MCMC for the analysis of genetic data on pedigrees: Tutorial Session 2

Elizabeth Thompson
University of Washington

- Genetic mapping and linkage lod scores
- Monte Carlo likelihood and likelihood ratio estimation
- Monte Carlo estimation of linkage lod scores

GENETIC MARKERS

- Human genome: 3×10^9 bp of DNA.

- DNA variants that can be typed in individuals.
  Allele — type of the DNA at position on chromosome

- Have been mapped: known locations on the genome.
  Locus— position on a chromosome, or DNA at that position

- Idea: map genes for traits relative to these markers.

- Microsatellites; lots of alleles; 350 in a genome scan
  One every 10^7bp
  SNPs: typically only two alleles; lots more exist; 1 per 1000 bp

THE STRUCTURE OF A GENETIC MODEL

• Population model: parameters $\mathbf{q}$, provide probabilities for latent $\mathcal{A}$– allelic types of FGL at each j

• Inheritance model: parameters ρ, provide probabilities for latent $\mathbf{S}$– inheritance of FGL at j, jointly over j.

• Individual genotypes $\mathbf{G}$ is deterministic function of $(\mathbf{S}, \mathcal{A})$

• penetrance model, parameters β relates $\mathbf{G}$ (and perhaps observable covariates) to observable data $\mathbf{Y}$

• $\xi = (\mathbf{q}, \rho, \beta)$.

FROM RECOMBINATION TO LOCATION

- Recall model for $S_{i,\bullet} = (S_{i,1}, ..., S_{i,L})$:
  $\Pr(S_{i,j} \neq S_{i,j+1}) = \rho_j$: assumed same $\forall\ i$ (convenience).
  $S_{i,j}$ assumed Markov in j: no genetic interference.

- Genetic distance d is expected number of crossover events on underlying chromosome: an additive measure.

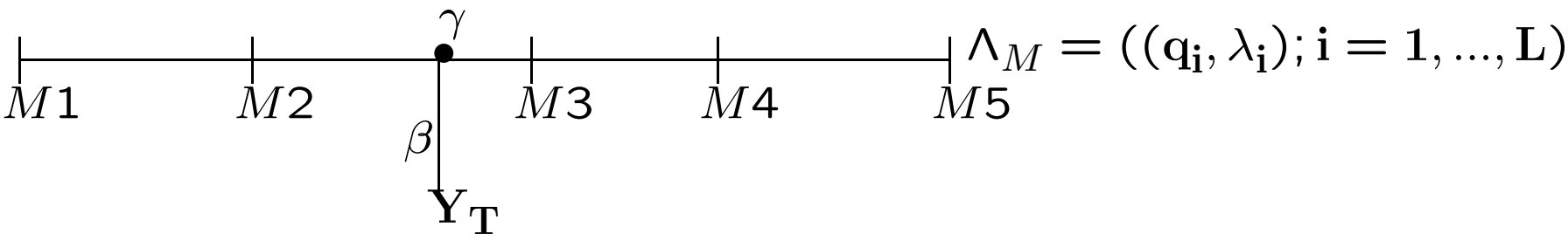
- Crossovers arise as a Poisson process rate 1 (per Morgan).

- There is a recombination between two loci if there is an odd number W of crossovers between them: $W(d) \sim \mathcal{P}(d)$.

- Hence the Haldane map function: $\rho(d) = (1/2)(1 - \exp(-2d))$.

- The key thing is the model: the map function just puts loci onto a linear location map. (See later: MCMC under interference.)

WHAT AND WHY THE LOCATION LOD SCORE



Parameter $\xi = (\beta, \gamma, \Lambda_M)$. Data $\mathbf{Y} = (\mathbf{Y_M}, \mathbf{Y_T})$

$$\text{lod}(\gamma) \;=\; \log_{10}\left(\frac{\Pr(\mathbf{Y}; \mathbf{\Lambda_M}, \beta, \gamma)}{\Pr(\mathbf{Y}; \mathbf{\Lambda_M}, \beta, \gamma = \infty)}\right)$$

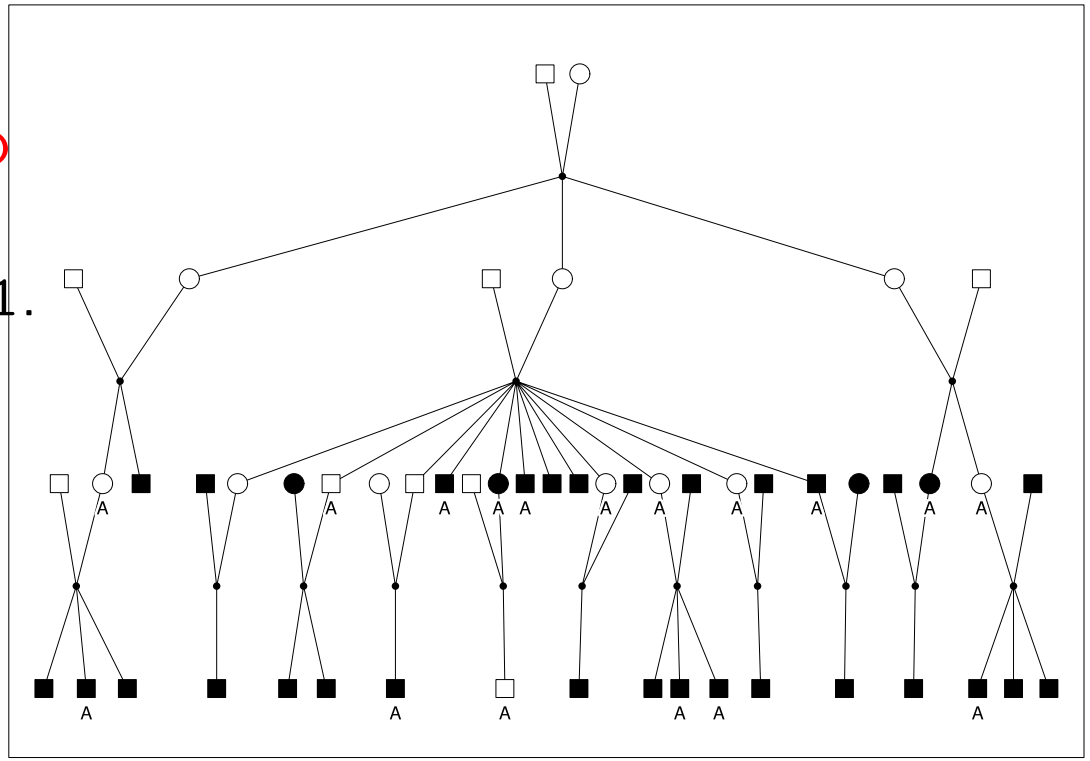
Trait locus location γ is parameter of interest:

$\gamma = \infty$ is no linkage.

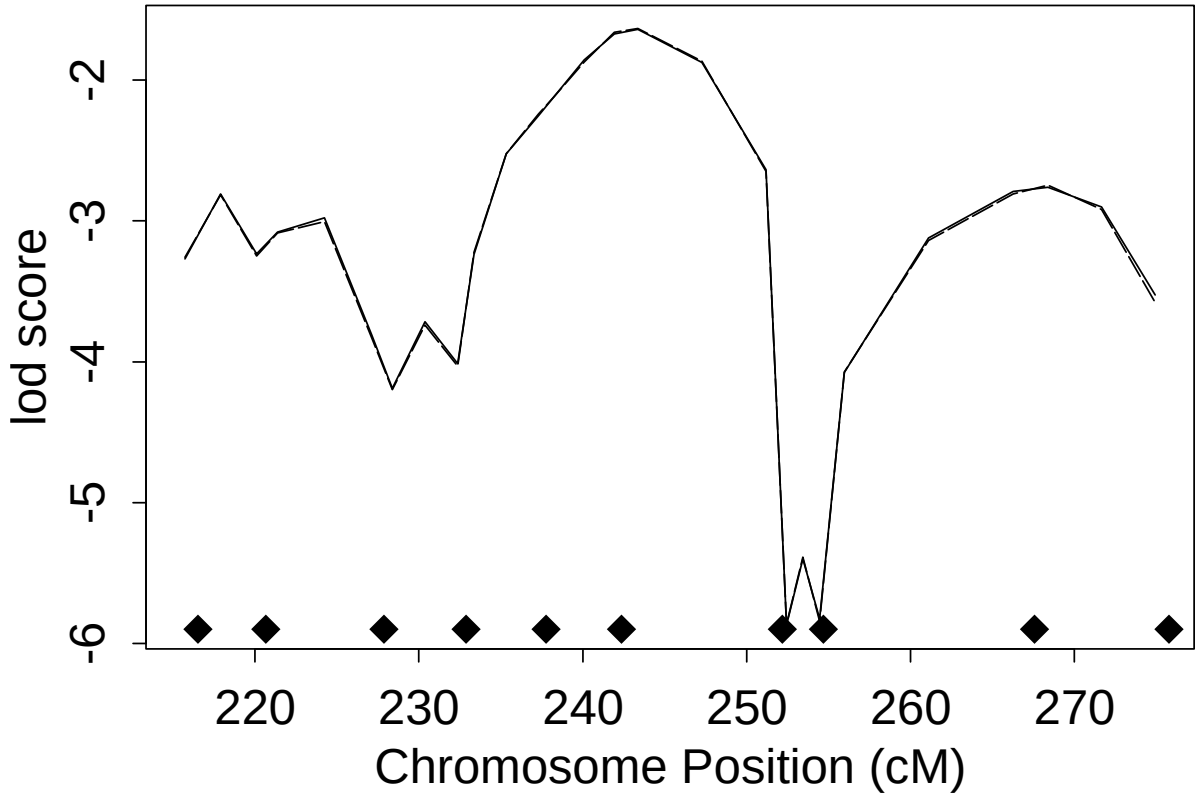
Exact computation is infeasible

AN EXAMPLE PEDIGREE: APPROXIMATED

Marker data
are SIMULATED
at 10 linked
markers on Chr 1.
Trait is close
to M6



AN EXAMPLE MULTIPOINT LOD SCORE



MONTE CARLO LIKELIHOODS ON PEDIGREES

- Monte Carlo estimates expectations.

- $$L(\xi) \;=\; P_\xi(\mathbf{Y}) \;=\; \sum_{\mathbf{S}} \mathbf{P}_\xi(\mathbf{S},\mathbf{Y}) \;=\; \sum_{\mathbf{S}} \mathbf{P}_\xi(\mathbf{Y} \mid \mathbf{S})\; \mathbf{P}_\xi(\mathbf{S})$$

for parameters ξ and latent variables $\mathbf{S}$.

- Simple (but not useful) example:

$$L(\xi) \;=\; \mathsf{E}_\xi(P_\xi(\mathbf{Y} \mid \mathbf{S}))$$

- More generally

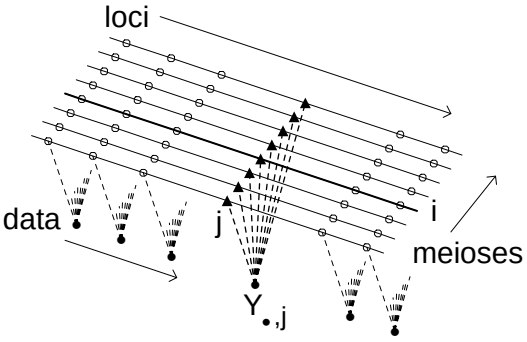
$$L(\xi) \;=\; \sum_{\mathbf{S}} \left(\frac{P_\xi(\mathbf{S},\; \mathbf{Y})}{P^*(\mathbf{S})}\right) P^*(\mathbf{S}) \;=\; \mathsf{E}_{P^*}\left(\frac{P_\xi(\mathbf{S},\; \mathbf{Y})}{P^*(\mathbf{S})}\right)$$

provided $P^*(\mathbf{S}) > 0$ if $P_\xi(\mathbf{S},\; \mathbf{Y}) > \mathbf{0}$.

SEQUENTIAL IMPUTATION OVER LOCI

Choose the sampling distribution:

$$\begin{aligned} P^*(S_{\bullet,j}) &= P_{\xi_0}(S_{\bullet,j} \mid S^{*(j-1)}, Y^{(j)}) \\ &= P_{\xi_0}(S_{\bullet,j} \mid S^*_{\bullet,1}, \ldots S^*_{\bullet,j-1}, \\ &\qquad Y_{\bullet,1}, \ldots, Y_{\bullet,j-1}, Y_{\bullet,j}) \\ &= P_{\xi_0}(S_{\bullet,j} \mid S^*_{\bullet,j-1}, Y_{\bullet,j}) \end{aligned}$$



Now:

$$\begin{aligned} P_{\xi_0}(S_{\bullet,j} \mid S^{*(j-1)}, Y^{(j)}) &= \frac{P_{\xi_0}(S_{\bullet,j}, Y_{\bullet,j} \mid S^{*(j-1)}, Y^{(j-1)})}{P_{\xi_0}(Y_{\bullet,j} \mid S^{*(j-1)}, Y^{(j-1)})} \\ &= \frac{P_{\xi_0}(S_{\bullet,j}, Y_{\bullet,j} \mid S^{*(j-1)}, Y^{(j-1)})}{w_j}. \end{aligned}$$

where, by pedigree-peeling, we can compute

$$w_j = P_{\xi_0}(Y_{\bullet,j} \mid Y^{(j-1)}, S^{*(j-1)}) = P_{\xi_0}(Y_{\bullet,j} \mid S^*_{\bullet,j-1}).$$

MONTE CARLO LIKELIHOOD ESTIMATE

Thus sequential imputation distribution is

$$P^*(\mathbf{S}^*) \;=\; \prod_{j=1}^{L} P_{\xi_0}(S_{\bullet,j} \mid S^{*(j-1)}, Y^{(j)}) \;=\; \frac{P_{\xi_0}(\mathbf{S}^*, \mathbf{Y})}{W_L(\mathbf{S}^*)}$$

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

Now

$$L(\xi_0) \;=\; P_{\xi_0}(\mathbf{Y}) \;=\; \mathsf{E}_{P^*}\left(\frac{P_{\xi}(\mathbf{S},\ \mathbf{Y})}{P^*(\mathbf{S})}\right) \;=\; \mathsf{E}_{P^*}(W_L(\mathbf{S}^*))$$

Given N realizations $\mathbf{S}^{(\tau)}$ the estimate of $L(\xi_0)$ is $N^{-1}\sum_{\tau} W_L(\mathbf{S}^{(\tau)})$.

THE IDEAL SAMPLING DISTRIBUTION

- We want $P^*(\mathbf{S})$ close to proportional to $P_{\xi_0}(\mathbf{Y},\mathbf{S})$
  – that is $P^*(\mathbf{S}) \approx P_{\xi_0}(\mathbf{S}|\mathbf{Y})$.

- Of course we cannot achieve this, else Monte Carlo would be unnecessary.

- Suppose we use MCMC to sample $\mathbf{S}$ from $P_{\xi_0}(\mathbf{S}|\mathbf{Y})$.

$$
\begin{aligned}
P_\xi(\mathbf{Y}) &= \sum_{\mathbf{S}} P_\xi(\mathbf{Y},\mathbf{S}) = \sum_{\mathbf{S}} \frac{P_\xi(\mathbf{Y},\mathbf{S})}{P_{\xi_0}(\mathbf{S} \mid \mathbf{Y})} \mathbf{P_{\xi_0}}(\mathbf{S} \mid \mathbf{Y}) \\
&= \mathsf{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y},\ \mathbf{S})}{P_{\xi_0}(\mathbf{S} \mid \mathbf{Y})} \mid \mathbf{Y} \right) \\
&= P_{\xi_0}(\mathbf{Y})\ \mathsf{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y},\mathbf{S})}{P_{\xi_0}(\mathbf{Y},\mathbf{S})} \mid \mathbf{Y} \right)
\end{aligned}
$$

LIKELIHOOD RATIO ESTIMATION

Thus we have

$$\frac{L(\xi)}{L(\xi_0)} = \frac{P_\xi(\mathbf{Y})}{P_{\xi_0}(\mathbf{Y})} = \mathsf{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y},\mathbf{S})}{P_{\xi_0}(\mathbf{Y},\mathbf{S})} \mid \mathbf{Y}\right)$$

$\mathbf{S}$ is the random variable, $\mathbf{Y}$ is fixed. $\mathbf{S} \sim P_{\xi_0}(\cdot|\mathbf{Y})$.
If $\mathbf{S}^{(\tau)}$, $\tau = 1, ..., N$, are realized from $P_{\xi_0}(\cdot|\mathbf{Y})$ then the likelihood ratio can be estimated by

$$\frac{1}{N}\sum_{\tau=1}^{N}\left(\frac{P_\xi(\mathbf{Y},\mathbf{S}^{(\tau)})}{P_{\xi_0}(\mathbf{Y},\mathbf{S}^{(\tau)})}\right)$$

LINKAGE LOCATION LIKELIHOOD RATIO

The form for linkage lod that follows directly from this is

$$\frac{L(\beta,\ \gamma_1,\ \Lambda_M)}{L(\beta,\ \gamma_0,\ \Lambda_M)} \ = \ \mathsf{E}_{\xi_0}\left(\frac{P_{\xi_1}(\mathbf{Y_T},\mathbf{Y_M},\mathbf{S_T},\mathbf{S_M})}{P_{\xi_0}(\mathbf{Y_T},\mathbf{Y_M},\mathbf{S_T},\mathbf{S_M})} \ \mid \ \mathbf{Y_T},\mathbf{Y_M}\right)$$

for two hypothesized trait locus positions γ_1 and γ_0.

Now $P_\xi(\mathbf{Y},\mathbf{S}) = \mathbf{P}_\beta(\mathbf{Y_T}|\mathbf{S_T})\mathbf{P}_{\mathbf{\Lambda_M}}(\mathbf{Y_M},\mathbf{S_M})\mathbf{P}_\gamma(\mathbf{S_T}|\mathbf{S_M})$ so ratio reduces to

$$\frac{L(\beta,\ \gamma_1,\ \Lambda_M)}{L(\beta,\ \gamma_0,\ \Lambda_M)} \ = \ \mathsf{E}_{\xi_0}\left(\frac{P_{\gamma_1}(\mathbf{S}_T \mid \mathbf{S}_M)}{P_{\gamma_0}(\mathbf{S}_T \mid \mathbf{S}_M)} \ \mid \ \mathbf{Y_T},\mathbf{Y_M}\right)$$

LOCAL ESTIMATE IS VERY SIMPLE: GLOBAL IS HARD

$$\frac{P_{\gamma_1}(\mathbf{S}_T \mid \mathbf{S}_M)}{P_{\gamma_0}(\mathbf{S}_T \mid \mathbf{S}_M)} = \prod_i \left[\left(\frac{\rho_{1l}}{\rho_{0l}}\right)^{|S_{i,T}-S_{i,l}|} \left(\frac{1-\rho_{1l}}{1-\rho_{0l}}\right)^{1-|S_{i,T}-S_{i,l}|} \left(\frac{\rho_{1r}}{\rho_{0r}}\right)^{|S_{i,T}-S_{i,r}|} \left(\frac{1-\rho_{1r}}{1-\rho_{0r}}\right)^{1-|S_{i,T}-S_{i,r}|} \right]$$

- The above works well only for $\gamma_1 \approx \gamma_0$, and for γ_0, γ_1 with same l and r.

- When likelihoods are not smooth, combining LR estimates does not work well – especially across markers.

AN MCMC IMPORTANCE SAMPLING ESTIMATE

Lange and Sobel (1996) write the likelihood in the form

$$\begin{aligned} L(\beta,\gamma,\Lambda_M) &= P_{\beta,\gamma,\Lambda_M}(\mathbf{Y_M},\mathbf{Y_T}) \;\propto\; \mathbf{P}_{\beta,\gamma,\mathbf{\Lambda_M}}(\mathbf{Y_T} \mid \mathbf{Y_M}) \\ &= \sum_{\mathbf{S}_M} P_{\beta,\gamma}(\mathbf{Y_T} \mid \mathbf{S_M})\mathbf{P}_{\mathbf{\Lambda_M}}(\mathbf{S_M} \mid \mathbf{Y_M}) \\ &= \mathsf{E}_{\Lambda_M}(P_{\beta,\gamma}(\mathbf{Y_T} \mid \mathbf{S_M}) \mid \mathbf{Y_M}). \end{aligned}$$

- Sample $\mathbf{S}_M$ given $\mathbf{Y_M}$: compute $P(Y_T|\mathbf{S}_M)\ \forall\ \beta,\gamma$
  a form of Rao-Blackwellization – integrate over S_T.
- Also importance sampling: maybe $P(S_M|Y_M) \approx P(S_M|Y_M,Y_T)$
- For "fuzzy" traits it works quite well.

METROPOLIS HASTINGS FOR INTERFERENCE

- Suppose we have interference model $P^{(I)}(\mathbf{S})$ in place of Haldane model $P^{(H)}(\mathbf{S})$ we have used so far.

- Use block-Gibbs update of meiosis i ($S_{i,\bullet}$) to propose $\mathbf{S}^\dagger$.

- Hastings ratio is for current $\mathbf{S}$ and proposed $\mathbf{S}^\dagger$ is

$$\begin{aligned} h(\mathbf{S}^\dagger;\mathbf{S}) &= \frac{P^{(I)}(\mathbf{S}^\dagger,\mathbf{Y})}{P^{(I)}(\mathbf{S},\mathbf{Y})}\,\frac{P^{(H)}(S_{i,\bullet} \mid S_{k,\bullet}, k\neq i,\mathbf{Y})}{P^{(H)}(S^\dagger_{i,\bullet} \mid S_{k,\bullet}, k\neq i,\mathbf{Y})} \\ &= \frac{P^{(I)}(\mathbf{S}^\dagger,\mathbf{Y})P^{(H)}(\mathbf{S},\mathbf{Y})}{P^{(I)}(\mathbf{S},\mathbf{Y})P^{(H)}(\mathbf{S}^\dagger,\mathbf{Y})} \\ &= \frac{P(\mathbf{Y}|\mathbf{S}^\dagger)P^{(I)}(\mathbf{S}^\dagger)P(\mathbf{Y}|\mathbf{S})P^{(H)}(\mathbf{S})}{P(\mathbf{Y}|\mathbf{S})P^{(I)}(\mathbf{S})P(\mathbf{Y}|\mathbf{S}^\dagger)P^{(H)}(\mathbf{S}^\dagger)} \end{aligned}$$

INTERFERENCE ctd.

$$h(\mathbf{S}^\dagger;\mathbf{S}) = \prod_{k=1}^{m} \frac{P^{(I)}(S^\dagger_{k,\bullet})}{P^{(I)}(S_{k,\bullet})} \frac{P^{(H)}(S_{k,\bullet})}{P^{(H)}(S^\dagger_{k,\bullet})}$$

$$= \frac{P^{(I)}(S^\dagger_{i,\bullet})}{P^{(I)}(S_{i,\bullet})} \frac{P^{(H)}(S_{i,\bullet})}{P^{(H)}(S^\dagger_{i,\bullet})}.$$

- $\Pr(\mathbf{S}^* = \mathbf{S}^\dagger) = a = \min(1, h)$. $\Pr(\mathbf{S}^* = \mathbf{S}) = 1 - a$.

- Question: better to sample under H and reweight, or use M-H to sample under model I?