Double EM Algorithm for Estimation of the Constrained IBD Probabilities

The constrained IBD probabilities may be estimated using a double EM algorithm, where the M-step is an internal EM algorithm.

In the internal algorithm we consider a discrete unobservable variable, X_k, for sib-pair k, $k = 1, \ldots, n$, taking nine values $\{G_{ij} : i, j = 0, 1, 2\}$ with probability q_{ij} (cf. the nonsex-specific case in section 3.4.2). Hence X_k is related to IBD status through representation (3.19).

6.2.1 Introduction to the EM Algorithm

The EM (Expectation-Maximisation) algorithm is a general-purpose iterative procedure for computation of maximum-likelihood estimates from incomplete data. The concept 'incomplete data' includes observed data with missing observations as well as the situation where the observed data are a function of some 'complete data' which cannot be observed. The algorithm was generally formulated by Dempster et al. 1977 [16]. Before that it had been used by geneticists in many special cases under the name 'gene counting' (see e.g. Ott [61]).

In the general formulation (McLachlan and Krishnan [55]) let $\boldsymbol{y}$ be a vector of incomplete (observed) data, with probability density function $f(\boldsymbol{y}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Further let $\boldsymbol{x}$ be the complete (unobservable) data vector with p.d.f. $f_C(\boldsymbol{x}, \boldsymbol{\theta})$. Then the complete data log-likelihood is

$$\log L_c(\boldsymbol{\theta}) = \log f_C(\boldsymbol{x}, \boldsymbol{\theta}).$$

In the *E-step* the conditional expectation of the complete data log likelihood, given the observed data,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{\theta}^{(t)}}[\log L_c(\boldsymbol{\theta})|\boldsymbol{y}],$$

is computed. Here $\boldsymbol{\theta}^{(t)}$ is the parameter value obtained in the previous iteration, or some initial value when $t = 0$.

In the *M-step* a new estimate $\boldsymbol{\theta}^{(t+1)}$ is obtained as the θ that maximises $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

The E-step and M-step are alternately repeated until the difference between the incomplete data log likelihood of two successive iterations,

$$\log L(\boldsymbol{\theta^{(t+1)}}) - \log L(\boldsymbol{\theta^{(t)}}),$$

is sufficiently small. It can be shown (see e.g. McLachlan and Krishnan [55]) that, as a consequence of Jensen's inequality, the incomplete data likelihood is never decreased after an iteration.

6.2.2 Description of the Algorithm

Suppose that we have observed IBD counts, that is the number of sib-pairs that shares i alleles at locus 1 and j alleles at locus 2, for $i, j = 0, 1, 2$. Then an ML estimation of $\boldsymbol{z}$ can be performed, where $\boldsymbol{q}$ is thought of as probabilities in an indirectly observed multinomial distribution. However, since we do not observe the counts we will use expected counts $\hat{N}_{ij}$, $i, j = 0, 1, 2$. First we give an outline of the algorithm, then we elaborate the calculations in each step. The index t is used for the outer iterations, while t' is used for iterations of the internal algorithm.

Outline of the Algorithm

***E-step*(z)**

Given an estimate of the ASP IBD probabilities from the previous step, $\boldsymbol{z}^{(t)}$, and given ω_{kij}, expected IBD counts $\hat{N}_{ij}^{(t)}$ given marker data are computed.

***M-step*(z)**

***E-step*$(\boldsymbol{q})$**

The conditional expectation of the complete data likelihood given $\hat{N}_{ij}^{(t)}$, i.e. $Q(\boldsymbol{q}, \boldsymbol{q}^{(t')})$ is computed.

***M-step*$(\boldsymbol{q})$**

The new estimate, $\boldsymbol{q}^{(t'+1)}$, is obtained by maximisation of $Q(\boldsymbol{q}, \boldsymbol{q}^{(t')})$ with respect to $\boldsymbol{q}$.

E-step(q) and *M-step*(q) are repeated until $L(\boldsymbol{q}^{(t'+1)}) - L(\boldsymbol{q}^{(t')})$ is sufficiently small. Then let

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t'+1)}.$$

The new estimate of the IBD probability vector is

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{A_2}\boldsymbol{q}^{(t+1)}.$$

E-step(z) and *M-step*(z) are repeated until $L(\boldsymbol{z}^{(t+1)}) - L(\boldsymbol{z}^{(t)})$ is sufficiently small.

Details of the Algorithm

The expected sharing proportions for each affected sib-pair in *E-step*(z) are

$$\mathcal{P}_{\boldsymbol{z}^{(t)}}(IBD = (i,j)|M_k, \mathrm{ASP}) = \frac{\mathcal{P}_{\boldsymbol{z}^{(t)}}(M_k|IBD = (i,j), \mathrm{ASP})z_{ij}^{(t)}}{\sum_{u,v=0}^{2} \mathcal{P}_{\boldsymbol{z}^{(t)}}(M_k|IBD = (i,j), \mathrm{ASP})z_{uv}^{(t)}},$$

which under assumption (6.4) simplifies to

$$\frac{\omega_{kij} z_{ij}^{(t)}/\alpha_{ij}}{\sum_{u,v=0}^{2} \omega_{kuv} z_{uv}^{(t)}/\alpha_{uv}} =: e_{ijk}^{(t)}.$$

Here α_{ij} is the unconditional IBD probability, defined in section 2.2 and ω_{kij} factorise into conditional IBD probabilities for each locus (se equation (6.6)). The expected counts, $\hat{N}_{ij}^{(t)}$, are obtained by summing $e_{ijk}^{(t)}$ over k.

Following the general formulation in section 6.2.1, the E-step involves computation of

$$\begin{aligned} Q(\boldsymbol{z}, \boldsymbol{z}^{(t)}) &= E_{\boldsymbol{z}^{(t)}}[\log L_C(\boldsymbol{z})|M_1, \dots, M_n, \mathrm{ASP}] \\ &= E_{\boldsymbol{z}^{(t)}}[\sum_{k=1}^{n} \log(z_{ibd_k})|M_1, \dots, M_n, \mathrm{ASP}] \\ &= \sum_{k=1}^{n} \sum_{i,j=1}^{2} \log(z_{ij}) \mathcal{P}_{\boldsymbol{z}^{(t)}}(\mathit{IBD} = (i,j)|M_k, \mathrm{ASP}) \\ &= \sum_{i,j=1}^{2} \log(z_{ij}) \sum_{k=1}^{n} \mathcal{P}_{\boldsymbol{z}^{(t)}}(\mathit{IBD} = (i,j)|M_k, \mathrm{ASP}). \end{aligned}$$

Thus $Q(\boldsymbol{z}, \boldsymbol{z}^{(t)})$ is maximised, with respect to $\boldsymbol{z}$, by

$$\frac{1}{n} \sum_{k=1}^{n} \mathcal{P}_{\boldsymbol{z}^{(t)}}(\mathit{IBD} = (i,j)|M_k, \mathrm{ASP}),$$

which under equation (6.4) is equal to $\frac{\hat{N}_{ij}^{(t)}}{n}$. Hence, we see that the expected counts from *E-step*(z), are the counts based on the result of the M-step of an EM-algorithm for unconstrained maximisation of the IBD probabilities.

Next consider *E-step*(q). The unobservable variable, X_k, may take the value G_{rs}, $r, s = 0, 1, 2$ with probability q_{rs}. From the interpretation in section 3.4, it is clear that the corresponding events certainly are unobservable, unless the location of the disease loci, the penetrance model and the parental genotypes are known. Under our representation we have

$$\mathcal{P}(\mathit{IBD} = (i,j)|X_k = G_{rs}) = a_{ir} a_{js},$$

for all k, where a_{ij} are the elements of matrix (3.11), cf. the nonsex-specific case in 3.4.2. Let

$$\begin{aligned} p(i,j,r,s;\boldsymbol{q}) &= \mathcal{P}_{\boldsymbol{q}}(X_k = G_{rs}|ibd_k = (i,j)) \\ &= \frac{a_{ir} a_{js} q_{rs}}{\sum_{u,v=0}^{2} a_{iu} a_{jv} q_{uv}}. \end{aligned}$$

In *E-step*(q) the complete data likelihood of $\boldsymbol{x} = x_1, \dots, x_n$ is

$$L_C(\boldsymbol{q}) = \prod_{k=1}^{n} q_{x_k},$$

where $q_{x_k} = q_{ij}$ when $x_k = G_{ij}$. Now, given the expected IBD counts, $\hat{N}_{ij}^{(t)}$,

$$\begin{aligned}
Q(\boldsymbol{q}, \boldsymbol{q}^{(t')}) &= E_{\boldsymbol{q}^{(t')}}[\sum_{k=1}^{n} \log(q_{x_k}) | ibd_1, \ldots, ibd_n] \\
&= \sum_{k=1}^{n} E_{\boldsymbol{q}^{(t')}}[\log(q_{x_k}) | ibd_k] \\
&= \sum_{k=1}^{n} \sum_{r,s=0}^{2} \log(q_{rs}) \mathcal{P}_{\boldsymbol{q}^{(t')}}(X_k = G_{rs} | ibd_k), \\
&= \sum_{k=1}^{n} \sum_{r,s=0}^{2} \log(q_{rs}) \sum_{i,j=0}^{2} \mathcal{P}_{\boldsymbol{q}^{(t')}}(X_k = G_{rs} | ibd_k = (i,j)) I_{\{ibd_k = (i,j)\}} \\
&= \sum_{r,s=0}^{2} \log(q_{rs}) \sum_{i,j=0}^{2} p(i,j,r,s;\boldsymbol{q}^{(t')}) \hat{N}_{ij}^{(t)}.
\end{aligned}$$

In *M-step*(q), $Q(\boldsymbol{q}, \boldsymbol{q}^{(t')})$ is maximised with respect to $\boldsymbol{q}$, which results in

$$q_{rs}^{(t'+1)} = \frac{\sum_{i,j=0}^{2} p(i,j,r,s;\boldsymbol{q}^{(t')}) \hat{N}_{ij}^{(t)}}{n}.$$

When considering the likelihood of $\boldsymbol{z}$, the IBD counts are the complete data. On the other hand, in the internal algorithm the expected IBD counts are the incomplete data. Hence, the incomplete data log-likelihood to be maximised in the iterations of the internal algorithm is the log of $L_C(\boldsymbol{z}^{(t')})$ in equation (6.2), where $\boldsymbol{z}^{(t')} = \boldsymbol{A_2 q}^{(t')}$.

The incomplete data log-likelihood of the outer loop is the log of the likelihood (6.1), which under equation (6.4) is proportional to

$$\prod_{k=1}^{n} \sum_{i,j=0}^{2} \frac{\omega_{kij} z_{ij}^{(t+1)}}{\alpha_{ij}}.$$

Chapter 7

Discussion and Open Problems

We have proved a set of constraints on two-locus ASP IBD probabilities that is a natural extension of the single-locus possible triangle. This has been done under two genetic models, one with general mating type frequencies and monotonicity assumptions on the penetrances and one with no penetrance assumptions but instead the population genetic assumptions HWE, RM and LE. The nonsex-specific IBD proportions for models satisfying the constraints are conveniently expressed in terms of nine parameters, which are unconstrained, except for being positive and summing to one. We have showed that the constraints are satisfied also at markers or other candidate loci linked to the disease loci. We described how the IBD probabilities may be estimated from marker data by the method of maximum likelihood, using a double EM algorithm.

It is our expectation that the parameterisation may be used to define new statistical methods, or to improve existing methods, for detection of genetic effects and interactions in IBD data. However, this is outside the scope of this thesis. In this chapter we indicate the use of the parameterisation, refer to a few available two-locus methods and list some remaining questions and open problems.

7.1 IBD Interaction Versus Gene Interaction

The aim of two-locus linkage analysis is to detect and model gene interaction between loci, or to increase the power to detect disease loci by conditioning on the effect of a previously identified disease locus (for examples in the non-parametric case, see Cordell et al. [10, 11], Cox et al [13]). By gene interaction we mean any deviation (on some scale) from additivity of the separate effects of the two genes, see for instance Frankel and Schork [31] for a discussion.

Consider the nonsex-specific two-locus ASP IBD probabilities in the form of the 3×3 table (7.1), with marginals z_{i+} and z_{+j}. Naturally, we de-

z_{00}	z_{01}	z_{02}	z_{0+}
z_{10}	z_{11}	z_{12}	z_{1+}
z_{20}	z_{21}	z_{22}	z_{2+}
z_{+0}	z_{+1}	z_{+2}	

Table 7.1: Joint nonsex-specific IBD probabilities and marginals.

fine interaction between the IBD probabilities at the two loci as dependence of rows and columns, that is $z_{ij} \neq z_{i+}z_{+j}$ for some i, j. A first important point is that physical gene interaction between two loci does not necessarily result in interaction in the joint IBD proportions. A good example of this is the multiplicative penetrance model, described in section 2.3.1. This is often used to represent epistasis between loci, but the joint IBD probabilities are, assuming LE between the loci, the product of the marginals (cf. Risch [65]).

For models, where the IBD probabilities may be expressed using representation (3.19), we cannot have interaction in the IBD probabilities without marginal effects at both loci. To see this, let $z_{0+} = \frac{1}{4}$, which means no IBD distortion at locus 1. Under representation (3.19), $z_{0+} = \frac{1}{4}(q_{00} + q_{01} + q_{02})$, and it follows that $q_{ij} = 0$ for $i \neq 0$. It may be seen that $z_{ij} = z_{i+}z_{+j}$, where $(z_{0+}, z_{1+}, z_{2+}) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ and $(z_{+0}, z_{+1}, z_{+2}) = (\frac{1}{4}q_{00}, \frac{1}{2}(q_{00} + q_{01}), \frac{1}{4}(q_{00} + q_{01} + q_{02}))$. However, it remains to be seen if small deviations in the marginals may be compatible with large, detectable interaction.

For models, where the IBD probabilities do not satisfy the inequalities, it is not true that no marginal effects imply no interaction. Consider for example the 3×3 table of IBD probabilities with $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ in the diagonal.

It can be shown that no interaction in q_{ij}, that is $q_{ij} = q_{i+}q_{+j}$ for $i, j = 0, 1, 2$, is equivalent to no interaction in the IBD probabilities. Hence, it may be profitable to study the interaction issues in terms of the parameters q_{ij}, $i, j = 0, 1, 2$, since they are unconstrained. However it is unclear how deviations from independence of different magnitude are translated between the two sets of parameters.

The modifier model in table 2.4 (a) is interesting as another example of the fact that gene interactions may be difficult to detect in IBD data. In an example, Ott [62] has calculated the IBD proportions for this model, with disease allele frequencies of 0.04 and 0.05 at locus 1 and locus 2 respectively.

This leads to IBD proportions close to the null proportions at locus 2, and a relatively small interaction, while there is a large deviation from the null ($z_2 = 0.89$) at locus 1. Thus, locus 2 may not be detected although it plays a role in the genetic mechanism.

Given the marginals of the IBD table 7.1, there are four degrees of freedom left for interaction, which may be summarised in some measure of association. One obvious measure is the chi-square statistic,

$$\sum_{i=0}^{2}\sum_{j=0}^{2}\frac{(z_{ij} - z_{i+}z_{+j})^2}{z_{i+}z_{+j}}, \tag{7.1}$$

where a term is replaced by zero if $z_{i+}z_{+j} = 0$. Another measure is the correlation between the number of alleles IBD at the two loci, which has been examined by for instance Nicolae [58] and Li and Reich [53] in their enumeration of two-locus models (see section 2.3.1).

We have not studied the interaction issues in depth, and would welcome some, perhaps more general, work on how interaction in the penetrances force interaction into the table of joint IBD probabilities.

7.2 Tests

One natural statistic for test of interaction is the chi-square statistic (7.1), with the IBD probabilities estimated subject to the constraints. This has to our knowledge not been suggested in the literature.

Another candidate is a likelihood ratio test, with the likelihood of the estimated IBD probabilities versus the likelihood of the null proportions α_{ij} (cf. section 2.2 and the one-locus MLS statistic (1.5)). To this end, Cordell et al. [10, 11] and Farrall [25, 27] have implemented ML estimation of the IBD probabilities, subject to the constraints in proposition 1. They also suggest likelihood ratio tests of model-specific hypothesis, for instance test of the hypothesis that the first locus is linked, but the second is not, that is the hypothesis $\alpha_{ij} = z_{i+}\alpha_j$ for $i, j = 0, 1, 2$. Furthermore, Cordell et al. [10, 11] suggest maximisation of the likelihood with respect to the variance components. Tests of certain disease models can be accomplished by for instance setting some components to zero. Maximisation in the general model, over all variance components, should be equivalent to the estimation procedure described in chapter 6.

Holmans [41] showed that the single-locus likelihood ratio statistic (1.5), with the IBD probabilities restricted to the possible triangle, has a chi-square distribution with a mixture of one and two degrees of freedom. It might be possible to derive the distribution of the corresponding two-locus statistic, which we conjecture to have a mixture of one to eight degrees of freedom, but this has as far as we know not been accomplished. Cordell et al. [10, 11] assess statistical significance by simulation.

Given the many degrees of freedom of the two-locus likelihood ratio test, models with fewer parameters may supply increased power. In the single-locus case, one-parameter models have been suggested (see e.g. Whittemore and Tu [76], Kong and Cox [46], Nicolae [58]). Tests based on these models have been shown to have the same or better power than the likelihood ratio test. As one example, Whittemore and Tu [76], considered a one-parameter model corresponding to $(q_0, q_1, q_2) = (1-\lambda, \lambda a, \lambda(1-a))$. For $0 \leq \lambda \leq 1$ and fixed $0 \leq a \leq 1$ the IBD probabilities follow a line between the null point and the point $(0, \frac{a}{2}, 1-\frac{a}{2})$, as shown in fig. 7.1. They showed that the

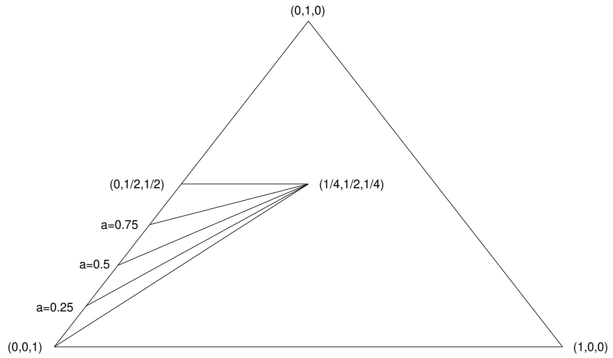

Figure 7.1: IBD probabilities under the one-parameter model of Whittemore and Tu [76] for three values of a.

score statistic for this model has the form $\hat{z}_2 + w\hat{z}_1$, where w is a function of a. A score statistic, for test of the null hypothesis $\lambda = 0$, is based on the derivative of the log-likelihood, $l'(\lambda)$. It is locally most powerful for λ close to zero (see Cox and Hinkley [12]) and

$$\frac{l'(0)}{\sqrt{E_{H_0}[l''(0)]}}$$

is approximately standard normally distributed under the null. If for instance the true penetrance model is additive (see equation (4.9)), then the IBD probabilities lie on the line given by $a = 1$, i.e. $z_1 = \frac{1}{2}$, and it may be proved that the weight $w = \frac{1}{2}$ is optimal. The weight w was chosen by Whittemore and Tu [76] to minimise the maximum penalty in terms of increased sample size compared to the true model, which resulted in $w = 0.355$.

It would be desirable to find a suitable lower-dimension parameterisation of the two-locus space, resulting in a robust test. Alternatively a score test in one or more of the parameters q_{ij}, $i, j = 0, 1, 2$ may be derived. This will be a weighted combination of some $\hat{z}_{ij}$, $i, j = 0, 1, 2$, where the weights involve some q_{ij}, $i, j = 0, 1, 2$. This imposes two questions. Which parameter or parameters are suitable to base the score test on and how should the weights be chosen? The answers to both questions rely on the issues discussed in section 7.1.

Cox et al. [13] suggested the Pearson correlation between family IBD scores, i.e. some functions of IBD states, at two loci as a means to detect interaction. Their ideas have been extended by Nicolae [58], who considered a model for the distribution of IBD scores including three parameters; one for each marginal and one for association. However, we have yet to explore this approach in detail and perhaps see how it fits into our parameterisation.

Whatever statistic we choose we will have to face issues of multiple testing. Tests of interactions between all unlinked markers during a genome scan would include tens of thousands of tests. In fact, the inheritance distribution may be calculated at every point in the genome and we are not restricted to the marker loci. Strategies such as performing two-locus analysis at locations with a single-locus statistic above some threshold may be appropriate. Cordell et al. [11] suggest a stepwise strategy where the likelihood at a locus is calculated conditional on evidence of linkage at a number of other loci. Development of such testing strategies is an important issue.

7.3 Miscellaneous Topics

- The most crucial assumption used in the derivation of the constraints on the IBD probabilites is assumption 2, stating conditional independence of sibling phenotypes, given the disease genotypes. This assumption rules out covariance between the sibling phenotypes caused by environmental factors or effects of other loci not accounted for in the two-locus model. It would be useful to see an investigation of how the constraints fail under some models with environmental covariates and gene-environment interaction, and also to understand how assumption 2 perhaps may be relaxed to allow for some forms of environmetal influence.

  Greenwood and Bull [34] argued that the possible triangle constraints (1.6) and (1.7) will hold under gene-environment interactions in which the size of a genetic effect may be modified by environmental factors, but the direction of the effect remains the same. Furthermore, they demonstrated that the IBD probabilities can fall outside the triangle when the sample contains sib-pairs with different exposures to an environmetal factor that changes the direction of the effect of a disease gene. Dizier et al. [17] suggested a test based on the departure from the triangle constraints, as a means to detect heterogeneity such as presence or absence of an associated trait or an environmetal factor, or differences in the characteristic of the trait itself (e.g. severe versus mild form).

- Given the definition of the two-locus representation (3.19) as the (Kronecker) square of the single-locus matrix (3.11), the representation should readily extend to more than two disease loci by Kronecker power of higher order (cf. Cordell et al. [11]), but we have not worked out any details. In fact most results in this thesis should have analogues for models with any number of loci.

- To our knowledge, constraints on ASP IBD probabilities for two *linked* disease loci have not been thoroughly addressed. The term $\mathcal{P}(x|pg)$ in equation (2.7) will in this case involve the recombination fraction and does not cancel out of the equation. Furthermore, linked loci imply additional complexity since the phase of the two-locus genotypes, discussed in section 2.1 now cannot be disregarded. The approach of Cordell et al. [10], with maximisation of the likelihood expressed in terms of genetic variance components has been extended to linked loci by Farrall [25] and Olson [60]. Note that in equation (4.6) only K_S depends on the recombination fraction. Farrall [25] showed how K_S is expressed in terms of variance components and recombination fractions.

- In a sib-pair study families with multiple affected siblings may be sampled. In the single locus case constraints on IBD probabilities for affected sib trios have been derived (se e.g. Whittemore and Tu [76]). In the two-locus case it may be impractical to work with joint IBD probabilities of sib-trios. In a sibship of more than two sibs, multiple sib-pairs can be formed, but these are not statistically independent. Selection of independent pairs only, may be a waste of information (see e.g. Hodge [40]). Thus, the treatment of large sibships is an important issue.

- We have proved that the two-locus analogue of the possible triangle is covered under a jointly recessive model with general mating type frequencies. However, we have not shown that the parameter space is covered under HWE, RM and LE, as was done in the single locus case by Dudoit and Speed [21]. While the possible triangle is conveniently displayed by the barycentric plot or a regular plot in two of the coordinates, we lack in the two-locus case a tool to visualise the coverage of different models.

- We have implemented the double EM algorithm as described in chapter 6, but it has been only tentatively evaluated. It needs to be tuned in terms of selection of start values and convergence criteria. The resulting estimates and the convergence properties should be evaluated.

Bibliography

[1] M. Bahlo, K.W. Broman, and T.P. Speed. Identification and adjustment for genotyping errors in data on sibpairs when parental genotypes are unavailable. *Am. J. Hum. Genet.*, 65:A241, 1999.

[2] M.S. Bartolomei and S.M. Tilghman. Genomic imprinting in mammals. *Annu. Rev. Genet.*, 31:493–525, 1997.

[3] W. C. Blackwelder and R. C. Elston. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet. Epidemiol.*, 2:85–97, 1985.

[4] M. Boehnke and N.J. Cox. Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.*, 61:423–429, 1997.

[5] K. W. Broman. Cleaning genotype data. In L. Goldin, C.I. Amos, G. A. Chase, A. M. Goldstein, G.P Jarvik, M.M. Martines, B.K. Suarez, D.E. Weeks, E.M. Wisjman, and J.W. MacCluer, editors, *Genetic Analysis Workshop 11: Analysis of Genetic and Environmental Factors in Common Disease*. Genetic Epidemiology, 1999.

[6] K.W. Broman. Estimation of allele frequencies with data on sibships. *Genet. Epidemiol.*, in press.

[7] K.W. Broman, J.C. Murray, V.C. Sheffield, and R.L. White. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, 63:861–869, 1998.

[8] K.W. Broman and J.L. Weber. Characterization of human crossover interference. *Am. J. Hum. Genet.*, 66:1911–1926, 2000.

[9] F.S. Collins. Shattuck lecture - medical and societal consequences of the human genome project. *N. Engl. J. Med.*, 341:28–37, 1999.

[10] H.J. Cordell, J.A. Todd, S.T. Bennet, Y. Kawaguchi, and M. Farrall. Two-locus maximum lod score analysis of a multifactorial trait: Joint

consideration of IDDM2 and IDDM4 with IDDM1 in type I diabetes. *Am. J. Hum. Genet.*, 57:920–934, 1995.

[11] H.J. Cordell, G. C. Wedig, K. B. Jacobs, and R. C. Elston. Multilocus linkage tests based on affected relative pairs. *Am. J. Hum. Genet.*, 66:1273–1286, 2000.

[12] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman & Hall, 1974.

[13] N. J. Cox, M. Frigge, D. L. Nicolae, P. Concannon, C. L Hanis, G. I. Bell, and A. Kong. Loci on chromosome 2 (niddm1) and 15 interact to increase susceptibiliy in mexican americans. *Nature Genet.*, 21:213–215, 1999.

[14] S. Davis and D. E. Weeks. Comparison of nonparametric statistics for detection of linkage in nuclear families: Single marker evaluation. *Am. J. Hum. Genet.*, 61:1431–1444, 1997.

[15] N.E. Day and M.J. Simons. Disease susceptibility genes - their identification by multiple case family studies. *Tissue Antigens*, 8:109–119, 1976.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal. Stat. Soc. B*, 39:1–38, 1977.

[17] M.H. Dizier, H. Quesneville, B. Prum, H. Selinger-Leneman, and F. Clerget-Darpoux. The triangle test statistic (tts): a test of genetic homogneity using departure from the triangle constraints in ibd distribution among affected sib-pairs. *Abb. Hum. Genet.*, 64:433–442, 2000.

[18] J. A. Douglas, M. Boehnke, and K. Lange. A multipoint method for detection of genotyping errors and mutations in sibling pair data. *Am J. Hum. Genet.*, 66:1287–1297, 2000.

[19] S. Dudoit. *Linkage analysis of complex traits using identity by descent data.* PhD thesis, University of California, Berkeley, 1999.

[20] S. Dudoit and T.P. Speed. A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data on sib-pairs. *Biostatistics*, 1:1–26, 1999.

[21] S. Dudoit and T.P. Speed. Triangle constraints for sib-pair identity by descent probabilities under a general multilocus model for disease

susceptibility. In Halloran M.E and Geisser S., editors, *Statistics In Genetics, the IMA Volumes in mathematics and its applications. vol 112*. Springer, 1999.

[22] M. Durner, V. J. Vieland, and D. A. Greenberg. Further evidence for the increased power of lod scores compared with nonparametric methods. *Am. J. Hum. Genet.*, 64:281–289, 1999.

[23] K. R. Ewen, M. Bahlo, S. A. Treloar, D. F. Levinson, B. Mowry, J. W. Barlow, and S. J. Foote. Identification and analysis of error types in high-throughput genotyping. *Am. J. Hum. Genet.*, 67:727–736, 2000.

[24] J.J. Faraway. Improved sib-pair linkage test for disease susceptibility loci. *Genet. Epid.*, 10:225–233, 1993.

[25] M. Farrall. Affected sibpair linkage tests for multiple linked susceptibility genes. *Genet. Epidemiol.*, 14:103–115, 1997.

[26] M. Farrall. Lod wars: The affected-sib-pair paradigm strikes back! *Am. J. Hum. Genet.*, 60:735–738, 1997.

[27] M. Farrall and H. Cordell. TWOLOC, a software package for analysing two-locus susceptibility gene models in affected sib-pair data. *Available by anonymous FTP from ftp.well.ox.ac.uk/pub/genetics/twoloc/*, 1998.

[28] E. Feingold, P. O. Brown, and D. O. Siegmund. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.*, 53:234–251, 1993.

[29] E. Feingold and D. O. Siegmund. Strategies for mapping heterogeneous recessive traits by allele-sharing methods. *Am. J. Hum. Genet.*, 60:965–978, 1997.

[30] R.A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Trans.R.Soc.Edinburgh*, 52:399–433, 1918.

[31] W. N. Frankel and N. J. Schork. Who's afraid of epistasis? *Nature Genet.*, 14:371–373, 1996.

[32] H. H. H. Goring and J. Ott. Relationship estimation in affected sib pair analysis of late onset diseases. *Eur. J. Hum. Genet.*, 5:69–77, 1997.

[33] D. A. Greenberg, S. E. Hodge, V. J. Vieland, and M. A. Spence. Affecteds-only linkage methods are not a pancea. *Am. J. Hum. Genet.*, 58:892–895, 1996.

[34] C. M. T. Greenwood and S. B. Bull. Analysis of affected sib pairs, with covariates - with and without constraints. *Am. J. Hum. Genet.*, 64:871–885, 1999.

[35] G. R. Grimmet and D. R. Stirzaker. *Probability and Random Processes.* Clarendon Press, Oxford, 1992.

[36] D.L. Hartl and T. Maruyama. Phenogram enumeration: The number of regular genotype-phenotype correspondences in genetic systems. *J.Theoret. Biol.*, 20:129–163, 1968.

[37] J.K. Haseman and R.C Elson. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, 2:3–19, 1972.

[38] E. R. Hauser and M. Boehnke. Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics*, 54:1238–1246, 1998.

[39] S.E. Hodge. Some epistatic two-locus models of disease. *Am.J.Hum.Genet.*, 33:381–395, 1981.

[40] S.E. Hodge. The information contained in multiple sibling pairs. *Genet. Epidemiol.*, 1:109–122, 1984.

[41] P. Holmans. Asymptotic properties of affected-sib-pair linkage analysis. *Am.J.Hum.Genet.*, 52:362–374, 1993.

[42] S. Horvath and M. P. Baur. Future directions of research in statistical genetics. *Statistics in Medicine*, 19:3337–3343, 2000.

[43] J.W. James. Frequency in relatives for an all-or-none trait. *Ann.Hum.Genet.*, 35:47–49, 1971.

[44] O. Kempthorne. The correlation between relatives in a random mating population. *Proc. Royal. Soc. Lond. B*, 143:103–113, 1954.

[45] O. Kempthorne. *An introduction to genetic statistics.* John Wiley & sons, 1957.

[46] A. Kong and N. J. Cox. Allele-sharing models: Lod scores and accurate linkage tests. *Am. J. Hum. Genet.*, 61:1179–1188, 1997.

[47] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am.J.Hum.Genet.*, 58:1347–1363, 1999.

[48] L. Kruglyak and E.S. Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am.J.Hum.Genet.*, 57:439–454, 1995.

[49] M. Lalande. Parental imprinting and human disease. *Annu. Rev. Genet.*, 30:173–195, 1996.

[50] E.S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA*, 84:2363–2367, 1987.

[51] E.S. Lander and L. Kruglyak. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11:241–247, 1995.

[52] A. Lernmark and J. Ott. Sometimes it's hot, sometimes it's not. *Nature Genetics*, 19:213–214, 1998.

[53] W. Li and J. Reich. A complete enumeration and classification of two-locus disease models. *Hum. Hered.*, 50:334–349, 2000.

[54] M. Lynch and B. Walsh. *Genetics and analysis of quantitative traits.* Sinauer Associates, Inc., 1998.

[55] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions.* John Wiley & sons, 1997.

[56] S.O. Moldin. Detection and replication of linkage to a complex human disease. *Genet. Epidemiol.*, 14:1023–1028, 1997.

[57] U. Motro and G. Thomson. The affected sib method. 1: Statistical features of the affected sib-pair method. *Genetics*, 110:525–538, 1985.

[58] D. L. Nicolae. *Allele sharing models in gene mapping: A likelihood approach.* PhD thesis, University of Chicago, 1999.

[59] J. R. O'Connell and D. E. Weeks. Pedcheck: A program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, 63:259–266, 1998.

[60] J. M. Olson. Likelihood-based models for genetic linkage analysis using affected sib-pairs. *Hum. Hered*, 47:110–120, 1997.

[61] J. Ott. Counting methods (em algorithm) in human pedigree analysis: Linkage and segregation analysis. *Ann. Hum. Genet., Lond.*, 40:443–454, 1977.

[62] J. Ott. *Analysis of human genetic linkage, third edition.* John Hopkins University Press, 1999.

[63] L. S. Penrose. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann. Eugen.*, 6:133–138, 1935.

[64] N. Risch. Linkage strategies for genetically complex traits. I multilocus models. *Am.J.Hum.Genet.*, 46:222–228, 1990.

[65] N. Risch. Linkage strategies for genetically complex traits. II the power of affected relative pairs. *Am.J.Hum.Genet.*, 46:229–241, 1990.

[66] S.B. Roberts, C. J. MacLean, M. C. Neale, L. J. Eaves, and K. S. Kendler. Replication of linkage studies of complex traits: An examination of variation in location estimates. *Am. J. Hum. Genet.*, 65:876–884, 1999.

[67] P. Sham. *Statistics in human genetics.* Arnold, 1998.

[68] T.P. Speed. What is a genetic map function? In T.P Speed and M.S. Waterman, editors, *Genetic Mapping and DNA Sequencing, the IMA Volumes in mathematics and its applications. vol 81.* Springer, 1996.

[69] R.S. Spielman, L. Baker, and C.M. Zmijewski. Gene dosage and susceptibility to insulin-dependent diabetes. *Ann. Hum. Genet Lond.*, 44:135–150, 1980.

[70] M.W. Strickberger. *Genetics, third edition.* Macmillan, New York, 1985.

[71] B:K: Suarez, T. Reich, and J. Trost. Limits of the general two-allele single locus model with incomplete penetance. *Ann. Hum. Genet. Lond.*, 40:231–244, 1976.

[72] J. Teng and D. Siegmund. Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics*, 54:1247–1265, 1998.

[73] G. Thomson and M.S. Esposito. The genetics of complex diseases. *Trends Cell. Biol.*, 9:M17–M20, 1999.

[74] H.K. Tiwari and R.C. Elston. Derivation of genetic variance components for multilocus models. *Genet. Epidemiol*, 14:1131–1136, 1997.

[75] J. M. Vance and K. B. Othmane. Methods of genotyping. In J.L. Haines and M. A. Pericak-Vance, editors, *Approaches to Gene Mapping in Complex Human Disease.* Wiley-Liss, 1998.

[76] A. S. Whittemore and I-P. Tu. Simple robust linkage tests for affected sibs. *Am. J. Hum. Genet.*, 62:1228–1242, 1997.

[77] B. R. Winkelmann and J. Hager. Genetic variation in coronary heart disease and myocardial infarction: Methodological overview and clinical evidence. *Pharmacogenomics*, 1:73–94, 2000.

[78] J. Xu, D.A. Meyers, and M.A. Pericak-Vance. Lod score analysis. In J.L. Haines and M. A. Pericak-Vance, editors, *Approaches to Gene Mapping in Complex Human Disease*. Wiley-Liss, 1998.

[79] H. Zhao, M.S. McPeek, and T.P. Speed. Statistical analysis of chromatid interference. *Genetics*, 139:1057–1065, 1995.