Chapter 6

Likelihoods on Pedigrees

6.1 The Baum algorithm and "Peeling"

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

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

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

Now define

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

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

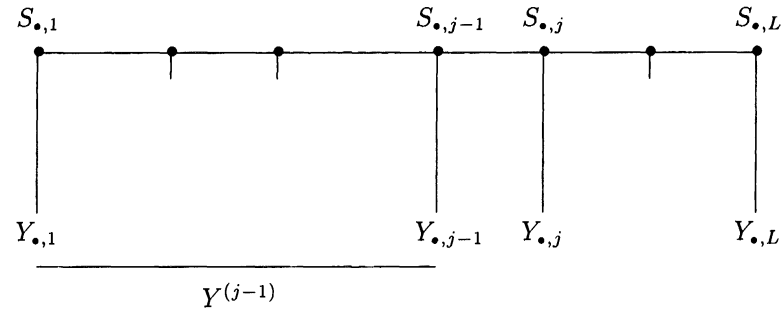

FIGURE 6.1. *The conditional independence structure of data, in the absence of genetic interference*

Thus,

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

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

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

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

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

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

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

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

it is now more convenient to define the joint probability

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

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

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

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

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

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

6.2 Exact likelihoods for multiple markers

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

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

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

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

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

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

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

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

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

6.3 Computations on large but simple pedigrees

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

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

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

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

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

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

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

6.4 Example of peeling a zero-loop pedigree

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

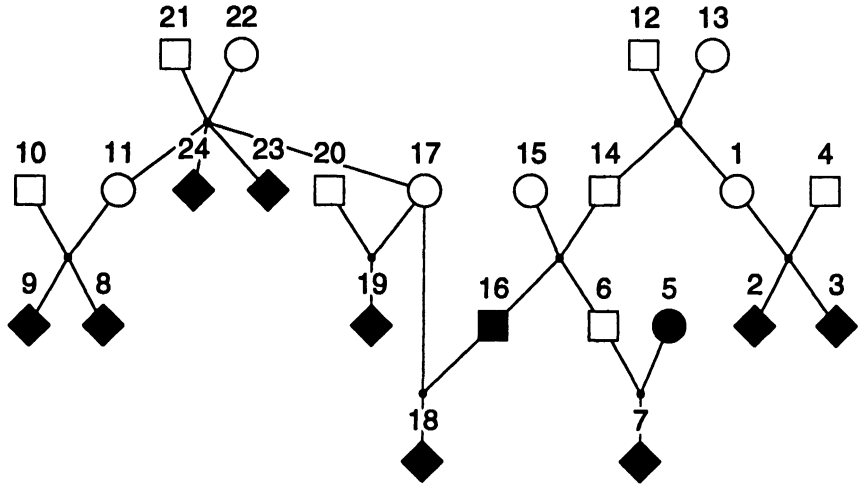

FIGURE 6.2. *Pedigree without loops. Shaded individuals are those for whom phenotypic data are assumed to be available*

Starting with the family to the right, we may compute

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

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

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and finally the overall likelihood is

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

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

6.5 Computations on complex pedigrees

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

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

First we peel the final individual 531:

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

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

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

and then over 334 and her founder parent 235 to obtain

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

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

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

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

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

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

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

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

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

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

6.6 Models with Gaussian random effects

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

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

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

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

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

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

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

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