Chapter 9

Likelihood Ratios for Genetic Analysis

9.1 Monte Carlo likelihood ratio estimation

The MCMC methods of Chapter 8 provide methods for obtaining realizations from $P_\theta(\mathbf{X} \mid \mathbf{Y})$, the probability distribution of latent variables $\mathbf{X}$ conditional on data $\mathbf{Y}$ under a model indexed by parameters θ. In this chapter, we discuss methods of using such realizations in Monte Carlo methods for linkage and segregation analysis, focusing on likelihood methods.

Recall again (equation (7.8)) that, for phenotypic data $\mathbf{Y}$,

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

where latent variables $\mathbf{X}$ are genotypes $\mathbf{G}$ or meiosis indicators $\mathbf{S}$. We again use θ to denote the general set of parameters of a genetic model. These include the recombination or gene location parameters. From equation (7.12), efficient Monte Carlo estimation of $L(\theta)$ will result from sampling from a distribution $P^*(\mathbf{X})$ close to proportional to the joint probability $P_\theta(\mathbf{X}, \mathbf{Y})$:

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

One possible choice is thus to simulate, by the methods of Chapter 8, not from $P_\theta(\mathbf{X} \mid \mathbf{Y})$ but from $P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})$, where $\theta_0 \approx \theta$. Then

$$\begin{aligned}
P_\theta(\mathbf{Y}) &= \sum_{\mathbf{X}} P_\theta(\mathbf{X}, \mathbf{Y}) \;=\; \sum_{\mathbf{X}} \frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})} P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \\
&= \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X},\ \mathbf{Y})}{P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})} \;\middle|\; \mathbf{Y} \right) \;=\; P_{\theta_0}(\mathbf{Y})\, \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_{\theta_0}(\mathbf{X}, \mathbf{Y})} \;\middle|\; \mathbf{Y} \right).
\end{aligned}$$

Hence in genetic analysis, or in any missing-data context, we have the key formula of Thompson and Guo (1991)

$$\text{(9.1)} \qquad \frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X},\mathbf{Y})}{P_{\theta_0}(\mathbf{X},\mathbf{Y})} \;\middle|\; \mathbf{Y}\right).$$

In this expectation, $\mathbf{X}$ is the random variable, $\mathbf{Y}$ is fixed. The distribution of $\mathbf{X}$ is $P_{\theta_0}(\cdot|\mathbf{Y})$. If $\mathbf{X}^{(\tau)}$, $\tau = 1, \ldots, N$, are realized from this distribution then the likelihood ratio can be estimated by

$$\frac{1}{N}\sum_{\tau=1}^{N}\left(\frac{P_\theta(\mathbf{X}^{(\tau)},\mathbf{Y})}{P_{\theta_0}(\mathbf{X}^{(\tau)},\mathbf{Y})}\right).$$

In section 8.1 we saw how MCMC can be used to realize $\mathbf{X}$ from $P_{\theta_0}(\cdot|\mathbf{Y})$.

Simulation at a single model θ_0 provides an estimate of the relative likelihood $L(\theta)/L(\theta_0)$ as a function of θ. This will be a satisfactory estimator only for those θ close to θ_0; specifically, for those θ for which $P_\theta(\mathbf{X}|\mathbf{Y})$ is close to proportional to $P_{\theta_0}(\mathbf{X},\mathbf{Y})$. Sometimes, primary interest is in the shape of the likelihood surface in the neighborhood of some specific point, such as the maximum likelihood estimate (MLE). In this case, preliminary MCMC runs and likelihood ratio function estimates can be used to obtain a ballpark value of the MLE (Geyer and Thompson, 1992). Alternatively, Monte Carlo EM can be used (see section 9.3). Once a ballpark estimate of the parameter values is found, one very large MCMC run can provide an accurate estimate of the MLE and of the likelihood in the region. However, this approach has limitations. One may be interested in the likelihood surface, or in log-likelihood differences, over large regions in the parameter space. Or, the large MCMC run may reveal that one's initial estimate was not sufficiently close to the MLE, and additional large runs may be necessary. It is desirable to find a method that combines realizations from all the runs, and provides an estimate of the likelihood surface over a range of parameter values.

9.2 Monte Carlo relative likelihood surfaces

One way of combining realizations from different MCMC samplers was provided by Geyer (1991*b*). MCMC samplers are run at many models, covering the range of interest, say at $\theta_0, \theta_1, \ldots, \theta_K$. The sets of N_j realizations from $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$, $j = 0, 1, \ldots, K$, give a combined set of realizations from

$$\begin{aligned} P^*(\mathbf{X}) &= \frac{1}{\sum_j N_j}\sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{X}|\mathbf{Y}) \\ &= \frac{1}{\sum_j N_j}\sum_{j=0}^{K} N_j P_{\theta_j}(\mathbf{X},\mathbf{Y})/L(\theta_j) \end{aligned}$$

and writing the likelihood estimation formula as an expectation with respect to this P^*

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

Now, although we have a sample from P^*, the denominator $P^*(\mathbf{X})$ cannot be explicitly computed, since it depends on the unknown $L(\theta_j)$, but we have the implicit Monte Carlo estimating equations

$$
\begin{aligned}
L(\theta_j) &= \sum_{\mathbf{X}^*} \left(\frac{P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{X}^*, \mathbf{Y})/L(\theta_l)} \right) \\
\text{(9.2)} \qquad &= \sum_{\mathbf{X}^*} \left(\sum_{l=0}^{K} N_l \frac{P_{\theta_l}(\mathbf{X}^*, \mathbf{Y})}{P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})} \frac{1}{L(\theta_l)} \right)^{-1}
\end{aligned}
$$

for $j = 0, ..., K$, where the sum is over the total set of realizations $\mathbf{X}^*$. These equations determine only the relative values of $L(\theta_j)$, but can be solved iteratively for these relative values. For example, one may iterate equation (9.2) directly, renormalizing after each cycle, to keep one value, say $L(\theta_0)$ fixed (=1). This iterative procedure is globally convergent to the unique solution of equation (9.2). Once the relative values of $L(\theta_j)$ are found, then, for any other value of θ in the range spanned by the set of θ_j, $L(\theta)$ can be estimated by

$$
\text{(9.3)} \qquad L(\theta) = \sum_{\mathbf{X}^*} \left(\sum_{l=0}^{K} N_l \frac{P_{\theta_l}(\mathbf{X}^*, \mathbf{Y})}{P_{\theta}(\mathbf{X}^*, \mathbf{Y})} \frac{1}{L(\theta_l)} \right)^{-1}
$$

where the sum is over the same total set of realizations as before. (Again, the estimate is relative to $L(\theta_0) = 1$.) Geyer (1991*b*) named this method *reverse logistic regression.*

There are two requirements for this approach to be an effective solution to the likelihood estimation problem. First, each sampler $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$, $j = 0, \ldots, K$ must cover well that part of the space of $\mathbf{X}$-values that has high total probability mass under that probability distribution — for an MCMC sampler on a large and structured space of latent variables, this is a non-trivial consideration (section 8.1). Second, even if the separate samplers are behaving "well", in this sense, for the mixture estimates to be effective we need good "overlap" between adjacent models. The conditional probability that a particular observation $\mathbf{X}$ derives from the sample P_{θ_j} is

$$
\frac{N_j P_{\theta_j}(\mathbf{X}|\mathbf{Y})}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{X}|\mathbf{Y})}.
$$

For every j, the values of these probabilities should not be too close to 1 for too large a proportion of the sampled $\mathbf{X}$-values. Thus adjacent parameter values θ_j must be chosen not too far apart, where the relevant measure of distance is in terms of the probability distributions $P_{\theta_j}(\mathbf{X}|\mathbf{Y})$ of the $\mathbf{X}$-values generated.

Other difficulties with using the reverse logistic regression method concern computational resources. Either the realized $\mathbf{X}^*$, or at least the values $P_\theta(\mathbf{X}^*, \mathbf{Y})$ for each θ of interest, must be saved, in order for equations (9.2) and (9.3) to be implemented. This can demand massive amounts of storage. An alternative is to use block averages of the ratios of $P_{\theta_j}(\mathbf{X}, \mathbf{Y})/P_{\theta_l}(\mathbf{X}, \mathbf{Y})$ in equation (9.2)

(Thompson, 1994*b*). In the extreme case, this block might be the average over a full run of the sampler at a given θ_j. Let

$$R_j(\theta_l, \theta_j) \;=\; N_j^{-1} \sum_{\mathbf{X}^{*(j)}} \frac{P_{\theta_l}(\mathbf{X}^{*(j)}, \mathbf{Y})}{P_{\theta_j}(\mathbf{X}^{*(j)}, \mathbf{Y})}$$

be the likelihood ratio estimate of $L(\theta_l)/L(\theta_j)$ from N_j realizations $\mathbf{X}^{*(j)}$ at θ_j. Here the chosen values of l may vary with j. We define $R_j(\theta_l, \theta_j)$ to be 0 if $L(\theta_l)/L(\theta_j)$ is not estimated from realizations under model θ_j. At a minimum, for each j, values for $R_j(\theta_l, \theta j)$ should be computed for the values θ_l adjacent to θ_j. Then the estimating equation (9.2) becomes

$$L(\theta_j) \;=\; \left(\sum_{l=0}^{K} R_j(\theta_l, \theta_j) \frac{1}{L(\theta_l)} \right)^{-1} . \tag{9.4}$$

Writing $\nu_j = 1/L(\theta_j)$, $R_{jl} = R_j(\theta_l, \theta_j)$, $\nu = (\nu_j)$, and $\mathbf{R} = (R_{jl})$, equation (9.4) becomes

$$\nu \;=\; \mathbf{R}\nu.$$

That is, the vector of ν_j-values is a right eigenvector of the matrix $\mathbf{R}$. Asymptotically, for large Monte Carlo runs, each computed R_{jl}-value converges to $L(\theta_l)/L(\theta_j) = \nu_j/\nu_l$. Thus, if, for each j, R_{jl} is evaluated for t other θ_l values, then each evaluated $R_{jl}\nu_l$ is approximately ν_j, and the corresponding eigenvalue should be t. This provides one check on the performance of the method, although in practice it is a weak criterion. The eigenvalue can be close to t even when performance is poor.

There are many open questions in the statistical properties of estimators such as those resulting from equation (9.4). If sufficient realizations can be stored, then equation (9.2) may provide the more satisfactory estimate. Suppose, however, only one in 1000 samples $\mathbf{X}^*$ or resulting probabilities $P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})$ can be stored. Then should one use the estimate (9.2), or one that uses the block averages over each block of 1000 steps? The latter would require more computation (evaluations of $P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})$), but the same amount of store. The Monte-Carlo variance of the block-average will be less than that of individual values $P_{\theta_j}(\mathbf{X}^*, \mathbf{Y})$, but possibly not by much if the autocorrelation in the Markov chain is very high. Clearly these questions are related also to issues of computational efficiency in sub-sampling and spacing in the MCMC (Geyer, 1992), discussed briefly in section 8.1.

9.3 Monte Carlo EM for the mixed model

For some models, exact computation of the conditional expectations required to implement an EM algorithm may be impractical or infeasible, particularly if the model is complex, or there are missing data. Penetrance parameters may not be simple functions of genotypic counts. Even the bivariate case of the simple polygenic model (section 2.6) may be complicated, if some individuals are observed for just one of the two traits (Thompson and Shaw, 1992). Chiasmata patterns are not so readily imputed if the recombination patterns of some gametes are not fully observable (section 5.3), due to missing typings or parental homozygosity at some loci. However, if latent genotypes or meiosis indicators and missing phenotypes can be realized from their conditional distributions given the observed data $\mathbf{Y} = \mathbf{y}$ under current values of the parameters, a Monte Carlo EM (MCEM) is easily implemented.

In section 2.6, the simple polygenic model was introduced, and the EM-algorithm for the variance component parameters σ_a^2 and σ_e^2 was outlined. In section 6.6, the univariate trait model was generalized to the mixed model, including both Mendelian genotypes and Gaussian polygenic effects (see equation (6.5)). The parameters then include also the frequency of the alleles at the diallelic Mendelian trait locus, and the vector of genotypic means $\boldsymbol{\mu} = (\mu(g))$ for the genotypes g at the locus. As before, we index the members of the pedigree by i, $i = 1, \ldots, n_{tot}$. Suppose that the n_{obs} observed members of the pedigree are those indexed by $i \in \mathcal{D}$. Then, for i in $\mathcal{D}$, we have equation (6.5):

$$Y_i \quad = \quad \mu(G_i) \ + \ Z_i \ + \ \epsilon_i.$$

The vector $\mathbf{Z} = (Z_i)$ is defined over all n_{tot} members of the pedigree, and has the multivariate Gaussian distribution $\mathbf{Z} \sim N(\mathbf{0}, \sigma_a^2\mathbf{A})$, where $\mathbf{A}$ is a matrix determined by the pedigree structure (section 2.6).

If $I\{E\}$ is the indicator function of the event E, the complete-data sufficient statistics of this exponential family model for $(\mathbf{G}, \mathbf{Z}, \mathbf{Y})$ are:
the number of observed individuals of each genotype g, or $\sum_{i \in \mathcal{D}} I\{G_i = g\}$
the total trait effect in those individuals, $\sum_{i \in \mathcal{D}} (Y_i - Z_i) I\{G_i = g\}$
the quadratic residual term, for observed individuals,

$$\epsilon'\epsilon \ = \ (\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z})'(\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z}), \text{ and}$$

the total genetic variance over all pedigree members, $\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}$.

If genotypes G_i and polygenic values Z_i were observable, then the MLEs of the parameters would be straightforward. For each discrete genotype g

$$\widehat{\mu(g)} \ = \ \frac{\sum_{i \in \mathcal{D}} (Y_i - Z_i) I\{G_i = g\}}{\sum_{i \in \mathcal{D}} I\{G_i = g\}}.$$

For the variance component parameters (see equation (2.17))

$$\begin{aligned} \widehat{\sigma_e^2} &= (\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z})'(\mathbf{Y} - \mu(\mathbf{G}) - \mathbf{Z})/n_{obs} \\ \widehat{\sigma_a^2} &= (\mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z})/n_{tot} \end{aligned}$$

where $\mu(\mathbf{G})$ denotes the vector of genotypic values of observed individuals $(\mu(G_i);\ i \in \mathcal{D})$. However, for this mixed model, both exact implementation of an EM algorithm and exact evaluation of the likelihood are infeasible. Monte Carlo methods can, however, be implemented.

For example, the conditional expectations of the statistics in the above equations, given the data $\mathbf{Y}$, may be estimated by averaging the values given by N realizations

$(\mathbf{Z}^{(\tau)}, \mathbf{G}^{(\tau)})$. At current parameter values, $(\sigma_e^2, \sigma_a^2, \mu)$, realizations are obtained from the conditional distribution $P_{\sigma_e^2,\sigma_a^2,\mu}(\mathbf{Z}, \mathbf{G} \mid \mathbf{Y})$, leading to Monte Carlo EM update equations

$$\mu(g)^* = (N)^{-1} \frac{\sum_{\tau=1}^{N} \sum_{i \in \mathcal{D}} (Y_i - Z_i^{(\tau)}) I\{G_i^{(\tau)} = g\}}{\sum_{\tau=1}^{N} \sum_{i \in \mathcal{D}} I\{G_i^{(\tau)} = g\}}$$

$$\text{(9.5)} \quad \sigma_e^{2*} = (N n_{obs})^{-1} \sum_{\tau=1}^{N} (\mathbf{Y} - \mu(\mathbf{G}^{(\tau)}) - \mathbf{Z}^{(\tau)})'(\mathbf{Y} - \mu(\mathbf{G}^{(\tau)}) - \mathbf{Z}^{(\tau)})$$

$$\text{(9.6)} \quad \sigma_a^{2*} = (N n_{tot})^{-1} \sum_{\tau=1}^{N} \mathbf{Z}^{(\tau)\prime} \mathbf{A}^{-1} \mathbf{Z}^{(\tau)}.$$

Equations (9.5) and (9.6) should be compared with the exact EM equations for the parameters of a polygenic model (equation (2.17)). With a Monte Carlo approach, the conditional variance of $\mathbf{Z}$ given $\mathbf{Y}$ and $\mathbf{G}$ need not be computed, since the variance is subsumed into the realized variability of the quadratic expressions. Note, however, that this variance is an intrinsic part of the iterative procedure. Just as in section 2.6, it is insufficient to use only the estimate $\tilde{\mathbf{a}} = N^{-1} \sum_{\tau=1}^{N} \mathbf{Z}^{(\tau)}$ of the conditional mean $\mathbf{a} = \mathrm{E}_{\sigma_e^2,\sigma_a^2,\mu}(\mathbf{Z} \mid \mathbf{G}, \mathbf{Y})$.

Returning to single-locus models, if genotypes $\mathbf{G}$ can be realized given the data $\mathbf{Y}$ and current parameter values, MCEM equations for parameters of penetrance densities are straightforward. The use of MCEM also permits extension to more complex models. One example is that of a more general mixed model for a quantitative trait, including also the effects of observed covariates and other variance component effects, such as those due to shared environment. This model assumes the trait value y_i is the sum of these effects together with the effect of a single-locus genotype G_i, a polygenic value z_i, and a residual with mean 0 and variance σ_e^2. Provided genotypes and polygenic values $(\mathbf{G}, \mathbf{Z})$ can be realized, conditional upon data $\mathbf{Y}$ and current parameter values, MCEM is again feasible. Achieving these realizations is not, in general, straightforward. We can do so by using Markov chain Monte Carlo (MCMC). Guo and Thompson (1992; 1994) have used MCEM for the mixed model and for joint linkage and segregation analysis. Generally, MCEM is as effective as EM at getting a ball-park estimate, and is remarkably robust even when quite small Monte Carlo samples are used. However, it is of little use in obtaining a precise final MLE—a large number of very large samples would be required.

9.4 Likelihood estimators for complex models

The mixed model also provides an example of Rao-Blackwellization (section 3.8) of Monte Carlo estimates of likelihood ratios. Applying the formula (9.1) directly to the mixed model with latent variables $(\mathbf{G}, \mathbf{Z})$, we have

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{G}, \mathbf{Z}, \mathbf{Y})}{P_{\theta_0}(\mathbf{G}, \mathbf{Z}, \mathbf{Y})} \;\middle|\; \mathbf{Y} \right).$$

However, considering only the latent variables **G**, it is also the case that

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y}|\mathbf{G})P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{Y}|\mathbf{G})P_{\theta_0}(\mathbf{G})} \;\middle|\; \mathbf{Y}\right), \tag{9.7}$$

while considering only latent variables **Z**

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y}|\mathbf{Z})P_\theta(\mathbf{Z})}{P_{\theta_0}(\mathbf{Y}|\mathbf{Z})P_{\theta_0}(\mathbf{Z})} \;\middle|\; \mathbf{Y}\right). \tag{9.8}$$

Since **Y** and **Z** are continuous random variables, we have now probability density functions rather than probability mass functions. However, we retain the notation $P_\theta(\cdot)$, to avoid introducing additional notation for this one example.

As shown in section 6.6, either integration over **Z** or summation over **G** is possible in the mixed-model likelihood (equation (6.6)), providing for exact computation of the probabilities

$$P_\theta(\mathbf{Y}|\mathbf{G}) \;=\; \int_{\mathbf{z}} \Pr(\mathbf{Y}|\mathbf{z},\mathbf{G})P_{\sigma_a^2}(\mathbf{z})d\mathbf{z} \tag{9.9}$$

in equation (9.7), or of the probabilities

$$P_\theta(\mathbf{Y}|\mathbf{Z}) \;=\; \sum_{\mathbf{G}} P_\theta(\mathbf{Y}|\mathbf{Z},\mathbf{G})P_\theta(\mathbf{G})$$

in equation (9.8). Equations (9.7) and (9.8) provide two alternative Rao-Blackwellized estimators. To implement the estimate based on (9.7) or on (9.8), only the realizations of **G** or of **Z** would be used. However, if using a Markov chain Monte Carlo (MCMC) sampler (Chapter 8), it will normally be necessary to generate both. For example, from N Monte Carlo realizations $(\mathbf{G}^{(\tau)},\mathbf{Z}^{(\tau)})$ generated from $P_{\theta_0}(\mathbf{G},\mathbf{Z}|\mathbf{Y})$, the estimate based on equation (9.7) would be

$$\widehat{\frac{L(\theta)}{L(\theta_0)}} \;=\; \frac{1}{N}\sum_{\tau=1}^{N} \frac{P_\theta(\mathbf{Y}|\mathbf{G}^{(\tau)})P_\theta(\mathbf{G}^{(\tau)})}{P_{\theta_0}(\mathbf{Y}|\mathbf{G}^{(\tau)})P_{\theta_0}(\mathbf{G}^{(\tau)})}. \tag{9.10}$$

The hope is that the reduction in Monte Carlo variance due to the partial exact computation (9.9) will compensate for this increased computation (see section 3.8).

Note also that, for some model comparisons, additional exact integration or summation over latent variables **Z** or **G** may be unnecessary. If under models indexed by θ and by θ_0,

$$P_\theta(\mathbf{Y}|\mathbf{G}) \;=\; P_{\theta_0}(\mathbf{Y}|\mathbf{G})$$

these probabilities need not be computed, and the estimate (9.10) reduces to a ratio of the prior genotype probabilities $P_\theta(\mathbf{G})/P_{\theta_0}(\mathbf{G})$ averaged over the realized $\mathbf{G}^{(\tau)}$. Similarly, if

$$P_\theta(\mathbf{Y}|\mathbf{Z}) \;=\; P_{\theta_0}(\mathbf{Y}|\mathbf{Z}),$$

the estimator based on equation (9.8) reduces to the ratio of population densities of $\mathbf{Z}$. By careful choice of models to be compared, procedures can be made more computationally efficient.

Reduction in Monte Carlo variance by Rao-Blackwellization is guaranteed only for independent realizations $(\mathbf{G}^{(\tau)}, \mathbf{Z}^{(\tau)})$ of the latent variables (Geyer, 1992). For dependent realizations, Geyer (pers.comm.) has provided a simple counterexample based on latent variables consisting the odd and even terms of a first order Gaussian autoregressive process. However, in many practical instances the Rao-Blackwellization procedure works well, even when MCMC realizations are used. Estimators based on (9.7) and (9.8) were introduced by Thompson and Guo (1991) and compared by Thompson (1994*c*) in likelihood analyses of genetic models with several latent heritable components. It was found that the estimator (9.10) works very well, leading to substantial gains in computational efficiency, whereas the estimator based on (9.8) is very inefficient. The summation over $\mathbf{G}$ required for the latter is generally computationally more intensive than integration over $\mathbf{Z}$. More importantly, the data $\mathbf{Y}$ and variables $\mathbf{Z}$ together constrain $\mathbf{G}$ very much more than $\mathbf{Y}$ and $\mathbf{G}$ constrain $\mathbf{Z}$. Since the conditional variance of $\mathbf{Z}$ given $\mathbf{Y}$ and $\mathbf{G}$ is relatively high, exact integration over $\mathbf{Z}$ reduces Monte Carlo variance substantially.

Note that equation (9.1), or forms thereof such as (9.7) and (9.8), are not the only possible ways to obtain Monte Carlo estimates of likelihood ratios. In particular, Meng and Wong (1996) have considered a variety of forms of importance sampling and Rao-Blackwellization, noting that (in the notation of this chapter)

$$(9.11) \qquad \frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \frac{\mathrm{E}_{\theta_0}(P_\theta(\mathbf{X},\mathbf{Y})\alpha(\mathbf{X}) \mid \mathbf{Y})}{\mathrm{E}_\theta(P_{\theta_0}(\mathbf{X},\mathbf{Y})\alpha(\mathbf{X}) \mid \mathbf{Y})}$$

where $\alpha(\mathbf{X})$ is an arbitrary function on the space of $\mathbf{X}$ values (provided the expectations exist, and the distributions have the same support). If $\alpha(\mathbf{X}) = 1/P_{\theta_0}(\mathbf{X},\mathbf{Y})$, equation (9.11) reduces to equation (9.1). Note that whereas use of equation (9.1) requires MCMC only at θ_0, the expectation in the denominator of equation (9.11) requires MCMC at the value θ. Various other choices of $\alpha(\mathbf{X})$ have been investigated in the recent MCMC literature. Jensen and Kong (1999) have used a version of equation (9.11) in their MCMC estimation of a single-marker linkage lod score on a complex pedigree.

As for the ratio estimator (7.17), the expression (9.11) is a ratio of expectations, and thus the Monte Carlo estimator is a ratio of averages over two sets of Monte Carlo realizations. For the estimator based on (7.17), the sampling distribution is the same in numerator and denominator, and thus Monte Carlo variance could be reduced, and computational efficiency enhanced, by using the same Monte Carlo realizations in the estimates of numerator and denominator. However, for the likelihood ratio estimator based on (9.11), different sampling distributions are required, so different Monte Carlo Markov chains must be run for the numerator and denominator. If MCMC is being done in any case at a set of values θ_j, for example as in section 9.2, this does not impose any increased computational burden for the Monte Carlo itself. However, long runs may be needed to reduce the Monte Carlo variance of the estimate of $L(\theta)/L(\theta_0)$ to acceptable levels.

9.5 Likelihood estimation of gene locations

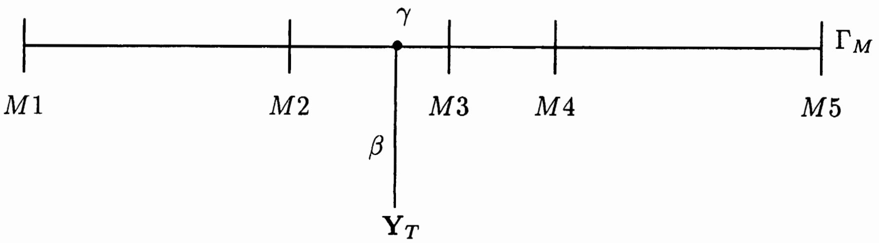

FIGURE 9.1. *Model parameters for estimation of a location likelihood curve*

In modern genetic analysis, a primary goal is the localization of trait genes. Genetic markers have been mapped throughout the genome at a scale suitable for multipoint linkage analysis. Thus, estimation of *location lod score curves* (sections 6.2, 7.6) is an important goal. Here we denote the marker model parameters by Γ_M. For a complex trait, the trait model parameters are also unknown. These parameters, β, determine the probabilities of phenotypes given the latent genes. While in some analyses, joint maximization of the likelihood with respect to trait model β and trait locus position γ may be attempted, often the *location lod score curve* is computed for fixed β. The likelihood (or a profile likelihood) is evaluated as a function of a hypothesized trait-locus location γ, against a fixed marker map Γ_M. The parametrization of the overall model is shown in Figure 9.1 The overall model is indexed by parameter $\theta = (\beta, \gamma, \Gamma_M)$. As before, the likelihood is

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

(equation (7.8)) which may take the form (1.5) if $\mathbf{X} = \mathbf{G}$, the underlying genotypes, or (4.11) if $\mathbf{X} = \mathbf{S}$, the inheritance patterns of genes. For computations with multiple marker loci, the Lander-Green paradigm (4.11) is more natural and more effective, but exact computation is limited to small pedigrees.

As in the discussion of *Elods* (equations (4.8) and (7.15)), for convenience we partition the data $\mathbf{Y}$ into the trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$. The corresponding latent variables are partitioned into $\mathbf{X}_T$ and $\mathbf{X}_M$. Monte Carlo estimation of the location likelihood ratio is always feasible. The form that follows directly from equation (9.1) is

$$\frac{L(\beta,\ \gamma_1,\ \Gamma_M)}{L(\beta,\ \gamma_0,\ \Gamma_M)} = \mathrm{E}_{\theta_0}\left(\frac{P_{\theta_1}(\mathbf{Y}_T, \mathbf{Y}_M \mid \mathbf{X}_T, \mathbf{X}_M) P_{\theta_1}(\mathbf{X}_T, \mathbf{X}_M)}{P_{\theta_0}(\mathbf{Y}_T, \mathbf{Y}_M \mid \mathbf{X}_T, \mathbf{X}_M) P_{\theta_0}(\mathbf{X}_T, \mathbf{X}_M)} \;\middle|\; \mathbf{Y}_T, \mathbf{Y}_M \right)$$

for two hypothesized trait locus positions γ_1 and γ_0. Noting the fact that only the position of the trait locus differs between numerator and denominator, the above equation reduces to

$$\text{(9.12)} \qquad \frac{L(\beta,\ \gamma_1,\ \Gamma_M)}{L(\beta,\ \gamma_0,\ \Gamma_M)} = \mathrm{E}_{\theta_0}\left(\frac{P_{\gamma_1}(\mathbf{X}_T \mid \mathbf{X}_M)}{P_{\gamma_0}(\mathbf{X}_T \mid \mathbf{X}_M)} \;\middle|\; \mathbf{Y}_T, \mathbf{Y}_M\right).$$

Thus only the conditional probability of trait-locus latent variables given marker-loci latent variables appears explicitly in the estimator. Although realization of the latent variables is complex, and requires MCMC methods, computation of the estimate from the realizations is generally very straightforward (Thompson and Guo, 1991).

One practical difficulty of the above approach is accurate estimation of log-likelihood differences for trait locations in different marker intervals. The likelihood-ratio estimate (9.1) works well in comparing locations within an interval, and in principle the mixtures method (9.2) facilitates estimation between intervals. However, in practice, values of $\mathbf{X}_T$ realized at γ_0 may have very small probabilities under γ_1 if there is a marker locus between the two positions. Additionally, the usual objective is to estimate the lod-score relative to the base-point in which the trait locus position γ is not within the marker map Γ_M. That is, the null hypothesis that the trait locus is unlinked. Again this can be accomplished by using the mixtures method (9.2), but several intervening positions γ, linked to but not within the marker map, may be required for effective estimation (Thompson, 1994*b*). The procedure becomes computationally intensive.

Another disadvantage of the approach of section 9.1 and this section is the fact that it allows estimation only of likelihood ratios, not of likelihoods. A modification is due to Lange and Sobel (1991) for the particular case of Monte Carlo estimation of location likelihoods. Their procedure also avoids the problems of sampling of the trait locus variables. Again, we assume the marker map and parameters Γ_M known, so that $P_{\Gamma_M}(\mathbf{Y}_M)$ is a constant factor in the likelihood. Then Lange and Sobel (1991) write the likelihood in a form which, using our current notation, becomes

$$\begin{aligned} L(\beta,\gamma,\Gamma_M) &= P_{\beta,\gamma,\Gamma_M}(\mathbf{Y}_M,\mathbf{Y}_T) \\ &\propto P_{\beta,\gamma,\Gamma_M}(\mathbf{Y}_T \mid \mathbf{Y}_M) \\ &= \sum_{\mathbf{X}_M} P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M) P_{\Gamma_M}(\mathbf{X}_M \mid \mathbf{Y}_M) \\ \text{(9.13)} \qquad &= \mathrm{E}_{\Gamma_M}(P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M) \mid \mathbf{Y}_M). \end{aligned}$$

Now latent variables $\mathbf{X}_M$ are sampled from their conditional distribution given the marker data $\mathbf{Y}_M$. Provided exact computation of $P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M)$ is possible for alternative trait models (β) and locations (γ), we have a Monte Carlo estimate of $L(\beta,\gamma,\Gamma_M)$. Comparison to the unlinked base-point requires only computation of $P_\beta(\mathbf{Y}_T)$, the probability of trait data under the parameters β of the trait locus model. This can be accomplished by single-locus peeling methods of Chapter 6. Since Γ_M is fixed, the Monte Carlo requires only a single set of realizations $\{\mathbf{X}_M^{(\tau)}, \tau = 1,\ldots,N\}$. The disadvantage is that $P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{X}_M^{(t)})$ must be computed for each such realization; this requires a single-locus peeling computation for the trait-locus data under the trait model. Further, this computation must be done, not only for each realization $\mathbf{X}_M^{(\tau)}$, but also for each β and γ at which a likelihood estimate is required.

In many cases, however, the gains outweigh the costs, except when the simulation distribution $P_{\Gamma_M}(\mathbf{X}_M \mid \mathbf{Y}_M)$ is not close to proportional to the ideal importance-sampling target distribution $P_{\beta,\gamma,\Gamma_M}(\mathbf{X}_M \mid \mathbf{Y}_M, \mathbf{Y}_T)$. This is particularly so for models (trait locations) γ which are not close to the truth, and for a trait which provides substantial information about the inheritance patterns of genes at the underlying trait locus, and hence also at linked marker loci. In fact, the cases where the Monte Carlo estimator based on equation (9.13) performs poorly are precisely those in which the likelihood ratio estimator (9.12) also has difficulties. There continue to be interesting open questions in the estimation of multilocus linkage likelihoods.

9.6 Marker *ibd* and complete-data log-likelihoods

Again suppose that, as in sections 7.5 and 9.5, we have trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$. Further, suppose that the marker map Γ_M, marker allele frequencies and marker population genotype frequencies are known, Consider also the case where the latent variables $\mathbf{X}_M$ are the meiosis indicators $\mathbf{S}_M$. As described above, MCMC methods, and in particular the M-sampler of section 8.4, provide effective methods for sampling from the conditional distribution $P_{\Gamma_M}(\mathbf{S}_M \mid \mathbf{Y}_M)$. Among pedigree members, the patterns of gene *ibd* at marker loci are functions of $\mathbf{S}_M$; $\mathbf{J} = \mathbf{J}(\mathbf{S})$ (section 3.6). Thus we have MCMC estimates of the conditional probabilities of gene *ibd* at marker loci, given the marker data.

Neither trait data, $\mathbf{Y}_T$ nor trait model enter into this sampling of marker latent variables conditional on marker data. However, under any trait model with some genetic component, related affected individuals or related individuals exhibiting extreme trait values will share genes *ibd* at trait loci with some increased probability. Hence also they will share genes *ibd* with increased probability at marker loci linked to those trait loci. In so-called "non-parametric" computations for linkage detection, marker data on a pedigree are analyzed to detect regions of the genome in which there is evidence for excess gene *ibd* among affected individuals, or individuals exhibiting extreme trait values. Such regions provide evidence for linkage.

The Monte Carlo sampling of $\mathbf{S}_M$ given marker data $\mathbf{Y}_M$ provides direct estimates of conditional probabilities of patterns of gene *ibd* $\mathbf{J}(\mathbf{S}_M)$. These gene *ibd* probabilities at locus j are computed dependent on all the marker data $\mathbf{Y}_M$, as, for example, are the probabilities $Q_j(S_{\bullet,j})$ of section 7.1. Here we have only Monte Carlo estimates of these probabilities, but MCMC realization on larger or more complex pedigrees is feasible in cases for which exact computation is not. Moreover, the resulting gene *ibd* patterns $\mathbf{J}(\mathbf{S}_M)$ may be scored jointly over haplotypes, and over loci. The example of section 4.5 showed the importance of considering both individuals and loci jointly. For the case where only marker data are considered, many of the problems of the Monte Carlo estimation procedures are much reduced, provided a good MCMC sampler is used (see sections 8.3, 8.4).

The statistical problem becomes one of development of appropriate test statistics, to detect linkage on the basis of estimated conditional *ibd* probabilities. Although most current methods involve statistics computed marginally over loci, and pairwise over individuals, there is an increasing literature in this area; see for example Whittemore and Tu (1998).

Another readily computed by-product of MCMC on pedigrees, or in any latent-variable problem, is the expected complete-data log-likelihood, $H_{\mathbf{y}}(\theta;\theta_0)$ (section 2.4). Returning again to the full data $\mathbf{Y}$ and latent variables $\mathbf{S}$, at a general model indexed by parameters θ_0 we have

$$\begin{aligned} H_{\mathbf{y}}(\theta;\theta_0) &= \mathrm{E}_{\theta_0}(\log_e P_\theta(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y}) \\ &= \mathrm{E}_{\theta_0}(\log_e P_\theta(\mathbf{Y} \mid \mathbf{S}) + \log_e P_\theta(\mathbf{S}) \mid \mathbf{Y}). \end{aligned} \tag{9.14}$$

For easier comparison with statistical results, we use natural (base-e) logarithms throughout this section. Due to the *a priori* independence of meioses

$$\log P_\theta(\mathbf{S}) = \sum_{i=1}^{m} \log P_\theta(S_{i,\bullet}) \tag{9.15}$$

and, provided data are locus-specific,

$$\log P_\theta(\mathbf{Y} \mid \mathbf{S}) = \sum_{j=1}^{L} \log P_\theta(Y_{\bullet,j} \mid S_{\bullet,j}) \tag{9.16}$$

(see equation (4.11)). Thus the expectation partitions into terms for each locus and for each meiosis. These terms must be computed in any case in the course of the MCMC, making accumulation of values for the estimated expectation particularly straightforward. Note that (9.15) depends only on the genetic map parameters, while (9.16) depends on the penetrance aspects of the model. In expectation, under the conditional distribution $P_{\theta_0}(\cdot|\mathbf{Y})$, each term depends, of course, on all the parameters in θ_0. The expected complete-data log-likelihood, with its component parts, proves to be a useful diagnostic measure of the performance of the MCMC.

The above discussion depends on the decomposition

$$\log P_\theta(\mathbf{S},\mathbf{Y}) = \log P_\theta(\mathbf{Y} \mid \mathbf{S}) + \log P_\theta(\mathbf{S}).$$

Reversing the decomposition of the complete-data log-likelihood

$$\log P_\theta(\mathbf{S},\mathbf{Y}) = \log P_\theta(\mathbf{Y}) + \log P_\theta(\mathbf{S} \mid \mathbf{Y}).$$

Thus, as in equation (2.9), differences in expected complete-data log-likelihoods depend on the true log-likelihood difference and the Kullback-Leibler information (section 2.2) in the distribution of $\mathbf{S}$ given $\mathbf{Y}$. That is,

$$K_{\mathbf{y}}(\theta;\theta_0) = \mathrm{E}_{\theta_0}(\log_e P_{\theta_0}(\mathbf{S}|\mathbf{Y}) - \log_e P_\theta(\mathbf{S}|\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}),$$

so the difference in expected complete-data log-likelihoods is

$$\begin{aligned}
H_{\mathbf{y}}(\theta_0;\theta_0) & - H_{\mathbf{y}}(\theta;\theta_0) \\
= & \quad \mathrm{E}_{\theta_0}(\log P_{\theta_0}(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y}) - \mathrm{E}_{\theta_0}(\log P_{\theta}(\mathbf{S},\mathbf{Y}) \mid \mathbf{Y}) \\
= & \quad \mathrm{E}_{\theta_0}(\log P_{\theta_0}(\mathbf{Y}) + \log P_{\theta_0}(\mathbf{S}|\mathbf{Y}) - \log P_{\theta}(\mathbf{Y}) - \log P_{\theta}(\mathbf{S}|\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}) \\
= & \quad \log P_{\theta_0}(\mathbf{Y}) \ - \ \log P_{\theta}(\mathbf{Y}) \ + \ \mathrm{E}_{\theta_0}(P_{\theta_0}(\mathbf{S}|\mathbf{Y}) - \log P_{\theta}(\mathbf{S}|\mathbf{Y}) \mid \mathbf{Y} = \mathbf{y}) \\
& \quad = \ \ell(\theta_0) \ - \ \ell(\theta) \ + \ K_{\mathbf{y}}(\theta;\theta_0). \qquad (9.17)
\end{aligned}$$

The extent to which this identity can be exploited in making inferences from MCMC output is also an area of ongoing research.