Bayesian Model Selection
for Multiple QTL

Brian S. Yandell
University of Wisconsin-Madison
www.stat.wisc.edu/~yandell/statgen

Animal Genetics Seminar, October 2006

outline

1. Bayesian vs. classical QTL study
2. Bayesian priors & posteriors
3. model search using MCMC
   - Gibbs sampler and Metropolis-Hastings
4. model assessment
   - Bayes factors & model averaging
5. data examples in detail
   - simulation & hyper data
1. Bayesian vs. classical QTL study

- **classical study**
  - maximize likelihood over unknowns
  - test for presence/absence of QTL at loci
  - model selection in stepwise fashion

- **Bayesian study**
  - sample unknowns from posterior
  - estimate QTL loci directly
  - sample simultaneously across models

Bayesian QTL: key players

- observed measurements
  - \( y \) = phenotypic trait
  - \( m \) = markers & linkage map
  - \( i = 1, \ldots, n \) = individual index

- 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
  - \( H \) = QTL model/genetic architecture

- \( p(r(q|m, \lambda, H)) \) genotype model
  - grounded by linkage map, experimental cross
  - recombination yields multinomial for \( q \) given \( m \)

- \( p(r(y|q, \mu, H)) \) phenotype model
  - distribution shape (assumed normal here)
  - unknown parameters \( \mu \) (could be non-parametric)

after
Sen Churchill (2001)
Bayes posterior vs. maximum likelihood

- **LOD**: classical Log ODds
  - maximizes likelihood
  - mixture over missing QTL genotypes $q$
  - maximize phenotype model parameters $\mu$
  - scan over possible loci $\lambda$
  - R/qtl scanone/scantwo: method = "em"

- **LPD**: Bayesian Log Posterior Density
  - averages over unknowns
  - average over missing QTL genotypes $q$
  - average phenotype model parameters $\mu$
  - scan over possible loci $\lambda$
  - R/qtl scanone/scantwo: method = "imp"

**suppose genetic architecture is known**
- $H = 1$ QTL or 2 QTL model
- available in R/qtl via scanone and scantwo routines

- $\text{LOD}(\lambda) = \log_{10} \{\max_{\mu} \text{pr}(y \mid m, \mu, \lambda)\} + c$
- $\text{LPD}(\lambda) = \log_{10} \{\text{pr}(\lambda \mid m) \int \text{pr}(y \mid m, \mu, \lambda) \text{pr}(\mu) d\mu\} + C$

with mixture over missing QTL genotypes:
- $\text{pr}(y \mid m, \mu, \lambda) = \sum_q \text{pr}(y \mid q, \mu) \text{pr}(q \mid m, \lambda)$
LOD & LPD: 1 QTL
n.ind = 100, 10 cM marker spacing

LOD & LPD: 1 QTL
n.ind = 100, 1 cM marker spacing
Bayesian strategy for QTL study

- augment data \((y,m)\) with missing genotypes \(q\)
- study unknowns \((\mu, \lambda, H)\) given augmented data \((y,m,q)\)
  - find better genetic architectures \(H\)
  - 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) (Yi et al. 2005)

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

\[
\text{posterior for } q, \mu, \lambda, H = \frac{\text{phenotype likelihood} \times \text{prior for } q, \mu, \lambda, H}{\text{constant}}
\]

\[
\text{pr}(q, \mu, \lambda, H \mid y, m) = \frac{\text{pr}(y \mid q, \mu, H) \times \text{pr}(q \mid m, \lambda, H) \times \text{pr}(\mu \mid H) \times \text{pr}(\lambda \mid m, H) \times \text{pr}(H)}{\text{pr}(y \mid m)}
\]

2. 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 model \(H\)
  - 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
what are likely QTL genotypes $q$?
how does phenotype $y$ improve guess?
(consider locus $\lambda$ halfway between D4Mit41 & D4Mit214)

Pr($q=AA$) = 0.5 at $\lambda$ for recombinants AA:AB
how does $y = bp$ alter posterior?
same math as E step of classical EM algorithm

how to estimate the genotypic means $\mu_q$?
(shrink sample means toward prior to reduce selection bias)
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
- no easy way to write down posterior
  - proportional to priors for genotypes $q$ and loci $\lambda$
  - constant averaged over all genotypes $q$ and loci $\lambda$

what is the genetic architecture $H$?

- 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
3. QTL Model Search using MCMC

- trick: Markov chains are samples from a stable distribution
  - Markov chain: the future depends on the past only through the present
  - artificially construct Markov chain with distribution we want
  - alter one thing (unknown) at a time—make the chain easy to construct
- sample QTL model components from full conditionals
  - sample locus \( \lambda \) given \( q,H \) (using Metropolis-Hastings step)
  - sample genotypes \( q \) given \( \lambda,\mu,y,H \) (using Gibbs sampler)
  - sample effects \( \mu \) given \( q,y,H \) (using Gibbs sampler)
  - sample QTL model \( H \) given \( \lambda,\mu,y,q \) (using Gibbs or M-H)

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

initial guess: no QTL

\[k^{th}\] sample draw \((\lambda,q,\mu,H)_{k+1}\) given \((\lambda,q,\mu,H)_k\)

stop after \(N = 100,000\) steps

---

Gibbs sampler idea

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

\[
\begin{pmatrix}
\mu_1 \\
\mu_2
\end{pmatrix}
\sim
N
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
1 & \rho \\
\rho & 1
\end{pmatrix}
\right)
\]

\[
\mu_1 \sim N(\rho \mu_2, 1 - \rho^2)
\]

\[
\mu_2 \sim N(\rho \mu_1, 1 - \rho^2)
\]
Gibbs sampler samples: $\rho = 0.6$

$N = 50$ samples

$N = 200$ samples

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 samples

$N = 200$ samples

$narrow g$  
$wide g$

$N = 1000$ samples

$narrow g$  
$wide g$

MCMC realization

added twist: occasionally propose from whole domain
Gibbs sampler with loci indicators

- partition genome into intervals
  - at most one QTL per interval
  - interval = 1 cM in length
  - assume QTL in middle of interval
- use loci to indicate presence/absence of QTL in each interval
  - $\gamma = 1$ if QTL in interval
  - $\gamma = 0$ if no QTL
- Gibbs sampler on loci indicators
  - see work of Nengjun Yi (and earlier work of Ina Hoeschele)
  - Yi, Yandell et al. (2005); R/qtlbim (2006)

$$Y = \beta_0 + \gamma_1 \beta_1(q_1) + \gamma_2 \beta_2(q_1) + e$$

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 $H$
- Satagopan, Yandell (1996); Gaffney (2001); R/bim (2002)

$nqtl = 1 : Y = \beta_0 + \beta_1(q_1) + e$

$nqtl = 2 : Y = \beta_0 + \beta_1(q_1) + \beta_2(q_2) + e$
Bayesian shrinkage estimation

- soft loci indicators
  - strength of evidence for $\lambda_j$ depends on variance of $\beta_j$
  - similar to $\gamma > 0$ on grey scale
- include all possible loci in model
  - pseudo-markers at 1cM intervals
- Wang et al. (2005 *Genetics*)
  - Shizhong Xu group at U CA Riverside

\[
Y = \beta_0 + \beta_1 (q_1) + \beta_2 (q_1) + \ldots + e \\
\beta_j (q_j) \sim N(0, \sigma_j^2), \sigma_j^2 \sim \text{inverse - chisquare}
\]

epistatic interactions

- model space issues
  - 2-QTL interactions only?
  - Fisher-Cockerham partition vs. tree-structured?
  - general interactions among multiple QTL
- model search issues
  - epistasis between significant QTL
    - check all possible pairs when QTL included?
    - allow higher order epistasis?
  - epistasis with non-significant QTL
    - whole genome paired with each significant QTL?
    - pairs of non-significant QTL?
4. Model Assessment

- balance model fit against model complexity

<table>
<thead>
<tr>
<th>aspect</th>
<th>smaller model</th>
<th>bigger model</th>
</tr>
</thead>
<tbody>
<tr>
<td>model fit</td>
<td>miss key features</td>
<td>fits better</td>
</tr>
<tr>
<td>prediction</td>
<td>may be biased</td>
<td>no bias</td>
</tr>
<tr>
<td>interpretation</td>
<td>easier</td>
<td>more complicated</td>
</tr>
<tr>
<td>parameters</td>
<td>low variance</td>
<td>high variance</td>
</tr>
</tbody>
</table>

- information criteria: penalize $L$ by model size $|H|$ 
  - compare $IC = -2 \log L(H \mid y) + \text{penalty}(H)$

- Bayes factors: balance posterior by prior choice 
  - compare $\text{pr}(\text{data} \mid \text{model} H)$

Bayes factors and BIC

- Bayesian interpretation 
  - $\text{pr}(\text{data} \mid \text{model}) = \text{pr}(\text{model} \mid \text{data}) / \text{pr}(\text{model})$ 
  - $\text{pr}(\text{data} \mid \text{model}) = \text{model posterior} / \text{model prior}$ 
  - marginal model averaged over all parameters

- Bayes Information Criteria 
  - $BIC = 2\log(\text{likelihood}) + \text{d.f.} \ast \log(\text{n.ind})$
  - downweight data likelihood by complexity 
  - complexity penalty matches Bayesian idea
Bayes factors and BIC

• Bayes factor ($BF$) for model comparison
  – ratio of $\text{pr(data} \mid \text{model})$ for 2 models
  – often reported as $2\log(BF)$
  – weak/moderate/strong evidence: 3/10/30

• $BIC$ comparison
  – difference of two $BIC$ values
  – same as $LR$ statistic with penalty when
    • comparing two nested models
    • simple hypotheses (e.g. 1 vs 2 QTL)

• $BF = BIC$ comparison for nested models

marginal LOD or LPD

• compare two architectures at locus
  – with ($H_2$) or without ($H_1$) QTL at $\lambda_2$
    • preserve model hierarchy (e.g. drop any epistasis with QTL at $\lambda_2$)
  – with ($H_2$) or without ($H_1$) epistasis at $\lambda_2$
  – allow for QTL at all other loci $\lambda_1$ in architecture $H_1$

• use marginal LPD or other diagnostic
  – posterior, Bayes factor, heritability

\[
\text{LOD}(\lambda_1, \lambda_2 \mid H_2) - \text{LOD}(\lambda_1 \mid H_1)
\]

\[
\text{LPD}(\lambda_1, \lambda_2 \mid H_2) - \text{LPD}(\lambda_1 \mid H_1)
\]
5. simulations and data analyses

- revisit 1 QTL simulation
  - refining position by marginal scan
    - single QTL vs. marginal on multi-QTL
    - $2\log(BF)$
  - substitution effect: 1-QTL vs. multi-QTL
- R/qtl hyper dataset (Sugiyama et al. 2001)
  - higher LPD with multi-QTL
  - detecting epistasis and linked QTL
substitution effect: 1 QTL vs. multi-QTL
single QTL effect vs. marginal effect from QTL at $\lambda$

scan of marginal Bayes factor

2logBF of phenotype for main

cellmean of phenotype for A+H
hyper data: scanone

LPD of bp for main+epistasis+sum

2log(BF) scan with 50% HPD region

2logBF of bp for main+epistasis+sum

cellmean of bp for A+H
hyper: number of QTL posterior, prior, Bayes factors

pattern of QTL on chromosomes
relative importance of epistasis

2-D plot of $2\log BF$: chr 6 & 15
1-D Slices of 2-D scans: chr 6 & 15

Cockerham epistatic effects

% of samples with each epistatic pair
1-D Slices of 2-D scans: chr 4 & 15

diagnostic summaries
QTL for Bayesian Interval Mapping
R/qtlbim: our software

• publication
  – Yi, Yandell, Churchill, Allison, Eisen, Pomp (2005 *Genetics*)
  – Yi et al. Yandell (in review)
  – CRAN release Fall 2006

• properties
  – new MCMC algorithms
    • Gibbs with loci indicators; no reversible jump
  – epistasis, fixed & random covariates, GxE
  – extensive graphics

R/qtlbim: our software

• R/qtlbim is cross-compatible with R/qtl
• Bayesian module within WinQTLCart
  – WinQTLCart output can be processed using R/bim

• Software history
  – initially designed (Satagopan Yandell 1996)
  – major revision and extension (Gaffney 2001)
  – R/bim to CRAN (Wu, Gaffney, Jin, Yandell 2003)
  – R/qtlbim to CRAN (Yi, Yandell, Mehta, Banerjee, Shriner, Neely, von Smith 2006)
other Bayesian software for QTLs

- R/bim*: Bayesian Interval Mapping
  - Satagopan Yandell (1996; Gaffney 2001) CRAN
  - no epistasis; reversible jump MCMC algorithm
  - version available within WinQTLCart (statgen.ncsu.edu/qtlcart)
- R/qt1*
  - Broman et al. (2003 Bioinformatics) CRAN
  - multiple imputation algorithm for 1, 2 QTL scans & limited mult-QTL fits
- Bayesian QTL / Multimapper
  - Sillanpää Arjas (1998 Genetics) www.rni.helsinki.fi/~mjs
  - no epistasis; introduced posterior intensity for QTLs
  - (no released code)
  - Stephens & Fisch (1998 Biometrics)
  - no epistasis
- R/bqtl
  - C Berry (1998 TR) CRAN
  - no epistasis, Haley Knott approximation

* Jackson Labs (Hao Wu, Randy von Smith) provided crucial technical support

many thanks

Jackson Labs
Gary Churchill
Hao Wu
Randy von Smith

U AL Birmingham
David Allison
Nengjun Yi
Tapan Mehta
Samprit Banerjee

USDA Hatch, NIH/NIDDK (Attie), NIH/R01 (Yi)

October 2006
Animal Genetics © Brian S. Yandell 47