NSF-CBMS Regional Conference Series
in Probability and Statistics
Volume 6

STATISTICAL INFERENCE FROM GENETIC DATA ON PEDIGREES

Elizabeth A. Thompson
University of Washington

Institute of Mathematical Statistics
Beachwood, Ohio

American Statistical Association
Alexandria, Virginia

Conference Board of the Mathematical Sciences

Regional Conference Series
in Probability and Statistics

Supported by the
National Science Foundation

The production of the *NSF-CBMS Regional Conference Series in Probability and Statistics* is managed by the Institute of Mathematical Statistics: Barry Arnold, IMS Managing Editor, Statistics; Patrick Kelly, IMS Production Editor; Julia Norton, IMS Treasurer; and Elyse Gustafson, IMS Executive Director.

Library of Congress Control Number: 00-134575

International Standard Book Number 0-940600-49-8

Printed in the United States of America

Contents

List of Tables

List of Figures

Preface

This monograph is based primarily on material presented at the CBMS Summer Course on **Inferences from genetic data on pedigrees** given at Michigan Technical University, Hougton, Michigan, in July 1999. This monograph is not a textbook; it contains no exercises, and is insufficiently detailed for that purpose. However, it could be used as a textbook, either in conjunction with the excellent texts of Weir (1996), Lange (1997) and Ott (1999), or by advanced students who will consult the cited literature for details.

The notes used at the Summer Course have been augmented by material from two lecture classes given at the University of Washington. A Special Topics class was given in January-March, 1999, and additional background on Markov chain Monte Carlo and Monte Carlo EM are included from that class. Some details were also first presented at a SEMSTAT workshop in Eindhoven in March 1999 (Thompson, 2000*b*). Although material has been added, the examples in Chapter 10 and on identity by descent under interference (section 11.2) were first presented at a Royal Statistical Society Meeting in London, in March 1999 (Thompson, 2000*a*). Versions of Figures 9.1, 10.1, 10.2, 10.6, and 10.7 first appeared in Thompson (2000*a*). However, the 11-chapter monograph follows closely the ten sessions of the Summer Course presentations, with chapter 2 being the only addition, providing statistical background with genetic examples. The order of Chapters 8 and 9 has been reversed from the Summer Course; a case can be made for either ordering.

A more basic Statistical Genetics class was given in Fall 1999, at University of Washington, and led to extensive revision of Chapters 1-4. It is hoped that the monograph can thus serve two purposes. For example, a more introductory course could cover of Chapters 1-4, with final material taken from sections 6.1, 6.2, 7.1, and 7.2. More advanced students could skip Chapters 1-2, skim Chapters 3-5, and study the later chapters more thoroughly.

I would like to thank Dr.Anant Godbole and Dr.JianPing Dong, for their excellent organization of the CBMS Regional Research Conference at Michigan Technical University. I am also grateful to the many students who attended this course, and to students attending the two University of Washington courses, for their helpful comments and criticisms. In particular, I would like to thank Eric Anderson, Nicky Chapman and Dr.Ellen Wijsman for help with LaTeX, BibTeX, Xfig, and GENEHUNTER, and for many discussions. I am grateful to Amy Anderson for her thorough and critical reading of Chapters 1 to 5, and to Eric Anderson, Dr.Erin Conlon, Dr.Mary Kuhner, Anne-Louise Leutenegger, and Jessica

Maia, who all read and commented on other chapters.

Some of the MCMC work was undertaken in collaboration with Dr.Simon Heath. In particular, the implementation of the algorithm described in section 3.6 and the initial incorporation of the L-sampler of Heath (1997) into our M-sampler software to create the LM-sampler (section 10.6) are both due to Dr.Heath. Figures 1.1, 1.2, 3.4, 3.5, and 10.3, first appeared in Thompson and Heath (1999), and are also due to Dr.Heath. I am grateful to Dr.Heath for our continuing collaboration.

The CBMS Regional Research Conference was funded by NSF grant number 98-13767 to Dr.Jianping Dong and Dr.Anant Godbole of The Mathematical Sciences Department of Michigan Technical University, Houghton, MI.

Table of Notation

Since there are an insufficient number of user-friendly letters and symbols, some must be used for more than one purpose. However, for convenience, we summarize the principal usages here

Notation	Usage
Parameters	
θ	the general (set of) parameters of a model
ρ	a recombination frequency parameter
γ	a (trait) locus location
Γ_M	a marker map; set of arker locations
β	a trait model penetrance parameter
r	number of multinomial outcomes (or phenotypes)
$p_1, \ldots, p_r$	probabilities of multinomial outcomes
k	number of alleles at a locus
$q_1, \ldots, q_k$	population allele frequencies at this locus
q	an allele frequency, often for a recessive allele
ψ	a kinship coefficient
ϕ	chiasmata avoidance function
$\kappa_i, i = 0, 1, 2$	gene-identity probabilities
Indices and labels	
i	an index used primarily for individuals or meioses
j	an index used primarily for alleles or loci
k, k_i	a label for a gene
L	a number of loci ordered on a chromosome
m	a count, often of the number of meioses
v	miscellaneous other counts, of genes for example
n	sample size
F	father, or paternal, often as subscript
M	mother, or maternal, often as subscript also marker, as in marker data $\mathbf{Y}_M$
N	Monte Carlo sample size also (Chapter 5) the random number of chiasmata)
τ	an index of Monte Carlo or MCMC realizations
$\mathcal{T}$	a set of indices of latent variables
$\mathcal{D}$	a set of indices of data observations
Variables	
$A_1, ..., A_k$	the alleles at a locus
$\mathbf{Y}$, value $\mathbf{y}$	the data random variables (usually phenotypes)
$\mathbf{Y}_M$	phenotypes at marker loci, in linkage mapping
$\mathbf{Y}_T$	trait phenotypes; $\mathbf{Y} = (\mathbf{Y}_T, \mathbf{Y}_M)$
$\mathbf{X}$, value $\mathbf{x}$	latent variables
$\mathbf{X}^\dagger$	a proposed value of $\mathbf{X}$ in Monte Carlo sampling
$\mathbf{X}^*$	a sampled or resampled value of $\mathbf{X}$ in Monte Carlo
$\mathbf{G} = \{G_i\}$	the set of genotypes of individuals i
g	a genotype — a possible value of G_i

Notation	Usage
Variables continued	
$\mathbf{S} = \{S_{i,j}\}$	set of meiosis indicators for meioses i and loci j
$S_{\bullet,j}$	the vector of $S_{i,j}$ at given locus j
$S_{i,\bullet}$	the vector of $S_{i,j}$ at given meiosis i
$G_{\bullet,j}, G_{i,\bullet}$	similarly for genotypes, locus j, individual i
$Y_{\bullet,j}, Y_{i,\bullet}$	similarly for phenotypes, locus j, individual i
$Y^{(j)}$	the data $\{Y_{\bullet,1}, Y_{\bullet,2}, \ldots, Y_{\bullet,j}\}$; $\mathbf{Y} = Y^{(L)}$
$\mathbf{J} = \mathbf{J}(\mathbf{S})$	a gene *ibd* pattern, a function of $\mathbf{S}$
$I_1, ..., I_{L-1}$	the intervals between L ordered loci
$\mathbf{R} = (R_j; j = 1, ..., L-1)$	the recombination indicators in intervals I_j
$\mathbf{r}$	a vector of recombination indices; value of $\mathbf{R}$
$\mathbf{C} = (C_j; j = 1, ..., L-1)$	the chiasmata presence/absence indicators in intervals I_j
$\mathbf{c}$	a vector of chiasma indices; value of $\mathbf{C}$
T, value t	a count (often binomial)
t_j	a multinomial count, e.g. of latent genotypes; also (Chapter 5) a set of binary indicators
n_{jl}, n_j	multinomial data counts, of observable phenotypes or genotypes
m_j	multinomial counts, often of alleles
Functions and probabilities	
Pr	probability, when not indexed by a parameter
$\Pr(E; \theta)$	probability of event E under model θ
$P_\theta(\cdot)$	a probability distribution, indexed by θ
$P^*(\cdot)$	a probability distribution, used for the sampling or resampling distribution in Monte Carlo methods
$\mathrm{E}_\theta(\cdot)$	Expectation, under a model indexed by θ
$\Phi(\cdot)$	the standard Normal (Gaussian) cumulative distribution function
$I(\cdot)$	the indicator function of an event
$L(\theta)$ or $L_{\mathbf{y}}(\theta)$	the likelihood for parameter θ given data $\mathbf{y}$
$L(\theta; \mathbf{Y})$	the likelihood function, considered also as a function of data random variables $\mathbf{Y}$
ℓ or $\ell(\theta)$	the log-likelihood function for parameter θ
$K_n(\theta; \theta_0)$	Kullback-Leibler information in a sample size n
$K_{\mathbf{y}}(\theta; \theta_0)$	K-L information in latent $\mathbf{X}$ given data $\mathbf{y}$
$H_{\mathbf{y}}(\theta; \theta_0)$	expected complete-data log-likelihood given $\mathbf{Y} = \mathbf{y}$: $\mathrm{E}_{\theta_0}(\log P_\theta(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$
$R(\cdot)$ and $R^*(\cdot)$	cumulative probabilities of data used in computing probabilities on graphs or pedigrees
$Q(\cdot)$, $Q^*(\cdot)$, $Q^\dagger(\cdot)$	cumulative conditional probabilities of latent variables given data on graphs or pedigrees
$h(\mathbf{X}^\dagger; \mathbf{X})$	Hastings ratio for proposed $\mathbf{X}^\dagger$ when at state $\mathbf{X}$
$q(\mathbf{X}^\dagger; \mathbf{X})$	proposal probability for $\mathbf{X}^\dagger$ when at state $\mathbf{X}$
a	the Metropolis-Hastings acceptance probability

Chapter 1

Genes, Pedigrees and Genetic Models

1.1 DNA, alleles, loci, genotypes, and phenotypes

The *DNA* in the nuclei of cells of an individual consists of about 3×10^9 base pairs (bp). This *DNA* is packaged into *chromosomes* each of which has a linear DNA sequence in a twisted double-helical structure. There are 46 chromosomes in the nucleus of each normal human cell, 22 pairs of *autosomes* and a pair of *sex chromosomes*. Of the two chromosomes of a pair, one derives from the DNA of the mother of the individual and the other derives from the DNA of the father. In this book, we will restrict attention to the autosomes, which contain the majority of the DNA coding for the proteins and affecting the characteristics of individuals. Similar approaches would apply to the sex chromosomes, but the details differ. There is additional DNA in the mitochondria, which are located in the cytoplasm of the cell; mitochondrial DNA is maternally inherited.

Any small segment of the DNA of the chromosome is known as a *locus*. Typically, a locus used to refer to the segment of DNA coding for some functional protein, but it is now used to refer to any position characterized by a specific DNA sequence, or by specific forms of variation in the sequence. These loci exhibiting observable variation in the DNA are *DNA marker loci*, and a *locus* simply indicates a particular position on a particular one of the pairs of chromosomes. The DNA at a locus may come in a variety of forms, or *alleles*. Any individual has two chromosomes of a given pair, and thus has two (possibly identical) alleles at each locus. The unordered pairs of alleles that an individual has is the individual's *genotype* at this locus. If the locus is one relating to a functional gene, the resulting potentially observable characteristic of the individual is the *phenotype*. A locus exhibiting non-negligible variation in a population is known as a *polymorphism*, or the locus is said to be *polymorphic*. Classically, the frequency of the the most frequent genotype should

be less than 99% for a locus to qualify as a polymorphism.

For example, the DNA which codes for the antigens that determine an individual's *ABO* blood type is at a certain position on Chromosome 9. This is a chromosome in mid-range size; chromosomes are numbered in approximately decreasing size order. This position is the *ABO locus.* There are three major alleles at the human *ABO* locus, A, B, and O, although these allelic types can be subdivided. The *ABO* locus is polymorphic in almost every human population. There are thus six genotypes; AA, AO, BB, BO, OO, and AB. However there are only four phenotypes (*ABO* blood types), type-A, type-B, type-O and type-AB. Individuals with genotype AA or AO have type-A blood type; individuals with genotype BB or BO have type-B blood type. For each of the phenotypes O and AB, there is a single corresponding genotype.

A genotype for which the two alleles are the same, such as AA, BB or OO are known as the *homozygous* genotypes. The individual is a *homozygote* or is *homozygous* at this locus. Where the two alleles are different (AO, BO or AB), the individual is a *heterozygote* or is *heterozygous* at this locus. Where a heterozygous genotype exhibits the same phenotype as one of the two homozygotes, the allele carried by this homozygote is said to be *dominant* to the other allele. At the *ABO* blood type locus, for example, individuals of genotype BO have type-B blood. The B allele is dominant to the O allele; the O allele is *recessive* to the B allele. Likewise, the A allele is dominant to the O allele; the O allele is recessive to the A allele. Individuals of genotype AB have type-AB blood, distinct from the phenotypes of both the AA (type-A) and BB (type-B) genotypes. The alleles A and B are said to be *codominant.*

Initially, genetic markers used in genetic analysis were blood type or enzyme markers such as the *ABO* locus. The first DNA markers were restriction fragment polymorphisms or RFLPs (Botstein et al., 1980). These often had several alleles, or at least two alleles with substantial frequency. These were followed by current microsatellite markers, where alleles correspond to different numbers of tandem repeats of a small number (2, 3, or 4) of base pairs. Microsatellite markers are often highly polymorphic, with 10 or more alleles observed with non-negligible frequency in any given population. These have become the markers of choice for genetic mapping, but statistically have several disadvantages all due to the high degree of polymorphism. Mutation rates at some microsatellite markers are high, and typing errors also more frequent. Accurate estimation of population allele frequencies is harder, and inferences can be sensitive to allele frequency assumptions. The newest DNA markers are single-nucleotide polymorphisms or SNPs. These measure variation at a single base of DNA. Although there are many SNPs in the human genome, perhaps as many as 1 per 500 bp, or several million in total, most have only two alleles. In the future, genetic mapping analyses may be based on a much larger number of much less informative markers with consequent additional challenges.

1.2 Mendel's laws and meiosis indicators

Mendel's First Law (1866) states that each individual has two "factors" (or genes) controlling a given characteristic, one being a copy of a corresponding gene in the father of the individual, the other a copy of a gene in the mother of the individual. Further, a copy of a randomly chosen one of the two is copied to each child, independently for different children and independently of genes contributed by the spouse. The probabilistic process of the random choice of genes to be copied is known as Mendelian *segregation.* The biological process forming the chromosomes of the gamete (sperm or egg) cell is known as *meiosis* At a single locus, the *segregation* of genes is fully specified by *meiosis indicators*

$$\begin{aligned} S_i &= 0 \quad \text{if copied gene is parent's maternal gene} \\ &= 1 \quad \text{if copied gene is parent's paternal gene} \end{aligned} \tag{1.1}$$

where $i = 1, ..., m$ indexes the meioses (parent-child links) in the pedigree. Mendel's First Law then simply states that the indicators S_i are independent, and

$$\Pr(S_i = 0) = \Pr(S_i = 1) = \frac{1}{2}.$$

For multiple loci, j, $j = 1, \ldots, L$, we must specify the segregation of genes at each locus:

$$\begin{aligned} S_{i,j} &= 0 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's maternal gene} \\ &= 1 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's paternal gene.} \end{aligned} \tag{1.2}$$

Contrary to Mendel's second law (Mendel, 1866), which in effect stated that $S_{i,j}$ are independent for different loci j, the segregation of alleles at loci on the same chromosome are dependent. The collection of alleles at loci on a chromosome in the maternal [paternal] gamete, is the maternal [paternal] *haplotype* of the offspring individual.

The word "gene" is overused in modern genetics, often referring to the locus (as in "the *ABO* gene"), or to an allele predisposing the individual to a particular disease or trait (as in "the cystic fibrosis gene"). Here we reserve the word "gene" for Mendel's original "factors"; the gene is the entity transmitted from parent to offspring. The meiosis indicators $S_{i,j}$ have also attracted a variety of names and notations. Karlin and Liberman (1979) used them in the theoretical analysis of meiosis patterns at loci on a chromosome (Chapter 5). Their first use in the computation of probabilities of gene descent in pedigrees is due to Donnelly (1983), who called them *switches.* Thompson (1994*c*) retained the notation of Donnelly (1983), but called them *segregation indicators.* Lander and Green (1987) use the phrase *inheritance vectors* while Sobel and Lange (1996) use *descent graphs.* Together with a defined pedigree structure, the meiosis indicators do indeed determine the inheritance or descent patterns of genes in a pedigree (section 3.6). However, in considering the indicators alone we prefer the name *meiosis indicators.*

For later convenience we define the following notation

$$
\begin{aligned}
S_{\bullet,j} &= \{S_{i,j}; i = 1, \ldots, m\}, \quad j = 1, \ldots, L \\
S_{i,\bullet} &= \{S_{i,j}; j = 1, \ldots, L\}, \quad i = 1, \ldots, m
\end{aligned} \tag{1.3}
$$

where m is the number of meioses in the pedigree, and L the number of loci along the chromosome. The m vectors $S_{i,\bullet}$ are *a priori* independent, but the components $S_{i,j}$ are dependent. The pattern of dependence depends on the process of meiosis, which will be considered further in Chapter 5. However, under (untrue) assumptions of absence of genetic interference in meiosis, there is a simple conditional independence structure. Suppose the loci are ordered $1, \ldots, L$ along a chromosome. Then given $S_{i,j}$, $(S_{i,1}, \ldots, S_{i,j-1})$ is independent of $(S_{i,j+1}, \ldots, S_{i,L})$. Or, $(S_{i,j})$ is first-order Markov over j.

1.3 Pedigrees: the conditional independence structure

A *pedigree* is a specification of the genealogical relationships among a set of individuals. A convenient form of this specification is to identify the father and the mother of each individual. Individuals at the top of the pedigree, whose parents are unspecified, are the *founders* of the pedigree; other individuals are *non-founders*. Individuals in the pedigree, and without offspring, are referred to as *final* individuals; unless there are data on such an individual, he contributes no information. Relationships among individuals are defined relative to the specified pedigree; thus, by definition, the founders are unrelated.

The pedigree of Figure 1.1 will be used extensively in examples throughout this book. It is a true pedigree structure, and derives from a study of Werner's syndrome, a rare recessive trait (Goddard et al., 1996). The pedigree has five founders and ten non-founders. In accordance with standard notation, male individuals are shown as squares, and females as circles. The form in which the pedigree is shown here is a *marriage node graph*. Individuals who together produce offspring are connected to a single *marriage node*, which is in turn connected to the resulting offspring. This pedigree was ascertained because the final individual was known to be the offspring of a first-cousin marriage, and was affected by a rare recessive trait. Typically affectation status for a trait of interest is depicted by shading, as in Figure 1.1. It was later discovered that each of the parents of the final individual is also the offspring of a marriage between first cousins.

The meiosis indicators determine the descent of genes in a pedigree. Figure 1.2 gives an example, using the pedigree of Figure 1.1. The meiosis indicators shown under each individual are for the paternal and maternal meiosis to that individual. For easier visualization males and paternal meioses or genes are shown to the left, and females and maternal meioses and genes to the right. For example it is seen that the paternal gene of the final individual is the same gene (labeled "8") as the maternal gene of his maternal grandfather. The gene does not descend from

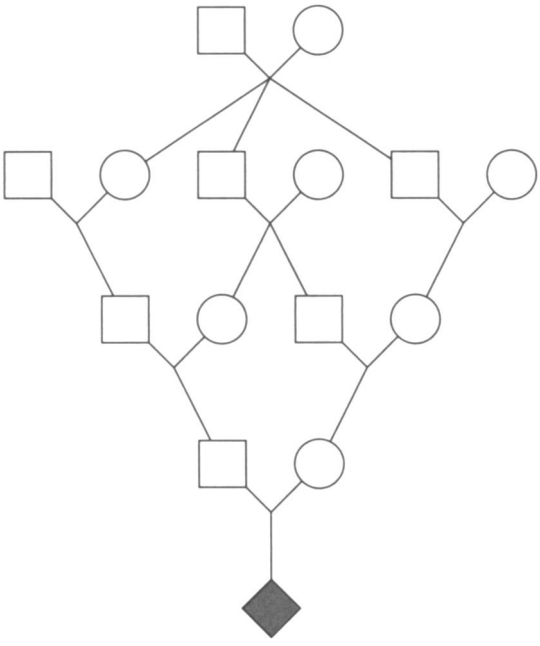

FIGURE 1.1. *An example pedigree from Goddard et al. (1996)*

grandparent to grandchild (the mother does not carry this gene), but both are copies of the same founder gene.

The meiosis indicators $\mathbf{S}$ determine the descent of founder genes. Together with the allelic type of each founder gene at each locus (the founder haplotypes), $\mathbf{S}$ determines the complete configuration of genotypes $\mathbf{G}$ on the pedigree. The independence of meioses then leads to a straightforward conditional independence structure for genotypes on a pedigree. Individuals receive copies of their parents' genes, and, together with their spouses, segregate genes to their offspring.

This conditional independence leads to one formulation of the probability of phenotypic data $\mathbf{Y}$ observed on a pedigree. The probability of the genotypes $\mathbf{G} = \{G_i\}$ is

$$\Pr(\mathbf{G}) = \prod_{\text{founders } i} \Pr(G_i) \prod_{\text{nonfounders } i} \Pr(G_i | G_{M_i}, G_{F_i}) \tag{1.4}$$

where G_i is the genotype of individual i, and M_i and F_i are the parents of individual

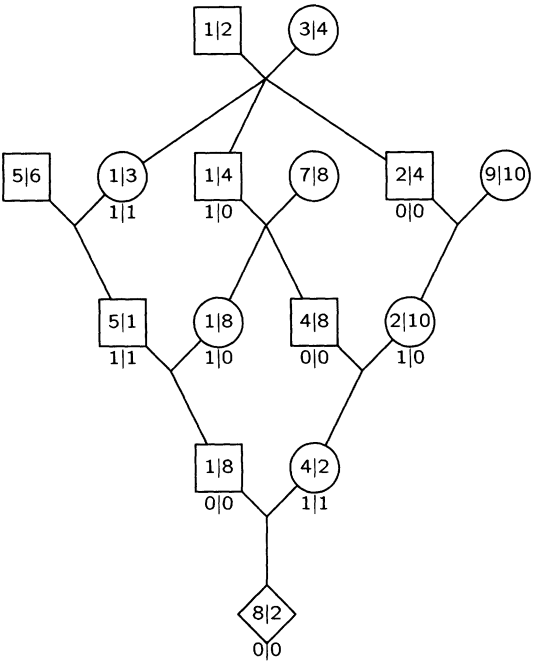

FIGURE 1.2. *Meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, at any given locus j. The indicators $S_{i,j}$ are shown under the offspring individual, while the resulting labeled founder genes are shown within each individual*

i. The data on individuals are determined by their underlying genotypes:

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{G}} \left(\prod_{\text{observed } i} \Pr(Y_i \mid G_i) \right) \Pr(\mathbf{G}) \tag{1.5}$$

The genotype G_i of individual i is the multilocus genotype: that is, a pair of haplotypes over all the relevant genetic loci. The phenotype Y_i may be a multivariate phenotype, with qualitative and/or quantitative components.

Two alternative views of the conditional independence structure are shown in Figure 1.3: this pedigree is slightly modified from our usual example, in order to have an individual with two spouses. As can be seen from equations (1.4) and (1.5), the conditional probability of a genotype G_i of individual i, given the genotypes of all other pedigree members, and given the data $\mathbf{Y}$, depends only on the data Y_i on individual i, and on the genotypes of parents, spouse(s), and offspring (Figure 1.3(a)). At a finer scale, provided paternal and maternal genes of individuals are distinguished, we may consider the dependence structure among

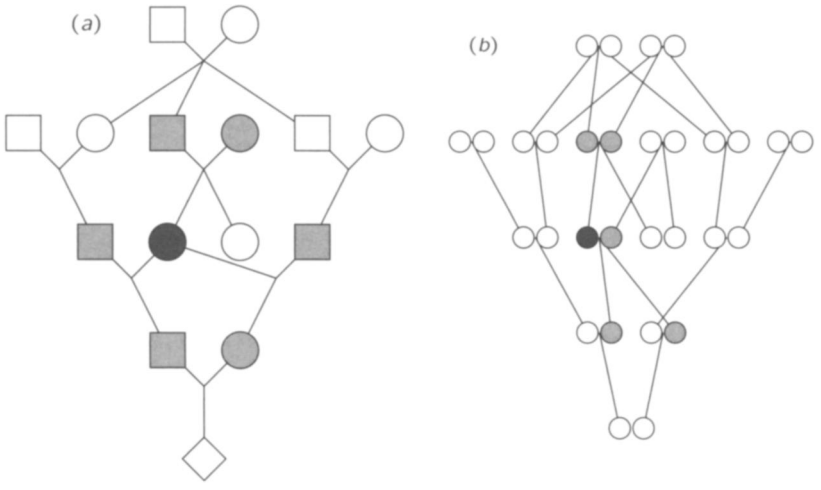

FIGURE 1.3. *The conditional independence neighborhood structure on a pedigree: (a) the individual neighborhood, and (b) the haplotype neighborhood. The reference individual (a) or haplotype (b) is dark shaded. The individuals [haplotypes] defining the local dependence structure for the reference individual [haplotype] are light shaded*

haplotypes (Figure 1.3(b)). Here, for example, for a paternal haplotype of a female individual i, the dependence is on the data Y_i, the maternal haplotype of i, the two haplotypes of the father of i, and the maternal haplotypes of the children of i. These are the haplotypes that segregate to, with, or from the paternal haplotype of i. The set of possible states of a haplotype neighborhood is smaller than of the genotypic neighborhood, since there are fewer haplotypes than multilocus genotypes, the latter being pairs of haplotypes; however there are more haplotype neighborhoods in the pedigree. Either can be more computationally efficient to consider in computing the probability of data on the pedigree (equation (1.5)).

1.4 Models, parameters, and inferences

Considered as a function of the parameters θ of the genetic model the probability $\Pr(\mathbf{Y})$ of equation (1.5) is the likelihood function $L(\theta)$. Broadly, there are three classes of parameters. First, there are *population parameters*, such as allele frequencies and allelic associations within and among loci. These index the probability distributions of founder genotypes and haplotypes in equation (1.4). Some examples of the estimation of these parameters from population data will be considered in Chapter 2. Other parameters such as those of assortative mating also enter into the probabilities of founder genotypes. Phenotypic correlations between

spouses impose genotypic dependencies which influence the probability distributions for offspring data. However, assortative mating will not be considered further in this monograph.

Second, there are *transmission parameters* which index the probability distributions of the meiosis indicators, and hence the probability of offspring genotypes, conditional on those of parents (equation (1.4)). Most importantly, there are parameters such as recombination frequencies, which characterize the dependence in meiosis among loci on a chromosome. The estimation of recombination frequencies will be addressed in Chapter 4, and linkage analysis more generally will be addressed throughout the monograph. The other major class of transmission parameters are those characterizing any deviations from Mendelian segregation. In some approaches to segregation analysis (Elston and Stewart, 1971), a test of Mendelian segregation proportions is performed. Theoretically, each heterozygote should transmit each allele with equal probability, and the probability distribution of the meiosis indicators should not depend upon the allelic types of the genes. However, caution is necessary in interpreting the results of such tests. Apparent distortion may result from selection; offspring individuals surviving to be typed, or to reproduce, may not be a random sample of those resulting from meiosis. In the case of crop plants, or domestic animals, there may be very strong artificial selection on certain loci, which affects the apparent segregation at loci on the same chromosome. In the case of studies of data on human pedigrees, similar apparent distortions can result from the ascertainment of pedigrees, or of parts of a pedigree, in which a particular trait is segregating. This ascertainment also leads to distorted segregation patterns at linked marker loci showing any allelic associations with the trait locus. Indeed, some tests for linkage in the presence of trait-marker allelic associations are tests of apparent segregation distortion in the meioses to affected offspring. Ascertainment is an important topic in the analysis of data on human pedigrees, and there is a large literature from Weinberg (1912) to Karunaratne and Elston (1998). However, it is outside the scope of this monograph.

Third, there are *penetrance parameters* indexing the relationship between genotype and phenotype. These enter only into $\Pr(\mathbf{Y} \mid \mathbf{G})$ (equation (1.5)). The probability that an individual carrying a certain allele is affected by a trait is known as the penetrance of the allele, which includes the degree of dominance. Another penetrance parameter is the probability of phenocopies (individuals exhibiting the phenotype of a genetic trait, but not having the predisposing genotype). More generally, penetrance parameters may characterize allelic and genotypic contributions to a quantitative trait, and the effects of individual environmental effects and covariates. Important covariates include gender and age; many complex traits are age and gender dependent. Also, different genotypes predisposing to the same disease may have different effects on age of onset. Additionally, there may be interaction effects, between alleles at different loci contributing to a given trait (epistasis), between multiple traits affected by alleles at a single locus (pleitropy), or between genetic effects and environmental covariates. The focus of this monograph is on Mendelian traits, such as DNA markers. We shall not consider the broad spectrum of parameters indexing the relationship between genotype and complex phenotypes.

The primary focus of this monograph is methods for inference from genetic data on pedigrees. We shall focus on inferences about the parameters of genetic models; that is, on segregation and linkage analysis. As data become increasingly available on a genomic array of markers, we focus on genetic mapping and the analysis of genetic maps. However, with a given genetic map, the probability of data $\Pr_{\gamma}(\mathbf{Y})$ (equation (1.5)) provides a likelihood for an hypothesized pedigree structure among individuals. Thus, pedigree validation and relationship estimation, using a genomic array of linked DNA markers, are methodologically analogous to segregation and linkage analysis. Other inference questions also require the computation of a probability $\Pr(\mathbf{Y})$, of phenotypes observed on some members of a pedigree structure. For example when both genetic model and pedigree structure are assumed correctly known, data provide information for the inference of ancestral origins of alleles (Geyer and Thompson, 1995), or of phenotypic risks for individuals.

Chapter 2

Likelihood, Estimation and Testing

2.1 Likelihood and log-likelihood.

In this and the following section, we review briefly the basic ideas and results of likelihood inference: details may be found in any standard mathematical statistics text for beginning graduate students. A vector of data random variables, $\mathbf{Y}$, whose value $\mathbf{y}$ is observed, has one of a family of probability distributions $\{P_\theta; \theta \in \Theta\}$, indexed by a *parameter* θ in *parameter space* Θ. The goals of estimation are to make inferences about which P_θ gave rise to the observed $\mathbf{y}$, and to assess the uncertainty associated with this inference.

The *likelihood* is $L_{\mathbf{y}}(\theta) = P_\theta(\mathbf{y})$, a function of θ. The likelihood provides the connection between the data $\mathbf{y}$ and the probability model P_θ. A *statistic* is a function of the data random variables $\mathbf{Y}$, an *estimator* $T = T(\mathbf{Y})$ is a statistic taking values in Θ, while the *estimate* is $T(\mathbf{y})$, the value taken by the estimator that is used to estimate θ.

For example, suppose Y_i, $i = 1, ..., n$ are independent identically distributed Bernoulli random variables, $B(1, \theta)$, the indicators of success in n independent trials, each with success probability θ. Then $P_\theta(y) = \theta^y(1-\theta)^{1-y}$ $(y = 0, 1)$ for each trial, and $L(\theta) = \prod_1^n(\theta^{y_i}(1-\theta)^{1-y_i})$. The log-likelihood is

$$\ell(\theta) = \log L(\theta) = (\sum_1^n y_i)\log(\theta) + (n - \sum_1^n y_i)\log(1-\theta). \tag{2.1}$$

Note that the (log)-likelihood depends only on the value of $T = \sum_1^n Y_i$, the total number of successes, which has a binomial $B(n, \theta)$ distribution. The likelihood based on the binomial probability of the observed value t of T is

$$\begin{aligned} L(\theta) &= P_\theta(T = t) = \frac{n!}{k!(n-k)!}\theta^t(1-\theta)^{n-t} \\ \ell(\theta) &= \log L(\theta) = \text{const} + t\log(\theta) + (n-t)\log(1-\theta). \end{aligned} \tag{2.2}$$

Up to an additive constant which does not depend on θ, the log-likelihood (2.2) is the same as that of equation (2.1). A statistic T for which this is the case is said to be *sufficient.* It is immaterial whether one considers the likelihood based on the full data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ or that based on a sufficient statistic such as T. Log-likelihoods are defined only up to an additive constant; only ratios of likelihoods are relevant for inference.

The *maximum likelihood estimate* (MLE) maximizes the likelihood as a function of θ, to give the value of θ that "best explains" the data $\mathbf{y}$. To obtain the MLE, we maximize $P_\theta(\mathbf{y})$, or $\log P_\theta(\mathbf{y})$ with respect to θ. For example, differentiating the log-likelihood (2.2) with respect to θ

$$\ell'(\theta) = \frac{t}{\theta} - \frac{n-t}{1-\theta}$$

Maximizing (2.2) by setting the derivative $\ell'(\theta)$ equal to 0 gives the MLE $\hat{\theta} = t/n$. In general, the equation $\ell'(\theta) = 0$ is known as the *likelihood equation.*

An estimator, $T(\mathbf{Y})$, is *unbiased* if, for any $\theta \in \Theta$, $\mathbf{Y} \sim P_\theta \implies \mathrm{E}(T(\mathbf{Y})) = \theta$, where $\mathrm{E}(\cdot)$ denotes expectation. We rewrite this definition as $\mathrm{E}_\theta(T(\mathbf{Y})) = \theta$ for all $\theta \in \Theta$, the subscript indicating the "true" θ-value—the value indexing the probability distribution with respect to which the expectations are evaluated. The *bias* of estimator $T(\mathbf{Y})$ is $b_T(\theta) = \mathrm{E}_\theta(T(\mathbf{Y})) - \theta$. An unbiased estimator is "correct, on average", over repetitions of the experiment. For example, if T is binomial $B(n, \theta)$, then $\mathrm{E}_\theta(T) = n\theta$, so the MLE is unbiased. However, unbiasedness alone is a very weak criterion. Some unbiased estimators may have poor properties, while many "good" estimators are biased. In particular many MLEs are biased, but under very broad conditions the bias decreases as the sample size increases.

A more important criterion is that an estimator should have small *mean square error* (mse). The mse of estimator $T(\mathbf{Y})$ is $\mathrm{mse}_\theta(T) = \mathrm{E}_\theta((T(\mathbf{Y}) - \theta)^2)$. If T is unbiased, $\mathrm{mse}_\theta(T) = \mathrm{var}_\theta(T)$, while, in general,

$$\mathrm{mse}_\theta(T) = \mathrm{var}_\theta(T) + (b_T(\theta))^2.$$

For example, for the unbiased maximum likelihood estimator T/n of the binomial parameter θ,

$$\begin{aligned} \mathrm{mse}(T/n) &= \mathrm{var}(T/n) = \mathrm{var}(T)/n^2 \\ &= n\theta(1-\theta)/n^2 = \theta(1-\theta)/n \end{aligned} \tag{2.3}$$

Consider an n-sample $\mathbf{Y}^{(n)} = (Y_1, \ldots, Y_n)$, where the components Y_i are independent and identically distributed, and a sequence of estimators (T_n) where $T_n = T(\mathbf{Y}^{(n)})$. Then the sequence of estimators (T_n) is *consistent for* θ if, for every $\theta \in \Theta$, and every $\epsilon > 0$, $P_\theta(|T_n - \theta| > \epsilon) \to 0$ as $n \to \infty$. In the binomial example, equation (2.3), the mse of the MLE, T/n, tends to 0 as $n \to \infty$, and hence the sequence of estimators, (T/n), is consistent.

Clearly, maximization of $L(\theta)$ is equivalent to maximization of $\ell(\theta) = \log(L(\theta))$. Moreover, if $\alpha(\theta)$ is a one-one function of θ then $\widehat{\alpha} = \alpha(\widehat{\theta})$. Likelihood is a pointwise function of θ; transformation of the parameter space Θ does not alter the likelihood.

2.2 Estimation, information, and testing

In likelihood inference, a key entity is the expected log-likelihood $\mathrm{E}_{\theta_0}(\log(P_\theta(\mathbf{Y})))$. This notation denotes that the true value of the parameter θ is θ_0, and it is the distribution under θ_0 with respect to which expectations are taken. The expected log-likelihood is thus a function both of the true θ_0 and the hypothesized θ. From the convexity of the function $-\log(\cdot)$, it follows by Jensen's inequality that

$$\begin{aligned}
\mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{Y}))) - \log(P_\theta(\mathbf{Y})) &= \mathrm{E}_{\theta_0}\left(-\log\left(\frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})}\right)\right) \\
&\geq -\log \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})}\right) \\
&= -\log\left(\sum_{\mathbf{y}} \frac{P_\theta(\mathbf{y})}{P_{\theta_0}(\mathbf{y})} P_{\theta_0}(\mathbf{y})\right) \\
&= -\log(\sum_{\mathbf{y}} P_\theta(\mathbf{y})) \\
&= -\log(1) = 0 \qquad (2.4)
\end{aligned}$$

Thus the expected log-likelihood is maximized with respect to θ by $\theta = \theta_0$: the expected log-likelihood is maximized at the true value of the parameter. The non-negative difference

$$K(\theta;\theta_0) = \mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{Y})) - \log(P_\theta(\mathbf{Y})))$$

is known as the *Kullback-Leibler information* (Kullback and Leibler, 1951). One of the fairly immediate consequences of equation (2.4) is that under very broad conditions MLEs are consistent.

A related result is the *Cramèr-Rao lower bound* which says that (subject to some regularity conditions) no unbiased estimator can have a variance smaller than

$$\left[\mathrm{E}_{\theta_0}\left(-\frac{\partial^2}{\partial\theta^2}\log(P_{\theta_0}(\mathbf{Y}))\right)\right]^{-1}$$

The quantity within the square brackets is known as the *Fisher information.* The larger the information, the smaller the variance can be. Subject to a few additional conditions, MLEs are asymptotically approximately Normal (Gaussian), with mean θ_0, the true parameter value, and variance the inverse of the Fisher information. This says that, *in large samples,* MLEs are the *best estimators.* The required regularity conditions will be satisfied for most of the examples discussed in this monograph. A condition which may sometimes fail is that the true value θ_0 should lie in the interior of the parameter space Θ.

Of course, the value of θ_0 is unknown, but at least for large samples, the MLE $\hat{\theta}$ is close to the true value θ_0. Thus, $\hat{\theta}(\mathbf{y})$ may be substituted for θ_0 in the Fisher information, to obtain an estimate of the variance of the MLE. In fact, the expectation in the Fisher information can be hard to compute. Then, at least

for large samples, an alternative is the *observed information*

$$-\frac{\partial^2}{\partial\theta^2}\log(P_{\theta_0}(\mathbf{Y}))$$

evaluated by substituting the observed $\mathbf{y}$ for $\mathbf{Y}$ and $\hat{\theta}(\mathbf{y})$ for θ_0. The theoretical details and justification may be found in a mathematical statistics text.

To provide an example which should be familiar to readers, we return to the case of a binomial random variable: T is $B(n,\theta)$. As before (equation (2.2))

$$\ell(\theta) \;=\; \text{const} \;+\; T\;\log(\theta) \;+\; (n-T)\;\log(1-\theta)$$

and the MLE is T/n which has expectation θ and variance $\theta(1-\theta)/n$. Now,

$$\ell''(\theta) = -\frac{T}{\theta^2} - \frac{(n-T)}{(1-\theta)^2}.$$

Since $\mathrm{E}_\theta(T) = n\theta$, and $\mathrm{E}_\theta(n-T) = n(1-\theta)$, the Fisher information is $n/\theta(1-\theta)$. Thus in this example, the MLE has the smallest possible variance.

In practice, we estimate the variance as

$$\hat{\theta}(1-\hat{\theta})/n \;=\; t(n-t)/n^3$$

where t is the observed value of T. In fact, the same result is given by substituting $\hat{\theta} = t/n$ for θ in $-1/\ell''(\theta)$, without going through the expectation step. It is not in general true that the two methods of obtaining an estimated variance of the MLE give identical formulae.

Just as the maximum likelihood estimate is the value of the parameter that best explains the observed data, the maximized value of the likelihood is a measure of how well this parameter value is supported by the data, relative to how well other parameter values are supported by the observation of these data. Accordingly, we define

$$L(\Theta_0) \;=\; \max_{\theta\in\Theta_0}(L(\theta))$$

as a measure of support for the hypothesis $\theta \in \Theta_0 \subset \Theta$, and

$$\Lambda(\Theta_1 : \Theta_0) \;=\; L(\Theta_1)/L(\Theta_0)$$

as a measure of the relative support for the two hypotheses $\theta \in \Theta_1$ and $\theta \in \Theta_0$.

In the case when $\Theta_0 \subseteq \Theta_1$, $\Lambda \geq 1$, and $2\log_e \Lambda \geq 0$. Again subject to regularity conditions, asymptotically, if $\theta \in \Theta_0$ is true, then $2\log_e \Lambda$ is approximately distributed as a chi-squared (χ^2) random variable, with degrees of freedom equal to $\dim(\Theta_1) - \dim(\Theta_0)$. If the true value θ_0 is not in the hypothesis space Θ_0 but is in Θ_1, then $2\log_e \Lambda \to \infty$ at a rate which depends on the minimum Kullback-Leibler information:

$$\inf_{\theta\in\Theta_0} K(\theta;\theta_0) \;=\; \inf_{\theta\in\Theta_0} (\mathrm{E}_{\theta_0}(\log(P_{\theta_0}(\mathbf{Y})) \;-\; \log(P_\theta(\mathbf{Y}))\;))$$

The regularity conditions in order that these results hold are essentially the same as the ones needed for the asymptotic results about MLEs. They will hold in the examples we discuss.

In particular, much of the data in genetics is multinomial, consisting of counts of outcomes of various types. It is therefore useful to consider the case of the general multinomial model. Suppose there are r possible outcomes, having probabilities $p_i, i = 1, \ldots, r$, and a vector of parameters θ, so p_i is $p_i(\theta)$. The log-likelihood is

$$\ell = \text{const} + \sum_{i=1}^{r} n_i \log p_i \tag{2.5}$$

For the general model, $\sum_{i=1}^{r} p_i = 1$ with no other constraints:

$$\ell = \sum_{i=1}^{r} n_i \log p_i = \sum_{i=1}^{r-1} n_i \log p_i + n_r \log(1 - \sum_{i=1}^{r-1} p_i)$$

$$\frac{\partial \ell}{\partial p_i} = \frac{n_i}{p_i} - \frac{n_r}{p_r}$$

for $i = 1, \ldots, r-1$, giving the MLE $\widehat{p_i} = n_i/n$. The maximized value of the log-likelihood is

$$\hat{\ell} = \sum_{i=1}^{r} n_i \log \widehat{p_i} = \sum_{i=1}^{r} n_i \log n_i - n \log n \tag{2.6}$$

The dimension of the general hypothesis space is $r - 1$ since the p_i are constrained to sum to 1.

Under a constrained model, the outcome probabilities p_i will be functions of some parameters θ_j, where normally the dimensionality of θ is less than $r - 1$. To estimate θ we must solve the equations

$$\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{r} \frac{n_i}{p_i} \frac{\partial p_i}{\partial \theta_j} \quad \text{for all } j$$

It is also possible to find the Fisher information:

$$\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{r} \frac{n_i}{p_i^2} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} + \sum_{i=1}^{r} \frac{n_i}{p_i} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k}$$

Now $\mathrm{E}(n_i) = np_i$, so

$$\mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k}\right) = \sum_{i=1}^{r} \frac{np_i}{p_i^2} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} - \sum_{i=1}^{r} \frac{np_i}{p_i} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k}$$

$$= n \sum_{i=1}^{r} \frac{1}{p_i} \frac{\partial p_i}{\partial \theta_j} \frac{\partial p_i}{\partial \theta_k} - n \sum_{i=1}^{r} \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k}$$

$$= n\sum_{i=1}^{r}\frac{1}{p_i}\frac{\partial p_i}{\partial\theta_j}\frac{\partial p_i}{\partial\theta_k} - n\frac{\partial^2\sum_{i=1}^{r}p_i}{\partial\theta_j\partial\theta_k}$$

$$= n\sum_{i=1}^{r}\frac{1}{p_i}\frac{\partial p_i}{\partial\theta_j}\frac{\partial p_i}{\partial\theta_k} \tag{2.7}$$

since $\sum_{i=1}^{r} p_i \equiv 1$. Equation (2.7) is sometimes known as Fisher's formula.

2.3 Population allele frequencies

In this section, we consider three examples of the above formulation, in the context of estimation of population allele frequencies. Consider a single genetic locus, with k alleles A_j and having population frequencies q_j, $j = 1,\ldots,k$. Now in a random-mating population, the allelic types of the maternal and paternal genes in an individual are independent. Thus the probability an individual is homozygous A_jA_j is q_j^2, while the probability the individual is heterozygous A_jA_l $(j < l)$ is $2q_jq_l$. These genotype frequencies are known as Hardy-Weinberg proportions, and a population exhibiting genotypes in these proportions is said to be in Hardy-Weinberg equilibrium.

First suppose the alleles A_j are codominant, and a random sample of n individuals is taken from a population assumed to be in Hardy-Weinberg proportions. Suppose that n_{jl} $(j \le l)$ individuals are observed to be of genotype A_jA_l. As above (equation (2.5)), the log-likelihood is

$$\ell = \text{const} + \sum_{j=1}^{k} n_{jj}log(q_j^2) + \sum_{1\le j<l\le k} n_{jl}\log(2q_jq_l)$$

$$= \text{const} + \sum_{j=1}^{k} m_j log(q_j)$$

where $m_j = 2n_{jj} + \sum_{l<j} n_{lj} + \sum_{j<l} n_{jl}$, is the number of A_j alleles among the $2n$ alleles of th n sampled individuals. Hence the MLE of q_j is $m_j/2n$, the sample proportions of the allelic types. The MLE has variance $q_j(1-q_j)/2n$.

Most natural populations show some degree of subdivision or structure, and so do not exhibit Hardy-Weinberg equilibrium. The deviation from Hardy-Weinberg proportions may be small and detectable only from large samples. Testing Hardy-Weinberg proportions is straightforward in the case of a random sample of individuals typed at a locus with codominant alleles. Under the general model, there are $\frac{1}{2}k(k+1)$ genotypic counts n_{jl} with the maximum log-likelihood given by equation (2.6), while assuming Hardy-Weinberg proportions, there are k allelic counts m_j with the same multinomial form of maximum log-likelihood. The dimension of the larger hypothesis space is $\frac{1}{2}k(k+1)-1$, and of the smaller is $k-1$. If Hardy-Weinberg proportions do hold in the population, then twice the difference in log-likelihoods is distributed as a chi-squared random variable on $\frac{1}{2}k(k-1)$ degrees of freedom ($\chi^2_{\frac{1}{2}k(k-1)}$).

A locus with two alleles is said to be *diallelic*, although the alternative *biallelic* is now used increasingly in the literature. As an example of the use of Fisher information, consider the case of a diallelic trait locus, with a recessive allele with allele frequency q. Assuming Hardy-Weinberg proportions, there are two phenotypic categories ($r = 2$), with population frequencies $p_1(q) = q^2$, $p_2(q) = 1 - q^2$. Suppose n individuals are sampled, and t are found to be of the recessive phenotype. Since $p_1(q)$ is a 1-1 transformation of q, over the parameter space $0 \le q \le 1$, the MLE of q is $\widehat{q} = \sqrt{(\widehat{p_1})} = \sqrt{t/n}$. This may also be verified by direct differentiation of the log-likelihood

$$\ell(q) \;=\; t\,\log(q^2) \;+\; (n-t)\,\log(1-q^2)$$

Now also

$$\frac{\partial p_1}{\partial q} \;=\; 2q \;\text{ and }\; \frac{\partial p_2}{\partial q} \;=\; -\,2q,$$

so using Fisher's equation (2.7)

$$\begin{aligned} \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial q^2}\right) &= n\left(\frac{1}{q^2}(2q)^2 \;+\; \frac{1}{1-q^2}(-2q)^2\right) \\ &= \frac{4n}{(1-q^2)} \end{aligned}$$

Thus the large-sample variance of the MLE is $(1-q^2)/4n$, which is $(1+q)/2q$ times larger than the variance $q(1-q)/2n$ obtained if the genotypes were observable. Of course, when there are only two phenotypes, there are no degrees of freedom to test for Hardy-Weinberg proportions.

As another example, consider the estimation of allele frequencies at a diallelic locus, when, instead of random individuals, we sample parent-offspring pairs. This might arise, for example, if our sample was of mothers with new-born infants. Table 2.1 shows the conditional and joint probabilities of feasible mother-child combinations.

parent genotype	probability	Pr(child|parent) for child genotype			Pr(parent, child) for child genotype		
		A_iA_i	A_iA_j	A_iA_l	A_iA_i	A_iA_j	A_iA_l
A_iA_i	q_i^2	q_i	q_j	q_l	q_i^3	$q_i^2 q_j$	$q_i^2 q_l$
A_iA_j	$2q_iq_j$	$\frac{1}{2}q_i$	$\frac{1}{2}(q_i+q_j)$	$\frac{1}{2}q_l$	$q_i^2 q_j$	$q_iq_j(q_i+q_j)$	$q_iq_jq_l$

TABLE 2.1. *Conditional and joint probabilities of feasible mother-child genotype combinations*

In the case $k = 2$, let n_{ij} be the number of mother-offspring pairs in which the mother has genotype g_i and the offspring has genotype g_j, where $g_0 = A_1A_1$, $g_1 = A_1A_2$ and $g_2 = A_2A_2$. Since $q_1 + q_2 = 1$, every term in Table 2.1 is a product of allele frequencies, and the multinomial log-likelihood reduces to

$$\ell \;=\; \sum_{(i,j)} n_{ij} \log \Pr(g_i, g_j)$$

$$
\begin{aligned}
&= n_{00}\log(q_1^3) + n_{01}\log(q_1^2 q_2) + n_{10}\log(q_1^2 q_2) + n_{11}\log(q_1 q_2) \\
&\quad + n_{12}\log(q_1 q_2^2) + n_{21}\log(q_1 q_2^2) + n_{22}\log(q_2^3) \\
&= (3n_{00} + 2(n_{01} + n_{10}) + n_{11} + n_{12} + n_{21})\log q_1 + \\
&\quad (3n_{22} + 2(n_{21} + n_{12}) + n_{11} + n_{10} + n_{01})\log q_2 \\
&= m_1 \log q_1 + m_2 \log q_2 \qquad (2.8)
\end{aligned}
$$

where m_1 is the number of distinct A_1 alleles, and m_2 is the number of distinct A_2 alleles, in the set of pairs. (By "distinct" we mean that we do not count both copies of an allele which segregates from parent to offspring.) The MLE of q_1 is thus $m_1/(m_1 + m_2)$. Note that

$$
\begin{aligned}
m_1 + m_2 &= 3(n_{00} + n_{01} + n_{10} + n_{21} + n_{12} + n_{22}) + 2n_{11} \\
&= 3n - n_{11}
\end{aligned}
$$

where n is the number of parent-offspring pairs. Although finding the MLE is a matter of "gene-counting", the total number of distinct genes to be counted is not $4n$, since parent and offspring share one gene, nor even $3n$. For each $(g_1, g_1) = (A_1A_2, A_1A_2)$ pair, one gene of allelic type A_1 and one of type A_2 can be counted, but the third distinct gene may be of either type, and does not contribute to the likelihood.

	factor freq.		phenotype frequencies			
	A	B	A	B	AB	0
Data			0.422	0.206	0.078	0.294
H_1 theory	p	q	$p(1-q)$	$p(1-q)$	pq	$(1-p)(1-q)$
H_1 fitted	0.500	0.284	0.358	0.142	0.142	0.358
H_2 theory	p	q	$p^2 + 2pr$	$q^2 + 2qr$	$2pq$	r^2
H_2 fitted	0.295	0.155	0.411	0.194	0.091	0.303

TABLE 2.2. *Data and estimated frequencies for Bernstein's analysis of ABO blood type determination*

As a final example in this section, we consider the classic analysis of Bernstein (1925) who established the mode of determination of the *ABO* blood types using data on population phenotype frequencies. The development in terms of likelihood ratio tests is due to Edwards (1972). Bernstein reported *ABO* blood types on a sample of 502 individuals: 42.2% type A, 20.6% type B, 7.8% type AB and 29.4% type O (Table 2.2). It is a minor mystery of Bernstein's data that these proportions do not give integer counts with a sample of $n = 502$; however we ignore that question here.

Now there were two prevailing hypotheses for the determination of the *ABO* blood types, the first, H_1 being that A and B are independently inherited factors, The frequency of individuals in the sample having the factor A is 0.500 (blood types A or AB), and B is 0.284 (blood types B or AB). As pointed out by Bernstein, independence of the factors would give an AB frequency of $0.500 \times$

$0.284 = 0.142$ much larger than the 0.078 observed. More rigorously, we can perform a likelihood ratio test of H_1 against the general multinomial alternative. For the general alternative, the fitted frequencies are the observed frequencies, and the log-likelihood is

$$\begin{aligned}\widehat{\ell} &= 502(.422\log.422 + .206\log.206 + .078\log.078 + .294\log.294)\\ &= -626.71\end{aligned}$$

Under the hypothesis H_1 the estimated frequencies are as shown in Table 2.2, and the log-likelihood is

$$\begin{aligned}\ell_1 &= 502(.422\log.358 + .206\log.142 + .078\log.142 + .294\log.358)\\ &= -647.50\end{aligned}$$

Twice the log-likelihood difference is 41.58, and would be the value of a χ^2_1 random variable if H_1 were true. Clearly, H_1 is rejected.

The second hypothesis, H_2 is that A and B are the two non-null alleles of a single system. If the three alleles A, B and O have frequencies p, q and r ($p+q+r=1$), and if Hardy-Weinberg proportions hold, then the frequencies of the four blood types are p^2+2pr, q^2+2qr, $2pq$ and r^2 (Table 2.2). Bernstein pointed out that the sum of the A and O blood type frequencies is $(p+r)^2$, or one minus the square root of this frequency is $(1-p-r)=q$. Similarly one minus the square root of the sum of the B and O blood type frequencies is p, and the square root of the O blood type frequency is r. The sum of these three numbers should be one. For his data

$$(1-\sqrt{0.422+0.294}) + (1-\sqrt{0.206+0.294}) + \sqrt{0.294} = 0.99$$

which is close to one, suggesting a good fit. Again, more formally, we may perform a likelihood ratio test. However, finding the MLEs of the parameters p, q and r is not simple; in fact, we shall discover in section 2.5 that these MLEs are $\widehat{p}=0.2945$ and $\widehat{q}=0.1547$, with the resulting fitted frequencies given in Table 2.2. The fitted frequencies are all close to the observed ones, and the log-likelihood is

$$\begin{aligned}\ell_2 &= 502(.422\log.4114 + .206\log.1942 + .078\log.0911 + .294\log.3033)\\ &= -627.52\end{aligned}$$

Twice the log-likelihood difference between this and the general alternative is now only 1.62. Again, this is the value of a χ^2_1 random variable if H_2 is true, and this hypothesis is not rejected.

Of course, there is also evidence on the ABO blood type determination in the transmission of genes from parents to children. For example, under H_2 an AB parent cannot have an O child, while under H_1 this may happen. Both inheritance patterns and population frequencies can provide information on genetic mechanisms. Bernstein's analysis is perhaps the first example of determination of the genetic model underlying a trait from population frequency data, rather than from the inheritance patterns in pedigrees.

2.4 The EM algorithm; general formulation

Many of the problems in genetic analysis fall within the classical *missing data* or *latent variable* framework. Many data may be missing, in the sense that some pedigree members may be unobserved, or not all marker phenotypes observed even for some available pedigree members. We therefore prefer the term *latent variables* for unobservable features, such as the multilocus haplotypes of individuals (equation (1.5)), or the meiosis indicators that specify the descent of genes in pedigrees (equation (1.2)). An important approach to likelihood analysis, and specifically to maximum likelihood estimation, in such latent variable problems was provided by Dempster et al. (1977). Although their approach had been developed previously in many special cases, they provided the overall framework, giving it the name the *EM algorithm*, or *expectation-maximization* algorithm.

For generality, we denote latent variables by $\mathbf{X}$, bearing in mind that for our examples, these will generally be meiosis indicators, genotypes, indicators of genotypic status or linkage phase, or genotypic or allelic counts. For simplicity, we use summation rather than integration over latent variables, since for the majority of our examples, the latent variables are discrete. The structure of any latent variable problem is that the likelihood $L(\theta)$ from observed data values $\mathbf{y}$ of the data random variables $\mathbf{Y}$ is

$$L(\theta) \;=\; P_\theta(\mathbf{Y}=\mathbf{y}) \;=\; \sum_{\mathbf{x}} P_\theta((\mathbf{X},\mathbf{Y})=(\mathbf{x},\mathbf{y}))$$

Now the joint probability of data and latent variables is

$$P_\theta((\mathbf{X},\mathbf{Y})=(\mathbf{x},\mathbf{y})) \;=\; P_\theta((\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y}))P_\theta(\mathbf{Y}=\mathbf{y}).$$

This joint probability, considered as a likelihood of parameter θ, is known as the *complete-data likelihood.* Taking logs and rearranging,

$$\log L(\theta) \;=\; \log P_\theta((\mathbf{X},\mathbf{Y})=(\mathbf{x},\mathbf{y})) \;-\; \log P_\theta((\mathbf{X}=\mathbf{x}|\mathbf{Y}=\mathbf{y})). \tag{2.9}$$

Now define

$$\begin{aligned} H_{\mathbf{y}}(\theta;\theta^*) &= \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y}=\mathbf{y}) \\ G_{\mathbf{y}}(\theta;\theta^*) &= \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X} \mid \mathbf{Y}=\mathbf{y}) \mid \mathbf{Y}=\mathbf{y}) \end{aligned}$$

The function $H_{\mathbf{y}}(\theta;\theta^*)$ is the *expected complete-data log-likelihood*, while the Kullback-Leibler information (section 2.2) in the conditional distribution of $\mathbf{X}$ given $\mathbf{Y}=\mathbf{y}$ is

$$K_{\mathbf{y}}(\theta;\theta^*) \;=\; G_{\mathbf{y}}(\theta^*;\theta^*) - G_{\mathbf{y}}(\theta;\theta^*).$$

Taking expectations over $\mathbf{X}$, under model θ^*, conditional upon $\mathbf{Y}=\mathbf{y}$, in equation (2.9) we obtain

$$\log L(\theta) \;=\; H_{\mathbf{y}}(\theta;\theta^*) \;-\; G_{\mathbf{y}}(\theta;\theta^*)$$

since $L(\theta)$ does not depend on the random variable $\mathbf{X}$. Now suppose that $\tilde{\theta}$ maximizes $H_{\mathbf{y}}(\theta;\theta^*)$ over θ, and consider

$$\log L(\tilde{\theta}) - \log L(\theta^*) = (H_{\mathbf{y}}(\tilde{\theta};\theta^*) - H_{\mathbf{y}}(\theta^*;\theta^*)) + (G_{\mathbf{y}}(\theta^*;\theta^*) - G_{\mathbf{y}}(\tilde{\theta};\theta^*)) \tag{2.10}$$

Now $\tilde{\theta}$ maximizes $H_{\mathbf{y}}(\theta;\theta^*)$. Also, for any probability distributions $P_\theta(\cdot)$ indexed by parameter θ, $\mathrm{E}_{\theta^*}(P_\theta(\mathbf{X}))$ is maximized by $\theta = \theta^*$ (equation (2.4)). Thus

$$H_{\mathbf{y}}(\tilde{\theta};\theta^*) \geq H_{\mathbf{y}}(\theta^*;\theta^*) \quad \text{and} \quad G_{\mathbf{y}}(\theta^*;\theta^*) \geq G_{\mathbf{y}}(\tilde{\theta};\theta^*) \tag{2.11}$$

$$\text{so} \quad \log L(\tilde{\theta}) \geq \log L(\theta^*) \tag{2.12}$$

with equality only if $\tilde{\theta}$ and θ^* provide the same conditional distribution for $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$.

Thus we have the EM algorithm for finding MLEs (Dempster et al., 1977).
E-step (expectation):

At the current estimate θ^* compute $H_{\mathbf{y}}(\theta;\theta^*) = \mathrm{E}_{\theta^*}(\log P_\theta(\mathbf{X},\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y})$
M-step (maximization):

Maximize $H_{\mathbf{y}}(\theta;\theta^*)$ with respect to θ to obtain a new estimate $\tilde{\theta}$.
E-steps and M-steps are alternated, and, in accordance with equation (2.12) the likelihood is non-decreasing over the process. Where the likelihood surface is unimodal, convergence to the MLE is assured, although it may be slow.

In the case when the complete-data joint probability $P_\theta((\mathbf{X},\mathbf{Y}) = (\mathbf{x},\mathbf{y}))$ is an exponential family of full rank, the EM equations take a particularly simple form. If $T_j(\mathbf{X},\mathbf{Y})$, $j = 1,...,k$ are the natural sufficient statistics, with corresponding natural parameters $\alpha_j(\theta)$, $j = 1,...,k$,

$$\begin{aligned}
P_\theta((\mathbf{X},\mathbf{Y}) = (\mathbf{x},\mathbf{y})) &= c(\theta)\ \exp\Big(\sum_{j=1}^{k} T_j(\mathbf{x},\mathbf{y})\alpha_j(\theta)\Big) \\
H_{\mathbf{y}}(\theta;\theta^*) &= \log c(\theta) - \sum_{j=1}^{k} \mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y} = \mathbf{y})\alpha_j(\theta) \\
\frac{\partial H_{\mathbf{y}}}{\partial \alpha_j} &= \frac{\partial \log c(\theta)}{\partial \alpha_j} - \mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y} = \mathbf{y}) \\
&= \mathrm{E}_\theta(T_j(\mathbf{X},\mathbf{Y})) - \mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y} = \mathbf{y}).
\end{aligned}$$

Thus to implement EM in this case we compute the conditional expectations of the natural sufficient statistics T_j, give the data $\mathbf{Y}$, under the current estimate θ^* and set them equal to their unconditioned expectations to obtain the new estimates $\tilde{\theta}$. Thus the EM algorithm is often discussed in terms of the E-step "imputing" the latent variables conditional upon the data $\mathbf{Y}$ under the currrent estimates θ^*, and the M-step being the maximization of the complete-data log-likelihood, using these imputed variables. Although for many practical cases this is so, some care is needed. Only in the case of an exponential family of full rank is the expected

complete-data log-likelihood a linear function of the natural sufficient statistics T_j. Even in this case, T_j may not be linear in the latent variables $\mathbf{X}$, so that

$$\mathrm{E}_{\theta^*}(T_j(\mathbf{X},\mathbf{y}) \mid \mathbf{Y}=\mathbf{y}) \quad \neq \quad T_j(\mathrm{E}_{\theta^*}(\mathbf{X} \mid \mathbf{Y}=\mathbf{y}),\ \mathbf{y})$$

An example is given in section 2.6.

This monograph will take a likelihood approach to inference, but some of the methods are closely related to those of Bayesian inference. In Bayesian inference, parameters θ are given a *prior* probability distribution $\pi(\theta)$ which expresses information or belief about parameter values before data $\mathbf{Y}$ are observed. After data are observed, beliefs about θ are expressed via the *posterior* distribution

$$\pi(\theta|\mathbf{Y}) \quad = \quad \frac{\pi(\theta)\mathrm{Pr}(\mathbf{Y};\theta)}{\int_\theta \pi(\theta)\mathrm{Pr}(\mathbf{Y};\theta)d\theta}$$

Bayesian inferences are based on the posterior probability distribution for parameters of interest. Clearly the likelihood $L(\theta) = \mathrm{Pr}(\mathbf{Y};\theta)$ is closely related to the Bayesian posterior.

Bayesian inference is often useful where there are many parameters, only a few of which are of interest. The nuisance parameters are integrated over to provide a marginal posterior distribution for a parameter of interest. This is thus often a convenient way to view a multi-parameter likelihood surface, integrating over nuisance parameters with respect to some prior distribution, rather than maximizing over them to obtain a profile likelihood. From the Bayesian viewpoint, there is no difference between latent variables $\mathbf{X}$ and parameters θ, and the conditional probability distribution of $\mathbf{X}$ given observed data $\mathbf{Y}$ would be referred to as a posterior distribution for $\mathbf{X}$, whereas the probability unconditioned on data would be the prior distribution for $\mathbf{X}$ at a given value of θ. To avoid confusion, we shall refer to the distribution of $\mathbf{X}$ given $\mathbf{Y}$, indexed by parameter θ as the *conditional* distribution, and reserve the word *posterior* for a Bayesian posterior for model parameters θ. We shall, however, refer to the model-based distribution for latent variables $\mathbf{X}$ as a *prior* distribution for $\mathbf{X}$. This should not be confused with a Bayesian prior distribution for model parameters θ.

2.5 Gene counting and the ABO blood types

We have seen in the examples of section 2.3 that, where genotypes are observable, estimating allele frequencies is just a matter of *counting the genes.* In a slightly more general sense, the same is true when genotypes cannot be fully observed. "Counting methods" have been used to estimate allele frequencies since the approach was first introduced by Ceppelini et al. (1955). In fact, these methods are particular instances of the EM-algorithm (section 2.4).

Given a sample of n individuals, the phenotypic counts n_j, $i = 1,\ldots,r$, are multinomial, with probabilities $p_i(\mathbf{q})$, where $\mathbf{q} = (q_1,\ldots,q_k)$ is the vector of underlying allele frequency parameters to be estimated:

$$\ell \quad = \quad \log \mathrm{Pr}(n_1,\ldots,n_r) \quad = \quad \sum_{i=1}^{r} n_i \log p_i(\mathbf{q}) \tag{2.13}$$

The complete-data, consisting of the counts m_j of allelic types of all distinct genes in the sample, are also multinomial:

$$\log \Pr(m_1, \ldots, m_k) = \sum_{j=1}^{k} m_j \log q_j.$$

Determining the conditional expected complete-data log-likelihood (E-step), is simply a matter of determining the expectations e_j of allele counts m_j given the phenotypic counts n_i and current estimates of the allele frequencies q_j. The M-step is even simpler: the new estimate of q_j is the proportion e_j/m^*. Here, m^* is the number of distinct genes in the n individuals: for the case of samples of unrelated individuals, $m^* = 2n$.

current q	current $2q/(1+q)$	recessive phenotype $t_1 = 36$ AA	dominant phenotype $t_2 + t_3 = 64$ AB	BB	new $q = (2t_1 + t_2)/2n$
0.5	0.667	36	42.67	21.33	0.573
0.573	0.729	36	46.64	17.36	0.593
0.593	0.745	36	47.66	16.34	0.598
0.598	0.749	36	47.91	16.09	0.600
0.600	0.750	36	48.00	16.00	0.600

TABLE 2.3. *Sequence of EM iterates for the example of estimation of the frequency of a recessive allele*

We consider two examples of the above, the first being the case of a recessive allele, with allele frequency q. Suppose in a sample size $n = 100$ there are $n_1 = 36$ of the recessive type AA. As seen in section 2.3, the MLE of q is $\sqrt{0.36} = 0.6$. Although the EM algorithm is unnecessary here, it provides a useful example.

The three genotypes are AA, AB and BB, with counts say t_i, $(i = 1, 2, 3)$. Now, $n_1 = t_1$, but the counts of AB and BB are unobservable since B is dominant to A. If these counts, t_2 and t_3, were known, then the number of A alleles is $m_1 = 2t_1 + t_2$, and the MLE of q would be $(2t_1 + t_2)/2n$. Further,

$$\Pr(AB \mid AB \text{ or } BB) = \frac{2q(1-q)}{1-q^2} = \frac{2q}{1+q}$$

so

$$\mathrm{E}_q(t_2 \mid n_2 = t_2 + t_3 = 64) = 64\frac{2q}{1+q}.$$

So now the EM-algorithm implements the sequence of iterates shown in Table 2.3. Starting from an arbitrary initial value $q = 0.5$, the proportion $2q/(1+q)$ is computed, and the 64 individuals of dominant phenotype divided into the expected numbers t_2 and t_3 that are that are AB and BB, respectively (E-step). Then a

current values				phenotype A		phenotype B		...
p	q	$\frac{2r}{p+2r}$	$\frac{2r}{q+2r}$	$\Pr(A) = 0.422$		$\Pr(B) = 0.206$		...
				AA	AO	BB	BO	...
0.3	0.3	0.73	0.73	0.115	0.307	0.056	0.150	...
0.308	0.170	0.77	0.86	0.096	0.326	0.029	0.177	...
0.298	0.156	0.79	0.87	0.091	0.331	0.026	0.180	...
0.295	0.155	0.79	0.88	0.089	0.333	0.025	0.181	...

...	phen AB	phen O	new values	
...	$\Pr(AB) = 0.078$	$\Pr(OO) = 0.294$	p	q
...	AB	OO		
...	0.078	0.294	0.308	0.170
...	0.078	0.294	0.298	0.156
...	0.078	0.294	0.295	0.155
...	0.078	0.294	0.295	0.155

TABLE 2.4. *EM iterates for the estimation of ABO allele frequencies. The iterates of allele frequencies, and the resulting conditional probabilities of genotype AO and BO, given phenotypes A and B, respectively, are shown in the upper left panel. Then are shown the resulting expected genotype frequencies, given the observed phenotype frequencies and current allele frequency estimates (E-step). Finally, in the lower right are shown the new iterates of the allele frequencies (M-step)*

new value of q is estimated as $(2t_1 + t_2)/2n$ (M-step). The process is repeated, and convergence to the MLE $\widehat{q} = 0.6$ is obtained within five steps.

The second example provides the MLEs of the ABO blood group allele frequencies discussed in section 2.3. Here the EM-algorithm is in fact one of the easiest ways to find the MLEs, since there is no explicit solution of the likelihood equation. Now, we must partition both the count of A phenotypes into expected counts of AA and AO genotypes, and the B phenotype into BB and BO genotypes:

$$\begin{aligned} \Pr(AO \mid \text{type } A) &= \frac{2pr}{p^2 + 2pr} = \frac{2r}{p + 2r} \\ \Pr(BO \mid \text{type } B) &= \frac{2qr}{q^2 + 2qr} = \frac{2r}{q + 2r}. \end{aligned}$$

Once the counts are partitioned, according to current estimates of allele frequencies, the new estimate of the A allele frequency p is $\Pr(AA) + (\Pr(AO) + \Pr(AB))/2$, and the new estimate of the B allele frequency q is $\Pr(BB) + (\Pr(BO) + \Pr(AB))/2$. Recall, Bernstein (1925) reported a sample of 502 individuals, with frequencies of the four types, A, B, AB and O, 0.422, 0.206, 0.078, and 0.294, respectively. Table 2.4 shown the sequence of EM-iterates, with convergence being obtained, from starting values $p = q = 0.3$ in four iterations. Again, the details of this example are due to Edwards (1972).

One interesting feature of the sequence of iterates in this example is that the value of p does not change monotonely; there is no reason why it should. What is

guaranteed to change monotonely is the value of the log-likelihood, which, for given allele frequencies may be easily evaluated (section 2.3 and equation (2.13)). For this example, over the iterations, the values of the log-likelihood are -687.1242, -628.9991, -627.5693, -627.5262, -627.5246. Note that, typically of the EM algorithm, the log-likelihood increases rapidly in the first steps, and the parameter values move rapidly to the neighborhood of the MLE, whereas the final convergence is much slower. In examples such as this, where evaluation of the log-likelihood is possible, this provides a better check on convergence than a criterion based on the changes in parameter estimates.

2.6 EM estimation for quantitative trait data

For simple qualitative or quantitative traits, were genotypes observable, estimation of penetrance parameters would also be primarily a matter of "counting". However, even in the simplest cases, explicit EM equations are not readily obtained. There may be no single statistic; the complete-data sufficient statistics may be functions of the genotypes G_i of every individual i. Consider, for example, the simplest possible model for a quantitative trait determined by alleles at a single diallelic locus. (For example, the trait value might be an enzyme level.) The phenotypic value is assumed to have a Gaussian distribution, with mean depending on the genotype at the locus, and variance σ_e^2. The penetrance parameters are the three genotypic means, and the residual variance σ_e^2. The only additional parameter is the allele frequency at determining the trait-locus genotype frequencies. The model for the phenotype Y_i of individual i having genotype G_i may be specified as

$$Y_i = \mu(G_i) + \epsilon_i. \tag{2.14}$$

If sampling unrelated individuals, then the Y_i are independent and identically distributed and this is a simple mixture estimation problem, which can be addressed by EM (see for example, Redner and Walker (1984)). Of greater interest, in the context of genetic analysis, are data observed for members of a pedigree structure. To implement an EM algorithm, we would need to estimate the conditional probabilities that each member of the pedigree is of each of the three genotypes, given current parameter values and the data $\mathbf{Y}$. For related individuals, estimation of the conditional probabilities of genotypes, $\mathbf{G}$, given the observed phenotypic data $\mathbf{Y}$ on the pedigree, is a complex computation equivalent to computation of the total likelihood $\Pr(\mathbf{Y})$. We return to this problem in section 7.4.

Estimation for a genetically more complex model turns out to be simpler, statistically. We consider briefly the classical *polygenic model*, where discrete genotypes G_i are replaced by Gaussian random effects Z_i, known as *polygenic values*. Rather than a single-locus trait, we are now considering a phenotype such as height, probably influenced by a very large number of genes throughout the genome. The genotype configuration $\mathbf{G}$ becomes a vector of polygenic values $\mathbf{z}$, and sums become integrals. The founder probabilities $\Pr(G_i)$ of equation (1.4) are replaced by $N(0, \sigma_a^2)$ population densities for Z_i, where the parameter σ_a^2 is known as the *additive genetic variance*. The transmission probabilities $\Pr(G_i|G_{M_i}, G_{F_i})$

(equation (1.4)) become a transmission density for Z_i given $Z_{M_i} = z_{M_i}$ and $Z_{F_i} = z_{F_i}$:

$$Z_i = \frac{1}{2}(z_{M_i} + z_{F_i}) + \eta_i \tag{2.15}$$

where the η_i are independent, identically distributed segregation residuals, $\eta_i \sim N(0, v_\eta)$, independent of Z_{M_i} and Z_{F_i}. If Z_{M_i} and Z_{F_i} are uncorrelated, then

$$\text{var}(Z_i) = (1/4)(\text{var}(Z_{M_i}) + \text{var}(Z_{F_i})) + v_\eta$$

so to maintain constant population variance σ_a^2 of the Z_i over the generations $v_\eta = \sigma_a^2/2$. The transmission equation (2.15) for the offspring value Z_i, given the parental values, may then be rewritten as $Z_i \sim N((z_{M_i} + z_{F_i})/2, \sigma_a^2/2)$. The joint probability of $\mathbf{Z}$ is Gaussian, with mean $\mathbf{0}$ and variance-covariance matrix $\sigma_a^2\mathbf{A}$, where $\mathbf{A}$ is a matrix determined by the pedigree structure, and known as the numerator-relationship-matrix (Henderson, 1976). In fact, $\mathbf{A}$ is the matrix $2\mathbf{\Psi}$, where the (i,k) component $\Psi_{i,k}$ is the coefficient of kinship $\psi(i,k)$ between individuals i and k (section 3.2).

The simplest penetrance model for a quantitative phenotypic value Y_i of individual i is that it is a direct reflection of the polygenic value Z_i. Ignoring all other possible fixed/random effects, the penetrances $\Pr(Y_i \mid G_i)$ become the density $Y_i \sim N(z_i, \sigma_e^2)$, given $Z_i = z_i$, or

$$Y_i = Z_i + \epsilon_i. \tag{2.16}$$

The variance σ_e^2 of the independent, identically distributed residuals ϵ_i is known as the residual or (individual) environmental variance. In this simplest version of the model, there are just two parameters, σ_e^2 and σ_a^2. In a pedigree (or a collection of pedigrees), suppose there are a total of n_{tot} individuals, and that for n_{obs} of them a value of the quantitative phenotype is observed. The complete-data log-likelihood is

$$\begin{aligned}\log\Pr(\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z}) &= \log\Pr(\mathbf{Y}=\mathbf{y} \mid \mathbf{Z}=\mathbf{z}) + \log\Pr(\mathbf{Z}=\mathbf{z}) \\ &= -\frac{1}{2}\left(n_{obs}\log(2\pi\sigma_e^2) + (\mathbf{y}-\mathbf{z})'(\mathbf{y}-\mathbf{z})/\sigma_e^2\right. \\ &\quad \left. + n_{tot}\log(2\pi\sigma_a^2) + \log(|\mathbf{A}|) + \mathbf{z}'\mathbf{A}^{-1}\mathbf{z}/\sigma_a^2\right).\end{aligned}$$

This is again of exponential family form, with two complete-data sufficient statistics $(\mathbf{y}-\mathbf{z})'(\mathbf{y}-\mathbf{z})$ and $\mathbf{z}'\mathbf{A}^{-1}\mathbf{z}$, which leads to EM equations

$$\begin{aligned}\sigma_e^{2*} &= \mathrm{E}_{\sigma_e^2,\sigma_a^2}((\mathbf{Y}-\mathbf{Z})'(\mathbf{Y}-\mathbf{Z}) \mid \mathbf{Y}=\mathbf{y})/n_{obs} \\ \sigma_a^{2*} &= \mathrm{E}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z} \mid \mathbf{Y}=\mathbf{y})/n_{tot}.\end{aligned} \tag{2.17}$$

If $\mathrm{E}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z} \mid \mathbf{Y}=\mathbf{y}) = \mathbf{a}$ and $\mathrm{Var}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z} \mid \mathbf{Y}=\mathbf{y}) = \mathbf{V}$, the equations reduce to

$$\begin{aligned}\sigma_e^{2*} &= (n_{obs})^{-1}((\mathbf{y}-\mathbf{a})'(\mathbf{y}-\mathbf{a}) + tr(\mathbf{V})) \\ \sigma_a^{2*} &= (n_{tot})^{-1}(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + tr(\mathbf{V}\mathbf{A}^{-1})).\end{aligned}$$

We do not pursue this further here. There is a large literature on the use of EM in polygenic models, particularly in plant and animal breeding. For additional details in the context of simple models on complex pedigrees, see Thompson and Shaw (1990; 1992). For more general work in this area, see the references therein. The point of this example is to show that, even in an exponential family of full rank, the natural sufficient statistics may not be linear functions of latent genotypic counts or values. Estimation of $\mathbf{a} = \mathrm{E}_{\sigma_e^2,\sigma_a^2}(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y})$ is straightforward but insufficient. Since the sufficient statistics are quadratic functions of $\mathbf{Z}$, the conditional variances $\mathbf{V}$ are also needed to implement the EM equations.

Chapter 3

Gene Identity by Descent

3.1 Kinship and inbreeding coefficients

A *gene*, as opposed to an allele or a locus, is the *DNA segment* that is copied from parents to offspring. Underlying the patterns of phenotypes observed on related individuals are the *genotypes*, but underlying the genotypes are the patterns of gene identity by descent. Phenotypes of relatives are similar because they have similar genotypes and may share a common environment. Genotypes are similar because relatives share genes that are identical by descent (*ibd*) — identical copies of a gene segregating from a common ancestor within the defined pedigree. Although for some microsatellite DNA markers mutation rates are non-negligible (section 1.1), for simplicity we disregard mutation throughout this book. In this case, genes that are *ibd* must be of the same allelic type, while genes that are not *ibd* are of independent allelic types.

Gene identity by descent is defined only within the context of a given pedigree structure. A pedigree specifies the two parents of every non-founder individual. A founder has neither parent specified, and by definition the genes in founders are not *ibd*. It will often be convenient if a pedigree is ordered in such a way that every individual is preceded in the listing by his parents; clearly, this is always possible.

Mendel's First Law (section 1.2) states that:

> a diploid individual receives at any given locus a copy of a randomly chosen one of the two genes in his father and (independently) a copy of a randomly chosen one of the two genes in his mother, and will pass on a copy of a randomly and independently chosen one of these two genes to each of his offspring.

This simple law leads to complex patterns of gene identity on an extended pedigree, due to the huge number of alternative events; 2^m for m meioses, at each locus. The segregating genes determine the patterns of gene identity by descent on the pedigree, and hence the patterns of similarity among relatives.

We start with coefficients of *inbreeding* and *kinship*, since these provide an introduction to the ideas of gene identity by descent, to alternative computational

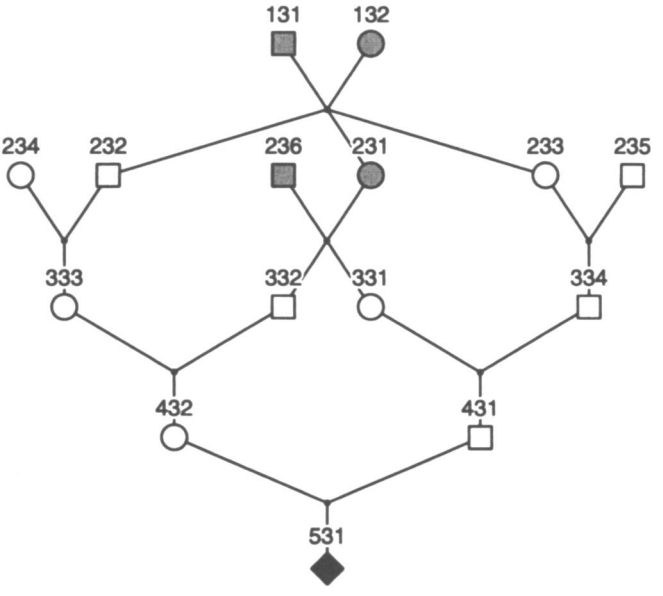

FIGURE 3.1. *An example pedigree. The structure is the same as that of Figure 1.1 of section 1.3. The four individuals shaded grey are bilateral ancestors of the final individual*

approaches, and to Monte Carlo estimation of expectations. Kinship and inbreeding are best thought of as relationships between gametes rather than between individuals. The coefficient of kinship between two individuals B and C, $\psi(B,C)$, is the probability that homologous genes on gametes segregating from B and from C are *ibd*, while the inbreeding coefficient of an individual B, f_B, is the probability that homologous genes on the two gametes uniting to form individual B are *ibd*. Hence

$$f_B = \psi(M_B, F_B)$$

where M_B and F_B are the parents of B. An individual is inbred if his parents are related. He is *autozygous* at a given locus if, at that locus, his two genes are *ibd*. His inbreeding coefficient is the *prior* probability of this event: that is, the probability based only on the pedigree structure.

3.2 Methods of computation

There are several methods for computing kinship and inbreeding coefficients. The early approach of *path-counting* (Wright, 1922) simply enumerates all the possibilities in an efficient way. In order for the two genes within an individual

B to be *ibd*, they must descend from a common ancestor A of his parents. The probability that genes segregating from A in two distinct meioses are *ibd* is 1 if A has two *ibd* genes and $1/2$ otherwise, or overall $f_A.1 + (1 - f_A).(1/2) = (1/2)(1 + f_A)$. If these two genes from A to two distinct offspring are *ibd*, then the probability the same genes descend to B gains a factor of $1/2$ at each successive meiosis. A *path*, $\mathcal{P}_A$, is defined as a sequence of individuals from B ascending to a common ancestor A of his two parents, and descending to B again via a disjoint sequence of individuals. Each such path contributes a term $2^{-(m_M+m_F+1)}(1 + f_A)$ to the inbreeding coefficient f_B, where m_M and m_F are the number of meioses in the path from A to B's mother M and father F respectively. (One may count the two meioses from M and F to B, or the two meioses from A to his two offspring, but not both.) Now, at a single locus, the genes of B can be *ibd* via at most one such path; the paths provide a set of mutually exclusive and exhaustive events leading to B having two *ibd* genes. Thus the inbreeding coefficient of B is

$$f_B = \sum_A \sum_{\mathcal{P}_A} 2^{-(m_M(\mathcal{P}_A)+m_F(\mathcal{P}_A)+1)}(1 + f_A). \tag{3.1}$$

For example, for the offspring of a first cousin marriage, there are 2 paths, one via each of the two grandparents shared by his parents, each having $m_M = m_F = 2$, providing an inbreeding coefficient of $2 \times 2^{-5} = 1/16$. As a more complex example, consider again the pedigree of Figure 1.1 in section 1.3. The pedigree is shown again in Figure 3.1, with the common ancestors of the parents of the final individual shaded grey. The final individual is the offspring of a first cousin marriage, but so also is each of his parents. Here there are two paths via his great-grandparents, each having $m_M = m_F = 2$ as for the simple cousin marriage, and 3 paths via each of his parents' two shared great-grandparents, each with $m_M = m_F = 3$, providing a total inbreeding coefficient of $2 \times 2^{-5} + 2 \times 3 \times 2^{-7} = 7/64$.

Although the path-counting method is the simplest for small pedigrees, it becomes impractical on very large and complex pedigrees. For example, in a segment of a Hutterite pedigree considered by Thompson and Morgan (1989), there are over 1000 ancestral paths connecting the two parents of one individual. Other approaches to computation of inbreeding and kinship follow from equations based on the properties of Mendelian segregation. We use the meiosis indicators introduced in section 1.2 and consider the kinship coefficient $\psi(B, C)$ between two individuals B and C. Provided B is not an ancestor of C, we may condition on the segregation S from B, where

$$\Pr(S = 0) = \Pr(S = 1) = \frac{1}{2}.$$

If $S = 0$, the segregating gene is B's maternal gene; that is, a gene from the mother of B. If $S = 1$, the gene is B's paternal gene. Thus we obtain immediately

$$\begin{aligned} \psi(B, C) &= \psi(M_B, C)P(S = 0) + \psi(F_B, C)P(S = 1) \\ &= (\psi(M_B, C) + \psi(F_B, C))/2 \end{aligned} \tag{3.2}$$

where M_B and F_B are the mother and the father of B. Also, from the definition, we have symmetry: $\psi(B, C) = \psi(C, B)$. Thus the only additional equation needed

is for the case $B = C$. In this case, we must consider two independent segregations from B, S_1 and S_2:

$$\Pr(S_1 = S_2) \;=\; \Pr(S_1 \neq S_2) \;=\; \frac{1}{2}.$$

If $S_1 = S_2$, the segregating genes are *ibd*. If $S_1 \neq S_2$, the genes comprise both the maternal and paternal genes of B. Thus

$$\begin{aligned} \psi(B,B) &= P(S_1 = S_2) \;+\; \psi(M_B, F_B)P(S_1 \neq S_2) \\ &= (1 \;+\; \psi(M_B, F_B))/2. \end{aligned}$$

Together with the boundary conditions

$$\begin{aligned} \psi(B,B) &= \tfrac{1}{2} \quad \text{for any founder } B, \\ \text{and } \psi(B,C) &= 0 \quad \text{if } B \text{ is a founder not an ancestor of } C, \end{aligned}$$

these equations determine the function $\psi(\cdot)$ on the pedigree.

A recursive algorithm based on these equations is very easily implemented, and works well even on large and complex pedigrees. However, it is not necessarily computationally efficient; the same expansion may be repeated many times. In principle, this can be avoided, by saving $\psi(B,\ C)$, for key pairs of individuals (B, C) in the ancestry of the pedigree, but the simplicity of the method is then lost. An alternative way to implement these equations is via a top-down sequential method, computing kinship coefficients between all pairs of ancestors arriving finally at the descendant individuals of interest. This is computationally trivial, but expensive on store. All computation is a trade-off between time and store.

3.3 Data on inbred individuals

Kinship and inbreeding coefficients measure only *ibd* between two gametes, at a single locus. However, this suffices for a consideration of data on unrelated inbred individuals. At a single locus, with alleles $A_1, \ldots, A_k$, having population frequencies $q_1, \ldots, q_k$, an individual having two *ibd* genes has genotype A_jA_j with probability q_j, while an individual who is not autozygous at this locus has genotype probabilities of Hardy-Weinberg form (section 2.3). Thus an individual who has inbreeding coefficient f has genotype probabilities

$$\begin{aligned} \Pr(A_jA_j) &= q_jf \;+\; q_j^2(1-f) \\ &= q_j(q_j \;+\; f(1-q_j)), \quad j = 1, \ldots, k \\ \Pr(A_jA_l) &= 2(1-f)q_jq_l, \quad 1 \leq j < l \leq k. \end{aligned} \tag{3.3}$$

Since an individual who is autozygous at a particular locus must be homozygous at that locus, inbreeding is of particular interest in the study of rare recessive traits. If the recessive allele has frequency q, the probability that an individual with inbreeding coefficient f is affected is $q(q+f(1-q))$. If the population consists

of a proportion α_i of individuals with inbreeding coefficient f_i, then the overall proportion of affected individuals is

$$\sum_i \alpha_i(q(q + f_i(1 - q))) = q(q + f(1 - q))$$

where $f = \sum_i \alpha_i f_i$ is the mean inbreeding coefficient in the population, or the expected inbreeding coefficient of an individual randomly chosen from the population. The conditional probability that an affected individual derives from the group with inbreeding coefficient f_i is

$$\frac{\alpha_i(q + f_i(1 - q))}{q + f(1 - q)}.$$

The probability that an affected individual with inbreeding coefficient f_i is autozygous at this locus is

$$\frac{f_i}{q + f_i(1 - q)}$$

while the overall probability an affected individual is autozygous at this locus is

$$\frac{f}{q + f(1 - q)}. \tag{3.4}$$

Note that for a very rare recessive trait ($q \approx 0$), a high proportion of the affected individuals will have non-zero inbreeding coefficients. Indeed, the groups i then contribute to the affected individuals in the same proportions $\alpha_i f_i / f$ as they contribute to the mean population inbreeding. Moreover, a high proportion of the affected individuals are not only inbred, but in fact autozygous at the locus in question. We return to these probabilities in section 4.6.

In a population in which the mean inbreeding coefficient is f, the genotype frequencies are given by equation (3.3). There are two points to note about this homozygote excess and heterozygote deficiency, relative to Hardy-Weinberg proportions. The first is that these are frequencies in an infinite population. In a finite population, individuals of necessity marry their relatives, and allele frequencies change over time. Whether or not there is a homozygote excess, relative to Hardy-Weinberg proportions with the current allele frequencies, depends on whether individuals are, on average, marrying an individual who is more or less closely related to them than is a randomly chosen member of the population. Second, the homozygote excess due to inbreeding is a particular special case of the homozygote excess due to subdivision of a population; inbreeding is a form of subdivision. However, under the inbreeding scenario, there is no differentiation among alleles. Under subdivision, different alleles may show differing patterns of variation in frequency among subdivisions. This leads to genotype frequencies in which each homozygote shows an excess frequency, but in an amount dependent on the variation of the frequency of that allele among subdivisions. Although in total there is a heterozygote deficiency, patterns of covariation of allele frequency may lead to increased frequencies of some heterozygote genotypes (Weir, 1996).

As an additional example of the use of the EM algorithm (section 2.4) to estimate parameters underlying genotype frequencies, we consider estimation of f under the model of equation (3.3). Suppose that a random sample of individuals is taken from the population, and that there are n_{jl} individuals of genotype A_jA_l for $j \leq l$. Then the likelihood for the parameters $\mathbf{q} = (q_1, \ldots, q_k)$ and f is

$$\begin{aligned} L(\mathbf{q}, f) &= P_{\mathbf{q},f}(\{n_{jl}\}) \\ &\propto \prod_j (q_j(q_j + f(1-q_j)))^{n_{jj}} \prod_{j<l} (2q_jq_l(1-f))^{n_{jl}}. \end{aligned}$$

Clearly, this is not an easy expression to maximize.

Let X_j be the number of homozygous A_jA_j individuals in the sample who have two identical-by-descent (*ibd*) genes at this locus. With X_j as the latent variables, the complete-data likelihood is

$$\begin{aligned} L^*(\mathbf{q}, f) &= P_{\mathbf{q},f}(\{n_{jl}\}, \{X_j\}) \\ &= \prod_j q_j^{2n_{jj}-X_j} \prod_{j<l} (2q_jq_l)^{n_{jl}} f^T (1-f)^{n-T} \end{aligned}$$

where $T = \sum_j X_j$. Let $m_j = 2n_{jj} + \sum_{l<j} n_{lj} + \sum_{l>j} n_{jl}$ be the number of A_j alleles observed in the sample. Then the complete-data log-likelihood reduces to

$$\begin{aligned} \ell^*(\mathbf{q}, f) &= \log P_{\mathbf{q},f}(\{n_{jl}\}, \{X_j\})) \\ (3.5) \qquad &= \text{const} + \sum_j (m_j - X_j) \log q_j + T \log f + (n-T) \log(1-f). \end{aligned}$$

The complete-data log-likelihood (3.5) is thus linear in the functions of the latent variables X_j and T. Computation of the expected complete-data log-likelihood requires only

$$\mathrm{E}_{\mathbf{q},f}(X_j \mid \{n_{jl}\}) = \frac{f n_{jj}}{f + q_j(1-f)}$$

using equation (3.4). Moreover, if X_j were observed, the MLEs based on (3.5) would be $\widehat{f} = T/n$ and $\widehat{q_j} = (m_j - x_j)/\sum_l (m_l - x_l)$. An EM algorithm for this problem is thus to iterate:

$$\begin{aligned} \mathrm{E-step}: \quad x_j &= f n_{jj}/(f + q_j(1-f)), \quad t = \textstyle\sum_j x_j \\ \mathrm{M-step}: \quad q_j &= (m_j - x_j)/\textstyle\sum_l (m_l - x_l), \quad f = t/n. \end{aligned}$$

As in the examples of section 2.5, the algorithm is easily implemented, and converges quickly.

3.4 Multi-gamete kinship and gene *ibd*

Kinship and inbreeding provide results only concerning a pair of genes, and thus a single genotype. Analysis even of data on a pair of related individuals, at a single

locus, requires consideration of four genes. An important extension to section 3.2 was made by Karigl (1981), who considered the probability of simultaneous identity by descent, $\psi(B_1, ..., B_m)$, of m genes segregating from a set of (not necessarily distinct) individuals $B_1, B_2, ..., B_m$. As in equation (3.2), if B_1 is not an ancestor of any of $B_2, ..., B_m$, conditioning on the segregation from B_1 gives

$$\psi(B_1, B_2, ..., B_m) \;=\; \frac{1}{2}\Big(\psi(M_{B_1}, B_2, ..., B_m) \;+\; \psi(F_{B_1}, B_2, ..., B_m)\Big) \tag{3.6}$$

where M_{B_1} and F_{B_1} are the parents of individual B_1. The symmetry of the definition provides that we may collect the arguments for some B_1 who is not an ancestor of any of the others to the first v arguments of ψ. Then, considering the v independent segregations from B_1, either the segregating gene is the same in every case, being a random gene from B_1, or both the maternal and the paternal genes of B_1 are among the v genes. Since

$$\Pr(S_1 = S_2 = ... = S_t) \;=\; 2^{-v+1},$$

we obtain

$$\begin{aligned}\psi(B_1, ..., B_1, B_2, ..., B_m) \;&=\; 2^{-v+1}\Big(\psi(B_1, B_2, ..., B_m) \;+ \\ &\qquad (2^{v-1} - 1)\;\psi(M_{B_1}, F_{B_1}, B_2, ..., B_m)\Big) \\ &=\; 2^{-v}\Big(\psi(M_{B_1}, B_2, ..., B_m) \;+\; \psi(F_{B_1}, B_2, ..., B_m) \\ &\qquad +\; (2^v - 2)\;\psi(M_{B_1}, F_{B_1}, B_2, ..., B_m)\Big). \end{aligned} \tag{3.7}$$

Together with symmetry and boundary conditions, these equations determine the multiple kinship coefficients on any pedigree. Note that the number of arguments of ψ is never increased by recursion, although the number of terms may be doubled at each step. Practical implementation can therefore be problematic on a large multi-generation pedigree if the initial number m of genes or individuals considered is more than about 7.

The m-gamete kinship coefficients can be used to determine probabilities of patterns of gene *ibd* among a set of m genes. First, however, a specification of such patterns (gene *ibd* states) is needed. Among a set of genes in given individuals, a gene *ibd* state is a partition of the genes into subsets that are *ibd*. We denote such a pattern by **J**, and refer to it as the *pattern* of gene identity by descent among the individuals. A partition of m ordered genes may be specified by a set of m integers as follows. Let $k_1 = 1$. Suppose genes $1, 2, \ldots, r$ have been assigned v distinct labels $k_1, \ldots, k_r$. If gene $r + 1$ is *ibd* to a previous gene l, $k_{r+1} = k_l$. Otherwise, $k_{r+1} = v + 1$. (For the case $m = 4$, this labeling is shown in Table 3.1.) As m increases, the number of possible states of gene *ibd* increases rapidly. For the 12 genes of 6 individuals, there are more than 4 million gene identity states (partitions of 12 ordered objects). However, for the analysis of phenotypic data on individuals, one need not distinguish the paternal and maternal genes of an individual. The interchange of labels on the two genes within each member of any

subset of the individuals groups the *ibd* states into *genotypically distinct* classes of states. For the case of two individuals, this grouping is also shown in Table 3.1. This grouping substantially decreases the number of patterns of gene *ibd* that must be considered. For example, for six individuals there are only just over 198,000 genotypically distinct classes of states (Thompson, 1974). Although this is not a small number, with modern computers and an efficient indexing of state classes it is not impossible to consider all the possible state classes given data on 6 individuals.

Returning to the relationship between multi-gamete kinship and gene *ibd* state probabilities, consider any specified (detailed or grouped) *ibd* state among the genes of a set of individuals. For example, for five individuals $(B_1, B_2, B_3, B_4, B_5)$ the state $(1,2,1,3,4,4,2,4,2,5)$. This state contributes 0.25 to $\psi(B_1, B_2)$, 0.5 to $\psi(B_3, B_4)$ and 0.125 to $\psi(B_1, B_4, B_5)$. Conversely, any multi-gamete kinship coefficient among individuals, say $\psi(B_1, \ldots, B_m)$ can be written as a weighted sum of *ibd* state probabilities:

$$\psi(B_1, \ldots, B_m) = \sum_{\mathbf{J}} \Pr(\text{segregating genes } ibd \mid \mathbf{J}) \Pr(\mathbf{J}).$$

If multi-gamete kinship coefficients are computed for all subsets of the individuals of interest, the linear equations may be inverted to give the *ibd* state probabilities, $\Pr(\mathbf{J})$ among the genes of the individuals. Karigl (1981) was interested primarily in the determination of the probabilities of patterns of *ibd* among the four genes of two individuals, at a single genetic locus. He gives details of the equations for this case.

3.5 Patterns of gene *ibd* in pairs of individuals

Among the four genes of two individuals at a single autosomal locus, there are 15 states of gene identity (Cotterman, 1974). These are shown in Table 3.1, and correspond simply to the number of partitions of the four genes into classes of genes that are *ibd*. However, there are only 9 groups of genotypically distinct classes of states, since with regard to genotypes the maternal and paternal origins of genes are irrelevant, so the identities of the two genes within each individual can be interchanged. For the case of two individuals, the state classes can be characterized by specifying the autozygous individual(s), and the number of genes shared *ibd* between the two individuals (Table 3.1).

For two non-inbred diploid individuals, there are only three possible genotypically distinct *gene identity states* at a single autosomal locus. That is, the individuals can share neither of their genes *ibd*, or one, or both. These events have probabilities $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2)$ say, $(\kappa_0 + \kappa_1 + \kappa_2 = 1)$, determined by the pedigree. Individuals are related if $\kappa_0 < 1$. Each relationship may thus be represented by a point in an equilateral triangle of unit height, the vertices corresponding to unrelated pairs $(\kappa_0 = 1)$, parent-offspring $(\kappa_1 = 1)$, and the identity (monozygous twin) relationship $(\kappa_2 = 1)$. (Care should be taken in applying the standard equations to monozygous twins, since they result from a single maternal and a single paternal meiosis.) The triangle representation is shown in Figure 3.2 and

ibd pattern B_1 p m	*ibd* pattern B_2 p m	*ibd* label	*ibd* group	state description individuals autozygous	state description genes shared
•	• •	1 1 1 1	1 1 1 1	B_1, B_2	4 genes *ibd*
•	• ○	1 1 1 2	1 1 1 2	B_1	3 genes *ibd*
•	○ •	1 1 2 1			
•	○ •	1 2 1 1	1 2 1 1	B_2	3 genes *ibd*
•	○ ○	1 2 2 2			
•	○ ○	1 1 2 2	1 1 2 2	B_1, B_2	none
•	○ †	1 1 2 3	1 1 2 3	B_1	none
•	† †	1 2 3 3	1 2 3 3	B_2	none
•	• ○	1 2 1 2	1 2 1 2	none	2 genes
•	○ •	1 2 2 1			shared
•	• †	1 2 1 3	1 2 1 3	none	1 gene
•	† •	1 2 3 1			shared
•	○ †	1 2 2 3			
•	† ○	1 2 3 2			
•	† ⋆	1 2 3 4	1 2 3 4	none	none

TABLE 3.1. *States of gene ibd among the four genes of two individuals*

the values of $\boldsymbol{\kappa}$ for some standard relationships are give in Table 3.2. The kinship coefficient is the probability that homologous genes segregating from two individuals are identical by descent and thus $\psi = (2\,\kappa_2 + \kappa_1)/4$. Lines of constant kinship are orthogonal to the line $\kappa_1 = 0$. Sibs, with $\boldsymbol{\kappa} = (1/4, 1/2, 1/4)$ have the same kinship coefficient as a parent-offspring relationship. Half-sibs, with $\boldsymbol{\kappa} = (1/2, 1/2, 0)$ have the same kinship coefficient as double-first-cousins ($\boldsymbol{\kappa} = (9/16, 3/8, 1/16)$).

Pairwise relationship	κ_0	κ_1	κ_2	ψ
Unrelated	1.00	0	0	0
Parent-offspring	0	1.00	0	0.25
Monozygous twin	0	0	1.00	0.50
Full Sib	0.25	0.50	0.25	0.25
Half sib, grandparent, aunt	0.50	0.50	0.00	0.125
First cousin	0.75	0.25	0	0.0625
Double first cousin	0.5625	0.375	0.0625	0.125
Quadruple half first cousin	0.5312	0.4375	0.0312	0.125

TABLE 3.2. *Values of $\boldsymbol{\kappa}$, and kinship coefficient ψ, for some standard relationships between two non-inbred individuals*

While each relationship determines a point $\boldsymbol{\kappa}$, the converse is not true. Several relationships give the same probabilities $\boldsymbol{\kappa}$; the simplest example is the three pairwise relationships grandparent-grandchild, half-sibs, and aunt-niece, all of which have $\boldsymbol{\kappa} = (1/2, 1/2, 0)$. Moreover, some points in the triangle are not (even in

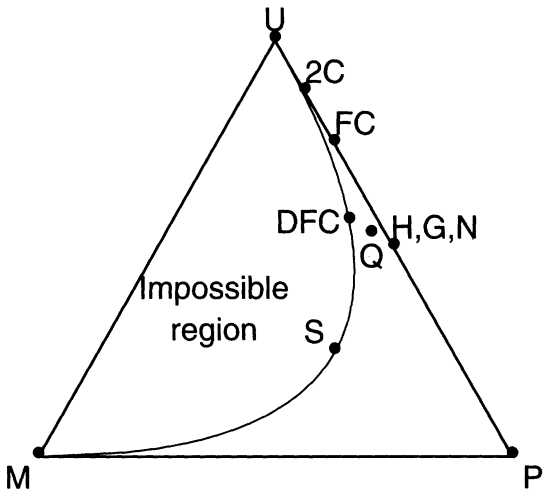

FIGURE 3.2. *The relationship triangle for non-inbred relatives*

the limit) attainable by any relationship. In fact, it can be shown that $\kappa_1^2 \geq 4\kappa_0\kappa_2$ (Thompson, 1986). This result follows from the fact that, for non-inbred individuals

$$\begin{aligned} \psi &= (1/4)(\psi_{MM} + \psi_{FF} + \psi_{MF} + \psi_{FM}) \\ \text{and } \kappa_2 &= (\psi_{MM}\psi_{FF} + \psi_{MF}\psi_{FM}) \end{aligned} \tag{3.8}$$

where the subscripted kinship coefficients are those between a parent (mother (M) or father (F)) of one individual, and a parent of the other. Then the arithmetic-geometric mean inequality gives

$$\begin{aligned} 4\kappa_2 &\leq (\psi_{MM} + \psi_{FF})^2 + (\psi_{MF} + \psi_{FM})^2 \\ &\leq (\psi_{MM} + \psi_{FF} + \psi_{MF} + \psi_{FM})^2 \\ &= (4\psi)^2 = (\kappa_1 + 2\kappa_2)^2 \\ &= \kappa_1^2 + 4\kappa_2(\kappa_1 + \kappa_2) \quad \text{or} \\ 4\kappa_2\kappa_0 &= 4\kappa_2(1 - (\kappa_1 + \kappa_2)) \leq \kappa_1^2. \end{aligned}$$

In order for equality to hold in this inequality, one pair of the crossparental kinship coefficients must be 0, and the other pair equal. Such relationships include full sibs ($\psi_{MM} = \psi_{FF} = 1/4$, $\psi_{MF} = \psi_{FM} = 0$) and double-cousins of any degree v, for which $\psi_{MM} = \psi_{FF} = (1/2)^{v+2}$, $\psi_{MF} = \psi_{FM} = 0$ or $\psi_{MF} = \psi_{FM} = (1/2)^{v+2}$, $\psi_{MM} = \psi_{FF} = 0$. These relationships give values of $\boldsymbol{\kappa}$ falling on the boundary parabola.

It is possible for the mother and father of each individual to be related to both the mother and the father of the other, without either individual being inbred. That is, all four of the cross-parental kinship coefficients in the above equation may

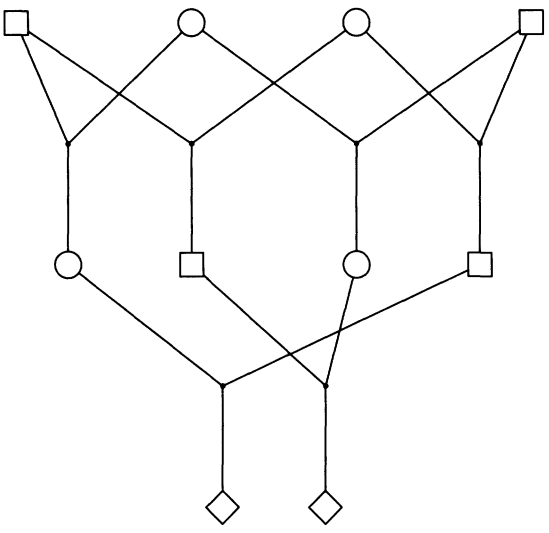

FIGURE 3.3. *The relationship of quadruple-half-first-cousins*

be non-zero. The simplest example is that of quadruple-half-first-cousins, shown in Figure 3.3. For this relationship, the mother and the father of each individual is a half-sib of both the mother and the father of the other, so $\psi_{MM} = \psi_{FF} = \psi_{MF} = \psi_{FM} = (1/8)$. Hence, using equation (3.8), $\kappa_2 = 1/32$, $\kappa_1 = 7/16$, $\kappa_0 = 17/32$ and $\psi = 1/8$. The point in the triangle lies midway between that for half-sibs and for double-first-cousins, which also each have $\psi = 1/8$.

More details of the material of this section, and references to earlier work, can be found in Chapter 2 of Thompson (1986).

3.6 Observations on related individuals

Phenotypic similarities among relatives result from the genes they share *ibd*. Among an ordered set of genes, a partition of the set may be used to specify which subsets of the genes are *ibd* (section 3.4). Again we denote such a pattern of gene *ibd* by **J**. In section 1.3, the meiosis indicators were defined (equations (1.2) and (1.3)), and it was seen how the meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, and patterns of gene identity by descent, at any given locus j. Thus, the passage of genes in pedigrees provides the connection between observable genetic characteristics and the pedigree structure, whether we are estimating relationships from genetic data, estimating the genetic basis of traits knowing the pedigree, or inferring the ancestry and descent of particular genes, knowing both the genetic model and the data (section 1.4).

In particular, we consider a currently observed set of individuals, and the pattern,

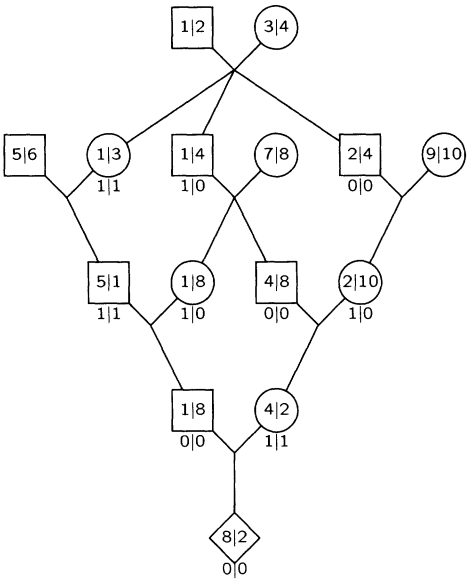

FIGURE 3.4. *Meiosis indicators $S_{\bullet,j}$ determine descent of founder genes, and patterns of gene identity by descent, at any given locus j: see Figure 1.2*

J, of genes *ibd* among them, at a single locus. We therefore drop the locus index j, and write $\mathbf{S} = \{S_i; i = 1, ..., m\}$ (equation (1.1)), for the m meioses of the pedigree. The example of Figure 1.2 is shown again in Figure 3.4. The meiosis indicators shown under each individual are for the paternal and for the maternal meiosis to that individual, respectively. Then **S** determines the pattern, **J**, of genes *ibd* in this currently observed set of individuals; $\mathbf{J} = \mathbf{J}(\mathbf{S})$. The probability of any phenotypic data **Y** (i.e. observed characteristics of the individuals) depends on **S** only through $\mathbf{J}(\mathbf{S})$, and so

$$\begin{aligned} \Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{J}} \Pr(\mathbf{Y} \mid \mathbf{J}) \Pr(\mathbf{J}). \end{aligned} \tag{3.9}$$

Equation (3.9) may be compared with equation (1.5) of Chapter 1. In equation (1.5) the latent variables were the genotypes G_i of individuals, whereas here they are the meiosis indicators. In both cases, the form of the likelihood is that of a latent variable problem (section 2.4), and either may be the more convenient for

likelihood computation and inference (Chapter 6).

In partitioning the likelihood as in equation (3.9), the "genetic model" is separated from the effects of genealogical and genetic structure. The probability of a set of meiosis indicators $\mathbf{S}$ at a single locus is trivial; the components are independent, each 0 or 1 with probability 1/2. The probability of a given pattern $\mathbf{J}(\mathbf{S})$ depends on the genealogical relationship among the observed individuals: in principle it may be computed by the methods of sections 3.4 or 3.8. Given the gene identity pattern, $\mathbf{J}(\mathbf{S})$, the probability of data depends on the different types of genes, their frequencies, and how they affect observable phenotypes.

Now consider the probability $\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))$, for a specified pattern of gene *ibd* among the observed individuals. The probability any distinct gene, k, is of allelic type $a(k)$ is the population frequency, $q_{a(k)}$, of the allele. Distinct genes k have independent allelic types. Thus, $\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))$ is the sum over all possible assignments $\mathcal{A}$ of allelic types to genes of the product of allele frequencies $q_{a(k)}$ of assigned alleles $a(k)$:

$$\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) = \sum_{\mathcal{A}} \prod_{k} q_{a(k)}. \tag{3.10}$$

This equation was given by Thompson (1974) who gave an example of *ABO* blood types on three individuals. The special case of two individuals (9 states $\mathbf{J}$) is discussed in Chapter 2 of Thompson (1986).

In general, efficient determination of all allocations $\mathcal{A}(j)$ at locus j compatible with data Y_{*j} is straightforward for genotypic data (for example, DNA marker phenotypes). An algorithm for this determination of is given by Kruglyak et al. (1996). The implementation we use is due to Simon Heath (personal communication) and is described in more detail by Thompson and Heath (1999). We use the same example pedigree, with the values of $S_{\bullet,j}$ given in Figure 1.2, and assume five individuals observed with the genotypes shown in Figure 3.5(a). The method rests first on the fact that only founder genes having copies in observed individuals are constrained in allelic type: in our example, the genes labeled $\{1, 2, 4, 5, 8, 10\}$. Further two genes constrain each other's allelic type only when both are present in an observed individual. The *gene graph* (Figure 3.5(b)) connects all such pairs of genes. Allocation of allelic types may be considered separately for each component subgraph of connected genes. In our example, the genes $\{1, 5\}$ may be considered independently of $\{2, 4, 8, 10\}$. This assignment is readily accomplished, even on a much larger example. For given $S_{\bullet,j}$ there are in general only 2, 1 or 0 possible assignments of allelic types to the genes of a component subgraph. For our example, there are two possible assignments for the first component and one for the second: $(a(1), a(5)) = (A, C)$ or (C, A) and $(a(2), a(4), a(8), a(10)) = (C, D, C, B)$. The algorithm can in principle be generalized to more complex phenotypes, using the conditional independence structure of the gene graph (Figure 3.5(b)), but the procedure becomes far more computationally intensive.

For completeness, and as an example of the above general formula, consider again the case of a non-inbred pair of relatives. There are then three *ibd* states J_0, J_1 and

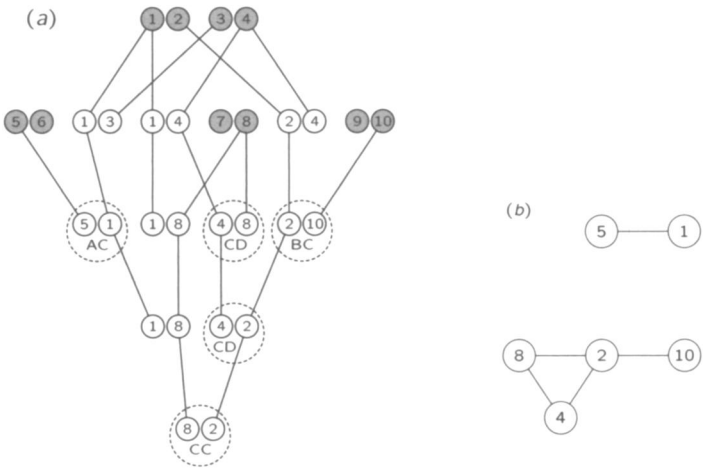

FIGURE 3.5. *Determination of probabilities* $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$. *The gene descent pattern is assumed to be that of Figure 1.2, and the pairs of genes are shown, rather than the individuals. Five individuals, shown as dashed circles, are assumed to be observed, with marker genotypes as indicated: see text for details. (a) Only genes present in observed individuals are constrained in type. (b) Two genes in a single observed individual are jointly constrained*

J_2, with probabilities κ_0, κ_1, and κ_2, these being determined by the relationship R between the individuals (section 3.5). The state J_k denotes that k genes are shared *ibd* between the two individuals. Suppose at a given locus the ordered genotypes of the pair are (G_1, G_2). Then the analogue of equation (3.9) is

$$\begin{aligned}\Pr(G_1, G_2; R) \;=\; & \kappa_0(R)\Pr(G_1, G_2; J_0) \;+\; \kappa_1(R)\Pr(G_1, G_2; J_1) \\ & +\; \kappa_2(R)\Pr(G_1, G_2; J_2).\end{aligned}$$

Now, $\Pr(G_1, G_2; J_2) \;=\; \Pr(G_1)$, the population frequency of the genotype, if $G_1 = G_2$, and 0 otherwise. This is the probability for a pair of monozygous twins. Also, $\Pr(G_1, G_2; J_0) \;=\; \Pr(G_1)\Pr(G_2)$, the probability for an unrelated pair of relatives. Finally, $\Pr(G_1, G_2; J_1)$ is the probability for a parent-offspring pair; these probabilities were given in Table 2.1 (section 2.3). For a pair of relatives, in most cases equation (3.10) take form too trivial to be illuminating. The one non-trivial case is $\Pr(G_1 = A_1A_2, G_2 = A_1A_2; J_1)$. Here the *ibd* gene may be either the A_1 or the A_2 allele; there are two feasible allocations $\mathcal{A}$ of allelic types to the three distinct genes in the two individuals (A_1, A_2, A_1) or (A_1, A_2, A_2) giving a total probability $p_1p_2p_1 + p_1p_2p_2 \;=\; p_1p_2(p_1 + p_2)$ as given in the Table 2.1.

Thus, to obtain the probability of genotypes (and hence of phenotypes) for any pair of non-inbred relatives, it is enough to know the probabilities for monozygous-twin, parent-offspring, and unrelated pairs. For a general pair of relatives, however,

Gene *ibd* state for two sibs with	Prior (pedigree) probability (a) an aunt	(b) a niece or half-sib
Sibs sharing 2 *ibd*		
1 2 1 2 1 3	1/8	1/8
1 2 1 2 3 4	1/8	1/8
Sibs sharing 1 *ibd*		
1 2 1 3 1 4	1/8	1/8
1 2 1 3 2 3	1/16	0
1 2 1 3 2 4	1/16	1/8
1 2 1 3 3 4	1/16	1/8
1 2 1 3 4 5	3/16	1/8
Sibs sharing 0 *ibd*		
1 2 3 4 1 3	1/16	0
1 2 3 4 1 5	1/16	1/8
1 2 3 4 3 5	1/16	1/8
1 2 3 4 5 6	1/16	0

TABLE 3.3. *Gene ibd state probabilities at a single locus for a pair of sisters with an aunt, niece, or half-sib. The states are given in the reduced genotypic state-class form, in which the paternal and maternal genes of the three individuals are not distinguished*

the nine genotypically distinct *ibd* patterns of Table 3.1 are required. The probabilities of the states must be computed (see, for example, Karigl (1981)), and also the probabilities of genotypes under each *ibd* state. Again, the latter are special cases of equation (3.10), and are given by Thompson (1986).

Finally, in this section, note that joint analysis of data on a set of relatives is always more powerful than pairwise analysis. A simple example which derives from an actual study is that of a pair of full sibs and their aunt or niece (Browning and Thompson, 1999). Due to the symmetry of the pairwise aunt-niece relationship, pairwise analysis cannot distinguish these relationships; nor distinguish the possibility that the third individual is a half-sister to the pair of sibs. However, two sibs and a aunt can carry six distinct genes at a locus, but sibs with a niece or half-sib cannot. The probabilities of the *ibd* states among the three individuals at a single locus are shown in Table 3.3. Loci at which the two sibs share both their genes *ibd* give the same probabilities of sharing with the third individual under the three possibilities of aunt, niece or half-sib. However, the other state probabilities differ, with the greatest power to distinguish an aunt from a niece or half-sib coming from those loci at which the full sibs do not share any genes *ibd*. Note that data at unlinked loci remains insufficient to distinguish the possibilities that the third individual is a niece or half-sister. As will be seen in section 4.5, data at linked loci results in identifiability of these two alternatives.

3.7 Monte Carlo estimation of expectations

Although the methods of section 3.4 are easily implemented, where large numbers of individuals are considered jointly computation may become impractical or even infeasible. Where exact probabilities cannot be computed, Monte Carlo estimation is an alternative. We use this section to introduce some important ideas in the Monte Carlo estimation of sums, integrals, or expectations. We shall use these methods to estimate probabilities of gene *ibd* patterns in section 3.8. These methods will be important for Chapters 7 and 8. Since, in this section, the latent variables are general, we use the notation $\mathbf{X}$ instead of $\mathbf{S}$.

To estimate $\sum_{\mathbf{x}} g(\mathbf{x})$ the sum may be written as an expectation

$$\sum_{\mathbf{x}} g(\mathbf{x}) = \sum_{\mathbf{x}} \frac{g(\mathbf{x})}{\Pr(\mathbf{X}=\mathbf{x})}\Pr(\mathbf{X}=\mathbf{x}) = \mathrm{E}_P\left(\frac{g(\mathbf{X})}{P(\mathbf{X})}\right)$$

where $P(\cdot)$ is some distribution over $\mathcal{X}$, the space of values of $\mathbf{X}$. The distribution $P(\cdot)$ must assign positive probability to every value $\mathbf{x}$ of $\mathbf{X}$ for which $g(\mathbf{x}) > 0$. If $X^{(1)}, ..., X^{(N)}$ are simulated from the distribution $P(\cdot)$,

$$\frac{1}{N}\sum_{\tau=1}^{N}\left(\frac{g(\mathbf{X}^{(\tau)})}{P(\mathbf{X}^{(\tau)})}\right) \tag{3.11}$$

is an unbiased estimator of the sum $\sum_{\mathbf{x}} g(\mathbf{x})$. Of course, it may not be a very good estimator; in fact, it may be a very bad estimator. The art of Monte Carlo is finding good distributions to simulate from, and good ways of simulating from them, in order to get good estimators. A "good" estimator is one with small variance. Note this is not the standard statistical paradigm where parameters are estimated from data. In that case, variances are over (hypothetical) repetitions of the experiment or random process giving rise to the data. In Monte Carlo, the relevant variances are Monte Carlo variances.

The simplest form of Monte Carlo is where we simulate independent, identically distributed realizations from some distribution $P(\cdot)$. Note that any sum of terms $g(\mathbf{x})$ is an expectation of $g^*(\mathbf{X}) = g(\mathbf{X})/P(\mathbf{X})$ with respect to the probability distribution $P(\cdot)$. The estimator (3.11) is then an average of terms $g^*(\mathbf{X})$, and, for independent realizations, the Monte Carlo variance of this estimator is

$$N^{-1}\left(\mathrm{E}_P((g^*(\mathbf{X}))^2) - (\mathrm{E}_P(g^*(\mathbf{X})))^2\right)$$

$$\text{or} \quad \left(\sum_{\mathbf{x}}(g^*(\mathbf{x})^2 P(\mathbf{x})) - \left(\sum_{\mathbf{x}} g^*(\mathbf{x})P(\mathbf{x})\right)^2\right)$$

which, substutituting $g^*(\mathbf{x}) = g(\mathbf{x})/P(\mathbf{x})$, is

$$N^{-1}\left(\sum_{\mathbf{x}}\left(\frac{g(\mathbf{x})^2}{P(\mathbf{x})}\right) - \left(\sum_{\mathbf{x}} g(\mathbf{x})\right)^2\right).$$

This may be estimated by the sample variance from the Monte Carlo:

$$(N(N-1))^{-1}\left(\sum_{\tau=1}^{N}\left(g^*(\mathbf{X}^{(\tau)})\right)^2 - N^{-1}\left(\sum_{\tau=1}^{N}\left(g^*(\mathbf{X}^{(\tau)})\right)\right)^2\right)$$

$$\text{or } (N(N-1))^{-1}\left(\sum_{\tau=1}^{N}\left(\frac{g(\mathbf{X}^{(\tau)})}{P(\mathbf{X}^{(\tau)})}\right)^2 - N^{-1}\left(\sum_{\tau=1}^{N}\left(\frac{g(\mathbf{X}^{(\tau)})}{P(\mathbf{X}^{(\tau)})}\right)\right)^2\right).$$

On pedigrees, the simplest distribution to simulate from is the prior distribution on genotypes, which is done by "gene dropping". Genes are assigned to the founders of the pedigree, segregation of genes down the pedigree is simulated, and the required statistics relating to the resultant current genes are computed. Such Monte Carlo estimates have been used by Edwards (1967) to estimate inbreeding coefficients, by MacCluer et al. (1986) to study the loss of genes in pedigrees of endangered species, and by Thompson et al. (1978) to study the potential power of a pedigree study.

Using equation (3.11) is often ineffective. Methods of more effective simulation normally involve some form of *importance sampling*. Note that

$$\begin{aligned} \mathrm{E}_P(g^*(\mathbf{X})) &= \sum_{\mathbf{x}} g^*(\mathbf{x})\ P(\mathbf{x}) \\ &= \sum_{\mathbf{x}} g^*(\mathbf{x})\frac{P(\mathbf{x})}{P^*(\mathbf{x})}\ P^*(\mathbf{x}) \\ &= \mathrm{E}_{P^*}\left(g^*(\mathbf{X})\frac{P(\mathbf{X})}{P^*(\mathbf{X})}\right) \end{aligned} \tag{3.12}$$

provided

$$P^*(\mathbf{X}) > 0 \quad \text{if} \quad g^*(\mathbf{X})P(\mathbf{X}) > 0. \tag{3.13}$$

Thus realizations from $P^*(\cdot)$ can be reweighted in order to estimate expectations under P. Where this is done in such a way that terms making larger contributions to the sum are realized with larger probabilities, this is *importance sampling*. Such sampling decreases the Monte Carlo variance of the estimator of the sum. The effectiveness of this approach depends on the choice of $P^*(\cdot)$. It works best when the summand $g^*(\mathbf{X})\ P(\mathbf{X})$ is the "same shape" as $P^*(\mathbf{X})$, since then the variance of $g^*(\mathbf{X})\ P(\mathbf{X})/P^*(\mathbf{X})$ is small. Ideally, if $P^*(\mathbf{X}) \propto g^*(\mathbf{X})P(\mathbf{X})$, the variance of $g^*(\mathbf{X})\ P(\mathbf{X})/P^*(\mathbf{X})$ is zero. However, this would mean

$$P^*(\mathbf{X}) = \frac{g^*(\mathbf{X})P(\mathbf{X})}{\sum_{\mathbf{x}} g^*(\mathbf{x})\ P(\mathbf{x})} = \frac{g(\mathbf{X})}{\sum_{\mathbf{x}} g(\mathbf{x})},$$

and if the denominator were known the Monte Carlo would be pointless! (Hammersley and Handscomb, 1964). The "same shape" criterion is most important in the tails of the distribution $P^*(\cdot)$; it is a problem if $P^*(\mathbf{X})$ is very small when $g^*(\mathbf{X})P(\mathbf{X})$ is not, since then with low probability there will be very large terms in the estimator, and the Monte Carlo variance will be high. In order to

be able to use a given $P^*(\cdot)$ we need first to be able to simulate from it, and second to compute $g^*(\mathbf{x})P(\mathbf{x})/P^*(\mathbf{x})$ at the realized values $\mathbf{x}$ of $\mathbf{X}$. This is sometimes far from straightforward, but we defer further discussion to Chapter 7.

Note the difference between a "simulation study" and a "Monte Carlo analysis". Simulation studies are typically undertaken to discover empirically the distribution of a test statistic, or to assess the potential power of a study design. It involves the simulation of data random variables under a model of interest. In a Monte Carlo analysis, integrals, sums, or expectations are estimated by simulating random variables from some distribution, but the random variables are not normally the data random variables (often, the data are fixed) and the distribution is simply a tool to provide good estimates of the required expectations. In practice, the difference may be slight. The probability distribution we simulate from in a Monte Carlo estimation problem may often be closely related to the probability model underlying the data in a statistical problem. Conversely, the probability distribution we use in a simulation study could equally be a convenient tool, with reweighting used to adjust the realizations to the distribution of interest (equation (3.12)). In a Monte Carlo analysis we shall normally simulate conditional on fixed data, but in a simulation study it may sometimes also be desirable to simulate potential data conditional on partial data already obtained.

3.8 Reduction of Monte Carlo variance

The earliest use of Monte Carlo estimation on pedigrees was to estimate inbreeding coefficients. Before digital computers were available, Wright and McPhee (1925) traced random paths up pedigrees. By random choice of a male or female parent, one is realizing the ancestry of a particular allele, and hence realizations of the *ibd* status of, for example, the two genes within a current individual. Much more recently, using a computer, Edwards (1967) realized the descent of genes down pedigrees to estimate inbreeding coefficients. In effect, both Wright and McPhee (1925) and Edwards (1967) are realizing latent variables $\mathbf{S}$. To estimate the probability of a specified *ibd* pattern, $\mathbf{J}^*$, define

$$g^*(\mathbf{S}) = 1 \text{ if } \mathbf{J}(\mathbf{S}) = \mathbf{J}^* \qquad (3.14)$$
$$= 0 \text{ otherwise.}$$

Then the probability of the pattern $\mathbf{J}^*$ is the expectation of $g^*(\mathbf{S})$ under the distribution of the random descent of genes in pedigrees.

Any probability can be estimated as the expectation of an indicator variable in this way, but the method is often not very efficient, if only the probability of a particular $\mathbf{J}^*$ is needed. On the other hand, if the probabilities of all *ibd* patterns among a given set of current genes are desired, this may be an effective approach; each realization of $\mathbf{S}$ contributes to some *ibd* pattern $\mathbf{J}(\mathbf{S})$. Different realizations $\mathbf{S}$ are, of course, independent, but the probabilities of different *ibd* patterns $\mathbf{J}$ estimated from the same set of realizations are dependent. It is important to recognize this dependence, but it is seldom a practical problem; multinomial covariances are small for large Monte Carlo samples.

Another key idea in effective Monte Carlo is "Rao-Blackwellization" of estimators. This procedure is named for the classic Rao-Blackwell Theorem in Statistics, whereby the statistical variance of an estimator $g(\mathbf{X})$ is reduced by replacing it by its conditional expectation given some statistic T: if $h(T) = \mathrm{E}(g(\mathbf{X})|T)$,

$$\mathrm{E}(h(T)) \;=\; \mathrm{E}(g(X)) \text{ and } \mathrm{var}(h(T)) \leq \mathrm{var}(g(\mathbf{X})).$$

Here we replace a part of the Monte Carlo by exact computation of a (conditional) probability or expectation. Formally, suppose the latent variables $\mathbf{X}$ are divided into two sets of components $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. As before, we wish to estimate $\mathrm{E}_P(g^*(\mathbf{X})) = \mathrm{E}_P(g^*(\mathbf{X}_1, \mathbf{X}_2))$, where each of $\mathbf{X}_1$ and $\mathbf{X}_2$ is a (possibily vector) variable. If pairs $(\mathbf{X}_1^{(\tau)}, \mathbf{X}_2^{(\tau)})$, $i = 1, ..., N$ are independently realized from the probability distribution $P(\cdot)$, one estimator is (see equation (3.11))

$$T_N^* \;=\; \frac{1}{N}\sum_{i=1}^{N} g^*(\mathbf{X}_1^{(\tau)}, \mathbf{X}_2^{(\tau)}).$$

Suppose it is possible to compute $h(\mathbf{X}_1) = \mathrm{E}_P(g^*(\mathbf{X}_1, \mathbf{X}_2) \mid \mathbf{X}_1)$. Another Monte Carlo estimator is then

$$T_N = \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{X}_1^{(\tau)}).$$

Then the Monte Carlo variance of T_N is easily shown to be no larger than that of T_N^*, and usually strictly smaller. Whether such Rao-Blackwellization is computationally effective depends on whether the increased cost of computing $h(\mathbf{X}_1)$ rather than $g^*(\mathbf{X}_1, \mathbf{X}_2)$ is outweighed by the reduction in the number of the Monte Carlo realizations required to achieve a given precision. There is no general rule; see also section 9.4.

Returning to realizations of gene descent in pedigrees, suppose we wished to estimate by Monte Carlo the inbreeding coefficient of the offspring of double first cousins: in fact, the answer is 0.125 (Table 3.2). If we use the estimator of equation (3.14), scoring 1 for each realization of $\mathbf{S}$ in which the final offspring individual is autozygous (has two *ibd* genes), the Monte Carlo variance is that of a binomial proportion for probability 1/8: $(1/8)(7/8)(1/N) = 0.1094/N$. If instead, we score *ibd* patterns in the double-first cousins, we have a trinomial realization of $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_0) = (9/16, 6/16, 1/16)$. Then the inbreeding coefficient of the offspring is estimated by $\widehat{\psi} = (2\widehat{\kappa_2} + \widehat{\kappa_1})/4$, which has Monte Carlo variance

$$\begin{aligned}(1/4)\mathrm{var}(\widehat{\kappa_2}) + (1/4)\mathrm{cov}(\widehat{\kappa_2}, \widehat{\kappa_1}) + (1/16)\mathrm{var}(\widehat{\kappa_1}) \;&= \\ N^{-1}(0.01465 - 0.00586 + 0.01465) \;&=\; 0.02344/N\end{aligned}$$

which is almost 5 times smaller. In this case, $\mathbf{S}_1$ corresponds to the meioses down to the double first cousins, and $\mathbf{S}_2$ to the meioses from the double-first-cousins to their offspring. The original estimator scores 1 or 0 depending on whether or not $(\mathbf{S}_1, \mathbf{S}_2)$

implies autozygosity of the offspring individual. The conditional expectation $h(\mathbf{S}_1)$ is simply the probability of autozygosity in the final individual, given the particular $\mathbf{S}_1$ realized.

As another example, consider estimation of the inbreeding coefficient of the final individual of the pedigree of Figure 3.1. The actual value is $7/64 = 0.1094$ (section 3.2), so using direct gene-drop, the Monte-Carlo standard error is $\sqrt{(7/64)(57/64)/N} = 0.3121/\sqrt{N}$. Alternatively, we may use Monte Carlo only to the parents of the individual. In this case, there are nine possible states of gene *ibd* among four genes of these two parent individuals (Table 3.1). For each *ibd* pattern in the parents, the conditional expectation of the indicator of autozygosity of the offspring is simply the conditional probability given the parental *ibd* state. These probabilities that the final individual, B, receives two *ibd* genes, range from 1.0, for the parental pattern 1111, down to 0.0 for the pattern 1234:

$$\begin{aligned} f_B = \psi(M_B, F_B) = {} & \Pr(1111) + 0.5(\Pr(1112) + \Pr(1121) + \Pr(1211) \\ & + \Pr(1222) + \Pr(1212) + \Pr(1221)) + \\ & 0.25(\Pr(1213) + \Pr(1231) + \Pr(1223) + \Pr(1232)). \end{aligned}$$

Here $\Pr(k_1k_2k_3k_4)$ is the probability of that pattern among the four parental genes, the first two being the genes of one parent and the last two of the other. The Monte Carlo standard error of this estimate is approximately $0.17/\sqrt{N}$, in this case estimated empirically. There are two sources in the gain in efficiency, one replacing a part of the Monte-Carlo by an exact computation of an expectation (Rao-Blackwellization), and second the negative covariances of the Monte-Carlo multinomial proportions providing the estimated *ibd* pattern probabilities in the parents. Since $\psi(M, F)$ is a positive linear combination of these *ibd* pattern probabilities, the negative covariance reduces the Monte Carlo standard error of the estimate of $\psi(M, F)$. This idea is a little different from the use of *antithetic* variates (Hammersley and Handscomb, 1964), but of similar effect. Antithetic variates are negatively correlated realizations used to reduce the variance of a sum or average. Here the realizations of $\mathbf{S}$ are independent, but the component events are negatively correlated.

Chapter 4

Genetic Linkage

4.1 Linkage and recombination: genetic distance

Contrary to Mendel's second law (Mendel, 1866), there is dependence in the inheritance of genes at syntenic loci (that is, loci on the same chromosome pair). Such loci are said to be *linked.* Where the data are affected by the alleles at more than one locus on a chromosome pair, it is no longer sufficient to consider the inheritance of genes at each locus separately.

Recall the meiosis indicators of (equation (1.2)):

$$\begin{aligned} S_{i,j} &= 0 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's maternal gene} \\ &= 1 \quad \text{if copied gene at meiosis } i \text{ locus } j \text{ is parent's paternal gene.} \end{aligned}$$

Here $i = 1, ..., m$ indexes the meioses of the pedigree, and $j = 1, ..., L$ indexes the genetic loci. The marginal distribution of each $S_{i,j}$ is as before (section 1.2):

$$\Pr(S_{i,j} = 0) \;=\; \Pr(S_{i,j} = 1) \;=\; \frac{1}{2}.$$

For different meioses i, the $S_{i,j}$ are independent.

We say that, in a given meiosis, *recombination* has occurred between two loci j and l, if the genes segregating to the gamete at these two loci are from different parental chromosomes. That is, they derive from different grandparents. For two loci, we do not need a full model for the vector $S_{i,\bullet}$ (equation 1.3). The pairwise distribution of $(S_{i,j}, S_{i,l})$ is determined by the *recombination frequency*, which is a measure of the dependence in inheritance between the two loci. For two given loci (l and j) the recombination frequency ρ between them is

$$\rho \;=\; \Pr(S_{i,l} \neq S_{i,j}) \quad \text{for each } i, \quad 0 \leq \rho \leq \frac{1}{2}. \tag{4.1}$$

For loci that are close together on a chromosome, ρ is close to 0. For independently segregating loci, $\rho = \frac{1}{2}$. Note that, although θ is the notation often used for the

recombination parameter in genetic analysis, we here use ρ and reserve θ for the more general set of all parameters of the genetic model.

A point on a gamete chromosome at which the DNA switches from being a copy of the parent's maternal [paternal] chromosome to being a copy of the parent's paternal [maternal] chromosome is known as a *crossover.* Haldane (1919) defined *genetic map distance* between any two loci as the expected number of crossovers occurring between them on a gamete. The unit of genetic distance is the Morgan, but it is often more convenient to use centiMorgans (cM). Since expectations are additive, regardless of dependence of random variables, genetic map distances are always additive. They also subsume any positional variation in recombination rates such as recombination hot-spots: they say nothing about the relationship between physical and genetic distances. A recombination occurs between two loci, if, in that meiosis, there are an odd number of crossovers between them.

In equation (4.1), we assume that the recombination frequency ρ does not vary with the meiosis i. In practice, recombination frequencies vary among meioses, a major factor in this variation being the sex of the parent. The expected number of crossovers between two locations can be quite different for a gamete from a male than for a gamete from a female. Thus genetic maps are sex-specific, where the sex in question is that of the parent producing the gamete. For ease of presentation, sex-differences in genetic maps will be ignored in this monograph. Computationally, such variation can be easily accommodated.

Haldane's original meiosis model, and other early models, were *two-strand* models. That is, the locations of crossovers between the two parental chromosomes were modeled. This is sufficient to determine the joint probabilities $\Pr(S_{i,1}, \ldots, S_{i,L})$, hence, in principle, probabilities of L-locus gene *ibd* patterns among a set of observed related individuals, and hence probabilities of observed data. In Haldane's model, these crossovers were assumed to occur as a Poisson process, rate 1 (per Morgan). Thus there is no *interference.* The number of crossovers in a given genetic distance has a Poisson distribution, the numbers of crossovers in disjoint intervals are independent, and, conditionally on the number occurring, their locations are uniformly and independently distributed, all measures being, of course, with respect to genetic (not physical) distance. The recombination frequency at genetic distance d Morgans, $\rho(d)$, as a function of d is known as the *map function.* Under the no-interference model, $\rho(d)$ is the probability that a Poisson random variable with mean d is odd:

$$\begin{aligned} \rho(d) &= \sum_{k \text{ odd}} e^{-d}\frac{d^k}{k!} = \frac{1}{2}e^{-d}\sum_{k=0}^{\infty}\left(\frac{d^k}{k!} - \frac{(-d)^k}{k!}\right) \\ &= \frac{1}{2}(1 - \exp(-2d)). \end{aligned} \tag{4.2}$$

Note that, under this model, $\rho(d)$ is an increasing function of d, $\rho(d) \to \frac{1}{2}$ as $d \to \infty$, and $\rho(d) \approx d$ as $d \to 0$. These are basic properties of map functions under most models for meiosis (see Chapter 5).

In modeling crossovers Fisher (1922) went to the other extreme: he assumed complete interference in the region of *Drosophila willistoni* chromosome he

considered. That is, at most one crossover in this chromosome region can occur in any meiosis. In this case, genetic distance and recombination frequency are equivalent. Although this model does not make sense over large chromosomal segments, current mouse data (King et al., 1991) suggest almost complete interference over regions of about 10cM.

4.2 Haplotypes, linkage, and association

The vector of alleles at loci on a chromosome is a *haplotype*, and a *multilocus genotype* is a pair of haplotypes. Note that the set of single-locus genotypes do not determine the multilocus genotype. The multilocus genotype includes a specification of *phase*; that is, which alleles (one at each locus) are on the same chromosome. Some modern literature does refer to the set of single-locus genotypes (without phase) as the multilocus genotype, but this terminology is confusing. For clarity, we refer to the potentially observable set of (single-locus) genotypes at any set of DNA marker loci as *marker phenotypes*, even when these loci do not correspond to functional genes.

For simplicity in this section we restrict attention to two diallelic loci, one with codominant alleles A_1 and A_2, and the other with codominant alleles B_1 and B_2. There are then four haplotypes A_1B_1, A_1B_2, A_2B_1 and A_2B_2. Suppose the haplotype frequencies are q_1, q_2, q_3 and q_4. There are 10 two-locus genotypes, but only 9 phenotypes. Genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1 both have the double-heterozygote phenotype A_1A_2, B_1B_2. The notation A_1B_1/A_2B_2 denotes that alleles A_1 and B_1 are on a single haplotype, and alleles A_2 and B_2 are on the other. Just as for the single-locus *ABO* blood type example (section 2.5), haplotype frequencies can be estimated from phenotype frequencies via the EM algorithm, under the general model of unconstrained patterns of association among the loci. Each phenotypic observation on an individual consists of a set of single-locus genotypes.

For the case of two loci, haplotypes are unobservable only for the double-heterozygote phenotype A_1A_2, B_1B_2. Each individual who is A_1A_2, B_1B_2 is of genotype A_1B_1/A_2B_2 with probability $q_1q_4/(q_1q_4 + q_2q_3)$ and of genotype A_1B_2/A_2B_1 with probability $q_2q_3/(q_1q_4 + q_2q_3)$. Thus, given a set of current haplotype frequency estimates q_i, $i = 1, \ldots, 4$ and the phenotypic counts, the conditional expected genotypic counts are easily obtained. New haplotype estimates then are the expected multinomial proportions of each haplotype.

Clearly, this method can be extended to any number of loci. Thus, for example, population data can be used to estimate haplotype frequencies at a set of tightly linked SNP markers (section 1.1). However, an individual heterozygous at l loci can have any of 2^{l-1} multilocus genotypes (pairs of haplotypes). The observation is partitioned among the 2^{l-1} possible pairs, in accordance with current haplotype frequency estimates. Performance of the EM algorithm can be poor when there are many linked polymorphic marker loci, particularly when many haplotypes may not occur in the sample. Thus, for microsatellite markers with many alleles or for many tightly linked SNP markers (section 1.1), population marker phenotype data alone

may not serve to provide accurate haplotype frequencies. Better performance of the EM algorithm is obtained by constraining some haplotype frequencies to zero, when the estimates of their frequencies appear to be approaching zero.

An individual who is homozygous at both loci can pass on only one haplotype to an offspring; for example an A_1A_1, B_2B_2 individual must pass on an A_1B_2 haplotype. An individual who is homozygous at one locus can pass either of two haplotypes. Each possibility has probability $1/2$ regardless of the recombination frequency ρ between the two loci; for example, an A_1A_1, B_1B_2 individual passes on A_1B_1 or A_1B_2 each with probability $1/2$. Only the double heterozygote A_1A_2, B_1B_2 provides meioses which are *informative for linkage.* That is, this individual passes each of the four haplotypes A_1B_1, A_1B_2, A_2B_1 and A_2B_2, with probabilities $(1-\rho)/2$, $\rho/2$, $\rho/2$ and $(1-\rho)/2$ if his genotype is A_1B_1/A_2B_2, and with probabilities $\rho/2$, $(1-\rho)/2$, $(1-\rho)/2$, and $\rho/2$ if his genotype is A_1B_2/A_2B_1.

A measure of allelic association between the two loci is

$$\begin{aligned}\Delta &= \Pr(A_1B_1) - \Pr(A_1)\Pr(B_1)\\ &= q_1 - (q_1+q_2)(q_1+q_3)\\ &= (q_1q_4 - q_2q_3)\end{aligned}$$

since $q_1+q_2+q_3+q_4=1$. This measure is due to Robbins (1918) and is known as the coefficient of *linkage disequilibrium.* This name is confusing, but the term is too well established to change. In the absence of selection, allelic associations between loci arise from population structure, admixture and history. They are, however, maintained by tight linkage. Suppose the current haplotype frequencies are q_1, q_2, q_3 and q_4, as above. In expectation, in the absence of selection, allele frequencies are unchanged at the next generation. Suppose the haplotype frequencies are q_1^*, q_2^*, q_3^* and q_4^*. Now, for example, an A_1B_1 haplotype in an offspring can arise in three ways. It can be transmitted from a parental A_1B_1 without recombination. It can also be transmitted from a parental A_1B_1 with recombination, if the second parental haplotype is A_1B_1, A_1B_2, or A_2B_1. Finally, with recombination, an A_1B_2/A_2B_1 parent may transmit an A_1B_1 haplotype. Thus

$$\begin{aligned}q_1^* &= (1-\rho)q_1 + \rho q_1(q_1+q_2+q_3) + \rho q_2q_3\\ &= q_1 - \rho(q_1q_4 - q_2q_3) = q_1 - \rho\Delta.\end{aligned}$$

Analogously, $q_2^* = q_2 + \rho\Delta$, $q_3^* = q_3 + \rho\Delta$ and $q_4^* = q_4 - \rho\Delta$. Thus

$$\begin{aligned}\Delta^* &= q_1^*q_4^* - q_2^*q_3^*\\ &= (q_1-\rho\Delta)(q_4-\rho\Delta) - (q_2+\rho\Delta)(q_3+\rho\Delta)\\ &= \Delta - \rho\Delta(q_1+q_2+q_3+q_4) + \rho^2(\Delta-\Delta)\\ &= (1-\rho)\Delta.\end{aligned}$$

In the absence of any maintaining force, such as selection, or continuing population subdivision and admixture, allelic associations decay in expectation over the generations, by a factor $(1-\rho)$. For unlinked loci ($\rho = \frac{1}{2}$) this decay is rapid, but for tightly linked loci ($\rho \approx 0$) allelic associations may be maintained over

hundreds of generations. Actual population are finite, and mating is non-random; allelic associations are often seen in small natural populations. For a more detailed discussion, see Weir (1996).

4.3 Lod scores for two-locus linkage analysis

In the absence of genetic interference (equation (4.2)), and in fact under most models for meiosis (Chapter 5), the recombination frequency, ρ, is an increasing function of genetic distance. Genetic mapping involves the ordering of loci on a chromosome, the detection of linkage, and the estimation of recombination frequencies. Some loci determine traits: others are DNA markers. Typically, a map constructed of DNA markers is then used to map the loci controlling a trait of interest. For unlinked loci, $\rho = \frac{1}{2}$. For loci that are genetically *linked*, $\rho < \frac{1}{2}$. *Linkage analysis* is concerned with estimating ρ and with testing the null hypothesis $H_0 : \rho = \frac{1}{2}$ against the alternative $H_1 : \rho < \frac{1}{2}$. Estimates and tests are based on likelihoods and likelihood ratios (Chapter 2).

If the genes (one at each of two loci) descending from given parent to a given offspring derive from different parental chromosomes, and hence from different grandparents, the offspring is said to be *recombinant* with respect to these two loci. In the simplest cases, whether an offspring i is a recombinant ($X_i = 1$) or not ($X_i = 0$) is observable. Then $P(X_i = 1) = \rho$ and the number of recombinants T in n independent meioses has the binomial $B(n, \rho)$ distribution.

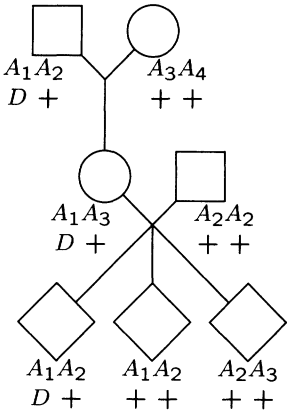

FIGURE 4.1. *Example of recombination in a three-generation family*

For example, at a DNA marker locus, suppose two grandparents have types A_1A_2 and A_3A_4, and their daughter has type A_1A_3. Suppose she marries someone of type A_2A_2 and their three children are of types A_1A_2, A_1A_2 and A_2A_3. Suppose also the grandparent of type A_1A_2, the daughter, and the first of the three children

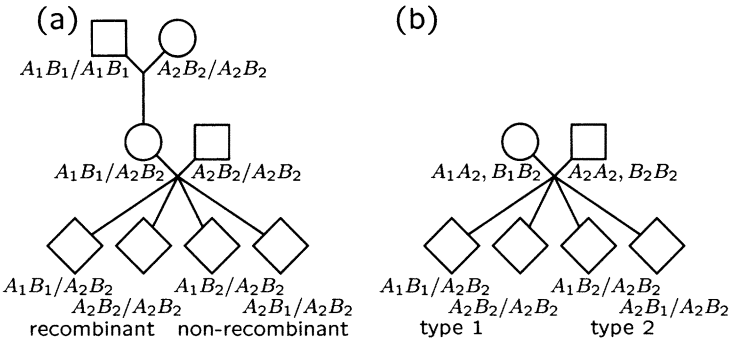

FIGURE 4.2. *Examples of (a) phase-known and (b) phase-unknown backcross linkage designs*

all carry an allele D causing some trait of interest, and the other individuals carry only normal alleles, denoted + (Figure 4.1). Then we know the trait allele D segregates with the A_1 marker allele from the grandparent to his daughter, and that the normal allele + segregates with A_3 from her other parent. To the three children from their mother, we have segregation of A_1 with D, of A_1 with +, and of A_3 with +. Thus children 1 and 3 are non-recombinant ($X_1 = X_3 = 0$) and child 2 is recombinant ($X_2 = 1$). So $n = 3$, the number of recombinants $T \sim B(3, \rho)$, and in this example T takes the value $t = 1$.

In the case where we can classify each offspring as recombinant or non-recombinant, as above, the number of recombinants in n observed offspring is $T \sim B(n, \rho)$. This type of data arises in a *backcross experiment*, where two inbred lines are crossed, and the hybrid is crossed back to either of the two lines. An example of this linkage design is shown in Figure 4.2(a). Suppose one line has only alleles A_1 at one locus and B_1 at the other (genotype A_1B_1/A_1B_1), while the other line has only A_2 and B_2 (genotype A_2B_2/A_2B_2). Then the cross will produce hybrid individuals who have genotype A_1B_1/A_2B_2. If we then cross back to the A_2B_2/A_2B_2 line, all the offspring will get A_2B_2 from that parent, and we can tell which combination A_1B_1, A_2B_2, A_1B_2 or A_2B_1 they get from their hybrid parent, and so whether or not they are recombinant.

Suppose n offspring of such matings are scored, and t are recombinant. To test for linkage, we compare the likelihood to its value in the absence of linkage ($\rho = \frac{1}{2}$). The log-likelihood difference is

$$\text{(4.3)} \quad \text{lod}(\rho) \;=\; \ell(\rho) - \ell(\frac{1}{2}) \;=\; t\log(\rho) \;+\; (n-t)\log(1-\rho) \;+\; n\log(2).$$

In linkage analysis it is traditional to use logs to base 10, and to refer to (4.3) as the *lod score* (Morton, 1955). In our numerical examples we shall use natural logarithms except where specified, for easier comparison with standard statistical results.

The maximum likelihood estimate of ρ is $\widehat{\rho} = t/n$, provided $2t \leq n$: note only values of $\rho \leq \frac{1}{2}$ have meaning under the model (4.2). Then to test $\rho = \frac{1}{2}$ against $\rho < \frac{1}{2}$, we may consider the maximized value of the lod score:

$$\text{lod}(\widehat{\rho}) \quad = \quad t \log t \; + \; (n-t)\log(n-t) - n\log(n/2) \tag{4.4}$$

provided $2t \leq n$, and 0 otherwise. This maximized lod score is a decreasing function of t, and we reject the null hypothesis $\rho = \frac{1}{2}$ if $t < t_0$. The critical value t_0 may be chosen to give a specified size of the test (type I error).

In many linkage experiments, however, or in human genetics where we do not have designed crosses, we often cannot classify all individuals as recombinant and non-recombinant. There are many possibilities, but a typical one is the *phase-unknown backcross*. This arises if one parent is A_1A_2, B_1B_2 and the other is A_2A_2, B_2B_2 as above, but now we do not know whether the first parent received A_1B_1 and A_2B_2 (type 1 combinations) from her parents, or A_1B_2 and A_2B_1 (type 2 combinations). This design is shown in Figure 4.2(b). Suppose we have families of this kind, and in each family we type just two offspring. Since each offspring gets A_2B_2 from the father, we can, as before, determine what each got from the mother. Either both offspring get the same "type" of combination (type 1 or type 2), or there is one of each. If there is one of each, then one offspring must be a recombinant and the other not; so this event has probability $\rho^* = 2\rho(1-\rho)$. If they get the same "type" of combination, then either both are recombinant, or neither is, so this event has probability $1 - \rho^* = \rho^2 + (1-\rho)^2$. So instead of a $T \sim B(n, \rho)$ count of recombinants, we have a $W \sim B(n, \rho^*)$ count of families.

Note however, that for $0 \leq \rho \leq \frac{1}{2}$, ρ^* is a 1-1 monotone increasing function of ρ, and when $\rho = \frac{1}{2}$ $\rho^* = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$. So testing $H_0 : \rho = \frac{1}{2}$ against the one-sided alternative $H_1 : \rho < \frac{1}{2}$, is exactly equivalent to testing $H_0^* : \rho^* = \frac{1}{2}$ against the one-sided alternative $H_1^* : \rho^* < \frac{1}{2}$. Thus the test follows exactly as before; we reject $\rho^* = \frac{1}{2}$ and conclude there is linkage if $W < w_0$, where again the critical value w_0 is determined by the desired size of the test.

4.4 Power, information and *Elod*s

For simplicity, consider the case of the phase-known backcross, where $T \sim B(n, \rho)$. Now when n is large, T is approximately $N(n\rho, n\rho(1-\rho))$, and under $H_0 : \rho = \frac{1}{2}$ it is a *good* approximation to take $T \sim N(\frac{n}{2}, \frac{n}{4})$. So $\frac{2}{\sqrt{n}}(T - \frac{n}{2}) \sim N(0,1)$ and for a test size α we reject H_0 in favor of $H_1 : \rho < \frac{1}{2}$ if $\frac{2}{\sqrt{n}}(T - \frac{n}{2}) \leq \Phi^{-1}(\alpha)$ where Φ is the standard Normal cumulative distribution function. For example, for $\alpha = 0.025$, $\Phi^{-1}(\alpha) = -1.96 \approx -2$, so H_0 is rejected if $T \leq \frac{n}{2} - \sqrt{n} = k^*$ (Table 4.1).

Using equation (4.4), we find that the (base 10) lod score is around 1 for a number of recombinants at the critical value for a test of size $\alpha = 0.025$ of $H_0 : \rho = \frac{1}{2}$ (Table 4.1). Traditionally, a base-10 lod score of 3 is required to infer linkage (Morton, 1955). This is a more stringent test, the idea being that if two arbitrary locations in the genome are chosen the prior probability of linkage is small. Also given in the table is the upper bound on the number of recombinants that will provide a lod score of 3.

offspring sampled n	critical value k^*	recombinant proportion k^*/n	lod score $\text{lod}_{10}(k^*/n)$	recombinants for lod score 3
25	≈ 7	≈ 0.3	1.088	≤ 3
100	≈ 40	≈ 0.4	0.874	≤ 31
625	≈ 287	≈ 0.46	0.905	≤ 267
1024	≈ 480	≈ 0.48	0.869	≤ 452

TABLE 4.1. *Critical values for a test size $\alpha = 0.025$ and base-10 lod scores for binomial samples*

Type	genotypes	number	each prob
I	$A_1A_1, B_2B_2,\ A_2A_2, B_1B_1$	2	$\rho^2/4$
II	A_1A_1, B_1B_2	1	$\frac{1}{2}(\rho^2 + (1-\rho)^2)$
III	A_1A_1, B_1B_2 etc.	4	$\frac{1}{2}\rho(1-\rho)$
IV	$A_1A_1, B_1B_1,\ A_2A_2, B_2B_2$	2	$(1-\rho)^2/4$

TABLE 4.2. *The groups of offspring genotypes in an intercross design. Note the A_1A_1, B_1B_2 type includes both double-heterozygote two-locus genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1. The third group includes the four types heterozygous at one of the two loci: A_1A_1, B_1B_2, A_1A_2, B_1B_1, A_2A_2, B_1B_2 and A_1A_2, B_2B_2*

Now if ρ is the true value, the probability H_0 is rejected is

$$\begin{aligned} \Pr(T < k^*; \rho) &= \Pr\left(\frac{T - n\rho}{\sqrt{n\rho(1-\rho)}} < \frac{k^* - n\rho}{\sqrt{n\rho(1-\rho)}}\right) \\ &\approx \Phi\left(\frac{k^* - n\rho}{\sqrt{n\rho(1-\rho)}}\right) = \Phi\left(\frac{\Phi^{-1}(\alpha) + \sqrt{n}(1-2\rho)}{2\sqrt{\rho(1-\rho)}}\right) \end{aligned} \tag{4.5}$$

again using the Normal approximation to the Binomial distribution. This is the power function of the test, and decreases over $0 \leq \rho \leq \frac{1}{2}$. Clearly, for a given sample size, linkage is more easily detected when ρ is small. Conversely, for given ρ, one may use (4.5) to determine the sample size n required for given power. The case of the phase-unknown backcross is analogous, with ρ being replaced by ρ^*, and n now denoting the number of two-child families.

In order to get more information, an *intercross* experiment may be performed, instead of a *backcross*. In this case two phase-known hybrid parents, each of type A_1B_1/A_2B_2 are mated. There are nine types of offspring, but these fall into four groups, shown in Table 4.2. Each type within a group has the same probability, as a function of ρ, and hence the total count of offspring in each group contains all the available information for linkage. (These total counts are the sufficient statistics for ρ.)

Consider a sample of size n, with n_j in class j, $j = 1, 2, 3, 4$. As in equation (2.5),

Types	H_2 : general	H_1:total prob	H_0 : $\rho = \frac{1}{2}$
I	q_1	$\frac{1}{2}\rho^2$	0.125
II	q_2	$\frac{1}{2}(\rho^2 + (1-\rho)^2)$	0.25
III	q_3	$2\rho(1-\rho)$	0.5
IV	q_4	$\frac{1}{2}(1-\rho)^2$	0.125

TABLE 4.3. *Probabilities of data observations in an intercross design. Given are the total probabilities of each group of types shown in Table 4.2, under the three alternative hypotheses*

the log-likelihood for these multinomial data is, up to an additive constant,

$$\ell_n(\mathbf{q}) = \sum_{j=1}^{4} n_j \log_e q_j(\rho).$$

The probabilities of each phenotype group are shown in Table 4.3, under the general multinomial model H_2, the general linkage model H_1, and in the absence of linkage H_0.
For example, suppose $\mathbf{n} = (1, 72, 42, 85)$.
Under H_2 : *general* q_j, $\sum_{j=1}^{4} q_j = 1$, $\widehat{q_j} = n_j/n$,
or $\hat{\mathbf{q}} = (0.005, 0.36, 0.21, 0.425)$. The dimension of H_2 is 3.
Under H_1 : *general* ρ, for these data we find, by evaluating the log-likelihood, that $\hat{\rho} = 0.12$ giving $\mathbf{q}(\hat{\rho}) = (0.007, 0.394, 0.211, 0.387)$. The dimension of H_1 is 1.
The null hypothesis is of no linkage; H_0 : $\rho = \frac{1}{2}$. This has dimension 0, and the fixed probabilities $\mathbf{q}(\frac{1}{2}) = (0.125, 0.25, 0.5, 0.125)$ of the four classes of types.

We see that the estimated cell probabilities under H_1 and H_2 are in good agreement, but quite different from those under H_0. Computing the maximized log-likelihoods for H_i, $i = 0, 1, 2$, we find that they are -307.76, -217.87, and -217.14 respectively. For testing null H_0 against alternative H_1, the (base e) lod score is 89.9. Twice this value (179.8) has approximately a χ_1^2 if H_0 is true. So clearly H_0 is rejected.
For testing null H_1 against alternative H_2, the lod score is 0.73, and twice this value (1.46) is χ_2^2 if H_1 is true. So H_1 is not rejected.

For multinomial data in general, we can find the form of the Kullback-Leibler information (section 2.2). Suppose $\mathbf{q}$ is the true value of $\mathbf{q}$, and $\mathbf{q_0}$ is some hypothesized value.

$$\ell_n(\mathbf{q}) = \sum_{j=1}^{4} n_j \log_e q_j.$$

So for a sample size n

$$\begin{aligned} K_n(\mathbf{q_0}; \mathbf{q}) &= \mathrm{E}_{\mathbf{q}}(\ell_n(\mathbf{q}) - \ell_n(\mathbf{q_0})) \\ &= n\sum_{j=1}^{4} q_j \log_e q_j - n\sum_{j=1}^{4} q_j \log_e q_{0j} \end{aligned}$$

True ρ	0.0	0.1	0.2	0.3	0.4	0.5
Intercross data	1.04	0.479	0.226	0.089	0.021	0.0
Backcross (phase known)	0.69	0.368	0.193	0.082	0.021	0.0
Backcross (phase unknown)	0.35	0.111	0.033	0.006	0.0004	0.0

TABLE 4.4. *Comparison of the information in linkage designs per offspring individual sampled: Kullback Leibler information for testing $\rho = 1/2$ as a function of the true value of ρ*

or, for a single observation,

$$K_1(\mathbf{q_0}; \mathbf{q}) = \sum_{j=1}^{4} q_j \log_e \left(\frac{q_j}{q_{0j}} \right).$$

(Note the notation is reversed from section 2.2. Here $\mathbf{q}$ is the true parameter value, and $\mathbf{q_0}$ is the hypothesized value.) In the case of linkage analysis data, $q_j = q_j(\rho)$ and the null hypothesis is $H_0 : \rho = \frac{1}{2} : \ q_{0j} = q_j(\frac{1}{2})$. Evaluating K_1 for the above *phase-known intercross* experiment, and for the previous binomial *phase-known* and *phase unknown* backcross experiments, we obtain the measures of information per offspring individual shown in Table 4.4.

This is a measure of information, per offspring sampled, for detecting linkage when ρ is the true value. We see that the more ρ differs from $\frac{1}{2}$ the more information there is, as expected. Also each phase-known offspring contributes at least twice as much as each of the two offspring in the phase-unknown case. Particularly when ρ is close to $1/2$, the phase-unknown two-offspring design has low power. We see that each *intercross* offspring contains more information than a *backcross* offspring, also as expected. However, note that there is *not* twice as much information in the intercross offspring, as there would be if we could tell the difference between the A_1B_1/A_2B_2 and A_1B_2/A_2B_1 offspring (see Table 4.3). As ρ gets closer to $\frac{1}{2}$ there is almost no additional information in doing an intercross design rather than a backcross.

Note that for $\rho = \frac{1}{2}$, the Kulback-Leibler information is the expected base-e lod score at the true value ρ_T of the recombination frequency. This measure of information is very widely used in linkage analysis, and is known as the *Elod* (Thompson et al., 1978). Note that we expect the base-e lod score to be approximately nK_1 when n is large. For our previous data with $n = 200$, we had $\hat{\rho} = 0.12$; in fact, the data were simulated at $\rho = 0.1$. Then 200×0.479 is about 95, in good agreement with the lod score value of 90 which we obtained. This also tells us that if we had realized that ρ might be around 0.1, it was very wasteful to breed 200 mice. When $\rho = 0.1$, about 20 mice are expected to give a lod score (base e) of more than 9; this is plenty to detect that $\rho \neq \frac{1}{2}$. (Note again that we have used natural logarithms in these examples, contrary to standard practice in genetics.)

The material of sections 4.3 and 4.4 extends readily to the estimation and testing of two recombination frequencies ρ_m in males, and ρ_f in females. Similar likelihood ratio tests may be used to test equality of male and female recombination frequencies. For a much more detailed account of classical linkage analysis and more

modern developments, the reader may consult the excellent text of Ott (1999).

4.5 Two-locus kinship and gene identity

The recursive equations for multiple kinship coefficients of section 3.4 (equations (3.6) and (3.7)) extend to multiple loci, conditioning on the meiosis indicators in a given meiosis, over the loci in question. Consider, for example, the case of $\psi_2(L_1(B^{(1)}, C), L_2(B^{(1)}, E))$. This expression denotes the two locus kinship probability, that, in a single gamete segregating from B, the gene at locus L_1 is *ibd* to that on a gamete segregating from individual C, while the gene at locus L_2 is *ibd* to that on a gamete segregating from individual E. The identical superscript "(1)" on the individual B indicates that we are considering here a single meiosis i from B, rather than two separate meioses to different offspring. Now if B is not an ancestor of C or E, we may condition on the four events $(S_{i,1},\ S_{i,2}) = (0,0),\ (0,1),\ (1,1),\ (1,0)$ with probabilities $\frac{1}{2}(1-\rho),\ \frac{1}{2}\rho,\ \frac{1}{2}(1-\rho),\ \frac{1}{2}\rho$ respectively, where ρ is the recombination frequency between locus 1 and locus 2. Thus we obtain

$$\begin{aligned}\psi_2(L_1(B^{(1)},\ C), L_2(B^{(1)},\ E)) \ &= \ \frac{1}{2}(1-\rho)\psi_2(L_1(M_B^{(B)}, C), L_2(M_B^{(B)}, E)) \ + \\ &\frac{1}{2}\rho\psi_2(L_1(M_B, C), L_2(F_B, E)) + \frac{1}{2}(1-\rho)\psi_2(L_1(F_B^{(B)}, C), L_2(F_B^{(B)}, E)) \\ &+ \ \frac{1}{2}\rho\psi_2(L_1(F_B, C), L_2(M_B, E)).\end{aligned}$$

Again, the superscript specifies which meiosis from an individual is considered—here the ones from M_B and F_B to B. In the case of two loci it is necessary to distinguish the meioses from a given parent. The full set of equations for determining two-locus gene identity probabilities between genes segregating from up to four individuals are given by (Thompson, 1988). These equations can be used to determine two-locus *ibd* state probabilities, even on a large and complex pedigree.

At two linked loci, there are also many more possible gene identity patterns (Denniston, 1975). Some relationships which have identical gene *ibd* probabilities at a single locus can, in principle, be distinguished by data at linked loci. The simplest example is for the three unilineal ($\kappa_2 = 0$) pairwise relationships of grandmother-granddaughter (G), half-sisters (H), and aunt-niece (N). Each of these relationships has $\boldsymbol{\kappa} = (\frac{1}{2}, \frac{1}{2}, 0)$, and hence they are indistinguishable on the basis of data at independently segregating loci. For such relationships, gene identity at two linked loci is summarized by

$$\kappa_{1,1}(\rho) \ = \ P(\textit{share 1 gene ibd at each of 2 loci at recombination } \rho).$$

For the three relationships above, we have

$$G \ : \ \kappa_{1,1}(\rho) \ = \ \frac{1}{2}(1-\rho)$$

$$H \ : \ \kappa_{1,1}(\rho) = \frac{1}{2}(\rho^2 + (1-\rho)^2) = \frac{1}{2}R \text{ say}$$
$$N \ : \ \kappa_{1,1}(\rho) = \frac{1}{2}((1-\rho)R + \rho/2).$$

Thus the relationships are identifiable of the basis of data at two linked loci ($0 < \rho < \frac{1}{2}$), but not on the basis of data at unlinked loci. All the three relationships have $\kappa_{1,1}(0) = \frac{1}{2}$ and $\kappa_{1,1}(\frac{1}{2}) = \frac{1}{4}$.

Note that, although $\kappa_{1,1}(\rho)$ is sufficient to specify pairwise genotype and phenotype distributions, it does not determine the two-locus kinship of the individuals, unlike at a single locus where $\psi = (\kappa_1 + 2\kappa_2)/4$. The shared genes at the two loci may be on the same haplotype in the individual, or on different ones. In fact, in H they are necessarily on the same (maternal) haplotype in the two half-sibs, while in G they may be on either haplotype of the grandmother. For N, for the first term they are on the same haplotype in the aunt, while the last term corresponds to the case where the genes at the two loci are on two different haplotypes in the aunt. In fact, G and H have the same two-locus kinship, $(1/8)(1-\rho)^2R$.

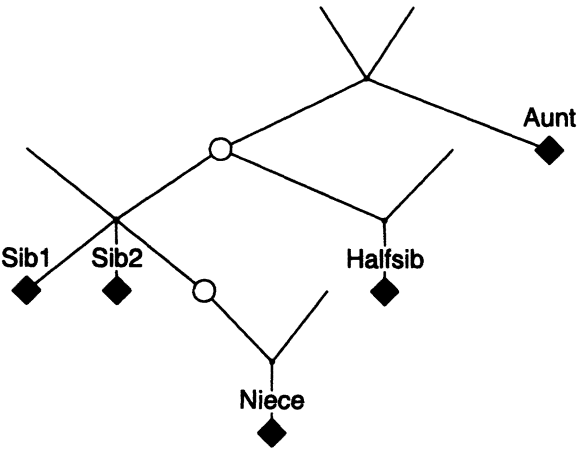

FIGURE 4.3. *Multi-locus genetic marker data are available on a pair of sibs, and on a third related individual, who may be an aunt, niece, or half-sister of the pair*

Returning again to the example of section 3.6, consider three individuals consisting of a pair of individuals who are putative full sibs, and a third who may be the aunt, niece, or half-sib of the sib pair (Figure 4.3). This example arose in a real example of inference of relationships considered by Browning (1999). Only with joint analysis of the data at linked loci on all three individuals are the alternative three relationships identifiable (Table 4.5). In the real-data example, the most likely

	Individuals	
	Pairwise	Joint
Loci unlinked	$H \equiv N \equiv A$	$H \equiv N$
Loci linked	$N \equiv A$	H, N, A identifiable

TABLE 4.5. *Distinguishing relationships among three individuals who are putatively a pair of sisters with an aunt, niece, or half-sib*

relationship is that the third individual is a niece of the sib pair (Browning and Thompson, 1999). Due to one member of the sib pair having data at relatively few markers, the inference is not conclusive. However, with data on a 10cM genome scan, for example, there would be no difficulty in distinguishing the relationships provided analysis is performed jointly both over individuals and over loci.

Some pairwise relationships which provide identical two-locus kinship coefficients have different three-locus kinship coefficients (Thompson, 1988). Thus, there are relationships that are non-identifiable on the basis of gametes observed at pairs of loci (whatever the values of the recombination frequencies between them), but that are identifiable on the basis of gametes observed at trios of loci. One may conjecture that there are relationships non-identifiable on the basis of L-locus kinship, but identifiable on the basis of $L + 1$-locus kinship.

4.6 Homozygosity mapping with a single marker

We introduce the ideas both of linkage analysis for linkage detection and of association analysis for the fine-scale localization of trait genes via *homozygosity mapping* using the ideas of two-locus gene *ibd* already encountered. Homozygosity mapping was developed by Lander and Botstein (1987) to detect linkage for the loci determining rare recessive disease traits, but as noted by Smith (1953) the principle is the same as in any linkage analysis: a likelihood for the recombination frequency ρ, or more generally for the trait locus location, is computed. With a single marker locus, the maximized likelihood under the hypothesis of linkage $\rho < \frac{1}{2}$ is compared with the likelihood under the hypothesis that the trait locus is not linked to the marker locus or loci $\rho = \frac{1}{2}$. In the case of homozygosity mapping, the linkage inference is based on data on unrelated affected inbred individuals. It relies on the fact that an inbred affected individual has high probability of carrying two *ibd* genes at the the trait (disease) locus (section 3.3), and hence also at any closely linked marker locus. Since *ibd* genes are necessarily of the same allelic type, such individuals will show a patch of homozygosity in the neighborhood of the trait locus; where the same markers are homozygous across multiple inbred affected individuals, the evidence for linkage accumulates.

Suppose the frequency of the recessive disease allele is q, and at the marker locus alleles A_i have frequencies p_i. Suppose that the affected individual has inbreeding coefficient f, and probability $f_2(\rho)$ of carrying genes *ibd* at both of two loci between which the recombination frequency is ρ. Then the probability the individual is

autozygous at a specific one of the two loci but not the other is $f - f_2(\rho)$, and the probability he is autozygous at neither is $(1 - 2f + f_2(\rho))$. If the individual has marker phenotype A_jA_l, he cannot be autozygous at the marker locus, and we have likelihood ratio

$$\begin{aligned}\frac{L(\rho)}{L(\rho=\frac{1}{2})} &= \frac{\Pr(\text{data}\ ;\ \rho)}{\Pr(\text{data}\ ;\ \rho=\frac{1}{2})} \\ &= \frac{2p_jp_l(q(f-f_2(\rho))+q^2(1-2f+f_2(\rho)))}{2p_jp_l(q(f-f^2)+q^2(1-f)^2)} \\ &= \frac{(f-f_2(\rho))+q(1-2f+f_2(\rho)))}{(1-f)(f+q(1-f))}. \qquad (4.6)\end{aligned}$$

Since sampling is through an affected individual, the data probabilities required here are those of the marker phenotypes, conditional on the affected trait phenotype. However, since the marginal probability of an affected individual, $qf + q^2(1-f)$, does not depend on ρ, the likelihood ratio is also the ratio of the joint probabilities of marker and trait phenotypes. The joint probabilities are slightly more easily considered.

Since $f_2(\rho)$ is a decreasing function of ρ, with value f^2 at $\rho = \frac{1}{2}$, the likelihood ratio (4.6) is always less than one. A heterozygous marker phenotype provides evidence against linkage. However, even at $\rho = 0$, where the value is $q(1-f)/(f + q(1-f))$ the evidence against linkage is not strong unless q is very small. Affected individuals may not carry *ibd* genes at the trait locus.

If the individual has homozygous marker phenotype A_jA_j the likelihood ratio is

$$\begin{aligned}\frac{L(\rho)}{L(\rho=\frac{1}{2})} &= \frac{\Pr(\text{data}\ ;\ \rho)}{\Pr(\text{data}\ ;\ \rho=\frac{1}{2})} \\ &= \frac{qp_jf_2(\rho) + q^2p_j(f-f_2(\rho)) + qp_j^2(f-f_2(\rho)) + q^2p_j^2(1-2f+f_2(\rho))}{qp_jf^2 + q^2p_jf(1-f) + qp_j^2f(1-f) + q^2p_j^2(1-f)^2} \\ &= \frac{f_2(\rho) + q(f-f_2(\rho)) + p_j(f-f_2(\rho)) + qp_j(1-2f+f_2(\rho))}{f^2 + qf(1-f) + p_jf(1-f) + qp_j(1-f)^2}. \qquad (4.7)\end{aligned}$$

The coefficient of the decreasing function of ρ, $f_2(\rho)$, is $(1-q)(1-p_j)$, and thus this likelihood ratio is maximized at $\rho = 0$. At this value, $f_2(\rho) = f$, so the likelihood ratio is

$$\frac{f + (1-f)qp_j}{(f+(1-f)q)(f+(1-f)p_j)}.$$

This is always greater than one, and is larger if q or p_j is small.

Likelihood ratios are multiplicative over unrelated pedigrees i, or log-likelihoods are additive. The base-10 log-likelihood ratio, or lod score is

$$\text{lod}(\rho) = \sum_i \log_{10}\left(\frac{L_i(\rho)}{L_i(\rho=\frac{1}{2})}\right)$$

where $L_i(\cdot)$ is the likelihood contributed by pedigree i. The maximized lod score is

$$\max_{0\le\rho\le\frac{1}{2}} (\mathrm{lod}(\rho)).$$

Of course, in combining over pedigrees, the maximizing ρ may be neither 0 nor $\frac{1}{2}$. In this case, the form of $f_2(\rho)$ is also relevant, not merely the value of f. Again, a useful measure of information for linkage analysis is the expected lod score or *Elod* (section 4.4):

$$\textit{Elod}(\rho) \quad = \quad \mathrm{E}_\rho(\mathrm{lod}(\rho)). \tag{4.8}$$

The *Elod* is additive over independent pedigrees. Each affected individual with inbreeding coefficient f has probability $f/(f+(1-f)q)$ of having two *ibd* genes at the disease locus. Hence, at $\rho = 0$, the contribution of each such affected individual to the *Elod* is

$$\frac{f}{f+(1-f)q}\sum_j p_j \log\left(\frac{f+(1-f)qp_j}{(f+(1-f)q)(f+(1-f)p_j)}\right) + \frac{(1-f)q}{f+(1-f)q}\log\left(\frac{(1-f)q}{f+(1-f)q}\right).$$

As $q \to 0$, this has limiting value

$$-\sum_j p_j \log(f+(1-f)p_j).$$

For example, for the affected offspring of first-cousin marriages ($f = 1/16$), and a polymorphic marker locus (for example, $p_j = 0.1$ for each of 10 alleles) the value is log(6.4). A small number of unrelated affected individuals all homozygous at the same polymorphic marker locus provides strong evidence for linkage.

Homozygosity mapping, and linkage analysis in general, can provide good evidence for linkage. With sufficient data, the loci determining simple Mendelian traits can be localized down to 1 cM ($\rho = 0.01$) (Boehnke, 1994). However, even with data at multiple linked loci, finer localization is normally impossible; there are insufficient informative meioses in the set of pedigrees to resolve loci that are too tightly linked. The above development of homozygosity mapping assumed, as do most linkage analyses, absence of allelic association between the trait and marker loci. However, most current copies of a rare recessive disease allele may trace to a single mutation, say on a haplotype carrying marker allele A_j. Then, as seen in section 4.2, at tight linkage, the allelic association between the loci decays slowly. In this case, not only will the majority of affected inbred individuals be homozygous at the marker locus, but most "unrelated" affected inbred individuals will be homozygous A_jA_j, due to remote coancestry of the disease alleles, not modeled by the analysis of the separate pedigrees. In effect, the analysis makes use of the absence of recombination at a large number of ancestral meioses from the original disease mutation to the current affected individuals. Such allelic associations have been used to assist in the fine-scale mapping of many rare recessive diseases including cystic fibrosis (Cox et al., 1989) and Werner's syndrome (Goddard et al., 1996).

4.7 Meiosis at multiple linked loci

We now introduce notation for a chromosome with L ordered loci, $1, \ldots, L$. For ease of notation, we assume that the loci are ordered $1, \ldots, L$ along the chromosome. We consider again the meiosis indicators of equation (1.2), and the vector notation of equation (1.3). Different meioses are independent, but the components of the meiosis indicator vector for meiosis i, $S_{i,\bullet} = (S_{i,1}, \ldots, S_{i,L})$, are dependent. Recall also the notation $S_{\bullet,j} = (S_{1,j}, \ldots, S_{M,j})$ for the set of all meiosis indicators on the pedigree, at locus j (equation (1.3)). Let the intervals between successive loci be $I_1, \ldots, I_{L-1}$. Let $R_j = 1$ if a gamete is recombinant on interval I_j, and $R_j = 0$ otherwise ($j = 1, \ldots, L-1$). Then, in a given meiosis i,

$$\begin{aligned} R_j &= 1 \text{ if } S_{i,j} \neq S_{i,j+1}, \text{ and} \\ R_j &= 0 \text{ if } S_{i,j} = S_{i,j+1}, \quad j = 1, \ldots, L-1. \end{aligned} \tag{4.9}$$

Each vector $(R_1, \ldots, R_{L-1})$ determines two equiprobable vectors $S_{i,\bullet} = (S_{i,1}, \ldots, S_{i,L})$. A model for $S_{i,\bullet}$ is equivalent to a model for $(R_1, \ldots, R_{L-1})$. One simple model for the distributions of $S_{i,\bullet}$ over more than two loci is considered in this section. More general models for $(R_1, \ldots, R_{L-1})$ will be considered in Chapter 5.

In considering the probability of data on related individuals in a pedigree (equation (3.9)):

$$L = \Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}). \tag{4.10}$$

Often (although not always), data observations will be specific to a given locus. For example, for DNA marker loci we observe phenotypes of individuals at given loci. Let $Y_{\bullet,j}$ denote the all data pertaining to locus j, so the full data pertaining to this chromosomal region is $\mathbf{Y} = (Y_{\bullet,1}, \ldots, Y_{\bullet,L})$, and

$$\Pr(\mathbf{Y} \mid \mathbf{S}) = \prod_j \Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j}))$$

where $\mathbf{J}(S_{\bullet,j})$ is the pattern of gene identity by descent among observed individuals, at locus j, which is determined by $S_{\bullet,j}$. Since meioses i are independent, equation (4.10) becomes

$$L = \Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \left(\prod_j \Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j})) \right) \left(\prod_i \Pr(S_{i,\bullet}) \right). \tag{4.11}$$

To proceed further, we need a model for the vector $S_{i,\bullet}$. Such models may derive from our model for the process of meiosis (Chapter 5) or may be based on computationally convenient assumptions. In either case, it is the binary meiosis indicators (1.2) which provide a means to trace the descent and ancestry of genes, at multiple linked loci. Just as for a single locus (section 3.6), they determine patterns of gene-identity-by-descent (gene *ibd*), which in turn determine patterns of phenotypic similarity among relatives.

The simplest models for meiosis assume *no interference*: this implies that the R_j are independent. Under this model, the dependence structure of the $S_{i,j}$ takes a simple form, with a first-order Markov property over loci j, and with meioses i being independent. The probability of any given indicator $S_{i,j}$ conditional on all the others, $\mathbf{S}_{-(i,j)} = \{S_{k,l}; (k,l) \neq (i,j)\}$, depends only on the indicators for the same meiosis and the two neighboring loci:

$$\begin{aligned} \Pr(S_{i,j} = s \mid \mathbf{S}_{-(i,j)}) &= \Pr(S_{i,j} = s \mid S_{i,j+1}, S_{i,j-1}) \\ &= \rho_{j-1}^{|s-S_{i,j-1}|}(1-\rho_{j-1})^{1-|s-S_{i,j-1}|} \\ &\qquad \rho_j^{|s-S_{i,j+1}|}(1-\rho_j)^{1-|s-S_{i,j+1}|} \end{aligned} \tag{4.12}$$

for $s = 0, 1$, where $\rho_j = \Pr(R_j = 1) = \Pr(S_{i,j} \neq S_{i,j+1})$ is the recombination frequency between locus j and locus $j+1$. Note that equation (4.12), is just counting the recombination/non-recombination events in intervals I_{j-1} and I_j, implied by the three indicators $(S_{i,j-1}, S_{i,j} = s, S_{i,j+1})$.

4.8 Multi-locus kinship and gene identity

Under the assumptions of conditional independence or absence of genetic interference, computation and Monte Carlo are, in principle, straightforward. The meiosis indicators, $\mathbf{S} = \{S_{i,j}\}$, are independent over meioses i, and are Markov over a sequence of loci j along a chromosome. The recursive equations for two-locus kinship generalize to the multilocus case, although becoming progressively more complicated. The probability of a *recombination pattern* in the intervals between marker loci is straightforward, being the product of the probabilities of recombination or non-recombination in successive intervals (equation (4.12)).

However, it is the resulting patterns of gene identity by descent among observed individuals that determine probabilities of observed data (equation (4.11)). Although the component $S_{i,j}$ are Markov over loci j, this is not usually so for the resulting patterns of gene *ibd*, $\mathbf{J}(S_{\bullet,j})$, among observed individuals. Different values of $S_{\bullet,j}$ may give rise to the same *ibd* pattern. Along the chromosome, the *ibd* process is an agglomeration of the $S_{\bullet,j}$ process. Grouping the states of a Markov chain does not, in general, produce a Markov chain.

As a specific example, consider again the pedigree of Figure 3.1, and suppose we are interested only in autozygosity of the final individual. Marginally at each locus the autozygosity probability is 7/64 or 0.1094 (section 3.2). Consider three loci, separated by a recombination frequencies of $\rho_1 = \rho_2 = 0.1$. The two-locus inbreeding coefficient of the final individual at recombination frequency 0.1 is 0.0566. This may be computed exactly by the recursive method outlined in section 4.5. Between the outer loci, in the absence of interference, the recombination frequency is

$$\begin{aligned} \rho &= \rho_1(1-\rho_2) + \rho_2(1-\rho_1) \\ &= 0.1 \times 0.9 + 0.1 \times 0.9 = 0.18. \end{aligned}$$

ibd state			Exact	True	Markov
N	N	N	$0.8915 - \delta$	0.7901	0.7881
N	N	I	$0.0183 + \delta$	0.0478	0.0497
N	I	N	$\delta - 0.0038$	0.0257	0.0255
N	I	I	$0.0566 - \delta$	0.0271	0.0273
I	N	N	$0.0183 + \delta$	0.0478	0.0497
I	N	I	$0.0345 - \delta$	0.0050	0.0031
I	I	N	$0.0566 - \delta$	0.0271	0.0273
I	I	I	δ	0.0295	0.0293

TABLE 4.6. *Prior autozygosity probabilities over three linked loci for the final individual of the pedigree of Figure 3.1*

At recombination frequency 0.18, the two-locus inbreeding coefficient of the final individual is 0.0345. These one- and two-locus values determine the three-locus probabilities up to one degree of freedom. We have

$$\begin{aligned}
\Pr(I\ N\ N) + \Pr(I\ N\ I) + \Pr(I\ I\ N) + \Pr(I\ I\ I) &= \Pr(I) = 0.1094 \\
\Pr(N\ I\ N) + \Pr(N\ I\ I) + \Pr(I\ I\ N) + \Pr(I\ I\ I) &= \Pr(I) = 0.1094 \\
\Pr(N\ N\ I) + \Pr(I\ N\ I) + \Pr(N\ I\ I) + \Pr(I\ I\ I) &= \Pr(I) = 0.1094.
\end{aligned}$$

Also, by symmetry, since $\rho_1 = \rho_2$,

$$\Pr(I\ I\ N) = \Pr(N\ I\ I) \text{ and } \Pr(N\ N\ I) = \Pr(I\ I\ N).$$

Then also

$$\begin{aligned}
\Pr(I\ I\ N) + \Pr(I\ I\ I) &= \Pr(I\ I\ N) + \Pr(I\ I\ I) \\
&= \Pr(I\ I; \rho = 0.1) = 0.0566 \\
\Pr(I\ N\ I) + \Pr(I\ I\ I) &= \Pr(I\ I; \rho = 0.18) = 0.345.
\end{aligned}$$

Fixing $\Pr(I\ I\ I) = \delta$, these equations determine all the probabilities, as given in the first column of Table 4.6, under the heading "exact". The values in the column labeled "true" are in fact obtained by Monte Carlo (section 3.7), using 10^8 independent realizations of genes on the pedigree, and are accurate to 10^{-4}. They are fully consistent with the exact probabilities. These probabilities may also be estimated using Markov chain Monte Carlo (Chapter 8). A comparison of the alternative Monte Carlo procedures in this example is given by Thompson (1994*a*).

The final column of Table 4.6 shows the probabilities that would be obtained, using the two-locus transition probabilities, and assuming the process to be first-order Markov. For this (assumed) Markov process of identity (I) and non-identity (N) the transition probabilities, and hence the three-locus probabilities, are determined as follows:

$$\Pr(I \to I) = 0.0566/0.1094 = 0.5174,$$

$$\begin{aligned} \Pr(I \to N) &= 1 - 0.5174 = 0.4826, \\ \Pr(N \to I) &= (0.1094 - 0.0566)/(1.0 - 0.1094) = 0.0593, \\ \text{and } \Pr(N \to N) &= 1 - 0.0593 = 0.9407. \end{aligned}$$

The resulting probabilities patterns of I and N over the three loci are shown in the final column of Table 4.6, labeled "Markov". None of the probabilities computed using the Markov assumption is completely accurate, but those having I at the second locus are close to Markov. The state I acts approximately (but not exactly) as a renewal state of the process. Proportionately, the probability that under the Markov assumption deviates most from the true value is that for the trio of states (I, N, I). Conditional on non-*ibd* at the center locus, the probability of I at the third locus is substantially increased by knowledge of state I at the first. The reason for this is that the states of $\mathbf{S}$ resulting in I are few and clustered in the total space of $\mathbf{S}$-values. For a fuller discussion of this see Thompson (1994*a*). The non-Markovian nature of I and N holds even for simpler pedigrees. It may seem that the differences in the probabilities are small, and substantial only for the state of very small probability. However, depending on the phenotypic data, states of low prior (pedigree) probability may have high probability conditional on the phenotypic data.

Chapter 5

Models for Meiosis

5.1 The meiosis process

In section 4.1, we introduced recombination as the process of crossing over between the two homologous parental chromosomes in the formation of an offspring gamete, and we have considered multilocus segregation probabilities under the assumption of no interference (section 4.7). In order to develop better models of multilocus segregation, it is necessary to consider the processes of *mitosis* and *meiosis* in greater detail. Mitosis is the normal process of cell division during somatic growth: meiosis is the process of gamete formation. Both processes involve chromosome duplication and separation, but only meiosis involves recombination. A chromosome is a doubled strand of helical DNA, with complementary bases on the two strands. Chromosomes of the shape often depicted in texts, or seen in an amniocentesis photograph, exist only just prior to mitosis or meiosis. These are actually doubled chromosomes. Each chromosome is thus two double strands of DNA. Each double-strand is known as a *chromatid*: the two chromatids of a single duplicate chromosome are known as *sister* chromatids. In the pair of chromosomes just prior to mitosis or meiosis, there are thus four chromatids, or eight strands of DNA in total. In our modeling here, we consider the four chromatids, or the *chromatid tetrad*, rather than all eight DNA strands.

Just after the previous mitotic division, each chromosome exists as a concentrated double-strand of DNA in the nucleus of the cell (Figure 5.1(a)). In the next stage, *interphase*, the chromosomes elongate (Figure 5.1(b)), and duplicate; at this stage the length of DNA in the nucleus of a cell is 2 meters. The DNA then re-concentrates to form the chromatid tetrad (Figure 5.1(c)). In mitosis, each chromosome divides to give two daughter cells (Figure 5.1(d)), each with a nucleus with the identical chromosome complement as the parent cell nucleus (Figure 5.1(a)). In the first meiotic division, however, one of each homologous pair of chromosomes must go to each daughter cell. In order to achieve this, the pair of chromosomes must become tightly aligned, and in so doing *chiasmata* occur, resulting in an exchange of DNA between two non-sister chromatids (Figure 5.1(e)).

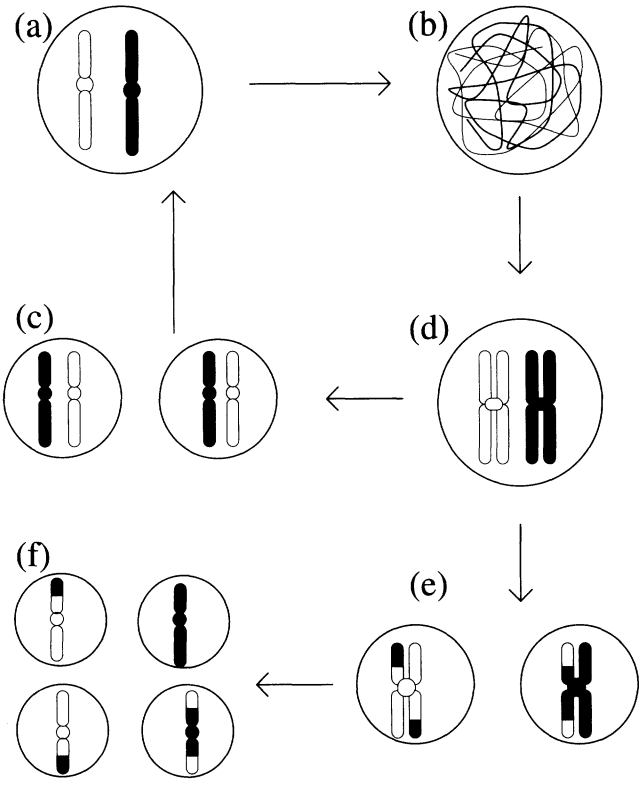

FIGURE 5.1. *The processes of mitosis and meiosis, shown for a single pair of homologous chromosomes in the nucleus of a cell of a diploid organism. See text for details*

The chromosomes separate; each daughter cell nucleus now contains only 50% of the DNA of the parent cell, but still in duplicate chromatid form (Figure 5.1(f)). Finally in the second meiotic division, in a process analogous to mitosis, these chromosomes divide, providing potential gamete cells (Figure 5.1(g)). Each potential gamete now contains 50% of the parental DNA, in the haploid form of one chromosome from each chromosome pair.

The crossover process is shown in more detail in Figure 5.2. Figure 5.2(a) shows the tetrad on which, in this example, two chiasmata are formed, and Figure 5.2(b) shows the four resulting gamete chromosomes. In mammalian organisms, for male meioses all four become gametes (sperm), while in female meioses three are discarded and one becomes a gamete (egg cell). However, only for certain non-mammalian species (such as fungi), or by carefully designed experiments (Hulten et al., 1990), is it possible to retrieve the four sperm from a given meiosis. In the analysis of data on an offspring individual, we observe only one paternal and one maternal meiotic product.

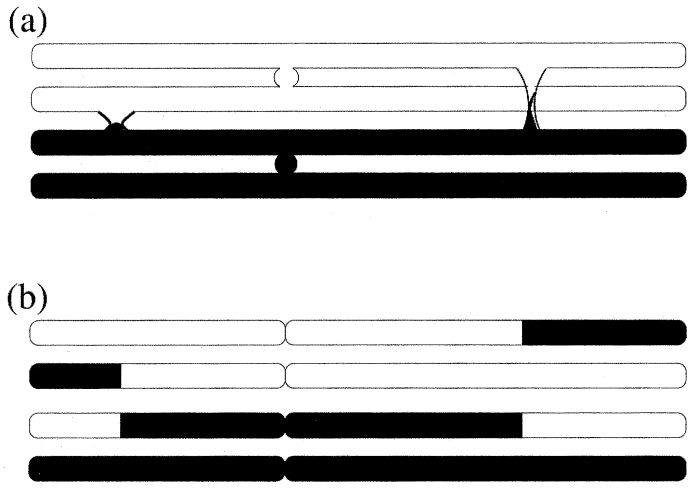

FIGURE 5.2. *The formation of chiasmata, and the crossovers resulting in the chromosomes of the four offspring gametes. The crossovers occurring are the same as in Figure 5.1(e)*

5.2 From chromatids to crossovers

Instead of modeling the crossover locations in a gamete (section 4.1), we now consider the occurrence of chiasmata locations at which crossovers between non-sister chromatids occur. Models for chiasmata formation are known as *four-strand* models, since the four chromatids are considered. Since each chiasma involves one paternal and one maternal chromatid, (paternal and maternal referring to the grandparental origins of the two homologous parental chromosomes as in equation (1.2)), each chiasma exists as a crossover in a resulting gamete with marginal probability $\frac{1}{2}$. Recall that the definition of genetic distance, provides for an expected one crossover per Morgan (section 4.1): this corresponds to an expectation of two chiasmata per Morgan, or one per 50 centiMorgans (cM).

Where only one meiotic product is observed, obtaining evidence for chromatid interference is practically impossible (Zhao et al., 1995) (but see also section 5.5). It is therefore often assumed that there is no *chromatid interference*: that is, that each chiasma involves two randomly chosen non-sister chromatids, independently of other the chromatids involved in other chiasmata. In this case, each chiasma results in a crossover in a given gamete, independently with probability $\frac{1}{2}$. Or the crossover process is just a thinned (probability= $\frac{1}{2}$) version of the chiasma process. Since a thinned Poisson process is also a Poisson process, this has no impact on the Haldane (1919) no-interference model. The chiasma process is Poisson, rate 2 per Morgan. The crossover process is Poisson, rate 1 per Morgan.

More generally, in a given chromosome interval of genetic length d, suppose there are $N(d)$ chiasmata, making now no assumptions about the probability distribution of N. If $N(d) = 0$, there are no chiasmata, no crossovers, and hence no recombination. For any non-zero value n of $N(d)$, in the absence of chromatid

interference, the probability of an odd number of crossovers is 1/2. (This is left as an exercise to the reader: it may be easier to think about tossing a fair coin n times, and the probability of an odd number of "heads".) Thus we have the formula of Mather (1938) for the recombination probability $\rho(d)$ at genetic distance d:

$$\rho(d) = \frac{1}{2}\Pr(N(d) > 0) = \frac{1}{2}(1 - \Pr(N(d) = 0)). \tag{5.1}$$

The only assumption here is the absence of chromatid interference: under this assumption $\rho(d)$ is an increasing function of d, and is bounded above by $\frac{1}{2}$. Note also that under Haldane's model $\Pr(N(d) = 0) = \exp(-2d)$, and Mather's formula applies (see equation (4.2)).

5.3 From chiasmata to recombination patterns

There is a multilocus version of Mather's formula (5.1). As in section 4.7, consider a chromosome with L ordered loci, $1, \ldots, L$, and label the intervals $I_1, \ldots, I_{L-1}$ and let $R_j = 1$ if a gamete is recombinant on interval I_j, and $R_j = 0$ otherwise ($j = 1, \ldots, L-1$). The recombination pattern is a function of the meiosis indicators $S_{i,j}$ for the given meiosis i, and provides a simpler representation for the current discussion:

$$\begin{aligned} R_j &= 0 \quad \text{if } S_{i,j} = S_{i,j+1} \\ R_j &= 1 \quad \text{if } S_{i,j} \neq S_{i,j+1} \end{aligned}$$

for $j = 1, \ldots, L-1$.

Now also let the (random) number of chiasmata in the intervals, in a meiosis, be $N_1, \ldots, N_{L-1}$. Let $C_j = 0$ if $N_j = 0$, and $C_j = 1$ otherwise ($j = 1, \ldots, L-1$): C_j is an indicator of presence of chiasmata in interval I_j. If $C_j = 0$, then $R_j = 0$. If $C_j = 1$, then $\Pr(R_j = 1) = \Pr(R_j = 0) = \frac{1}{2}$. In the absence of chromatid interference, the R_j are conditionally independent given C_j. Thus

$$\begin{aligned} \Pr((R_1, \ldots, R_{L-1}) = \mathbf{r}) &= \sum_{\mathbf{c} \geq \mathbf{r}} (\frac{1}{2})^{|\mathbf{c}|} \Pr((C_1, \ldots, C_{L-1}) = \mathbf{c}) \\ &= (\frac{1}{2})^{L-1} \sum_{\mathbf{c} \geq \mathbf{r}} 2^{|\mathbf{1}-\mathbf{c}|} \Pr((C_1, \ldots, C_{L-1}) = \mathbf{c}) \end{aligned} \tag{5.2}$$

where $|\mathbf{c}| = \sum_1^{L-1} c_j$ is the number of unit indicators in $\mathbf{c}$, and $\mathbf{1}$ is a vector of ones. This equation is (in essence) due to Weinstein (1936). Karlin and Liberman (1979) give a version in terms of the meiosis indicators rather than the recombination indicators. A recent discussion, using slightly different notation, is given by Speed (1996).

The estimation of chiasmata presence and absence patterns from recombination data provides another example of use of the EM algorithm. Consider again equation (5.2), and the estimation of patterns of chiasmata presence and absence,

from a sample of n completely observed patterns, $\mathbf{r}$, of recombination and non-recombination. An unconstrained estimate of $\Pr(\mathbf{R} = \mathbf{r})$ is $n(\mathbf{r})/n$ where $n(\mathbf{r})$ is the number of meioses exhibiting recombination patterns $\mathbf{r}$. However, if the equation

$$n(\mathbf{r}) = n \sum_{\mathbf{c} \geq \mathbf{r}} (\frac{1}{2})^{|\mathbf{c}|} \Pr((C_1, \ldots, C_{L-1}) = \mathbf{c}) \tag{5.3}$$

is inverted, negative values of $\Pr(\mathbf{C} = \mathbf{c})$ may result. An EM algorithm (section 2.4) avoids this, providing estimates of the probabilities of the underlying chiasmata presence/absence patterns, $q(\mathbf{c}) = \Pr((C_1, \ldots, C_{L-1}) = \mathbf{c})$, subject only to the constraint of no chromatid interference. In fact, this EM algorithm is very similar to that of section 4.2. There a phenotypic observation was partitioned in expectation among the possible multilocus genotypes (pairs of haplotypes) providing that phenotype. Here observation of a recombination pattern is subdivided among the chiasmata presence/absence patterns that could give rise to the recombination pattern:

$$\begin{aligned} \Pr((C_1, \ldots, C_{L-1}) = \mathbf{c} \mid \mathbf{r}) &= \frac{(\frac{1}{2})^{|\mathbf{c}|}\, q(\mathbf{c})}{\sum_{\mathbf{c}^* \geq \mathbf{r}} (\frac{1}{2})^{|\mathbf{c}^*|}\, q(\mathbf{c}^*)} \quad \text{if } \mathbf{c} \geq \mathbf{r} \\ &= 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

Thus, given current estimates $q(\mathbf{c})$ and the data counts $n(\mathbf{r})$, the conditional expected number of meioses exhibiting chiasmata pattern $\mathbf{c}$ is

$$\sum_{\mathbf{r} \leq \mathbf{c}} n(\mathbf{r}) \frac{(\frac{1}{2})^{|\mathbf{c}|}\, q(\mathbf{c})}{\sum_{\mathbf{c}^* \geq \mathbf{r}} (\frac{1}{2})^{|\mathbf{c}^*|}\, q(\mathbf{c}^*)}$$

and the new estimate is simply n^{-1} times this expected number. This EM algorithm, although very simply implemented, has poor convergence if there are many loci, or very tightly linked loci, since then many patterns $\mathbf{c}$ do not occur in the sample. Moreover, the resulting constrained MLEs differ from the inversion of (5.3) only when some $\Pr(\mathbf{C} = \mathbf{c})$ have MLE 0. In this case, unfortunately, convergence of the EM algorithm can be very slow. However, again as in the case of section 4.2, some frequencies $q(\mathbf{c})$ may be constrained to zero, and estimation of other chiasmata pattern frequencies continued in the subspace.

5.4 The chiasmata avoidance process

The vector $(C_1, \ldots, C_{L-1})$, specifies the avoidance and non-avoidance probabilities of the *chiasma process* on intervals of the chromosome. It is slightly neater, although of course equivalent, to express $\Pr((R_1, \ldots, R_{L-1}) = \mathbf{r})$ (or the probability of gametic types $\Pr(S_{i,\bullet})$) in terms of the avoidance probabilities alone, as in Mather's formula (5.1). We specify a subset $\mathcal{T}$ of the intervals $\{I_1, ..., I_{L-1}\}$ as follows. Let $t_j = 1$ if $I_j \in \mathcal{T}$, and $t_j = 0$ otherwise. Let $\phi_{\mathbf{t}}$ be the probability of no chiasmata

in $\mathcal{T}$. The set of $\phi_{\mathbf{t}}$, for all binary vectors $\mathbf{t}$ length $(L-1)$, is the set of avoidance probabilities of the chiasma process. If $t_j = 1$ there are no chiasmata in I_j, but if $t_j = 0$ the presence/absence of chiasmata in I_j is unspecified. There is thus a one-one relationship between the avoidance probabilities $\phi_{\mathbf{t}}$ and the presence/absence probabilities $\Pr(C_1, \ldots, C_{L-1})$:

$$\begin{aligned} \phi_{\mathbf{t}} &= \Pr(\text{no chiasmata in } \mathcal{T}) \\ &= \sum_{\mathbf{c} \leq (\mathbf{1}-\mathbf{t})} \Pr((C_1, \ldots, C_{L-1}) = \mathbf{c}). \end{aligned} \tag{5.4}$$

Lange (1997) derives an expression

$$\Pr((R_1, \ldots, R_{L-1}) = \mathbf{r}) = (\frac{1}{2})^{L-1} \sum_{\mathbf{t}} (-1)^{<\mathbf{r},\mathbf{t}>} \phi_{\mathbf{t}} \tag{5.5}$$

by a different method, again with notation differing slightly from ours. (Here, $<\mathbf{r},\mathbf{t}>$ is the inner product $\sum_j r_j t_j$.) Rather than deriving this equation directly, we use equation (5.4) to show that (5.2) and (5.5) are equivalent. Substituting (5.4) into (5.5) we obtain

$$\begin{aligned} \sum_{\mathbf{t}} (-1)^{<\mathbf{r},\mathbf{t}>} \phi_{\mathbf{t}} &= \sum_{\mathbf{t}} (-1)^{<\mathbf{r},\mathbf{t}>} \left(\sum_{\mathbf{c} \leq (\mathbf{1}-\mathbf{t})} \Pr(\mathbf{C} = \mathbf{c}) \right) \\ &= \sum_{\mathbf{c}} \left(\sum_{\mathbf{t} \leq (\mathbf{1}-\mathbf{c})} (-1)^{<\mathbf{r},\mathbf{t}>} \right) \Pr(\mathbf{C} = \mathbf{c}). \end{aligned}$$

Equating coefficients of $\Pr(\mathbf{C} = \mathbf{c})$ from (5.2), to complete the proof we need only show that for each $\mathbf{r}$ and $\mathbf{c}$

$$2^{|\mathbf{1}-\mathbf{c}|} I\{\mathbf{c} \geq \mathbf{r}\} = \sum_{\mathbf{t} \leq (\mathbf{1}-\mathbf{c})} (-1)^{<\mathbf{r},\mathbf{t}>}$$

where $I\{\mathbf{c} \geq \mathbf{r}\} = 1$ if $\mathbf{c} \geq \mathbf{r}$, and 0 otherwise. Consider first the case $\mathbf{c} \geq \mathbf{r}$. Then $r_j = 1 \Rightarrow c_j = 1 \Rightarrow t_j = 0$, so $<\mathbf{r},\mathbf{t}>= 0$ and we sum terms $(+1)$ over $2^{|\mathbf{1}-\mathbf{c}|}$ values of $\mathbf{t}$, confirming this case. Now consider any other $\mathbf{c}$, and consider any one component j for which $r_j = 1$ but $c_j = 0$. Thus $1 - c_j = 1$, and we sum over $t_j = 0$ and $t_j = 1$. For each set of values of the other $t_{j'}$, the two values of t_j give opposite signs to $(-1)^{<\mathbf{r},\mathbf{t}>}$. The coefficients cancel, and the overall coefficient is 0, as required. This completes the proof.

Given a model which determines either $\Pr(\mathbf{C} = \mathbf{c})$ or $\phi_{\mathbf{t}}$, exact computation of probabilities $\Pr(\mathbf{R} = \mathbf{r})$ of all patterns of recombination and non-recombination in a set of $L-1$ marker intervals is practical for L up to about 12. Two methods of likelihood evaluation under interference have been proposed: both rely on efficient computation of these probabilities. Weeks et al. (1993) provides an approach for models of count interference (see section 5.6), while Lin and Speed (1996) provides a method for the renewal process chi-square models of position interference

(section 5.7). For a fixed marker map, it is feasible to precompute and store these probabilities for up to about 12 markers ($2^{11} = 2048$). However, using these probabilities in any exact computation of a likelihood on a pedigree usually entails highly computationally intensive procedures, further limiting L and/or the pedigree sizes and structures that can be considered.

5.5 Chromatid interference

Where only one of the four gametic products of meiosis can be observed, it is hard to find evidence for chromatid interference. However, the non-negativity of probabilities $P(\mathbf{C} = \mathbf{c})$ in equation (5.2) does impose constraints on feasible recombination pattern probabilities $P(\mathbf{R} = \mathbf{r})$. Conversely, observed frequencies of patterns of recombination can provide evidence for the existence of chromatid interference. We consider now one specific constraint implied by Mather's formula, whose violation may provide evidence for chromatid interference (Fisher, 1948). Mather's formula (5.1) implies that recombination probabilities are an increasing function of genetic distance, bounded above by $\frac{1}{2}$. Under chromatid interference this is no longer so. Consider, in particular, the case of *complete positive chromatid interference*: in that case, successive chiasmata involve alternating disjoint pairs of non-sister chromatids. Then the recombination probability at genetic distance d is

$$\rho(d) = \frac{1}{2}\Pr(N(d) \text{ odd}) + \Pr(N(d) \text{ even but not divisible by } 4).$$

In the case when the chiasma process is a Poisson process rate 2, this becomes

$$\begin{aligned}\rho(d) &= \frac{1}{2}\exp(-2d)\left(\sum_{k=0}^{\infty}\left(\frac{(2d)^{2k+1}}{(2k+1)!} + 2\,\frac{(2d)^{4k+2}}{(4k+2)!}\right)\right) \\ &= \frac{1}{2}(1 - \exp(-2d)\,\cos(2d)).\end{aligned}$$

In this case, $\rho(d)$ is greater than $\frac{1}{2}$ at certain distances, and is not monotone. Fisher (1948) discusses possible evidence for $\rho(d) > \frac{1}{2}$ in the case of the pseudoautosomal region of the mammalian sex chromosomes in mice; Weinstein provides an interesting contribution to the discussion.

Another possibility is *complete negative chromatid interference*: in this case every chiasma on the tetrad involved the same pair of chromatids. Then half the gametes would show no recombination, but in the other potential two gametes from a meiosis every chiasma results in a crossover. Again, when chiasmata occur as a Poisson process rate 2,

$$\rho(d) = \frac{1}{2}\Pr(N(d) \text{ odd}) = \frac{1}{4}(1 - \exp(-4d)).$$

Note that when d is small, $\rho(d) \approx d$, as usual. However, at large genetic distances, only one half of the gametes show independent segregation, the other half apparently showing tight linkage. With multilocus data, such an extreme pattern of

recombination would be detectable. A less extreme pattern might simply be thought to be due to heterogeneity of recombination among meioses. Chromatid interference is very much confounded both with other forms of interference, and interference in general may be confounded with heterogeneity in recombination. For the remainder of this chapter we consider only models with no chromatid interference.

To ensure a biologically feasible interference model, a model of chiasma formation in the chromatid tetrad at meiosis is desirable. Under the no-interference model (Haldane, 1919), chiasmata, and hence crossovers, arise as a Poisson process; the count on a chromosome arm has a Poisson distribution, and conditionally on the count their positions are independently and uniformly distributed (all distributions being in terms of genetic, not physical, distance). Thus, in the absence of chromatid interference, there are, broadly, two classes of interference model: count interference and position interference.

5.6 Count-location models for chiasmata

In a count-location model, the count of chiasmata on a chromosome arm is no longer necessarily Poisson, but conditional upon the count, they are independently and uniformly distributed. In such models

$$\phi_{\mathbf{t}} \quad = \quad \phi(< \mathbf{t}, \mathbf{d} >)$$

where d_j is the genetic length of interval I_j. That is, the chiasma avoidance function depends only on the total length of chromosome avoided. Such models have been considered by Liberman and Karlin (1984), who call the corresponding map functions $\rho(d)$ *multilocus feasible.*

Suppose that the probability mass function of the total number of chiasmata N on a chromosome arm length Λ Morgans has probability generating function $g_N(\cdot)$. Then, given $N = n$, the probability of no chiasmata in length d is $(1 - d/\Lambda)^n$, and

$$\phi(d) \quad = \quad \sum_{n=0}^{\infty} \Pr(N = n)(1 - \frac{d}{\Lambda})^n \quad = \quad g_N(1 - d/\Lambda) \tag{5.6}$$

with corresponding map function, from Mather's formula,

$$\rho(d) \quad = \quad \frac{1}{2}(1 - \phi(d)) \quad = \quad \frac{1}{2}(1 - g_N(1 - d/\Lambda)). \tag{5.7}$$

Note that the expected number of chiasmata N in length Λ of chromosome is, by the definition of genetic length, 2Λ.

Consider now some simple examples:
(1) Suppose N has a Poisson distribution with mean 2Λ: $N \sim \mathcal{P}(2\Lambda)$
Then $g_N(w) = \mathrm{E}(w^N) = \exp(2\Lambda(w - 1))$ and from equation (5.6),

$$\phi(d) \quad = \quad \exp(2\Lambda(1 - \frac{d}{\Lambda} - 1)) \quad = \quad \exp(-2d)$$

and from equation (5.7) we have again the no-interference equation (4.2).

(2) Another tractable count-location model is given by assuming a fixed maximum number K of chiasmata on a chromosome, and that $N \sim \mathcal{B}(K, \frac{1}{2})$, with $2\Lambda = \mathrm{E}(N) = K/2$.
Then $g_N(w) = \mathrm{E}(w^N) = (\frac{1}{2}(1+z))^K$ and from equation (5.6),

$$\phi(d) \;=\; (\frac{1}{2}(2 - \frac{d}{\Lambda}))^K \;=\; (1 - \frac{d}{2\Lambda})^{4\Lambda}.$$

For large chromosomes, there is little interference: $\phi(d)$ becomes close to the non-interference value $\exp(-2d)$. On small chromosomes there is stronger interference. For example, if $\Lambda = \frac{1}{2}$, $\phi(d) = (1-d)^2$, $\rho(d) = d(1 - \frac{1}{2}d)$; the avoidance probability is smaller, and the recombination probability larger, than in the absence of interference.

(3) It appears to be be a biological reality, that for correct division of the chromosomes in the first meiotic division (Figure 5.1(d) to Figure 5.1(e)), each chromosome pair should have at least one chiasma. Note that under any such model $N \geq 1$ so that $\Lambda = \frac{1}{2}\mathrm{E}(N) \geq \frac{1}{2}$; in fact, even the smallest human autosomes have genetic length estimates just over 0.5 Morgans. One example of a model which incorporates this restriction is the truncated Poisson model, in which N has a Poisson distribution ($N \sim \mathcal{P}(\alpha)$) conditioned on $N \geq 1$. Then $2\Lambda = \mathrm{E}(N) = \alpha/(1 - \exp(-\alpha))$, and Λ is an increasing function of α, increasing from $\frac{1}{2}$ when $\alpha = 0$. Then

$$g_N(w) \;=\; \frac{\exp(\alpha(w-1)) - \exp(-\alpha)}{1 - \exp(-\alpha)} \quad \text{and} \quad \phi(d) \;=\; \frac{\exp(\alpha(1 - \frac{d}{\Lambda})) - 1}{\exp(\alpha) - 1}.$$

(4) An alternative model incorporating the restriction $N \geq 1$ is that due to Sturt (1976), in which N has a shifted, rather than truncated, Poisson distribution: $(N-1) \sim \mathcal{P}(2\Lambda - 1)$. Then

$$g_N(w) \;=\; w\exp((2\Lambda - 1)(w-1)) \quad \text{and} \quad \phi(d) \;=\; (1 - \frac{d}{\Lambda})\exp(-(2\Lambda - 1)\frac{d}{\Lambda}).$$

The Sturt model has been found to fit existing data well (Weeks et al., 1993).

All the models (2),(3) and (4) are close to the Haldane model on large chromosomes, but show different departures on small chromosomes. It is an unfortunate feature of count-location models that the recombination probability at genetic distance d is determined by the length of the chromosome and the distribution of N on the entire chromosome.

5.7 Renewal process models of chiasma formation

Although count-location models are convenient, mathematically, it is implausible that, given N, chiasmata are independently located. In particular, the consequence

that the chiasma avoidance function depends only on total length avoided is unrealistic. Consider two intervals, lengths d_1 and d_3 separated by an interval length d_2. Then, for a count location model, the probability

$$\phi_{(1,0,1)} = P(C_1 = 0, C_3 = 0) = \phi(d_1 + d_3)$$

and is independent of d_2. Position interference models allow for more general meiotic processes; we will consider only those where chiasmata arise as a stationary renewal process (Speed, 1996; Lange, 1997). This imposes certain restrictions on the map function $\rho(d)$, which are discussed by Speed (1996); subject to these restrictions, the renewal density is $-\rho''(d)$.

We consider briefly some examples: more details are given by Speed (1996) and references therein.

(1) Suppose chiasmata occur along the tetrad bundle as a Poisson process, rate 2, so that the interarrival time distribution is exponential with mean $\frac{1}{2}$, and has probability density function $2\exp(-2d)$. Integrating twice, and imposing the conditions $\rho(0) = 0$, and $\rho'(0) = 1$, we obtain again equation (4.2), confirming this interpretation of the no-interference model.

(2) Kosambi (1944) proposed a map function

$$\rho(d) = \frac{1}{2}\tanh(2d) = \frac{1}{2}\left(\frac{\exp(4d) - 1}{\exp(4d) + 1}\right)$$

which satisfies the conditions detailed by Speed (1996) and results in a renewal density

$$16\frac{(\exp(2d) - \exp(-2d))}{(\exp(2d) + \exp(-2d))^3}.$$

Although this map function is not *multilocus feasible* in the sense of Liberman and Karlin (1984), it has a valid interpretation as the result of a renewal process model for chiasmata. The renewal process class of models includes almost all of the map functions proposed in the literature, but not the Sturt map function.

(3) Although the Sturt count-location model has no renewal process analogue, the *truncated Poisson* distribution does (Browning, 1999). This shows that two quite different processes can lead to same map function (Speed, 1996). Further, Browning (1999) has shown that a zero-modified Poisson distribution is the *unique* model that is both a count-location and a stationary renewal-process. (This includes, of course, both the Poisson model and the truncated Poisson model.)

(4) A flexible and simple renewal-process model is the *chi-square model* (Zhao et al., 1995). The renewal density is a scaled $\chi^2_{2(m+1)}$, with the scaling $(4(m+1))^{-1}$ such that the expected inter-arrival distance is $\frac{1}{2}$. One interpretation of this model is that potential chiasmata occur as a Poisson process and that every $(m+1)^{\text{th}}$ such potential chiasma becomes an actual chiasma. These models fit data well

(Zhao et al., 1995), and have properties that make recombination probabilities over several loci, and hence likelihood computations on pedigrees, somewhat tractable (Lin and Speed, 1996). A generalization of the chi-squared model is the *Poisson-skip model* (Lange, 1997). In this case, the r th potential chiasmata becomes one with probability β_r. The renewal density is a mixture of chi-squared (χ^2) distributions, with the scaling of genetic distance again chosen such the mean inter-arrival time of the chiasma process on the tetrad is $\frac{1}{2}$.

Chapter 6

Likelihoods on Pedigrees

6.1 The Baum algorithm and "Peeling"

We review here the algorithm given by Baum (1972) for the computation of the likelihood in a hidden Markov model. The procedure is general to any stochastic system with discrete-valued latent variables $S_{\bullet,j}$ with a first-order Markov structure, and outputs $Y_{\bullet,j}$ depending only on $S_{\bullet,j}$. However, for convenience, we retain the notation of section 4.7 with meiosis indicators $S_{\bullet,j}$ and phenotypic data $Y_{\bullet,j}$ for locus j, with loci ordered $j = 1, \ldots, L$ along a chromosome. The dependence structure is shown in Figure 6.1. The Baum algorithm can proceed in either direction, and both formulations will be given. For closer analogy with pedigree peeling (section 6.3), we consider first the backwards computation, which is less natural for time series. On a pedigree, data are usually on the final generations. In time series or signal processing, on the other hand, data are observed forwards in time and prediction is often the question of interest.

For data observations $\mathbf{Y} = (Y_{\bullet,j}, j = 1, \ldots, L)$, we want to compute $\Pr(\mathbf{Y})$. Due to the first-order Markov dependence of the $S_{\bullet,j}$, equation (4.10) can be written

$$\begin{aligned} \Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{S}, \mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{L} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \prod_{j=1}^{L} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right). \end{aligned} \tag{6.1}$$

Now define

$$R_j(s) = \Pr(Y_{\bullet,k}, k = (j+1), \ldots, L \mid S_{\bullet,j} = s)$$

with $R_L(s) = 1$ for all s. The conditional independence structure (Figure 6.1), provides that $\{Y_{\bullet,k}, k = (j+1), \ldots, L\}$, $Y_{\bullet,j}$, and $S_{\bullet,j-1}$ are mutually independent given $S_{\bullet,j}$.

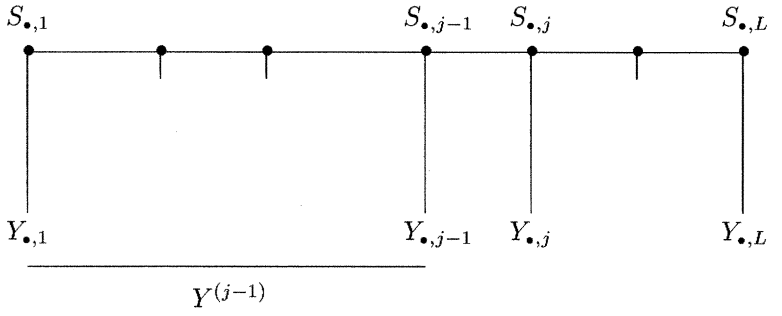

FIGURE 6.1. *The conditional independence structure of data, in the absence of genetic interference*

Thus,

$$R_{j-1}(s) = \sum_{s^*} [\Pr(S_{\bullet,j} = s^* \mid S_{\bullet,j-1} = s) \Pr(Y_{\bullet,j} \mid S_{\bullet,j} = s^*)\, R_j(s^*)] \tag{6.2}$$

for $j = 2, ..., L$, while at the final step

$$\Pr(\mathbf{Y}) = \sum_{s^*} [\Pr(S_{\bullet,1} = s^*) \Pr(Y_{\bullet,1} \mid S_{\bullet,1} = s^*) R_1(s^*)]$$

Thus the L-dimensional sum (6.1) may be computed as a telescoping series of one-dimensional sums over the possible values s^* of each $S_{\bullet,j}$ in turn, computed for each possible value s of $S_{\bullet,j-1}$. Where each $S_{\bullet,j}$ can take only a small number of possible values, this makes practical and feasible the computation, even for very large values of L. In fact, the computation is linear in L.

In the case of meiosis indicators, the direction along a chromosome is irrelevant and

$$\Pr(S_{\bullet,j} = s^* \mid S_{\bullet,j-1} = s) = \Pr(S_{\bullet,j-1} = s^* \mid S_{\bullet,j} = s)$$

However, in general only the forward transitions $\Pr(S_{\bullet,j} = s^* \mid S_{\bullet,j-1} = s)$ may be readily available. Even in this case, peeling in the direction from 1 to L is also possible. For convenience, we define $Y^{(j)} = (Y_{\bullet,1}, \ldots, Y_{\bullet,j})$, the data along the chromosome up to and including locus j. Note $\mathbf{Y} = Y^{(L)}$. Instead of the conditional probability

$$R_j(s) = \Pr(Y_{\bullet,k}, k = (j+1), \ldots, L) \mid S_{\bullet,j} = s)$$

it is now more convenient to define the joint probability

$$\begin{aligned} R_j^*(s) &= \Pr(Y_{\bullet,k}, k = 1, \ldots, j-1,\ S_{\bullet,j} = s) \\ &= \Pr(Y^{(j-1)},\ S_{\bullet,j} = s) \end{aligned}$$

with $R_1^*(s) = \Pr(S_{\bullet,1} = s)$. Now equation (6.2) is replaced by

$$R_{j+1}^*(s) = \sum_{s^*} [\Pr(S_{\bullet,j+1} = s \mid S_{\bullet,j} = s^*) \Pr(Y_{\bullet,j} \mid S_{\bullet,j} = s^*)\ R_j^*(s^*)] \tag{6.3}$$

for $j = 1, 2, \ldots, L-1$, with

$$\Pr(\mathbf{Y}) = \sum_{s^*} \Pr(Y_{\bullet,L} \mid S_{\bullet,L} = s^*)\ R_L^*(s^*).$$

We return to these equations in sections 6.2 and 6.4 in the context of likelihood computations on the basis of data observed on members of a pedigree. We note here only that efficient computation of the penetrance probabilities $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ (section 3.6) is key to the implementation.

6.2 Exact likelihoods for multiple markers

Exact likelihood computations on pedigrees rely on algorithms analogous to the Baum-type peeling algorithms of the previous section. One form in which the approach applies quite directly is the methods of Lander and Green (1987). The likelihood of equation (3.9) of section 3.6 is

$$L = \Pr(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))\ \Pr(\mathbf{S})$$

where $\mathbf{J}(\mathbf{S})$ is the gene ibd pattern among observed individuals determined by meiosis indicators (inheritance vectors), $\mathbf{S}$. Since the inheritance vectors $\mathbf{S} = \{S_{\bullet,j}\}$ (equation (1.2)) are first-order Markov over loci j, and the data $\mathbf{Y}$ typically partition into data Y_j relating to each locus j (see section 4.7), the likelihood takes the form equivalent to equation (4.11):

$$L = \sum_{\mathbf{S}} \left(\prod_j \Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j})) \right) \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{L} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \right),$$

which is directly analogous to equation (6.1) of section 6.1. Thus, either the forwards (equation (6.3)) or backwards (equation (6.2)) computation method can be applied.

Note however that this exact computation is limited to very small pedigrees. If there are m meioses on the pedigree, then $S_{\bullet,j}$ can take 2^m values, and in moving along the chromosome, we must consider transitions from the 2^m values of $S_{\bullet,j}$ to the 2^m values of $S_{\bullet,j+1}$. For a pedigree with n individuals, f of whom are founders, $m = 2n - 3f$. In practice we are limited to pedigrees where m is no more than 16. Additionally, for each locus j, and for each value of $S_{\bullet,j}$, we must compute $\Pr(Y_{\bullet,j} \mid \mathbf{J}(S_{\bullet,j}))$. For marker loci, computation is straightforward for given $S_{\bullet,j}$,

but again this limits the number of $S_{\bullet,j}$ that can be considered, and hence the size of the pedigree.

With data increasingly available at multiple linked marker loci, calculation of likelihoods using such data is desirable. While there may be uncertainties about marker locations, or other aspects of the marker model such as allele frequencies, these are normally assumed known. Rather than the linkage lod-scores of section 4.3, a *location score curve* is computed (Lathrop et al., 1984; Lange, 1997). This is equivalent to the curve of lod scores for linkage of the trait plotted as a function of hypothesized trait-locus location d against a fixed map of markers. Specifically, the map-specific lod score is $log_{10}(L(d)/L(\infty))$, where d is the hypothesized chromosomal location measured in genetic distance, and $d = \infty$ corresponds to $\rho = \frac{1}{2}$, or absence of linkage. The *location score* is defined as $2\log_e(L(d)/L(\infty))$. Under appropriate conditions, this statistic has approximately a chi-squared distribution in the absence of linkage (see section 2.2). Clearly, the location score is simply $2log_e(10)$ or about 4.6 times the map-specific lod score. In this book, we shall consider lod scores for gene location, rather than location scores. The location lod score curve differs from the linkage detection lod scores of section 4.3 in that the likelihood is considered as a function of trait locus position, and not maximized over this parameter. Other parameters of the trait model, such as penetrances or allele frequencies, may be assumed known, or may be maximized over to obtain a profile log-likelihood curve for the trait locus location. We return to location lod score curves in later chapters, noting here only that fast computation of many multipoint linkage likelihoods is needed to obtain such a curve.

Efficient methods using the algorithm of this section have been developed over the last few years by Kruglyak and co-workers. Kruglyak et al. (1995) show how to use the dependencies in the Markov transitions to reduce the computational burden from order $2^m \times 2^m$ to order $m2^m$, almost doubling the size of pedigree that can be considered. Kruglyak et al. (1996) give an algorithm for the efficient computation of the penetrance probabilities $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$: see section 3.6. Most recently, Kruglyak and Lander (1998) have used a discrete Fourier transform representation to achieve greater efficiencies. While these methods have greatly increased applicability of the algorithm, procedures are intrinsically exponential in pedigree size, and thus limited to pedigrees of moderate size. Moreover, increased efficiency comes at the expense of decreased flexibility. Use of parental symmetries restricts the programs to equal male and female genetic maps, and efficient computation is possible only where single-locus marker genotypes are observed without ambiguity or error.

6.3 Computations on large but simple pedigrees

In section 1.3, equation (1.5) gave the form of the probability of data observations on a pedigree:

$$\Pr(\mathbf{Y}) = \sum_{\mathbf{G}} \left(\prod_{\text{observed } i} \Pr(Y_i \mid G_i) \right) \Pr(\mathbf{G}).$$

This probability is the likelihood for the genetic model underlying the phenotypic data $\mathbf{Y}$. How is this likelihood to be computed? While each term of the product can be easily evaluated, the difficulty is in the sum over $\mathbf{G}$. On a very small pedigree it may be possible to enumerate all possible genotypic configurations $\mathbf{G}$, and to compute the sum directly. In other special cases it may be possible to use a recursive algorithm to compute the gene identity pattern probabilities in the observed individuals, and hence to compute the marginal probability $P(\mathbf{G})$ for these individuals alone. However, in general this is impractical. Independently, Hilden (1970), Elston and Stewart (1971), and Heuch and Li (1972) laid the foundations of the approach that has been widely used over the last 20 years, and has made it possible to compute likelihoods of genetic models given data on large pedigrees.

The approach formalized by Elston and Stewart (1971), for simple pedigrees, was a generalization of the backwards Baum algorithm (equation 6.2). The approach uses the approach of section 6.1 but generalized to pedigree structures, using individual genotypes as the latent variables. The summation proposed by Elston and Stewart (1971) was sequential, and used only the functions $R(\cdot)$, so that pedigree structures were limited to those where summation can proceed always up a pedigree. Hilden (1970) used joint probabilities, analogous to the functions $R^*(\cdot)$, and identified individual genes, so his procedure was, in principle, more general. The program of Heuch and Li (1972) was recursive, using functions both analogous to $R(\cdot)$ and to $R^*(\cdot)$, but was limited to simpler genetic models. The approach was generalized to arbitrary pedigree structures by Cannings et al. (1978), who gave it the name "peeling" and the functions $R(\cdot)$ and $R^*(\cdot)$ the name "R-functions". However, the idea of conditioning in this way when computing probabilities on pedigrees can be traced at least to Haldane and Smith (1947).

The basic idea is simply one of efficient sequential summation. The number of terms in which a specific G_i, the genotype of individual i, appears is limited to the penetrance term for that individual, and to segregation terms from the parents and to the offspring of individual i. Thus performing a summation over the possible values of G_i results in a function of (at worst) the genotypes of i's parents, spouses and offspring. Of course, this is only useful if implemented sensibly. By starting at the edges (top/bottom/side) of the pedigree, one limits the number of individuals whose genotypes must be considered jointly. For a pedigree without loops, there are (many) sequences of nuclear families such that each is connected to the as yet unprocessed part of the pedigree via a single individual, the *pivot*. In this case, summation over the non-pivot members of each family leads to a function of only the pivot genotypes, which may be incorporated into the summation for that individual in due course. This sequential summation process has come to be known as "peeling", and the specification of the order of individuals (normally of nuclear families) in which summation will be carried out as the "peeling sequence". We work through an example in detail in the following section.

The procedure is just the same for linked loci. The (multilocus) genotype of an individual is an unordered pair of multilocus haplotypes. That is, it is a specification of not only the single-locus genotypes, but also phase information. The segregation probabilities $\Pr(G_i|G_{M_i}, G_{F_i})$ are functions of the recombination fractions. If there are two diallelic loci, there are 4 haplotypes, and hence 10 genotypes; computation

is quite possible for a pedigree without loops. With more loci, or more alleles, computation rapidly becomes infeasible. The programs using this approach have greatly improved (Cottingham et al., 1993), and computer speed increases also. However, the algorithm is intrinsically constrained by the number of multilocus segregation probabilities $\Pr(G_i|G_{M_i}, G_{F_i})$, and hence depends on the cube of the number of possible genotypes per individual, which is exponential in the number of loci to be considered jointly.

6.4 Example of peeling a zero-loop pedigree

As an example of the peeling method of section 6.3, consider the pedigree of figure 6.2. This pedigree is a general zero-loop pedigree, in that it contains multiple founder couples and an individual with two spouses.

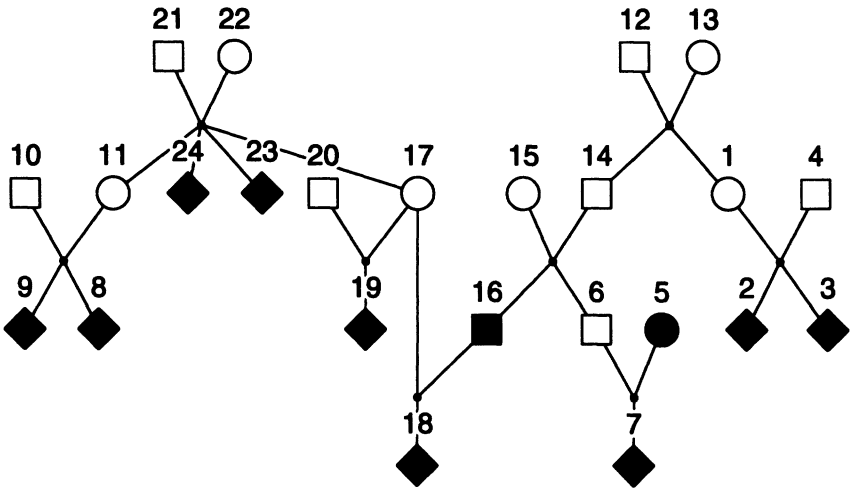

FIGURE 6.2. *Pedigree without loops. Shaded individuals are those for whom phenotypic data are assumed to be available*

Starting with the family to the right, we may compute

$$\begin{aligned} R_1(g) &= \Pr(Y_2, Y_3 \mid G_1 = g) \\ &= \sum_{g^*} \Pr(G_4 = g^*) \left(\sum_{g'} \Pr(Y_2|G_2 = g')\Pr(G_2 = g'|G_1 = g, G_4 = g^*) \right) \\ &\qquad \left(\sum_{g''} \Pr(Y_3|G_3 = g'')\Pr(G_2 = g''|G_1 = g, G_4 = g^*) \right) \end{aligned}$$

This is a generalized version of equation (6.2), where now there are two offspring nodes (2 and 3) and one parent node (4) to be summed over, whereas previously the structure was linear. Here the individual 1 is the *pivot* connecting this nuclear family to the remainder of the pedigree. Note that we do not need to include a term for the phenotypes of individual 4, since this individual is unobserved. Similarly for the family $\{5, 6, 7\}$, 6 is the pivot and

$$\begin{aligned} R_6(g) &= \Pr(Y_5, Y_7 \mid G_6 = g) \\ &= \sum_{g^*} \Pr(Y_5|G_5 = g^*)\Pr(G_5 = g^*) \\ &\qquad \left(\sum_{g'} \Pr(Y_7|G_7 = g')\Pr(G_7 = g'|G_6 = g, G_5 = g^*) \right). \end{aligned}$$

The other two peripheral families with a parent pivot may be handled similarly:

$$\begin{aligned} R_{11}(g) &= \Pr(Y_8, Y_9 \mid G_{11} = g) \\ &= \sum_{g^*} \Pr(G_{10} = g^*) \\ &\qquad \left(\sum_{g'} \Pr(Y_8|G_8 = g')\Pr(G_8 = g'|G_{11} = g, G_{10} = g^*) \right) \\ &\qquad \left(\sum_{g''} \Pr(Y_9|G_9 = g'')\Pr(G_9 = g''|G_{11} = g, G_{10} = g^*) \right) \end{aligned}$$

and

$$\begin{aligned} R_{17}^{(1)}(g) &= P(Y_{19}|G_{17} = g) \\ &= \sum_{g^*} \Pr(G_{20} = g^*) \\ &\qquad \left(\sum_{g'} \Pr(Y_{19}|G_{19} = g')\Pr(G_{19} = g'|G_{17} = g, G_{20} = g^*) \right). \end{aligned}$$

Note that for this last family, this is only a part of the information connecting to individual 17 via her offspring. The superscript indicates that only her first family (spouse 20 and offspring 19) is included. Individual 17's other family is not yet a peripheral family; it will be considered below. Where an individual is a parent in multiple families, the families may be considered separately; appropriate book-keeping must ensure that every term in equations (1.4) and (1.5) is entered once and once only.

Now no remaining peripheral family has a parent pivot. Thus, to proceed further across the pedigree, we must consider an R^*-function. For example, since $R_1(g)$

has been computed, the family $\{1, 12, 13, 14\}$ is now peripheral, and has pivot 14. First, summing conditionally upon the parents' genotypes,

$$\Pr(Y_2, Y_3 | G_{12} = g^*, G_{13} = g') = \sum_g \Pr(G_1 = g | G_{12} = g^*, G_{13} = g') R_1(g).$$

Then we may sum over these parental genotypes (G_{12}, G_{13}) to obtain

$$\begin{aligned} R^*_{14}(g) &= \Pr(Y_2, Y_3, G_{14} = g) \\ &= \sum_{g^*, g'} (\Pr(G_{12} = g^*) \Pr(G_{13} = g') \\ &\qquad \Pr(Y_2, Y_3 | G_{12} = g^*, G_{13} = g') \Pr(G_{14} = g | G_{12} = g^*, G_{13} = g')). \end{aligned} \tag{6.4}$$

Because this function is the probability of data connected to individuals 14 via his parents, we now have a joint probability of G_{14} rather than one conditional on G_{14}. However, the transition probabilities are still the downwards transition probabilities of offspring conditional upon parents. The terms are simply the relevant terms of equations (1.4) and (1.5). Note also that the data on this part of the pedigree remains (Y_2, Y_3); these are the only observed individuals in this part. Finally, note that, although the segment of pedigree is "above 14" in the sense of being connected to him through his parents, it includes his nephew and niece, 2 and 3.

At the next step, we combine the data on 2 and 3, with that on 5 and 7. First, conditional on parental genotypes (G_{14}, G_{15})

$$\Pr(Y_5, Y_7 \mid G_{14} = g^*, G_{15} = g') = \sum_g \Pr(G_6 = g | G_{14} = g^*, G_{15} = g') R_6(g).$$

Then, summing over (G_{14}, G_{15}) and including the probabilities computed in equation (6.4),

$$\begin{aligned} R^*_{16}(g) &= \Pr(Y_2, Y_3, Y_5, Y_7, G_{16} = g) \\ &= \sum_{g^*, g'} (R^*_{14}(g^*) \Pr(G_{15} = g') \Pr(Y_5, Y_7 \mid G_{14} = g^*, G_{15} = g')) . \end{aligned}$$

At this point, we again have a peripheral family, with a parent pivot, and we may include the data on 16 and 18 to obtain

$$\begin{aligned} R^{(2)}_{17}(g) &= \Pr(Y_2, Y_3, Y_5, Y_7, Y_{16}, Y_{18} \mid G_{17} = g) \\ &= \sum_{g^*} \Bigg(R^*_{16}(g^*) \Pr(Y_{16} | G_{16} = g^*) \\ &\qquad \sum_{g'} \Pr(Y_{18} | G_{18} = g') \Pr(G_{18} = g' | G_{17} = g, G_{15} = g^*) \Bigg) . \end{aligned}$$

The penetrance probability $\Pr(Y_{16} | G_{16} = g^*)$ is included only when individual 16 is to be summed out of the expression. This is just the convention we employ; the

important thing is that this term is included once and once only for each possible genotype of 16. In programming, where there are many zero penetrances, it may be desirable to incorporate the penetrance where an individual such as 16 is first encountered, since this will reduce the number of non-zero terms that must be carried forward. Note also that the individuals 2 and 3, who are not biologically related to 17 are "below" her, in the sense that the information their phenotypes provide on the genotype of 17, is through her offspring, 18. We may now combine the information from 17's two families:

$$\begin{aligned} R_{17}(g) &= \Pr(Y_2, Y_3, Y_5, Y_7, Y_{16}, Y_{18}, Y_{19} \mid G_{17} = g) \\ &= R_{17}^{(2)}(g) R_{17}^{(1)}(g) \end{aligned}$$

Now finally there is only one remaining family; any member of this family may serve as the final pivot. For example, with a parent pivot

$$\begin{aligned} R_{21}(g) &= \Pr(Y_2, Y_3, Y_5, Y_7, Y_8, Y_9, Y_{16}, Y_{18}, Y_{19}, Y_{23}, Y_{24} \mid G_{21} = g) \\ &= \sum_{g^*} \Bigg(\Pr(G_{22} = g^*) \\ &\qquad \left(\sum_{g'} \Pr(Y_{23} | G_{23} = g') \Pr(G_{23} = g' | G_{21} = g, G_{22} = g^*) \right) \\ &\qquad \left(\sum_{g'} \Pr(Y_{24} | G_{24} = g') \Pr(G_{24} = g" | G_{21} = g, G_{22} = g^*) \right) \\ &\qquad \left(\sum_{g'} R_{17}(g') \Pr(G_{17} = g' | G_{21} = g, G_{22} = g^*) \right) \\ &\qquad \left(\sum_{g'} R_{11}(g') \Pr(G_{11} = g' | G_{21} = g, G_{22} = g^*) \right) \Bigg) \end{aligned}$$

and finally the overall likelihood is

$$\Pr(Y_2, Y_3, Y_5, Y_7, Y_8, Y_9, Y_{16}, Y_{18}, Y_{19}, Y_{23}, Y_{24}) = \sum_g R_{21}(g) \Pr(G_{21} = g).$$

Into this final sum, all founder probabilities, all parent-pair to offspring transmission probabilities, and all penetrance probabilities for observed individuals have been included once and only once. Also, all R-functions computed in the course of the procedure have been included at a subsequent stage. Note also that each summation is over the genotypes of a single individual, and the maximum number of terms that must be computed at an intermediate stage is the number of possible ordered genotype pairs for a pair of parents. Even for simple pedigrees, peeling becomes infeasible if there are multiple loci with multiple alleles. The number of ordered pairs of genotypes to be considered can be too large, each genotype being an unordered pair of multilocus haplotypes.

6.5 Computations on complex pedigrees

The Elston-Stewart approach was generalized to complex pedigrees and more complex genetic models by Cannings et al. (1978; 1980). The Hilden (1970) approach also dealt, in principle, with arbitrarily complex pedigrees. For a pedigree with loops, functions on the genotypes of a *cutset* of individuals may have to be considered. This is a set of individuals who divide a processed segment of pedigree, from the unprocessed part. The processing therefore results in a function over the set of all possible genotype combinations for the individuals in the cutset. Even for a single autosomal diallelic locus, with 3 possible genotypes for each individuals, there are 3^n potential genotype combinations for n individuals. (In general, K^n, for K genotypes.) In this case, the objective of a good peeling sequence is to limit the cutset sizes as much as possible. Even so, on very complex pedigrees, with multiple intersecting loops, peeling becomes infeasible, particularly if there are more alleles, or more loci.

As an example, we outline a sequence of peeling operations to compute a likelihood on our standard example pedigree (figure 3.1), using the labeling of individuals of that figure. As in the case of a zero-loop pedigree, there are many alternative ways to work through a pedigree. Indeed, in principle summations may be done in any desired order. The order we give here is straightforward in that terms relating to a single whole marriage node are dealt with at each step. It is complicated, in that we traverse the pedigree partly upward and partly downward, to show the range of possibilities. For greater generality, we assume phenotypic data may be available on any of the individuals. We give the sequence of functions computed, but not the details of the equations. Within a given family the equations are of similar form to those of the previous section.

First we peel the final individual 531:

$$R_{432,431}(g_1, g_2) \quad = \quad \Pr(Y_{531} \mid G_{432} = g_1, G_{431} = g_2).$$

Next we might sum over the genotypes of individual 431 to obtain

$$R_{432,331,334}(g_1, g_2, g_3) \quad = \quad \Pr(Y_{531}, Y_{431} \mid G_{432} = g_1, G_{331} = g_2, G_{334} = g_3)$$

and then over 334 and her founder parent 235 to obtain

$$R_{432,331,233}(g_1, g_2, g_3) \quad = \\ \Pr(Y_{531}, Y_{431}, Y_{334}, Y_{235} \mid G_{432} = g_1, G_{331} = g_2, G_{233} = g_3).$$

At this point, there is no way to avoid a cutset of size four after the next step. The current members $\{432, 331, 233\}$ are offspring of three different nuclear families. To show the method, we choose to deal next with the founding family of the pedigree, so that 233 is replaced by her two siblings 231 and 232 in the cutset. The resulting function is in part conditional, and in part joint, since the section of the pedigree whose contribution to the likelihood has been computed connects to 432 and 331 through their offspring, but to 231 and 232 through their parents. Finally, since

the segment of pedigree analyzed is growing unwieldy, we introduce the notation $\mathbf{Y}_{\mathcal{D}}$ for the phenotypic data on a set of individuals $\mathcal{D}$. Then we have

$$R^*_{432,331,232,231}(g_1, g_2, g_3, g_4) = \Pr(\mathbf{Y}_{\mathcal{D}_1}, G_{232} = g_3, G_{231} = g_4 \mid G_{432} = g_1, G_{331} = g_2)$$

where $\mathcal{D}_1 = \{531, 431, 334, 235, 233, 131, 132\}$. Now since both 231 and 331 are in the cutset, we can reduce the cutset size by peeling the nuclear family of which they are both members, to obtain

$$R^*_{432,332,232}(g_1, g_2, g_3) = \Pr(\mathbf{Y}_{\mathcal{D}_2}, G_{332} = g_2, G_{232} = g_3 \mid G_{432} = g_1)$$

where $\mathcal{D}_2 = \mathcal{D}_1 \bigcup \{231, 331, 236\}$. Then

$$R^*_{432,332,333}(g_1, g_2, g_3) = \Pr(\mathbf{Y}_{\mathcal{D}_3}, G_{332} = g_2, G_{333} = g_3 \mid G_{432} = g_1)$$

where $\mathcal{D}_3 = \mathcal{D}_2 \bigcup \{234, 232\}$. Finally, incorporating the genotypic transmissions and phenotypic data on this 3-member nuclear family, and summing, we have the overall probability of all the data observed on the pedigree.

The scheme presented here, of peeling one nuclear family at a time, is a special case of more general procedures. Clearly, summations may be carried out in any order. Sometimes, it is more effective to peel several nuclear families simultaneously. Sometimes, some of the parent-pair offspring relationships within a family may be incorporated, leaving the others for later. Generally, whenever there is an R-function on two or more offspring of a nuclear family, it is efficient peel them, replacing them in the cutset by their two parents. It is also not necessary to peel by genotypes. Instead it can be more efficient to distinguish the maternal and paternal genes of individuals, and sum separately over these. This increases the number of genotypes, but can simplify the dependence structure of the data. Methods of gene-peeling were considered by Harbron and Thomas (1994) and by Harbron (1995). The simplification of the neighborhood structure due to considering genes rather than genotypes was shown in Figure 1.3.

6.6 Models with Gaussian random effects

We return briefly to the polygenic model of equations (2.15) and (2.16), introduced in section 2.6. Elston and Stewart (1971) noted that, since for a multivariate Gaussian distribution all marginal and conditional distributions are also Gaussian, and since a Gaussian form is specified by its mean and variance, the peeling process can also be used to compute the likelihood for σ_a^2, and for other parameters, such as an environmental variance σ_e^2. In this case, the sequential summation is just successive integration of latent additive genetic effects. Also, the inverse of the variance-covariance matrix $\mathbf{A}^{-1}$ of effects $\mathbf{z}$ is sparse, involving only terms for members within a nuclear family.

Additional Gaussian latent effects can be incorporated, for example effects of shared environment (Cannings et al., 1980). Also complex pedigrees are no

problem, in principle. In fact, the computational process is simpler than for discrete genotypes. In place of K^n discrete genotype combinations for n cutset individuals each of whom may have any of K genotypes, we now have a n-variate Gaussian distribution, specified by n means, and $n(n+1)/2$ covariance terms. The sequential Elston-Stewart summation method becomes a sequential integration of Gaussian densities.

A more general model for a quantitative trait is the *mixed model* (Morton and MacLean., 1974), which combines the Mendelian and polygenic models of section 2.6. The model for the quantitative phenotype, Y_i of individual i becomes

$$Y_i = \mu(G_i) + Z_i + \epsilon_i \tag{6.5}$$

where G_i is the genotype, and Z_i is the polygenic value (see equations (2.14) and (2.16)). The transmission model for Z_i is as in equation (2.15). Even for this simplest version of the mixed model, without other Gaussian or discrete components, peeling is infeasible. The overall likelihood is a mixture of multivariate Gaussian components, the number being the number of possible configurations of major genotypes $\mathbf{G}$ on the pedigree:

$$\begin{aligned} L = \Pr(\mathbf{Y}) &= \sum_{\mathbf{G}} \left(\Pr(\mathbf{G}) \int_{\mathbf{z}} \Pr(\mathbf{Y}|\mathbf{z},\mathbf{G}) dP_{\sigma_a^2}(\mathbf{z}) \right) \\ &= \int_{\mathbf{z}} \left(\sum_{\mathbf{G}} \Pr(\mathbf{Y}|\mathbf{z},\mathbf{G}) \Pr(\mathbf{G}) \right) dP_{\sigma_a^2}(\mathbf{z}) \end{aligned} \tag{6.6}$$

where $P_{\sigma_a^2}(\mathbf{z})$ is the multivariate Gaussian distribution of $\mathbf{z}$ (equation (2.15)). These forms for the likelihood show that for given $\mathbf{G}$ it is possible to integrate over $\mathbf{z}$ for the Gaussian form $\Pr(\mathbf{Y}|\mathbf{z},\mathbf{G})$, and that for given $\mathbf{z}$ it is possible to sum over $\mathbf{G}$ using the Elston-Stewart algorithm or its generalizations. A general discussion of propagation of probabilities on graphs, for both continuous and discrete latent variables, is given by Lauritzen (1992). It is of interest that the dependence structure of discrete and continuous variables of the genetic mixed model falls within the framework of Lauritzen (1992) for full exact computation of a likelihood. However, the pattern of dependence among the components of $\mathbf{Y}$, $\mathbf{G}$ and $\mathbf{z}$ means that, wherever data are observed on the pedigree, it may be necessary to compute separately the contribution from each component of the mixture of Gaussian distributions, one for each value of $\mathbf{G}$. Generally, in the context of data on extended pedigrees, it is impossible both to integrate over $\mathbf{z}$ and sum over $\mathbf{G}$ to obtain an exact value for the likelihood L. We return to this in section 9.4, where Monte Carlo methods of estimation of mixed model likelihoods are presented.

Chapter 7

Monte Carlo Estimates on Pedigrees

7.1 Baum algorithm for conditional probabilities

While the above method of likelihood computation was known to Baum (1972), his primary aim was estimation of the transition probabilities of the Markov chain, and of the probability relationship between input and output (Baum and Petrie, 1966; Baum et al., 1970). Here, these are transition probabilities $P(S_{\bullet,j+1} = s \mid S_{\bullet,j} = s^*)$ and penetrance probabilities $P(Y_{\bullet,j} \mid S_{\bullet,j})$. If the latent variables $\mathbf{S}$ were observed, the sufficient statistics for estimation of these transition and penetrance parameters would be simple functions of $\mathbf{Y}$ and $\mathbf{S}$. Thus, to estimate parameters of the model, for example by using an EM algorithm (Dempster et al., 1977), one must impute these functions of the underlying $\mathbf{S}$ conditional on $\mathbf{Y}$. Again, here we use the notation of meiosis indicators of section 4.7, but the framework is general to any hidden Markov model.

Thus, the forward-backward algorithms of Baum et al. (1970) address *inter alia* the computation of marginal probabilities

$$Q_j(s) = \Pr(S_{\bullet,j} = s \mid \mathbf{Y}),\ j = 1, \ldots, L.$$

We define two functions

$$\begin{aligned} Q_j^{\dagger}(s) &= \Pr(S_{\bullet,j} = s \mid Y^{(j)}) \\ Q_{j+1}^{*}(s) &= \Pr(S_{\bullet,j+1} = s \mid Y^{(j)}). \end{aligned}$$

The function $Q_j^{\dagger}(\cdot)$ provides the imputation of $S_{\bullet,j}$ given data $Y^{(j)}$ up to and including locus j, while $Q_{j+1}^{*}(\cdot)$ is the predictor of $S_{\bullet,j+1}$ also given $Y^{(j)} = (Y_{\bullet,1}, \ldots, Y_{\bullet,j})$.

Then $Q_1^{\dagger}(s) = \Pr(S_{\bullet,1} = s \mid Y_{\bullet,1})$,

$$
\begin{aligned}
Q^*_{j+1}(s) &= \Pr(S_{\bullet,j+1} = s \mid Y^{(j)}) \\
&= \sum_{s^*} \Pr(S_{\bullet,j+1} = s | S_{\bullet,j} = s^*)\, Q^\dagger_j(s^*) \qquad (7.1)
\end{aligned}
$$

and

$$
\begin{aligned}
Q^\dagger_{j+1}(s) &= \sum_{s^*} \Pr(S_{\bullet,j+1} = s,\ S_{\bullet,j} = s^* \mid Y^{(j+1)}) \\
&= \frac{\sum_{s^*} \Pr(S_{\bullet,j+1} = s,\ S_{\bullet,j} = s^*,\ Y_{\bullet,j+1} \mid Y^{(j)})}{\Pr(Y_{\bullet,j+1} \mid Y^{(j)})} \\
&\propto \sum_{s^*} \Pr(S_{\bullet,j+1} = s,\ S_{\bullet,j} = s^*,\ Y_{\bullet,j+1} \mid Y^{(j)}) \\
&= \sum_{s^*} \Big(\Pr(Y_{\bullet,j+1} | S_{\bullet,j+1} = s) \\
&\qquad \Pr(S_{\bullet,j+1} = s | S_{\bullet,j} = s^*)\, \Pr(S_{\bullet,j} = s^* \mid Y^{(j)}) \Big) \\
&= \Pr(Y_{\bullet,j+1} | S_{\bullet,j+1} = s) \sum_{s^*} \Big(\Pr(S_{\bullet,j+1} = s | S_{\bullet,j} = s^*)\, Q^\dagger_j(s^*) \Big). \qquad (7.2)
\end{aligned}
$$

Provided $S_{\bullet,j+1}$ takes only a limited number of values s, the probabilities may be normalized, giving each function $Q^\dagger_j(s)$, $j = 2, \ldots, L$, in turn, the final one being

$$
Q_L(s) = Q^\dagger_L(s) = \Pr(S_{\bullet,L} = s \mid \mathbf{Y}) \qquad (7.3)
$$

the desired distribution of $S_{\bullet,L}$ given $\mathbf{Y}$.

Now we may proceed backwards to obtain $Q_j(\cdot)$ for $j = L-1, \ldots, 3, 2, 1$:

$$
\begin{aligned}
\Pr(S_{\bullet,j-1} &= s,\ S_{\bullet,j} = s^* \mid \mathbf{Y}) \\
&= \Pr(S_{\bullet,j} = s^* \mid \mathbf{Y})\, \Pr(S_{\bullet,j-1} = s \mid S_{\bullet,j} = s^*, \mathbf{Y}) \\
&= Q_j(s^*)\, \Pr(S_{\bullet,j-1} = s \mid S_{\bullet,j} = s^*, Y^{(j-1)}) \\
&= \frac{Q_j(s^*) \Pr(S_{\bullet,j} = s^* \mid S_{\bullet,j-1} = s) \Pr(S_{\bullet,j-1} = s \mid Y^{(j-1)})}{\Pr(S_{\bullet,j} = s^* \mid Y^{(j-1)})} \\
&= Q_j(s^*)\, \Pr(S_{\bullet,j} = s^* \mid S_{\bullet,j-1} = s) Q^\dagger_{j-1}(s) / Q^*_j(s^*). \qquad (7.4)
\end{aligned}
$$

The second step uses conditional independence of $S_{\bullet,j-1}$ and $Y_{\bullet,j}, \ldots, Y_{\bullet,L}$ given $S_{\bullet,j}$, and the third is an application of Bayes Theorem, using the conditional independence of $S_{\bullet,j}$ and $Y^{(j-1)}$ given $S_{\bullet,j-1}$. Note that this backward step involves both the forward probability function $Q^\dagger_{j-1}(\cdot)$ of equation (7.2) and the predictive probability $Q^*_j(\cdot)$ of equation (7.1). Now the marginal probabilities $Q_{j-1}(s) = \Pr(S_{\bullet,j-1} = s \mid \mathbf{Y})$ are readily obtained by summing over s^*:

$$
\begin{aligned}
Q_{j-1}(s) &= \Pr(S_{\bullet,j-1} = s \mid \mathbf{Y}) \\
&= \sum_{s^*} \Pr(S_{\bullet,j-1} = s, S_{\bullet,j} = s^* \mid \mathbf{Y}) \qquad (7.5)
\end{aligned}
$$

In the context of time series, equation (7.1) is known as the predictor, and (7.2) as the filter, while the backward equations (7.4) is the smoother, incorporating all data $\mathbf{Y}$ into the imputation of each $S_{\bullet,j}$.

Finally, instead of computing the marginal distributions $Q_j(s)$, we may prefer a realization from the joint distribution $\Pr(\mathbf{S} \mid \mathbf{Y})$. The Baum algorithm provides this also. The forward computation is exactly as before (equation (7.2)). The backward computation is replaced by sampling. First, $S_{\bullet,L}$ is sampled from $Q_L(\cdot)$ (equation (7.3)). Then, similarly to equation (7.4), given a realization of $(S_{\bullet,j} = s^*, S_{\bullet,j+1}, \ldots, S_{\bullet,L})$, a straightforward application of Bayes Theorem gives

$$\begin{aligned} &\Pr(S_{\bullet,j-1} = s \mid S_{\bullet,j} = s^*, S_{\bullet,j+1}, \ldots, S_{\bullet,L}, \mathbf{Y}) \\ &\quad = \Pr(S_{\bullet,j-1} = s \mid S_{\bullet,j} = s^*, Y^{(j-1)}) \\ &\quad \propto \Pr(S_{\bullet,j} = s^* \mid S_{\bullet,j-1} = s) Q^{\dagger}_{j-1}(s) \end{aligned} \tag{7.6}$$

where proportionality is with respect to s. Normalizing these probabilities, we can realize $S_{\bullet,j-1}$. This is done for each $j = L, L-1, \ldots, 4, 3, 2$ in turn, providing an overall realization $\mathbf{S} = (S_{\bullet,1}, \ldots, S_{\bullet,L})$ from $\Pr(\mathbf{S} \mid \mathbf{Y})$.

7.2 An EM algorithm for map estimation

Suppose, as above we have L marker loci along a chromosome, with recombination frequencies $\rho_{m,j-1}$ and $\rho_{f,j-1}$ in male and female meioses, respectively, between locus $j-1$ and locus j. With data $\mathbf{Y}$ and latent variables $\mathbf{S}$ consider the complete-data log-likelihood

$$\begin{aligned} \log \Pr(\mathbf{S}, \mathbf{Y}) &= \log(\Pr(S_{\bullet,1})) + \sum_{j=2}^{L} \log(\Pr(S_{\bullet,j} \mid S_{\bullet,j-1})) \\ &\quad + \sum_{j=1}^{L} \log(\Pr(Y_{\bullet,j} \mid S_{\bullet,j})) \end{aligned} \tag{7.7}$$

(see equation 6.1). Now, in the absence of interference, the recombination probabilities $\rho_{m,j-1}$ and $\rho_{f,j-1}$ enter only into the term $\log(\Pr(S_{\bullet,j} \mid S_{\bullet,j-1}))$ which takes the form

$$\begin{aligned} \log(\Pr(S_{\bullet,j} \mid S_{\bullet,j-1})) &= R_{m,j-1} \log(\rho_{m,j-1}) + (M_m - R_{m,j-1}) \log(1 - \rho_{m,j-1}) \\ &\quad + R_{f,j-1} \log(\rho_{f,j-1}) + (M_f - R_{f,j-1}) \log(1 - \rho_{f,j-1}) \end{aligned}$$

where $R_{m,j-1} = \sum_{i \text{ male}} |S_{i,j} - S_{i,j-1}|$ is the number of recombinations in interval $(j-1, j)$ in male meioses, and M_m is the total number of male meioses scored in the pedigree. The recombination counts $R_{f,j-1}$, for $j = 2, \ldots, L$, and total meioses M_f are similarly defined for the female meioses. Thus computation of the expected complete-data log-likelihood requires only computation of

$$\begin{aligned} \tilde{R}_{m,j-1} &= \mathrm{E}(R_{m,j-1} \mid \mathbf{Y}) \\ &= \sum_{i \text{ male}} \mathrm{E}(|S_{i,j} - S_{i,j-1}|) \end{aligned}$$

and similarly $\tilde{R}_{f,j-1}$, which are easily computed from equation (7.4). Since this is a simple binomial log-likelihood, the M-step sets the new estimate of $\rho_{m,j-1}$ to $\tilde{R}_{m,j-1}/M_m$, and similarly for all intervals $j = 2, 3, \ldots, L$ and for both the male and female meioses. The EM algorithm is thus readily implemented to provide estimates of recombination frequencies for all intervals and for both sexes.

An alternative is Monte-Carlo EM. Instead of computing the bivariate distributions of $(S_{\bullet,j-1}, S_{\bullet,j})$ (equation (7.4)), N realizations of $\mathbf{S}$, $\{\mathbf{S}^{(\tau)}; \tau = 1, \ldots, N\}$, are obtained from the conditional distribution of $\Pr(\mathbf{S} \mid \mathbf{Y})$ under the current parameter values, as described above (equations (7.3) and (7.6)). These are scored exactly as above:

$$R^{(\tau)}_{m,j-1} = \sum_{i \text{ male}} |S^{(\tau)}_{i,j} - S^{(\tau)}_{i,j-1}|.$$

A Monte Carlo estimate of $\tilde{R}_{m,j-1}$ is $\sum_{\tau=1}^{N} R^{(\tau)}_{m,j-1}/N$, and the new estimate of $\rho_{m,j-1}$ is $\tilde{R}_{m,j-1}/M_m$ as before, again with analogous formulae for all intervals and both sexes. This Monte Carlo EM is readily implemented, and, like many Monte Carlo EM procedures, performs as well as the deterministic version. Initially, the Monte Carlo sample size N need not be large, although for the final EM steps it should be increased. We return to Monte Carlo EM in section 9.3.

7.3 Importance sampling for likelihoods

The primary aim in computation of $\Pr(\mathbf{Y})$ on a pedigree is normally segregation or linkage analysis. For segregation analysis, or for linkage analyses where trait loci are explicitly modeled, computations using the Elston-Stewart framework is more straightforward, but computations are then limited to few loci, and to relatively simple pedigrees. For computations for multiple markers, the Lander-Green paradigm is more natural and more effective, but is limited to small pedigrees. Despite increasing computational power, the feasibility of exact computations on pedigrees remains limited. A pedigree may often be too large for computation of the likelihood using the methods of section 6.2, there may be too many linked loci for the method of section 6.3, or the pedigree may be too complex for the methods of section 6.5. Where exact computation is infeasible, Monte Carlo estimation (section 3.7) offers an alternative.

Given phenotypic data $\mathbf{Y}$ on a pedigree, the likelihood for parameters θ specifying a genetic model can be written

$$L(\theta) = P_\theta(\mathbf{Y}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X})\, P_\theta(\mathbf{X}) \tag{7.8}$$

where $\mathbf{X}$ are latent variables, either the genotypes $\mathbf{G}$ or the meiosis indicators $\mathbf{S}$. Thus

$$L(\theta) = \mathrm{E}_\theta(P_\theta(\mathbf{Y} \mid \mathbf{X})). \tag{7.9}$$

This is the form given by Ott (1979), and in principle we could estimate $L(\theta)$ by simulating $\mathbf{X}$ from the prior genotype distribution under model θ and averaging the value of the penetrance probabilities $P_\theta(\mathbf{Y} \mid \mathbf{X})$ for the realized values of $\mathbf{X}$. This does not work well, except on very small pedigrees, since the realized $\mathbf{X}$ are almost certain to be inconsistent with data $\mathbf{Y}$, or at best to make infinitesimal contribution to the likelihood.

Of course, realizations may be made from any distribution $P^*(\mathbf{X})$ (equation (3.12)):

$$L(\theta) = \mathrm{E}_{P^*}\left(\frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P^*(\mathbf{X})}\right) \tag{7.10}$$

provided (equation (3.13))

$$P^*(\mathbf{X}) > 0 \quad \text{if} \quad P_\theta(\mathbf{X}, \mathbf{Y}) > 0. \tag{7.11}$$

An advantage of this approach is that a single set of realizations from $P^*(\mathbf{X})$ will provide a Monte Carlo estimate of $L(\theta)$ over a range of models θ. That is, one set of realizations provides an estimate of the likelihood function, not only the likelihood at a single point. The first use of Monte Carlo likelihood function estimation in the context of pedigree analysis is due to K. Lange in Ott (1979). In this case, $P^*(\mathbf{X})$ was taken to be $P_{\theta_0}(\mathbf{X})$. However, this is no more effective than the original form (7.9). Again almost all realizations may be incompatible with $\mathbf{Y}$ or provide only infinitesimal contributions to the likelihood.

Recall again the brief discussion of importance sampling in section 3.7. In addition to the requirement (7.11), one must be able to realize from the distribution $P^*(\mathbf{X})$, and one must be able to evaluate $P^*(\mathbf{x})$ at the realized values $\mathbf{x}$ in order to compute the estimate. Finally, in order to reduce the Monte Carlo variance (section 3.7), $P^*(\mathbf{X})$ should be approximately proportional to the summand $P_\theta(\mathbf{X}, \mathbf{Y})$. In order to meet this requirement note:

$$P_\theta(\mathbf{X} \mid \mathbf{Y}) \propto P_\theta(\mathbf{X}, \mathbf{Y}). \tag{7.12}$$

However, simulation from $P_\theta(\mathbf{X} \mid \mathbf{Y})$ would be useless, even if possible, since we must also be able to evaluate it in our Monte Carlo estimate, and to evaluate it we need to know the denominator $P_\theta(\mathbf{Y})$, which is what we are trying to estimate. One alternative is to realize from a distribution close to $P_\theta(\mathbf{X} \mid \mathbf{Y})$, which can be evaluated.

A disadvantage of the likelihood function estimation approach (7.10) is that the range of models for which this estimation is effective is likely to be small, given the requirement that the single $P^*(\mathbf{X})$ must be approximately proportional to all the $P_\theta(\mathbf{X}, \mathbf{Y})$.

7.4 Risk probabilities and reverse peeling

In analyses of data on a pedigree, under a model indexed by known values of the parameters θ, quantities of interest include the conditional genotype probabilities

$P_\theta(G_{i,\bullet}|\mathbf{Y})$ for individuals i. These probabilities are known as *risk probabilities*, since the genotypes of interest are often those conferring a disease risk. In sections 6.1 and 7.1 we saw that, for a first-order Markov structure for latent variables $S_{\bullet,j}$, sequential computation of the likelihood $P_\theta(\mathbf{Y})$ using the functions

$$R_j(s) = P_\theta(Y_{\bullet,k}, k = (j+1), \ldots, L \mid S_{\bullet,j} = s)$$

had the same computational complexity as computation of conditional probabilities $Q_j(s) = \Pr(S_{\bullet,j} = s \mid \mathbf{Y})$ using the two functions

$$\begin{aligned} Q_j^\dagger(s) &= \Pr(S_{\bullet,j} = s \mid Y_{\bullet,1}, \ldots, Y_{\bullet,j}) \text{ and} \\ Q_{j+1}^*(s) &= \Pr(S_{\bullet,j+1} = s \mid Y_{\bullet,1}, \ldots, Y_{\bullet,j}) \end{aligned}$$

The latter computation requires two passes along the chromosome (forward and backward), while the likelihood computation requires only one (forward or backward), but in both cases the computation is of order $4^m L$ where m is the number of meioses in the pedigree.

The same applies to latent variables $G_{i,\bullet}$ on a pedigree structure. If $P_\theta(\mathbf{Y})$ can be computed, using the peeling method outlined in section 6.3, so also can the risk probabilities $P_\theta(G_{i,\bullet} \mid \mathbf{Y})$. This can be done by taking each individual i in turn, as the final individual L in a peeling sequence (equation (7.3)). However, it is more effectively accomplished by saving, for each possible value g of $G_{i,\bullet}$, the probabilities $R_i(g)$ (equation (6.2)), obtained in peeling up the pedigree. These probabilities are then combined with the functions $R_i^*(g)$ (equation (6.3)) obtained by progressing back down the pedigree. For example, if individual i divides the pedigree into two parts, the set $D(i)$ connected through his spouses and offspring, and the set $A(i)$ connected through his parents (including his siblings and their descendants), then in proceeding up the pedigree

$$R_i(g) = P_\theta(\mathbf{Y}_{D(i)} \mid G_{i,\bullet} = g)$$

while in proceeding down, relative to individual i,

$$R_i^*(g) = P_\theta(\mathbf{Y}_{A(i)}, G_{i,\bullet} = g)$$

so that

$$P_\theta(G_{i,\bullet} = g \mid \mathbf{Y}) \propto P_\theta(Y_{i,\bullet} \mid G_{i,\bullet} = g) R_i(g) R_i^*(g)$$

and these probabilities may be normalized to give the required probabilities $P_\theta(G_{i,\bullet} \mid \mathbf{Y})$. This procedure of working back down the pedigree to obtain risk probabilities is sometimes known as *reverse peeling*. In the case where peeling always up the pedigree is computationally feasible, all risk probabilities on a large pedigree can be computed in two passes through the pedigree. Even on a complex pedigree, with multiple interconnecting loops, few passes through the pedigree are required to obtain all the marginal (over individuals i) conditional (on $\mathbf{Y}$) risk probabilities (Thompson, 1981).

7.5 Elods and SIMLINK

Simulation of data random variables $\mathbf{Y}$ is often undertaken as part of a power study. For example, simulation of latent genotypes $\mathbf{G}$ and resulting marker and trait phenotypes $\mathbf{Y}$ can be used to assess the power of a potential linkage study. Before the times of readily available genome-wide marker data, linkage detection was primarily a question analyzing the coinheritance of observed trait phenotypes $\mathbf{Y}_T$ and marker locus phenotypes $\mathbf{Y}_M$, for a single trait locus, T, and single marker locus,M. If the two loci are linked, the recombination frequency is $\rho < \frac{1}{2}$, while if they are unlinked inheritance is independent at the two loci ($\rho = \frac{1}{2}$). Thus we have the *lod score* (Morton, 1955);

$$\text{lod}(\rho) = \log\left(\frac{P_\rho(\mathbf{Y}_M, \mathbf{Y}_T)}{P_{\rho=\frac{1}{2}}(\mathbf{Y}_M, \mathbf{Y}_T)}\right) \tag{7.13}$$

which is the logarithm of the likelihood ratio comparing the two hypotheses (see equation (4.3)). The expected lod score is then

$$\begin{aligned} \textit{Elod}(\rho) &= \mathrm{E}_\rho(\log(P_\rho(\mathbf{Y}_M, \mathbf{Y}_T)) - \log(P_{\rho=\frac{1}{2}}(\mathbf{Y}_M, \mathbf{Y}_T))) \\ &= \mathrm{E}_\rho(\log(P_\rho(\mathbf{Y}_M, \mathbf{Y}_T)) - \log(P(\mathbf{Y}_M)) - \log(P(\mathbf{Y}_T))). \end{aligned} \tag{7.14}$$

In advance of a study, one may compute the expected lod score to be obtained, given the sizes and counts of pedigree structures available, as was previously done in the case of homozygosity mapping (equation (4.8)). As discussed in section 4.4, if base-e lod scores are used, *Elod*ρ is also the Kullback-Leibler information $K(\rho = \frac{1}{2}; \rho)$ for testing $\rho = \frac{1}{2}$ when the true value of the recombination frequency is ρ. Thompson et al. (1978) first developed these *Elod*s in the context of linkage analysis, and they have become quite widely used (Ott, 1999). In fact, Thompson et al. (1978) produced Monte Carlo estimates of the expectation in equation (7.14), by simulating the underlying trait and marker genotypes from $P_\rho(\mathbf{G}_M, \mathbf{G}_T)$, and then the associated phenotypes, and then computing the lod score (7.13) for each realized set of phenotypes.

As data at multiple DNA markers became potentially available, there was a rush to map Mendelian traits, using previously collected trait data. The *Elod* became an important tool in assessing whether there were sufficient trait data for probable linkage detection if the marker typing were to be undertaken. One problem in using the *Elod* (7.14) is that the expectation is over both trait and marker phenotypes. Normally, however, there was already information on the trait phenotypes $\mathbf{Y}_T$ that would be available to researchers. Ploughman and Boehnke (1989) addressed this case. Given a single-locus trait model, and trait data $\mathbf{Y}_T$, it is possible to simulate the underlying inheritance patterns or genotypes, $\mathbf{G}_T$, at the trait locus. This is accomplished by a Monte Carlo version of reverse peeling (section 7.4) analogous to that given by equation (7.6) in section 7.1. Once trait genotypes $\mathbf{G}_T$ are realized, conditional on the available trait data $\mathbf{Y}_T$, marker latent genotypes $\mathbf{G}_M$ and potentially observable marker phenotypes $\mathbf{Y}_M$ are readily obtained:

$$P_\rho(\mathbf{Y}_M, \mathbf{G}_M, \mathbf{G}_T \mid \mathbf{Y}_T) = P(\mathbf{Y}_M \mid \mathbf{G}_M) P_\rho(\mathbf{G}_M \mid \mathbf{G}_T) \quad (7.15)$$
$$P(\mathbf{G}_T \mid \mathbf{Y}_T)$$

The dependence structure here is a special case of that shown in Figure 6.1. The combined realizations $(\mathbf{Y}_T, \mathbf{Y}_M)$ may be used to estimate a *Elod*, conditional upon the fixed $\mathbf{Y}_T$. These conditional *Elod*s became an essential tool in applied studies, particularly during the 1980s when many Mendelian traits were mapped, and marker typing remained the most expensive component of studies.

7.6 Sequential imputation

We turn now to a use of reverse peeling (section 7.4) in the Monte Carlo estimation of likelihoods (section 7.3). Recall that efficient Monte Carlo estimation of the likelihood $L(\theta) = P_\theta(\mathbf{Y})$ will result from sampling latent genotypes $\mathbf{G}$ from a distribution $P^*(\mathbf{G})$ close to proportional to the joint probability $P_\theta(\mathbf{G}, \mathbf{Y})$

$$P^*(\mathbf{G}) \approx P_\theta(\mathbf{G} \mid \mathbf{Y}) \propto P_\theta(\mathbf{G}, \mathbf{Y})$$

(equation (7.12)). The following approach to choice of $P^*(\mathbf{G})$ is due to Kong et al. (1994) and Irwin et al. (1994).

Suppose, as before, there are data at L genetic loci (say a disease and $L-1$ markers) on a chromosome, and assume absence of genetic interference. Let $Y_{\bullet,j}$ again denote the data for locus j and $G_{\bullet,j}$ the underlying genotypes at that locus for all members of the pedigree. Note that, provided paternal and maternal alleles are distinguished, genotypes $G_{\bullet,j}$ satisfy the same first-order Markov dependence over loci as do the meiosis indicators $S_{\bullet,j}$ (Figure 6.1). For any specified θ_0 of interest, a realization $G^*_{\bullet,j}$ is obtained for each locus in turn from the distribution

$$\begin{aligned} P^*(G_{\bullet,j}) &= P_{\theta_0}(G_{\bullet,j} \mid G^{*(j-1)}, Y^{(j)}) \\ &= P_{\theta_0}(G_{\bullet,j} \mid G^*_{\bullet,1}, \ldots G^*_{\bullet,j-1}, Y_{\bullet,1}, \ldots, Y_{\bullet,j-1}, Y_{\bullet,j}) \\ &= P_{\theta_0}(G_{\bullet,j} \mid G^*_{\bullet,j-1}, Y_{\bullet,j}) \end{aligned}$$

where as in section 6.1, $Y^{(j)} = (Y_{\bullet,1}, \ldots, Y_{\bullet,j})$, $G^{(j)}$ is analogously defined, and θ_0 indexes the genetic model. Predictive weights w_j are also computed:

$$w_j = P_{\theta_0}(Y_{\bullet,j} \mid Y^{(j-1)}, G^{*(j-1)}) = P_{\theta_0}(Y_{\bullet,j} \mid G^*_{\bullet,j-1}).$$

Due to the conditional independence structure, each of the realizations of $G_{\bullet,j}$ and each computation of w_j is computationally equivalent to a single-locus peeling computation analogous to those of section 7.4.
Now

$$\begin{aligned} P_{\theta_0}(G_{\bullet,j} \mid G^{*(j-1)}, Y^{(j)}) &= \frac{P_{\theta_0}(G_{\bullet,j}, Y_{\bullet,j} \mid G^{*(j-1)}, Y^{(j-1)})}{P_{\theta_0}(Y_{\bullet,j} \mid G^{*(j-1)}, Y^{(j-1)})} \\ &= \frac{P_{\theta_0}(G_{\bullet,j}, Y_{\bullet,j} \mid G^{*(j-1)}, Y^{(j-1)})}{w_j}. \end{aligned}$$

Thus the joint simulation distribution for $\mathbf{G}^* = (G^*_{\bullet,1}, ..., G^*_{\bullet,L})$ is

$$P^*(\mathbf{G}^*) = \prod_{j=1}^{L} P_{\theta_0}(G_{\bullet,j} \mid G^{*(j-1)}, Y^{(j)}) = \frac{P_{\theta_0}(\mathbf{G}^*, \mathbf{Y})}{W_L(\mathbf{G}^*)}$$

where $W_L(\mathbf{G}^*) = \prod_{j=1}^{L} w_j$. Thus

$$\begin{aligned} \mathrm{E}_{P^*}(W_L(\mathbf{G}^*)) &= \sum_{\mathbf{G}^*} W_L(\mathbf{G}^*) P^*(\mathbf{G}^*) \\ &= \sum_{\mathbf{G}^*} P_{\theta_0}(\mathbf{G}^*, \mathbf{Y}) = P_{\theta_0}(\mathbf{Y}). \end{aligned} \tag{7.16}$$

A Monte Carlo estimate of $L(\theta_0) = P_{\theta_0}(\mathbf{Y})$ is given by the mean value of $W_L(\mathbf{G}^*)$, over repeated independent repetitions of the sequential imputation process. Repeating the process for different trait locus positions on the chromosome, one obtains an estimated likelihood curve for the location of the trait locus. That is, we have a Monte Carlo estimate of the *location lod score curve* (section 6.2).

In genetic analyses, given the data, conditional expectations with respect to some particular model $P_{\theta_0}(\cdot)$ are often needed. These address such questions as: In which meioses and at what locations are the recombinations? Who should be sampled to obtain most additional information about the trait model or trait locus position? Where are the biggest uncertainties in underlying marker genotypes? How would it affect inferences to reduce such uncertainty? In principle, such expectations can be readily estimated, using the sequential imputation probability distribution P^* and computed weights W_L. For any function g^* of $\mathbf{G}$ and $\mathbf{Y}$,

$$\begin{aligned} \mathrm{E}_{\theta_0}(g^*(\mathbf{G}, \mathbf{Y}) \mid \mathbf{Y}) &= \sum_{\mathbf{G}} g^*(\mathbf{G}, \mathbf{Y}) P_{\theta_0}(\mathbf{G} \mid \mathbf{Y}) \\ &= \sum_{\mathbf{G}} g^*(\mathbf{G}, \mathbf{Y}) \frac{P^*(\mathbf{G}) W_L(\mathbf{G})}{P_{\theta_0}(\mathbf{Y})} \\ &= \frac{\mathrm{E}_{P^*}(g^*(\mathbf{G}, \mathbf{Y}) W_L(\mathbf{G}))}{P_{\theta_0}(\mathbf{Y})}. \end{aligned}$$

The normalizing factor $P_{\theta_0}(\mathbf{Y})$ is the unknown likelihood. Equation (7.16) provides a Monte Carlo estimate of $P_{\theta_0}(\mathbf{Y})$, so that

$$\mathrm{E}_{\theta_0}(g^*(\mathbf{G}, \mathbf{Y}) \mid \mathbf{Y}) = \frac{\mathrm{E}_{P^*}(g^*(\mathbf{G}, \mathbf{Y}) W_L(\mathbf{G}))}{\mathrm{E}_{P^*}(W_L(\mathbf{G}))}. \tag{7.17}$$

In this ratio estimator (7.17), each expectation in numerator and denominator is estimated by averaging values of each argument over independent realizations of $\mathbf{G}$ from the distribution $P^*(\mathbf{G})$. The same realizations may be used in estimating both the numerator and demoninator. This is often advantageous, since often there will then be positive correlation between the two Monte Carlo estimates, with consequent reduction in the Monte Carlo variance of the ratio.

Chapter 8

Markov chain Monte Carlo on Pedigrees

8.1 Simulation conditional on data: MCMC

Equation (7.10) gave the likelihood for a genetic model on a pedigree as an expectation over latent variables $\mathbf{X}$, and hence, in principle, provided a method for Monte Carlo estimation of the likelihood. We need to estimate

$$L(\theta) \;=\; P_\theta(\mathbf{Y}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{X}, \mathbf{Y}).$$

As previously, any suitable latent variables may be used, normally either meiosis indicators $\mathbf{S}$ or genotypes $\mathbf{G}$. For convenience, we use the general notation $\mathbf{X}$ for the general formulation.

However, unless the simulation distribution $P^*(\mathbf{X})$ is conditioned in some way on data $\mathbf{Y}$, equation (7.10) is often useless. Genotypes or gene descent patterns simulated from the prior probability distribution given only the model and the pedigree structure will rarely even be consistent with the observed data. Importance sampling considerations dictate that the sampling distribution should be close to proportional to $P_\theta(\mathbf{X}, \mathbf{Y})$, or as a function of latent variables $\mathbf{X}$ to $P_\theta(\mathbf{X} \mid \mathbf{Y})$ (equation (7.12)). Intuitively also, to obtain realizations that have better than infinitesimal probability of giving a non-negligible contribution to the likelihood we must simulate conditional on the data. However

$$P_\theta(\mathbf{X} \mid \mathbf{Y}) \;=\; \frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_\theta(\mathbf{Y})}, \tag{8.1}$$

and the normalizing factor $P_\theta(\mathbf{Y})$ is unknown. If we could compute $L(\theta) = P_\theta(\mathbf{Y})$, Monte Carlo estimation of likelihoods would be unnecessary.

Enter Markov chain Monte Carlo, or MCMC. We review briefly the Metropolis-Hastings class of algorithms (Hastings, 1970) for generating dependent realizations from a target probability distribution known only up to a normalizing factor. For

consistency of notation, we denote the target distribution by $P_\theta(\mathbf{X} \mid \mathbf{Y})$. The space of possible values of $\mathbf{X}$ is denoted $\mathcal{X}$. For each $\mathbf{X}$ in $\mathcal{X}$ a *proposal distribution* $q(\cdot; \mathbf{X})$ is defined. Then, if the process is now at $\mathbf{X}$ the next value is generated as follows:

1. Generate $\mathbf{X}^\dagger$ from the proposal distribution $q(\cdot; \mathbf{X})$
2. Compute the Hastings ratio

$$h(\mathbf{X}^\dagger; \mathbf{X}) = \frac{q(\mathbf{X}; \mathbf{X}^\dagger) P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})}{q(\mathbf{X}^\dagger; \mathbf{X}) P_\theta(\mathbf{X} \mid \mathbf{Y})}. \tag{8.2}$$

Note that h depends only on the ratio of densities $P_\theta(\cdot \mid \mathbf{Y})$, so that any normalizing factor need not be computed.

3. The resampled X^* is then determined from the Hastings ratio as follows:

$$\begin{aligned} P^*(\mathbf{X}^* = \mathbf{X}^\dagger) &= a = \min(1, h(\mathbf{X}^\dagger; \mathbf{X})) \\ P^*(\mathbf{X}^* = \mathbf{X}) &= (1 - a). \end{aligned}$$

Thus a is the *acceptance probability* for the proposed $\mathbf{X}^\dagger$.
Clearly, given the current value of $\mathbf{X}$, the probability distribution of $\mathbf{X}^*$ is determined, independently of the past of the process: a Markov chain on the space $\mathcal{X}$ of values of $\mathbf{X}$ has been defined.

It remains to show that the desired distribution $P_\theta(\mathbf{X}|\mathbf{Y})$ is an equilibrium distribution of the Markov chain. Hence, if the chain is aperiodic and irreducible, $P_\theta(\mathbf{X}|\mathbf{Y})$ is the unique equilibrium distribution. In this case, the ergodic theorem provides that time averages over realizations of the chain converge to expectations under the equilibrium distribution. These time-averages may then be used as Monte Carlo estimates of these expectations, just as previously in sections 3.7 and 7.6 simple averages of independent realizations were used.

The net resampling distribution $P^*(\mathbf{X}^*)$ is compounded from the proposal $q(\mathbf{X}^\dagger; \mathbf{X})$ and the acceptance or rejection step. Since the process is symmetric in $\mathbf{X}$ and a proposed $\mathbf{X}^\dagger$, with $h(\mathbf{X}^\dagger; \mathbf{X}) = (h(\mathbf{X}; \mathbf{X}^\dagger))^{-1}$, without loss of generality we can assume $h(\mathbf{X}^\dagger; \mathbf{X}) \geq 1$ or

$$q(\mathbf{X}; \mathbf{X}^\dagger) P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y}) \geq q(\mathbf{X}^\dagger; \mathbf{X}) P_\theta(\mathbf{X} \mid \mathbf{Y}).$$

Then a proposed transition from $\mathbf{X}$ to $\mathbf{X}^\dagger$ is accepted ($a = 1$) and the probability of the move is the proposal probability:

$$P^*(\mathbf{X}^\dagger; \mathbf{X}) = q(\mathbf{X}^\dagger; \mathbf{X}).$$

For the reverse move, from $\mathbf{X}^\dagger$, $\mathbf{X}$ must be both proposed and accepted. Thus, the probability, $P^*(\mathbf{X}; \mathbf{X}^\dagger)$, of the reverse transition is

$$\begin{aligned} q(\mathbf{X}; \mathbf{X}^\dagger) h(\mathbf{X}; \mathbf{X}^\dagger) &= q(\mathbf{X}; \mathbf{X}^\dagger) \frac{q(\mathbf{X}^\dagger; \mathbf{X}) P_\theta(\mathbf{X} \mid \mathbf{Y})}{q(\mathbf{X}; \mathbf{X}^\dagger) P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})} \\ &= q(\mathbf{X}^\dagger; \mathbf{X}) \frac{P_\theta(\mathbf{X} \mid \mathbf{Y})}{P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})}. \end{aligned}$$

Combining these two equations, we have

$$P^*(\mathbf{X}^\dagger; \mathbf{X}) P_\theta(\mathbf{X} \mid \mathbf{Y}) = P^*(\mathbf{X}; \mathbf{X}^\dagger) P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y}). \tag{8.3}$$

In words, under the defined Markov chain and distribution $P_\theta(\cdot \mid \mathbf{Y})$, the probability of being at $\mathbf{X}$ and moving to $\mathbf{X}^\dagger$ is the same as the probability of being at $\mathbf{X}^\dagger$ and moving to $\mathbf{X}$. This *detailed balance* condition holds for all $\mathbf{X}$ and $\mathbf{X}^\dagger$, which is a sufficient condition for $P_\theta(\cdot \mid \mathbf{Y})$ to be an equilibrium distribution of the Markov chain.

The algorithm of Metropolis et al. (1953) is a special case; if $q(\mathbf{X}^\dagger; \mathbf{X}) = q(\mathbf{X}; \mathbf{X}^\dagger)$ the Hastings ratio reduces to the odds ratio of the proposal state $\mathbf{X}^\dagger$ versus the current state $\mathbf{X}$. An alternative version of MCMC sampling is the Gibbs sampler (Geman and Geman, 1984). We consider here the general case in which, at a given step, $\mathbf{X}$ is partitioned into two sets of components, $\mathbf{X} = (\mathbf{X}_u, \mathbf{X}_f)$, the subscripts u denoting *updated* and f denoting *fixed*. These subsets change at each step, so that every component of $\mathbf{X}$ is sometimes updated. The sampled $\mathbf{X}^*$ differs from $\mathbf{X}$ only in the set of components $\mathbf{X}_u$, and $\mathbf{X}_u^*$ is sampled from the distribution $P_\theta(\mathbf{X}_u|\mathbf{X}_f, \mathbf{Y})$. Suppose $\mathbf{X}$ is currently from the desired distribution $P_\theta(\mathbf{X} \mid \mathbf{Y})$, so that the marginal distribution of the current $\mathbf{X}_f$ is $P_\theta(\mathbf{X}_f \mid \mathbf{Y})$. Thus the distribution of the resampled $\mathbf{X}^*$ is

$$\begin{aligned} P^*(\mathbf{X}_u^*, \mathbf{X}_f^*) &= P^*(\mathbf{X}_u^* \mid \mathbf{X}_f^*) P^*(\mathbf{X}_f^*) \\ &= P_\theta(\mathbf{X}_u^*|\mathbf{X}_f, \mathbf{Y}) P_\theta(\mathbf{X}_f \mid \mathbf{Y}) \\ &= P_\theta(\mathbf{X}^* \mid \mathbf{Y}). \end{aligned} \tag{8.4}$$

Thus the Gibbs sampler also maintains the equilibrium distribution $P_\theta(\cdot \mid \mathbf{Y})$.

The Gibbs sampler is, in fact, a special case of a Metropolis-Hastings sampler. Consider a Metropolis-Hastings sampler in which the proposal distribution is the resampling distribution of the Gibbs sampler:

$$q(\mathbf{X}^\dagger; \mathbf{X}) = P_\theta(\mathbf{X}_u^\dagger|\mathbf{X}_f, \mathbf{Y}) I(\mathbf{X}_f^\dagger \equiv \mathbf{X}_f)$$

where $I(\cdot)$ is the indicator function. Then the Hastings ratio is

$$\begin{aligned} h(\mathbf{X}^\dagger; \mathbf{X}) &= \frac{q(\mathbf{X}; \mathbf{X}^\dagger) P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})}{q(\mathbf{X}^\dagger; \mathbf{X}) P_\theta(\mathbf{X} \mid \mathbf{Y})} \\ &= \frac{P_\theta(\mathbf{X}_u|\mathbf{X}_f, \mathbf{Y}) P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})}{P_\theta(\mathbf{X}_u^\dagger|\mathbf{X}_f, \mathbf{Y}) P_\theta(\mathbf{X} \mid \mathbf{Y})} \\ &= \frac{P_\theta(\mathbf{X} \mid \mathbf{Y})}{P_\theta(\mathbf{X}_f \mid \mathbf{Y})} \frac{P_\theta(\mathbf{X}_f \mid \mathbf{Y})}{P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})} \frac{P_\theta(\mathbf{X}^\dagger \mid \mathbf{Y})}{P_\theta(\mathbf{X} \mid \mathbf{Y})} \\ &= 1. \end{aligned}$$

In this case $a \equiv h(\mathbf{X}^\dagger; \mathbf{X}) \equiv 1$, and no rejection step is necessary. Although, in the Gibbs sampler there is no rejection step, $\mathbf{X}^* = \mathbf{X}$ is possible, since $\mathbf{X}_u$ is a possible value for the resampled $\mathbf{X}_u^*$.

In order for the time-average over the chain to converge to the expectation under the equilibrium distribution, the ergodic theorem must apply. For discrete Markov chains, we need irreducibility. However, in practice, too much attention is paid to irreducibility. Any chain can be made irreducible, using Metropolis rejection, but irreducibility *per se* is useless. For example, one might decide that once in

a million trials one will propose a new realization from the prior distribution of latent variables. Once in a million million realizations one might get something compatible with the data. Once in a million, million, million trials one might get an accepted realization. Obviously nothing has changed with regard to realizations from the chain, but the sampler is irreducible. Metropolis rejected restarts are often a good idea — one of several key ideas in getting better samplers, and in assessing how good they are. However, it has to be done with the practical goal of more efficient Monte Carlo estimation.

There are two (related) sorts of convergence which often get confused. One is convergence of the marginal distribution of each $\mathbf{X}^{(\tau)}$ to the equilibrium distribution of the Markov chain as τ becomes large. The other relates to the convergence of a time-average over the chain to the expectation of the function under the equilibrium distribution. Both depend on the mixing properties of the Markov chain, and parameters such as the largest non-unit eigenvalue of the transition matrix, but the first can (in principle) be addressed by burn-in (discarding enough realizations before starting to accumulate the time-average) and is not normally a practical problem. The second class of questions remain even if we could start in the equilibrium probability distribution. This is a much bigger problem; all parts of the space contributing substantially to the target probability distribution must be sampled. Although shorter runs in different parts of the space may be helpful in diagnosing a problem, Monte Carlo estimation must be done using a time-average of a single realization of the Markov chain process. Runs in different parts of the space cannot be combined, without knowledge of how to weight the realizations from the different starts. (See Geyer (1992) for more discussion.)

Estimation of the standard deviations of Monte Carlo estimates of expectations is essential. Several easily implemented estimators have been proposed, but assessment of the estimates is hard, in practice. Again, Geyer (1992) is a good reference. One of the simplest methods of estimating Monte Carlo variances is by using batch means (Hastings, 1970). One divides the realizations into sufficiently large batches so that the batch means are "almost independent", and relates the variance of independent batch means to the variance of the overall mean (the estimator of the expectation). The variance of independent batch means can be estimated from the empirical variance as in section 3.7. One can test for autocorrelation between the batch means. This is quite effective if the sampler is doing well, but can severely underestimate variance if the sampler is not getting around the space. However, other variance estimators have the same deficiency, and the empirical variance of the batch means is easily computed.

Variance estimation also relates to the choice of spacing in sampling realizations from an MCMC. The optimal spacing is the one that achieves minimum computational cost for given precision of the resulting estimator. This optimal spacing depends on the relative costs of generating the samples and of evaluating the contribution to the estimator at the realized values, but is seldom large (Geyer, 1992).

This section has aimed only to outline the main principles and issues in MCMC. For those who wish to pursue the topic, Gilks et al. (1996) is a good starting point, while there is already a large more recent literature.

8.2 Single-site updating methods

As in other areas of application, the earliest MCMC samplers that were used to realize latent variables on pedigrees conditional on phenotypic data were mainly single-site updating methods. The proposed changes to the latent variable configurations were thus very small. Lange and Matthysse (1989) used as their latent variables both the genotypes and inheritance patterns of genes, and used a Metropolis algorithm to propose changes. Sheehan (1990) and Thompson and Guo (1991) used a Gibbs sampling approach, using the genotypes as the latent variables, while Thompson (1994*a*; 1994*b*) used a Metropolis algorithm to update a single meiosis indicator $S_{i,j}$ for meiosis i and locus j.

Unfortunately, in genetic examples the constraints on genotypes $\mathbf{G}$ or meiosis indicators $\mathbf{S}$ imposed by Mendelian segregation and discrete marker phenotypes mean that any proposal that makes multiple changes to the current value of $\mathbf{G}$ or $\mathbf{S}$ has a high probability of proposing a configuration inconsistent with the data $\mathbf{Y}$. By contrast, although proposed changes are small, single-site updates are easily proposed and often accepted. The genes and heritable effects in an individual are determined by those in his parents, and jointly with those in his spouse, influence those in his offspring (Figure 1.3(a)). This neighborhood structure means that a single-site Gibbs sampler is easy to implement. Each genetic effect in each individual is successively updated, conditional upon the remainder.

Specifically, where genotypes $\mathbf{G}$ are the latent variables, underlying genotypes for both trait and marker loci are sampled individual by individual and locus by locus. For a single-site update to component $G_{i,j}$, the genotype of individual i at locus j, the proposal distribution for the Gibbs sampler (equation (8.4)) is

$$\begin{aligned} q_{i,j}(\mathbf{G}^*;\mathbf{G}) &= P_\theta(G^*_{i,j} \mid \mathbf{G}_{-(i,j)}, \mathbf{Y}) \text{ for component } (i,j) \\ G^*_{k,l} &= G_{k,l} \text{ for } (k,l) \neq (i,j), \quad \text{or } \mathbf{G}^*_{-(i,j)} = \mathbf{G}_{-(i,j)}. \end{aligned} \tag{8.5}$$

As for $\mathbf{S}$ in section 4.7, we use the standard notation $\mathbf{G}_{-(i,j)}$ for the set of all components of $\mathbf{G}$ other than $G_{i,j}$. This full conditional distribution for $G_{i,j}$ is easily computed, but only small changes to $\mathbf{G}$ are possible at each step. On the other hand, the full conditionals for larger blocks of components $\mathbf{G}_\mathcal{T} = \{G_{i,j}; (i,j) \in \mathcal{T}\}$ are more computationally intensive or even infeasible.

For certain data configurations, the single-site genotypic Gibbs sampler is not irreducible when a locus is multiallelic. However, theoretical irreducibility can always be easily achieved. The practical problem is failure of the sampler to mix adequately. This can be a problem on large pedigrees even for diallelic loci, particularly if underlying genotypes are highly constrained (but not determined) by the data. The reducibility of the Gibbs sampler for genetic loci with more than two alleles was first addressed by Sheehan and Thomas (1993), in the context of a single-genotype Gibbs sampler. Their method used modification of either the segregation probabilities or the penetrance probabilities, so that the sampler was no longer irreducible. For example, modifying the penetrances

$$\begin{aligned} P^*(Y_{i,j} \mid G_{i,j}) &= P_\theta(Y_{i,j} \mid G_{i,j}) \quad \text{if } P_\theta(Y_{i,j} \mid G_{i,j}) > 0 \\ P^*(Y_{i,j} \mid G_{i,j}) &= c \qquad\qquad \text{if } P_\theta(Y_{i,j} \mid G_{i,j}) = 0. \end{aligned} \tag{8.6}$$

Then

$$\begin{aligned}\frac{P_\theta(\mathbf{G},\mathbf{Y})}{P^*(\mathbf{G},\mathbf{Y})} &= 1 \text{ if } P_\theta(\mathbf{Y} \mid \mathbf{G}) > 0 \\ &= 0 \text{ if } P_\theta(\mathbf{Y} \mid \mathbf{G}) = 0.\end{aligned}$$

Thus no reweighting is required in order for the realizations to represent the distribution of genotypes under the true genetic model. All realizations consistent with the true model have equal weight; those inconsistent with it are just dropped from the output sample. Lin et al. (1993) used similar penetrance modifications to achieve irreducibility, but used Metropolis-coupled samplers (Geyer, 1991*a*), coupling a sampler under the true model to samplers which were not only irreducible, but also moved more quickly around the space. Rather than a uniform penetrance modification for all individuals, only individual-specific changes necessary to achieve irreducibility are made. The expansion of the space that is sampled is therefore limited.

Several methods for more efficient sampling of the space of feasible underlying genotype configurations have been developed. Some of these are due to Shili Lin (Lin et al., 1993; Lin et al., 1994). Others are due to Eric Sobel (Sobel and Lange, 1993) and to Charles Geyer (Geyer and Thompson, 1995). We briefly outline here only the methods of Lin et al. (1993; 1994), directed specifically towards sampling of genotypes at polymorphic marker loci where there are many unsampled individuals in the pedigree. These methods use a form of "heated proposals", resulting in samplers that move around the space of genotypic configurations far more effectively.

One possibility is to base a Metropolis-Hastings sampler on the local conditional distribution for the single component $G_{i,j}$ (equation (8.5)), but in a way that enhances movement around the space. The method of Lin et al. (1994) "flattens" the proposal distribution in a manner similar to simulated annealing, using a "temperature" parameter T:

$$\begin{aligned}q_{i,j}(\mathbf{G}^*;\mathbf{G}) &\propto (P_\theta(G^*_{i,j} \mid \mathbf{G}_{-(i,j)},\mathbf{Y}))^{1/T} \text{ for component } (i,j) \\ G^*_{k,l} &= G_{k,l} \text{ for } (k,l) \neq (i,j), \quad \text{or } \mathbf{G}^*_{-(i,j)} = \mathbf{G}_{-(i,j)}.\end{aligned}$$

The Hastings ratio is then

$$\begin{aligned}h(\mathbf{G}^*;\mathbf{G}) &= \frac{q(\mathbf{G};\mathbf{G}^*)P_\theta(\mathbf{G}^* \mid \mathbf{Y})}{q(\mathbf{G}^*;\mathbf{G})P_\theta(\mathbf{G} \mid \mathbf{Y})} \\ &= \frac{(P_\theta(G_{i,j} \mid \mathbf{G}^*_{-(i,j)},\mathbf{Y}))^{1/T}P_\theta(\mathbf{G}^* \mid \mathbf{Y})}{(P_\theta(G^*_{i,j} \mid \mathbf{G}_{-(i,j)},\mathbf{Y}))^{1/T}P_\theta(\mathbf{G} \mid \mathbf{Y})} \\ &= \frac{(P_\theta(\mathbf{G} \mid \mathbf{Y}))^{1/T}P_\theta(\mathbf{G}^* \mid \mathbf{Y})(P_\theta(\mathbf{G}_{-(i,j)} \mid \mathbf{Y}))^{1/T}}{(P_\theta(\mathbf{G}^* \mid \mathbf{Y}))^{1/T}P_\theta(\mathbf{G} \mid \mathbf{Y})(P_\theta(\mathbf{G}^*_{-(i,j)} \mid \mathbf{Y}))^{1/T}} \\ &= \frac{(P_\theta(\mathbf{G}^* \mid \mathbf{Y}))^{1-1/T}}{(P_\theta(\mathbf{G} \mid \mathbf{Y}))^{1-1/T}} \\ &= \frac{(P_\theta(G^*_{i,j} \mid \mathbf{G}_{-(i,j)},\mathbf{Y}))^{1-1/T}}{(P_\theta(G_{i,j} \mid \mathbf{G}_{-(i,j)},\mathbf{Y}))^{1-1/T}}\end{aligned}$$

using, in several steps, the fact that $\mathbf{G}_{-(i,j)} = \mathbf{G}^*_{-(i,j)}$. The Hastings ratio is thus as easily computed as the local conditionals $P_\theta(G_{i,j} \mid \mathbf{G}_{-(i,j)}, \mathbf{Y})$. An interesting feature of this system is that, with $T > 1$, the probability of change in $\mathbf{G}$ is reduced from that for the Gibbs sampler, where $T = 1$ (C. Jennison, pers. comm. 1992). However, because this increases the probability that the sampler remains in low-probability states, it increases the overall probability of a succession of changes that moves $\mathbf{G}$ to a different part of the space. The probabilities of single-step changes are not necessarily indicative of overall performance of the sampler, particularly in high-dimensional spaces.

Under the assumption that $S_{\bullet,j}$ are first-order Markov over loci j (section 4.7), the single-site meiosis indicator sampler is also easily implemented (Thompson, 1994*a*). Since $S_{i,j}$ is binary, a Metropolis algorithm is natural. A meiosis i and locus j are selected at random, and a change from $S_{i,j} = s$ to $S_{i,j} = (1 - s)$ is proposed. This proposal changes only the recombinant/non-recombinant status in the two intervals adjoining locus j, and the conditional probability of marker data at locus j:

$$
\begin{aligned}
h(\mathbf{S}^*;\mathbf{S}) &= \frac{P_\theta(\mathbf{Y} \mid \mathbf{S}^*)P_\theta(\mathbf{S}^*)}{P_\theta(\mathbf{Y} \mid \mathbf{S})P_\theta(\mathbf{S})} \\
&= \frac{P_\theta(Y_{\bullet,j} \mid S^*_{\bullet,j})P_\theta(S^*_{i,j} \mid S_{i,j-1}, S_{i,j+1})}{P_\theta(Y_{\bullet,j} \mid S_{\bullet,j})P_\theta(S_{i,j} \mid S_{i,j-1}, S_{i,j+1})} \\
&= \frac{P_\theta(Y_{\bullet,j} \mid S^*_{\bullet,j})}{P_\theta(Y_{\bullet,j} \mid S_{\bullet,j})} \left(\frac{\rho_{j-1}}{1-\rho_{j-1}}\right)^{T_{j-1}} \left(\frac{\rho_j}{1-\rho_j}\right)^{T_j}, \qquad (8.7)
\end{aligned}
$$

for $j = 1, \ldots, L$ (see equation 4.12). Here $\rho_{j-1} = \Pr(S_{i,j-1} \neq S_{i,j})$ is the recombination frequency between locus $j-1$ and locus j, and $T_{j-1} = (|S_{i,j-1} - s| - |S_{i,j-1} - 1 + s|)$ is the indicator of whether the proposal places ($T_{j-1} = +1$) or removes ($T_{j-1} = -1$) a recombination between locus $j-1$ and j. The values ρ_j and T_j are analogously defined for the interval j to $j+1$, and $\rho_0 = \rho_L = \frac{1}{2}$. The first term in the Hastings ratio $h(\mathbf{S}^*;\mathbf{S})$ is given by equation (3.10) and is easily computed by the methods outlined in that section, provided there are not too many data $S_{\bullet,j}$ on the pedigree. Generally, the space of latent variables is smaller for $\mathbf{S}$ than for $\mathbf{G}$, and hence MCMC is more effective. The sampler may not be irreducible (Sobel and Lange, 1996), but there are many fewer constraints than with a genotypic sampler and irreducibility is often provable on a locus-by-locus basis (Thompson, 1994*a*; Thompson and Heath, 1999). Note that, provided recombination frequencies between adjacent loci are strictly positive, irreducibility is a single-locus issue.

8.3 Combining exact computation and Monte Carlo

A major difficulty with MCMC methods is to ensure proper mixing of the samplers, and hence efficient Monte Carlo estimation. On large pedigrees, with models

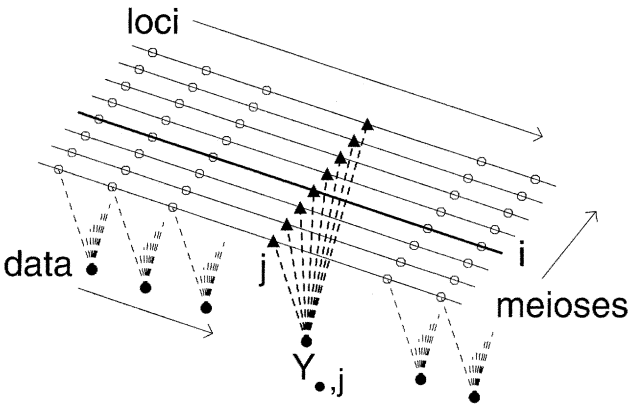

FIGURE 8.1. *The conditional independence structure for MCMC sampling*

or data involving multiple linked loci, single-variable MCMC updating methods are not effective. Some approaches to improving Monte Carlo estimates involve some combination of exact and Monte Carlo computation. One straightforward idea is simply to compute exactly on those parts of the pedigree on which this is possible (Thompson, 1991). The results from peeling peripheral parts of the pedigree enter as potentials on nodes of the remaining core (Geyer and Thompson, 1995), and the space over which MCMC sampling is required is reduced. Rao-Blackwellized estimators for mixed-model likelihoods (section 9.3) also combine exact computation with MCMC sampling. However, the sampling used for these estimators by Thompson and Guo (1991) was single-site updating. Major improvements can be gained only by improved MCMC samplers.

Recently a variety of joint-updating schemes have been developed. For example, Jensen et al. (1995) update genotypes of blocks of individuals jointly at several loci. Jensen and Kong (1999) update arbitrary collections of the latent variables in the pedigree, selected using the HUGIN Bayesian expert system software (Andersen et al., 1989). Heath (1997) and Thompson and Heath (1999) use the meiosis indicators $\mathbf{S} = \{S_{i,j}\}$. Heath (1997) updates jointly the components of $S_{\bullet,j}$, the indicators at a single locus j: the *L-sampler*. Thompson and Heath (1999) update jointly the components of $S_{i,\bullet}$, the meiosis indicators for all loci in a single meiosis i: the *M-sampler*. All these MCMC methods provide, directly or indirectly, realizations of the descent of genes in pedigrees and the genotypes of individuals, and hence Monte Carlo estimates of likelihoods for linkage and segregation analysis (sections 6.2, 6.3 and 7.6), and the probabilities of gene identity by descent and haplotype sharing conditional on observed trait and marker data $\mathbf{Y}$ (section 3.6).

In a Bayesian framework, the segregation and linkage parameters of genetic models are assigned prior probability distributions (see section 2.4). In this case, the same MCMC methods provide estimates of the posterior probability distributions of linkage and trait gene effects and locations.

In the locus-by-locus sampler (L-sampler) first developed by Kong (1991), all genotypes $G_{\bullet,j} = \{G_{i,j}\}$ at a single locus j are updated conditionally upon those at neighboring loci. Computationally the approach is analogous to the sequential imputation method of section 7.5, except that sampling is from the full conditional of $G_{\bullet,j}$. Heath (1997) has further developed the L-sampler, and widened its scope, using $S_{\bullet,j}$ rather than $G_{\bullet,j}$. Because of the structure, this full conditional distribution of $S_{\bullet,j}$ given the data $\mathbf{Y}$ and the meiosis indicators $\mathbf{S}_{-j} = \{S_{\bullet,l}, l \neq j\}$ is

$$P_\theta(S_{\bullet,j} \mid \mathbf{S}_{-j}, \mathbf{Y}) \quad = \quad P_\theta(S_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1}, Y_{\bullet,j}).$$

That is, the distribution depends only on current values of $S_{\bullet,j-1}$ and $S_{\bullet,j+1}$ and data $Y_{\bullet,j}$. Thus, the calculation of $P_\theta(S_{\bullet,j} \mid \mathbf{S}_{-j}, \mathbf{Y})$ is a single-locus peeling computation analogous to those of section 6.3, and is often feasible. The developments of Heath (1997) are in the context of Bayesian analyses of quantitative traits, under models of several loci contributing additively to the trait value. His approach uses a variety of improved sampling and computational ideas, including more efficient peeling algorithms, integrated proposal distributions (Besag et al., 1995) and reversible jump MCMC (Green, 1995). The output consists of realizations of putative trait loci from a Bayesian posterior; no likelihood or lod score is obtained. One great advantage of the L-sampler is that it is irreducible, provided only that recombination probabilities between adjacent loci are strictly positive. Moreover, this MCMC sampling is a great improvement over single-site methods. However, when there are multiple tightly linked marker loci, mixing can be poor.

8.4 Tightly-linked loci: the M-sampler

The single-site ($S_{i,j}$) or single-locus ($S_{\bullet,j}$) update has mixing problems when loci are tightly linked. An alternative form of block-updating is to update jointly the meiosis indicators for all loci in a given meiosis ($S_{i,\bullet}$). The M-sampler is a whole-meiosis Gibbs sampler (Thompson and Heath, 1999) for $S_{i,\bullet}$. At each step a random meiosis is selected for updating; alternatives in which meioses are updated sequentially are also possible. Note also that, for an unobserved founder with only one offspring in the pedigree, the meiosis from the founder parent to the offspring can be ignored (and not sampled), since there is no information on the haplotypes transmitted.

To implement the M-sampler we must compute

$$\Pr(S_{i,\bullet} \mid \{S_{k,\bullet}, k \neq i\}, \mathbf{Y}).$$

As previously (section 6.2), we suppose that the marker data $\mathbf{Y}$ can be partitioned into data relating to each locus $j = 1, 2, ..., L$, and that the loci are numbered in

order along the chromosome. Then

$$\mathbf{Y} \quad = \quad (Y_{\bullet,1}, ..., Y_{\bullet,L}).$$

As in section 6.2, let

$$Y^{(j)} \quad = \quad (Y_{\bullet,1}, ..., Y_{\bullet,j}), \quad \text{so} \quad \mathbf{Y} = Y^{(L)}.$$

We have seen in section 3.6 that $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ can be easily computed.

Now define

$$Q_j^\dagger(s) \quad = \quad \Pr(S_{i,j} = s \mid \{S_{k,\bullet}, k \neq i\},\ Y^{(j)})$$

for $s = 0, 1$. Note that this function $Q_j^\dagger(\cdot)$ is analogous, but not identical, to the function $Q_j^\dagger(\cdot)$ of section 7.1. There the probability considered was the joint distribution for all components of $S_{\bullet,j}$, conditional on $Y^{(j)}$; here the probability is for $S_{i,j}$ conditioning additionally on indicators at other meioses $\{S_{k,\bullet}, k \neq i\}$. Meiosis indicators $S_{i,\bullet}$ are *a priori* independent over i, and become dependent only through conditioning on the data $\mathbf{Y}$ (Figure 8.1). Thus, $Q_j^\dagger(s)$ is the probability for the meiosis indicator $S_{i,j}$, given the data $Y^{(j)}$ and other $(k \neq i)$ meiosis indicators at loci up to and including locus j. (The components $S_{k,l}$ for $l > j$ are irrelevant, since $Y_{\bullet,l}$ is not conditioned upon.) Thus, by analogy with section 7.1, $Q_j^\dagger(s)$ may be computed sequentially just as in equation (7.2). The only difference is that now, rather than considering all 2^m possible values of $S_{\bullet,j}$, we consider only values of the single binary indicator $S_{i,j}$, conditioning on the remainder $(k \neq i)$ which remain fixed. In meiosis i, there is no recombination between locus $(j-1)$ and locus j if the value $(s = 0,\ 1)$ of $S_{i,j}$ is the same as at locus $(j-1)$, and there is recombination if the values differ. That is

$$Q_1^\dagger(s) \quad \propto \quad \Pr(Y_{\bullet,1} \mid S_{\bullet,1})$$

and

$$\begin{aligned} Q_j^\dagger(s) \quad \propto \quad & \Pr(Y_{\bullet,j} \mid S_{\bullet,j})\ (Q_{j-1}^\dagger(s)(1-\rho_{j-1}) \\ & + \ Q_{j-1}^\dagger(1-s)\rho_{j-1}) \end{aligned} \tag{8.8}$$

for $j = 2, ..., L$. In this equation, $S_{\bullet,j}$ takes the current value at meioses k other than i, and the value s for meiosis i. As before, ρ_{j-1} is the recombination frequency between locus $j-1$ and locus j. Thus we may compute (8.8) for each j in turn, working forwards sequentially along the chromosome.

Finally we have computed

$$Q_L^\dagger(s) \quad = \quad \Pr(S_{i,L} = s \mid \{S_{k,\bullet}, k \neq i\},\ \mathbf{Y} = Y^{(L)})$$

and thus $S_{i,L}$ may be sampled from this desired conditional distribution. Suppose now each $S_{i,l}$ has been successively sampled from the required distribution for $l = L, L-1, \ldots, j+1, j$. Then

$$\begin{aligned} & \Pr(S_{i,j-1} = s \mid \{S_{k,\bullet}, k \neq i\}, \{S_{i,l}, l = j, ..., L\}, \mathbf{Y}) \\ & \qquad \propto Q_{j-1}^\dagger(s)\ (T_j\ \rho_{j-1} \ + \ (1-T_j)(1-\rho_{j-1})) \end{aligned} \tag{8.9}$$

where $T_j = |S_{i,j} - s|$ is the indicator of recombination in the interval $j-1$ to j. Thus we may work backwards along the chromosome, sampling each $S_{i,j}$ in turn ($j = L, ..., 1$), obtaining overall a joint realization of $S_{i,j}$, $j = 1, ..., L$ from its full conditional distribution given $\{S_{k,\bullet}, k \neq i\}$ and $\mathbf{Y}$. Again, this is directly analogous to equation (7.6) of section 7.1.

Throughout this chapter we have ignored the fact that genetic maps differ between males and females: the order of loci is the same, but the recombination frequencies can differ quite widely. Linkage analysis computations should accommodate different values of recombination frequencies for males and females. For the M-sampler this is particularly straightforward, since each meiosis is in a male or in a female. As will be shown in section 11.2, the M-sampler can also incorporate more general meiosis models, including genetic interference, by using a Metropolis-Hastings acceptance/rejection step (Thompson, 2000*a*).

Implementations of almost all the computational algorithms referred to in this chapter are freely available by ftp. The Rockefeller Genetic Linkage Software list at http://linkage.rockefeller.edu/soft/list.html is an excellent reference. The software of our group is implemented primarily in our MORGAN package, which is available by ftp at . The most recent release of MORGAN (MORGAN_VF1, shortly to be replaced by MORGAN_V2.3) includes L-sampler and M-sampler implementations. The site www.stat.washington.edu/thompson/Genepi/pangaea.shtml also includes the *Loki* package for MCMC linkage analysis of quantitative traits (Heath, 1997).

Chapter 9

Likelihood Ratios for Genetic Analysis

9.1 Monte Carlo likelihood ratio estimation

The MCMC methods of Chapter 8 provide methods for obtaining realizations from $P_\theta(\mathbf{X} \mid \mathbf{Y})$, the probability distribution of latent variables $\mathbf{X}$ conditional on data $\mathbf{Y}$ under a model indexed by parameters θ. In this chapter, we discuss methods of using such realizations in Monte Carlo methods for linkage and segregation analysis, focusing on likelihood methods.

Recall again (equation (7.8)) that, for phenotypic data $\mathbf{Y}$,

$$L(\theta) \;=\; P_\theta(\mathbf{Y}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{X}, \mathbf{Y}),$$

where latent variables $\mathbf{X}$ are genotypes $\mathbf{G}$ or meiosis indicators $\mathbf{S}$. We again use θ to denote the general set of parameters of a genetic model. These include the recombination or gene location parameters. From equation (7.12), efficient Monte Carlo estimation of $L(\theta)$ will result from sampling from a distribution $P^*(\mathbf{X})$ close to proportional to the joint probability $P_\theta(\mathbf{X}, \mathbf{Y})$:

$$P^*(\mathbf{X}) \;\approx\; P_\theta(\mathbf{X} \mid \mathbf{Y}) \;\propto\; P_\theta(\mathbf{X}, \mathbf{Y}).$$

One possible choice is thus to simulate, by the methods of Chapter 8, not from $P_\theta(\mathbf{X} \mid \mathbf{Y})$ but from $P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})$, where $\theta_0 \approx \theta$. Then

$$\begin{aligned} P_\theta(\mathbf{Y}) \;&=\; \sum_{\mathbf{X}} P_\theta(\mathbf{X}, \mathbf{Y}) \;=\; \sum_{\mathbf{X}} \frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})} P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \\ &=\; \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X},\, \mathbf{Y})}{P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})} \;\middle|\; \mathbf{Y} \right) \;=\; P_{\theta_0}(\mathbf{Y})\, \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_{\theta_0}(\mathbf{X}, \mathbf{Y})} \;\middle|\; \mathbf{Y} \right). \end{aligned}$$

Hence in genetic analysis, or in any missing-data context, we have the key formula

of Thompson and Guo (1991)

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X},\mathbf{Y})}{P_{\theta_0}(\mathbf{X},\mathbf{Y})} \;\middle|\; \mathbf{Y}\right). \tag{9.1}$$

In this expectation, $\mathbf{X}$ is the random variable, $\mathbf{Y}$ is fixed. The distribution of $\mathbf{X}$ is $P_{\theta_0}(\cdot|\mathbf{Y})$. If $\mathbf{X}^{(\tau)}$, $\tau = 1,\ldots,N$, are realized from this distribution then the likelihood ratio can be estimated by

$$\frac{1}{N}\sum_{\tau=1}^{N}\left(\frac{P_\theta(\mathbf{X}^{(\tau)},\mathbf{Y})}{P_{\theta_0}(\mathbf{X}^{(\tau)},\mathbf{Y})}\right).$$

In section 8.1 we saw how MCMC can be used to realize $\mathbf{X}$ from $P_{\theta_0}(\cdot|\mathbf{Y})$.

Simulation at a single model θ_0 provides an estimate of the relative likelihood $L(\theta)/L(\theta_0)$ as a function of θ. This will be a satisfactory estimator only for those θ close to θ_0; specifically, for those θ for which $P_\theta(\mathbf{X}|\mathbf{Y})$ is close to proportional to $P_{\theta_0}(\mathbf{X},\mathbf{Y})$. Sometimes, primary interest is in the shape of the likelihood surface in the neighborhood of some specific point, such as the maximum likelihood estimate (MLE). In this case, preliminary MCMC runs and likelihood ratio function estimates can be used to obtain a ballpark value of the MLE (Geyer and Thompson, 1992). Alternatively, Monte Carlo EM can be used (see section 9.3). Once a ballpark estimate of the parameter values is found, one very large MCMC run can provide an accurate estimate of the MLE and of the likelihood in the region. However, this approach has limitations. One may be interested in the likelihood surface, or in log-likelihood differences, over large regions in the parameter space. Or, the large MCMC run may reveal that one's initial estimate was not sufficiently close to the MLE, and additional large runs may be necessary. It is desirable to find a method that combines realizations from all the runs, and provides an estimate of the likelihood surface over a range of parameter values.

9.2 Monte Carlo relative likelihood surfaces

One way of combining realizations from different MCMC samplers was provided by Geyer (1991*b*). MCMC samplers are run at many models, covering the range of interest, say at $\theta_0, \theta_1, \ldots, \theta_K$. The sets of N_j realizations from $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$, $j = 0, 1, \ldots, K$, give a combined set of realizations from

$$\begin{aligned} P^*(\mathbf{X}) &= \frac{1}{\sum_j N_j}\sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{X}|\mathbf{Y}) \\ &= \frac{1}{\sum_j N_j}\sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{X},\mathbf{Y})/L(\theta_j) \end{aligned}$$

and writing the likelihood estimation formula as an expectation with respect to this P^*

$$L(\theta_j) = \mathrm{E}_{P^*}\left(\frac{P_{\theta_j}(\mathbf{X},\mathbf{Y})}{P^*(\mathbf{X})}\right).$$

Now, although we have a sample from P^*, the denominator $P^*(\mathbf{X})$ cannot be explicitly computed, since it depends on the unknown $L(\theta_j)$, but we have the implicit Monte Carlo estimating equations

$$\begin{aligned} L(\theta_j) &= \sum_{\mathbf{X}^*} \left(\frac{P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{X}^*, \mathbf{Y})/L(\theta_l)} \right) \\ &= \sum_{\mathbf{X}^*} \left(\sum_{l=0}^{K} N_l \frac{P_{\theta_l}(\mathbf{X}^*, \mathbf{Y})}{P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})} \frac{1}{L(\theta_l)} \right)^{-1} \end{aligned} \tag{9.2}$$

for $j = 0, ..., K$, where the sum is over the total set of realizations $\mathbf{X}^*$. These equations determine only the relative values of $L(\theta_j)$, but can be solved iteratively for these relative values. For example, one may iterate equation (9.2) directly, renormalizing after each cycle, to keep one value, say $L(\theta_0)$ fixed (=1). This iterative procedure is globally convergent to the unique solution of equation (9.2). Once the relative values of $L(\theta_j)$ are found, then, for any other value of θ in the range spanned by the set of θ_j, $L(\theta)$ can be estimated by

$$L(\theta) = \sum_{\mathbf{X}^*} \left(\sum_{l=0}^{K} N_l \frac{P_{\theta_l}(\mathbf{X}^*, \mathbf{Y})}{P_{\theta}(\mathbf{X}^*, \mathbf{Y})} \frac{1}{L(\theta_l)} \right)^{-1} \tag{9.3}$$

where the sum is over the same total set of realizations as before. (Again, the estimate is relative to $L(\theta_0) = 1$.) Geyer (1991*b*) named this method *reverse logistic regression.*

There are two requirements for this approach to be an effective solution to the likelihood estimation problem. First, each sampler $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$, $j = 0, \ldots, K$ must cover well that part of the space of $\mathbf{X}$-values that has high total probability mass under that probability distribution — for an MCMC sampler on a large and structured space of latent variables, this is a non-trivial consideration (section 8.1). Second, even if the separate samplers are behaving "well", in this sense, for the mixture estimates to be effective we need good "overlap" between adjacent models. The conditional probability that a particular observation $\mathbf{X}$ derives from the sample P_{θ_j} is

$$\frac{N_j P_{\theta_j}(\mathbf{X}|\mathbf{Y})}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{X}|\mathbf{Y})}.$$

For every j, the values of these probabilities should not be too close to 1 for too large a proportion of the sampled $\mathbf{X}$-values. Thus adjacent parameter values θ_j must be chosen not too far apart, where the relevant measure of distance is in terms of the probability distributions $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$ of the $\mathbf{X}$-values generated.

Other difficulties with using the reverse logistic regression method concern computational resources. Either the realized $\mathbf{X}^*$, or at least the values $P_\theta(\mathbf{X}^*, \mathbf{Y})$ for each θ of interest, must be saved, in order for equations (9.2) and (9.3) to be implemented. This can demand massive amounts of storage. An alternative is to use block averages of the ratios of $P_{\theta_j}(\mathbf{X}, \mathbf{Y})/P_{\theta_l}(\mathbf{X}, \mathbf{Y})$ in equation (9.2)

(Thompson, 1994*b*). In the extreme case, this block might be the average over a full run of the sampler at a given θ_j. Let

$$R_j(\theta_l, \theta_j) = N_j^{-1} \sum_{\mathbf{X}^{*(j)}} \frac{P_{\theta_l}(\mathbf{X}^{*(j)}, \mathbf{Y})}{P_{\theta_j}(\mathbf{X}^{*(j)}, \mathbf{Y})}$$

be the likelihood ratio estimate of $L(\theta_l)/L(\theta_j)$ from N_j realizations $\mathbf{X}^{*(j)}$ at θ_j. Here the chosen values of l may vary with j. We define $R_j(\theta_l, \theta_j)$ to be 0 if $L(\theta_l)/L(\theta_j)$ is not estimated from realizations under model θ_j. At a minimum, for each j, values for $R_j(\theta_l, \theta j)$ should be computed for the values θ_l adjacent to θ_j. Then the estimating equation (9.2) becomes

$$L(\theta_j) = \left(\sum_{l=0}^{K} R_j(\theta_l, \theta_j) \frac{1}{L(\theta_l)} \right)^{-1}. \tag{9.4}$$

Writing $\nu_j = 1/L(\theta_j)$, $R_{jl} = R_j(\theta_l, \theta_j)$, $\boldsymbol{\nu} = (\nu_j)$, and $\mathbf{R} = (R_{jl})$, equation (9.4) becomes

$$\boldsymbol{\nu} = \mathbf{R}\boldsymbol{\nu}.$$

That is, the vector of ν_j-values is a right eigenvector of the matrix $\mathbf{R}$. Asymptotically, for large Monte Carlo runs, each computed R_{jl}-value converges to $L(\theta_l)/L(\theta_j) = \nu_j/\nu_l$. Thus, if, for each j, R_{jl} is evaluated for t other θ_l values, then each evaluated $R_{jl}\nu_l$ is approximately ν_j, and the corresponding eigenvalue should be t. This provides one check on the performance of the method, although in practice it is a weak criterion. The eigenvalue can be close to t even when performance is poor.

There are many open questions in the statistical properties of estimators such as those resulting from equation (9.4). If sufficient realizations can be stored, then equation (9.2) may provide the more satisfactory estimate. Suppose, however, only one in 1000 samples $\mathbf{X}^*$ or resulting probabilities $P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})$ can be stored. Then should one use the estimate (9.2), or one that uses the block averages over each block of 1000 steps? The latter would require more computation (evaluations of $P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})$), but the same amount of store. The Monte-Carlo variance of the block-average will be less than that of individual values $P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})$, but possibly not by much if the autocorrelation in the Markov chain is very high. Clearly these questions are related also to issues of computational efficiency in sub-sampling and spacing in the MCMC (Geyer, 1992), discussed briefly in section 8.1.

9.3 Monte Carlo EM for the mixed model

For some models, exact computation of the conditional expectations required to implement an EM algorithm may be impractical or infeasible, particularly if the model is complex, or there are missing data. Penetrance parameters may not be simple functions of genotypic counts. Even the bivariate case of the simple polygenic

model (section 2.6) may be complicated, if some individuals are observed for just one of the two traits (Thompson and Shaw, 1992). Chiasmata patterns are not so readily imputed if the recombination patterns of some gametes are not fully observable (section 5.3), due to missing typings or parental homozygosity at some loci. However, if latent genotypes or meiosis indicators and missing phenotypes can be realized from their conditional distributions given the observed data $\mathbf{Y} = \mathbf{y}$ under current values of the parameters, a Monte Carlo EM (MCEM) is easily implemented.

In section 2.6, the simple polygenic model was introduced, and the EM-algorithm for the variance component parameters σ_a^2 and σ_e^2 was outlined. In section 6.6, the univariate trait model was generalized to the mixed model, including both Mendelian genotypes and Gaussian polygenic effects (see equation (6.5)). The parameters then include also the frequency of the alleles at the diallelic Mendelian trait locus, and the vector of genotypic means $\boldsymbol{\mu} = (\mu(g))$ for the genotypes g at the locus. As before, we index the members of the pedigree by i, $i = 1, \ldots, n_{tot}$. Suppose that the n_{obs} observed members of the pedigree are those indexed by $i \in \mathcal{D}$. Then, for i in $\mathcal{D}$, we have equation (6.5):

$$Y_i \;=\; \mu(G_i) \;+\; Z_i \;+\; \epsilon_i.$$

The vector $\mathbf{Z} = (Z_i)$ is defined over all n_{tot} members of the pedigree, and has the multivariate Gaussian distribution $\mathbf{Z} \sim N(\mathbf{0}, \sigma_a^2 \mathbf{A})$, where $\mathbf{A}$ is a matrix determined by the pedigree structure (section 2.6).

If $I\{E\}$ is the indicator function of the event E, the complete-data sufficient statistics of this exponential family model for $(\mathbf{G}, \mathbf{Z}, \mathbf{Y})$ are:

the number of observed individuals of each genotype g, or $\sum_{i \in \mathcal{D}} I\{G_i = g\}$

the total trait effect in those individuals, $\sum_{i \in \mathcal{D}} (Y_i - Z_i) I\{G_i = g\}$

the quadratic residual term, for observed individuals,

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \;=\; (\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z})'(\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z}), \text{ and}$$

the total genetic variance over all pedigree members, $\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}$.

If genotypes G_i and polygenic values Z_i were observable, then the MLEs of the parameters would be straightforward. For each discrete genotype g

$$\widehat{\mu(g)} \;=\; \frac{\sum_{i \in \mathcal{D}} (Y_i - Z_i) I\{G_i = g\}}{\sum_{i \in \mathcal{D}} I\{G_i = g\}}.$$

For the variance component parameters (see equation (2.17))

$$\begin{aligned} \widehat{\sigma_e^2} &= (\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z})'(\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z})/n_{obs} \\ \widehat{\sigma_a^2} &= (\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z})/n_{tot} \end{aligned}$$

where $\mu(\mathbf{G})$ denotes the vector of genotypic values of observed individuals $(\mu(G_i);\ i \in \mathcal{D})$. However, for this mixed model, both exact implementation of an EM algorithm and exact evaluation of the likelihood are infeasible. Monte Carlo methods can, however, be implemented.

For example, the conditional expectations of the statistics in the above equations, given the data $\mathbf{Y}$, may be estimated by averaging the values given by N realizations

$(\mathbf{Z}^{(\tau)}, \mathbf{G}^{(\tau)})$. At current parameter values, $(\sigma_e^2, \sigma_a^2, \boldsymbol{\mu})$, realizations are obtained from the conditional distribution $P_{\sigma_e^2,\sigma_a^2,\boldsymbol{\mu}}(\mathbf{Z}, \mathbf{G} \mid \mathbf{Y})$, leading to Monte Carlo EM update equations

$$\mu(g)^* = (N)^{-1} \frac{\sum_{\tau=1}^{N} \sum_{i \in \mathcal{D}} (Y_i - Z_i^{(\tau)}) I\{G_i^{(\tau)} = g\}}{\sum_{\tau=1}^{N} \sum_{i \in \mathcal{D}} I\{G_i^{(\tau)} = g\}}$$

$$\sigma_e^{2*} = (N n_{obs})^{-1} \sum_{\tau=1}^{N} (\mathbf{Y} - \mu(\mathbf{G}^{(\tau)}) - \mathbf{Z}^{(\tau)})' (\mathbf{Y} - \mu(\mathbf{G}^{(\tau)}) - \mathbf{Z}^{(\tau)}) \tag{9.5}$$

$$\sigma_a^{2*} = (N n_{tot})^{-1} \sum_{\tau=1}^{N} \mathbf{Z}^{(\tau)\prime} \mathbf{A}^{-1} \mathbf{Z}^{(\tau)}. \tag{9.6}$$

Equations (9.5) and (9.6) should be compared with the exact EM equations for the parameters of a polygenic model (equation (2.17)). With a Monte Carlo approach, the conditional variance of $\mathbf{Z}$ given $\mathbf{Y}$ and $\mathbf{G}$ need not be computed, since the variance is subsumed into the realized variability of the quadratic expressions. Note, however, that this variance is an intrinsic part of the iterative procedure. Just as in section 2.6, it is insufficient to use only the estimate $\tilde{\mathbf{a}} = N^{-1} \sum_{\tau=1}^{N} \mathbf{Z}^{(\tau)}$ of the conditional mean $\mathbf{a} = \mathrm{E}_{\sigma_e^2,\sigma_a^2,\boldsymbol{\mu}}(\mathbf{Z} \mid \mathbf{G}, \mathbf{Y})$.

Returning to single-locus models, if genotypes $\mathbf{G}$ can be realized given the data $\mathbf{Y}$ and current parameter values, MCEM equations for parameters of penetrance densities are straightforward. The use of MCEM also permits extension to more complex models. One example is that of a more general mixed model for a quantitative trait, including also the effects of observed covariates and other variance component effects, such as those due to shared environment. This model assumes the trait value y_i is the sum of these effects together with the effect of a single-locus genotype G_i, a polygenic value z_i, and a residual with mean 0 and variance σ_e^2. Provided genotypes and polygenic values $(\mathbf{G}, \mathbf{Z})$ can be realized, conditional upon data $\mathbf{Y}$ and current parameter values, MCEM is again feasible. Achieving these realizations is not, in general, straightforward. We can do so by using Markov chain Monte Carlo (MCMC). Guo and Thompson (1992; 1994) have used MCEM for the mixed model and for joint linkage and segregation analysis. Generally, MCEM is as effective as EM at getting a ball-park estimate, and is remarkably robust even when quite small Monte Carlo samples are used. However, it is of little use in obtaining a precise final MLE—a large number of very large samples would be required.

9.4 Likelihood estimators for complex models

The mixed model also provides an example of Rao-Blackwellization (section 3.8) of Monte Carlo estimates of likelihood ratios. Applying the formula (9.1) directly to the mixed model with latent variables $(\mathbf{G}, \mathbf{Z})$, we have

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{G}, \mathbf{Z}, \mathbf{Y})}{P_{\theta_0}(\mathbf{G}, \mathbf{Z}, \mathbf{Y})} \;\middle|\; \mathbf{Y} \right).$$

However, considering only the latent variables $\mathbf{G}$, it is also the case that

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y}|\mathbf{G})P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{Y}|\mathbf{G})P_{\theta_0}(\mathbf{G})} \;\middle|\; \mathbf{Y}\right), \tag{9.7}$$

while considering only latent variables $\mathbf{Z}$

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y}|\mathbf{Z})P_\theta(\mathbf{Z})}{P_{\theta_0}(\mathbf{Y}|\mathbf{Z})P_{\theta_0}(\mathbf{Z})} \;\middle|\; \mathbf{Y}\right). \tag{9.8}$$

Since $\mathbf{Y}$ and $\mathbf{Z}$ are continuous random variables, we have now probability density functions rather than probability mass functions. However, we retain the notation $P_\theta(\cdot)$, to avoid introducing additional notation for this one example.

As shown in section 6.6, either integration over $\mathbf{Z}$ or summation over $\mathbf{G}$ is possible in the mixed-model likelihood (equation (6.6)), providing for exact computation of the probabilities

$$P_\theta(\mathbf{Y}|\mathbf{G}) = \int_{\mathbf{z}} \Pr(\mathbf{Y}|\mathbf{z},\mathbf{G})P_{\sigma_a^2}(\mathbf{z})d\mathbf{z} \tag{9.9}$$

in equation (9.7), or of the probabilities

$$P_\theta(\mathbf{Y}|\mathbf{Z}) = \sum_{\mathbf{G}} P_\theta(\mathbf{Y}|\mathbf{Z},\mathbf{G})P_\theta(\mathbf{G})$$

in equation (9.8). Equations (9.7) and (9.8) provide two alternative Rao-Blackwellized estimators. To implement the estimate based on (9.7) or on (9.8), only the realizations of $\mathbf{G}$ or of $\mathbf{Z}$ would be used. However, if using a Markov chain Monte Carlo (MCMC) sampler (Chapter 8), it will normally be necessary to generate both. For example, from N Monte Carlo realizations $(\mathbf{G}^{(\tau)},\mathbf{Z}^{(\tau)})$ generated from $P_{\theta_0}(\mathbf{G},\mathbf{Z}|\mathbf{Y})$, the estimate based on equation (9.7) would be

$$\widehat{\frac{L(\theta)}{L(\theta_0)}} = \frac{1}{N}\sum_{\tau=1}^{N} \frac{P_\theta(\mathbf{Y}|\mathbf{G}^{(\tau)})P_\theta(\mathbf{G}^{(\tau)})}{P_{\theta_0}(\mathbf{Y}|\mathbf{G}^{(\tau)})P_{\theta_0}(\mathbf{G}^{(\tau)})}. \tag{9.10}$$

The hope is that the reduction in Monte Carlo variance due to the partial exact computation (9.9) will compensate for this increased computation (see section 3.8).

Note also that, for some model comparisons, additional exact integration or summation over latent variables $\mathbf{Z}$ or $\mathbf{G}$ may be unnecessary. If under models indexed by θ and by θ_0,

$$P_\theta(\mathbf{Y}|\mathbf{G}) = P_{\theta_0}(\mathbf{Y}|\mathbf{G})$$

these probabilities need not be computed, and the estimate (9.10) reduces to a ratio of the prior genotype probabilities $P_\theta(\mathbf{G})/P_{\theta_0}(\mathbf{G})$ averaged over the realized $\mathbf{G}^{(\tau)}$. Similarly, if

$$P_\theta(\mathbf{Y}|\mathbf{Z}) = P_{\theta_0}(\mathbf{Y}|\mathbf{Z}),$$

the estimator based on equation (9.8) reduces to the ratio of population densities of $\mathbf{Z}$. By careful choice of models to be compared, procedures can be made more computationally efficient.

Reduction in Monte Carlo variance by Rao-Blackwellization is guaranteed only for independent realizations $(\mathbf{G}^{(\tau)}, \mathbf{Z}^{(\tau)})$ of the latent variables (Geyer, 1992). For dependent realizations, Geyer (pers.comm.) has provided a simple counter-example based on latent variables consisting the odd and even terms of a first order Gaussian autoregressive process. However, in many practical instances the Rao-Blackwellization procedure works well, even when MCMC realizations are used. Estimators based on (9.7) and (9.8) were introduced by Thompson and Guo (1991) and compared by Thompson (1994*c*) in likelihood analyses of genetic models with several latent heritable components. It was found that the estimator (9.10) works very well, leading to substantial gains in computational efficiency, whereas the estimator based on (9.8) is very inefficient. The summation over $\mathbf{G}$ required for the latter is generally computationally more intensive than integration over $\mathbf{Z}$. More importantly, the data $\mathbf{Y}$ and variables $\mathbf{Z}$ together constrain $\mathbf{G}$ very much more than $\mathbf{Y}$ and $\mathbf{G}$ constrain $\mathbf{Z}$. Since the conditional variance of $\mathbf{Z}$ given $\mathbf{Y}$ and $\mathbf{G}$ is relatively high, exact integration over $\mathbf{Z}$ reduces Monte Carlo variance substantially.

Note that equation (9.1), or forms thereof such as (9.7) and (9.8), are not the only possible ways to obtain Monte Carlo estimates of likelihood ratios. In particular, Meng and Wong (1996) have considered a variety of forms of importance sampling and Rao-Blackwellization, noting that (in the notation of this chapter)

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \frac{\mathrm{E}_{\theta_0}(P_\theta(\mathbf{X},\mathbf{Y})\alpha(\mathbf{X}) \mid \mathbf{Y})}{\mathrm{E}_\theta(P_{\theta_0}(\mathbf{X},\mathbf{Y})\alpha(\mathbf{X}) \mid \mathbf{Y})} \tag{9.11}$$

where $\alpha(\mathbf{X})$ is an arbitrary function on the space of $\mathbf{X}$ values (provided the expectations exist, and the distributions have the same support). If $\alpha(\mathbf{X}) = 1/P_{\theta_0}(\mathbf{X},\mathbf{Y})$, equation (9.11) reduces to equation (9.1). Note that whereas use of equation (9.1) requires MCMC only at θ_0, the expectation in the denominator of equation (9.11) requires MCMC at the value θ. Various other choices of $\alpha(\mathbf{X})$ have been investigated in the recent MCMC literature. Jensen and Kong (1999) have used a version of equation (9.11) in their MCMC estimation of a single-marker linkage lod score on a complex pedigree.

As for the ratio estimator (7.17), the expression (9.11) is a ratio of expectations, and thus the Monte Carlo estimator is a ratio of averages over two sets of Monte Carlo realizations. For the estimator based on (7.17), the sampling distribution is the same in numerator and denominator, and thus Monte Carlo variance could be reduced, and computational efficiency enhanced, by using the same Monte Carlo realizations in the estimates of numerator and denominator. However, for the likelihood ratio estimator based on (9.11), different sampling distributions are required, so different Monte Carlo Markov chains must be run for the numerator and denominator. If MCMC is being done in any case at a set of values θ_j, for example as in section 9.2, this does not impose any increased computational burden for the Monte Carlo itself. However, long runs may be needed to reduce the Monte Carlo variance of the estimate of $L(\theta)/L(\theta_0)$ to acceptable levels.

9.5 Likelihood estimation of gene locations

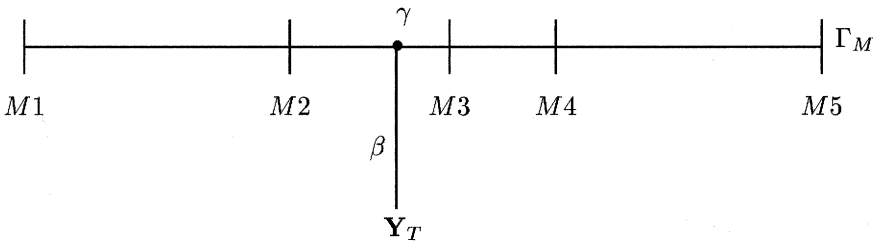

FIGURE 9.1. *Model parameters for estimation of a location likelihood curve*

In modern genetic analysis, a primary goal is the localization of trait genes. Genetic markers have been mapped throughout the genome at a scale suitable for multipoint linkage analysis. Thus, estimation of *location lod score curves* (sections 6.2, 7.6) is an important goal. Here we denote the marker model parameters by Γ_M. For a complex trait, the trait model parameters are also unknown. These parameters, β, determine the probabilities of phenotypes given the latent genes. While in some analyses, joint maximization of the likelihood with respect to trait model β and trait locus position γ may be attempted, often the *location lod score curve* is computed for fixed β. The likelihood (or a profile likelihood) is evaluated as a function of a hypothesized trait-locus location γ, against a fixed marker map Γ_M. The parametrization of the overall model is shown in Figure 9.1 The overall model is indexed by parameter $\theta = (\beta, \gamma, \Gamma_M)$. As before, the likelihood is

$$L(\theta) \;=\; P_\theta(\mathbf{Y}) \;=\; \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X})\, P_\theta(\mathbf{X})$$

(equation (7.8)) which may take the form (1.5) if $\mathbf{X} = \mathbf{G}$, the underlying genotypes, or (4.11) if $\mathbf{X} = \mathbf{S}$, the inheritance patterns of genes. For computations with multiple marker loci, the Lander-Green paradigm (4.11) is more natural and more effective, but exact computation is limited to small pedigrees.

As in the discussion of *Elods* (equations (4.8) and (7.15)), for convenience we partition the data $\mathbf{Y}$ into the trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$. The corresponding latent variables are partitioned into $\mathbf{X}_T$ and $\mathbf{X}_M$. Monte Carlo estimation of the location likelihood ratio is always feasible. The form that follows directly from equation (9.1) is

$$\frac{L(\beta,\ \gamma_1,\ \Gamma_M)}{L(\beta,\ \gamma_0,\ \Gamma_M)} \;=\; \mathrm{E}_{\theta_0}\left(\frac{P_{\theta_1}(\mathbf{Y}_T, \mathbf{Y}_M \mid \mathbf{X}_T, \mathbf{X}_M) P_{\theta_1}(\mathbf{X}_T, \mathbf{X}_M)}{P_{\theta_0}(\mathbf{Y}_T, \mathbf{Y}_M \mid \mathbf{X}_T, \mathbf{X}_M) P_{\theta_0}(\mathbf{X}_T, \mathbf{X}_M)} \;\middle|\; \mathbf{Y}_T, \mathbf{Y}_M \right)$$

for two hypothesized trait locus positions γ_1 and γ_0. Noting the fact that only the position of the trait locus differs between numerator and denominator, the above

equation reduces to

$$\frac{L(\beta,\ \gamma_1,\ \Gamma_M)}{L(\beta,\ \gamma_0,\ \Gamma_M)} = \mathrm{E}_{\theta_0}\left(\frac{P_{\gamma_1}(\mathbf{X}_T \mid \mathbf{X}_M)}{P_{\gamma_0}(\mathbf{X}_T \mid \mathbf{X}_M)} \;\middle|\; \mathbf{Y}_T, \mathbf{Y}_M\right). \tag{9.12}$$

Thus only the conditional probability of trait-locus latent variables given marker-loci latent variables appears explicitly in the estimator. Although realization of the latent variables is complex, and requires MCMC methods, computation of the estimate from the realizations is generally very straightforward (Thompson and Guo, 1991).

One practical difficulty of the above approach is accurate estimation of log-likelihood differences for trait locations in different marker intervals. The likelihood-ratio estimate (9.1) works well in comparing locations within an interval, and in principle the mixtures method (9.2) facilitates estimation between intervals. However, in practice, values of $\mathbf{X}_T$ realized at γ_0 may have very small probabilities under γ_1 if there is a marker locus between the two positions. Additionally, the usual objective is to estimate the lod-score relative to the base-point in which the trait locus position γ is not within the marker map Γ_M. That is, the null hypothesis that the trait locus is unlinked. Again this can be accomplished by using the mixtures method (9.2), but several intervening positions γ, linked to but not within the marker map, may be required for effective estimation (Thompson, 1994*b*). The procedure becomes computationally intensive.

Another disadvantage of the approach of section 9.1 and this section is the fact that it allows estimation only of likelihood ratios, not of likelihoods. A modification is due to Lange and Sobel (1991) for the particular case of Monte Carlo estimation of location likelihoods. Their procedure also avoids the problems of sampling of the trait locus variables. Again, we assume the marker map and parameters Γ_M known, so that $P_{\Gamma_M}(\mathbf{Y}_M)$ is a constant factor in the likelihood. Then Lange and Sobel (1991) write the likelihood in a form which, using our current notation, becomes

$$\begin{aligned} L(\beta,\gamma,\Gamma_M) &= P_{\beta,\gamma,\Gamma_M}(\mathbf{Y}_M,\mathbf{Y}_T) \\ &\propto P_{\beta,\gamma,\Gamma_M}(\mathbf{Y}_T \mid \mathbf{Y}_M) \\ &= \sum_{\mathbf{X}_M} P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M) P_{\Gamma_M}(\mathbf{X}_M \mid \mathbf{Y}_M) \\ &= \mathrm{E}_{\Gamma_M}(P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M) \mid \mathbf{Y}_M). \end{aligned} \tag{9.13}$$

Now latent variables $\mathbf{X}_M$ are sampled from their conditional distribution given the marker data $\mathbf{Y}_M$. Provided exact computation of $P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M)$ is possible for alternative trait models (β) and locations (γ), we have a Monte Carlo estimate of $L(\beta,\gamma,\Gamma_M)$. Comparison to the unlinked base-point requires only computation of $P_\beta(\mathbf{Y}_T)$, the probability of trait data under the parameters β of the trait locus model. This can be accomplished by single-locus peeling methods of Chapter 6. Since Γ_M is fixed, the Monte Carlo requires only a single set of realizations $\{\mathbf{X}_M^{(\tau)}, \tau = 1,\ldots,N\}$. The disadvantage is that $P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M^{(t)})$ must be computed for each such realization; this requires a single-locus peeling computation for the

trait-locus data under the trait model. Further, this computation must be done, not only for each realization $\mathbf{X}_M^{(\tau)}$, but also for each β and γ at which a likelihood estimate is required.

In many cases, however, the gains outweigh the costs, except when the simulation distribution $P_{\Gamma_M}(\mathbf{X}_M \mid \mathbf{Y}_M)$ is not close to proportional to the ideal importance-sampling target distribution $P_{\beta,\gamma,\Gamma_M}(\mathbf{X}_M \mid \mathbf{Y}_M, \mathbf{Y}_T)$. This is particularly so for models (trait locations) γ which are not close to the truth, and for a trait which provides substantial information about the inheritance patterns of genes at the underlying trait locus, and hence also at linked marker loci. In fact, the cases where the Monte Carlo estimator based on equation (9.13) performs poorly are precisely those in which the likelihood ratio estimator (9.12) also has difficulties. There continue to be interesting open questions in the estimation of multilocus linkage likelihoods.

9.6 Marker *ibd* and complete-data log-likelihoods

Again suppose that, as in sections 7.5 and 9.5, we have trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$. Further, suppose that the marker map Γ_M, marker allele frequencies and marker population genotype frequencies are known, Consider also the case where the latent variables $\mathbf{X}_M$ are the meiosis indicators $\mathbf{S}_M$. As described above, MCMC methods, and in particular the M-sampler of section 8.4, provide effective methods for sampling from the conditional distribution $P_{\Gamma_M}(\mathbf{S}_M \mid \mathbf{Y}_M)$. Among pedigree members, the patterns of gene *ibd* at marker loci are functions of $\mathbf{S}_M$; $\mathbf{J} = \mathbf{J}(\mathbf{S})$ (section 3.6). Thus we have MCMC estimates of the conditional probabilities of gene *ibd* at marker loci, given the marker data.

Neither trait data, $\mathbf{Y}_T$ nor trait model enter into this sampling of marker latent variables conditional on marker data. However, under any trait model with some genetic component, related affected individuals or related individuals exhibiting extreme trait values will share genes *ibd* at trait loci with some increased probability. Hence also they will share genes *ibd* with increased probability at marker loci linked to those trait loci. In so-called "non-parametric" computations for linkage detection, marker data on a pedigree are analyzed to detect regions of the genome in which there is evidence for excess gene *ibd* among affected individuals, or individuals exhibiting extreme trait values. Such regions provide evidence for linkage.

The Monte Carlo sampling of $\mathbf{S}_M$ given marker data $\mathbf{Y}_M$ provides direct estimates of conditional probabilities of patterns of gene *ibd* $\mathbf{J}(\mathbf{S}_M)$. These gene *ibd* probabilities at locus j are computed dependent on all the marker data $\mathbf{Y}_M$, as, for example, are the probabilities $Q_j(S_{\bullet,j})$ of section 7.1. Here we have only Monte Carlo estimates of these probabilities, but MCMC realization on larger or more complex pedigrees is feasible in cases for which exact computation is not. Moreover, the resulting gene *ibd* patterns $\mathbf{J}(\mathbf{S}_M)$ may be scored jointly over haplotypes, and over loci. The example of section 4.5 showed the importance of considering both individuals and loci jointly. For the case where only marker data are considered, many of the problems of the Monte Carlo estimation procedures are much reduced, provided a good MCMC sampler is used (see sections 8.3, 8.4).

The statistical problem becomes one of development of appropriate test statistics, to detect linkage on the basis of estimated conditional *ibd* probabilities. Although most current methods involve statistics computed marginally over loci, and pairwise over individuals, there is an increasing literature in this area; see for example Whittemore and Tu (1998).

Another readily computed by-product of MCMC on pedigrees, or in any latent-variable problem, is the expected complete-data log-likelihood, $H_{\mathbf{y}}(\theta;\theta_0)$ (section 2.4). Returning again to the full data $\mathbf{Y}$ and latent variables $\mathbf{S}$, at a general model indexed by parameters θ_0 we have

$$\begin{aligned} H_{\mathbf{y}}(\theta;\theta_0) &= \mathrm{E}_{\theta_0}(\log_e P_\theta(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y}) \\ &= \mathrm{E}_{\theta_0}(\log_e P_\theta(\mathbf{Y} \mid \mathbf{S}) + \log_e P_\theta(\mathbf{S}) \mid \mathbf{Y}). \end{aligned} \tag{9.14}$$

For easier comparison with statistical results, we use natural (base-e) logarithms throughout this section. Due to the *a priori* independence of meioses

$$\log P_\theta(\mathbf{S}) = \sum_{i=1}^{m} \log P_\theta(S_{i,\bullet}) \tag{9.15}$$

and, provided data are locus-specific,

$$\log P_\theta(\mathbf{Y} \mid \mathbf{S}) = \sum_{j=1}^{L} \log P_\theta(Y_{\bullet,j} \mid S_{\bullet,j}) \tag{9.16}$$

(see equation (4.11)). Thus the expectation partitions into terms for each locus and for each meiosis. These terms must be computed in any case in the course of the MCMC, making accumulation of values for the estimated expectation particularly straightforward. Note that (9.15) depends only on the genetic map parameters, while (9.16) depends on the penetrance aspects of the model. In expectation, under the conditional distribution $P_{\theta_0}(\cdot|\mathbf{Y})$, each term depends, of course, on all the parameters in θ_0. The expected complete-data log-likelihood, with its component parts, proves to be a useful diagnostic measure of the performance of the MCMC.

The above discussion depends on the decomposition

$$\log P_\theta(\mathbf{S},\mathbf{Y}) = \log P_\theta(\mathbf{Y} \mid \mathbf{S}) + \log P_\theta(\mathbf{S}).$$

Reversing the decomposition of the complete-data log-likelihood

$$\log P_\theta(\mathbf{S},\mathbf{Y}) = \log P_\theta(\mathbf{Y}) + \log P_\theta(\mathbf{S} \mid \mathbf{Y}).$$

Thus, as in equation (2.9), differences in expected complete-data log-likelihoods depend on the true log-likelihood difference and the Kullback-Leibler information (section 2.2) in the distribution of $\mathbf{S}$ given $\mathbf{Y}$. That is,

$$K_{\mathbf{y}}(\theta;\theta_0) = \mathrm{E}_{\theta_0}(\log_e P_{\theta_0}(\mathbf{S}|\mathbf{Y}) - \log_e P_\theta(\mathbf{S}|\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}),$$

so the difference in expected complete-data log-likelihoods is

$$\begin{aligned} H_{\mathbf{y}}(\theta_0;\theta_0) &- H_{\mathbf{y}}(\theta;\theta_0) \\ &= \mathrm{E}_{\theta_0}(\log P_{\theta_0}(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y}) - \mathrm{E}_{\theta_0}(\log P_{\theta}(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y}) \\ &= \mathrm{E}_{\theta_0}(\log P_{\theta_0}(\mathbf{Y}) + \log P_{\theta_0}(\mathbf{S}|\mathbf{Y}) - \log P_{\theta}(\mathbf{Y}) - \log P_{\theta}(\mathbf{S}|\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}) \\ &= \log P_{\theta_0}(\mathbf{Y}) - \log P_{\theta}(\mathbf{Y}) + \mathrm{E}_{\theta_0}(P_{\theta_0}(\mathbf{S}|\mathbf{Y}) - \log P_{\theta}(\mathbf{S}|\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}) \\ &= \ell(\theta_0) - \ell(\theta) + K_{\mathbf{y}}(\theta;\theta_0). \end{aligned} \tag{9.17}$$

The extent to which this identity can be exploited in making inferences from MCMC output is also an area of ongoing research.

Chapter 10

Case studies using the M- and LM-samplers

10.1 Background to a study

Much of the material in sections 10.1 to 10.5 has recently been published (Thompson, 2000*a*). It was presented at a one-day Royal Statistical Society conference in March 1999, and was discussed again in July 1999 at the CBMS Summer Course. Section 10.6 is the result of more recent work.

First, the methods of the previous chapters are illustrated using data based on an extended Icelandic pedigree, provided by Dr. J. H. Edwards. The trait, apparent in three families, was thought to be a simple recessive, with an animal analogue suggesting a possible location on human Chromosome 1 (Remmers et al., 1996). However, findings were negative, and for purposes of illustration Heath and Thompson (1997) simulated marker data, conditional on a recessive trait locus in the chromosomal region. The resimulation of data assumed the same marker locations, population allele frequencies, and marker phenotype availability as in the original data. Marker data were simulated conditional on descent paths at the trait locus that implied that the four affected final individuals would be so. No phenotypic assumptions were imposed for other pedigree members. Using these simulated data, there was some evidence for excess gene identity by descent among the six parents of affected individuals (Heath and Thompson, 1997). However, in attempting to analyze these simulated data, under the assumption of a rare recessive trait, findings were ambiguous, primarily due to the fact that no founders were ancestral to more than three of the six parents of the affected individuals, even though the ancestry of the families was fully traced for seven generations. Accounting for the affected individuals required three separate origins of the recessive disease allele within the pedigree. For current purposes, we have therefore also modified the pedigree structure, making possible a single ancestral origin of the disease allele, and realized disease ancestry accordingly (Figure 10.1).

Conditional on the realized gene ancestry, we have resimulated marker data.

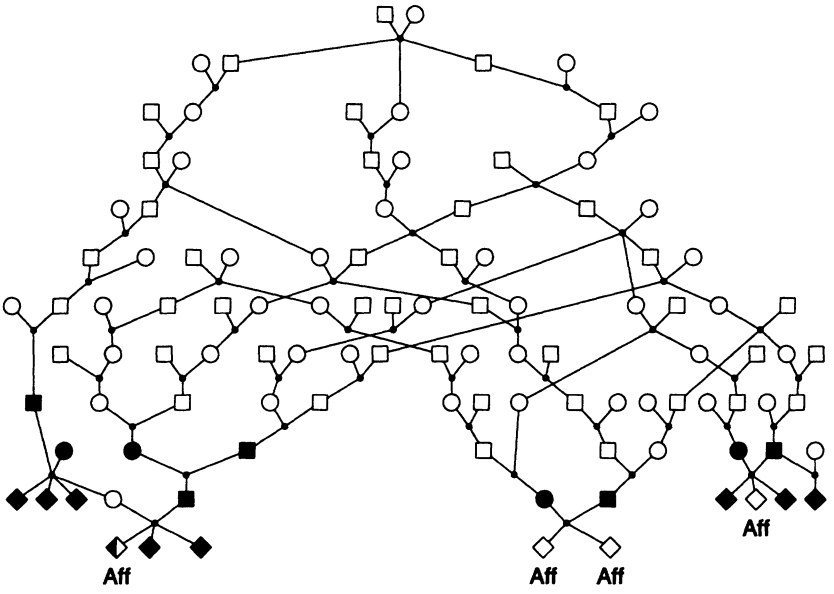

FIGURE 10.1. *The modified Icelandic pedigree. The four individuals marked "Aff" are affected. Those shaded black have marker data available at the majority of the 17 marker loci. The affected half-shaded individual is typed at only two of the marker loci*

Since the data are simulated, we avoid difficulties caused by errors in marker map or in meiosis model assumptions –for example we did not incorporate recombination heterogeneity between the sexes. The marker allele frequencies and data availability are as in the original data. Very few data are available on the affected individuals themselves (two markers on only one of the four cases), and overall the pedigree is quite sparsely observed, with the data being on the majority of the close relatives of affected individuals (Figure (10.1)).

There are data at 17 marker loci, some of which are quite polymorphic, exhibiting up to 7 alleles, even among the 18 observed individuals. Some were also tightly linked: indeed the ones adjacent to the putative trait locus were less than 2cM apart. In simulating marker data, using the original map, we obtained haplotype sharing among the three nuclear families containing affected individuals over 5 or 6 markers. To have data corresponding to modern linkage detection problems, we resimulated data using a genetic map with marker intervals at 10% recombination frequency, with the disease locus at the center of one interval (recombination 0.0528 to each flanking marker). Some realizations then gave almost no genes *ibd* among the three families at any marker locus, with obvious consequent problems for linkage detection. The necessary scale of the map is dependent on the pedigree structures

locus	number of alleles	true *ibd* state	true phenotypes
trait	2	a a a a a a	222222
M1	6	b d - m k z	226166
M2	7	b d g m k z	575373
M3	6	b d g m x z	656155
M4	6	b d g m x z	364442
M5	4	b e h m v z	333333
M6	3	b e h m v t	113311
M7	6	b e i a x t	643624
M8	7	b e i a n t	445426
M9	7	b a - a - a	576677
M10	6	a a a a a a	444444
M11	8	b a w w w y	378887
M12	4	b a v - w y	142123
M13	5	b f w - - y	312425
M14	6	b f w p x y	125323
M15	7	b e m p x y	156447
M16	7	b - n r x z	727324
M17	7	- - r w x z	136344

TABLE 10.1. *True gene identity by descent simulated on the modified Icelandic pedigree*

available for analysis (Thompson, 1997). However, our chosen data realization did exhibit a gene *ibd* in all six parents of affected individuals at one of the two markers flanking the disease locus. Since the data are simulated the "true" trait location is known; this is mid-way between markers M10 and M11.

The simulated data in three affected offspring individuals are shown in Table 10.1. For true *ibd* status, each letter indicates a different founder haplotype. A founder origin occurring once only in the set of six haplotypes is denoted "-". The disease allele at the trait locus is allele "2". Note that, apart from data at two loci for one individual, the marker types of affected individuals are not observed. Observations are available only on relatives of these individuals. Thus, at locus M4, although there are only three like alleles in the six haplotypes of affected individuals, the observed data permit the possibility that four of the six genes are *ibd*.

10.2 Conditional gene *ibd* probabilities

Given the trait and marker phenotypic data, we first analyzed conditional probabilities of gene identity by descent among haplotypes segregating from each member of each of the three parent couples with affected offspring. The marker allele frequencies and recombination probabilities used in simulating the data were assumed in the analysis. The trait allele frequency was assumed to be $q = 0.001$. This low value makes very probable a single origin of the disease allele in the

Locus	Probability ×1000 All non-*ibd*				Probability ×1000 4 or more *ibd*			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
trait	937	0	0	0	0	978	997	969
M1	879	988	907	746	0	0	0	0
M2	906	999	879	306	0	0	0	0
M3	975	925	906	10	0	0	0	0
M4	874	863	808	17	0	0	0	425
M5	924	843	755	263	0	0	1	39
M6	931	726	742	532	0	0	0	2
M7	901	971	682	689	0	0	0	0
M8	919	751	414	589	0	0	0	0
M9	864	685	28	458	0	2	508	18
M10	676	539	0	387	4	7	982	30
M11	872	434	13	532	0	0	0	0
M12	879	406	180	598	0	0	0	0
M13	870	643	370	672	0	0	0	0
M14	872	867	589	773	0	0	0	0
M15	988	894	978	988	0	0	0	0
M16	947	916	963	980	0	0	0	0
M17	993	894	981	990	0	0	0	0

TABLE 10.2. *Conditional probabilities of gene identity by descent given the marker data simulated on the modified Icelandic pedigree. Shown are probabilities* ×1000. *For details of the cases (1)–(4), see text*

pedigree. There are in all 39 founders in the pedigree, and hence 78 founder genes at the disease locus, but only the four in the original couple are ancestral to all the six carrier parents of affected individuals. The 203 (potential) patterns were scored marginally at each marker. Several cases were considered:

(1) All markers and the trait locus independently segregating (unlinked). A null trait locus provides the single-locus prior probability of *ibd* given only the pedigree structure

(2) Correct map for marker data. The correct recessive trait model and affected trait status of individuals is assumed, but the trait locus is modeled as unlinked to the markers.

(3) Correct map, with trait locus in correct location between M10 and M11.

(4) Correct marker map, with trait locus in incorrect position between M3 and M4.

For the trait locus, one member of the original couple was specified to be a heterozygous carrier, and the other a non-carrier. No assumptions were made about the trait genotypes of any other founders or ancestors. The affected individuals whose haplotypes were scored were assumed homozygous for the disease allele.

The results are summarized in Table 10.2. Each column refers to the specified one of the four cases (1)-(4) given above. The table consists of probabilities multiplied

by 1000 for ease of presentation. The first set of four columns gives the conditional probabilities of no *ibd* among the six haplotypes. The second block of four columns gives the total conditional probability of *ibd* patterns in which at least four of the six haplotypes are *ibd*. The MCMC incorporates jointly the information from all linked loci, although the conditional probabilities are here summarized marginally for each locus. Since the MCMC runs jointly over loci, scoring of joint realized patterns is also possible.

The sampler used here is the M-sampler (section 8.4), so one MCMC step consists of resampling the meiosis indicators $S_{i,j}$ jointly for all 18 loci for a randomly chosen meiosis i. Each run consists of 10^7 meiosis MCMC steps, and takes about 12 hours on a workstation running a shared LSBatch system. States of *ibd* are only output if the sum over loci of the estimated conditional probability is greater than 0.001. Given the marker data, more states are thus feasible than are given in the output summary: states which were realized in the MCMC with low frequencies do not appear.

Although the marker data alone do not suggest high levels of gene *ibd* among affecteds, the conditional probability of some *ibd* among the six haplotypes in the region of the true trait location (M9,M10,M11) is high. Even independently, column (1), there is evidence of some gene *ibd* in this region, particularly at marker M10. The values in this column may be compared with the single-locus prior. Based only on the pedigree structure, the probability of no gene *ibd* is 0.937. The trait locus itself contains a lot of the information on segregation (Table 10.2). Even in the absence of marker data, the trait information reduces the probability of no gene *ibd* among the six haplotypes from 0.937 to close to 0, and increases the probability of four or more haplotypes *ibd* from close to 0 to 0.978.

When the trait locus is hypothesized in its true position, very high levels of gene *ibd* are estimated at the adjacent marker M10, while the high levels at the trait locus itself are reinforced. Disconcertingly, when the trait locus is hypothesized in an incorrect position, *ibd* at the trait locus is only slightly decreased, while estimated *ibd* probabilities at loci in the region of this incorrect position (M3, M4, M5) are much increased. The strength of the information provided by the segregation pattern of this rare recessive trait makes inference of gene location difficult. Since marker data are very sparse on the pedigree, it is possible for the marker descent patterns to adapt to alternative hypothesized gene locations.

10.3 Likelihoods and log-likelihoods

We then attempted a Monte Carlo estimate of the full location lod score, assuming each of the six parents of affected individuals to be heterozygous for a very rare recessive trait allele. However, the Monte Carlo likelihood estimation methods of Chapter 9 failed to converge. A plot of the expected base-e complete-data log-likelihoods (equation (9.14)) from this same Monte-Carlo run reveals why (Figure 10.2). The MCMC was performed at hypothesized trait locus positions γ_0 in the center of each marker interval, at positions linked but outside the span of the markers, and also with the trait locus unlinked. The complete-data log-likelihood

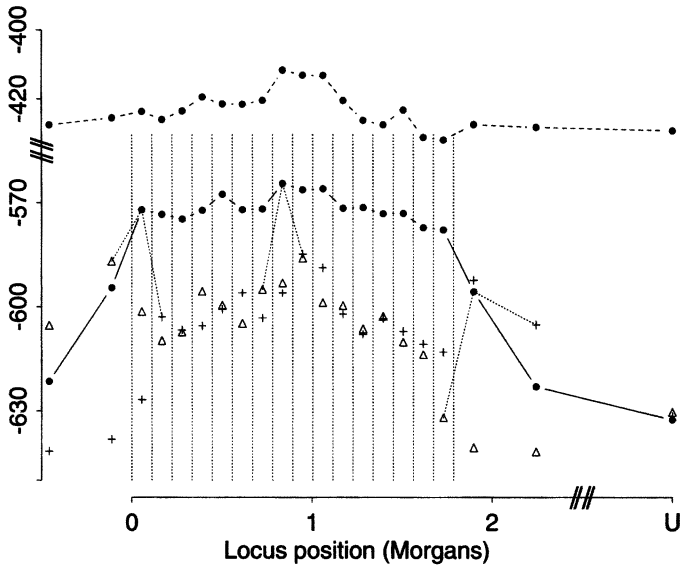

FIGURE 10.2. *Expected complete-data log-likelihood components for the simulated data on the modified Icelandic pedigree. Shown are* $\mathrm{E}_{\gamma_0}(\log_e \Pr(\mathbf{Y} \mid \mathbf{S}) \mid \mathbf{Y})$ *(upper curve), and* $\mathrm{E}_{\gamma_0}(\log_e P_\gamma(\mathbf{S}) \mid \mathbf{Y})$ *for* $\gamma = \gamma_0$ *(•, lower curve), and for* γ *to the left (△) and right (+) of* γ_0*. The location U denotes unlinked. For additional details see text*

is partitioned into segregation (equation (9.15)) and penetrance (equation (9.16)) parts. The figure shows first the penetrance contribution $\mathrm{E}_{\gamma_0}(\log \Pr(\mathbf{Y} \mid \mathbf{S}) \mid \mathbf{Y})$ to the expected complete-data log-likelihood for each simulation position (upper curve). This conditional probability does not directly depend on the hypothesized trait location, γ, although the expected log-probability does so through the realized $\mathbf{S}$. The segregation contribution $\mathrm{E}_{\gamma_0}(\log P_\gamma(\mathbf{S}) \mid \mathbf{Y})$ depends both on the simulation location γ_0, and on the evaluation location γ. The figure shows the values for each simulation position γ_0 (lower curve), with evaluations at γ_0 and at positions one step to the left (△) and to the right (+). Shown also are three example connections of realizations at a given γ_0, shown as •, with the same realizations evaluated to the left (△) and right (+). These log-likelihood differences are of order 25, indicating that $\mathbf{S}$-values realized at a given γ_0 are of order e^{25} less probable under neighboring values: it is unsurprising the Monte Carlo estimation of the likelihood is infeasible.

The expected complete-data log-likelihood is not only useful in diagnosing failure; it also provides some evidence regarding alternative models. For the four cases (1)–(4) considered in section 10.2, the complete-data base-e log-likelihoods averaged over each run are -1704, -1061, -982 and -998 respectively. Clearly, the assumption that the marker loci are unlinked (-1704) is unwarranted. The other three runs assume the correct marker map, with the trait locus unlinked, correctly positioned, and incorrectly positioned, respectively. The largest expected complete-date log-

likelihood is obtained when the model is correct, while the value under the model that the trait locus is unlinked is almost 80 units smaller. Summing the two curves, for the penetrances, $\log \Pr(\mathbf{Y}|\mathbf{S})$, and segegrations, $\log \Pr(\mathbf{S})$ in Figure 10.2, we see that the maximum expected complete-data log-likelihood is obtained for trait locations between marker M8 and marker M11. Within this range there is little discrimination, but outside these three marker intervals both segregation and penetrance contributions decrease markedly.

10.4 Gene *ibd* in a smaller example

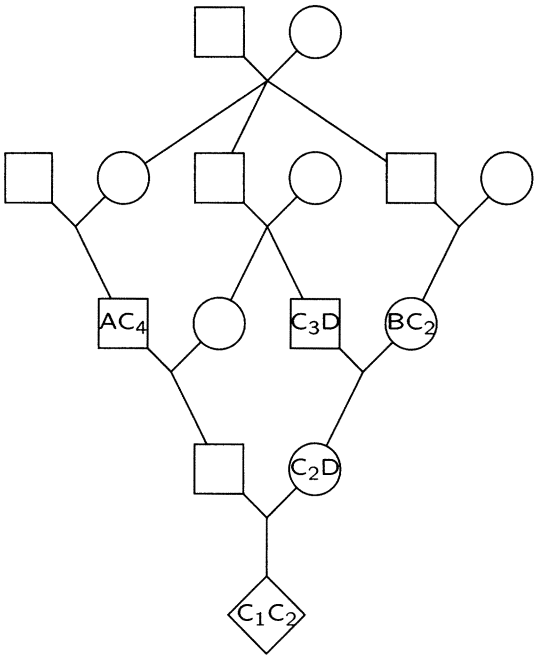

FIGURE 10.3. *Hypothetical phenotypic data assumed at each marker locus on the pedigree of Figure 1.1. The four potentially distinct C alleles are labeled C_1 to C_4*

To examine the performance of the MCMC method in more detail, we consider a smaller example, returning again to the pedigree of Figure 1.1. We suppose marker data as in Figure 3.5 at each of five marker loci, with recombination frequency 20% between adjacent markers (genetic distance 25.54 cM under a no-interference meiosis model). The trait data, for a rare recessive trait, is only that the final

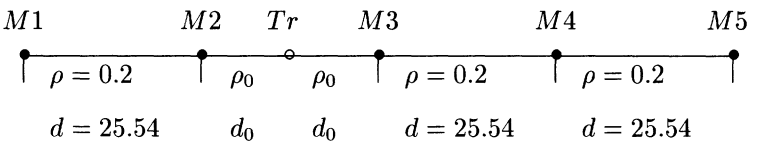

FIGURE 10.4. *Marker (M1 to M5) and trait (Tr) locations for the example of Figure 10.3. The trait locus is at the midpoint of the (M2, M3) interval, so* $d_0 = 12.77cM$ *and* $\rho_0 = 0.1187$

gene *ibd* pattern	pedigree prior	marker loci M3	marker loci M5	trait locus $q = 0.001$	trait locus $q = 0.05$
all 4 genes *ibd*	29	182	127	275	189
3 of 4 genes *ibd*	156	381	355	400	317
2 pairs of *ibd* genes	84	129	119	85	88
2 of 4 genes *ibd*	484	250	303	238	327
all 4 non-*ibd*	247	58	96	2	79
complete-data log-likelihood: segregations				-44.7	-45.1
complete-data log-likelihood: penetrances				-40.2	-37.1

TABLE 10.3. *Conditional probabilities* (×1000) *of gene ibd among the four C alleles on the pedigree of Figure 10.3, with five equally spaced marker loci, M1 to M5, and for a recessive trait unlinked to the markers*

	trait with $q = 0.001$			trait with $q = 0.05$		
gene *ibd* pattern	trait	M3	M5	trait	M3	M5
all 4 genes *ibd*	390	344	155	361	326	152
3 of 4 genes *ibd*	530	394	372	504	446	370
2 pairs of *ibd* genes	27	78	122	40	84	121
2 of 4 genes *ibd*	53	176	279	86	130	283
all 4 non-*ibd*	0	8	72	9	14	74
complete-data log-likelihood						
segregations		-38.5			-38.6	
penetrances		-39.1			-35.6	

TABLE 10.4. *Conditional probabilities* (×1000) *of gene ibd among the four C alleles on the pedigree of Figure 10.3, with five equally spaced marker loci, M1 to M5. The trait is now in the map, midway between M2 and M3*

individual of the pedigree is affected. Initially a trait allele frequency of $q = 0.001$ was assumed, although a value $q = 0.05$ was also considered.

At each of the five marker loci, frequencies 0.2, 0.2, 0.4 and 0.2 were assumed for alleles A, B, C, and D, respectively. Of particular interest in this test example is the potential for gene *ibd* among 4 potentially distinct C alleles, labeled C_1 to

C_4 in Figure 10.3. At each marker locus, given these marker phenotypes, all 15 possible patterns of gene *ibd* among these four C-alleles are possible. The C allele was given a relatively high frequency in order to give the possibility of four distinct origins non-negligible probability, while in contrast the trait was assumed rare to give high conditional probability that the affected individual is autozygous (has two *ibd* genes) at the trait locus.

Tables 10.3 and 10.4 summarize the conditional probabilities, when markers are run unlinked to the trait locus, and when the locus is in the mid-point of the second of the four marker intervals (Figure 10.4). Trait allele frequencies of $q = 0.001$ and $q = 0.05$ were each used. Each run consists of 10^7 M-sampler steps, each step selecting a random meiosis for update. We see a similar pattern to the example of section 10.2. Table 10.3 shows that each of the marker and trait data separately increases the conditional probability of gene *ibd* among the like alleles and decreases the probability of non-identity, relative to the prior based only on the pedigree structure. When the trait locus is within the marker map, the trait and marker data together reinforce the inference of gene *ibd* (Table 10.4). However, the effect of hypothesizing a trait location within the map on the inference of *ibd* at the marker loci is not nearly as strong as for the example of section 10.2. The effects are stronger for a rarer trait, both when unlinked (Table 10.3) and when linked (Table 10.4). However, the 50-fold change in allele frequency from $q = 0.001$ to $q = 0.05$ has a relatively minor effect. Of course, when the trait is unlinked, changing trait allele frequency does not impact marker *ibd*. When the trait is linked, the impact of trait data on marker *ibd* is larger for the adjacent marker M3 than for the terminal marker M5. The 50-fold difference in trait allele frequency (Table 10.4) has a moderate impact at the adjacent marker M3, but almost no impact at the terminal marker M5. The total complete-data log-likelihoods are larger when the trait is in the map, indicating evidence for linkage. The penetrance terms differ by about 3 between $q = 0.001$ and $q = 0.05$, the latter value giving higher probabilities.

10.5 MCMC lod score estimation

For the example of section 10.4, exact lod scores can be computed using GENEHUNTER 2 (Kruglyak et al., 1996; Kruglyak and Lander, 1998). These are shown in Figure 10.5. The two solid lines show the base-10 lod scores for trait locus position when the previous the marker data are assumed on five members of the pedigree. The higher curve corresponds to a trait allele frequency $q = 0.001$, and the lower to $q = 0.05$. The two broken curves show the base-10 lod scores when the marker data consist only of the final individual being homozygous for marker allele C with allele frequency 0.4, at each of the five marker loci. Again the upper curve is for $q = 0.001$ and the lower for $q = 0.05$. Note that the differences between the lod score curves for $q = 0.001$ and $q = 0.05$ are not large, although there is more evidence for linkage when a rarer trait frequency is assumed. This is a 50-fold change in allele frequency, and thus a 2500-fold change in the frequency of the recessive phenotype. Since there are only five founders in the pedigree, even

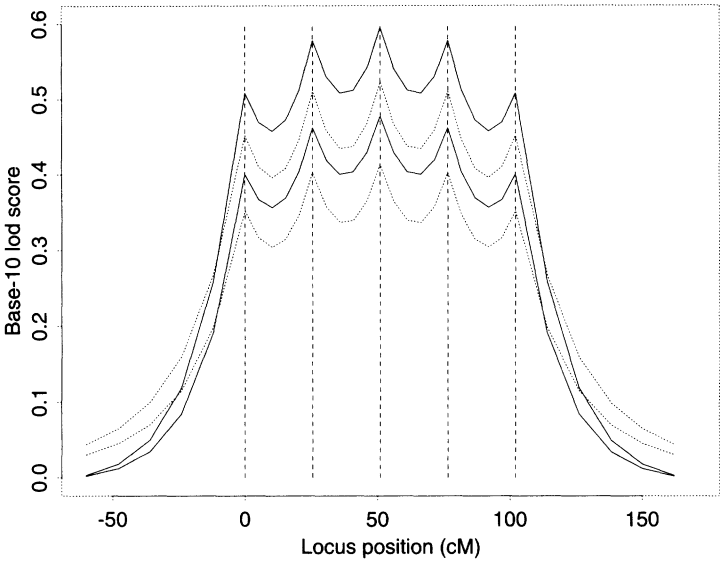

FIGURE 10.5. *Exact base-10 location lod scores computed using GENEHUNTER 2. The solid lines correspond to having marker data on five pedigree members, and the broken lines to having marker data on only the final affected inbred individual. In each pair, the upper curve corresponds to a trait allele frequency* $q = 0.001$, *and the lower to* $q = 0.05$

at this much higher trait allele frequency the probability of two separate origins of the allele in the pedigree is small.

We note that the shape of this location lod score curve is atypical, with maxima at the markers due to the assumption of the same marker data at each locus. This complete concordance of the data, and its consistency with absence of recombination between trait and markers, leads to the symmetry of the curve and to local maxima of the lod score which occur at rather than between the markers. We see that most of the information for linkage is in the data on the final individual; this is the power of homozygosity mapping for a rare recessive trait, as discussed in section 4.6. However, at loose linkage to the marker loci, the marker data on the additional four individuals do impact the lod score curve. Due to the particular marker data assumed, whereby the C_3 allele is known not to be transmitted to the final individual (Figure 10.3), lod scores are sharply decreased outside the map, and in fact are slightly negative at looser linkage.

Attempting estimation of this lod score curve, using the M-sampler as before, gave improvement over the example of section 10.3, although not fully satisfactory results. The expected complete-data base-e log-likelihoods for the case $q = 0.001$ are shown in Figure 10.6, again separated into the penetrance and segregation contributions. As before, the average log-probability of meioses sampled under hypothesized trait location γ_j is much larger under that location than under

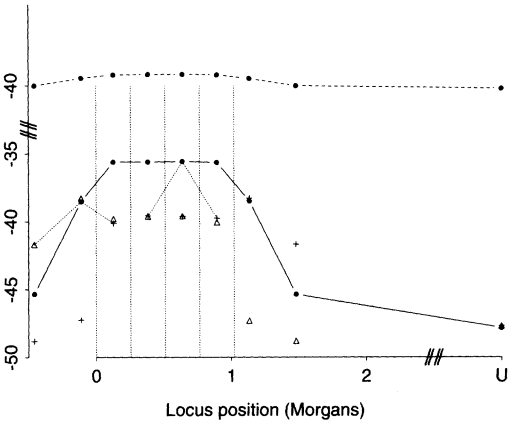

FIGURE 10.6. *Expected complete-data log-likelihoods with the hypothetical data of Figure 10.3 assumed at each of five equally spaced linked marker loci. The notation is as in Figure 10.2*

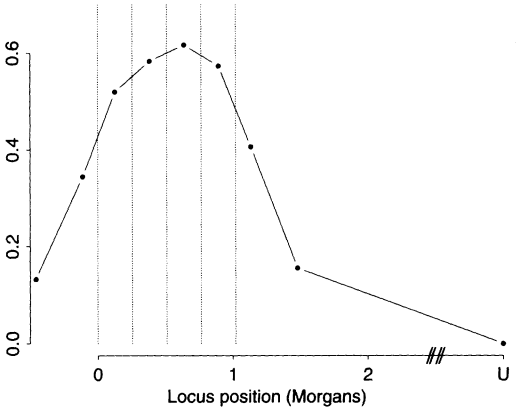

FIGURE 10.7. *Estimated Monte Carlo location base-10 lod score curve for the hypothetical data of Figure 10.3*

locations in the neighboring marker intervals to the left and right. However, now the difference is less than 5 rather than 25. Although e^5 is two orders of magnitude, estimation of the location score curve is now feasible in the sense that the methods converge to provide an estimate.

The method of equation (9.4) of section 9.2 is used, estimating only likelihood ratios at the two points adjacent to the simulation value in each estimating equation. Thus ideally, solution of equation (9.4) should provide an eigenvalue of 2. However, although gene *ibd* probabilities appear to be reliably estimated,

diagnostics suggested the sampler was not mixing well, and different runs gave substantially different lod score estimates. By comparison with exact values (Figure 10.5), the lod score was overestimated. One resulting lod score estimate is shown in Figure 10.7.

In the hope of improving performance, the assumed trait allele frequency was increased to $q = 0.05$. The true lod score is not much affected (Figure 10.5). Unfortunately, neither is the Monte Carlo estimate; curves very similar to that of Figure 10.7 were again obtained. However, the MCMC performance was more robust at the higher trait allele frequency, with much better agreement between runs. For a run giving final estimates indistinguishable from those of Figure 10.7, the relevant eigenvalue of equation (9.4) was 1.94, apparently close to the "perfect" value 2. This indicates indicates good agreement of the ratios provided by simulations at adjacent points.

Despite this apparent success, the absolute values of the log-likelihood differences are still overestimated. As seen in the next section this is primarily due to an insufficient number of simulation points for the MCMC. Additionally, the method of combining the likelihood ratio estimates into an overall lod score appears often to give a positive bias. The Monte Carlo estimator based on equation (9.1), of the ratio of the likelihood at an adjacent point to that at the simulation point, is unbiased. However, the statistical properties of the estimation method based on equation (9.4) are unclear. All that is guaranteed is that the resulting estimator of the lod score is consistent, as the number of realizations at each simulation point becomes infinite. Finally, the value 1.94, although "close" to 2, was less close than with better MCMC samplers sampling at more trait locations. Then, a value in the range 1.98 to 2.02 is typically obtained.

10.6 Better MCMC lod scores

The M-sampler (section 8.4) does not suffer poor mixing due to tightly linked loci, but can mix poorly where there are extended ancestral paths of descent in a pedigree. Conversely, the L-sampler (section 8.3) works well on extended pedigrees, but mixes poorly with multiple linked loci. Combining the two samplers, say in the ratio of 10 M-steps to 1 L-step, can achieve more robust and reliable estimates with higher Monte Carlo precision (Heath and Thompson, 1997). The estimation of conditional *ibd* probabilities of section 10.4 was repeated using the LM-sampler, with an L-sampler proportion of 20%. This means that every step updates either a randomly chosen meiosis (M-step), or a randomly chosen locus (L-step), and each step is independently chosen to be an L-step with probability 0.2. For an equal number of total steps (in this example, 10^7), the MCMC runs took three times as much CPU, but the results of Tables 10.3 and 10.4 were unchanged both for $q = 0.001$ and $q = 0.05$. However, it is likely that the LM-sampler would achieve the same results with a smaller number of total steps.

Using an LM-sampler, and assuming now the higher trait allele frequency of $q = 0.05$, more accurate Monte Carlo lod score estimates are obtained for the example of section 10.5. Shown in Figure 10.8 as a solid line is the exact base-10

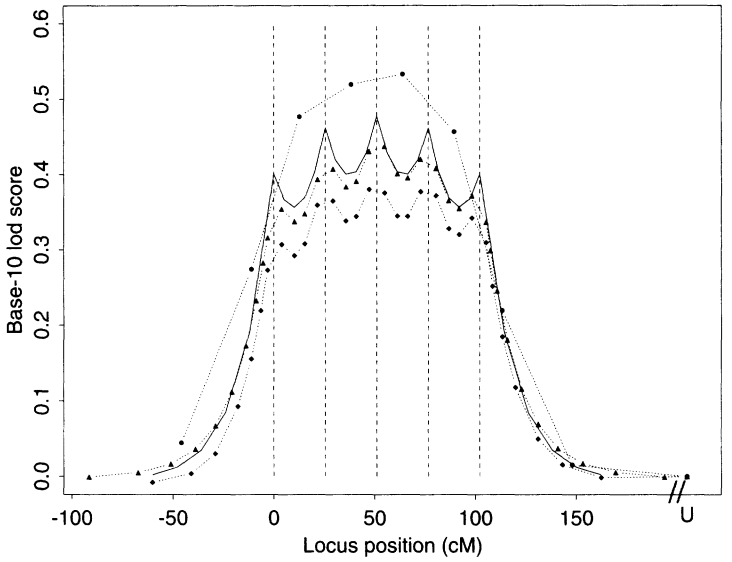

FIGURE 10.8. *Base-10 location score curves for the example of section 10.5 re-estimated, shown also with the exact value*

L-sampler probability	0.0	0.2	0.2	0.2	0.2
MCMC sample points					
unlinked	1	1	1	1	1
each end	2	2	3	7	10
each interval	1	1	3	4	4
Total	9	9	19	31	37
eigenvalue of (9.4)	1.942	1.979	1.993	1.999	2.002
MCMC realizations/point	10^6	10^6	10^6	10^6	10^6
CPU time (secs)	5,237	29,466	60,507	96,153	116,443
Shown in Figure	10.7	10.8	—	10.8 (*)	10.8

TABLE 10.5. *Summary of LM-sampler runs on the example of section 10.5. The penultimate run, designated (*), is the run also used for the results of Figures 10.9 and 10.10. The first column shows the M-sampler run discussed in section 10.5. The runs were done on a DEC alpha workstation 400-233, with 192 MB memory*

lod score computed using GENEHUNTER 2 (Kruglyak et al., 1996; Kruglyak and Lander, 1998). We note again that this lod score is atypical with local maxima at every marker, due to the assumption of the same marker data at each locus and its consistency with absence of recombination between trait and markers (Figure 10.5). This makes this lod score curve a challenge for Monte Carlo estimation, even though this pedigree is small. Also shown are three Monte Carlo estimates of the lod score,

with the MCMC done using the LM-sampler. As in previous sections, likelihood ratios were estimated only relative to adjacent trait locations, and the lod score estimation method of equation (9.4) was used to combine these into a single set of lod scores. Only lod scores at the simulation points are estimated. There is no attempt to interpolate between these points, which are connected by broken lines in Figure 10.8 for clarity only.

For clarity and easier comparison, we show here only three curves, each done with an L-sampler proportion of 20%. The MCMC is performed with the trait locus in each of the positions indicated, starting with the trait locus unlinked. When the hypothesized trait locus location is changed, the first step is to update the trait locus meiosis indicators. The initial set-up is done using the L-sampler set-up for unlinked loci (Heath, 1997). On a large pedigree, with extensive marker data, some burn-in for the linked marker loci should therefore be included, but this was ignored in this example. The marker loci were not tightly linked (see Figure 10.4).

The run characteristics and results are summarized in Table 10.5. The first column shows the M-sampler run of section 10.5 for comparison, but this curve is not shown in the figure. As can be seen from a comparison of Figure 10.7 and Figure 10.8, the results are similar when the same simulation points are chosen. This wide point spacing, with only a single point in each marker interval, leads to an overestimate of the lod score. With the LM-sampler (second column of Table 10.5), the upward bias is less, and the eigenvalue of the estimating equation increases from 1.942 to 1.979—closer to the idealized value of 2. Of greater relevance may be that the run takes almost 6 times as much CPU. On the positive side, the LM-sampler gives more consistent results. In fact, both runs were the first run at these computational settings. However, there was greater variability among runs using the M-sampler alone. With 10^6 MCMC steps at each simulation position for the trait locus, results using the LM-sampler were almost identical in repeat runs.

The three curves shown in Figure 10.8 correspond to the second and to the last two columns of Table 10.5. Other runs, including some not listed in this table gave comparable results. Using the sample LM-sampler settings, but increasing the number of points for MCMC and likelihood-ratio estimation (Table 10.5), we obtain much better lod score estimates (Figure 10.8). With more points for estimation and evaluation, the bias in the estimated lod score is reduced or even eliminated. The eigenvalue of the estimating equation becomes increasingly close to the ideal value of 2.000. All curves with several points within the marker intervals managed to mimic the atypical dips of the true curve. More difficulty was encountered in getting the precise level of the curve, relative to the null hypothesis that the trait locus is unlinked. Even with seven linked evaluation and simulation points at each end of the map (Table 10.5), there are still adjacent simulation points at which the likelihood ratio is too large to be well estimated. The final run, with 10 evaluation points at each end of the map did well, even mimicking the true very slightly negative lod scores at each end of the map when the trait locus is close to unlinked. However, even here, there is a slight asymmetry and downward bias as the trait locus crosses the first marker. All the runs show this asymmetry when the trait locus is moved from across the map from left to right, and it is reversed when the direction is reversed. Possibly, more burn-in as the trait locus gets close to the

marker loci would resolve this.

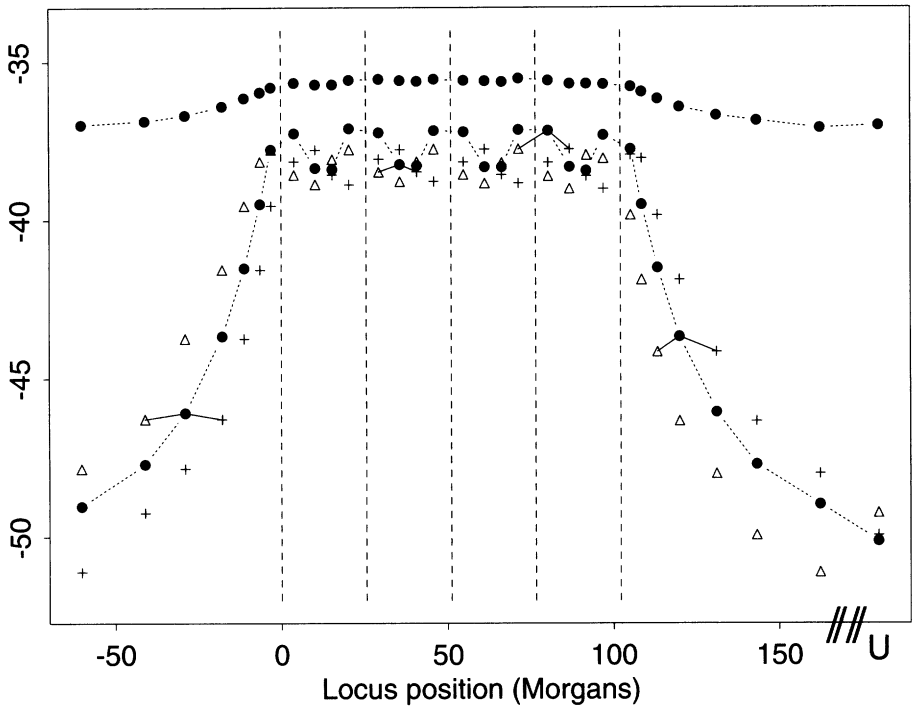

FIGURE 10.9. *Expected complete-data log-likelihoods for the example of section 10.5, shown for the penultimate run of Table 10.5. The notation is as in Figure 10.2. As in that figure, the contribution from penetrance terms is shown separately from that for segregation terms*

For a given L-sampler proportion, the CPU time is almost directly proportional to the number of simulation points, or more generally to the total number of MCMC steps. An L-step appears to take about 20 times as long as an M-step; of course, this ratio is highly data-set and pedigree dependent. For comparison purposes, all runs were done on a 1995 DEC alpha workstation 400-233, upgraded to have 192 MB memory. This machine is about four times slower than newer single-processor DEC alpha workstations with 1GB memory. In addition to computing the likelihood ratios, the program also produced the expected complete-data log-likelihoods (section 9.6) and the conditional probabilities of recombination in all intervals, in both male and female meioses. The added computational cost of producing these useful diagnostics is slight.

The expected complete-data log-likelihoods are shown in Figure 10.9, for the penultimate run shown in Table 10.5. The notation is the same as in Figure 10.2. The contribution from the penetrance terms $\Pr(\mathbf{Y}|\mathbf{S})$ is the upper curve, while

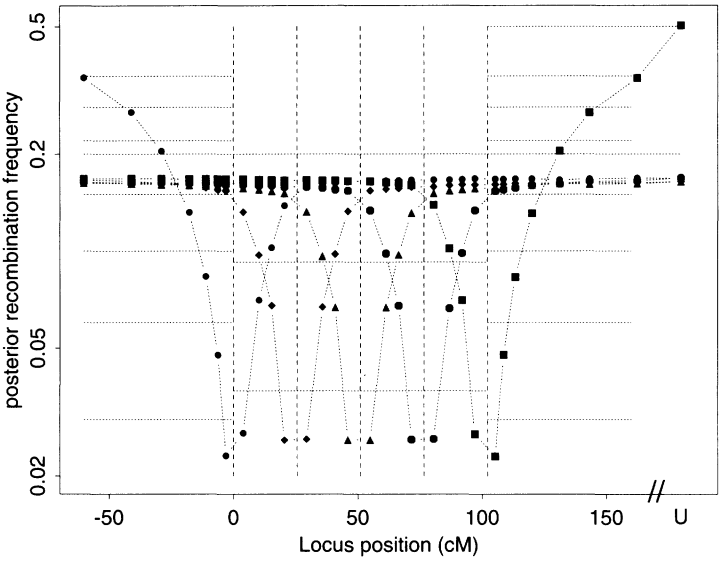

FIGURE 10.10. *Estimated conditional probabilities of recombination in the five map intervals for the example of section 10.5, shown for the penultimate run of Table 10.5. For details, see text*

the lower curve gives the expected value of $P_\gamma(\mathbf{S})$. Each point is plotted at the coordinate corresponding to the trait location γ for which the probability is evaluated. For the penetrance curve, and the main segregation-probability curve (indicated by $\bullet$), the simulation point and evaluation point are the same. A $\triangle$ indicates an evaluation point to the left of the simulation point and a $+$ indicates an evaluation point to the right. As in Figures 10.2 and 10.6, a few corresponding $(\triangle - \bullet - +)$ triplets are connected by lines. By comparison with the Figure 10.6, we see that differences are now small between evaluations at adjacent locations of the log-probabilities of realizations at a given point: the $(\triangle - \bullet - +)$ triplets. As expected, the log-probabilities are highest where the simulation point is also the evaluation point. However, for some evaluation points outside the marker map, we see that the probability is up to seven times (e^2) larger for realizations at an adjacent point than at the point itself—the vertical $(\triangle - \bullet - +)$ differences in the figure. Ideally, for accurate estimation of Monte Carlo lod score curves, both sets of log-probability differences should be small. The results suggested that more simulation points outside the marker map are needed, as also suggested by a comparison of the estimated and exact lod score curves of Figure 10.8. This led to subsequent production of the final run shown in the figure, and as the last column of Table 10.5.

Figure 10.10 shows the conditional probabilities of recombination, given the marker and trait data, for each trait location, in each of the five intervals of the marker and trait locus map. For consistency, these are shown for the same penultimate run of Table 10.5. Each symbol represents a different interval; the

interval containing the trait locus changes as the trait locus moves across the marker map. For greater clarity the frequencies are shown on a log scale. The program estimates frequencies for male and female meioses separately, but these have been combined in the current figure. Even where, as here, the prior recombination frequencies are the same in male and female meioses, the frequencies conditional on data are not. The conditional probabilities depend on the specific marker data and the gender of individuals in whose meioses recombinations are imputed. Also shown in the figure, by broken horizontal lines, are the prior recombination frequencies between markers (20%), at trait locations outside the map, and for two of the four locations for the trait locus within each marker interval. Except for an unlinked trait locus, or very loose linkage, the concordant data at all the markers and at the trait locus depresses the conditional probabilities of recombination below their prior expectation. Even with these fully concordant data, however, the conditional probabilities are not small: each is about 85% to 90% of the prior value.

Chapter 11

Other Monte Carlo Likelihoods in Genetics

11.1 Improving pedigree samplers

The ways in which MCMC samplers can be extended, combined, and improved, are almost limitless. One method has been discussed in section 10.6. Where the pedigree is not too complex, so that the L-sampler is feasible (and practical), combining the L-sampler and M-sampler on extended pedigrees can achieve more robust and reliable results with higher Monte Carlo precision (Heath and Thompson, 1997). The M-sampler (section 8.4) does not suffer poor mixing due to tightly linked loci, but can mix poorly where there are extended ancestral paths of descent in a pedigree. Additionally, the M-sampler may not be irreducible. Since the L-sampler is irreducible (section 8.3), combination of the L-sampler and M-sampler can ensure irreducibility, as well as improve mixing. Whereas the L-sampler is often the more computationally intensive, and seems to take longer to achieve stable probability estimates, the M-sampler may simply fail to sample the part of the space containing the majority of the probability mass (Table 11.1). The examples of section 10.6 all combined L and M steps with the same probability (20%) that any given step is an L-step. Obviously, there is scope for other patterns of systematic or random resampling.

There are ways to improve the meiosis sampling itself. Updating all indicators at a meiosis jointly shows much improved performance over single-site updating (Thompson and Heath, 1999). Moreover, updating by meiosis avoids problems of poor mixing due to tight linkage. However, clearly there would be greater improvement if the vectors $S_{i,\bullet}$ for several meioses i were to be updated jointly. Likewise the L-sampler can be improved. For very tightly linked loci, single-locus updates are ineffective. However, where feasible, the L-sampler might update jointly $S_{\bullet,j}$ for several loci j. For the L-sampler, on a complex pedigree, usually no more than two or three loci can be updated jointly.

One case where updating several meioses jointly is effective and easily done is

		Update by locus	
		singly	jointly
Update by meiosis	singly	single-site: Update $S_{i,j}$. Performance poor	L-sampler: Update $S_{\bullet,j}$. Performance poor for tight linkage
	jointly	M-sampler: Update $S_{i,\bullet}$. Performance poor for extended pedigrees	LM-sampler. Improved mixing and more robust estimation

TABLE 11.1. *Single-site and joint updating schemes on a pedigree*

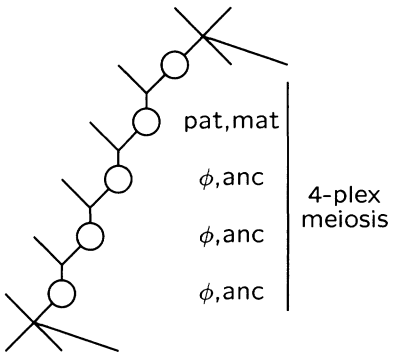

FIGURE 11.1. *A multiplex meiosis consisting of an ancestral chain of four meioses. These meioses may be jointly updated. For additional details, see text*

where there is a succession of several ancestral meioses over several generations with no phenotypic data, in each case there being one founder parent with a single offspring in the pedigree (Figure 11.1). A number of such chains may be seen in the pedigree of Figure 10.1. Recall (section 8.4) that, for such a founder parent, the meiosis to the single offspring is not scored. The relevant gene in each offspring (in this example, the paternal gene) is, in effect, a founder gene. We refer to the chain of meioses from the pedigree (non-founder) parents as a *multiplex* meiosis. For the first meiosis of the chain, we score, as usual, whether the offspring receives the parent's maternal or paternal gene. For subsequent meioses, the state is characterized by whether, at a given locus, the transmitted gene is the parent's gene from a peripheral (ϕ) or pedigree (anc) parent. The state of the multiplex meiosis is characterized by the number of meiosis indicators in the chain that currently point to a gene from one of the peripheral founders (0,1,2,3 in Figure 11.1), and the state (0=maternal, 1=paternal) of the first meiosis. With this specification, the transition probabilities remain first-order Markov along the chromosome. A pre-processing of the pedigree,

assigning each multiplex meiosis its appropriate Markov transition probabilities, can greatly improve efficiency of the MCMC. There are fewer (multiplex) meioses to resample: in the example we have replaced four meioses with a total of $2^4 = 16$ states by a single multiplex meiosis with $2 \times 4 = 8$ states. For example, the state $(1, 2)$ would denote that the first individual of the chain receives her mother's paternal gene, and that for 2 of the subsequent meioses the offspring receive their mother's founder gene. (In Figure 11.1, the founder parents are male, and the pedigree parents are female.) Although this single factor of two in the number of states is not large, repeated over a large pedigree this can lead to a significant reduction. More important than the number of states is the mixing of the MCMC. Even when transition probabilities for descent down the chain are small, with joint updating alternative descent paths for an allele are more readily sampled. The ability to change the descent path down the whole chain in a single MCMC step greatly improves mixing.

The joint updating of meioses may be carried further. The Lander-Green algorithm for exact computation can be readily performed on up to 15 meioses. While one might not want to incorporate such an intensive computation into an MCMC, computation is quite feasible for, say, a subset of $m^* = 10$ of the total set of m meioses in the pedigree. The procedure is exactly as in equations (8.8) and (8.9), except now that, instead of the two values of s, $Q_t(\mathbf{s})$ must be evaluated and normalized for each of the 2^{m^*} vectors of the indicators for these m^* meioses, say 1024 values to be stored for each locus along the chromosome. Additionally the penetrance probabilities $P(Y^{(j)} \mid S_{\bullet,j})$ would be needed for each of the 1024 values. The extent to which improved mixing compensates for the increased computation remains to be investigated, but there is no doubt that joint updating will help in some cases. When the L-sampler in infeasible due to extreme complexity of the pedigree, joint updating of several meioses could ensure irreducibility of the meiosis sampler. However, this is an area where many open questions remain. In particular, on an extended pedigree, appropriate choice of the meioses to be updated jointly is far from obvious.

11.2 Interference by Metropolis-Hastings

In the absence of interference, but where different meioses exhibit different recombination probabilities, the procedure of resampling a whole meiosis jointly over loci (section 8.4) is more convenient than other forms of MCMC. Sex-specific maps can be routinely incorporated, provided they are known, and no assumptions regarding the relationship between male and female recombination frequencies are necessary. Each meiosis is resampled, and the relevant computations made, under the map appropriate to that meiosis. For multiplex meioses (section 11.1), which may contain individuals of different sexes, male and female meioses must be accounted separately, and the transition probabilities must be pre-computed, but there is no intrinsic computational difficulty.

Genetic interference in meiosis (Chapter 5) is a more complex issue, since it destroys the first-order Markov conditional-independence structure of the meiosis

indicators along a chromosome. The assumption of first-order dependence in the $S_{\bullet,j}$ is crucial to the computations of sections 6.1 and 7.1 and to the M-sampler as developed in section 8.4. There is no general computational algorithm for exact computation of multilocus linkage likelihoods under interference, although Weeks et al. (1993) and Lin and Speed (1996) have shown how to incorporate interference in some cases. However, as for any multilocus problem, exact likelihood computation on an extended or complex pedigree remains computationally infeasible. In fact, for exact computations under interference, the numbers of markers, and/or pedigree structures, are severely limited, and computation is cumbersome. The erstwhile practice of transforming recombination frequencies between markers using a genetic map function, and then performing a no-interference computation becomes increasingly futile as maps become denser, and marker data on observed individuals more complete.

Although genetic interference is very seldom incorporated into linkage computations, it exists in human meiosis (Broman and Weber, 2000). Failure to incorporate it can reduce the power to detect linkage (Goldstein et al., 1995). In an analysis of data at multiple tightly linked markers from actual meioses, Thompson and Meagher (1998) have shown that interference can have a significant impact on patterns of joint segregation of genes at distances of 20cM to 30cM. Using our whole-meiosis M-sampler (section 8.4), since all the meiosis indicators for all the linked loci in an entire meiosis are resampled jointly, incorporation of an interference model is feasible.

In the M-sampler, given marker data $\mathbf{Y}$ at loci $j = 1, \ldots, L$, meiosis indicators at meiosis i, $S_{i,\bullet} = (S_{i,1}, \ldots, S_{i,L})$ are realized from

$$P(S_{i,\bullet} \mid S_{k,\bullet}, k \neq i, \mathbf{Y}) \quad \propto \quad P(\mathbf{Y} \mid \mathbf{S})\, P^{(H)}(\mathbf{S}) \tag{11.1}$$

where $\mathbf{S}$ is the total set of meiosis indicators for all loci at all meioses of the pedigree, and the super-script (H) denotes the Haldane (no-interference) model. We now continue to use equation (11.1) as our proposal distribution, and add a Metropolis acceptance step (Metropolis et al., 1953), to provide the correct conditional distribution of $S_{i,\bullet}$ under interference (denoted $P^{(I)}(\cdot)$). The required Hastings-ratio $h(\mathbf{S}^\dagger; \mathbf{S})$ (equation 8.2) for current $\mathbf{S}$ and proposed $\mathbf{S}^\dagger$ is

$$\begin{aligned}
h(\mathbf{S}^\dagger; \mathbf{S}) &= \frac{P^{(I)}(\mathbf{S}^\dagger, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y})} \frac{P^{(H)}(S_{i,\bullet} \mid S_{k,\bullet}, k \neq i, \mathbf{Y})}{P^{(H)}(S^\dagger_{i,\bullet} \mid S_{k,\bullet}, k \neq i, \mathbf{Y})} \\
&= \frac{P^{(I)}(\mathbf{S}^\dagger, \mathbf{Y}) P^{(H)}(\mathbf{S}, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y}) P^{(H)}(\mathbf{S}^\dagger, \mathbf{Y})} \\
&= \frac{P(\mathbf{Y}|\mathbf{S}^\dagger) P^{(I)}(\mathbf{S}^\dagger) P(\mathbf{Y}|\mathbf{S}) P^{(H)}(\mathbf{S})}{P(\mathbf{Y}|\mathbf{S}) P^{(I)}(\mathbf{S}) P(\mathbf{Y}|\mathbf{S}^\dagger) P^{(H)}(\mathbf{S}^\dagger)} \\
&= \prod_{k=1}^{m} \frac{P^{(I)}(S^\dagger_{k,\bullet})}{P^{(I)}(S_{k,\bullet})} \frac{P^{(H)}(S_{k,\bullet})}{P^{(H)}(S^\dagger_{k,\bullet})} \\
&= \frac{P^{(I)}(S^\dagger_{i,\bullet})}{P^{(I)}(S_{i,\bullet})} \frac{P^{(H)}(S_{i,\bullet})}{P^{(H)}(S^\dagger_{i,\bullet})}.
\end{aligned}$$

recombination patterns	prob under model I $d = 25.54$cM $\rho = 0.2554$	prob under model II $d = 0.2$cM $\rho = 0.2$	prob under model 0 $d = 25.54$cM $\rho = 0.2$	prob ratio (I)	prob ratio (II)
r r r r	0.0	0.0	0.0016	0.0	0.0
r r r n, n r r r	0.0	0.0	0.0064	0.0	0.0
r r n r, r n r r	0.0027	0.0	0.0064	0.422	0.0
r r n n, n n r r	0.0027	0.0	0.0256	0.106	0.0
r n r n, n r n r	0.1196	0.05	0.0256	4.672	1.953
r n n r	0.0054	0.05	0.0256	0.212	1.953
n r r n	0.0054	0.0	0.0256	0.212	0.0
n n n r, r n n n	0.1223	0.1	0.1024	1.194	0.977
n n r n, n r n n	0.125	0.15	0.1024	1.221	1.465
n n n n	0.2446	0.35	0.4096	0.597	0.854

TABLE 11.2. *Probabilities of recombination (r) and non-recombination (n) in four equal marker intervals, under interference models I and II and under the Haldane model of no interference (model 0)*

The acceptance probability is then $\alpha = \min(1, h(\mathbf{S}^\dagger; \mathbf{S})$. This considerable reduction in the expression for $h(\mathbf{S}^\dagger; \mathbf{S})$, and consequent ease of computation of the acceptance probability relies on three facts:
(1) the probability of data $\mathbf{Y}$ given meiosis pattern $\mathbf{S}$ or $\mathbf{S}^\dagger$ does not depend on the interference process (I) or (H), giving rise to $\mathbf{S}$,
(2) the independence of meiosis patterns $S_{k,\bullet}$ at different meioses k (when not conditioned on data $\mathbf{Y}$), and
(3) $S^\dagger_{k,\bullet} = S_{k,\bullet}$ for $k \neq i$.

As an example, consider again the standard test pedigree (Figure 1.1): this example was also given in Thompson (2000*a*). As in section 10.4, consider five equispaced marker loci, 25.54cM apart (recombination frequency 20% under the Haldane no-interference model). We consider the case of extreme position interference, but no chromatid interference, in which chiasmata on the underlying tetrad are equispaced at 50cM spacing. Then using the notation of section 5.3 for the indicator vectors $\mathbf{C}$ of presence (1) or absence (0) of chiasmata in the four intervals there are only 5 possible values: $\mathbf{C}$ = (1,0,1,1), (1,1,0,1), (0,1,1,0), (1,0,1,0) or (0,1,0,1). Under a model which places the first marker uniformly in an interval between two chiasmata, these five possible chiasmata indicator vectors have probabilities 0.0216, 0.0216, 0.0216, 0.4676 and 0.4676 respectively. Using equation (5.2), these translate to the probabilities of patterns of recombination (r) or non-recombination (n) given in under model (I) in Table 11.2. In this table, pairs of vectors having the same probability under any model are listed together. For example, for equispaced markers, patterns *rrrn* and *nrrr* have the same probability, by symmetry. The tabulated probability refers to the probability of each of the two patterns. We see there are substantial differences in the probabilities under this interference model (I) and under no interference (model 0; Haldane). However,

the ratios are not so extreme as to make the MCMC ineffective. All probabilities are strictly positive under the proposal (Haldane) distribution, and non-zero ratios differ by a factor of at most 22 (0.212 to 4.672).

It is not clear that the correct assessment of interference effects should be through imposing equal genetic distance; that is, total expected numbers of crossovers. If instead we constrain the recombination frequency between adjacent markers to be 20%, the distance under our model of complete position interference is 20cM. Again there are five possible indicator vectors $\mathbf{C}$ of chiasmata presence/absence, but this time these are $(0,0,1,0)$, $(0,1,0,0)$, $(0,1,0,1)$, $(1,0,0,1)$, and $(1,0,1,0)$, each having probability 0.2. Since chiasmata have an exact 50cM spacing and the marker intervals are 20cM, it is not longer possible for there to be chiasmata in two adjacent intervals. Again assuming the first marker is randomly and uniformly placed relative to the chiasmata, and using equation (5.2), the corresponding probabilities of patterns of recombination/non-recombination are as given for model (II) in Table 11.2. Again the ratios, for model (II) relative to the proposal (Haldane) model are not extreme; this time the non-zero ratios range only from 0.854 t0 1.953. Although both models (I) and (II) have some recombination vector events of probability 0, this does not lead to invalid estimates. If proposed, these vectors will not be accepted. The total probability under the Haldane model of recombination vectors that cannot be accepted under the interference models is not large (0.014 under model I, 0.104 under model II).

Gene *ibd* pattern	single-locus prior	single-locus conditional	no interference marker M5	no interference marker M3
All 4 genes *ibd* genes	29	133	127	180
3 of 4 genes *ibd*	156	286	356	381
2 pairs of *ibd* genes	84	154	118	130
2 of 4 genes *ibd*	484	354	303	251
all 4 non-*ibd*	247	73	96	58
mean log-probability $\log P_\theta(\mathbf{S})$)			-44.69	
mean log-probability $\log P_\theta(\mathbf{Y}\|\mathbf{S})$			-33.23	
MCMC steps (accepted %)			10^7(100%)	

TABLE 11.3. *Gene ibd probabilities* (×1000) *for single loci, and under no interference (Haldane model)*

The marker data at each locus assumed are as in sections 10.4 and 10.6 for the five individuals of the pedigree with marker phenotypes observed (Figure 10.3). The allele frequencies are again assumed to be 0.2, 0.2, 0.4, and 0.2 for the four alleles at each locus. Again, sampling latent meiosis indicators $\mathbf{S}$ conditional on the marker data, we score gene *ibd* probabilities among the four potentially distinct C alleles. In Table 11.3 are shown the gene *ibd* probabilities for single loci, and for linked markers under the Haldane model of no interference for the central marker M3 and an end marker M5. These are the same values seen for marker loci in

Gene *ibd* pattern	Model I marker M5	Model I marker M3	Model II marker M5	Model II marker M3
All 4 genes *ibd*	74	104	101	152
3 of 4 genes *ibd*	305	339	334	370
2 pairs of *ibd* genes	94	114	106	126
2 of 4 genes *ibd*	349	321	327	276
all 4 non-*ibd*	178	122	132	76
mean log-probability $\log_e P_\theta(\mathbf{S})$	-50.74		-45.23	
mean log-probability $\log_e P_\theta(\mathbf{Y}\|\mathbf{S})$	-34.12		-33.58	
MCMC steps (accepted %)	10^7(68.6%)		10^7(80.4%)	

TABLE 11.4. *Gene ibd probabilities* ($\times$1000) *under the recombination pattern probabilities given for interference models (I) and (II) in Table 11.2. Each run consisted of 10,000,000 whole-meiosis Gibbs/Metropolis updates, and took about 1 hour CPU on a DEC Alpha 400-233 work-station with 256MB memory*

Table 10.3 in the case of a trait locus unlinked to the markers: they are shown again here for easier reference in the context of interference effects. The *prior* is the probability given by the pedigree alone, without marker data. The *conditional* is the probability when the marker phenotypes are assumed for a single locus. The table shows that the data increase probability of gene *ibd*—not surprisingly since the four genes scored are of the same allelic type. Having data at five linked markers reinforces the inference of gene *ibd*, particularly for the marker $M3$ in the center of the map. Note that the marker spacing is 25.54cM, so that the five loci extend over 1 Morgan. In every meiosis of the pedigree there is a probability 0.5904 of at least one recombination among these five markers. Even so, the concordant data at these linked markers reinforces probabilities of gene *ibd*.

The results of 10^7 MCMC meiosis resamples are given in Table 11.4. We see a substantial effect of interference on the conditional probabilities of gene *ibd*. In particular, probabilities that all four C alleles are *ibd* are reduced, and that all are distinct are increased. The percentage of MCMC proposals accepted and the expected base-e complete-data log-likelihoods both provide an indication of the effect of interference. In comparison to the non-interference case, matching recombination frequencies (model (II)) provides closer results than does matching genetic distances (model (I)).

The interference models considered in this section are extreme, assuming complete position interference, although no chromatid interference. Other less extreme examples still show substantial impact on genome sharing among relatives at distances of 20 to 30cM. For example, Browning and Thompson (1999) considered the aunt-niece-sibs example of section 4.5, using a chi-square model with parameter $m = 2$ for the interference process (example (4) of section 5.7). Although the impact of interference on genetic inferences remains a little studied area, the results here suggest that further study is warranted. Although interference will have little impact on mapping Mendelian traits when markers are highly informative,

it will affect the resolution of genes contributing to quantitative traits or to disease liability. Its impact will also be greater in using tightly linked but less informative markers, such as Single Nucleotide Polymorphisms (SNPs): see section 1.1. For such markers, haplotypes cannot be readily inferred, even with data on pedigrees, and interference will affect the imputation probabilities for such haplotypes.

11.3 Inference of typing or pedigree error

Throughout this monograph, we have focused on the case where the pedigree relationship among individuals is known, and often where the marker map and other parameters of the model for the marker data are assumed known without error. We have also not explicitly considered the possibility of errors in marker genotypes. However, as noted in section 1.4, the probability of data under a known genetic model is a likelihood for the pedigree relationships among the individuals. Also, on an assumed pedigree it is clearly possible to address other aspects of the model for the data, including possible typing errors. In analyses of real data, errors, uncertainty, or heterogeneity in the marker model often arise and may have an impact on inference. Traditionally, the approach has been to correct for errors in advance of other analyses, usually on a marker-by-marker basis. This can be unsatisfactory (Broman, 1999), and with greater automation of marker genotyping it becomes important to have methods of analyzing multilocus marker data, and allowing within the analysis for possible errors in typing or specification of individual relationships.

For inference of possible data error, the general method is simply one of generalizing the model for the relationship between underlying latent variables $S_{\bullet,j}$ or genotypes $G_{\bullet,j}$ at locus j, and the observable data $Y_{\bullet,j}$ on the individuals. The likelihood is most easily considered as in equation (3.9) or (6.1):

$$\begin{aligned} \Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{S},\mathbf{Y}) = \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \left(\Pr(S_{\bullet,1}) \prod_{j=2}^{L} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \prod_{j=1}^{L} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right). \end{aligned} \tag{11.2}$$

The dependence structure (Figure 6.1), and hence general Baum-algorithm computational approach (section 6.1) remain unchanged. The generalization is only in $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ for each locus j. It may be easier to consider likelihood computation with an additional layer of latent variable—the true genotypes determined by the underlying pattern of gene *ibd* (Kumm et al., 1999):

$$\Pr(Y_{\bullet,j} \mid S_{\bullet,j}) = \sum_{G_{\bullet,j}} \Pr(Y_{\bullet,j} \mid G_{\bullet,j}) \Pr(G_{\bullet,j} \mid S_{\bullet,j}). \tag{11.3}$$

Assuming typing errors are individual-specific

$$\Pr(Y_{\bullet,j} \mid G_{\bullet,j}) = \prod_{i} \Pr(Y_{i,j} \mid G_{i,j})$$

a product over individuals i. Nonetheless, computation of (11.3) and hence the likelihood (11.2) can become computationally intensive for general error models and more than a very few individuals. In principle, the likelihood (11.2) can be used to estimate the form and the parameters of the error model. More practically, the reverse Baum algorithm (section 7.1) can be used to to determine the loci at which there is a high probability of error given all the observed data: that is, $\Pr(G_{i,j} \neq Y_{i,j} \mid \mathbf{Y})$ is large.

Under a given penetrance model, the likelihood of alternative relationships can be compared. Boehnke and Cox (1997) used the Baum algorithm to compute likelihoods for alternative sib and half-sib relationships from multilocus marker data. Browning (1999) extended this to a variety of extended-family relationships, up to second cousins. On larger pedigrees, in principle at least, MCMC may be used to obtain a Monte Carlo likelihood ratio. Since the likelihood is given by equation (11.2), we have the likelihood ratio equation (9.1) which, in the present context becomes

$$\frac{P_1(\mathbf{Y})}{P_2(\mathbf{Y})} = \mathrm{E}_1\left(\frac{P_2(\mathbf{Y},\mathbf{S})}{P_1(\mathbf{Y},\mathbf{S})} \;\middle|\; \mathbf{Y}\right)$$

where the subscripts on probabilities and expectations designate two alternative relationship hypotheses. Any of the MCMC samplers of earlier sections can be used to sample from

$$P_1(\mathbf{S} \mid \mathbf{Y}) = \frac{P_1(\mathbf{Y},\mathbf{S})}{P_1(\mathbf{Y})} \propto P_1(\mathbf{Y} \mid \mathbf{S})P_1(\mathbf{S}).$$

Care is needed in implementing these likelihood ratio estimators, since different relationships may imply a different number of relevant meioses. Unlike in the comparison of different genetic models, the penetrance probabilities $P_1(\mathbf{Y} \mid \mathbf{S})$ may depend on the hypothesized relationship. Nonetheless, we must consider MCMC sampling of $\mathbf{S}$ not of *ibd* patterns $J(\mathbf{S})$, although the latter are more readily compared for alternative relationships. In the assumed absence of interference, the segregation process $\mathbf{S}$ is Markov along the chromosome, but the agglomerated process $J(\mathbf{S})$ is not (section 4.8).

In any give meiosis, there are relatively few changes in $S_{i,j}$ as j changes. As the number of linked marker loci becomes very large and they are thus tightly linked, it becomes inefficient to use the complete set of components of $\mathbf{S}$ as the latent variables, and also difficult to get effective samplers on this space. Instead, one may consider a set of latent processes $S_i(z)$ where z is the position on the chromosome measured in terms of genetic distance. This framework was first developed by Donnelly (1983), and used by Bickeboller and Thompson (1996*a*; 1996*b*) to study the descent of genome in small pedigrees. Browning (1998) used the same underlying model to develop importance-sampling methods of estimating Monte Carlo likelihoods for alternative pedigree relationships. Browning (1999) extended the approach to the development of Monte Carlo likelihood methods to distinguish between alternative models of meiosis and genetic interference, including the models discussed in sections 5.6 and 5.7.

11.4 Other Monte-Carlo procedures for linkage analysis

Another broad area of linkage analysis not addressed in this monograph is the mapping of loci contributing to quantitative traits, or quantitative trait loci (QTL). For linkage designs in experimental organisms there are well developed methods for detecting, mapping, and resolving the QTL contributing to increasingly complex traits (Knott and Haley, 1992; Zeng, 1994; Long et al., 1995). Increasingly, on larger or more complex problems MCMC is used (Hoeschele, 1994; Sorensen et al., 1995; Satagopan et al., 1996). Heath (1997) developed methods of segregation and linkage analysis on extended pedigrees, for models involving multiple QTL contributing additively to a complex quantitative trait.

There are two main differences between MCMC methods for QTL analysis and the methods developed in this monograph. First, a Bayesian approach (section 2.4) is normally taken. For complex models, with many nuisance parameters, a likelihood approach has limitations. The traditional likelihood approach has been to maximize over these parameters, obtaining a profile likelihood for the parameters of interest. However, a Bayesian approach which integrates or samples (in the case of Monte Carlo) over nuisance parameters may provide a better reflection of the true information regarding parameters of interest. Using MCMC, samples are realized from the posterior probability distributions of parameters. A disadvantage of a Bayesian approach is that there is no exact computational approach against which MCMC results can be compared. As seen in section 10.6, even our best MCMC samplers need tuning to produce accurate likelihood estimates. For a Bayesian posterior probability distribution for parameters of a complex model, there is no way to assess the accuracy of Monte Carlo results. There is also no standard interpretation of findings. Whereas there may not be unanimity regarding the exact meaning of a base-10 lod score of 3.5, say, there is no collective experience at all regarding, say, a finding of 97% probability that at least two QTL contribute to a trait.

A second major difference between the methods of this monograph and MCMC methods for QTL analysis also relates to the model complexity, but to its effect on the MCMC methods used. For a model such as that of Heath (1997) in which the number of QTL contributing to a trait can vary, the dimension of the model is not fixed. In sampling over the parameters of a varying number of QTL, the number of parameters sampled changes. Thus methods of reversible-jump MCMC (Green, 1995) must be used, to sample between models of varying dimension.

11.5 Monte-Carlo likelihoods in population genetics

One of the earliest uses of MCMC in genetic analysis was not on pedigrees, but on the evolutionary history of populations and species. Since data are normally observed in the present, forwards simulation of the evolutionary process is of limited

use in developing Monte Carlo inference procedures. Just as on a pedigree, effective realizations must be conditioned on the data. Kingman (1982) developed the theory of the coalescent, which allows for study of the ancestry of a current sample from a population. Kuhner et al. (1995) developed Monte Carlo likelihood methods for estimating evolutionary parameters, based on MCMC resampling of coalescent ancestries of the current population sample. Griffiths and Tavaré (1994*b*; 1994*a*) also developed a Monte Carlo likelihood approach to similar problems. Their approach uses importance sampling (section 7.3) rather than MCMC, and they realize successive events in the ancestry of a current sample. Stephens and Donnelly (2000) have given a recent synthesis, discussion, and extension of these two approaches.

More recently, Monte Carlo likelihood approaches have been used in a wide variety of population-genetic areas. One of these is the development of Monte Carlo likelihood methods for fine-scale mapping. Due to the limited number of meioses, the resolution of loci from pedigree data is no finer than about 1 cM (Boehnke, 1994). As described briefly in section 4.6, allelic associations resulting from slow decay of initial linkage disequilibrium between a new mutation and a tightly linked marker locus can provide evidence for linkage and for precise localization of a disease locus. This has been a recent focus of several successful mappings of loci with rare recessive disease alleles (Cox et al., 1989; Hästbacka et al., 1992; Goddard et al., 1996). The current marker haplotypes of chromosomes carrying disease alleles are the outcome of their patterns of shared ancestry, and recombination events occurring in the meioses of that ancestry.

The first attempt at Monte Carlo likelihood analysis for this problem (Kaplan et al., 1995) used forwards simulation of the population, but suffered again from the disadvantage of being unable to condition effectively on current data. The methods of Rannala and Slatkin (1998) and Graham and Thompson (1998) use Monte Carlo realization of the coalescent ancestry of the disease sample as the basic tool in obtaining a Monte Carlo likelihood for fine-scale localization of a rare allele. Note that the ancestry of a sample ascertained for a rare allele is quite different from that of a random sample from the population. There is a very strong ascertainment effect: Griffiths and Tavaré (1998) provide applicable results.

In the case of Graham and Thompson (1998), recombinations relative to the putative disease locus are then realized on the ancestry, and exact computational methods used to compute the likelihood contribution of a given recombination history. For a single marker at recombination frequency ρ to the disease locus

$$L(\rho) = P_{\mathbf{q},\rho,\Pi}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X}) P_{\rho,\Pi}(\mathbf{X}) \tag{11.4}$$

where $\mathbf{X}$ denotes the latent variables of coalescent ancestry at the disease locus, and recombination events between disease locus and marker occurring in the meioses of that ancestry. The nuisance parameters are marker allele frequencies $\mathbf{q}$ which enter only into the penetrance probability $P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})$, and Π the parameters of the demographic history of the population. Note the similarity of equation (11.4) to those of likelihoods on pedigrees, for example equations (1.5), (3.9), or (7.8). However, unlike the Monte Carlo likelihoods based on those equations, here

$P_{\mathbf{q}}(\mathbf{Y} \mid \mathbf{X})$ is computed exactly, while the latent variables $\mathbf{X}$ are realized from their distribution under the given population model and hypothesized recombination frequency ρ. Thus a direct Monte Carlo estimate of the likelihood (11.4) is obtained.

Between the time-scale of evolution and coalescent ancestry and that of meioses in a defined pedigree, are the population-genetic models that provide probability distributions for the change of allele frequencies over generations, due to migration, population admixture, and random genetic drift. Here also, Monte Carlo methods of likelihood computation may be applied, the data $\mathbf{Y}$ being allele sample counts for different alleles, at different generations, and the latent variables $\mathbf{X}$ being the underlying true allele counts. Parameters of interest are those that determine the rate of change of allele frequencies, including the effective population size. Estimation of effective population size is of interest in the assessment of endangered populations. The dependence structure is identical to that of Figure 6.1. Instead of first-order Markov dependence of meioses at loci along a chromosome, we have first-order Markov generation-to-generation transitions of allele frequencies. The samples $\mathbf{Y}_j$ taken at a given generation j depend only on the allele frequencies $\mathbf{X}_j$ at that time. Equation (6.1) gives the form of the likelihood. Anderson and Thompson (1999) have used MCMC to obtain Monte Carlo likelihoods for the problem of estimating effective population size.

At every level, genetics provides examples of clearly defined highly structured probability models. The latent variables of genetics are "real": meioses, genotypes, recombination events, allele counts, and ancestral history. Monte Carlo methods are well suited to these problems, and often exact computation of likelihoods and probability distributions is infeasible. This final chapter has described a number of areas in which these methods are being applied, beyond those of linkage analysis from pedigree data which has been the focus of earlier chapters. These are only a few current examples; doubtless others will follow.

Bibliography

Andersen, S. K., Olesen, K. G., Jensen, F. V. and Jensen, F. (1989), HUGIN—a shell for building Bayesian belief universes for expert systems, *in* N. S. Sridharan, ed., 'Proceedings on the Eleveth International Joint Conference on Artificial Intellience', Morgan Kaufmann, San Mateo, CA, pp. 1080–1085.

Anderson, E. C. and Thompson, E. A. (1999), MCMC likelihoods for population genetics, *in* 'Proceedings of the 52nd Session of the International Statistical Institute', Vol. 3, pp. 347–348.

Baum, L. E. (1972), An inequality and associated maximization technique in statistical estimation for probabilistic functions on Markov processes, *in* O. Shisha, ed., 'Inequalities-III; Proceedings of the Third Symposium on Inequalities. University of California Los Angeles, 1969', Academic Press, New York, pp. 1–8.

Baum, L. E. and Petrie, T. (1966), Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics* **37**, 1554–1563.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions on Markov chains, *Annals of Mathematical Statistics* **41**, 164–171.

Bernstein, F. (1925), Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen, *Zeitschrift für induktiv Abstammungs- und VererbungsLehre* **37**, 237–270. English translation in: *Selected contributions to the literature of blood groups and immunology (Dunsford Memorial). I. The ABO system.* U.S.Army Medical Research Laboratory; Fort Knox (1966).

Besag, J. E., Green, P., Higdon, D. and Mengerson, K. (1995), Bayesian computation and stochastic systems, *Statistical Science* **10**, 3–66.

Bickeböller, H. and Thompson, E. A. (1996*a*), Distribution of genome shared IBD by half sibs: approximation via the Poisson clumping heuristic, *Theoretical Population Biology* **50**, 66–90.

Bickeböller, H. and Thompson, E. A. (1996*b*), The probability distribution of the amount of an individual's genome surviving to the following generation, *Genetics* **143**, 1043–1049.

Boehnke, M. (1994), Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes, *Am J Hum Genet* **55**, 379–390.

Boehnke, M. and Cox, N. J. (1997), Accurate inference of relationships in sib-pair linkage studies, *American Journal of Human Genetics* **61**, 423–429.

Botstein, D., White, R. L., Skolnick, M. H. and Davis, R. W. (1980), Construction of a linkage map in man using restriction fragment polymorphism, *American Journal of Human Genetics* **132**, 314–331.

Broman, K. W. (1999), Cleaning genotypic data, *Genetic Epidemiology* **17**, Suppl. 79–83.

Broman, K. W. and Weber, J. (2000), Characterization of human crossover interference, *American Journal of Human Genetics* **66**, 1911–1926.

Browning, S. (1998), Relationship information contained in gamete identity by descent data, *Journal of Computational Biology* **5**, 323–334.

Browning, S. (1999), Monte Carlo Likelihood Calculation for Identity by Descent Data, Ph.D. thesis, Department of Statistics, University of Washington.

Browning, S. and Thompson, E. A. (1999), Interference in the analysis of genetic marker data, *American Journal of Human Genetics* **65**, Suppl. A244.

Cannings, C., Thompson, E. A. and Skolnick, M. H. (1978), Probability functions on complex pedigrees, *Advances of Applied Probability* **10**, 26–61.

Cannings, C., Thompson, E. A. and Skolnick, M. H. (1980), Pedigree analysis of complex models, *in* J. Mielke and M. Crawford, eds, 'Current Developments in Anthropological Genetics', Plenum Press, New York, pp. 251–298.

Ceppelini, R., Siniscalco, M. and Smith, C. A. B. (1955), The estimation of gene frequencies in a random mating population, *Annals of Human Genetics* **20**, 97–115.

Cotterman, C. W. (1974), A Calculus for Statistico-Genetics. Ph.D. Thesis 1940. Ohio State University, *in* P. A. Ballonoff, ed., 'Genetics and Social Structure', Academic Press, New York.

Cottingham, R. W., Idury, R. M. and Schäffer, A. A. (1993), Faster sequential genetic linkage computations, *American Journal of Human Genetics* **53**, 252–263.

Cox, T. K., Kerem, B., Rommens, J., Iannuzzi, M. C., Drumm, M., Collins, F. S., Dean, M. and et al. (1989), Mapping of the cystic fibrosis gene using putative ancestral recombinants, *American Journal of Human Genetics* **45 Suppl**, A136.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with Discussion), *Journal of the Royal Statistical Society, B* **39**, 1–37.

Denniston, C. (1975), Probability and genetic relationship: two loci, *Annals of Human Genetics* **39**, 89–104.

Donnelly, K. P. (1983), The probability that related individuals share some section of genome identical by descent, *Theoretical Population Biology* **23**, 34–63.

Edwards, A. W. F. (1967), Automatic construction of genealogies from phenotypic information (AUTOKIN), *Bulletin of the European Society of Human Genetics* **1**, 42–43.

Edwards, A. W. F. (1972), *Likelihood*, Cambridge University Press, Cambridge, UK.

Elston, R. C. and Stewart, J. (1971), A general model for the analysis of pedigree data, *Human Heredity* **21**, 523–542.

Fisher, R. A. (1922), The systematic location of genes by means of crossover observations, *American Naturalist* **56**, 406–411.

Fisher, R. A. (1948), A quantitative theory of genetic recombination and chiasma formation, *Biometrics* **4**, 1–13.

Geman, S. and Geman, D. (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geyer, C. J. (1991*a*), Markov chain Monte Carlo maximum likelihood., *in* E. M. Keramidas and S. M. Kaufman, eds, 'Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface', Interface Foundation of North America, Fairfax Station, VA, pp. 156–163.

Geyer, C. J. (1991*b*), Reweighting Monte Carlo Mixtures, Technical Report 568, School of Statistics, Univ. of Minn.

Geyer, C. J. (1992), Practical Markov chain Monte Carlo (with discussion), *Statistical Science* **7**, 473–511.

Geyer, C. J. and Thompson, E. A. (1992), Constrained Monte Carlo maximum likelihood for dependent data, (with dicussion), *Journal of the Royal Statistical Society, B* **54**, 657–699.

Geyer, C. J. and Thompson, E. A. (1995), Annealing Markov chain Monte Carlo with applications to ancestral inference, *Journal of the American Statistical Association* **90**, 909–920.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds (1996), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, New York.

Goddard, K. A., Yu, C. E., Oshima, J., Miki, T., Nakura, J., Piussan, C., Martin, G. M. and et al. (1996), Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers, *American Journal of Human Genetics* **58**, 1286–1302.

Goldstein, D. R., Zhao, H. and Speed, T. P. (1995), Relative efficiencies of chi 2 models of recombination for exclusion mapping and gene ordering, *Genomics* **27**, 265–273.

Graham, J. and Thompson, E. A. (1998), Disequilibrium likelihoods for fine-scale mapping of a rare allele, *American Journal of Human Genetics* **63**, 1517–1530.

Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711–732.

Griffiths, R. C. and Tavaré, S. (1994*a*), Ancestral Inference in Population Genetics, *Statistical Science* **9**, 307–319.

Griffiths, R. C. and Tavaré, S. (1994*b*), Simulating probability distributions in the coalescent, *Theoretical Population Biology* **46**, 131–159.

Griffiths, R. C. and Tavaré, S. (1998), The age of a mutation in a general coalescent tree, *Stochastic Models* **14**, 273–295.

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

Guo, S. W. and Thompson, E. A. (1994), Monte Carlo estimation of mixed models for large complex pedigrees, *Biometrics* **50**, 417–432.

Haldane, J. B. S. (1919), The combination of linkage values and the calculation of distances between the loci of linked factors, *Journal of Genetics* **8**, 229–309.

Haldane, J. B. S. and Smith, C. A. B. (1947), A new estimate of the linkage between the genes for colour-blindness and haemolphilia in man, *Annals of Eugenics* **14**, 10–31.

Hammersley, J. M. and Handscomb, D. C. (1964), *Monte Carlo Methods*, Methuen and Co., London, UK.

Harbron, C. (1995), A pedigree based algorithm for finding efficient peeling sequences, *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **12**, 13–27.

Harbron, C. and Thomas, A. W. (1994), Alternative graphical representations of genotypes in a pedigree, *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **11**, 217–228.

Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. and Lander, E. (1992), Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland, *Nature Genetics* **2**, 204–211.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

Heath, S. C. (1997), Markov chain Monte Carlo segregation and linkage analysis for oligogenic models, *American Journal of Human Genetics* **61**, 748–760.

Heath, S. and Thompson, E. A. (1997), MCMC samplers for multilocus analyses on complex pedigrees, *American Journal of Human Genetics* **61**, A278.

Henderson, C. R. (1976), A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values, *Biometrics* **32**, 69–83.

Heuch, I. and Li, F. M. H. (1972), PEDIG—A computer program for calculation of genotype probabilities, using phenotypic information, *Clinical Genetics* **3**, 501–504.

Hilden, J. (1970), GENEX—An algebraic approach to pedigree probability calculus, *Clinical Genetics* **1**, 319–348.

Hoeschele, I. (1994), Bayesian QTL mapping via the Gibbs sampler, *in* 'Proc. 5th World Congr. Genet. Appl. Livst. Prod.', Vol. 21, Guelph, Canada, pp. 241–244.

Hulten, M., Lawrie, N. M. and Laurie, D. A. (1990), Chiasma-based genetic map of chromosome 21, *American Journal of Medical Genetics* **Suppl. 7**, 148–154.

Irwin, M., Cox, N. and Kong, A. (1994), Sequential imputation for multilocus linkage analysis, *Proceedings of the National Academy of Sciences (USA)* **91**, 11684–11688.

Jensen, C. S., Kjaerulff, U. and Kong, A. (1995), Blocking Gibbs sampling in very large probabilistic expert systems., *International Journal of Human-Computer Studies* **42**, 647–666.

Jensen, C. S. and Kong, A. (1999), Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops, *American Journal of Human Genetics* **65**, 885–901.

Kaplan, N. L., Hill, W. G. and Weir, B. S. (1995), Likelihood methods for locating disease genes in nonequilibrium populations, *American Journal of Human Genetics* **56**, 18–32.

Karigl, G. (1981), A recursive algorithm for the calculation of gene identity coefficients, *Annals of Human Genetics* **45**, 299–305.

Karlin, S. and Liberman, U. (1979), A natural class of recombination processes and related measures of crossover interference, *Advances of Applied Probability* **11**, 479–501.

Karunaratne, P. and Elston, R. (1998), Likelihood calculation conditional on observed pedigree structure, *American Journal of Human Genetics* **62**, 738–739.

King, T. R., Dove, W. F., Guénet, J., Hermann, B. G. and Shedlovsky, A. (1991), Meiotic mappinge of murine chromosome 17: The string of loci around *l(17)-2-Pas.*, *Mammalian Genome* **1**, 37–46.

Kingman, J. F. C. (1982), The Coalescent, *Stochastic Processes* **13**, 235–248.

Knott, S. and Haley, C. (1992), Maximum Likelihood Mapping of Quantitative Trait Loci Using Full–Sib Families, *Genetics* **132**, 1211–1222.

Kong, A. (1991), Analysis of pedigree data using methods combining peeling and Gibbs sampling, *in* E. M. Keramidas and S. M. Kaufman, eds, 'Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface', Interface Foundation of North America, Fairfax Station, VA, pp. 379–385.

Kong, A., Liu, J. and Wong, W. H. (1994), Sequential imputations and Bayesian missing data problems, *Journal of the American Statistical Association* **89**, 278–288.

Kosambi, D. D. (1944), The estimation of map distances from recombination values, *Annals of Eugenics* **12**, 172–175.

Kruglyak, L., Daly, M. J. and Lander, E. S. (1995), Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping, *American Journal of Human Genetics* **56**, 519–527.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996), Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics* **58**, 1347–1363.

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

Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995), Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling, *Genetics* **140**, 1421–1430.

Kullback, S. and Leibler, R. A. (1951), On information and sufficiency, *Annals of Statistics* **22**, 79–86.

Kumm, J., Browning, S. and Thompson, E. A. (1999), Validation of pedigree data in the presence of genotyping error, *American Journal of Human Genetics* **65**, Suppl. A???

Lander, E. S. and Botstein, D. (1987), Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children, *Science* **236**, 1567–1570.

Lander, E. S. and Green, P. (1987), Construction of multilocus genetic linkage maps in humans., *Proceedings of the National Academy of Sciences (USA)* **84**, 2363–2367.

Lange, K. (1997), *Mathematical and Statistical Methods for Genetic Analysis*, Statistics for Biology and Health, Springer Verlag, New York.

Lange, K. and Matthysse, S. (1989), Simulation of pedigree genotypes by random walks, *American Journal of Human Genetics* **45**, 959–970.

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

Lathrop, G. M., Lalouel, J. M. and Julier, C. amd Ott, J. (1984), Strategies for multilocus linkage analysis in humans, *Proceedings of the National Academy of Sciences (USA)* **81**, 3443–3446.

Lauritzen, S. J. (1992), Propagation of probabilities, means and variances in mixed graphical association models, *Journal of the American Statistical Association* **87**, 1098–1108.

Liberman, U. and Karlin, S. (1984), Theoretical models of genetic map functions, *Theoretical Population Biology* **25**, 331–346.

Lin, S. and Speed, T. P. (1996), Incorporating crossover interference into pedigree analysis using the χ^2 model, *Human Heredity* **46**, 315–322.

Lin, S., Thompson, E. A. and Wijsman, E. M. (1993), Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data, *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **10**, 1–17.

Lin, S., Thompson, E. A. and Wijsman, E. M. (1994), An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees, *Annals of Human Genetics* **58**, 343–357.

Long, A., Mullaney, S., Reid, L., Fry, J., Langley, C. and Mackay, T. (1995), High Resolution Mapping of Genetic Factors Affecting Abdominal Bristle Number in *Drosophila melanogaster*, *Genetics* **139**, 1273–1291.

MacCluer, J. W., VandeBerg, J. L., Read, B. and Ryder, O. A. (1986), Pedigree Analysis by Computer Simulation, *Zoo Biology* **5**, 147–160.

Mather, K. (1938), Crossing-over, *Biological Reviews of the Cambridge Philosophical Society* **13**, 252–292.

Mendel, G. (1866), Experiments in Plant Hybridisation, *in* J. H. Bennett, ed., 'English translation and commentary by R. A. Fisher', Oliver and Boyd, Edinburgh, 1965.

Meng, X. L. and Wong, W. H. (1996), Simulating ratios of normalizing constants via simple identity: A theoretical exploration, *Statistica Sinica* **6**, 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.

Morton, N. E. (1955), Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**, 277–318.

Morton, N. E. and MacLean., C. J. (1974), Analysis of family resemblance. III. Complex segregation of quantitative traits, *American Journal of Human Genetics* **26**, 489–503.

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

Ott, J. (1999), *Analysis of Human Genetic Linkage*, 3 rd. ed., The Johns Hopkins University Press, Baltimore, MD.

Ploughman, L. M. and Boehnke, M. (1989), Estimating the power of a proposed linkage study for a complex genetic trait, *American Journal of Human Genetics* **44**, 543–551.

Rannala, B. and Slatkin, M. (1998), Likelihood analysis of disequilibrium mapping, and related problems, *American Journal of Human Genetics* **62**, 459–473.

Redner, R. A. and Walker, H. F. (1984), Mixture Densities, Maximum Likelihood and the EM Algorithm, *SIAM Review* **26**, 195–202.

Remmers, E. F., Du, Y., Ding, Y. P., Kotake, S., Ge, L., Zha, H., Goldmuntz, E. A., Hansen, C. and Wilder, R. L. (1996), Localization of the gene responsible for the op (osteopetrotic) defect in rats on chromosome 10, *Journal of Bone and Mineral Research* **11**, 1856–1861.

Robbins, R. B. (1918), Some applications of mathematics to breeding problems. III, *Genetics* **3**, 375–389.

Satagopan, J. M., Yandell, B. S., Newton, M. A. and Osborn, T. C. (1996), A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo., *Genetics* **144**, 805–816.

Sheehan, N. A. (1990), Genetic Restoration on Complex Pedigrees, Ph.d. thesis, Department of Statistics, University of Washington.

Sheehan, N. A. and Thomas, A. W. (1993), On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme, *Biometrics* **49**, 163–175.

Smith, C. A. B. (1953), Detection of linkage in human genetics, *Journal of the Royal Statistical Society, B* **15**, 153–192.

Sobel, E. and Lange, K. (1993), Metropolis sampling in pedigree analysis, *Statistical Methods in Medical Research* **2**, 263–282.

Sobel, E. and Lange, K. (1996), Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics, *American Journal of Human Genetics* **58**, 1323–1337.

Sorensen, D. A., Andersen, S., Gianola, D. and Korsgaard, I. R. (1995), Bayesian inference in threshold models using Gibbs sampling, *Genetics Selection Evolution* **27**, 229–249.

Speed, T. (1996), What is a genetic map function?, *in* T. Speed and M. S. Waterman, eds, 'Genetic Mapping and DNA Sequencing', Vol. 81 of *IMA Volumes in Mathematics and its Applications*, Springer-Verlag, New York, pp. 65–88.

Stephens, M. and Donnelly, P. (2000), Inference in molecular population genetics (with Discussion), *Journal of the Royal Statistical Society, B* **62**, in press.

Sturt, E. (1976), A mapping function for human chromosomes, *Annals of Human Genetics* **40**, 147–163.

Thompson, E. A. (1974), Gene identities and multiple relationships, *Biometrics* **30**, 667–680.

Thompson, E. A. (1981), Pedigree analysis of Hodgkin's disease in a Newfoundland genealogy, *Annals of Human Genetics* **45**, 279–292.

Thompson, E. A. (1986), *Pedigree Analysis in Human Genetics*, Johns Hopkins University Press, Baltimore.

Thompson, E. A. (1988), Two-locus and three-locus gene identity by descent in pedigrees, *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **5**, 261–280.

Thompson, E. A. (1991), Probabilities on complex pedigrees: the Gibbs sampler approach, *in* E. M. Keramidas and S. M. Kaufman, eds, 'Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface', Interface Foundation of North America, Fairfax Station, VA, pp. 321–328.

Thompson, E. A. (1994*a*), Monte Carlo estimation of multilocus autozygosity probabilities, *in* J. Sall and A. Lehman, eds, 'Proceedings of the 1994 Interface conference', Fairfax Station, VA, pp. 498–506.

Thompson, E. A. (1994*b*), Monte Carlo likelihood in genetic mapping, *Statistical Science* **9**, 355–366.

Thompson, E. A. (1994*c*), Monte Carlo likelihood in the genetic mapping of complex traits, *Philosophical Transactions of the Royal Society of London (Series B)* **344**, 345–351.

Thompson, E. A. (1997), Conditional gene identity in affected individuals, *in* I. H. Pawlowitzki, J. H. Edwards and E. A. Thompson, eds, 'Genetic Mapping of Disease Genes', Academic Press, London, pp. 137–146.

Thompson, E. A. (2000*a*), MCMC estimation of multi-locus genome sharing and multipoint gene location scores, *International Statistical Review* **68**, 53–73.

Thompson, E. A. (2000*b*), Monte Carlo methods on Genetic Structures, *in* O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg, eds, 'Complex Stochastic Systems', Séminaire Européan de Statistique, Chapman and Hall, London, UK, pp. 179–222:in press.

Thompson, E. A. and Guo, S. W. (1991), Evaluation of likelihood ratios for complex genetic models, *I.M.A. Journal of Mathematics Applied in Medicine & Biology* **8**, 149–169.

Thompson, E. A. and Heath, S. C. (1999), Estimation of conditional multilocus gene identity among relatives, *in* F. Seillier-Moiseiwitsch, ed., 'Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology', IMS Lecture Note–Monograph Series Volume 33, Institute of Mathematical Statistics, Hayward, CA, pp. 95–113.

Thompson, E. A., Kravitz, K., Hill, J. and Skolnick, M. H. (1978), Linkage and the power of a pedigree structure, *in* N. E. Morton, ed., 'Genetic Epidemiology', Academic Press, New York, pp. 247–253.

Thompson, E. A. and Meagher, T. R. (1998), Genetic linkage in the estimation of pairwise relationship, *Theoretical and Applied Genetics* **97**, 857–864.

Thompson, E. A. and Morgan, K. (1989), Recursive descent probabilities for rare recessive lethals, *Annals of Human Genetics* **53**, 357–374.

Thompson, E. A. and Shaw, R. G. (1990), Pedigree analysis for quantitative traits: Variance components without matrix inversion, *Biometrics* **46**, 399–414.

Thompson, E. A. and Shaw, R. G. (1992), Estimating polygenic models for multivariate data on large pedigrees, *Genetics* **131**, 971–978.

Weeks, D. E., Lathrop, G. M. and Ott, J. (1993), Multipoint mapping under genetic interference., *Human Heredity* **43**, 86–97.

Weinberg, W. (1912), Zur Verebung der Anlage der Bluterkrankheit mit methodol. Erganzungen meiner Geschwistermethode, *Arch. Rass. u. GesBiol.* **9**, 694–709.

Weinstein, A. (1936), The theory of multiple-strand crossing over, *Genetics* **21**, 155–199.

Weir, B. S. (1996), *Genetic Data Analysis II*, Sinauer Associates, Inc., Sunderland, MA.

Whittemore, A. S. and Tu, I.-P. (1998), Simple, Robust Linkage Tests for Affected Sibs, *American Journal of Human Genetics* **62**, 1228–1242.

Wright, S. (1922), Coefficients of inbreeding and relationship, *American Naturalist* **56**, 330–338.

Wright, S. and McPhee, H. C. (1925), An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees, *Journal of Agricultural Research* **31**, 377–383.

Zeng, Z. (1994), Precision Mapping of Quantitative Trait Loci, *Genetics* **136**, 1457–1468.

Zhao, H., Speed, T. P. and McPeek, M. S. (1995), Statistical analysis of crossover interference using the chi-square model, *Genetics* **139**, 1045–1056.

Cited web sites

http://linkage.rockefeller.edu/soft/list.html
The Rockefeller Genetic Linkage Software list
http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml
Software for Inferences from Genetic Data on Pedigrees.