Chapter 8

Markov chain Monte Carlo on Pedigrees

8.1 Simulation conditional on data: MCMC

Equation (7.10) gave the likelihood for a genetic model on a pedigree as an expectation over latent variables $\mathbf{X}$, and hence, in principle, provided a method for Monte Carlo estimation of the likelihood. We need to estimate

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

As previously, any suitable latent variables may be used, normally either meiosis indicators $\mathbf{S}$ or genotypes $\mathbf{G}$. For convenience, we use the general notation $\mathbf{X}$ for the general formulation.

However, unless the simulation distribution $P^*(\mathbf{X})$ is conditioned in some way on data $\mathbf{Y}$, equation (7.10) is often useless. Genotypes or gene descent patterns simulated from the prior probability distribution given only the model and the pedigree structure will rarely even be consistent with the observed data. Importance sampling considerations dictate that the sampling distribution should be close to proportional to $P_\theta(\mathbf{X}, \mathbf{Y})$, or as a function of latent variables $\mathbf{X}$ to $P_\theta(\mathbf{X} \mid \mathbf{Y})$ (equation (7.12)). Intuitively also, to obtain realizations that have better than infinitesimal probability of giving a non-negligible contribution to the likelihood we must simulate conditional on the data. However

$$P_\theta(\mathbf{X} \mid \mathbf{Y}) \;=\; \frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_\theta(\mathbf{Y})}, \tag{8.1}$$

and the normalizing factor $P_\theta(\mathbf{Y})$ is unknown. If we could compute $L(\theta) = P_\theta(\mathbf{Y})$, Monte Carlo estimation of likelihoods would be unnecessary.

Enter Markov chain Monte Carlo, or MCMC. We review briefly the Metropolis-Hastings class of algorithms (Hastings, 1970) for generating dependent realizations from a target probability distribution known only up to a normalizing factor. For