THE UNIVERSITY OF CHICAGO

SEQUENTIAL IMPUTATION AND MULTILOCUS LINKAGE ANALYSIS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

MARK EDWARD IRWIN

CHICAGO, ILLINOIS

MARCH 1995

Acknowledgements

I would like to thank my advisor, Augustine Kong, for his assistance in producing this work and for pushing to get it completed. I would also like to thank Nancy Cox for helping me to understand the genetics, Michael Frigge for helping to get the software working, and Fred Wright for helping with the genetics description.

This research was supported in part by National Institutes of Health grant GM-46800. The computation for this work was performed using computer facilities supported in part by the National Science Foundation Grants DMS 89-05292, DMS 87-03942 and DMS 86-01732, awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

Table of Contents

List of Figures

List of Tables

Abstract

Multilocus linkage analysis is an important tool in estimating the location of a gene of interest on a chromosome, giving a starting point for molecular geneticists to precisely locate and characterize the gene. The data for the analysis consists of a family structure, observed trait data influenced by the gene in question for some members of the family, and data from a collection of genetic markers. Then a model describing relationship between the observed trait information and the gene in question and the distances between the loci is adopted. Given this model, the data are used to construct a likelihood surface. There exist two general approaches for calculating likelihood surfaces, exact calculation by peeling algorithms, or approximation using Markov Chain Monte Carlo.

An alternative Monte Carlo approach called sequential imputation is proposed here for multilocus likelihood computations. This method is most useful in mapping situations where the data consists of large pedigrees with substantial missing information and it is desirable to perform linkage analysis utilizing data from many polymorphic markers simultaneously. For this type of problem, we believe that sequential imputation is more efficient than exact calculation or Markov Chain Monte Carlo methods. A pedigree with 155 individuals, 9 loci, and 155,520 haplotypes will be used to illustrate how sequential imputation can be used for approximating likelihood surfaces, parameter estimation, standard error calculation, and bayesian analysis.

Chapter 1

Introduction

1.1 Summary

With activities such as the Human Genome Project and the search for disease causing genes, there is great demand for efficient methods of mapping the human chromosome. When mapping disease genes or genetic markers, it is usually more efficient statistically to handle many linked loci simultaneously. However with the advances in laboratory techniques leading to highly polymorphic loci, the current approaches may not be able handle this more powerful data computationally. In particular, linkage analyses involving large pedigrees with many polymorphic markers can be extremely difficult to do.

Because the likelihood function cannot be obtained in closed form, one popular and standard approach is to evaluate the likelihoods point by point with exact calculations. Computationally efficient algorithms for calculating likelihoods are available for large pedigrees with a small number of loci, (Elston and Stewart, 1971, Lange

and Elston, 1975, Cannings, Thompson, and Skolnick, 1978, Lange and Boehnke, 1983, and Lathrop, Lalouel, Julier and Ott, 1984), and for small pedigrees with a large number of loci (Lander and Green, 1987). However, for large pedigrees with a large number of loci, especially those which have substantial missing data, exact evaluation of a single likelihood value by these peeling methods can be prohibitive. With these exact calculations, the memory requirements and computing time depend on the number of possible outcomes for each person given the observed data, which is the product of the number of possiblities for each locus. If, for example, a person in the pedigree has no observed marker data, the number of possibilities for that person can be huge. As many pedigrees currently used in linkage studies have little or no marker data from the members of top two or three generations, this is a common and serious problem. This difficulty was noted explicitly by investigators studying disorders having late age at onset, such as diabetes (Rothschild et al., 1993) or Alzheimer's disease (Schellenberg et al, 1992).

A second popular approach is based on Markov Chain Monte Carlo methods (Geman and Geman, 1984, and Gelfand and Smith, 1990). Instead of exact likelihood calculation, multiple correlated samples of the missing data are imputed conditional on the observed data, often by the Gibbs sampler (Kong, 1991b, and Guo and Thompson, 1992) or the Metropolis algorithm (Lange and Sobel, 1991). These approaches can appreciably reduce the the demand on computational resources as compared with the exact calculation methods, particularly the memory requirements. This makes the Markov Chain Monte Carlo methods particularly useful for dealing with highly

inbred pedigrees or complex traits. However for analyses involving large pedigrees and many loci, the resulting Markov Chain may take a very long time to converge to its stationary distribution due to the high correlations between the samples. In addition, if some of the loci have three or more possible alleles, the resulting chain may not be irreducible (Sheehan and Thomas, 1993), and thus the samples would not be generated from the desired distribution.

A novel Monte Carlo method called *sequential imputation* (Kong, Liu, and Wong, 1994) is proposed here to handle analyses with large pedigrees and many loci. It combines features of the exact computations and the Monte Carlo methods. Loci are processed one, or a few, at a time to reduce the computational burden. The result is a collection of complete independent data sets with associated weights. Instead of evaluating likelihood values individually, the whole likelihood surface can sometimes be obtained using results from a single simulation run. In addition to likelihood surface estimation, the simulated data sets and weights can be also be used for other purposes, such as simultaneous parameter estimation, sensitivity analysis, or approximation of bayesian posterior distributions. Also, unlike some other methods (Lander and Green, 1987), sequential imputation can deal with more complex models, such as incorporating genetic interference with no extra difficulties.

In the next section a brief description of genetic linkage analysis is given. For a more complete treatment of the topic, the books by Ott (1991) or Thompson (1986) are useful places to start.

In chapter two, sequential imputation in the genetic setting will be introduced.

The effect of the number of the imputations, the decomposition of the missing data and the order of processing on the efficiency of the procedure will be discussed. Two efficient methods for estimating the location of a single gene given a fixed marker map will be proposed. The advantage of these two procedures is that while the disease data will be processed many times for each set of imputations, the marker data only needs to be processed once. If any of the markers have a moderate number of alleles, processing the markers will take the bulk of the computing time, not the disease locus, and thus computing time can be greatly lowered.

Chapter three will discuss three additional applications where sequential imputation can be a useful tool: parameter estimation by Monte Carlo EM (MCEM), standard error estimation, and bayesian analysis. The sequential imputation implementation of MCEM has the big advantage that it is possible to run the procedure with only one set of imputed data. The other currently used MCEM approaches need to impute new sets of missing data each time the parameter estimate is updated. Also discussed in this section will be a hybrid procedure combining Monte Carlo sampling with direct calculation of conditional log likelihoods. The sequential imputation samples generated for the MCEM procedure can also be used to give standard error estimates for the MCEM parameter estimates by approximating the information matrix. This helps alleviate the problem of standard errors of parameter estimates not being reported due to computational difficulties. Finally a method of approximating posterior distributions and moments with the sequential imputation samples will be examined. Assuming a prior distribution which is conjugate for the complete

data, the required calculations are easy to do. These calculations are also useful for examining likelihood surfaces. To exhibit these three applications, the problem of simultaneous estimation of marker recombination fractions will be examined.

In chapter four, the methods of chapters two and three will be applied to a pedigree of 155 individuals segregating for Maturity Onset Diabetes of the Young (MODY). Data from 9 loci (8 markers plus the disease locus) leading to 155,520 haplotypes will be used in this example. This example is beyond the scope of any existing programs which do exact likelihood calculations. In this example, when it was possible to make the comparison, the accuracy of the estimates of the likelihood function given by sequential imputation was very good. In addition, it also appears, that at least for this example, the estimation procedures are relatively robust to moderate deviations from the model used for simulation.

Finally in chapter 5, a few concluding remarks are made, focusing on comparisons with the exact calculation methods and the Markov Chain Monte Carlo approaches.

1.2 Genetics Background

1.2.1 Chromosomes and Genes

In humans, genetic information is contained on 23 pairs of *chromosomes*. Chromosomes can be thought of as a strand of genetic material, and every cell in the body has an identical copy of the set of chromosome pairs. Each *homologous* chromosome within a pair (excluding the sex chromosomes) is of a similar physical length, though

the lengths can very greatly between different pairs of chromosomes. Chromosome pairs can be identified and ordered by length, usually from longest (chromosome 1) to shortest (chromosome 22). Within each chromosome pair, one chromosome was inherited from the father and one from the mother.

A location on a chromosome pair is usually referred to as a *locus* (plural: loci). If the location has a genetic function, it usually is known as a *gene*. The location of each gene influencing a given trait is fixed to a single location on a chromosome pair, say at a distance d relative to the *centromere*, the location where the two arms of an homologous chromosome join. For example, the gene causing Huntington's disease occurs near the end of the short arm of chromosome 4 and the gene for Cystic Fibrosis is roughly in the middle of the long arm of chromosome 7. The genetic information at each locus can be described by two finite discrete random variables (one for each chromosome) known as *alleles*. If an allele has k possible states, the locus has $\binom{k}{2} + k$ possible states, or *genotypes*, as the origin of the alleles usually has no observable effect (a 1 from the father and a 2 from the mother is equivalent to a 2 from the father and a 1 from the mother). The observable expression of the genotype is known as the *phenotype*. It is possible for different genotypes to lead to the same phenotype. For example, with the ABO blood group locus, there are 6 different genotypes but only 4 possible phenotypes. The relationship is shown in table 1.1.

Phenotype	Possible Genotypes
Type A	AA or AO
Type B	BB or BO
Type AB	AB
Type O	OO

Table 1.1 Relationship Between Phenotype and Genotype for the ABO Locus

1.2.2 Crossovers, Recombinations and Linkage

When a parent passes on genetic material to a child, each chromosome pair acts independently of the rest. In addition, the contribution from one parent is independent of the other. However, a homologous chromosome passed on to a child by one parent is usually not an exact copy of one of the parental pair. During meiosis, the homologous chromosomes pair up and lie next to each other. By a process known as *crossing over*, pieces of the homologous chromosomes are exchanged, giving a new pair, one of which is passed on the the child (each with a probability of 1/2). The edges of the exchanged sections are referred to as *crossover points* or just *crossovers*. The actual locations of the crossovers cannot be directly observed, but sometimes it is possible to infer that one may have occurred. If it is possible to observe the genotypes of two loci, sometimes it can be determined that a odd number of crossovers between the two locations has occurred. The situation where this occurs is known as a *recombination* between the two loci. Figure 1.1 displays the situation where a crossover may lead to a recombination. In (1), loci A and B have not recombined as the crossover did

not occur between them. However in (2), loci C and D have recombined as a single crossover occurred between them. Finally in (3), there has not been a recombination between E and F, even though there have been 2 crossovers between them, as alleles from the same homologous chromosome get passed on together. This approach to thinking about recombination does not consider the situation where loci occur on different chromosome pairs. A more general definition is that a recombination occurs if the two alleles of the two loci inherited from one parent come from different grandparents. It should be noted that the above description of the crossover process is a simplication as the actual process involves four strands (each homologous chromosome splits into two identical strands before crossing over occurs), but the result is essentially the same.

An example of the type of data available to the geneticist is shown in Figure 1.2, where the inheritance of two loci by two children is shown. In this example, it is assumed that not only the genotypes of the parents are observed, but also how the alleles of the two loci pair up. Often this additional information is unknown as it depends on knowing the genotypes of relatives of the two parents and the allele pattern in these relatives. For this example, assume that the father has alleles A_1 and A_2 at the first locus and alleles B_1 and B_2 at the second locus. Also assume that A_1 and B_1 are together on one homologous chromosome and A_2 and B_2 are on the other. The data on the mother and two children is similar. In this example there are no recombinations between the two loci for Child 1. However, Child 2 has a recombination between the two loci inherited from the father. This child received A_1

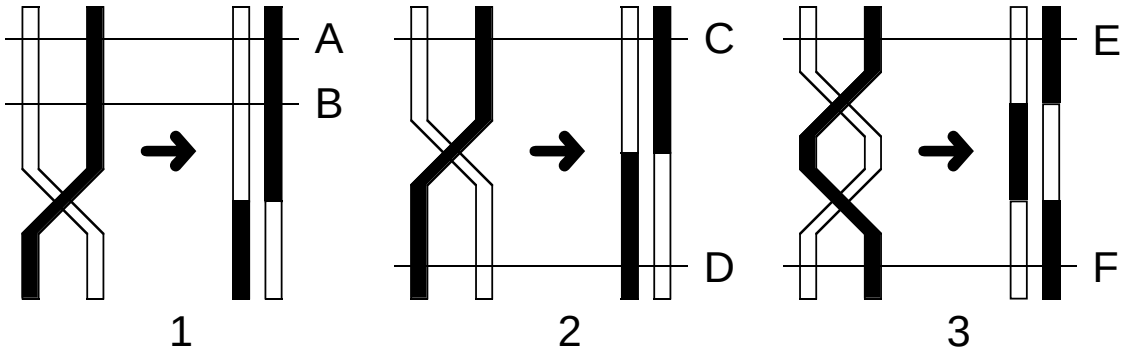

Figure 1.1 Examples of the relationship between crossovers and recombinations. The two chromosomes have been coloured to distinguish the genetic material. (1) No recombination and no crossovers between A and B. (2) A recombination between C and D with one crossover. (3) No recombination between E and F, as there are two crossovers between them.

at the first locus from the father, which was on one homologous chromosome, and B_2 at the second locus, which was on the other paternal chromosome. The probability that a parent will produce a recombination between loci A and B is known as the *recombination fraction*, usually denoted as θ_{AB}. This probability can be estimated by observating the fraction of chromosomes produced with this recombination. In this simple example, one of four meisoses observed lead to a recombination, giving a maximum likelihood estimate of the recombination fraction of $\hat{\theta}_{AB} = 0.25$.

Two loci are said to be *unlinked* if $\theta = 1/2$. If $\theta < 1/2$, the two loci are said to be *linked*. Loci with θ near zero are often referred to being *tightly linked*. It is possible for $\theta > 1/2$, though the situation usually is of little interest in linkage analysis. Generally, loci are linked if they occur on the same chromosome pair and unlinked

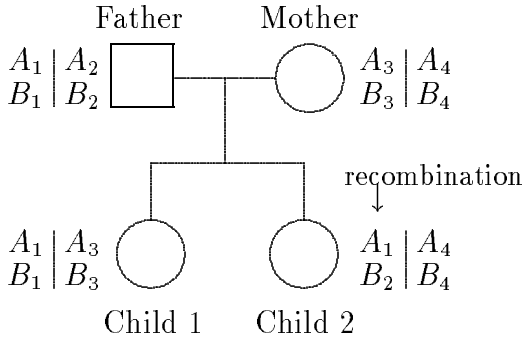

Figure 1.2 An example pedigree of a family with data from 2 loci. The alleles are given for each family member, with the chromosomes separated by vertical lines.

if they appear on different chromosome pairs. The closer two loci are physically, the smaller the recombination fraction between them, though the relationship between physical distance and recombination fractions is extremely complicated. *Linkage analysis* usually refers to estimating locations with recombination information.

Assume there are three ordered loci (A, B, and C) on a chromosome pair. It can be shown that the distance between loci A and C, θ_{AC} is not equal to $\theta_{AB} + \theta_{BC}$. Thus when examining more than two loci, it is often more convenient to use a different measure of distance. The most commonly used measure is the *map distance* d, the expected number of recombinations between two loci. The map distance has the desired property

$$d_{AC} = d_{AB} + d_{BC}.$$

To relate the map distance with the recombination fraction, a *map function* is often used. A number of common choices for map functions are given in Chapter 1 of Ott

(1991). Possibly the most important is Haldane's map function

$$d(\theta) = \begin{cases} -\frac{1}{2}\log(1-2\theta) & 0 \leq \theta < 0.5 \\ \infty & otherwise \end{cases} \tag{1.1}$$

with inverse

$$\theta(d) = \frac{1}{2}(1 - e^{-2d}). \tag{1.2}$$

If θ is small (say $\theta < 0.1$), $d \approx \theta$. This map function comes from the assumption that the crossovers follow a Poisson process, which implies that the recombinations are independent. The other popular map functions were introduced to model *interference*, a phenomenon where the recombinations are correlated. In many analyses, no interference is assumed, as it is easier to deal with computationally, and often has little effect on the accuracy of estimated gene locations, especially when the loci aren't tightly linked.

1.2.3 Genetic Markers and Likelihood

In the above example (Figure 1.2) is was possible to directly estimate the recombination fraction between the two loci as it was assumed it was known how the alleles for the two loci matched up in the parents. Normally this would not be known without outside information such as the genotypes of the siblings and parents of the parents in the pedigree. Also, when trying to locate a disease gene, the only data available is the disease phenotype, which often does not uniquely specify a disease genotype. In this situation, it is still possible to estimate the location of the disease gene by

using the information supplied by *marker* loci genotypes. A marker is a locus with an observable genotype and usually assumed to have a known location.

Given a model describing the relationship between phenotype and genotype for the gene of interest, the marker and phenotype data can be used to calculate the likelihood for the location of the gene (denoted θ). Let the observed phenotype and marker genotype data be denote by $\mathbf{y}$, and the complete genotype data be denoted by $\mathbf{z}$. Then the likelihood function is

$$\begin{aligned} L(\theta|\mathbf{y}) &= p_\theta(\mathbf{y}) = \int p_\theta(\mathbf{y},\mathbf{z})d\mathbf{z} \\ &= \int p_\theta(\mathbf{y}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z}. \end{aligned}$$

Note that $p_\theta(\mathbf{y}|\mathbf{z})$ does not depend on the locations of the loci. However it may depend on nuisance parameters describing the relationship between phenotype and genotype. The density $p_\theta(\mathbf{z})$ only depends on the locations of the loci.

Calculating this likelihood function may be difficult and time consuming. In human pedigree data, the family structures can be complicated and there may be appreciable missing data. The common approach to calculating the likelihood exactly is through the peeling approaches mentioned earlier.

Chapter 2

Sequential Imputation

2.1 Description

In multi-locus problems, if for each person and each locus, it is known exactly what allele type is inherited from the father and what allele type is inherited from the mother, the likelihood function is usually trivial to write down. Hence refer to the information that is desirable, but often not available (as least not entirely), as *missing data* and denote it by $\mathbf{z}$. The observed data, denoted by $\mathbf{y}$, usually include genotypes of each individual marker for some members of the pedigree. An individual may be typed for some but not all of the marker loci. Note that the genotype of a locus for an individual consists of the types of two alleles. When the locus is heterozygous, the genotype itself does not contain information on descent, i.e., which allele is inherited from which parent. In the case of disease mapping, $\mathbf{y}$ will also include available disease phenotypes of the members. The combination $(\mathbf{y}, \mathbf{z})$ is referred to as the *complete*

data. Let θ be the unknown parameter vector so that the likelihood function is

$$L(\theta) = p_\theta(\mathbf{y}).$$

In the case of disease mapping, θ is often a scalar which denotes the location of the disease gene relative to a set of markers whose locations are assumed to be known. In more complicated situations, θ may also incorporate other parameters such as population allele frequencies and nuisance parameters which appear in the model relating the disease genotype and phenotype. In linkage mapping of markers, θ is a vector which denotes the relative locations among a collection of markers. Sometimes, the order of the markers is known and the genetic distances between successive markers is of interest. In other situations, even the order is unknown and has to be estimated.

Let $\{y_1, \ldots, y_n\}$ and $\{z_1, \ldots, z_n\}$ be some decomposition of $\mathbf{y}$ and $\mathbf{z}$. At this time, assume that there are n loci so that for $t = 1, \ldots, n$, y_t and z_t are respectively the observed and missing data on locus t. Other decompositions will be considered in later sections. Note that the labels $t, t = 1, \ldots, n$, do not necessarily correspond to the *physical* ordering, assumed or real, of the loci. Given a certain value of θ, say θ_0, sequential imputation (Kong, Liu and Wong 1991) is a Monte Carlo method which allows us to obtain an unbiased estimate of $L(\theta_0)$ and generate weighted samples of $\mathbf{z} = \{z_1, \ldots, z_n\}$ from the conditional distribution $p_{\theta_0}(\mathbf{z}|\mathbf{y})$. The method involves first drawing z_1^* from $p_{\theta_0}(z_1|y_1)$ and computing $w_1 = p_{\theta_0}(y_1)$. Then the following two steps are applied for $t = 2, \ldots, n$, in increasing order of t:

(A) Draw z_t^* from the conditional distribution

$$p_{\theta_0}(z_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*, y_t). \tag{2.1}$$

Notice that the z_t^*'s have to be drawn sequentially since each z_t^* is drawn conditioned on the previous imputued missing parts $z_1^*, \ldots, z_{t-1}^*$.

(B) Sequentially compute the predictive probabilities $p_{\theta_0}(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*)$ and

$$w_t = w_{t-1} p_{\theta_0}(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*). \tag{2.2}$$

Let $w = w_n$ so that

$$w = p_{\theta_0}(y_1) \prod_{t=2}^{n} p_{\theta_0}(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*). \tag{2.3}$$

Given the decomposition described above, for each t, **(A)** and **(B)** are done simultaneously and involve a single locus peel (Ploughman and Boehnke 1989, Ott 1989). For details on simulating missing data for one locus conditioned on the imputed missing data of other loci see Kong (1991a). When steps **(A)** and **(B)** are done, the result is a set of imputed missing data $\mathbf{z}^* = (z_1^*, \ldots, z_t^*)$ with associated weight $w = w(\mathbf{y}, \mathbf{z}^*)$. This whole process is repeated independently m times. Let the results be denoted by $\mathbf{z}^*(1), \mathbf{z}^*(2), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$ where $\mathbf{z}^*(j) = (z_1^*(j), \ldots, z_n^*(j))$ and $w(j) = w(\mathbf{y}, \mathbf{z}^*(j))$ for $j = 1, \ldots, m$.

Theorem 2.1 *Let $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$ be the results of a sequential imputation run done at a parameter value θ_0. Then*

$$\overline{w} = \frac{1}{m} \sum_{j=1}^{m} w(j) \tag{2.4}$$

is an unbiased estimate of $L(\theta_0) = p_{\theta_0}(\mathbf{y})$.

Proof Note that $\mathbf{z}^*(j)$ is drawn from the density

$$\begin{aligned}
p^*_{\theta_0}(\mathbf{z}^*(j)|\mathbf{y}) &= p_{\theta_0}(z^*_1(j)|y_1)\prod_{t=2}^{n} p_{\theta_0}(z^*_t(j)|y_1, z^*_1(j), \ldots, y_{t-1}, z^*_{t-1}(j), y_t) \\
&= \frac{p_{\theta_0}(z^*_1(j), y_1)}{p_{\theta_0}(y_1)} \prod_{t=2}^{n} \frac{p_{\theta_0}(y_1, \ldots, y_t, z^*_1(j), \ldots, z^*_t(j))}{p_{\theta_0}(y_1, \ldots, y_t, z^*_1(j), \ldots, z^*_{t-1}(j))} \\
&= \frac{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(y_1)} \prod_{t=2}^{n} \frac{p_{\theta_0}(y_1, \ldots, y_{t-1}, z^*_1(j), \ldots, z^*_{t-1}(j))}{p_{\theta_0}(y_1, \ldots, y_t, z^*_1(j), \ldots, z^*_{t-1}(j))} \\
&= p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j) \frac{1}{p_{\theta_0}(y_1)\prod_{t=2}^{n} p_{\theta_0}(y_t|y_1, z^*_1(j), \ldots, y_{t-1}, z^*_{t-1}(j))} \\
&= \frac{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))}{w(\mathbf{y}, \mathbf{z}^*(j))} \\
&= p_{\theta_0}(\mathbf{z}^*(j)|\mathbf{y}) \frac{p_{\theta_0}(\mathbf{y})}{w(\mathbf{y}, \mathbf{z}^*(j))}. \qquad (2.5)
\end{aligned}$$

So

$$\begin{aligned}
E_{p^*}[w(\mathbf{y}, \mathbf{z}^*)|\mathbf{y}] &= \sum_{\mathbf{z}^*} w(\mathbf{y}, \mathbf{z}^*) p^*_{\theta_0}(\mathbf{z}^*|\mathbf{y}) \\
&= \sum_{\mathbf{z}^*} w(\mathbf{y}, \mathbf{z}^*) \frac{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))}{w(\mathbf{y}, \mathbf{z}^*)} \\
&= \sum_{\mathbf{z}^*} p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j)) \\
&= p_{\theta_0}(\mathbf{y}).
\end{aligned}$$

Thus $\overline{w}$ is an unbiased estimate of $L(\theta_0)$. □

In addition to getting an unbiased estimate of the likelihood, the samples $\mathbf{z}^*(j), j = 1, \ldots, m$ generated by sequential imputation can be treated as weighted samples (weight $\propto w(j)$) taken from the conditional distribution $p_{\theta_0}(\mathbf{z}|\mathbf{y})$. These samples can be used for many purposes. Suppose h is a function of $\mathbf{y}$ and $\mathbf{z}$, and suppose the conditional expectation

$$E_\theta[h(\mathbf{y}, \mathbf{z})|\mathbf{y}] \qquad (2.6)$$

is of interest for some value θ. However, suppose that (2.6) is difficult to compute directly, but $h(\mathbf{y},\mathbf{z})$ can be easily evaluated for any realization of $\mathbf{z}$. For example, consider the case where

$$h(\mathbf{y},\mathbf{z}) = \frac{p_{\theta_1}(\mathbf{y},\mathbf{z})}{p_{\theta_0}(\mathbf{y},\mathbf{z})} \tag{2.7}$$

where θ_0 and θ_1 are two values of θ, then as shown by Thompson and Guo,

$$\begin{aligned} E_{\theta_0}[h(\mathbf{y},\mathbf{z})|\mathbf{y}] &= \int \frac{p_{\theta_1}(\mathbf{y},\mathbf{z})}{p_{\theta_0}(\mathbf{y},\mathbf{z})} p_{\theta_0}(\mathbf{z}|\mathbf{y})d\mathbf{z} = \frac{1}{p_{\theta_0}(\mathbf{y})}\int p_{\theta_1}(\mathbf{y},\mathbf{z})d\mathbf{z} \\ &= \frac{p_{\theta_1}(\mathbf{y})}{p_{\theta_0}(\mathbf{y})} = \frac{L(\theta_1)}{L(\theta_0)} \end{aligned} \tag{2.8}$$

is the likelihood ratio. Note that both $p_{\theta_0}(\mathbf{y},\mathbf{z})$ and $p_{\theta_1}(\mathbf{y},\mathbf{z})$, the complete data likelihoods, can usually be easily evaluated for any $\mathbf{z}$.

In general,

$$E_\theta[h(\mathbf{y},\mathbf{z})|\mathbf{y}] = \int h(\mathbf{y},\mathbf{z})p_\theta(\mathbf{z}|\mathbf{y})d\mathbf{z}. \tag{2.9}$$

Thus if samples $\mathbf{z}(j), j = 1,\ldots,m$, can be drawn from $p_\theta(\mathbf{z}|\mathbf{y})$ for some value of θ, then

$$\frac{1}{m}\sum_{j=1}^{m} h(\mathbf{y},\mathbf{z}(j)) \tag{2.10}$$

is an unbiased estimate of (2.6) (Guo and Thompson, 1992). Unfortunately, sampling from the distribution $p_\theta(\mathbf{z}|\mathbf{y})$ directly requires peeling all n loci simultaneously. However, note that

$$\int h(\mathbf{y},\mathbf{z})p_\theta(\mathbf{z}|\mathbf{y})d\mathbf{z} = \int h(\mathbf{y},\mathbf{z})\frac{p_\theta(\mathbf{z}|\mathbf{y})}{p^*_\theta(\mathbf{z}|\mathbf{y})}p^*_\theta(\mathbf{z}|\mathbf{y})d\mathbf{z} \tag{2.11}$$

where p^* is as defined in (2.5). Therefore, by (2.11), if $\mathbf{z}^*(j)$ are samples generated

by sequential imputation, then

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m} h(\mathbf{y},\mathbf{z}^*(j))\frac{p_\theta(\mathbf{z}^*(j)|\mathbf{y})}{p^*_\theta(\mathbf{z}^*(j)|\mathbf{y})} &= \frac{1}{m}\sum_{j=1}^{m} h(\mathbf{y},\mathbf{z}^*(j))\frac{w(j)p_\theta(\mathbf{z}^*(j)|\mathbf{y})}{p_\theta(\mathbf{y},\mathbf{z}^*(j))} \\ &= \frac{1}{m}\sum_{j=1}^{m} h(\mathbf{y},\mathbf{z}^*(j))\frac{w(j)}{p_\theta(\mathbf{y})} \qquad (2.12)\end{aligned}$$

is an unbiased estimate of (2.6). Although $p_\theta(\mathbf{y})$ is unknown, it can be approximated by $\overline{w}$, so

$$\frac{1}{m\overline{w}}\sum_{j=1}^{m} w(j)h(\mathbf{y},\mathbf{z}^*(j)) = \sum_{j=1}^{m}\frac{w(j)}{W}h(\mathbf{y},\mathbf{z}^*(j)), \qquad (2.13)$$

where $W = \sum_k w(k)$, is a natural estimate of (2.6). The expression (2.13) is an *importance sampling* estimate using normalized importance sampling weights $w(j)/W$ (meaning the weights sum to one). Using normalized weights makes (2.13) a ratio estimate, which is biased. However the bias goes to zero as m approaches infinity and its contribution to the mean square error of the estimate is negligible for large m.

Theorem 2.2 *Let $\mathbf{z}^*(1),\ldots,\mathbf{z}^*(m)$ and $w(1),\ldots,w(m)$ be the results of a sequential imputation run done at a parameter value θ_0. Then*

$$\hat{L}(\theta_1) = \hat{p}_{\theta_1}(\mathbf{y}) = \frac{1}{m}\sum_{j=1}^{m}\frac{p_{\theta_1}(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*(j))}w(j) \qquad (2.14)$$

is an unbiased estimate of $L(\theta_1) = p_{\theta_1}(\mathbf{y})$.

Proof

$$\begin{aligned}E_{p^*_{\theta_0}}\left[\frac{p_{\theta_1}(\mathbf{y},\mathbf{z}^*)}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*)}w(\mathbf{y},\mathbf{z}^*)|\mathbf{y}\right] &= \sum_{\mathbf{z}^*}\frac{p_{\theta_1}(\mathbf{y},\mathbf{z}^*)}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*}w(\mathbf{y},\mathbf{z}^*)p^*_{\theta_0}(\mathbf{z}^*|\mathbf{y}) \\ &= \sum_{\mathbf{z}^*}\frac{p_{\theta_1}(\mathbf{y},\mathbf{z}^*)}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*)}w(\mathbf{y},\mathbf{z}^*)\frac{p_{\theta_0}(\mathbf{y},\mathbf{z}^*)}{w(\mathbf{y},\mathbf{z}^*)} \\ &= \sum_{\mathbf{z}^*}p_{\theta_1}(\mathbf{y},\mathbf{z}^*) \\ &= p_{\theta_1}(\mathbf{y}).\end{aligned}$$

Therefore $\hat{L}(\theta_1)$ is an unbiased estimate of $L(\theta_1)$. □

Therefore, by applying sequential imputation based on a parameter value θ_0, an unbiased estimate of the likelihood for any other parameter value can be obtained. However note that bias is not the issue here. If θ_1 is too far away from θ_0, the estimate (2.14) can have a very large variance.

2.2 Efficiency and Sample Size

2.2.1 Coefficients of Variation of the Weights

The coefficient of variation of $\overline{w}$, $C[\overline{w}]$, measures the relative standard error of $\overline{w}$ as an estimate of $p_\theta(\mathbf{y})$. We have

$$C[\overline{w}] = \frac{1}{\sqrt{m}} C[w(\mathbf{y}, \mathbf{z}^*)] = \frac{1}{\sqrt{m}} \frac{\sqrt{\mathrm{Var}_{p^*_{\theta_0}}[w(\mathbf{y}, \mathbf{z}^*)]}}{E_{p^*_{\theta_0}}[w(\mathbf{y}, \mathbf{z}^*)]} = \frac{1}{\sqrt{m}} \frac{\sqrt{\mathrm{Var}_{p^*_{\theta_0}}[w(\mathbf{y}, \mathbf{z}^*)]}}{p_\theta(\mathbf{y})}.$$

Its sample estimate is

$$\hat{C}[\overline{w}] = \frac{1}{\sqrt{m}} \hat{C}[w(\mathbf{y}, \mathbf{z}^*)] = \frac{1}{\sqrt{m}} \frac{s_w}{\overline{w}},$$

where s_w denotes the sample standard deviation of the $w(j)$'s. For $C[\overline{w}]$ to be some desirable value δ, m, the number of imputations needs to be

$$\frac{1}{\delta^2} \times (C[w(\mathbf{y}, \mathbf{z}^*)])^2.$$

For example, suppose that it is desired that $C[\overline{w}]$ be approximately 0.1. Based on a sample, this implies that m, the number of imputations needs to be about

$$100 \times \frac{s_w^2}{\overline{w}^2}.$$

In general, as a rule of thumb, the number of imputations should be at least 500, and not smaller than

$$25 \times (\hat{C}[w(\mathbf{y}, \mathbf{z}^*)])^2 = 25 \times \frac{s_w^2}{\overline{w}^2}. \tag{2.15}$$

This is not just because it is felt that $C[\overline{w}]$ should be less than 0.2. Since the distribution of $w(\mathbf{y}, \mathbf{z}^*)$ can be highly skewed, if m is less than (2.15), then the sample coefficient of variation $\hat{C}[w(\mathbf{y}, \mathbf{z}^*)]$ is not a relable estimate of the actual coefficient of variation $C[w(\mathbf{y}, \mathbf{z}^*)]$, and as a result, the standard error estimate of $\overline{w}$ can be badly off.

As demonstrated, the efficiency of sequential imputation is inversely proportional to $(C[w(\mathbf{y}, \mathbf{z}^*)])^2$. Note that $C[w(\mathbf{y}, \mathbf{z}^*)]$ depends on the trial distribution from which $\mathbf{z}^*$'s are drawn, which in turn depends on the decompositions of $\mathbf{y}$ and $\mathbf{z}$ used for sequential imputations. From (2.5)

$$w(\mathbf{y}, \mathbf{z}^*) = p_\theta(\mathbf{y}) \frac{p_\theta(\mathbf{z}^*|\mathbf{y})}{p_\theta^*(\mathbf{z}^*|\mathbf{y})}. \tag{2.16}$$

In importance sampling, p^* is referred to as the *trial distribution*, the ratio

$$\frac{p_\theta(\mathbf{z}^*|\mathbf{y})}{p_\theta^*(\mathbf{z}^*|\mathbf{y})}$$

is called the importance sampling weight, so $w(\mathbf{y}, \mathbf{z}^*)$ is the importance sampling weight multiplied by the unknown constant $p_\theta(\mathbf{y})$. Note that

$$(C[w(\mathbf{y}, \mathbf{z}^*)])^2 = \mathrm{Var}_{p_\theta^*}\left[\frac{p_\theta(\mathbf{z}^*|\mathbf{y})}{p_\theta^*(\mathbf{z}^*|\mathbf{y})}\right]$$

as

$$\mathrm{Var}_{p_\theta^*}\left[\frac{p_\theta(\mathbf{z}^*|\mathbf{y})}{p_\theta^*(\mathbf{z}^*|\mathbf{y})}\right] \quad = \quad \mathrm{Var}_{p_\theta^*}\left[\frac{w(\mathbf{y}, \mathbf{z}^*)}{p_\theta(\mathbf{y})}\right]$$

$$= \frac{\mathrm{Var}_{p^*_\theta}[w(\mathbf{y},\mathbf{z}^*)]}{(p_\theta(\mathbf{y}))^2}$$
$$= (C[w(\mathbf{y},\mathbf{z}^*)])^2.$$

Thus $C[w(\mathbf{y},\mathbf{z}^*)]$ can be considered as a measure of *distance* between the actual conditional distribution of $\mathbf{z}$, $p_\theta(\mathbf{z}|\mathbf{y})$, and the trial distribution, $p^*_\theta(\mathbf{z}^*|\mathbf{y})$. To keep this distance small, it is desirable to have $p^*_\theta(\cdot|\mathbf{y})$ as close to $p_\theta(\cdot|\mathbf{y})$ as possible.

2.2.2 Different Data Decompositions

In the description of sequential imputation, a special decomposition of the observed data $\mathbf{y}$ and the missing data $\mathbf{z}$, i.e., y_t and z_t denote respectively the observed and missing data of a single locus t. To improve efficiency, it is necessary to consider other decompositions. Two important criteria for choosing an appropriate decomposition are:

(I) Steps **(A)** and **(B)** can be performed cheaply, in terms of both computing time and memory requirement.

(II) The coefficient of variation $C[w(\mathbf{y},\mathbf{z}^*)]$ is kept small.

Note that **(I)** and **(II)** are often conflicting criteria. For example, under the trivial decomposition $\mathbf{y} = \{y_1\}$ and $\mathbf{z} = \{z_1\}$, $p_\theta(\mathbf{z}|\mathbf{y}) = p^*_\theta(\mathbf{z}^*|\mathbf{y})$ and $w(\mathbf{y},\mathbf{z}^*) = p_\theta(\mathbf{y})$ with zero variation. But this requires peeling all the loci jointly, which is exactly what is to be avoided. So compromises are necessary. A number of modifications to the basic proposed procedure which will help reduce the variation of $w(\mathbf{y},\mathbf{z}^*)$ *without*

increasing difficulties in computation will be discussed. While these modifications are trivial mathematically, their effects may be drastic in practice.

2.2.3 Redefining $\mathbf{y}_1$ to Incorporate More Information

Note that $p(\mathbf{z}|\mathbf{y})$ can be written as $p(z_1|\mathbf{y})\prod_{t=2}^{n} p(z_t|\mathbf{y}, z_1, \ldots, z_{t-1})$. So drawing z_1^* from $p(z_1|\mathbf{y})$ is obviously preferable to drawing z_1^* from $p(z_1|y_1)$ *if* the former can be done cheaply. Unfortunately this is not the case. However, this suggests that when drawing z_1^*, as much information as possible should be conditioned on as long as it doesn't increase computational cost. For each locus and each parent-offspring pair, define an identity by descent (IBD) variable as the indicator of whether the allele inherited by the offspring came from the grandfather or the grandmother. Note that given $\mathbf{z}$, an IBD variable is known if and only if the parent in question is heterozygous at the locus in question. So redefine y_1 to include the observed data on the first locus processed plus the IBD variables of the other loci that can be deduced from the observed data. Conditioning on these IBD variables is easy to do and has virtually no effect on the computations needed to perform steps **(A)** and **(B)**, but can reduce the variation of the weights significantly.

Up to this point, apart from the deducible IBD's, it is assumed that y_1 consists of observed data on a single locus. This is not necessary and it can be preferable to incorporate more than one locus into y_1. Note that the first step of sequential imputation involves computing $p_\theta(y_1)$ and drawing $z_1^*(j)$, $j = 1, \ldots, m$ from $p_\theta(z_1|y_1)$. This requires peeling the loci incorporated in y_1 jointly, but the key is that only a single

peel is required for all m imputations. However after this first step, computations for the m imputations are done separately. As long as the amount of computing time and memory required to perform this first peel are within acceptable limits, as many loci as possible should be incorporated into y_1. This will decrease the coefficient of variation of the weights and as a consequence can reduce the overall computing time by lowering the number of imputations required.

2.2.4 Imputing as Little as Possible

The z_t's are imputed only to help simplify computations. In the original description of sequential imputation, $\mathbf{z}$ includes every loci and every member in the pedigree. In some cases, some members of the pedigree may be typed for some, but not all of the loci. For a particular person and locus, call the missing data ignorable if the person is not typed for that locus nor are any of that person's descendents. Since there is absolutely no information in these ignorable data, imputing will only add noise and inflate the variance of the weights. Hence, for each $t, t = 1, \ldots, n$, redefine z_t to include only non-ignorable data. This redefinition does not make steps **(A)** and **(B)** any more difficult. In fact, the amount of computation needed for processing a locus with some ignorable data can be reduced. But more importantly, this redefinition can drastically reduce the variance of the importance weights.

2.2.5 Imputation Order

The order that the loci are processed affects the trial distribution p^* and hence the variance, but not the mean, of the weights $w(\mathbf{y}, \mathbf{z}^*)$. Thus an optimal order is one that minimizes the variance of the weights. Two guidelines for choosing an optimal, or near optimal, processing order are:

(a) Loci which have the least amount of missing information among the non-ignorable data should be processed first. So loci with more untyped individuals who are not ignorable should be processed late. For two loci which have the same individuals typed, the one with more alleles, and hence usually more informative, should be processed first.

(b) It is preferable to have the processed loci *physically* contiguous at any time t. For example, if there are five loci physically ordered (A, B, C, D, E), the processing orders (A, B, C, D, E), (E, D, C, B, A), and (C, B, D, A, E) are preferred over orderings such as (A, E, B, D, C).

Rule **(a)** takes precedence over rule **(b)**. It usually not too difficult to rank marker loci based on informativeness. However it can be more complicated when a disease locus is involved.

2.3 Computing Location Scores of a Disease Locus

Location scores for a disease gene relative to a number of marker loci with known locations can be estimated by a simple strategy. Set y_n to be the observed disease

data and process the markers first based on criteria **(a)** and **(b)**. The average of the weights before processing the disease data,

$$\overline{w_{n-1}} = \frac{1}{m}\sum_{j=1}^{m} w_{n-1}(j),$$

is an unbiased estimate of $p(y_1, \ldots, p_{n-1})$, the likelihood of the marker data. Hence

$$\hat{p}_{\infty}(\mathbf{y}) = \overline{w_{n-1}} \times p(y_n)$$

is an unbiased estimate of the likelihood for the scenario that the disease locus is unlinked to the markers. Then process the disease locus at various locations linked to the marker loci. This strategy has the advantage that only one set of marker imputations can be used to compute the likelihoods of all locations (Lange and Sobel 1991). Moreover, since the likelihood estimates at different locations result from a single simulation run, they tend to be positively correlated. As a consequence, standard error estimates of likelihood ratios among different locations are lower. In addition, this is the correct approach for computing the likelihood in the unlinked situation. It should be noted that recent experience suggests that for some data sets it may be more efficient to process the disease locus first, possibly jointly with one or two markers. This alternative strategy, even though it requires multiple simulation runs for the different disease gene locations, is preferred when the disease status is available for many individuals in the upper generations while marker genotypes of the same individuals are missing. This can occur, for example, when a highly penetrant disease can be diagnosed in three or more generations at the top of the pedigree for whom marker data is unavailable. Also, if the disease allele is very rare in the

population, disease genotypes of many individuals in the upper generations often can be deduced with little uncertainty.

Regardless of whether the disease locus is processed first or last, it is usually enough to apply sequential imputation to a single location, probably in the middle, within each interval spanned by two physically adjacent markers. Equation (2.14) can then be applied to approximate the likelihoods for other locations in the interval.

Assume that the markers define K intervals and that the disease is processed at positions $\theta_1, \ldots, \theta_K$, one location per interval. For the kth interval, let the weights be

$$w^k(j) = w_{n-1}(j) p_{\theta_k}(y_n | y_1, z_1^*(j), \ldots, y_{n-1}, z_{n-1}^*(j)),$$

the missing data be

$$\mathbf{z}_k^*(j) = (z_1^*(j), \ldots, z_{n-1}^*(j), z_{n,k}^*(j)),$$

where $z_{n,k}^*(j)$ is drawn from

$$p_{\theta_k}(z_{n,k} | y_1, z_1^*(j), \ldots, y_{n-1}, z_{n-1}^*(j), y_n)$$

For a position d Morgans from a fixed point, located in interval k, the likelihood $p_d(\mathbf{y})$ can be estimated by

$$\hat{p}_d(\mathbf{y}) = \frac{1}{m} \sum_{j=1}^{m} \frac{p_d(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_k}(\mathbf{y}, \mathbf{z}^*(j))} w^k(j).$$

Then the location score at position d can be estimated by

$$\hat{l}(d) = \log_{10} \frac{\hat{p}_d(\mathbf{y})}{\hat{p}_\infty(\mathbf{y})} = \log_{10} \hat{p}_d(\mathbf{y}) - \log_{10} \hat{p}_\infty(\mathbf{y}). \tag{2.17}$$

Then by the delta method, the variance of $\hat{l}(d)$ is approximately

$$\mathrm{Var}(\hat{l}(d)) = \frac{1}{m}(c(d)^2 + c(\infty)^2 - 2\rho(d)c(d)c(\infty)) \tag{2.18}$$

where

$$c(d) = C\left[\frac{p_d(\mathbf{y},\mathbf{z}^*)}{p_{\theta_k}(\mathbf{y},\mathbf{z}^*)}w^k(\mathbf{y},\mathbf{z}^*)\right],$$

$$\begin{aligned} c(\infty) &= C[w_{n-1}(\mathbf{y},\mathbf{z}^*)p(y_n)] \\ &= C[w_{n-1}(\mathbf{y},\mathbf{z}^*)], \end{aligned}$$

and

$$\rho(d) = \mathrm{Cor}\left(\frac{p_d(\mathbf{y},\mathbf{z}^*)}{p_{\theta_k}(\mathbf{y},\mathbf{z}^*)}w^k(\mathbf{y},\mathbf{z}^*), w_{n-1}(\mathbf{y},\mathbf{z}^*)p(y_n)\right).$$

$\mathrm{Var}(\hat{l}(d))$ can be estimated by

$$se(d)^2 = \frac{1}{m}(\hat{c}(d)^2 + \hat{c}(\infty)^2 - 2\hat{\rho}(d)\hat{c}(d)\hat{c}(\infty))$$

where

$$\hat{c}(d) = \hat{C}\left[\frac{p_d(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_k}(\mathbf{y},\mathbf{z}^*(j))}w^k(j)\right],$$

$$\hat{c}(\infty) = \hat{C}[w_{n-1}((j))p(y_n)],$$

and

$$\hat{\rho}(d) = \mathrm{Cor}\left(\frac{p_d(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_k}(\mathbf{y},\mathbf{z}^*(j))}w^k(j), w_{n-1}(j)p(y_n)\right).$$

If there are locations that have standard errors that are unacceptably large, processing the disease locus first, possibly with one or two close by markers, for that assumed location of the disease locus, can be contemplated.

2.4 Combining Runs

One possible problem with the above procedure is that the standard errors of location scores for positions near or on markers can be much larger than for locations away

from the markers. In addition to the possible solution of processing the disease locus first, the method of the previous section can be modified to allow for more efficient estimation of the likelihood function. As in the previous section, start by processing the markers first. However, instead of processing the information on the disease locus at one location between physically adjacent markers, process the disease information at multiple locations within the marker interval. Then it is possible to combine the information from these runs to give a more efficient estimate of the locations scores. Though it is possible to combine the information from three or more simulation locations, for most situations using only two locations is neccessary and that will be the situation discussed here. Note that

$$\begin{aligned} p^c_\theta(\mathbf{y}, a) &= \int \frac{p_{\theta_1}(\mathbf{z}|\mathbf{y})}{p^*_{\theta_1}(\mathbf{z}|\mathbf{y})} \frac{p_{\theta_1}(\mathbf{y})}{\tilde{p}_{\theta_1}(\mathbf{y})} \frac{p_\theta(\mathbf{y},\mathbf{z})}{\frac{p_{\theta_1}(\mathbf{y})}{\tilde{p}_{\theta_1}(\mathbf{y})} p_{\theta_1}(\mathbf{z}|\mathbf{y}) + a\frac{p_{\theta_2}(\mathbf{y})}{\tilde{p}_{\theta_2}(\mathbf{y})} p_{\theta_2}(\mathbf{z}|\mathbf{y})} p^*_{\theta_1}(\mathbf{z}|\mathbf{y}) d\mathbf{z} \\ &\quad + a \int \frac{p_{\theta_2}(\mathbf{z}|\mathbf{y})}{p^*_{\theta_2}(\mathbf{z}|\mathbf{y})} \frac{p_{\theta_2}(\mathbf{y})}{\tilde{p}_{\theta_2}(\mathbf{y})} \frac{p_\theta(\mathbf{y},\mathbf{z})}{\frac{p_{\theta_1}(\mathbf{y})}{\tilde{p}_{\theta_1}(\mathbf{y})} p_{\theta_1}(\mathbf{z}|\mathbf{y}) + a\frac{p_{\theta_2}(\mathbf{y})}{\tilde{p}_{\theta_2}(\mathbf{y})} p_{\theta_2}(\mathbf{z}|\mathbf{y})} p^*_{\theta_2}(\mathbf{z}|\mathbf{y}) d\mathbf{z} \\ &= \int p_\theta(\mathbf{y},\mathbf{z}) d\mathbf{z} \\ &= p_\theta(\mathbf{y}) \end{aligned}$$

where $\tilde{p}_{\theta_1}(\mathbf{y})$ is an estimate of $p_{\theta_1}(\mathbf{y})$ and $\tilde{p}_{\theta_2}(\mathbf{y})$ is an estimate of $p_{\theta_2}(\mathbf{y})$.

Now let

$$\begin{aligned} w_{\theta_1}(\mathbf{y},\mathbf{z}) &= \frac{p_{\theta_1}(\mathbf{y},\mathbf{z})}{p^*_{\theta_1}(\mathbf{z}|\mathbf{y})} \\ w_{\theta_2}(\mathbf{y},\mathbf{z}) &= \frac{p_{\theta_2}(\mathbf{y},\mathbf{z})}{p^*_{\theta_2}(\mathbf{z}|\mathbf{y})} \end{aligned}$$

Thus $p^c_\theta(\mathbf{y}, a)$ can be rewritten as

$$\begin{aligned} p^c_\theta(\mathbf{y}, a) &= \int w_{\theta_1}(\mathbf{y}, \mathbf{z}) \frac{p_\theta(\mathbf{y}, \mathbf{z})}{p_{\theta_1}(\mathbf{y}, \mathbf{z}) + a\frac{\tilde{p}_{\theta_1}(\mathbf{y})}{\tilde{p}_{\theta_2}(\mathbf{y})} p_{\theta_2}(\mathbf{y}, \mathbf{z})} p^*_{\theta_1}(\mathbf{z}|\mathbf{y}) d\mathbf{z} \\ &\quad + a \int w_{\theta_2}(\mathbf{y}, \mathbf{z}) \frac{p_\theta(\mathbf{y}, \mathbf{z})}{\frac{\tilde{p}_{\theta_2}(\mathbf{y})}{\tilde{p}_{\theta_1}(\mathbf{y})} p_{\theta_1}(\mathbf{y}, \mathbf{z}) + a p_{\theta_2}(\mathbf{y}, \mathbf{z})} p^*_{\theta_2}(\mathbf{z}|\mathbf{y}) d\mathbf{z} \end{aligned}$$

By the method of the previous section, let $w_1(1), \ldots, w_1(m)$ and $\mathbf{z}^*_1(1), \ldots, \mathbf{z}^*_1(m)$ and $w_2(1), \ldots, w_2(m)$ and $\mathbf{z}^*_2(1), \ldots, \mathbf{z}^*_2(m)$ be the results of sequential imputation runs done at parameter values θ_1 and θ_2 respectively. Then let

$$\begin{aligned} \overline{w}_1 &= \frac{1}{m} \sum_{j=1}^m w_1(j) \\ \overline{w}_2 &= \frac{1}{m} \sum_{j=1}^m w_2(j) \end{aligned}$$

In the earlier argument, the choice of $\tilde{p}_{\theta_1}(\mathbf{y})$ and $\tilde{p}_{\theta_2}(\mathbf{y})$ has been left arbitrary. One logical choice for these quantities is to set them to $\overline{w}_1$ and $\overline{w}_2$ respectively. Thus

$$\begin{aligned} \hat{p}^c_\theta(\mathbf{y}, a) &= \sum_{j=1}^m w_1(j) \frac{p_\theta(\mathbf{y}, \mathbf{z}^*_1(j))}{p_{\theta_1}(\mathbf{y}, \mathbf{z}^*_1(j)) + a\frac{\overline{w}_1}{\overline{w}_2} p_{\theta_2}(\mathbf{y}, \mathbf{z}^*_1(j))} \\ &\quad + a \sum_{j=1}^m w_2(j) \frac{p_\theta(\mathbf{y}, \mathbf{z}^*_2(j))}{\frac{\overline{w}_2}{\overline{w}_1} p_{\theta_1}(\mathbf{y}, \mathbf{z}^*_2(j)) + a p_{\theta_2}(\mathbf{y}, \mathbf{z}^*_2(j))} \end{aligned} \tag{2.19}$$

is an unbiased estimate of $p_\theta(\mathbf{y})$. The parameter a is a tuning paramter which determines which sample should have more weight in estimating $p_\theta(\mathbf{y})$. It should be chosen to try an minimize the variance of $\hat{p}^c_\theta(\mathbf{y}, a)$ or probably more importantly, to minimize the variance of $\log \hat{p}^c_\theta(\mathbf{y}, a) - \log \hat{p}_\infty(\mathbf{y})$, the estimated location score at θ.

This extension to the procedure of the previous section is most useful for estimating location scores near markers. This method suggests that in addition to processing the disease in the middle of the marker intervals, they should be also processed directly on top of the markers.

Chapter 3

Applications

3.1 Background

In addition to the estimation of likelihoods and location scores discussed in the previous chapter, the results of a sequential imputation run can be used for other applications. These applications are all based on equation (2.13), the importance sampling estimate of a conditional expectation. Three useful applications will be discussed in this chapter.

The first application to be discussed is a Monte Carlo EM procedure for parameter estimation. While direct calculations of the likelihood function are useful when only one or two parameters need to be estimated, such as estimating a disease gene's location given a fixed marker map as discussed in the previous chapter, this usually won't be feasible when a moderate or large number of parameters need to be jointly estimated. As shown by Wei and Tanner (1990) and Guo and Thompson (1992), Monte Carlo EM can be a useful estimation tool. As the major difference between

the sequential imputation, Wei and Tanner, and Guo and Thompson approaches is the method of simulating the missing data, each method has situations where it is preferable to the others. However for the type of linkage problems discussed earlier, the sequential imputation approach has advantages over the other two. In the sequential imputation approach, the missing data is simulated once and is reused for each iteration of the procedure, whereas in the other two approaches, the missing data need to be simulated each time the parameter estimate is updated. This can be a big advantage in the genetic setting, as simulation of new data sets can be extremely costly. Also to be discussed is a hybrid form of Monte Carlo EM. Instead of directly calculating conditional expectations as in the original EM procedure, or averaging over multiple sets of complete simulated data as in Monte Carlo EM, this hybrid procedure combines the two procedures. In this hybrid procedure only part of the desired missing data is simulated, the rest is integrated over. Assuming the necessary calculations can be done, this hybrid procedure should be more efficient.

The next application discussed will be the calculation of standard error estimates. Often in the analysis of pedigree data, parameter estimates are reported without associated standard errors. In particular, estimates of genetic distances in marker maps are usually given without any measure of uncertainty. Sequential imputation, as well as other Monte Carlo approaches can be used to help solve this problem. The multiple complete data set generated by sequential imputation can be used to estimate the information matrix at the maximum likelihood estimate in a similar fashion to the approaches by Wei and Tanner (1990) and Guo and Thompson (1992).

This estimate of the information matrix can then be inverted to give an estimate of the variance - covariance matrix of the maximum likelihood estimate.

The last application that will be discussed will be the application of sequential imputation to bayesian analysis. Under carefully chosen priors, the multiple data sets generated by sequential imputation can be used to approximate posterior distributions and their moments. These calculations are also useful in examining features of the likelihood surface, which can be particularly useful when approximate normality may not hold for the maximum likelihood estimates.

The problem of simultaneous estimation of marker recombination fractions will be used to illustrate the three sets of procedures.

3.2 Monte Carlo EM

For mapping marker loci, Lander and Green(1987) proposed using the EM algorithm (Dempster, Laird, and Rubin, 1977) to find maximum likelihood estimates of the recombination probabilities. When it is computationally infeasible to perform the E-step of the algorithm exactly, sequential imputation can be used for implementing a Monte Carlo version of the EM algorithm (MCEM). As in the original EM algorithm, the observed data is augmented by latent data such that it is easy to calculate and maximize the log-likelihood of the complete data. Assuming that the complete data generated by sequential imputation has this property, a modified version of the MCEM algorithm as described by Wei and Tanner (1990) can be easily implemented (also see Guo and Thompson, 1992 for another genetic application). Let $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$

and $w(1), \ldots, w(m)$ be the results of a sequential imputation run at parameter value θ_0. Let the initial value for the MCEM procedure be $\theta^{(0)}$. Usually $\theta^{(0)}$ should be set to θ_0, though this is not necessary. Then the MCEM procedure consists of interating the following three steps.

(1) Update the weights for iteration $i+1$:

$$w^{(i+1)}(j) = w^{(i)}(j)\frac{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta^{(i-1)}}(\mathbf{y}, \mathbf{z}^*(j))} = w(j)\frac{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))}.$$

(2) MC E-step

$$\begin{aligned}\hat{Q}_{i+1}(\theta, \theta^{(i)}) &= \frac{1}{m}\sum_{j=1}^{m}\frac{w^{(i+1)}(j)}{\overline{w^{(i+1)}}}\log p_\theta(\mathbf{y}, \mathbf{z}^*(j)) \\ &= \sum_{j=1}^{m}\frac{w^{(i+1)}(j)}{W^{(i+1)}}\log p_\theta(\mathbf{y}, \mathbf{z}^*(j))\end{aligned}$$

where

$$W^{(i+1)} = \sum_{j=1}^{m} w^{i+1}(j),\ \overline{w^{(i+1)}} = W^{(i+1)}/m.$$

Note that $\hat{Q}_{i+1}(\theta, \theta^{(i)})$ is a Monte Carlo estimate of $E_{\theta^{(i)}}[\log p_\theta(\mathbf{y}, \mathbf{z})|\mathbf{y}]$ by (2.13) and that $\overline{w^{(i+1)}} = \hat{p}_{\theta^{(i)}}(\mathbf{y})$ by (2.14).

(3) M-step

Set $\theta^{(i+1)}$ to be the maximizer of $\hat{Q}_{i+1}(\theta, \theta^{(i)})$.

Then the MCEM estimate is $\hat{\theta}$ where

$$\hat{\theta} = \lim_{i\to\infty}\theta^{(i)}$$

If the sequence $\theta^{(i)}, i = 1, 2, 3, \ldots$ get too far away from θ_0, the weights $w^{(i)}(j)$ may become poorly behaved. If at iteration $i+1$ the coefficient of variation of the weights

gets larger than desired, a new sequential imputation run with parameter value θ_0 set to $\theta^{(i)}$ should be run, possibly with more imputations.

This version of MCEM has two optimality properties similar to EM: the Monte Carlo log likelihood increases with each step and the procedure converges to a stationary point of the Monte Carlo log likelihood surface. The following two theorems describe these results more precisely.

Theorem 3.1 *Let $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$ be the results of a sequential imputation run at a parameter value θ_0. Let $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ be the sequence of maximizers of $\hat{Q}_{i+1}(\theta, \theta^{(i)})$. Then for all i,*

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) \geq \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).$$

Proof

$$\begin{aligned}
&\hat{Q}_{i+1}(\theta^{(i+1)}, \theta^{(i)}) - \hat{Q}_{i+1}(\theta^{(i)}, \theta^{(i)}) \\
&\quad = \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} (\log p_{\theta^{(i+1)}}(\mathbf{y}, \mathbf{z}^*(j)) - \log p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))) \\
&\quad = \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \log \frac{p_{\theta^{(i+1)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))} \\
&\quad \leq \log \left(\sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \frac{p_{\theta^{(i+1)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))} \right) \quad \text{by Jensen's Inequality} \\
&\quad = \log \left(\frac{1}{m \hat{p}_{\theta^{(i)}}(\mathbf{y})} \sum_{j=1}^{m} w(j) \frac{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} \frac{p_{\theta^{(i+1)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))} \right) \\
&\quad = \log \left(\frac{1}{m \hat{p}_{\theta^{(i)}}(\mathbf{y})} \sum_{j=1}^{m} w(j) \frac{p_{\theta^{(i+1)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} \right) \\
&\quad = \log \frac{\hat{p}_{\theta^{(i+1)}}(\mathbf{y})}{\hat{p}_{\theta^{(i)}}(\mathbf{y})} \\
&\quad = \log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) - \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).
\end{aligned}$$

Thus by the definition of $\theta^{(i)}$

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) - \log \hat{p}_{\theta^{(i)}}(\mathbf{y}) \geq \hat{Q}_{i+1}(\theta^{(i+1)}, \theta^{(i)}) - \hat{Q}_{i+1}(\theta^{(i)}, \theta^{(i)}) \geq 0.$$

Therefore

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) \geq \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).$$

□

Theorem 3.2 *Let* $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ *and* $w(1), \ldots, w(m)$ *be the results of a sequential imputation run at a parameter value* θ_0. *Let* $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ *be the sequence of maximizers of* $\hat{Q}_{i+1}(\theta, \theta^{(i)})$. *Then*

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) > \log \hat{p}_{\theta^{(i)}}(\mathbf{y}) \tag{3.1}$$

if

$$\theta^{(i)} \notin \Gamma = \left\{ \theta : \frac{D \log \hat{p}_\theta(\mathbf{y})}{D\theta} = 0 \right\}$$

Proof Let

$$\begin{aligned}
h(\theta) &= \frac{D \log \hat{p}_\theta(\mathbf{y})}{D\theta} \\
&= \frac{1}{\hat{p}_\theta(\mathbf{y})} \frac{D}{D\theta} \hat{p}_\theta(\mathbf{y}) \\
&= \frac{1}{m\hat{p}_\theta(\mathbf{y})} \sum_{j=1}^{m} w(j) \frac{\frac{D}{D\theta} p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} \\
&= \frac{1}{m\hat{p}_\theta(\mathbf{y})} \sum_{j=1}^{m} w(j) \frac{p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} \frac{D \log p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{D\theta}
\end{aligned}$$

and

$$\begin{aligned} k_{i+1}(\theta, \theta^{(i)}) &= \frac{D\hat{Q}_{i+1}(\theta, \theta^{(i)})}{D\theta} \\ &= \frac{1}{m\hat{p}_{\theta^{(i)}}(\mathbf{y})} \sum_{j=1}^{m} w(j) \frac{p_{\theta^{(i)}}(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} \frac{D \log p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{D\theta}. \end{aligned}$$

Thus

$$h(\theta^{(i)}) = k_{i+1}(\theta^{(i)}, \theta^{(i)}).$$

Then if $\theta^{(i)} \notin \Gamma$, $\theta^{(i)}$ is not a stationary point of $\hat{Q}_{i+1}(\theta, \theta^{(i)})$ so

$$\hat{Q}_{i+1}(\theta^{(i+1)}, \theta^{(i)}) > \hat{Q}_{i+1}(\theta^{(i)}, \theta^{(i)})$$

and

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) > \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).$$

□

In some cases, the complete data generated by sequential imputation isn't adequate for easy calculation or maximization of the log-likelihood of the complete data. Assume however, that the complete data $(\mathbf{y}, \mathbf{z})$ can be further augmented by missing data $\mathbf{x}$, such that:

(i) $E_{\theta_1}[\log p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z})|\mathbf{y}, \mathbf{z}]$ is easy to calculate

(ii) $\sum c_j E_{\theta_1}[\log p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z})|\mathbf{y}, \mathbf{z}]$ is easy to maximize,

where θ and θ_1 are two parameter values and the c_j's are non-negative numbers. This can occur when jointly estimating the recombination fractions between a set of markers under the assumptions of no interference. Assuming the complete data $(\mathbf{y}, \mathbf{z})$

is the set of haplotypes for each person in the pedigree, the complete data used in Chapter 2, steps **(2)** and **(3)** are both very difficult. However by augmenting $(\mathbf{y}, \mathbf{z})$ by the number of recombinations between pairs of adjacent markers, the calculations become easy to do. The hybrid procedure modifies steps **(2)** and **(3)** of the previous MCEM procedure as follows

(2*) MC E-step

$$\begin{aligned}\tilde{Q}_{i+1}(\theta, \theta^{(i)}) &= \frac{1}{m}\sum_{j=1}^{m}\frac{w^{(i+1)}(j)}{\overline{w^{(i+1)}}}E_{\theta^{(i)}}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)] \\ &= \sum_{j=1}^{m}\frac{w^{(i+1)}(j)}{W^{(i+1)}}E_{\theta^{(i)}}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)]\end{aligned}$$

where

$$W^{(i+1)} = \sum_{j=1}^{m} w^{i+1}(j),\ \overline{w^{(i+1)}} = W^{(i+1)}/m.$$

Note that $\tilde{Q}_{i+1}(\theta, \theta^{(i)})$ is a Monte Carlo estimate of $E_{\theta^{(i)}}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{y}]$ by (2.13) and that $\overline{w^{(i+1)}} = \hat{p}_{\theta^{(i)}}(\mathbf{y})$ by (2.14).

(3*) M-step

Set $\theta^{(i+1)}$ to be the maximizer of $\tilde{Q}_{i+1}(\theta, \theta^{(i)})$.

This hybrid procedure has similar properties to the originally described procedure. In particular, there are analogues to both of the previous theorems for this procedure.

Theorem 3.3 *Let $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$ be the results of a sequential imputation run at a parameter value θ_0. Let $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ be the sequence of maximizers of $\tilde{Q}_{i+1}(\theta, \theta^{(i)})$. Then for all i,*

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) \geq \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).$$

Proof $\tilde{Q}_{i+1}(\theta^{(i+1)},\theta^{(i)}) - \tilde{Q}_{i+1}(\theta^{(i)},\theta^{(i)})$

$$
\begin{aligned}
&= \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} (E_{\theta^{(i)}}[\log p_{\theta^{(i+1)}}(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)] \\
&\qquad - E_{\theta^{(i)}}[\log p_{\theta^{(i)}}(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)]) \\
&= \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} E_{\theta^{(i)}} \left[\log \frac{p_{\theta^{(i+1)}}(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))}{p_{\theta^{(i)}}(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))} |\mathbf{y},\mathbf{z}^*(j) \right] \\
&\leq \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \log E_{\theta^{(i)}} \left[\frac{p_{\theta^{(i+1)}}(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))}{p_{\theta^{(i)}}(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))} |\mathbf{y},\mathbf{z}^*(j) \right] \\
&\qquad\qquad \text{by Jensens's Inequality} \\
&= \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \log \frac{p_{\theta^{(i+1)}}(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta^{(i)}}(\mathbf{y},\mathbf{z}^*(j))} \\
&= \hat{Q}_{i+1}(\theta^{(i+1)},\theta^{(i)}) - \hat{Q}_{i+1}(\theta^{(i)},\theta^{(i)}) \\
&\leq \log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) - \log \hat{p}_{\theta^{(i)}}(\mathbf{y}) \\
&\qquad\qquad \text{by theorem 3.1.}
\end{aligned}
$$

Thus by the definition of $\theta^{(i)}$

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) \geq \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).$$

□

Before stating the analogue to theorem 3.2 a lemma (Serfling, 1980) is needed.

Lemma 3.1 *Assume that for each θ that the derivative*

$$\frac{D \log p_\theta(\mathbf{y})}{D\theta}$$

exists for all $\mathbf{y}$ *and that for each θ_0, there exists a function $g(\mathbf{y})$ (possibly depending on θ_0) such that for θ in a neighbourhood $N(\theta_0)$,*

$$\left| \frac{D p_\theta(\mathbf{y})}{D\theta} \right| \leq g(\mathbf{y})$$

for all $\mathbf{y}$ *and*

$$\int g(\mathbf{y})d\mathbf{y} < \infty.$$

Now let

$$l(\theta, \theta_0) = \frac{D}{D\theta} E_{\theta_0}[\log p_\theta(\mathbf{y})].$$

Then

$$l(\theta_0, \theta_0) = 0$$

Theorem 3.4 *Let* $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ *and* $w(1), \ldots, w(m)$ *be the results of a sequential imputation run at a parameter value* θ_0. *Let* $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$ *be the sequence of maximizers of* $\tilde{Q}_{i+1}(\theta, \theta^{(i)})$. *Assuming that* $p_\theta(\mathbf{x}|\mathbf{y}, \mathbf{z})$ *satisfies the conditions of the previous lemma,*

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) > \log \hat{p}_{\theta^{(i)}}(\mathbf{y}) \tag{3.2}$$

if

$$\theta^{(i)} \notin \Gamma = \left\{ \theta : \frac{D \log \hat{p}_\theta(\mathbf{y})}{D\theta} = 0 \right\}$$

Proof As before, let

$$\begin{aligned} h(\theta) &= \frac{D \log \hat{p}_\theta(\mathbf{y})}{D\theta} \\ &= \frac{1}{\hat{p}_\theta(\mathbf{y})} \frac{D}{D\theta} \hat{p}_\theta(\mathbf{y}) \\ &= \frac{1}{m\hat{p}_\theta(\mathbf{y})} \sum_{j=1}^{m} w(j) \frac{p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} \frac{D \log p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{D\theta} \end{aligned}$$

Let

$$\tilde{k}_{i+1}(\theta, \theta^{(i)}) = \frac{D\tilde{Q}_{i+1}(\theta, \theta^{(i)})}{D\theta}$$

$$
\begin{aligned}
&= \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \frac{D}{D\theta} E_{\theta^{(i)}}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)] \\
&= \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \frac{D}{D\theta} \{\log p_\theta(\mathbf{y},\mathbf{z}^*(j)) + E_{\theta^{(i)}}[\log p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)]\} \\
&= \frac{1}{m\hat{p}_{\theta^{(i)}}} \sum_{j=1}^{m} w(j) \frac{p_{\theta^{(i)}}(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*(j))} \frac{D \log p_\theta(\mathbf{y},\mathbf{z}^*(j))}{D\theta} \\
&\quad + \frac{1}{m\hat{p}_{\theta^{(i)}}} \sum_{j=1}^{m} w(j) \frac{p_{\theta^{(i)}}(\mathbf{y},\mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y},\mathbf{z}^*(j))} \frac{D}{D\theta} E_{\theta^{(i)}}[\log p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)]
\end{aligned}
$$

Thus

$$\tilde{k}_{i+1}(\theta^{(i)},\theta^{(i)}) = h(\theta^{(i)})$$

since by lemma 3.1

$$\frac{D}{D\theta} E_{\theta^{(i)}}[\log p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}^*(j))|\mathbf{y},\mathbf{z}^*(j)]|_{\theta=\theta^{(i)}} = 0$$

for each j.

Then if $\theta^{(i)} \notin \Gamma, \theta^{(i)}$ is not a stationary point of $\tilde{Q}_{i+1}(\theta,\theta^{(i)})$ so

$$\tilde{Q}_{i+1}(\theta^{(i+1)},\theta^{(i)}) > \tilde{Q}_{i+1}(\theta^{(i)},\theta^{(i)})$$

and

$$\log \hat{p}_{\theta^{(i+1)}}(\mathbf{y}) > \log \hat{p}_{\theta^{(i)}}(\mathbf{y}).$$

□

Instead of modifying the original procedure by calculating conditional expectations involving the additional level of missing data, the additional level of missing data could have been imputed as well. However there are a couple of disadvantages in sampling the additional missing data. By simulating data on a larger space, the estimate

of $p_\theta(\mathbf{y})$ becomes less precise, assuming the number of imputations doesn't change. As the two forms of sequential imputation MCEM try to find the maximum of the Monte Carlo estimate of the likelihood function, the one that gives a more precise estimate should give a better approximation to the maximum likelihood estimate. In addition, calculating the conditional expectation of the log likelihood may be faster than drawing $\mathbf{x}$ and calculating the log likelihood.

It should be noted that the MCEM procedures discussed by Wei and Tanner and Guo and Thompson can be modified in a similar fashion as this hybrid procedure is based on the relationship

$$E_{\theta_1}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{y}] = E_{\theta_1}[E_{\theta_1}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{y},\mathbf{z}]|\mathbf{y}].$$

In general, it is necessary that the conditional expectation $E_{\theta_1}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{y},\mathbf{z}]$ is calculate, $\sum c_j E_{\theta_1}[\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z})|\mathbf{y},\mathbf{z}]$ is easy to maximize, and that it is easy to get a sample (possibly weighted) from the distribution $p_{\theta_1}(\mathbf{z}|\mathbf{y})$.

There is an important difference between the sequential imputation version of MCEM and those previously proposed. After each iteration of the EM algorithm, instead of reweighting the data as in step **(1)**, they generate completely new equally weighted sets of missing data simulated at $\theta^{(i+1)}$ for the next MC E-step. This can be highly inefficient since updating the weights will usually take much less time than imputing new missing data.

3.2.1 Estimating Marker Recombination Fractions

One possible use of the MCEM procedure is to estimate the distances between K markers in a pedigree assuming that the order of the K markers is known. If all the identity by descent (IBD) variables of the non-founders (people with parents in the pedigree) as described in section 2.2.3 are observed, the log likelihood function has the form

$$\log L(\theta) = \sum_{k=1}^{K-1} x_k \log \theta_k + (2n - x_k) \log(1 - \theta_k) \tag{3.3}$$

where θ_k is the recombination probability between markers k and $k+1$ and x_k is the number of recombinations observed between markers k and $k+1$ for the n non-founders. The number of recombinations between two loci is easy to calculate given the IBD's. Thus the maximum likelihood estimates in this case are

$$\hat{\theta}_k = \frac{x_k}{2n}. \tag{3.4}$$

However, not all of the IBD variables can be determined from the complete data $(\mathbf{y}, \mathbf{z})$ used for sequential imputation as it is not possible to determine the IBD's at a locus for children of a parent who is homozygous at that locus. The modified MCEM procedure described above can be used in this case as the conditional expectations of the number of recombinations for any θ_1

$$\tilde{x}_k(j) = E_{\theta_1}[x_k | \mathbf{y}, \mathbf{z}^*(j)] \tag{3.5}$$

are easy to calculate. These conditional expectations are exactly what is needed for the evaluation and maximization of $E_\theta[p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z}^*(j)) | \mathbf{y}, \mathbf{z}^*(j)]$. After the modifica-

tion, the sequence of MCEM estimates of θ_k is given by

$$\hat{\theta}_k^{(i+1)} = \sum_{j=1}^{m} \frac{w^{(i+1)}(j)}{W^{(i+1)}} \frac{\tilde{x}_k(j)}{2n}, \tag{3.6}$$

the weighted average of the estimates of the recombination fractions from each sequential imputation sample.

3.3 Standard Error Estimation

In addition to the parameter estimates, the information matrix can be estimated with the results of a sequential imputation run. As shown by Louis (1982), the observed information matrix $I(\theta)$ can be decomposed into:

$$\begin{aligned} I(\theta) &= -E_\theta \left[\frac{D^2 \log p_\theta(\mathbf{y}, \mathbf{z})}{D^2\theta} | \mathbf{y} \right] - E_\theta \left[\left(\frac{D \log p_\theta(\mathbf{y}, \mathbf{z}(j))}{D\theta} \right)^2 | \mathbf{y} \right] \\ &\quad + \left(E_\theta \left[\frac{D \log p_\theta(\mathbf{y}, \mathbf{z})}{D\theta} | \mathbf{y} \right] \right)^2. \end{aligned} \tag{3.7}$$

By (2.13) this can be estimated by

$$\begin{aligned} \hat{I}(\theta) &= -\sum_{j=1}^{m} \frac{w_\theta(j)}{W_\theta} \frac{D^2 \log p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{D^2\theta} - \sum_{j=1}^{m} \frac{w_\theta(j)}{W_\theta} \left(\frac{D \log p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{D\theta} \right)^2 \\ &\quad + \left(\sum_{j=1}^{m} \frac{w_\theta(j)}{W_\theta} \frac{D \log p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{D\theta} \right)^2 \end{aligned} \tag{3.8}$$

where

$$w_\theta(j) = w(j) \frac{p_\theta(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))}$$

and

$$W_\theta = \sum_{j=1}^{M} w_\theta(j) = \hat{p}_\theta(\mathbf{y})$$

If the modified version of MCEM is run, an estimate of the information matrix is

$$\begin{aligned}\hat{I}(\theta) &= -\sum_{j=1}^{m}\frac{w_\theta(j)}{W_\theta}E_\theta\left[\frac{D^2\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))}{D^2\theta}|\mathbf{y},\mathbf{z}^*(j)\right] \\ &\quad -\sum_{j=1}^{m}\frac{w_\theta(j)}{W_\theta}E_\theta\left[\left(\frac{D\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))}{D\theta}\right)^2|\mathbf{y},\mathbf{z}^*(j)\right] \\ &\quad +\left(\sum_{j=1}^{m}\frac{w_\theta(j)}{W_\theta}E_\theta\left[\frac{D\log p_\theta(\mathbf{x},\mathbf{y},\mathbf{z}^*(j))}{D\theta}|\mathbf{y},\mathbf{z}^*(j)\right]\right)^2. \qquad (3.9)\end{aligned}$$

Both of these estimates are similar to the estimate discussed by Wei and Tanner (1990). The only major difference is the adjustment required for the unequally weighted samples. If $\hat{I}$ is evaluated at the MCEM estimate, and that estimate is in the interior of the parameter space, the last term for both estimators is 0, and does not need to be evaluated.

The variance-covariance matrix of the parameter estimates can then be approximated by $\hat{I}^{-1}(\hat{\theta})$. This matrix can then be used to construct of confidence sets for the parameters. However if the maximum likelihood estimates are close to the boundary of the parameter space, for example, a recombination fraction estimate very close to zero, the inverse of the information matrix may not be appropriate for determining standard errors.

3.3.1 Marker Recombination Fractions

In the case of joint estimation of marker recombination, the information matrix is easy to calculate. Assuming that all components of $\hat{\theta}$, the MCEM estimate, are bounded away from 0, $\hat{I}(\hat{\theta})$ has the form:

$$\begin{aligned}
\hat{I}_{k,k}(\hat{\theta}) &= \sum_{j=1}^{m} \frac{w_{\hat{\theta}}(j)}{W_{\hat{\theta}}} \left\{ \frac{\hat{x}_k(j)}{\hat{\theta}_k^2} + \frac{2n - \hat{x}_k(j)}{(1-\hat{\theta}_k)^2} - \left(\frac{\hat{x}_k(j)}{\hat{\theta}_k} - \frac{2n - \hat{x}_k(j)}{1-\hat{\theta}_k} \right)^2 \right\} \\
\hat{I}_{k,l}(\hat{\theta}) &= \sum_{j=1}^{m} \frac{w_{\hat{\theta}}(j)}{W_{\hat{\theta}}} \left(\frac{\hat{x}_k(j)}{\hat{\theta}_k} - \frac{2n - \hat{x}_k(j)}{1-\hat{\theta}_k} \right) \left(\frac{\hat{x}_l(j)}{\hat{\theta}_l} - \frac{2n - \hat{x}_l(j)}{1-\hat{\theta}_l} \right)
\end{aligned}$$

where

$$\hat{x}_k(j) = E_{\hat{\theta}}[x_k | \mathbf{y}, \mathbf{z}^*(j)]$$

If the information matrix is to be evaluated at a θ different than $\hat{\theta}$ or $\hat{\theta}_k$ is 0 for some k, then the terms involving the expectation of the first derivative of $\log p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z})$ needs to be included in the calculation.

3.4 Bayesian Analysis

In addition to the previously described frequentist procedures, sequential imputation can also be used to implement Bayesian procedures. In particular, sequential imputation can be used for the approximation of posterior distributions and moments. This can be particularly useful for marker mapping as using the inverse of information matrix for obtaining standard errors and construction of confidence sets may be poorly behaved if some the the estimated recombination fractions are close to zero.

As before, let $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$ be the results of a sequential imputation run at a parameter value θ_0. Then as shown by Kong, Liu, and Wong (1991), for a given prior distribution $p(\theta)$, the posterior distribution $p(\theta|\mathbf{y})$ can be

approximated by the importance sampling estimate

$$\hat{p}(\theta|\mathbf{y}) = \sum_{j=1}^{m} \frac{w^*(j)}{W^*} p(\theta|\mathbf{y}, \mathbf{z}^*(j)) \tag{3.10}$$

where

$$w^*(\mathbf{y}, \mathbf{z}^*(j)) = w^*(j) = \frac{p(\mathbf{y}, \mathbf{z}^*(j))}{p_{\theta_0}(\mathbf{y}, \mathbf{z}^*(j))} w(\mathbf{y}, \mathbf{z}^*(j)),$$

$$W^* = \sum_{j=1}^{m} w^*(j),$$

and

$$p(\mathbf{y}, \mathbf{z}^*(j)) = \int p_\theta(\mathbf{y}, \mathbf{z}^*(j)) p(\theta) d\theta$$

For computational purposes, it is necessary that the prior allows for $p(\theta|\mathbf{y}, \mathbf{z}^*(j))$ and $p(\mathbf{y}, \mathbf{z}^*(j))$ to be evaluated easily. That often means that a conjugate prior needs to be used.

3.4.1 Marker Mapping

In the case of marker mapping, the posterior distribution $p(\theta|\mathbf{y}, \mathbf{z}^*(j))$ and the predictive distributions $p(\mathbf{y}, \mathbf{z}^*(j))$ will be difficult to calculate given the missing data generated by sequential imputation. However if the number of recombinations between each of the markers are simulated conditioned on $(\mathbf{y}, \mathbf{z}^*(j))$, the conditional and predictive distributions can be easy to calculate. Assume that K recombination fractions are to be determined where θ_k is the recombination fraction between markers k and $k+1$. Let $\mathbf{z}^*(1), \ldots, \mathbf{z}^*(m)$ and $w(1), \ldots, w(m)$ be the results of a sequential imputation run performed at θ_0. Let $\mathbf{x}^*(j) = (x_1^*(j), \ldots, x_K^*(j))$ be drawn from

$p_{\theta_0}(\mathbf{x}|\mathbf{y}, \mathbf{z}^*(j))$ where $x_k^*(j)$ denotes the number of recombinations between markers k and $k+1$ in the jth dataset. Assume that the prior distribution of θ is

$$p(\theta) = \prod_{k=1}^{K} \frac{1}{B_{0.5}(\alpha_k, \beta_k)} \theta_k^{\alpha_k - 1}(1-\theta_k)^{\beta_k - 1} I_{\{0 \le \theta_k \le 0.5\}}$$

where

$$B_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt$$

is the incomplete beta function. Thus the prior is a product of K truncated beta distributions. The support of θ_k is truncated at 0.5 as recombination fractions greater than 0.5 do not make sense when all markers are on the same chromosome and there is no interference.

Then for a realization $(\mathbf{x}, \mathbf{y}, \mathbf{z})$

$$p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \prod_{k=1}^{K} \binom{2n}{x_k} \theta_k^{x_k}(1-\theta_k)^{2n-x_k},$$

$$p(\theta|\mathbf{x}, \mathbf{y}, \mathbf{z}) = \prod_{k=1}^{K} \frac{1}{B_{0.5}(x_k + \alpha_k, 2n - x_k + \beta_k)} \theta_k^{x_k + \alpha_k - 1}(1-\theta_k)^{2n - x_k + \beta_k - 1} I_{\{0 \le \theta_k \le 0.5\}},$$

and

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto \prod_{k=1}^{K} \binom{2n}{x_k} \frac{B_{0.5}(x_k + \alpha_k, 2n - x_k + \beta_k)}{B_{0.5}(\alpha_k, \beta_k)}$$

So

$$w^*(j) = \left\{ \prod_{k=1}^{K} \binom{2n}{x_k^*(j)} \frac{B_{0.5}(x_k^*(j) + \alpha_k, 2n - x_k^*(j) + \beta_k)}{B_{0.5}(\alpha_k, \beta_k)} \right\} \frac{w(j)}{p_{\theta_0}(\mathbf{x}^*(j), \mathbf{y}, \mathbf{z}^*(j))}$$

Then an estimate of the marginal posterior density of θ_k given the independent truncated beta prior is

$$\hat{p}(\theta_k|\mathbf{y}) = \sum_{j=1}^{m} p(\theta_k|\mathbf{x}^*(j), \mathbf{y}, \mathbf{z}^*(j)) \frac{w^*(j)}{W^*} \tag{3.11}$$

which is a weighted average of truncated beta distributions.

In general, the moments of $\hat{p}(\theta_k|\mathbf{y})$ can be easily calculated based on equation (3.10). In particular,

$$\hat{E}[\theta_k|\mathbf{y}] = \sum_{j=1}^{m} E[\theta_k|\mathbf{y}, \mathbf{z}^*(j)]\frac{w^*(j)}{W^*}, \tag{3.12}$$

$$\hat{E}[\theta_k^2|\mathbf{y}] = \sum_{j=1}^{m} E[\theta_k|\mathbf{y}, \mathbf{z}^*(j)]\frac{w^*(j)}{W^*}, \tag{3.13}$$

and

$$\widehat{\mathrm{Var}}(\theta_k|\mathbf{y}) = \hat{E}[\theta_k^2|\mathbf{y}] - (\hat{E}[\theta_k|\mathbf{y}])^2. \tag{3.14}$$

For the marker recombinations case, based on equation (3.11), the first two posterior moments can be estimated by

$$\hat{E}[\theta_k|\mathbf{y}] = \sum_{j=1}^{m} \frac{B_{0.5}(x_k^*(j)+\alpha_k+1, 2n-x_k^*(j)+\beta_k)}{B_{0.5}(x_k^*(j)+\alpha_k, 2n-x_k^*(j)+\beta_k)}\frac{w^*(j)}{W^*}$$

and

$$\hat{E}[\theta_k^2|\mathbf{y}] = \sum_{j=1}^{m} \frac{B_{0.5}(x_k^*(j)+\alpha_k+2, 2n-x_k^*(j)+\beta_k)}{B_{0.5}(x_k^*(j)+\alpha_k, 2n-x_k^*(j)+\beta_k)}\frac{w^*(j)}{W^*}.$$

Then the posterior variance can be estimated by

$$\widehat{\mathrm{Var}}(\theta_k|\mathbf{y}) = \hat{E}[\theta_k^2|\mathbf{y}] - (\hat{E}[\theta_k|\mathbf{y}])^2.$$

If n is large (so $B_{0.5}(x_k^*(j)+\alpha_k, 2n-x_k^*(j)+\beta_k) \approx B_1(x_k^*(j)+\alpha_k, 2n-x_k^*(j)+\beta_k)$), the posterior means and variances are approximately

$$\begin{aligned}\hat{E}[\theta_k|\mathbf{y}] &\approx \sum_{j=1}^{m} \frac{\alpha_k + x_k^*(j)}{\alpha_k+\beta_k+2n}\frac{w^*(j)}{W^*} \\ &= \mu_k(\mathbf{y}),\end{aligned}$$

$$\widehat{\mathrm{Var}}(\theta_k|\mathbf{y}) \approx \sum_{j=1}^{m} \frac{(\alpha_k+x_k^*(j))(\alpha_k+x_k^*(j)+1)}{(\alpha_k+\beta_k+2n)((\alpha_k+\beta_k+2n+1)}\frac{w^*(j)}{W^*} - (\mu_k(\mathbf{y}))^2$$

Note that even though the recombination fractions are independent given the complete data and the independence prior, they generally will be correlated under the posterior distributions. The posterior distribution of θ_k and θ_l can be approximated by

$$\hat{p}(\theta_k, \theta_l) = \sum_{j=1}^{m} p(\theta_k|\mathbf{x}^*(j), \mathbf{y}, \mathbf{z}^*(j)) p(\theta_l|\mathbf{x}^*(j), \mathbf{y}, \mathbf{z}^*(j)) \frac{w^*(j)}{W^*}, \tag{3.15}$$

a mixture of products of independent truncated beta distributions. The posterior covariance between θ_k and θ_l can be approximated by

$$\begin{aligned} \widehat{\mathrm{Cov}}(\theta_k, \theta_l|\mathbf{y}) &= \sum_{j=1}^{m} \left\{ \frac{B_{0.5}(x_k^*(j) + \alpha_k + 1, 2n - x_k^*(j) + \beta_k)}{B_{0.5}(x_k^*(j) + \alpha_k, 2n - x_k^*(j) + \beta_k)} \right. \\ &\qquad \left. \times \frac{B_{0.5}(x_l^*(j) + \alpha_l + 1, 2n - x_l^*(j) + \beta_l)}{B_{0.5}(x_l^*(j) + \alpha_l, 2n - x_l^*(j) + \beta_l)} \right\} \frac{w^*(j)}{W^*} \\ &\qquad - \hat{E}[\theta_k|\mathbf{y}] \hat{E}[\theta_l|\mathbf{y}] \end{aligned}$$

By setting the prior to the uniform distribution ($\alpha_k = \beta_k = 1$ for all k), the posterior distribution as estimated by (3.10) is an estimate of the likelihood function. Note however, it is not the same estimate as (2.14), as it is based on simulating from a larger space of missing data. However by simulating the number of recombinations in each marker interval, it is easier to examine the properties of the likelihood surface by looking at the integrated likelihood surface. For example, equation (3.11) estimates the likelihood integrated over all recombination fractions but θ_k. Plotting this over a range of θ_k's can be used to examine the uncertainty in the estimation of θ_k, ignoring the other recombination fractions. Using (3.15) allows the joint relationship of estimates of θ_k and θ_l to be examined. In particular, it is possible to see whether or not a pair of parameters are being estimated approximately independently, or

whether they appear to be highly correlated. This can be done quite easily by examining contour plots of (3.15) over a range of θ_k and θ_l. Examining the likelihood surface this way can be particularly useful when one or more of the recombination fractions is estimated to be or close to zero, the boundary of the parameter space. When this happens, the approximate normality of the maximum likelihood estimates breaks down, implying the variance - covariance matrix, possibly estimated by the previously described procedure, does not accurately describe the uncertainty in the estimates of the recombination fractions.

Chapter 4

Examples

4.1 RW Pedigree Data Set

The RW pedigree shown in figure 4.1 segregating for Maturity Onset Diabetes of the Young (MODY) (Bell et. al., 1991), a form of non-insulin-dependent diabetes mellitus, is used to illustrate different properties of sequential imputation. The form of MODY segregating in this pedigree has been linked to markers on 20q(Bell et. al., 1991). Note that the diagnostic information summarized in Figure 1 and used in these analyses are derived from both the clinical diagnosis of MODY and from biochemical studies. Thus, some individuals who do not have clinical disease are considered affected in these analyses. Because recombination events in some individuals without clinical disease, but considered as affected by our biochemical criteria, may be critical in localizing the MODY gene, results of these analyses are, of course, dependent on the diagnostic assumptions made. These analyses are not presented to justify any particular diagnostic criteria or localization of the MODY locus within this region,

Figure 4.1 The RW Pedigree, showing inheritance of MODY. The affected members of the pedigree are shown with the solid symbols. The letters A and B show where the three branches of the pedigree join together.

Of interest here is the location of the MODY gene relative to eight markers on the long arm of chromosome 20. The markers of interest, with the number of alleles possible, are shown in table 4.1. The list is given in the order that they occur, from centromere to telomere.

In the initial analyses discussed in this chapter, the distances between the markers

Marker	Number of Alleles
ADA1	5
ADA2	2
L127	6
S22	3
S4	2
RM292	12
GPR	6
GSA	3

Table 4.1 Markers Examined and Number of Alleles

(denoted as the CEPH distances) are assumed to be those given in table 4.2. The localizations of ADA, L127, S22, S4, GPR, and GSA are based on the CEPH (Centre D'Etude du Polymorphisme Humain) families (NIH/CEPH Collaborative Mapping Group, 1992). The location of the RM292 locus was estimated from the RW pedigree alone given the locations and distances of the other seven markers. Other distances between the markers will be considered later in this chapter. They will be based on intermediate analyses on the RW pedigree.

In all of the analysis that follow, the disease penetrances and the disease and marker allele frequences are held fixed. The disease penetrance values are set to

$$\begin{aligned} p(Affected|dd) &= 0 \\ p(Affected|dD) &= 0.95 \end{aligned}$$

Marker 1	Marker 2	Recombination Fraction
ADA1	ADA2	0
ADA1/ADA2	L127	0.034
L127	S22	0.050
S22	S4	0.121
S4	RM292	0.011
RM292	GPR	0.111
GPR	GSA	0.132

Table 4.2 CEPH based distances between adjacent markers

$$p(Affected|DD) \quad = \quad 1$$

No age dependent penetrance was accounted in this analysis since the affection status was partly dependent on biochemical studies. Age dependent penetrance could have been very easily incorporated in this analysis if desired. The frequency of the abnormal disease allele in the general population is assumed to be 0.0001. The allele frequencies for the eight markers are given in table 4.3.

4.2 Three and Four Point Analyses

The eight markers on the long arm of chromosome 20 define a set of six intervals. To examine the accuracy of the procedure, sequential imputation runs for each of the intervals were done by the procedure discussed in section 2.3. In each of the intervals, except for one, the processing order was the marker with the more alleles,

Marker	Allele Frequencies
ADA1	0.29 0.24 0.22 0.17 0.08
ADA2	0.62 0.38
L127	0.34 0.30 0.20 0.11 0.03 0.02
S22	0.50 0.33 0.17
S4	0.60 0.40
RM292	0.14 0.12 0.12 0.11 0.09 0.09 0.09 0.07 0.07 0.04 0.04 0.02
GPR	0.76 0.10 0.08 0.02 0.02 0.02
GSA	0.72 0.25 0.03

Table 4.3 Marker allele frequencies, listed from most to least frequent.

the marker with less alleles, and finally the MODY disease data. The one exception was the ADA1/ADA2 - L127 interval. For this interval, the processing order was ADA1, L127, ADA2, and finally MODY. ADA1 was processed before L127 since it was typed in many more family members and earlier examination suggested that it was a more informative marker than L127 in this family. In each of the six intervals, $m = 2,000$ sets of imputations were done with MODY processed at four locations, unlinked to the markers, in the middle of the interval, and at both ends of the interval and calculations were done assuming the CEPH distances. As two of the intervals, S22 - S4 and GPR - GSA, had large coefficients of variation, additional runs were done for these intervals. For the GPR - GSA interval, the new run was done by switching the processing order of GPR and GSA and processing MODY last. However for the S22

- S4 interval, separate runs were done at the midpoint of the interval and at the ends of the interval with a processing order of MODY, S22, and then S4. For comparison exact calculations were done with the LINKMAP program of the LINKAGE package (Lathrop et. al. 1984). Within each interval, exact likelihoods were calculated for 6 equally spaced locations.

Figure 4.2 shows the estimated location scores for each of the intervals as calculated by (2.17). Also shown are the true location scores as calculated by LINKMAP. Except for right on top of the markers, the location scores were estimated based on the part of the run where MODY was processed in the middle. Right on top of the markers, the location scores were estimated with the runs done on top of the markers. As can be seen in the plot, the sequential imputation estimates are very close to the true values. For points not on the markers, all of the estimates are within 2 standard errors of their true value. The estimates on top of the markers are also good (≤ 2.13 standard errors), except for the two estimates on top of GPR. However processing MODY first at this location reduces the two errors greatly. Since the location right on top of GPR is the least likely location in the region of interest, it is not surprising that sequential imputation has the most problems here.

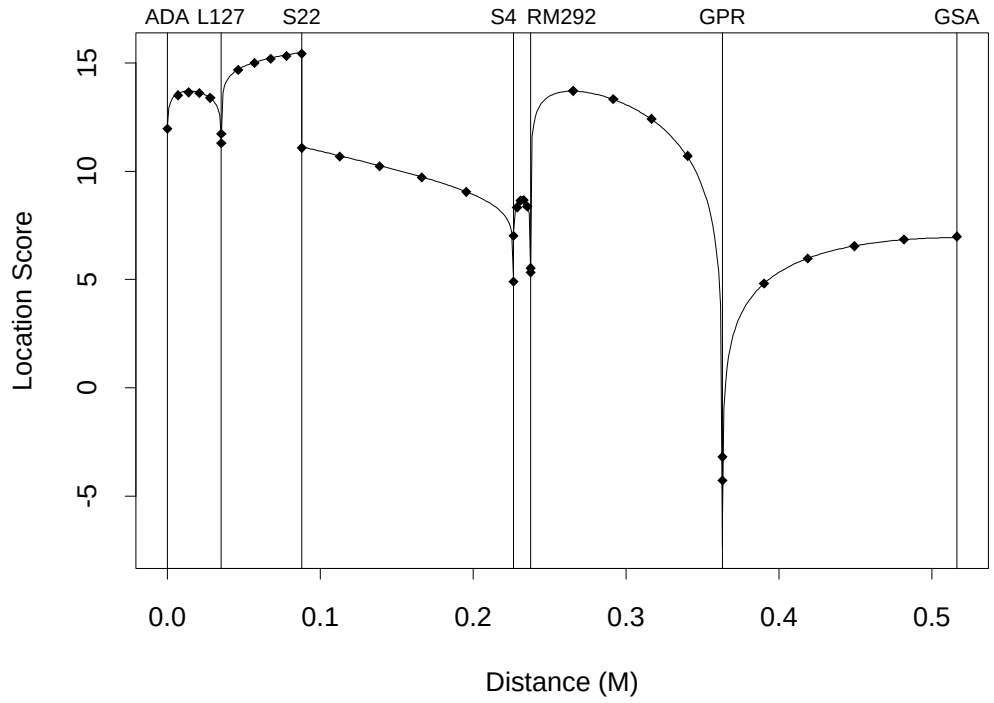

Figure 4.2 Three and four point location scores. The score estimated by sequential imputation are shown by the line and the exact scores as calculated by LINKMAP are shown by the diamonds.

4.3 Nine Point Analysis - CEPH Distances

To determine the most likely position of the MODY locus assuming the CEPH locations of the markers, the method of section 2.3 is used. The markers are processed in the order of RM292, ADA1, L127, GPR, S22, GSA, ADA2, and finally S4. The

disease data is processed at seven different locations, unlinked to the markers and at the midpoints of the six intervals defined by the markers. A total of $m = 10,000$ imputations are done. Nine point location scores are calculated for 104 locations between ADA and GSA (7 to 28 per interval) by equation (2.17). Note that this is a nine point analysis involving $155,520 = 5 \times 2 \times 6 \times 3 \times 2 \times 12 \times 6 \times 3 \times 2$ haplotypes.

A plot of the location scores is shown in figure 4.3. The location with the highest likelihood is between RM292 and GPR, 0.0235 M from RM292. This position has a likelihood estimated to be at least 5.85 times larger than positions in the other 5 intervals. The most likely position, outside of the RM292 - GPR interval is right on top of the S22 marker.

A plot of the standard errors of the location score estimates is shown in figure 4.4. As can be seen in the plot, the standard error are very small for most of the region of interest. The standard errors get large near some of the markers and between S4 and RM292. However for locations with estimated location scores greater than 14, the positions that are probably of most interest, the standard errors range from 0.0115 to 0.0260. Even though the standard errors for positions between S4 and RM292 are much larger, they still are fairly well behaved, with values between 0.081 and 0.105.

4.4 MCEM Analysis of Marker Distances

The procedure described in section 3.2.1 is used to estimate the marker recombination fractions from the RW family data. In the following analysis it should be noted that it is assumed that the distance between ADA1 and ADA2 is assumed to be zero.

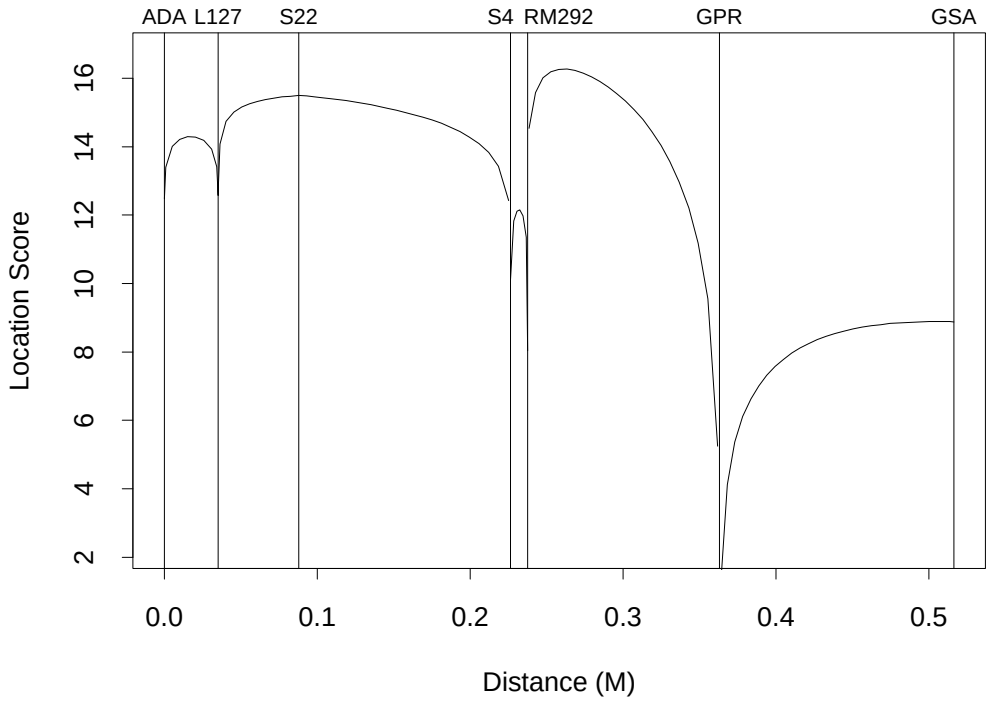

Figure 4.3 Nine Point Location Scores - CEPH Distances

A sequential imputation run of $4,000$ imputations is run on the marker data. The processing order of the sequential imputation run is RM292, ADA1, L127, GPR, S22, GPR, ADA2, and finally S4. The same order was used for this analysis as the analysis of the previous section as the coefficients of variation were well behaved. For this run, instead of using the CEPH distances, the marker distances are set to the value in table 4.4. These distances are based on preliminary MCEM analysis ignoring the ADA2 marker data and with the number of alleles reduced in L127 and GPR from 6 to 4. These distances should be closer to the maximum likelihood estimates

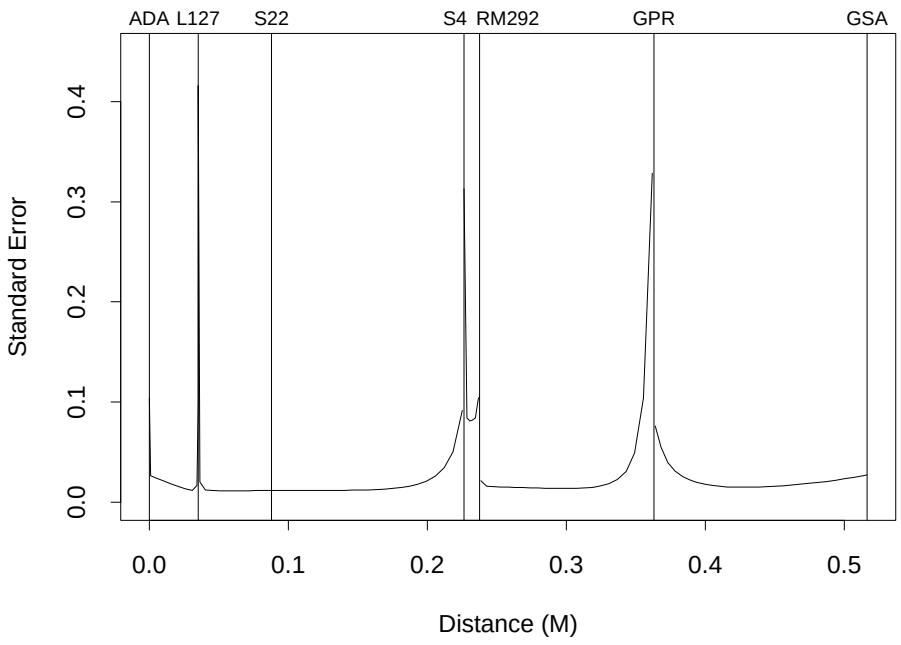

Figure 4.4 Standard Errors of Nine Point Location Scores - CEPH Distances

of the recombination fractions and should give a lower coefficent of variation in the sequential imputation run.

The initial values used in the MCEM procedure are the same as in the sequential imputation run. The stopping criterion of the procedure was to continue until $\max_k |\theta_k^{(i+1)} - \theta_k^{(i)}| < 1.0 \times 10^{-5}$ where $\theta_k^{(i)}, k = 1, \ldots, 6$ is the value of the kth recombination fraction from the ith iteration. The procedure converged in 39 iterations with the likelihood increasing by 1.98 times. The coefficient of variation of the weights increased from 2.99 to 3.17. As the coefficient of variation is still small when the procedure stops, it is not necessary to do another sequential imputation run. Thus

Marker 1	Marker 2	Recombination Fraction
ADA1	ADA2	0
ADA1/ADA2	L127	0.001
L127	S22	0.037
S22	S4	0.028
S4	RM292	0.027
RM292	GPR	0.152
GPR	GSA	0.044

Table 4.4 Distances between adjacent markers used in sequential imputation run for estimation of recombination fractions by MCEM

the MCEM estimates of the recombination fractions, which will be referred to as the MCEM distances, are given in table 4.5. The standard error estimates in this table were calculated by inverting the information matrix estimated by equation (3.9).

It is not necessary to use the values that the sequential imputation run was done at for the initial values of the MCEM procedure. For comparison, the MCEM procedure is rerun using the CEPH distances as initial values. With these initial values and the same stopping criterion as before, the procedure stopped after 44 iterations with the likelihood increasing by 54,000 times. Also the coefficient of variation of the weights decreased from 10.15 to 3.18. Even though CEPH distance are not particularly close to the distances the sequential imputation procedure was run at, the MCEM procedure stabilized quickly with a big increase in the likelihood and a lowering of the

Marker 1	Marker 2	Recombination Fraction	Standard Error
ADA1/ADA2	L127	0	
L127	S22	0.028	0.011
S22	S4	0.038	0.013
S4	RM292	0.028	0.011
RM292	GPR	0.130	0.022
GPR	GSA	0.066	0.016

Table 4.5 Estimated distances between adjacent markers with standard errors

coefficient of variation. The largest difference in the estimated recombination fractions between the two run is 0.0001.

4.5 Nine Point Analysis - MCEM Distances

With the results of the above analysis, it is possible to test whether the CEPH distances are consistent with the data. As mentioned earlier, the likelihood ratio of the MCEM distances to the CEPH distances is 54,000. This corresponds to a likelihood ratio statistic of 21.79 on 5 degrees of freedom (p-value = 0.0006). Note the the degrees of freedom is 5, not 6, since the location of RM292 under the CEPH distances is estimated from the RW pedigree. Thus there is strong evidence that the recombination data in the RW family is not consistent with the CEPH distances. Based on table 4.5 it appears that most of the inconsistency of the CEPH distances is with $\theta_{GPR-GSA}$. Therefore it is reasonable to calculate location scores for MODY

under the MCEM distances.

The sequential imputation run used for estimating location scores for MODY under the CEPH can be used for estimating location score for MODY under the MCEM distances by applying equation (2.14) to calculate an unbiased estimate of the likelihood under the new marker distances. The estimates of the location under the MCEM distances are shown in figure 4.5. Under these marker distances, the most likely location for MODY is still between RM292 and GPR, now at a distance of 0.0267 M from RM292. However this location has a likelihood over 50 times larger than positions in the other 5 intervals. As before the most likely position outside the RM292 - GPR interval is right on top of S22.

The standard errors of the estimated location scores under the MCEM distances are shown in figure 4.6. They generally range from 5 to 15 times larger than the standard errors under the CEPH distances. This is not surprising as the likelihood calculations are being done at a greater distance from the simulation conditions than before. However, this analysis is precise enough to determine the most likely interval containing the MODY locus as the standard errors for positions that have location scores greater than 14 are all less than 0.15. If a more precise estimate of the location of the MODY locus under the MCEM distances is desired, an efficient approach would be to perform a new set of simulations with the MODY locus placed at its most likely position based on this analysis, this time processing the disease locus early on.

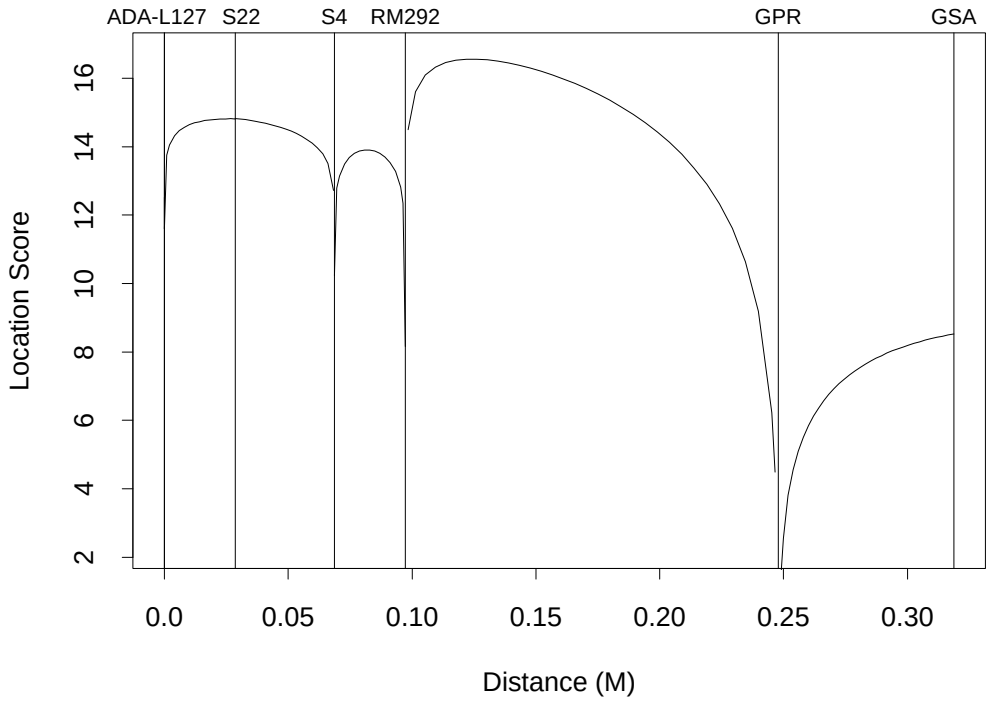

Figure 4.5 Nine Point Location Scores - MCEM Distances

4.6 Tighter Estimation of the Location of MODY

Using the suggestion at the end of the previous section, two sequential imputation runs were done to get more precise estimates of the most likely location of the MODY locus between RM292 and GPR under the two sets of marker distances. In each of the runs, MODY is placed at its most likely location based on the earlier analysis. In both runs the processing order is changed to RM292, MODY, L127, ADA1, ADA2, GPR, S22, GSA, and finishing with S4. The one big change from before is doing only

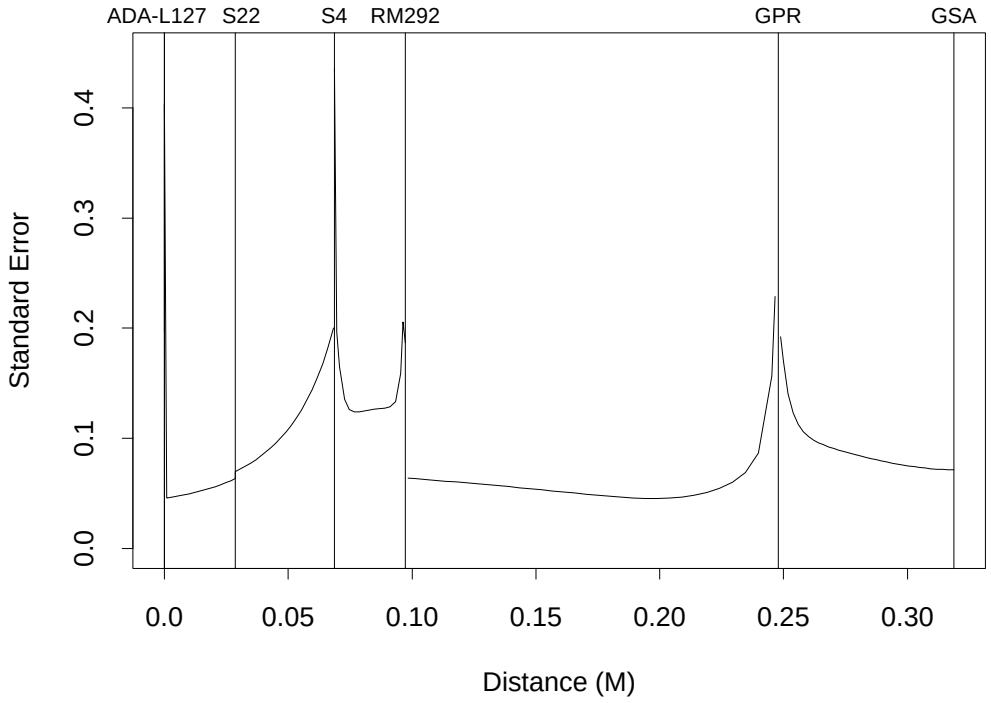

Figure 4.6 Standard Errors of Nine Point Location Scores - MCEM Distances

$m = 4,000$ imputations instead of 10,000. By processing MODY second instead of last, the variance of the weights should decrease by a large factor, implying that less imputations are needed.

Plots of the estimated location scores are shown in figures 4.7 (CEPH distances) and 4.8 (MCEM distances). To calculate the location scores for the new runs, the estimated probability of MODY unlinked to the markers from the earlier run was used. As can be seen in the plots, the estimates of the locations scores under the two processing orders are very similar. Also for both sets of marker distances, the most

likely location for MODY is the same for the new runs as for the original run, with a distance from RM292 of 0.0235 M for the CEPH distances and 0.0267 M for the MCEM distances.

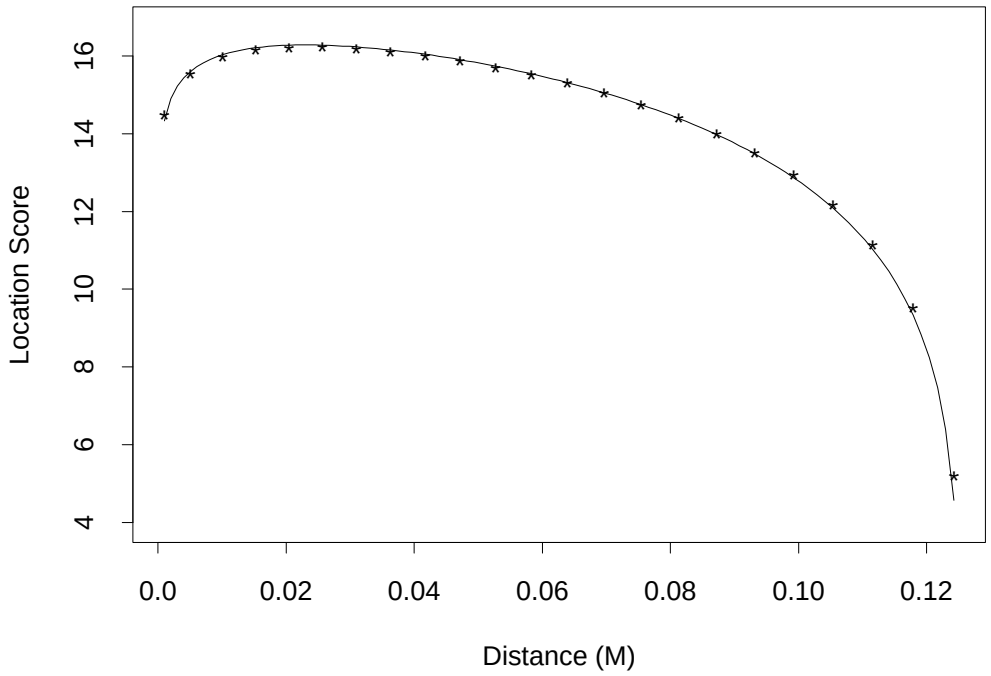

Figure 4.7 More precise estimation of location scores under the CEPH Distances. The estimates under the new run is shown by the solid line and the estimates from the run processing MODY last are shown by *.

To compare the relative efficiency of two sets of estimates of $p_d(\mathbf{y})$ from two

different runs, an appropriate statistic is

$$releff(d) = \frac{(\hat{C}_1[\hat{p}_d(\mathbf{y})])^2}{(\hat{C}_2[\hat{p}_d(\mathbf{y})])^2} \times \frac{m_1}{m_2}, \tag{4.1}$$

which is an approximation to the ratio of the variances of $\log \hat{p}_d(\mathbf{y})$ multiplied by the ratio of sample sizes. This statistic tells how many times more imputations are needed under run 1 to get the same precision in estimating $\log p_d(\mathbf{y})$ under run 2. Values of $releff(d) > 1$ imply that run 2 is more efficient, otherwise run 1 is more efficient.

As can be seen in figures 4.9 and 4.10, the runs processing MODY earlier are more efficient in estimating $\log p_d(\mathbf{y})$ than the runs that process it last. Under the CEPH distances, processing MODY early only requires about half the simulations of processing it last. With the MCEM distances, there are big gains in efficiency in processing MODY early and under the correct marker distances, the bulk of which come from doing the sequential imputation run under the correct marker distances. It is interesting to note that even though the estimate of $\log p_d(\mathbf{y})$ around the maximum is about 29 times more efficient in the second run, the estimate of where the likelihood function is maximized under the MCEM distances is the same for the two runs (based on a grid search of 0.001 on the recombination fraction scale).

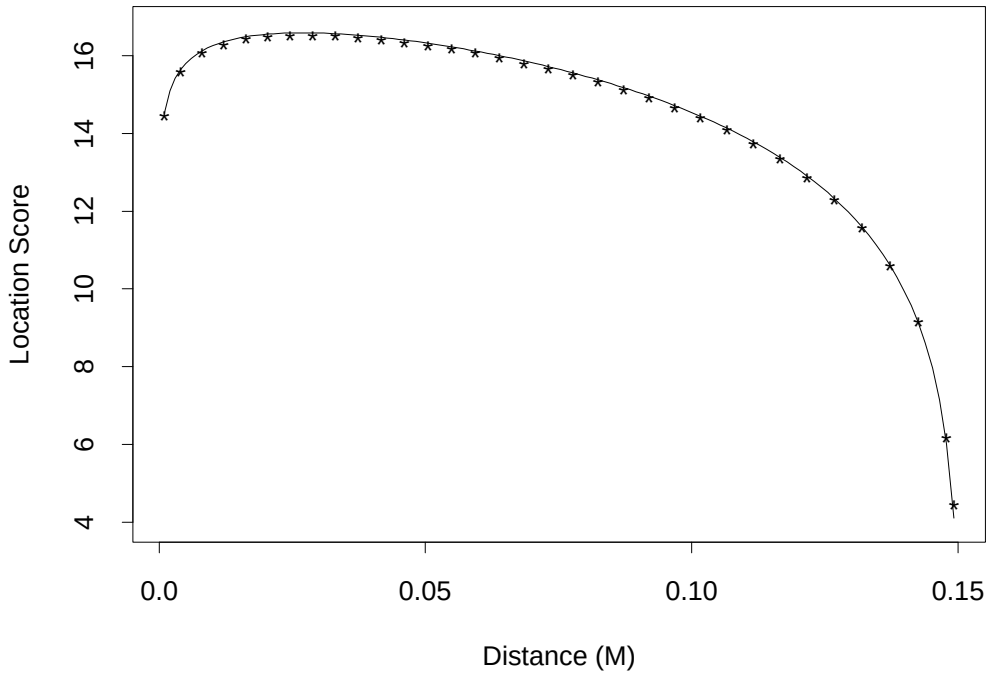

Figure 4.8 More precise estimation of location scores under the MCEM Distances. The estimates under the new run is shown by the solid line and the estimates from the run processing MODY last are shown by *.

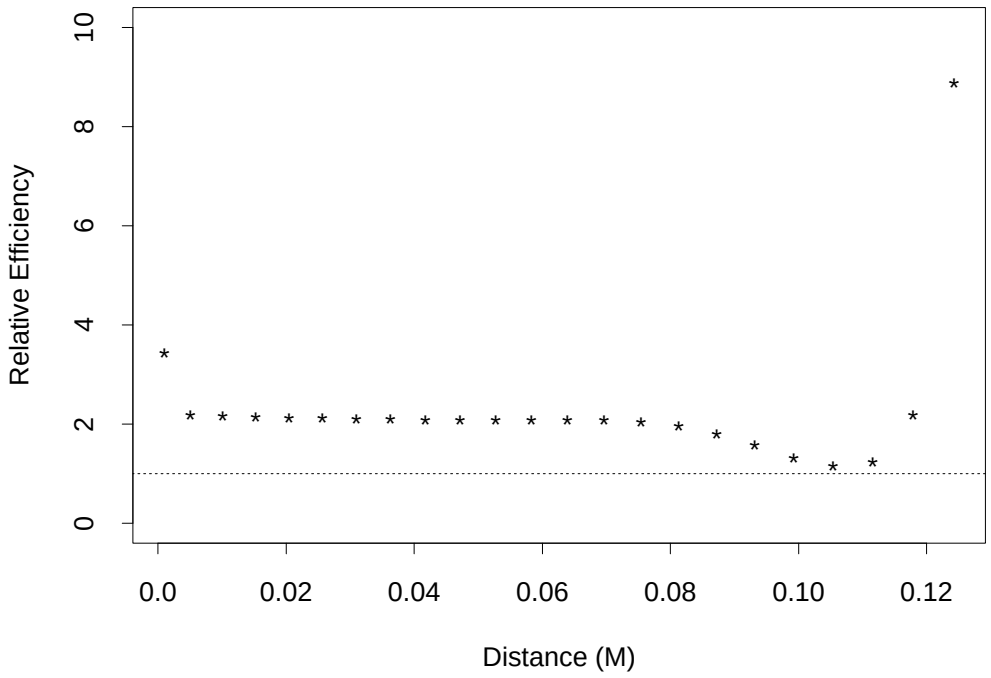

Figure 4.9 The relative efficiency of processing the disease locus second over processing it last under the CEPH distances.

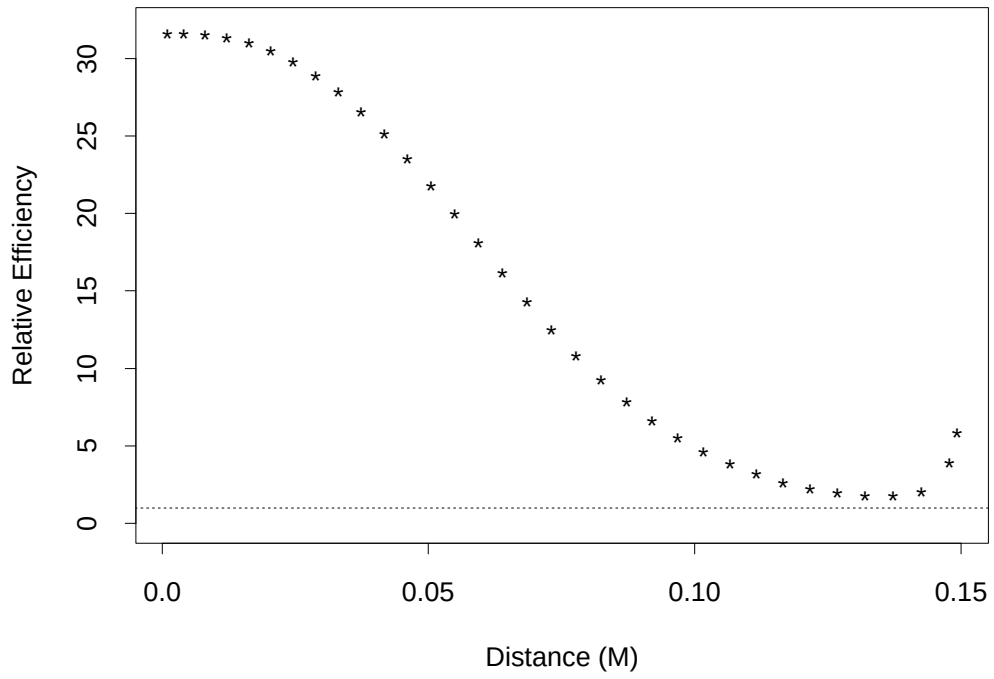

Figure 4.10 The relative efficiency of processing the disease locus second over processing it last under the MCEM distances.

4.7 Bayesian Analysis

A sequential imputation run with $m = 6,000$ was performed to examine the methods discussed in section 3.4. For this run, the recombination fractions were set to the MCEM estimates except for $\theta_{ADA1/ADA2-L127}$ which was set to 0.01 instead. This change was made as preliminary work suggested that setting this value too low would lead to poor behaviour of the weights $w^*(j)$, with large coefficients of variations and 1 or 2 of the iterates dominating the sample. The processing order of this run was kept the same as the run used to determine the MCEM estimates of the recombination fractions.

Assuming a uniform prior on the recombinations fractions (which corresponds to $\alpha_k = \beta_k = 1$ for $k = 1, \ldots, 6$), the estimates of the marginal posterior distributions as calculated by (3.11) are shown in figure 4.11. The posterior means, modes, and standard deviations of these distributions are shown in table 4.6. Also included in this table are the set of MCEM estimates based on this new data set of $m = 6,000$ iterations. It is important to note that the posterior marginal modes are not strictly comparable to the MCEM estimates as the former considers each recombination fraction separately where the latter considers all of the recombination fractions jointly.

Marker 1	Marker 2	Mean	Standard Deviation	Mode	MCEM
ADA1/ADA2	L127	0.007	0.007	0	0
L127	S22	0.030	0.015	0.023	0.028
S22	S4	0.043	0.021	0.036	0.036
S4	RM292	0.032	0.018	0.023	0.029
RM292	GPR	0.137	0.033	0.133	0.130
GPR	GSA	0.078	0.032	0.059	0.066

Table 4.6 Estimates of means, standard deviations and modes of the marginal posterior distributions of recombination fractions

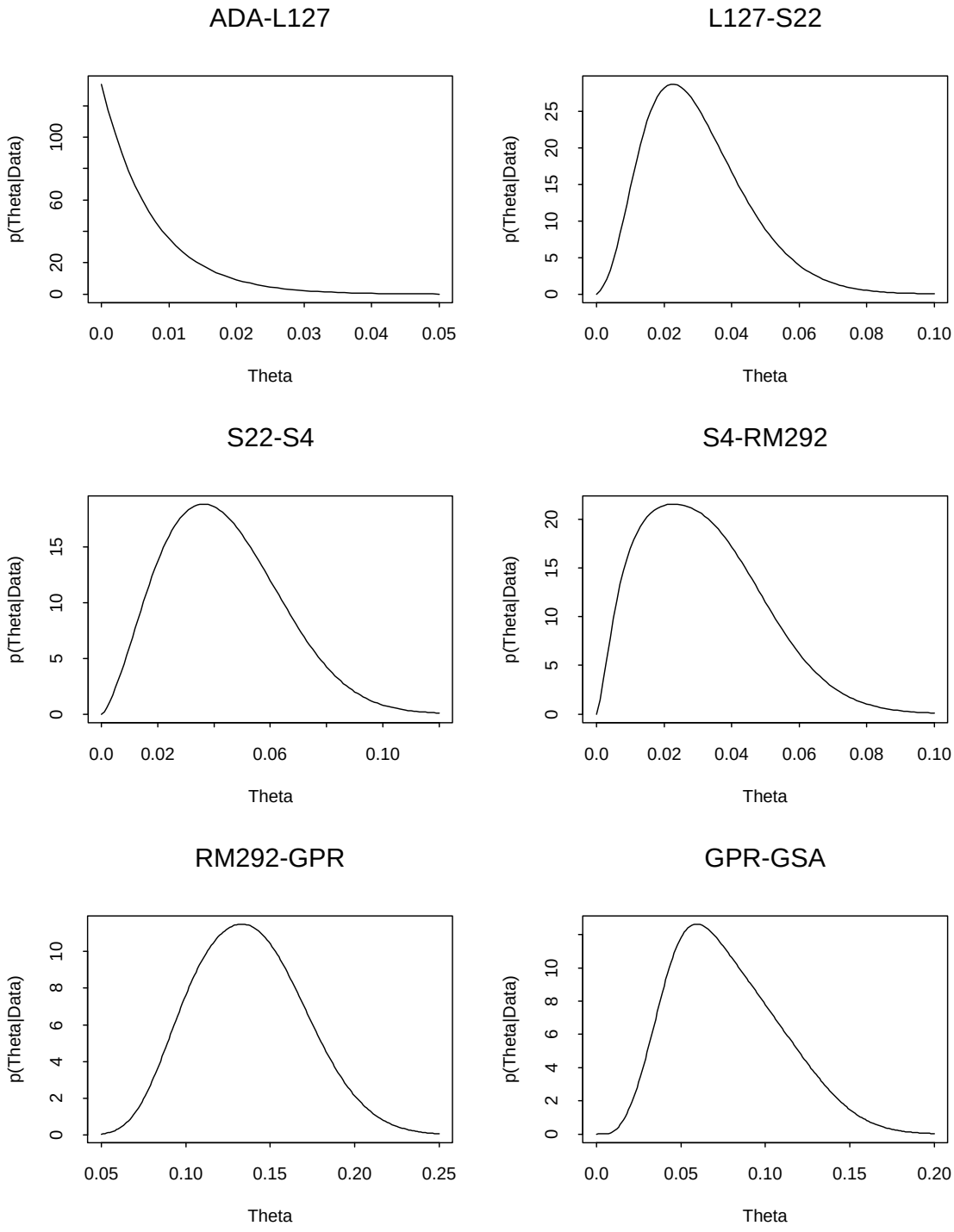

Figure 4.11 Marginal posterior distributions under uniform prior

There is some evidence of some correlation between a number of the recombination fractions. Table 4.7 shows the posterior correlation of all pairs of recombination fractions. Generally the posterior correlations between the recombinations are small, which is not surprising since there is relatively little missing marker data in this pedigree. The correlations that are largest in magnitude occur mainly with adjacent intervals. These correlations are all negative in sign, which is to be expected. Figure 4.12 shows contour plots of the bivariate posterior distributions for the five pairs of adjacent intervals. There are two additional values which stand out in table 4.7, the correlations of GPR-GSA with L127-S22 and with S22-S4. In particular the correlation of GPR-GSA with L127-S22 is particularly surprising since it is positive. The direction of the association for both of these bivariate distributions can be seen in figure 4.13.

	L127-S22	S22-S4	S4-RM292	RM292-GPR	GPR-GSA
ADA1/ADA2-L127	-0.029	-0.037	0.061	-0.064	0.042
L127-S22		-0.180	0.003	-0.019	0.095
S22-S4			-0.171	-0.057	-0.195
S4-RM292				-0.100	-0.006
RM292-GPR					-0.043

Table 4.7 Posterior correlations of recombination fractions under a uniform prior

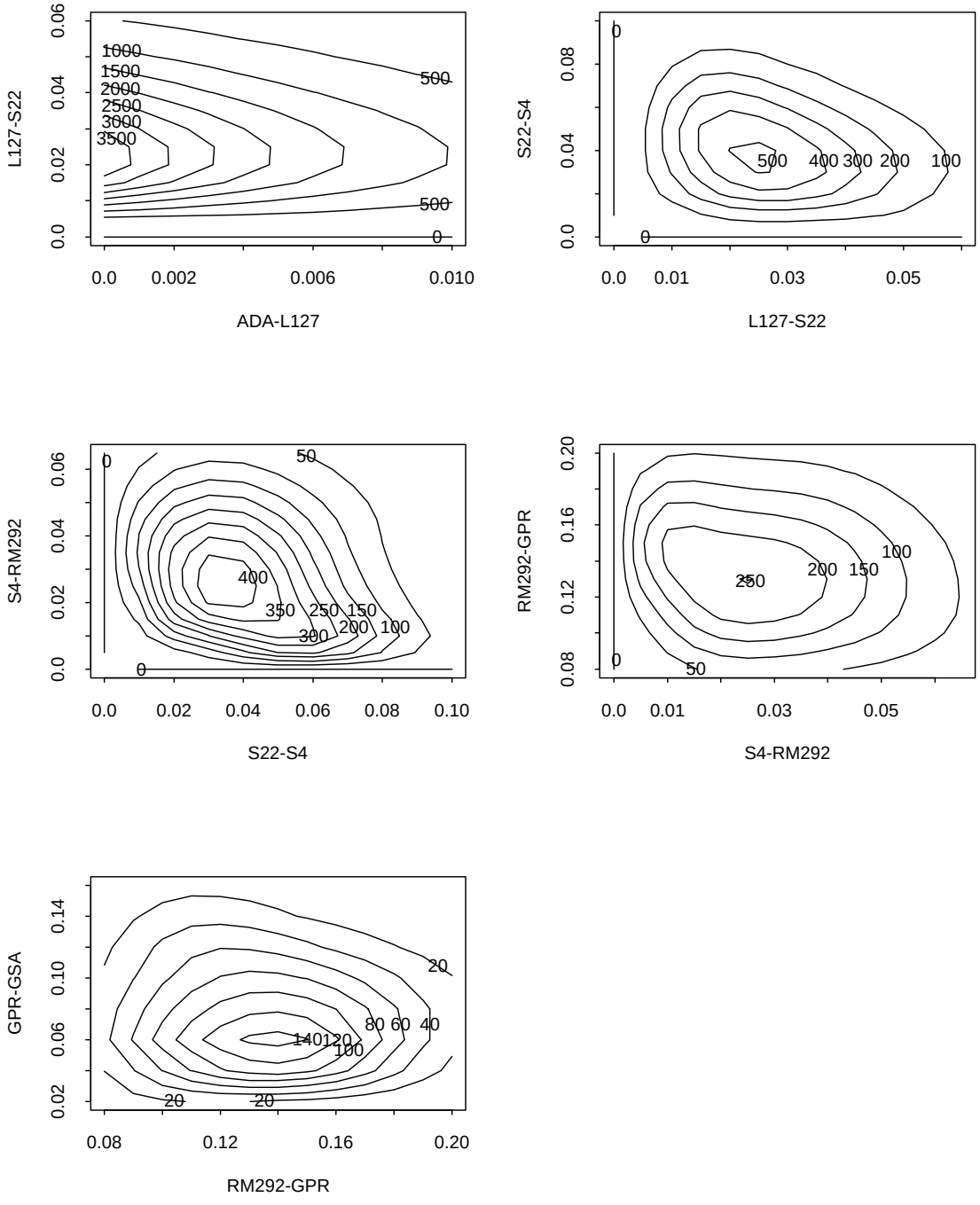

Figure 4.12 Contour plots of bivariate posterior distributions of recombination fractions for adjacent intervals under a uniform prior

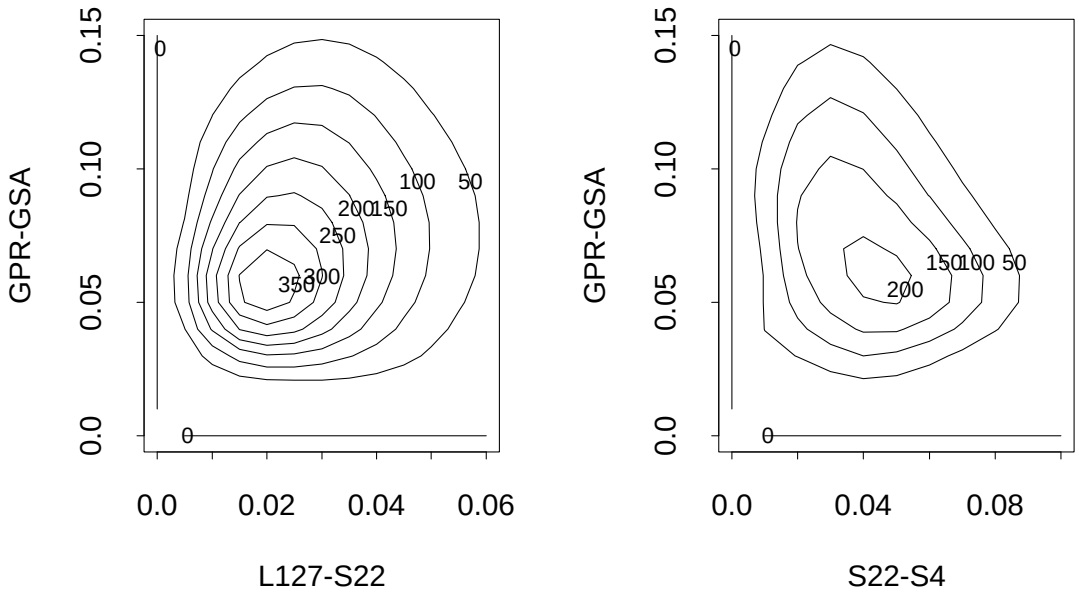

Figure 4.13 Contour plots of bivariate posterior distributions of the recombination fraction between GPR-GSA with L127-S22 and with S22-S4 under a uniform prior

4.8 Computing

To study the computational burden required in the preceding analyses, the CPU time required to run the nine point analysis of section 4.3 was estimated for five different workstations. The five machines examined were an HP 715/64, a Sun SPARC 10/512, a Sun SPARC 20/50, a DECstation 5000/240, and a DECstation 5000/25. Table 4.8 shows for each machine the amount of physical RAM available, the estimated CPU time for the 10000 imputations, and the total memory required for the run.

Machine	Physical RAM	CPU Time	Total Memory Required
HP 715/64	32 Meg	16.67 Hours	20232 K
Sun SPARC 10/512	96 Meg	25 Hours	20312 K
Sun SPARC 20/50	32 Meg	25 Hours	20312 K
DECstation 5000/240	64 Meg	36.25 Hours	33912 K
DECstation 5000/25	24 Meg	65.5 Hours	33912 K

Table 4.8 Estimated CPU time and required memory for section 4.3 analysis.

It needs to be noted that the above times are just for the time spent on running this program. As there are always other processes running, the actual time required to run the program will be longer. Assuming that, except for the required system processes, the sequential imputation program was the only job running on the machine, the actual running times on the HP, the two SPARCS, and possibly the DECstation 5000/240 would likely between two and five percent longer than the CPU time. The

actual time on the DECstation 5000/25 would likely be 20 to 100 percent longer than the CPU time. The problem with this machine is that it does not have enough physical memory, thus requiring a lot of swapping to disk. The problem will not be as severe with this machine when smaller analyses are run. Except for this low powered DECstation 5000/25, the time required to run the analysis is reasonable. It is important to note that various parts of the program, in particular the peeling algorithms, have not yet been optimized for speed. Some of the planned changes, in particular improving the way the program deals with ignorable missing data as discussed in section 2.2.4 should lead to big improvements in computing time and space requirements for some pedigrees. The amount of improvement will depend on the amount of ignorable missing data and the number of possible alleles at each locus.

In comparison, if we tried to evaluate the likelihood for one location of the MODY gene using one of the programs which does exact calculation, we would run into memory problems very quickly. For example, LINKAGE, one of the most popular programs for multipoint analysis was tried on the MODY data. On a Sun SPARC 1 with 32 megabytes of RAM, the program ran into memory problems with 192 haplotypes. Similar problems would occur if the program was run on the HP or the Sun SPARC 20 mentioned above. Even if the memory requirements weren't a problem, on the above machines, the calculation of one likelihood value would take months, possibly years to complete.

Chapter 5

Discussion

Because of the inherent limitations of existing computer software and algorithms which do exact calculations of likelihoods, investigators often have to reduce the size of the data set in their analysis. This is usually done by some combination of reducing the number of loci, reducing the number of alleles per locus, looking at a subset of a pedigree, or splitting a pedigree into two or more separate pedigrees. These actions may lead to a loss of information and may also create bias. For example, Rothschild et al (1993) also looked at the RW pedigree. In their analysis, they only looked at part of the pedigree (the lower two branches in figure 4.1), looked at only three markers at a time, reduced the number of alleles by recoding the haplotype data, and split the remaining part of the pedigree into two separate families. By doing this data reduction, their largest four point lod score was less than some of their two-point lod scores, probably most likely due to their pedigree splitting. Also looking at a subset of the loci at a time, as was done in figure 4.2 has an additional problem. Though it is often done, location scores from different intervals should not be compared.

These location scores measure how likely the hypothesized gene is to be found at that location relative to being on a different chromosome based on a subset of the markers. As different locations use different marker combinations in computing the location scores, they are not comparable. Jumps in the location scale curve or a shift of the most likely location from the full multipoint analysis can occur due to differing marker informativeness. For example, the marker combination of S22 and L127 is more informative than the pair S22 and S4. An additional inefficiency of computing likelihoods point by point, is that careful analysis of the data, which may include various sensitivity analyses, is discouraged. With a Monte Carlo approach, it is easy to examine how sensitive the analysis is to effects of assumed marked distances or disease penetrances.

Sequential imputation is a novel method for imputing the missing data conditioned on the observed data. Alternative methods for performing the imputations include the Gibbs sampler (Guo and Thompson, 1992, Sheehan and Thomas, 1993) and the closely related Metropolis algorithm (Lange and Sobel, 1991). These methods are based on Markov Chain theory (Geman and Geman, 1984, Gelfand and Smith, 1990). Instead of the weighted independent samples generated by sequential imputation, they produced correlated samples with equal weights. Because of the special character of pedigree analysis (Kong 1991b), these methods can sometimes be very inefficient because of the high correlations of the samples. In addition, the Markov chain methods may not work when there are three or more alleles occurring at a locus as the resulting Markov Chain may not be irreducible (Sheehan and Thomas, 1993).

A number of researchers are working on methods which ensure irreducibility (Lin et. al., 1994) and to increase the efficiency of these iterative methods (Geyer and Thompson, 1994). It should be emphasized that some of the problems being handled by the Gibbs sampler, such as inbred pedigrees with many loops and complex traits, do not fall into the area of applications of sequential imputation, the reason being that, in those cases, peeling a single locus can be impossible. However, when restricted to the class of multi-point problems described earlier, we believe that sequential imputation is more efficient than any other existing method. This includes a Gibbs sampling approach suggested in Kong (1991b). Furthermore, there is an additional advantage sequential imputation has over the other iterative methods. Sequential imputation gives direct estimates of likelihoods while these other methods give only estimates of the likelihood ratios between other values of the parameter and the value used for performing the imputations. This estimate of the likelihood ratio (2.8) with h as given in (2.7), is actually also an importance sampling estimate. Hence it can have a very large variance for values of the parameter vector very far from the one used for imputing. This creates problems if we are interested in comparing the likelihoods of different orderings of the loci. Even if separate imputations are performed for the different orders, there may not be dependable estimate of the likelihood ratios. By comparison, sequential imputation does not have the same problem since it provides direct estimates of the likelihoods for different orderings of the loci.

Monte Carlo methods such as sequential imputation are most useful for problems in disease mapping where analyses involving large pedigrees with a substantial amount

of missing data are unavoidable. For mapping of markers, since it is possible to concentrate on small nuclear families with little missing data, exact computations of likelihoods and the implementation of the EM algorithm can be very fast even with a large number of loci using packages developed based on the algorithm given in Lander and Green (1987). Hence, the need for Monte Carlo methods is not as apparent. However, sequential imputation can still be a useful, although less crucial, tool for marker mapping. As seen in chapters three and four, joint estimation of the recombination fractions in large pedigrees can be easily done by Monte Carlo EM. In addition, the Lander and Green algorithm depends critically on the assumption of the lack of genetic interference. In comparison, sequential imputation can allow for interference with no additional cost. While the effects of interference are likely to be small in most situations, the capability to compute probabilities and likelihoods under models which incorporate interference can only help to refine analyses. A related issue is that estimated genetic distances between markers are often published without associated standard errors. That can be partly explained by the technical difficulties in obtaining standard errors when the data are not complete. However, as shown earlier, sequential imputation can be used to estimate the information matrix at the maximum likelihood estimate. Then this estimate of the information matrix can be inverted to get the desired standard errors.

In the MODY example, it was possible to get good estimates of the location scores when processing the marker data first using the method of section 2.3. However when there is more missing data in the upper generations than in the MODY example,

this approach may be inefficient. In Kass et. al. (1994), they applied the method of sequential imputation to locate a gene causing Conduction System disease and Dilated Cardiomyopathy. As the disease mechanism was hypothesized to be a simple dominant disease with high penetrance, they found it more efficient to process the disease locus first even though multiple runs were required. The assumed disease model implied that the disease locus was highly informative and helped to make up for the multiple runs.

The efficiency of sequential imputation depends on the coefficient of variation of the importance sampling weights. In section 2.2, a number of ways to improve the efficiency were proposed. Except for processing two or more loci at a time, all of them were implemented for the analyses performed in chapter 4. While it does not seem to be necessary to process more than one locus at a time for the MODY example, that may not be the case with other data sets, particularly those which have missing marker data for three or more generations at the top of the pedigree. Going in the other direction, we can also contemplate splitting a locus in two or more artificially. For example, a locus with 12 alleles can be considered as two loci right on top of each other with 4 and 3 alleles each. The split can be chosen so that one of these half-loci carries a lot more information than the other half and is processed first. Moreover, two halves of two different loci can be combined during processing to reduce the variations of the weights. Finally, although the current software only handles pedigrees without loops, sequential imputation can be useful for pedigrees that have a few loops, as long as the peeling of a single locus can be efficiently performed.

References

Bell, G.I., Xiang, K.S., Newman, M.V., Wu, S.H., Wright, L.G., Fajans, S.S., Spielman, R.S., and Cox, N.J. (1991). Gene for the Non-insulin-dependent Diabetes Mellitus (Maturity Onset Diabetes of the Young) is Linked to DNA Polymorphism on Human Chromosome 20q. *Proc. Natl. Acad. Sci. USA* **88**: 1484-1488.

Cannings, C., Thompson, E.A., and Skolnick, M.H. (1978). Probability Functions on Complex Pedigrees. *Adv. Appl. Probab.* **10**: 26-61.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *J. R. Statist. Soc.* B **39**: 1-38.

Elston, R.C. and Stewart, J. (1971). A General Model for the Genetic Analysis of Pedigree Data. *Hum. Hered.* **21**: 523-542.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *J. Amer. Statis. Assoc.* **85**: 398-409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6**: 721-741.

Geyer, G. and Thompson, E. (1994). Annealing Markov Chain Monte Carlo with Applications to Pedigree Analysis. To appear in *J. Amer. Statis. Assoc.*

Guo, S.W. and Thompson, E.A. (1992). A Monte Carlo Method for Combined Segregation and Linkage Analysis. *Am. J. Hum. Genet.* **51**: 1111-1126.

Kass, S., MacRae, C., Graber, H.L., Sparks, E.A., McNamara, D., Boudoulas, H., Basson, C.T., Baker, P.B., Cody, R.J., Fishman, M.C., Cox, N., Kong, A., Wooley, C.F., Seidman, J.G., and Seidman, C.E. (1994). A Gene Defect that Causes Conduction System Disease and Dilated Cardiomyopathy Maps to Chromosome 1p1-1q1. *Nature Genet.* **7**:546-551.

Kong, A. (1991a). Efficient Methods for Computing Linkage Likelihoods of Recessive Diseases in Inbred Pedigrees. *Genet. Epidemiol.* **8**: 81-103.

Kong, A. (1991b). Analysis of Pedigree Data Using Methods Combining Peeling and Gibbs Sampling. *Procedings of the 23rd Symposium on the Interface.*

Kong, A., Liu, J.S., and Wong, W.H. (1991). Sequential Imputations and Bayesian Missing Data Problems. *J. Amer. Statis. Assoc.* **89**: 278-288.

Lander, E.S. and Green, P. (1987). Construction of Multilocus Genetic Linkage Maps in Humans. *Proc. Natl. Acad. Sci. USA* **84**: 2363-2367.

Lange, K. and Boehnke, M. (1983). Extensions to Pedigree Analysis V. Optimal Calculation of Mendelian Likelihoods. *Hum. Hered.* **33**: 291-301.

Lange, K. and Elston, R.C. (1975). Extensions to Pedigree Analysis I. Likelihood Calculations for Simple and Complex Pedigrees. *Hum. Hered.* **25**: 95-105.

Lange, K. and Sobel, E. (1991). A random walk method for computing genetic location scores. *Am. J. Hum. Genet.* **49**: 1320-1334.

Lathrop, G.M., Lalouel, Julier, C., and Ott, J. (1984). Strategies for Multilocus Linkage Analysis in Humans. *Proc. Natl. Acad. Sci. USA* **81**: 3443-3446.

Lin, S., Thompson, E., and Wijsman, E. (1994). Finding Noncommunicating Sets for Markov Chain Monte Carlo Estimations on Pedigrees. *Am. J. Hum. Genet.* **54**:695-704.

Louis, T.A. (1982). Finding the Observed Information Matrix when Using the *EM* Algorithm. *J. R. Statist. Soc.* B **44**: 226-233.

NIH/CEPH Collaborative Mapping Group (1992). A Comprehensive Genetic Linkage Map of the Human Genome. *Science* **258**: 67-86.

Ott, J. (1989). Computer-simulation Methods in Human Linkage Analysis. *Proc. Natl. Acad. Sci. USA* **86**: 4175-4178.

Ott, J. (1991). Analysis of Human Genetic Linkage. Revised Edition. The Johns Hopkins University Press, Baltimore.

Ploughman, L.M. and Boehnke, M. (1989). Estimating the Power of a Proposed Linkage Study for a Complex Genetic Trait. *Am. J. Hum. Genet.* **44**: 543-551.

Rothschild, C.B., Akots, G., Hayworth, R., Pettenati, M.J., Rao, P.N., Wood, P., Stolz, F., Hansmann, I., Serino, K., Keith, T.P., Fajans, S.S., and Bowden, D.W. (1993). A Genetic Map of Chromosome 20q12-q13.1: Multiple Highly Polymorphic Microsatellite and RFLP Markers Linked to the Maturity-Onset Diabetes of the Young (MODY) Locus. *Am. J. Hum. Genet.* **52**: 110-123.

Schellenberg, G.D., Bird, T.D., Wijsman, E.M., Orr, H.T., Anderson, L., Nemens, E., White, J.A., Bonnycastle, L., Weber, J.L., Alonso, M.E., Potter, H., Heston, L.L., and Martin, G.M. (1992). Genetic Linkage Evidence for a Familial Alzheimer's Disease Locus on Chromosome 14. *Science* **258**: 668-671.

Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York.

Sheehan, N. and Thomas, A. (1993). On the Irreducibility of a Markov Chain Defined on a Space of Genotype Configurations by a Sampling Scheme. *Biometrics* **49**: 163-175.

Thompson, E.A. (1986). Pedigree Analysis in Human Genetics. The Johns Hopkins University Press, Baltimore.

Wei, G.C.G and Tanner, M.A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *J. Amer. Statist. Assoc.* **85**: 699-704.