Chapter 7

Monte Carlo Estimates on Pedigrees

7.1 Baum algorithm for conditional probabilities

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

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

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

We define two functions

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

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

7.2 An EM algorithm for map estimation

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

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

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

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

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

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

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

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

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

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

7.3 Importance sampling for likelihoods

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

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

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

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

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

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

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

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

provided (equation (3.13))

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

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

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

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

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

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

7.4 Risk probabilities and reverse peeling

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

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

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

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

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

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

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

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

while in proceeding down, relative to individual i,

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

so that

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

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

7.5 Elods and SIMLINK

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

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

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

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

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

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

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

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

7.6 Sequential imputation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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