Chapter 4

QTL Mapping in Random Populations

13

Random Model Approach to QTL Mapping

The mapping populations we have discussed so far are all initiated from crosses of two or a few lines (breeds). As a result, the number of alleles is relatively small, and thus the conclusion is drawn based on narrow genetic variation. In addition, through control of the mating design, we can control the allele frequencies. Because the number of alleles is determined by the number of inbred lines involved in a line crossing experiment and the number of lines is small, we can estimate and test the allelic effects or the average effects of allelic substitution. The linear models that allow us to estimate and test the allelic effects are called fixed effect models. Therefore, all methods we have learned so far are based on the fixed model approach.

When designed matings are impossible, we must collect data as they exist and use such data to conduct QTL mapping. Because the number of alleles involved in the mapping population is unknown and we cannot control the allelic frequencies, the fixed model approach is hard to implement. In this chapter, we introduce an alternative approach of QTL mapping that involves multiple alleles, the random model approach to QTL mapping. Under the random model framework, rather than estimating and testing the effects of allelic substitution of QTL, we estimate and test the variances of the allelic effects for the QTL.

The mapping population may consist of a few large pedigrees or many small pedigrees (e.g., nuclear families). A pedigree is a collection of genetically related individuals descending from a few ancestors. Two types of pedigrees are commonly used in QTL mapping: complicated pedigrees and simple pedigrees. A complicated pedigree is a collection of relatives that expand for multiple generations. Members in a complicated pedigree can be inbred or outbred and their relationships can be arbitrarily complicated. A simple pedigree, however, consists of two outbred parents and their children, and thus it is also called a nuclear family. When the phenotypic values of the parents are excluded from the analysis, the method is called full-sib analysis. In this chapter, we only discuss the random model methodology using multiple full-sib families. Extension of the method to QTL mapping for complicated pedigrees will be mentioned briefly toward the end of this chapter.

A brief introduction to the milestones of the random model methodology of QTL mapping is presented in this paragraph. The random model methodology is also called variance component analysis (Searle et al., 1992). The parameters of interest are variances of the random effects rather than the effects themselves. In terms of QTL mapping, the parameter of interest is the variance of the allelic effects of the locus under investigation in the mapping population. Early works of random model QTL mapping include Goldgar (1990), Schork (1993),Amos (1994) and Xu and Atchley (1995). These studies laid the foundation for the popular QTL mapping procedures in human pedigrees (e.g., Almasy and Blangero (1998)). Goldgar (1990) partitioned the entire genome into many regions (chromosome segments) and used a multipoint method to estimate the identity-by-descent (IBD) value shared by pair of relatives (siblings) for the target region. Using the maximum likelihood method, Goldgar (1990) was able to estimate the genetic variance explained by that chromosome segment. Schork (1993) extended the method and proposed to estimate variance components of multiple segments simultaneously. In addition, Schork (1993) also included a common environmental effect shared by relatives and estimate the common environmental variance. Amos' (1994) model differs from Goldgar (1990) in that fixed effects not relevant to genetics are included in the model. Therefore, Amos' (1994) method is a linear mixed model approach. Another new feature of Amos' (1994) model is that he replaced the IBD of a chromosome region by the IBD of a marker. Xu and Atchley (1995) adopted the idea of interval mapping (Lander and Botstein, 1989) to estimate the genetic variance of a particular location of the genome using flanking markers. Using genome-wide markers, Xu and Atchley (1995) were able to scan the entire genome for QTL under the random model approach.

Prior to the maximum likelihood methods of QTL variance estimation, Haseman and Elston (1972) developed a sib-pair regression method for estimating genetic variance of a polymorphic marker. They found that the squared difference between the phenotypic values of a sib-pair is a linear function of the genetic variance of a marker. Therefore, they regressed the squared phenotypic difference on the IBD value of a sib-pair to obtain an estimation of the genetic variance. Many people believe that the regression model of Haseman and Elston (1972) is a genius. However, the real creativity of Haseman and Elston (1972) comes from the recognition of the variation of sib-pair IBD and the method to calculate the conditional expectation of IBD values given marker information. It is the variance of locus specific IBD that allows the separation of the QTL variance from the polygenic variance. The original sib-pair regression of Haseman and Elston (1972) is still a marker analysis. It is the sib-pair interval mapping of Fulker and Cardon (1994) that allows the entire genome to be scanned, and thus puts QTL mapping of random populations in the same framework as interval mapping of line crosses (Lander and Botstein, 1989).

13.1 Identity-by-descent (IBD)

Identity-by-descent is a special terminology in quantitative genetics to describe the relationship between two alleles. If two alleles are the same copy of an ancestral allele in the past, the two allele are said to be identical-by-descent (IBD). In contrast, two alleles are said to be identical-by-state (IBS) if they have the same allelic form, regardless of their origins. Without mutation, two alleles that are IBD must also be IBS, however, the reverse is not necessarily true. We now use Figure 13.1 (modified from Lynch and Walsh (1998)) to demonstrate the difference between IBD and IBS. This diagram shows the paths of the four alleles of the parents to the four alleles of the progeny. Such a diagram is called a descent graph. The progeny in the left (sib 1) has two A_1 alleles, but only A_1 in the left is IBD to the A_1 allele carried by the progeny in the right (sib 2). All three A_1 alleles in the progeny are IBS. The term IBD is an event describing the relationship between two alleles. In a diploid organism, however, an individual has two alleles at any given locus. To describe the relationship between two individuals, we define the IBD value as a proportion of the number of IBD alleles. Two individuals can share two IBD alleles, one IBD allele or non-IBD allele. Therefore, the IBD proportion between two individuals can be $2/2 = 1.0$, $1/2 = 0.5$ or $0/2 = 0.0$, depending on how many IBD alleles shared by the two individuals.

IBD is the key of QTL mapping under the random model methodology. We now discuss the IBD value between siblings and the properties of the IBD value. Let $A_1^s A_2^s$ and $A_1^d A_2^d$ be the genotypes of the father and the mother of a nuclear family, respectively, where A_1^s and A_2^s are the two alleles of the father and A_1^d and A_2^d are the two alleles of the mother. Note that the two alleles

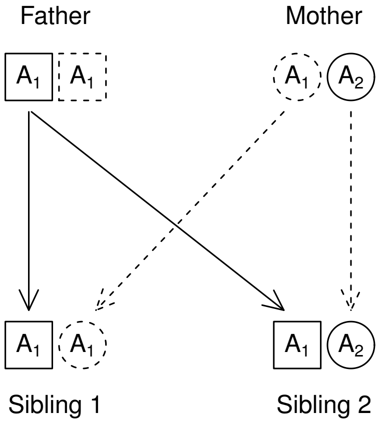

Fig. 13.1. Descent graph showing the allelic transmissions from the parents to the progeny. Sib 1 has two A_1 alleles, but only A_1 in the left is IBD to the A_1 allele carried by sib 2. All three A_1 alleles in the progeny are IBS.

carried by a parent are ordered. For example, A_1^s and A_2^s represent the paternal and maternal alleles of the father, respectively. Previously (Chapter 12), we used A_1, A_2, A_3 and A_4 to represent the four alleles carried by the two parents. Now, these four alleles are represented by A_1^s, A_2^s, A_1^d and A_2^d, respectively. We use this new notation simply to represent the four different origins of the alleles. These four alleles are not necessarily different in terms of the allelic forms (states). For example, if both A_1^s and A_2^s have the same allelic state, say A_1, then the actual genotype of the father is A_1A_1, a homozygote.

Four possible genotypes in terms of the allelic origins can be generated by the mating pair, which are $A_1^sA_1^d$, $A_1^sA_2^d$, $A_2^sA_1^d$ and $A_2^sA_2^d$, each with an equal probability. If we randomly sample a pair of siblings from the family, there will be 16 possible combinations of the sib-pairs. These 16 sib-pair combinations are listed in Table 13.1. We now evaluate each sib-pair to see how many alleles are shared by the siblings. Take sib pair $A_1^sA_1^d - A_1^sA_2^d$ for example, the two siblings share one common allele from their father, A_1^s, but the two alleles from their mother are different in origin. Since each individual has two alleles at any locus, the proportion of IBD alleles shared by the siblings is $\pi = 1/2 = 0.5$. However, sib-pair $A_1^sA_1^d - A_1^sA_1^d$ share both alleles IBD and thus their IBD value is $\pi = 2/2 = 1.0$. Although the two individuals are siblings, they behave like identical twins at the locus of interest. Some sib pairs, e.g., $A_1^sA_1^d - A_2^sA_2^d$, do not share any IBD allele and thus $\pi = 0/2 = 0.0$. For this locus, these two individuals act like strangers, although they are actually siblings. We can see that the IBD proportion shared by siblings, denoted by π, is a discrete variable, taking one of three possible values. Among the 16 possible sib pairs, four of them (on the major diagonals) share two IBD alleles ($\pi = 1.0$) , four of them (on the minor diagonals) share none IBD allele ($\pi = 0.0$), and the remaining eight sib pairs share one IBD allele ($\pi = 0.5$). Therefore, the expectation of the IBD value is

Table 13.1. The 16 possible pairs of siblings and the numbers of IBD alleles (in parentheses) shared by each sib-pair

	Sib two			
Sib one	$A_1^sA_1^d$	$A_1^sA_2^d$	$A_2^sA_1^d$	$A_2^sA_2^d$
$A_1^sA_1^d$	$A_1^sA_1^d$-$A_1^sA_1^d$(2)	$A_1^sA_1^d$-$A_1^sA_2^d$(1)	$A_1^sA_1^d$-$A_2^sA_1^d$(1)	$A_1^sA_1^d$-$A_2^sA_2^d$(0)
$A_1^sA_2^d$	$A_1^sA_2^d$-$A_1^sA_1^d$(1)	$A_1^sA_2^d$-$A_1^sA_2^d$(2)	$A_1^sA_2^d$-$A_2^sA_1^d$(0)	$A_1^sA_2^d$-$A_2^sA_2^d$(1)
$A_2^sA_1^d$	$A_2^sA_1^d$-$A_1^sA_1^d$(1)	$A_2^sA_1^d$-$A_1^sA_2^d$(0)	$A_2^sA_1^d$-$A_2^sA_1^d$(2)	$A_2^sA_1^d$-$A_2^sA_2^d$(1)
$A_2^sA_2^d$	$A_2^sA_2^d$-$A_1^sA_1^d$(0)	$A_2^sA_2^d$-$A_1^sA_2^d$(1)	$A_2^sA_2^d$-$A_2^sA_1^d$(1)	$A_2^sA_2^d$-$A_2^sA_2^d$(2)

$$\mathrm{E}(\pi) = \frac{1}{4} \times 1.0 + \frac{1}{2} \times 0.5 + \frac{1}{4} \times 0.0 = \frac{1}{2}. \tag{13.1}$$

The variance of the IBD value is

$$\mathrm{var}(\pi) = \mathrm{E}(\pi^2) - \mathrm{E}^2(\pi) = \frac{1}{4} \times 1.0^2 + \frac{1}{2} \times 0.5^2 + \frac{1}{4} \times 0.0^2 - 0.5^2 = \frac{1}{8}. \tag{13.2}$$

If we consider the whole genome, siblings share half of their genome (genetical material), and thus have a genome-wide IBD proportion of 0.5. However, if we consider a single locus, the IBD proportion is a variable with an expectation of $\frac{1}{2}$ and a variance of $\frac{1}{8}$. This variance is the key to the random model analysis of QTL for outbred populations.

13.2 Random effect genetic model

Consider a mapping population consisting of n full-sib families each with two siblings. Let y_{j1} and y_{j2} be the phenotypic values of a quantitative trait for the two siblings in family j for $j = 1, \ldots, n$. The phenotypes can be described by the following linear models,

$$\begin{aligned} y_{j1} &= \mu + a_{j1} + \gamma_{j1} + \epsilon_{j1} \\ y_{j2} &= \mu + a_{j2} + \gamma_{j2} + \epsilon_{j2} \end{aligned} \tag{13.3}$$

where μ is the population mean of the trait, γ_{j1} and γ_{j2} are the genetic effects of a putative QTL for the two siblings, a_{j1} and a_{j2} are the polygenic effects (collective effects of all loci of the genome), and ϵ_{j1} and ϵ_{j2} are the residual errors for the two siblings. These equations can be expressed in matrix notation as

$$\begin{bmatrix} y_{j1} \\ y_{j2} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} + \begin{bmatrix} a_{j1} \\ a_{j2} \end{bmatrix} + \begin{bmatrix} \gamma_{j1} \\ \gamma_{j2} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \tag{13.4}$$

The expectation of the array of phenotypic values is

$$\mathrm{E} \begin{bmatrix} y_{j1} \\ y_{j2} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \end{bmatrix} \tag{13.5}$$

and the variance-covariance matrix of the phenotypes is

$$\mathrm{var} \begin{bmatrix} y_{j1} \\ y_{j2} \end{bmatrix} = \mathrm{var} \begin{bmatrix} a_{j1} \\ a_{j2} \end{bmatrix} + \mathrm{var} \begin{bmatrix} \gamma_{j1} \\ \gamma_{j2} \end{bmatrix} + \mathrm{var} \begin{bmatrix} \epsilon_{j1} \\ \epsilon_{j2} \end{bmatrix}, \tag{13.6}$$

where

$$\mathrm{var} \begin{bmatrix} \gamma_{i1} \\ \gamma_{i2} \end{bmatrix} = \begin{bmatrix} 1 & \pi_j \\ \pi_j & 1 \end{bmatrix} \sigma_\gamma^2 \tag{13.7}$$

$$\mathrm{var} \begin{bmatrix} a_{j1} \\ a_{j2} \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \sigma_a^2 \tag{13.8}$$

and

$$\mathrm{var} \begin{bmatrix} \epsilon_{j1} \\ \epsilon_{j2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \sigma^2 \tag{13.9}$$

The three variance components, σ_a^2, σ_γ^2 and σ^2, are the the polygenic variance, the genetic variance of the QTL and the residual variance, respectively. Note that the covariance between the siblings at the QTL is $\mathrm{cov}(\gamma_{j1}, \gamma_{j2}) = \pi_j \sigma_\gamma^2$

while the covariance at the polygene is $\mathrm{cov}(a_{j1}, a_{j2}) = 0.5\sigma_a^2$. The siblings are assumed to share no common environmental effect and thus $\mathrm{cov}(\epsilon_{j1}, \epsilon_{j2}) = 0.0\sigma^2$. The three different coefficients, π_j, 0.5 and 0.0, in the covariance between siblings, allow us to separate the three variance components, and thus to estimate and test the QTL variance σ_γ^2. Let $y_j = \{y_{j1}, y_{j2}\}$, $\gamma_j = \{\gamma_{j1}, \gamma_{j2}\}$, $a_j = \{a_{j1}, a_{j2}\}$ and $\epsilon_j = \{\epsilon_{j1}, \epsilon_{j2}\}$ be vector presentations of the corresponding contents. Let $1 = \{1, 1\}$ be a unity vector. The linear models for the two siblings can be rewritten as

$$y_j = 1\mu + a_j + \gamma_j + \epsilon_j. \tag{13.10}$$

Let

$$\Pi_j = \begin{bmatrix} 1 & \pi_j \\ \pi_j & 1 \end{bmatrix} \tag{13.11}$$

be the IBD matrix for the QTL,

$$A_j = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{13.12}$$

be the additive relationship matrix for the polygene and

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{13.13}$$

be the identity matrix for the residual error. The compact matrix representations for the expectation and variance of y_j are

$$\mathrm{E}(y_j) = 1\mu \tag{13.14}$$

and

$$\mathrm{var}(y_j) = V_j = A_j\sigma_a^2 + \Pi_j\sigma_\gamma^2 + I\sigma^2, \tag{13.15}$$

respectively. Although A_j is family independent, i.e., all families have the same A_j, we still use a subscript j for notational consistency. In addition, when we deal with families with variable size, the dimension of A_j will be different from one family to another. Variable family size will be dealt with in a later section.

13.3 Sib-pair regression

The Haseman and Elston (1972) sib-pair regression method does not use the original phenotypic values as response variables; rather, the squared difference between the phenotypic values of a sib-pair is treated as the response variable. Define $s_j = (y_{j1} - y_{j2})^2$ as the squared difference of sib-pair j. The expectation of s_j is

$$
\begin{aligned}
E(s_j) =& E\left[(y_{j1} - y_{j2})^2\right] \\
=& \sigma_a^2 + 2\sigma_\gamma^2 + 2\sigma^2 - 2\sigma_\gamma^2\pi_j
\end{aligned}
\tag{13.16}
$$

Let $\beta = -2\sigma_\gamma^2$ be the regression coefficient and

$$
\alpha = \sigma_a^2 + 2\sigma_\gamma^2 + 2\sigma^2 \tag{13.17}
$$

be the intercept, the above model is rewritten as

$$
s_j = \alpha + \beta\pi_j + \varepsilon_j \tag{13.18}
$$

where ε_j is the residual for the linear model and the variance of ε_j is denoted by σ_ε^2. The residual is not normally distributed and the residual variance here is not the environmental variance. Let $\hat{\beta}$ be the least square estimate of the regression coefficient. The estimated genetic variance for the marker is

$$
\hat{\sigma}_\gamma^2 = -\frac{1}{2}\hat{\beta} \tag{13.19}
$$

One may notice that the intercept given in equation (13.17) is different from that of Haseman and Elston (1972), which is

$$
\alpha = 2\sigma_\gamma^2 + \sigma_e^2 \tag{13.20}
$$

The Haseman and Elston (1972) model assumes absence of polygenic effect and thus σ_a^2 is excluded. The environmental error defined in Haseman and Elston (1972) is $e_j = \epsilon_{j1} - \epsilon_{j2}$, the difference between the environmental errors of the sib-pair. Therefore, $\sigma_e^2 = 2\sigma^2$, twice the environmental error variance defined in this chapter.

This regression analysis looks strange because the regression coefficient $\beta = 2\sigma_\gamma^2$ also appears in the intercept (as a component). The model can be revised to the following form,

$$
s_j = \alpha + \beta(\pi_j - 1) + \varepsilon_j \tag{13.21}
$$

where α is redefined as $\alpha = \sigma_a^2 + 2\sigma^2$, which has excluded $2\sigma_\gamma^2$. However, the independent variable is $\pi_j - 1$, instead of π_j. The final estimation of the β remains the same.

13.4 Maximum likelihood estimation

We assume that π_j is known for all $j = 1, \ldots, n$. This represents a situation where we can observe the genotypes of the QTL. In QTL mapping, the genotype of a QTL is not observable but the distribution of the genotype can be inferred from marker information. The IBD value of a putative QTL is then replaced by the IBD value estimated from marker data. The likelihood function discussed in this section is based on known IBD values. Method dealing with inferred IBD will be deferred to a later section, where a multipoint method for estimating the IBD values will also be introduced. The IBD values for all the sib-pairs are treated as data. The data also include the phenotypic values of the siblings. The parameter vector is $\theta = \{\mu, \sigma_\gamma^2, \sigma_a^2, \sigma^2\}$. The log likelihood function is

$$L(\theta) = -\frac{1}{2}\sum_{j=1}^{n} \ln |V_j| - \frac{1}{2}\sum_{j=1}^{n} (y_j - 1\mu)^{\mathrm{T}} V_j^{-1} (y_j - 1\mu). \tag{13.22}$$

This likelihood function is complicated and the explicit solution for the MLE of the parameters is not available.

13.4.1 EM algorithm

Xu and Atchley (1995) used the simplex algorithm (Nelder and Mead, 1965) to find the MLE of θ. Here we introduce the EM algorithm to estimate θ. Under the random model framework, the parameters of interest are the variance components. Although QTL effects and polygenic effects both appear in the linear models, they are not the parameters of interest. Because of this, they do not appear in the log likelihood function. If these genetic effects were observed, the genetic variance components would have been estimated easily using a simple formula of variance. This reminds us the EM algorithm, in which we can take advantage of the simplicity of the variance formula by treating both the QTL effects and the polygenic effects as missing values. In this section, we introduce the EM algorithm for estimating variance components. Derivation of the method requires some complicated matrix algebra and thus will not be provided here. We will simply show the final iterative equation for each parameter. We start with all parameter values at iteration t, denoted by $\theta^{(t)} = \{\mu^{(t)}, \sigma_a^{2(t)}, \sigma_\gamma^{2(t)}, \sigma^{2(t)}\}$, and then proceed to update each parameter in sequence conditional on $\theta^{(t)}$. Let us denote V_j evaluated at $\theta^{(t)}$ by

$$V_j^{(t)} = \Pi_j \sigma_\gamma^{2(t)} + A_j \sigma_a^{2(t)} + I\sigma^{2(t)} \tag{13.23}$$

We now introduce the EM algorithm starting with $\theta = \theta^{(t)}$. If the QTL and the polygenic effects were observed for all individuals, the MLE's of the QTL and polygenic variances would be calculated using

$$\sigma_\gamma^2 = \frac{1}{2n}\sum_{j=1}^{n} \gamma_j^T \Pi_j^{-1} \gamma_j \tag{13.24}$$

and

$$\sigma_a^2 = \frac{1}{2n}\sum_{j=1}^{n} a_j^T A_j^{-1} a_j \tag{13.25}$$

respectively. The environmental error variance would be obtained using

$$\sigma^2 = \frac{1}{2n}\sum_{j=1}^{n} y_j^T(y_j - \mu - a_j - \gamma_j) \tag{13.26}$$

The EM algorithm takes advantage of these simple equations by replacing the terms involving a_j and γ_j by their expectations. However, to calculate the expectations, we need to use the phenotypic values and the current values of the parameters. Such expectations are called the posterior expectations. The EM algorithm starts with calculation of these expectations (the E-step) and then uses the simple equations to update the variance components (the M-step). In the M-step, the parameters at the $(t+1)$th iteration are calculated using the following iteration equations.

$$\mu^{(t+1)} = \left[\sum_{i=1}^{n} 1^{\mathrm{T}}(V_j^{(t)})^{-1}1\right]^{-1}\left[\sum_{j=1}^{n} 1^{\mathrm{T}}(V_j^{(t)})^{-1}y_j\right] \tag{13.27}$$

$$\begin{aligned}\sigma_\gamma^{2(t+1)} &= \frac{1}{2n}\sum_{j=1}^{n}\mathrm{E}(\gamma_j^T \mathit{\Pi}_j^{-1}\gamma_j)\\ &= \frac{1}{2n}\sum_{j=1}^{n}\left\{\mathrm{E}(\gamma_j^T)\mathit{\Pi}_j^{-1}\mathrm{E}(\gamma_j) + \mathrm{tr}\left[\mathit{\Pi}_j^{-1}\mathrm{var}(\gamma_j)\right]\right\}\end{aligned} \tag{13.28}$$

$$\begin{aligned}\sigma_a^{2(t+1)} &= \frac{1}{2n}\sum_{j=1}^{n}\mathrm{E}(a_j^T A_j^{-1}a_j)\\ &= \frac{1}{2n}\sum_{j=1}^{n}\left\{\mathrm{E}(a_j^T)A_j^{-1}\mathrm{E}(a_j) + \mathrm{tr}\left[A_j^{-1}\mathrm{var}(a_j)\right]\right\}\end{aligned} \tag{13.29}$$

$$\sigma^{2(t+1)} = \frac{1}{2n}\sum_{j=1}^{n} y_j^T\left[y_j - 1\mu^{(t+1)} - \mathrm{E}(a_j) - \mathrm{E}(\gamma_j)\right] \tag{13.30}$$

These equations are the steps required in the M-step. We can see that these equations contain the expectations and variances of QTL effects and the polygenic effects. Calculating these expectations and variances is called the E-step. For the QTL effect, we have

$$\begin{aligned}\mathrm{E}(\gamma_j) &= \mathit{\Pi}_j\sigma_\gamma^2 V_j^{-1}(y_j - \mu)\\ \mathrm{var}(\gamma_j) &= \mathit{\Pi}_j\sigma_\gamma^2(I - V_j^{-1}\mathit{\Pi}_j\sigma_\gamma^2)\end{aligned} \tag{13.31}$$

The expectation and variance for the polygenic effect are

$$\begin{aligned}\mathrm{E}(a_j) &= A_j\sigma_a^2 V_j^{-1}(y_j - \mu)\\ \mathrm{var}(a_j) &= A_j\sigma_a^2(I - V_j^{-1}A_j\sigma_a^2)\end{aligned} \tag{13.32}$$

13.4.2 EM algorithm under singular value decomposition

The EM algorithm is not always guaranteed to work because occasionally Π_j^{-1} may not exist. For example, if the two siblings share 2 IBD alleles at the QTL, $\pi_j = 1$, then Π_j will be singular. To avoid this problem, we can take a different approach. This approach requires a linear transformation of the genetic effects using singular value decomposition. Let us define Γ_j as an upper triangular matrix so that $\Gamma_j^T \Gamma_j = \Pi_j$. We call Γ_j the Cholesky decomposition. This decomposition exists even if Π_j is positive semidefinite (not necessarily positive definite). Using the Cholesky decomposition, we can avoid inverting Π_j because the inverse of Π_j does not exist if Π_j is not positive definite. Similarly, we can take the Cholesky decomposition for the additive relationship matrix, denoted by H_j, i.e., $H_j^T H_j = A_j$. For two siblings per family, both Γ_j and H_j have explicit expressions,

$$\Gamma_j = \begin{bmatrix} 1 & \pi_j \\ 0 & \sqrt{1-\pi_j^2} \end{bmatrix} \tag{13.33}$$

and

$$H_j = \begin{bmatrix} 1 & 0.5 \\ 0 & \sqrt{1-0.5^2} \end{bmatrix} \tag{13.34}$$

We now rewrite the linear model as

$$y_j = 1\mu + H_j^T a_j^* + \Gamma_j^T \gamma_j^* + \epsilon_j \tag{13.35}$$

where $a_j^* \sim N(0, I\sigma_a^2)$ and $\gamma_j^* \sim N(0, I\sigma_\gamma^2)$. Note the difference between a_j and a_j^* and the difference between γ_j and γ_j^*. After the transformation, $a_j = H_j^T a_j^*$, and $\gamma_j = \Gamma_j^T \gamma_j^*$, we now deal with vectors with independent elements. The expectation and variance-covariance matrix remain the same as given before, i.e.,

$$\mathrm{E}(y_j) = 1\mu \tag{13.36}$$

and

$$\begin{aligned} \mathrm{var}(y_j) &= H_j^T \mathrm{var}(a_j^*) H_j + \Gamma_j^T \mathrm{var}(\gamma_j^*) \Gamma_j + I\sigma^2 \\ &= H_j^T H_j \sigma_a^2 + \Gamma_j^T \Gamma_j \sigma_\gamma^2 + I\sigma^2 \\ &= A_j \sigma_a^2 + \Pi_j \sigma_\gamma^2 + I\sigma^2 \end{aligned} \tag{13.37}$$

The EM algorithm after the singular value decomposition consists of the following steps. For the maximization step, we have

$$\mu^{(t+1)} = \left[\sum_{i=1}^{n} 1^{\mathrm{T}} (V_j^{(t)})^{-1} 1 \right]^{-1} \left[\sum_{j=1}^{n} 1^{\mathrm{T}} (V_j^{(t)})^{-1} y_j \right] \tag{13.38}$$

$$\begin{aligned}\sigma_\gamma^{2(t+1)} &= \frac{1}{2n}\sum_{j=1}^{n}\mathrm{E}(\gamma_j^{*T}\gamma_j^*) \\ &= \frac{1}{2n}\sum_{j=1}^{n}\left\{\mathrm{E}(\gamma_j^{*T})\mathrm{E}(\gamma_j^*) + \mathrm{tr}\left[\mathrm{var}(\gamma_j^*)\right]\right\} \qquad (13.39)\end{aligned}$$

$$\begin{aligned}\sigma_a^{2(t+1)} &= \frac{1}{2n}\sum_{j=1}^{n}\mathrm{E}(a_j^{*T}a_j^*) \\ &= \frac{1}{2n}\sum_{j=1}^{n}\left\{\mathrm{E}(a_j^{*T})\mathrm{E}(a_j^*) + \mathrm{tr}\left[\mathrm{var}(a_j^*)\right]\right\} \qquad (13.40)\end{aligned}$$

$$\sigma^{2(t+1)} = \frac{1}{2n}\sum_{j=1}^{n} y_j^T\left[y_j - 1\mu^{(t+1)} - H_j^T\mathrm{E}(a_j^*) - \Gamma_j^T\mathrm{E}(\gamma_j^*)\right] \qquad (13.41)$$

For the expectation step, we have

$$\begin{aligned}\mathrm{E}(a_j^*) &= H_j\sigma_a^2V_j^{-1}(y_j - 1\mu) \\ \mathrm{var}(a_j^*) &= \sigma_a^2(I - H_jV_j^{-1}H_j^T\sigma_a^2) \qquad (13.42)\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(\gamma_j^*) &= \Gamma_j\sigma_\gamma^2V_j^{-1}(y_j - 1\mu) \\ \mathrm{var}(\gamma_j^*) &= \sigma_\gamma^2(I - \Gamma_jV_j^{-1}\Gamma_j^T\sigma_\gamma^2) \qquad (13.43)\end{aligned}$$

Clearly, we have avoided using A_j^{-1} and Γ_j^{-1} under the singular value decomposition approach. This approach is general because it works regardless whether Π_j is singular or not.

13.4.3 Multiple siblings

The sib-pair approach we have discussed so far can only handle two siblings per family. If a full-sib family contains more than two siblings, the extra siblings cannot be used. Additional information from the extra siblings will be wasted. Extension of the existing method to multiple siblings is quite straightforward under the general framework of random model methodology. This extension is different from the sib-pair regression approach proposed by Haseman and Elston (1972), which cannot be extended to multiple siblings. Let n_j be the number of siblings for the jth full-sib family. Assume that there are n full-sib families in the mapping population, the total sample size is $N = \sum_{j=1}^{n} n_j$. The sib-pair approach is a special case where

$n_j = 2, \forall j = 1, \ldots, n$. The only difference between multiple siblings and the sib-pair situations under the random model approach is the dimensionality of the matrices. For the sib-pair method, all matrices have a dimensionality of 2×2. For multiple siblings, vectors y_j, a_j, γ_j and ϵ_j all have a dimensionality of $n_j \times 1$ and matrices A_j and $\mathit{\Pi}_j$ have a dimensionality of $n_j \times n_j$. For example, for three siblings per family, the A_j and $\mathit{\Pi}_j$ matrices are

$$A_j = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \tag{13.44}$$

and

$$\mathit{\Pi}_j = \begin{bmatrix} 1 & \pi_{12} & \pi_{13} \\ \pi_{12} & 1 & \pi_{23} \\ \pi_{13} & \pi_{23} & 1 \end{bmatrix} \tag{13.45}$$

where π_{23} is the IBD value between sibs 2 and 3. The corresponding changes from sib-pair to multiple siblings in the EM algorithms occur in the following three places,

$$\begin{aligned} \sigma_\gamma^{2(t+1)} &= \frac{1}{N} \sum_{j=1}^{n} \mathrm{E}(\gamma_j^{*T} \gamma_j^*) \\ &= \frac{1}{N} \sum_{j=1}^{n} \left\{ \mathrm{E}(\gamma_j^{*T}) \mathrm{E}(\gamma_j^*) + \mathrm{tr}\left[\mathrm{var}(\gamma_j^*) \right] \right\} \end{aligned} \tag{13.46}$$

$$\begin{aligned} \sigma_a^{2(t+1)} &= \frac{1}{N} \sum_{j=1}^{n} \mathrm{E}(a_j^{*T} a_j^*) \\ &= \frac{1}{N} \sum_{j=1}^{n} \left\{ \mathrm{E}(a_j^{*T}) \mathrm{E}(a_j^*) + \mathrm{tr}\left[\mathrm{var}(a_j^*) \right] \right\} \end{aligned} \tag{13.47}$$

$$\sigma^{2(t+1)} = \frac{1}{N} \sum_{j=1}^{n} y_j^T \left[y_j - 1\mu^{(t+1)} - H_j^T \mathrm{E}(a_j^*) - \mathit{\Gamma}_j^T \mathrm{E}(\gamma_j^*) \right] \tag{13.48}$$

If the $\mathit{\Pi}_j$ matrix is the true IBD matrix for the n_j siblings, $\mathit{\Pi}_j$ is positive semidefinite. The EM algorithm under the singular value decomposition will works equally well as the sib-pair method. However, the $\mathit{\Pi}_j$ matrix for a QTL is always estimated using marker information and the method of estimation for $\mathit{\Pi}_j$ cannot guarantee the positive semidefinite property. Therefore, the EM algorithm may not always work. As a result, the simplex method is highly recommended because the method directly evaluates the log likelihood function, which only requires a positive definite $V_j = A_j \sigma_a^2 + \mathit{\Pi}_j \sigma_\gamma + I\sigma^2$. The property of positive definite for V_j is guaranteed as long as σ^2 is not too small. If Π_j is positive semidefinite, V_j is always positive definite because $\sigma^2 > 0$ always holds. Another reason for using the simplex method is that the EM algorithm is sensitive to the initial values of the parameters while the simplex is robust to the initial values of the parameters. For example, using the simplex method, the initial values for $\sigma_a^2 = \sigma_\gamma^2 = 0$ usually work very well but the EM algorithm cannot take such initial values. Unfortunately, programming the simplex algorithm is not as easy as writing the program code for the EM algorithm.

13.5 Estimating the IBD value for a marker

When a marker is not fully informative, the IBD values shared by siblings are ambiguous. We need a special algorithm to estimate the IBD values. We assume that every individual in a nuclear family has been genotyped for the marker, including the parents. If parents are not genotyped, their genotypes must be inferred first from the genotypes of the progeny. However, the method becomes complicated and will not be dealt with in this chapter. There are four possible allelic sharing states for a pair of siblings. The probabilities of the four states are denoted by $\Pr(s, d)$, where $s = \{1, 0\}$ indicate whether or not the siblings share their paternal alleles (alleles from the sire). If they share the paternal alleles, we let $s = 1$; otherwise, $s = 0$. Similarly, $d = \{1, 0\}$ indicates whether or not the siblings share the maternal alleles (alleles from the dam). Once we have the four probabilities, the estimated IBD value for the siblings is

$$\begin{aligned}\hat{\pi} &= \frac{1}{2}\{2\Pr(1,1) + 1[\Pr(1,0) + \Pr(0,1)]\} \\ &= \Pr(1,1) + \frac{1}{2}[\Pr(1,0) + \Pr(0,1)]. \end{aligned} \tag{13.49}$$

Therefore, the problem of estimating the IBD value becomes a problem of calculating the four probabilities of allelic sharing states. We first convert the numbers of shared alleles for the 16 possible sib-pair combinations listed in Table 13.1 into a 4×4 matrix,

$$S = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}, \tag{13.50}$$

where S_{ij} is the ith row and the jth column of matrix S. We then construct another 4×4 matrix, denoted by P, to represent the probabilities of the 16 possible sibling-pair combinations. To find matrix P, we need to find two 4×1 vectors, one for each sibling. Let $U = \{U_1, U_2, U_3, U_4\}$ be the vector for sib one and $V = \{V_1, V_2, V_3, V_4\}$ be the vector for sib two. Recall that each progeny can take one of four possible genotype configurations. Each element of vector U is the probability that the progeny takes that particular genotype. For example, if the sire and the dam of the nuclear family have genotypes of A_1A_2 and A_3A_3, respectively, then the four genotype configurations are A_1A_3, A_1A_3, A_2A_3 and A_2A_3. If sib one has a genotype of A_2A_3, it matches the third and the forth genotype configurations. Therefore, $U = \{0, 0, \frac{1}{2}, \frac{1}{2}\}$. If sib two has a genotype of A_1A_3, then $V = \{\frac{1}{2}, \frac{1}{2}, 0, 0\}$. Both U and V are defined as column vectors. The 4×4 matrix P is defined as $P = UV^T$, i.e.,

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \\ P_{41} & P_{42} & P_{43} & P_{44} \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{bmatrix} \begin{bmatrix} V_1 & V_2 & V_3 & V_4 \end{bmatrix}, \tag{13.51}$$

where $P_{ij} = U_iV_j$ is the ith row and the jth column of matrix P. In the above example, $U = \{0, 0, \frac{1}{2}, \frac{1}{2}\}$ and $V = \{\frac{1}{2}, \frac{1}{2}, 0, 0\}$, and therefore,

$$P = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 \end{bmatrix}. \tag{13.52}$$

The two matrices, S and P, contain all information for estimating the IBD value shared by the siblings. The probability that the siblings share both alleles IBD is

$$\Pr(1, 1) = P_{11} + P_{22} + P_{33} + P_{44}, \tag{13.53}$$

which is the sum of the diagonal elements of matrix P, called the trace of matrix P. The probability that the siblings share one IBD allele from the sire is

$$\Pr(1, 0) = P_{12} + P_{21} + P_{34} + P_{43}. \tag{13.54}$$

The probability that the siblings share one IBD allele from the dam is

$$\Pr(0, 1) = P_{13} + P_{24} + P_{31} + P_{42}. \tag{13.55}$$

Although $\Pr(0, 0)$ is not required in estimating the IBD value, it is the probability that the siblings share no IBD allele. It is calculated using

$$\Pr(0, 0) = P_{14} + P_{23} + P_{32} + P_{41}, \tag{13.56}$$

which is the sum of the minor diagonal elements of matrix P. In the above example, the estimated IBD value is

$$\hat{\pi} = \Pr(1, 1) + \frac{1}{2}[\Pr(1, 0) + \Pr(0, 1)] = 0 + \frac{1}{2}(0 + \frac{1}{2}) = \frac{1}{4}. \tag{13.57}$$

In fact, we do not need to calculate $\Pr(s, d)$ for estimating the IBD value for a marker. The IBD value may be simply estimated using

$$\hat{\pi} = \frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4} S_{ij}P_{ij}. \tag{13.58}$$

In matrix notation, $\hat{\pi} = \frac{1}{2}[J'(S\#P)J]$, where $J = \{1, 1, 1, 1\}$ is a column vector of unity and the symbol # represents element-wise matrix multiplication. The reason that we develop $\Pr(s, d)$ is to infer the IBD value of a putative QTL using $\Pr(s, d)$ of the markers.

13.6 Multipoint method for estimating the IBD value

Multipoint method for QTL mapping is more important under the IBD-based random model framework than that under the fixed model framework using a single line cross. The reason is that fully informative markers are rare in a random mating population, in which parents are usually randomly sampled. This is in contrast to a line crossing experiment, in which parents that initiate the cross are inbred and often selected based on maximum diversity in both marker and phenotype distributions. Consider five loci in the order of ABCDE, where locus C is a putative QTL and the other loci are markers. Let r_{AB}, r_{BC}, r_{CD} and r_{DE} be the recombination fractions between pairs of consecutive loci. The purpose of the multipoint analysis is to estimate the IBD value shared by siblings for locus C given marker information for loci A, B, D and E. It is not appropriate to use the estimated IBD values for markers to estimate the IBD value of the QTL. Instead, we should use the probabilities of the four possible IBD sharing states of the markers to estimated the four possible IBD share states of the QTL, from which the estimated IBD value of the QTL can be obtained.

Let us denote $\Pr(s, d)$ for a marker locus, say locus A, by $P_A(s, d)$. We now define $P_A = \{P_A(1, 1), P_A(1, 0), P_A(0, 1), P_A(0, 0)\}$ as a 4×1 vector for the probabilities of the four IBD sharing states of locus A. Because this locus is a marker, the four probabilities are calculated based on information of the marker genotypes. Let us define $D_A = \mathrm{diag}\{P_A(1, 1), P_A(1, 0), P_A(0, 1), P_A(0, 0)\}$ as a diagonal matrix for locus A. Similar diagonal matrices are defined for all other makers, i.e., D_B, D_D and D_E. These diagonal matrices represent the marker data, from which the multipoint estimate of the IBD of locus C is obtained. The transition matrix between loci A and B is

$$T_{AB} = \begin{bmatrix} \psi_{AB}^2 & (1-\psi_{AB})\psi_{AB} & \psi_{AB}(1-\psi_{AB}) & (1-\psi_{AB})^2 \\ (1-\psi_{AB})\psi_{AB} & \psi_{AB}^2 & (1-\psi_{AB})^2 & \psi_{AB}(1-\psi_{AB}) \\ \psi_{AB}(1-\psi_{AB}) & (1-\psi_{AB})^2 & \psi_{AB}^2 & (1-\psi_{AB})\psi_{AB} \\ (1-\psi_{AB})^2 & \psi_{AB}(1-\psi_{AB}) & (1-\psi_{AB})\psi_{AB} & \psi_{AB}^2 \end{bmatrix}, \tag{13.59}$$

where $\psi_{AB} = r_{AB}^2 + (1 - r_{AB})^2$. The four conditional probabilities of IBD sharing states for locus C are calculated using the following equations,

$$
\begin{aligned}
\Pr(1,1) &= \frac{J'D_AT_{AB}D_BT_{BC}D_{(1)}T_{CD}D_DT_{DE}D_EJ}{\sum_{k=1}^4 J'D_AT_{AB}D_BT_{BC}D_{(k)}T_{CD}D_DT_{DE}D_EJ}, \\
\Pr(1,0) &= \frac{J'D_AT_{AB}D_BT_{BC}D_{(2)}T_{CD}D_DT_{DE}D_EJ}{\sum_{k=1}^4 J'D_AT_{AB}D_BT_{BC}D_{(k)}T_{CD}D_DT_{DE}D_EJ}, \\
\Pr(0,1) &= \frac{J'D_AT_{AB}D_BT_{BC}D_{(3)}T_{CD}D_DT_{DE}D_EJ}{\sum_{k=1}^4 J'D_AT_{AB}D_BT_{BC}D_{(k)}T_{CD}D_DT_{DE}D_EJ}, \\
\Pr(0,0) &= \frac{J'D_AT_{AB}D_BT_{BC}D_{(4)}T_{CD}D_DT_{DE}D_EJ}{\sum_{k=1}^4 J'D_AT_{AB}D_BT_{BC}D_{(k)}T_{CD}D_DT_{DE}D_EJ}.
\end{aligned} \tag{13.60}
$$

The denominator can be simplified into $J'D_AT_{AB}D_BT_{BC}T_{CD}D_DT_{DE}D_EJ$ because $\sum_{k=1}^4 D_{(k)} = I$. Let π be the IBD value shared by the siblings at locus C. The multipoint estimate of π using marker information is

$$\hat{\pi} = \Pr(1,1) + \frac{1}{2}[\Pr(1,0) + \Pr(0,1)]. \tag{13.61}$$

For a family with n_j siblings, there are $\frac{1}{2}n_j(n_j-1)$ sib pairs. Therefore, for each locus, we need to calculate $\frac{1}{2}n_j(n_j-1)$ IBD values, one per sib pair. These IBD values are calculated one sib pair at a time. Because they are not calculated jointly, the IBD matrix obtained from these estimated IBD values may not be positive definite or positive semidefinite. As a result, the EM algorithm proposed early may not always work because it requires the positive semidefinite property.

Before we proceed to the next section, it is worthy to mention a possible extension of the additive variance component model to include the dominance variance. The modified model looks like

$$y_j = 1\mu + a_j + \gamma_j + \xi_j + \epsilon_j \tag{13.62}$$

where $\xi_j \sim N(0, \varDelta_j\sigma^2_\xi)$ is a vector of dominance effects, σ^2_ξ is the dominance variance and $\varDelta_j$ is the dominance IBD matrix for the siblings in the jth family. The dominance IBD matrix is

$$\varDelta_j = \begin{bmatrix} 1 & \delta_j \\ \delta_j & 1 \end{bmatrix} \tag{13.63}$$

where δ_j is a binary variable indicating the event of the siblings sharing IBD genotype. If the two siblings share both IBD alleles (i.e., the same IBD genotype), then $\delta_j = 1$; otherwise, $\delta = 0$. The estimated δ_j is

$$\hat{\delta}_j = \Pr(1,1) \tag{13.64}$$

the probability that the siblings share both IBD alleles.

13.7 Genome scanning and hypothesis tests

For each putative QTL position, we calculate the estimated IBD value for each sib-pair, denoted by $\hat{\pi}_j$ for the jth family. We then construct the IBD matrix

$$\hat{\Pi}_j = \begin{bmatrix} 1 & \hat{\pi}_j \\ \hat{\pi}_j & 1 \end{bmatrix} \tag{13.65}$$

for the jth family. These estimated IBD matrices are used in place of the true IBD matrices for QTL mapping.

The null hypothesis is $H_0 : \sigma_\gamma^2 = 0$. Again, a likelihood ratio test statistic is adopted here for the hypothesis test. The likelihood ratio test statistic is

$$\lambda = -2 \left[L_0(\hat{\hat{\mu}}, \hat{\hat{\sigma}}_a^2, \hat{\hat{\sigma}}^2) - L_1(\hat{\mu}, \hat{\sigma}_\gamma^2, \hat{\sigma}_a^2, \hat{\sigma}^2) \right] . \tag{13.66}$$

Note that the MLE of parameters under the null model differs from the MLE of parameters under the full model by wearing double hats.

The genome is scanned for every putative position with a one or two centiMorgan increment. The test statistic will form a profile along the genome. The estimated QTL position takes the location of the genome where the peak of the test statistic profile occurs, provided that the peak is higher than a predetermined critical level. The critical value is usually obtained with permutation test (Churchill and Doerge, 1994) that has been described in Chapter 7.

The random model approach to interval mapping utilizes a single QTL model. When multiple QTL are present, the QTL variance will be over estimated. This bias can be eliminated through fitting a multiple QTL model, which will be presented in the next section. If the multiple QTL are not tightly linked, the interval mapping approach still provides reasonable estimates for the QTL variances. Multiple peaks of the test statistic profile indicate the presence of multiple QTL. The positions of the genome where the multiple peaks occur represent estimated positions of the multiple QTL.

One difference between the random model approach and the fixed model approach to interval mapping is that the variances of QTL in other chromosomes will be absorbed by the polygenic variance in the random model analysis rather than by the residual variance in the fixed model analysis. Therefore, the random model approach of interval mapping handles multiple QTL better than the fixed model approach.

13.8 Multiple QTL model

Assume that there are p QTL in the model, the multiple QTL model is

$$y_j = \mu + \sum_{k=1}^{p} \gamma_{jk} + \epsilon_j \tag{13.67}$$

where $\gamma_{jk} \sim N(0, \mathit{\Pi}_{jk}\sigma_k^2)$ is an $n_j \times 1$ vector for the kth QTL effects, $\mathit{\Pi}_{jk}$ is an $n_j \times n_j$ IBD matrix for the kth QTL and σ_k^2 is the genetic variance for the kth QTL. Previously, we had a polygenic effect in the single QTL model to absorb effects of all other QTL not included in the model. With the multiple QTL model, the polygenic effect has disappeared (not needed). The expectation of the model remains the same, i.e., $\mathrm{E}(y_j) = \mu$. The variance matrix is

$$\mathrm{var}(y_j) = V_j = \sum_{k=1}^{p} \mathit{\Pi}_{jk}\sigma_k^2 + I\sigma^2 \tag{13.68}$$

The likelihood function and the maximum likelihood estimation of the parameters can follow what described previously in the single QTL model.

The complications with the multiple QTL model come from the genome locations of the QTL. Under the single QTL model, we can scan the genome to find the peaks of the test statistic profile. Under the multiple QTL model, the multiple dimensional scanning is hard to implement. Therefore, an entirely different method, called the Bayesian method, is required to simultaneously search for multiple QTL. The Bayesian method will be introduced in Chapter 15. An ad hoc method is discussed here for parameter estimation when the QTL positions are assumed to be known. We can put a finite number of QTL that evenly cover the entire genome, and hope that some proposed QTL will sit nearby a true QTL. We may put one QTL in every d cM, say $d = 20$, of the entire genome. This may end up with too many proposed QTL (more than the actual number of QTL). Some of the proposed QTL may be nearby one or more true QTL and thus they can absorb the QTL effects. These proposed QTL are useful and must be included in the model. Majority of the proposed QTL may not be able to pick up any information at all because they may not be close to any QTL. These proposed QTL are fake ones and, theoretically, should be excluded from the model. Therefore, model selection appears to be necessary to delete those false QTL. However, the random model methodology introduced in this chapter is special in the sense that it allows a model to include many proposed QTL without encountering much technical problem. The number of proposed QTL can even be larger than the sample size. This is clearly different from the fixed effect linear model where the number of model effects must be substantially smaller than the sample size to be able to produce any meaningful result.

Under the multiple QTL model, the parameter vector is $\theta = \{\mu, \sigma_1^2, \ldots, \sigma_p^2, \sigma^2\}$ with a dimensionality $(p + 2) \times 1$. The high dimensionality of θ is a serous problem with regard to choosing the appropriate algorithm for parameter estimation. The fact that majority of the proposed QTL are fake and thus their estimated variance components should be near zero. Since the EM algorithm does not allow the use of zero as the initial value for a variance component, it is not an option. The simplex algorithm, although allows the use of zero as initial values, can only handle a few variance components, say $p = 20$ or less. The easiest and simplest algorithm is the sequential search with one component at a time (Han and Xu, 2010). We search for the optimal value of one component, conditional on the values of all other components. When every variance component is sequentially optimized, another round of search begins. The iteration continues until a certain criterion of convergence is satisfied. The sequential search algorithm is summarized as follows Han and Xu (2010).

1. Set $t = 0$ and initialize $\theta = \theta^{(t)} = \{\mu^{(t)}, \sigma_1^{2(t)}, \cdots, \sigma_p^{2(t)}, \sigma^{2(t)}\}$, where

$$\begin{aligned}\mu^{(0)} &= \bar{y}\\ \sigma^{2(0)} &= \frac{1}{N}\sum_{j=1}^{n}(y_j - 1\bar{y})^T(y_j - 1\bar{y})\\ \sigma_k^{2(0)} &= 0, \forall k = 1, \cdots, p\end{aligned} \tag{13.69}$$

2. Define $V_j^{(t)} = \sum_{j=1}^{n} \Pi_{jk}\sigma_k^{2(t)} + I\sigma^{2(t)}$ and update μ by maximizing

$$L(\mu) = -\frac{1}{2}\sum_{j=1}^{n}\ln\left|V_j^{(t)}\right| - \frac{1}{2}\sum_{j=1}^{n}(y_j - 1\mu)^T\left(V_j^{(t)}\right)^{-1}(y_j - 1\mu) \tag{13.70}$$

3. Define $V_j(\sigma_k^2) = V_j^{(t)} - \Pi_{jk}(\sigma_k^{2(t)} - \sigma_k^2)$ and update σ_k^2 by maximizing

$$\begin{aligned}L(\sigma_k^2) = &-\frac{1}{2}\sum_{j=1}^{n}\ln\left|V_j(\sigma_k^2)\right|\\ &-\frac{1}{2}\sum_{j=1}^{n}\left(y_j - 1\mu^{(t)}\right)^T V_j^{-1}(\sigma_k^2)\left(y_j - 1\mu^{(t)}\right)\end{aligned} \tag{13.71}$$

   for all $k = 1, \ldots, p$
4. Define $V_j(\sigma^2) = V_j^{(t)} - I(\sigma^{2(t)} - \sigma^2)$ and update σ^2 by maximizing

$$\begin{aligned}L(\sigma^2) = &-\frac{1}{2}\sum_{j=1}^{n}\ln\left|V_j(\sigma^2)\right|\\ &-\frac{1}{2}\sum_{j=1}^{n}\left(y_j - 1\mu^{(t)}\right)^T V_j^{-1}(\sigma^2)\left(y_j - 1\mu^{(t)}\right)\end{aligned} \tag{13.72}$$

5. Set $t = t + 1$ and repeat from Steps 2 to 4 until a certain criterion of convergence is satisfied.

The solution for μ in step 2 is explicit, as shown below,

$$\mu = \left[\sum_{j=1}^{n} 1^T V_j^{(t)} 1\right]^{-1}\left[\sum_{j=1}^{n} 1^T V_j^{(t)} y_j\right] \tag{13.73}$$

However, the solutions for σ_k^2 within Step 3 and σ^2 within Step 4 must be obtained via some numerical algorithm. Since the dimensionality is low (a single variable), any algorithm will work well. The simplex algorithm, although designed for multiple variable, works very well for a single variable.

Although the multiple QTL model implemented via the sequential search algorithm can handle an extremely large number of QTL, most of the estimated variance components will be close to zero. A QTL with a zero variance component is equivalent to being excluded from the model. It is still a model selection strategy but only conducted implicitly rather than explicitly. The caveat of the ad hoc sequential search is that the estimated residual variance, σ^2, approaches to zero as p grows. This, however, does not affect the relative contribution of each identified QTL because the relative contribution of the kth QTL is expressed as

$$h_k^2 = \frac{\sigma_k^2}{\sum_{k'}^{p} \sigma_{k'}^2 + \sigma^2} \tag{13.74}$$

13.9 Complex pedigree analysis

The random model approach to QTL mapping can also be extended to handle large families with complicated relationships. Such families are called pedigrees. A pedigree may consist of relatives with arbitrary relationships and the members many expand for several generations. The model and algorithm remain the same as the nuclear family analysis except that methods for calculating the IBD matrix for a putative QTL are much more involved. The multiple regression method developed by Fulker and Cardon (1994) for sibpair analysis and later extended to pedigree analysis by Almasy and Blangero (1998) can be adopted. The regression method for calculating the IBD matrix, however, is not optimal. The Markov chain Monte Carlo method for calculating the IBD matrix implemented in the software package Lokie (Heath, 1997) and the program named in SimWalk2 (Sobel et al., 2001) are recommended.