Bayesian Interval Mapping

1. Bayesian strategy 3-19
2. Markov chain sampling 20-27
3. sampling genetic architectures 28-35
4. criteria for model selection 36-44

QTL model selection: key players

- observed measurements
  - \( y \) = phenotypic trait
  - \( m \) = markers & linkage map
  - \( i \) = individual index (1,…,n)
- missing data
  - missing marker data
  - \( q \) = QT genotypes
    - alleles QQ, Qq, or qq at locus
- unknown quantities
  - \( \lambda \) = QT locus (or loci)
  - \( \mu \) = phenotype model parameters
  - \( \gamma \) = QTL model/genetic architecture
- \( p(\gamma | m, \lambda, \gamma) \) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for \( q \) given \( m \)
- \( p(y | q, \mu, \gamma) \) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters \( \mu \) (could be non-parametric)

after Sen Churchill (2001)
1. Bayesian strategy for QTL study

- augment data \((y, m, q)\) with missing genotypes \(q\)
- study unknowns \((\mu, \lambda, \gamma)\) given augmented data \((y, m, q)\)
  - find better genetic architectures \(\gamma\)
  - find most likely genomic regions = QTL = \(\lambda\)
  - estimate phenotype parameters = genotype means = \(\mu\)
- sample from posterior in some clever way
  - multiple imputation (Sen Churchill 2002)
  - Markov chain Monte Carlo (MCMC)
    - (Satagopan et al. 1996; Yi et al. 2005, 2007)

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{constant}}
\]

posterior for \(q, \mu, \lambda, \gamma\) = \(\frac{\text{phenotype likelihood} \times \text{prior for } q, \mu, \lambda, \gamma}{\text{constant}}\)

\[
\Pr(q, \mu, \lambda, \gamma | y, m) = \frac{\Pr(y | q, \mu, \lambda, \gamma) \times \Pr(q | m, \lambda, \gamma) \Pr(\mu | \gamma) \Pr(\lambda | m, \gamma) \Pr(\gamma)}{\Pr(y | m)}
\]
Bayes posterior for normal data

model \[ y_i = \mu + e_i \]

environment \[ e \sim N(0, \sigma^2), \sigma^2 \text{ known} \]

likelihood \[ y \sim N(\mu, \sigma^2) \]

prior \[ \mu \sim N(\mu_0, \kappa \sigma^2), \kappa \text{ known} \]

posterior: mean tends to sample mean

single individual \[ \mu \sim N(\mu_0 + b_1(y_1 - \mu_0), \sigma^2) \]

sample of \( n \) individuals \[ \mu \sim N(b_n \overline{y}_n + (1 - b_n) \mu_0, b_n \sigma^2 / n) \]

with \( \overline{y}_n = \text{sum}_{i=1,...,n} y_i / n \)

shrinkage factor
(shrinks to 1)

\[ b_n = \frac{\kappa n}{\kappa n + 1} \rightarrow 1 \]

what values are the genotypic means?
phenotype model \( pr(y|q, \mu) \)
Bayes posterior QTL means

posterior centered on sample genotypic mean
but shrunken slightly toward overall mean

phenotype mean:

\[ E(y | q) = \mu_q \quad V(y | q) = \sigma^2 \]

genotypic prior:

\[ E(\mu_q) = \bar{y} \quad V(\mu_q) = \kappa\sigma^2 \]

posterior:

\[ E(\mu_q | y) = b_q \bar{y}_q + (1 - b_q)\bar{y} \quad V(\mu_q | y) = b_q \sigma^2 / n_q \]

\[ n_q = \text{count}\{q_i = q\} \quad \bar{y}_q = \text{sum} y_i / n_q \]

shrinkage:

\[ b_q = \frac{\kappa n_q}{\kappa n_q + 1} \to 1 \]

partition genotypic effects
on phenotype

- phenotype depends on genotype
- genotypic value partitioned into
  - main effects of single QTL
  - epistasis (interaction) between pairs of QTL

\[ \mu_q = \beta_0 + \beta_q = E(Y; q) \]
\[ \beta_q = \beta(q_2) + \beta(q_2) + \beta(q_1, q_2) \]
partition genotypic variance

- consider same 2 QTL + epistasis

- centering variance
  \[ V(\beta_0) = \kappa_0 \sigma^2 = s^2 \]

- genotypic variance
  \[ V(\beta_q) = \kappa_1 \sigma^2 = \sigma_q^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}^2 \]

- heritability
  \[ h_q^2 = \frac{\sigma_q^2}{\sigma_q^2 + \sigma^2} = h_1^2 + h_2^2 + h_{12}^2 \]

posterior mean \( \approx \) LS estimate

\[ \beta_q \mid y \sim N(b_q \hat{\beta}_q, b_q C_q \sigma^2) \]

\[ \approx N(\hat{\beta}_q, C_q \sigma^2) \]

LS estimate \( \hat{\beta}_q = \text{sum}_i [\text{sum}_j \hat{\beta}(q_{ij})] = \text{sum}_i w_{qi} y_i \)

variance \( V(\hat{\beta}_q) = \text{sum}_i w_{qi}^2 \sigma^2 = C_q \sigma^2 \)

shrinkage \( b_q = \kappa_1 / (\kappa_1 + C_q) \to 1 \)
\[ \text{pr}(q|m, \lambda) \text{ recombination model} \]

\[
\text{pr}(q|m, \lambda) = \text{pr(geno} | \text{ map, locus}) \approx \\
\text{pr(geno} | \text{ flanking markers, locus})
\]

\[ m_1 \quad m_2 \quad q? \quad m_3 \quad m_4 \quad m_5 \quad m_6 \]

\[ \lambda \quad \text{distance along chromosome} \]
what are likely QTL genotypes $q$? 
how does phenotype $y$ improve guess?

posterior on QTL genotypes $q$

- full conditional of $q$ given data, parameters
  - proportional to prior $\text{pr}(q | m, \lambda)$
    - weight toward $q$ that agrees with flanking markers
  - proportional to likelihood $\text{pr}(y | q, \mu)$
    - weight toward $q$ with similar phenotype values
  - posterior recombination model balances these two
- this is the E-step of EM computations

$$
\text{pr}(q | y, m, \mu, \lambda) = \frac{\text{pr}(y | q, \mu) \cdot \text{pr}(q | m, \lambda)}{\text{pr}(y | m, \mu, \lambda)}
$$
Where are the loci $\lambda$ on the genome?

- prior over genome for QTL positions
  - flat prior = no prior idea of loci
  - or use prior studies to give more weight to some regions
- posterior depends on QTL genotypes $q$
  \[
  \text{pr}(\lambda \mid m, q) = \frac{\text{pr}(\lambda) \cdot \text{pr}(q \mid m, \lambda)}{\text{constant}}
  \]
  - constant determined by averaging
    - over all possible genotypes $q$
    - over all possible loci $\lambda$ on entire map
- no easy way to write down posterior

what is the genetic architecture $\gamma$?

- which positions correspond to QTLs?
  - priors on loci (previous slide)
- which QTL have main effects?
  - priors for presence/absence of main effects
    - same prior for all QTL
    - can put prior on each d.f. (1 for BC, 2 for F2)
- which pairs of QTL have epistatic interactions?
  - prior for presence/absence of epistatic pairs
    - depends on whether 0, 1, 2 QTL have main effects
    - epistatic effects less probable than main effects
\[ \gamma = \text{genetic architecture:} \]

\text{loci:}
- main QTL
- epistatic pairs

\text{effects:}
- add, dom
- aa, ad, dd

\text{Bayesian priors & posteriors}

- augmenting with missing genotypes \( q \)
  - prior is recombination model
  - posterior is (formally) E step of EM algorithm
- sampling phenotype model parameters \( \mu \)
  - prior is “flat” normal at grand mean (no information)
  - posterior shrinks genotypic means toward grand mean
  - (details for unexplained variance omitted here)
- sampling QTL loci \( \lambda \)
  - prior is flat across genome (all loci equally likely)
- sampling QTL genetic architecture model \( \gamma \)
  - number of QTL
    - prior is Poisson with mean from previous IM study
  - genetic architecture of main effects and epistatic interactions
    - priors on epistasis depend on presence/absence of main effects
2. Markov chain sampling

- construct Markov chain around posterior
  - want posterior as stable distribution of Markov chain
  - in practice, the chain tends toward stable distribution
    - initial values may have low posterior probability
    - burn-in period to get chain mixing well
- sample QTL model components from full conditionals
  - sample locus $\lambda$ given $q, \gamma$ (using Metropolis-Hastings step)
  - sample genotypes $q$ given $\lambda, \mu, \gamma$ (using Gibbs sampler)
  - sample effects $\mu$ given $q, y, \gamma$ (using Gibbs sampler)
  - sample QTL model $\gamma$ given $\lambda, \mu, y, q$ (using Gibbs or M-H)

\[
(\lambda, q, \mu, \gamma) \sim p(\lambda, q, \mu, \gamma \mid y, m)
\]

\[
(\lambda, q, \mu, \gamma)_1 \rightarrow (\lambda, q, \mu, \gamma)_2 \rightarrow \cdots \rightarrow (\lambda, q, \mu, \gamma)_N
\]

MCMC sampling of unknowns $(q, \mu, \lambda)$ for given genetic architecture $\gamma$

- Gibbs sampler
  - genotypes $q$
  - effects $\mu$
  - not loci $\lambda$

\[
q \sim p(q \mid y, m, \mu, \lambda)
\]

\[
\mu \sim \frac{p(y \mid q, \mu)p(\mu)}{p(y \mid q)}
\]

\[
\lambda \sim \frac{p(q \mid m, \lambda)p(\lambda \mid m)}{p(q \mid m)}
\]

- Metropolis-Hastings sampler
  - extension of Gibbs sampler
  - does not require normalization
    - $p( q \mid m ) = \sum_\lambda p( q \mid m, \lambda ) \cdot p(\lambda)$
Gibbs sampler for two genotypic means

- want to study two correlated effects
  - could sample directly from their bivariate distribution
  - assume correlation $\rho$ is known
- instead use Gibbs sampler:
  - sample each effect from its full conditional given the other
  - pick order of sampling at random
  - repeat many times

$$
\begin{align*}
(\mu_1, \mu_2) &\sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \\
\mu_1 &\sim N \left( \rho \mu_2, 1 - \rho^2 \right) \\
\mu_2 &\sim N \left( \rho \mu_1, 1 - \rho^2 \right)
\end{align*}
$$

Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

$N = 200$ samples
full conditional for locus

- cannot easily sample from locus full conditional
  \[ \text{pr}(\lambda | y, m, \mu, q) = \text{pr}(\lambda | m, q) = \text{pr}(q | m, \lambda) \text{pr}(\lambda) / \text{constant} \]
- constant is very difficult to compute explicitly
  - must average over all possible loci \( \lambda \) over genome
  - must do this for every possible genotype \( q \)
- Gibbs sampler will not work in general
  - but can use method based on ratios of probabilities
  - Metropolis-Hastings is extension of Gibbs sampler

Metropolis-Hastings idea

- want to study distribution \( f(\lambda) \)
  - take Monte Carlo samples
    - unless too complicated
    - take samples using ratios of \( f \)
- Metropolis-Hastings samples:
  - propose new value \( \lambda^* \)
    - near (?) current value \( \lambda \)
    - from some distribution \( g \)
  - accept new value with prob \( a \)
    - Gibbs sampler: \( a = 1 \) always

\[
a = \min \left( 1, \frac{f(\lambda^*)g(\lambda^* - \lambda)}{f(\lambda)g(\lambda - \lambda^*)} \right)
\]
Metropolis-Hastings for locus $\lambda$

added twist: occasionally propose from entire genome

Metropolis-Hastings samples

$N = 200$ samples

$N = 1000$ samples
3. sampling genetic architectures

• search across genetic architectures $A$ of various sizes
  – allow change in number of QTL
  – allow change in types of epistatic interactions

• methods for search
  – reversible jump MCMC
  – Gibbs sampler with loci indicators

• complexity of epistasis
  – Fisher-Cockerham effects model
  – general multi-QTL interaction & limits of inference

reversible jump MCMC

• consider known genotypes $q$ at 2 known loci $\lambda$
  – models with 1 or 2 QTL

• M-H step between 1-QTL and 2-QTL models
  – model changes dimension (via careful bookkeeping)
  – consider mixture over QTL models $H$

\[
\gamma = 1 \text{QTL} : Y = \beta_0 + \beta(q_1) + e \\
\gamma = 2 \text{QTL} : Y = \beta_0 + \beta(q_1) + \beta(q_2) + e
\]
geometry of reversible jump

Move Between Models

Reversible Jump Sequence

c21 = 0.7

m=2

m=1

geometry allowing $q$ and $\lambda$ to change

a short sequence

first 1000 with $m<3$
collinear QTL = correlated effects

4-week

-0.6 -0.4 -0.2 0.0 0.2

-0.6 -0.4 -0.2 0.0

additive 1

additive 2

cor = -0.81

8-week

effect 1
effect 2

cor = -0.7

-0.2 -0.1 0.0 0.1 0.2

-0.3 -0.2 -0.1 0.0

• linked QTL = collinear genotypes
  ➢ correlated estimates of effects (negative if in coupling phase)
  ➢ sum of linked effects usually fairly constant

sampling across QTL models $\gamma$

$0 \rightarrow \lambda_1 \rightarrow \lambda_{m+1} \rightarrow \lambda_2 \rightarrow \cdots \lambda_m \rightarrow L$

action steps: draw one of three choices

• update QTL model $\gamma$ with probability $1-b(\gamma)\cdot d(\gamma)$
  – update current model using full conditionals
  – sample QTL loci, effects, and genotypes

• add a locus with probability $b(\gamma)$
  – propose a new locus along genome
  – innovate new genotypes at locus and phenotype effect
  – decide whether to accept the “birth” of new locus

• drop a locus with probability $d(\gamma)$
  – propose dropping one of existing loci
  – decide whether to accept the “death” of locus
Gibbs sampler with loci indicators

- consider only QTL at pseudomarkers
  - every 1-2 cM
  - modest approximation with little bias
- use loci indicators in each pseudomarker
  - $\gamma = 1$ if QTL present
  - $\gamma = 0$ if no QTL present
- Gibbs sampler on loci indicators $\gamma$
  - relatively easy to incorporate epistasis
  - Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
    - (see earlier work of Nengjun Yi and Ina Hoeschele)

\[
\mu_q = \mu + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_2), \quad \gamma_k = 0,1
\]

Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for $\lambda_j$ depends on $\gamma$
  - $0 \leq \gamma \leq 1$ (grey scale)
  - shrink most $\gamma$s to zero
- Wang et al. (2005 *Genetics*)
  - Shizhong Xu group at U CA Riverside

\[
\mu_q = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_2), \quad 0 \leq \gamma_k \leq 1
\]
4. criteria for model selection
balance fit against complexity

• classical information criteria
  – penalize likelihood $L$ by model size $|\gamma|$  
    
    \[ IC = -2 \log L(\gamma \mid y) + \text{penalty}(\gamma) \]  
    – maximize over unknowns

• Bayes factors
  – marginal posteriors $\text{pr}(y \mid \gamma)$  
    – average over unknowns

classical information criteria

• start with likelihood $L(\gamma \mid y, m)$  
  – measures fit of architecture ($\gamma$) to phenotype ($y$)
    • given marker data ($m$)
  – genetic architecture ($\gamma$) depends on parameters
    • have to estimate loci ($\mu$) and effects ($\lambda$)

• complexity related to number of parameters
  – $|\gamma|$ = size of genetic architecture
    • BC:  
      \[ |\gamma| = 1 + n.qtl + n.qtl(n.qtl - 1) = 1 + 4 + 12 = 17 \]  
    • F2:  
      \[ |\gamma| = 1 + 2n.qtl + 4n.qtl(n.qtl - 1) = 1 + 8 + 48 = 57 \]
classical information criteria

- construct information criteria
  - balance fit to complexity
  - Akaike: AIC = \(-2 \log(L) + 2 |\gamma|\)
  - Bayes/Schwartz: BIC = \(-2 \log(L) + |\gamma| \log(n)\)
  - Broman: BIC_δ = \(-2 \log(L) + \delta |\gamma| \log(n)\)
  - general form: IC = \(-2 \log(L) + |\gamma| D(n)\)

- compare models
  - hypothesis testing: designed for one comparison
    - 2 log[LR(\gamma_1, \gamma_2)] = L(y|m, \gamma_2) - L(y|m, \gamma_1)
  - model selection: penalize complexity
    - IC(\gamma_1, \gamma_2) = 2 log[LR(\gamma_1, \gamma_2)] + (|\gamma_2| - |\gamma_1|) D(n)

information criteria vs. model size

- WinQTL 2.0
- SCD data on F2
- A = AIC
- 1 = BIC(1)
- 2 = BIC(2)
- d = BIC(δ)
- models
  - 1, 2, 3, 4 QTL
    - 2 + 5 + 9 + 2
  - epistasis
    - 2:2 AD
Bayes factors

- ratio of model likelihoods
  - ratio of posterior to prior odds for architectures
  - averaged over unknowns

\[
B_{12} = \frac{\text{pr}(\gamma_1 \mid y, m) / \text{pr}(\gamma_2 \mid y, m)}{\text{pr}(\gamma_1) / \text{pr}(\gamma_2)} = \frac{\text{pr}(y \mid m, \gamma_1)}{\text{pr}(y \mid m, \gamma_2)}
\]

- roughly equivalent to BIC
  - BIC maximizes over unknowns
  - BF averages over unknowns

\[
-2 \log(B_{12}) = -2 \log(LR) - (|\gamma_2| - |\gamma_1|) \log(n)
\]
issues in computing Bayes factors

• $BF$ insensitive to shape of prior on $\gamma$
  – geometric, Poisson, uniform
  – precision improves when prior mimics posterior

• $BF$ sensitivity to prior variance on effects $\theta$
  – prior variance should reflect data variability
  – resolved by using hyper-priors
    • automatic algorithm; no need for user tuning

• easy to compute Bayes factors from samples
  – sample posterior using MCMC
  – posterior $\text{pr}(\gamma \mid y, m)$ is marginal histogram

Bayes factors & genetic architecture $\gamma$

• $|\gamma| =$ number of QTL
  – prior $\text{pr}(\gamma)$ chosen by user
  – posterior $\text{pr}(\gamma \mid y, m)$
    • sampled marginal histogram
    • shape affected by prior $\text{pr}(A)$

\[
BF_{\gamma_1, \gamma_2} = \frac{\text{pr}(\gamma_1 \mid y, m)/\text{pr}(\gamma_1)}{\text{pr}(\gamma_2 \mid y, m)/\text{pr}(\gamma_2)}
\]

• pattern of QTL across genome
• gene action and epistasis
BF sensitivity to fixed prior for effects

\[ \beta_{ij} \sim N \left( 0, \sigma_G^2 / m \right), \sigma_G^2 = h^2 \sigma_{total}^2, \quad h^2 \text{ fixed} \]

BF insensitivity to random effects prior

\[ \beta_{ij} \sim N \left( 0, \sigma_G^2 / m \right), \sigma_G^2 = h^2 \sigma_{total}^2, \quad \frac{1}{2} h^2 \sim \text{Beta}(a,b) \]