Module 4: Bayesian Methods
Lecture 9 A: Default prior selection

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Jeffreys prior Unit information priors Empirical Bayes priors

Outline

Independent binary sequence

Suppose researcher A has data of the following type:

$$M_A: \quad y_1, \ldots, y_n \sim \text{ i.i.d. binary}(\theta), \ \theta \in [0, 1].$$

A asks you to do a Bayesian analysis, but either

- doesn't have any prior information about θ, or
- wants you to obtain "objective" Bayesian inference for θ.

You need to come up with some prior $\pi_A(\theta)$ to use for this analysis.

Independent binary sequence

Suppose researcher B has data of the following type:

$$M_B: \quad y_1, \ldots, y_n \sim \text{ i.i.d. binary}(\tfrac{e^\gamma}{1+e^\gamma}), \ \gamma \in (-\infty, \infty).$$

B asks you to do a Bayesian analysis, but either

- doesn't have any prior information about γ, or
- wants you to obtain "objective" Bayesian inference for γ.

You need to come up with some prior $\pi_B(\gamma)$ to use for this analysis.

Prior generating procedures

Suppose we have a procedure for generating priors from models:

$$\text{Procedure}(M) \to \pi$$

Applying the procedure to model M_A should generate a prior for θ:

$$\text{Procedure}(M_A) \to \pi_A(\theta)$$

Applying the procedure to model M_B should generate a prior for γ:

$$\text{Procedure}(M_B) \to \pi_B(\gamma)$$

What should the relationship between π_A and π_B be?

Induced priors

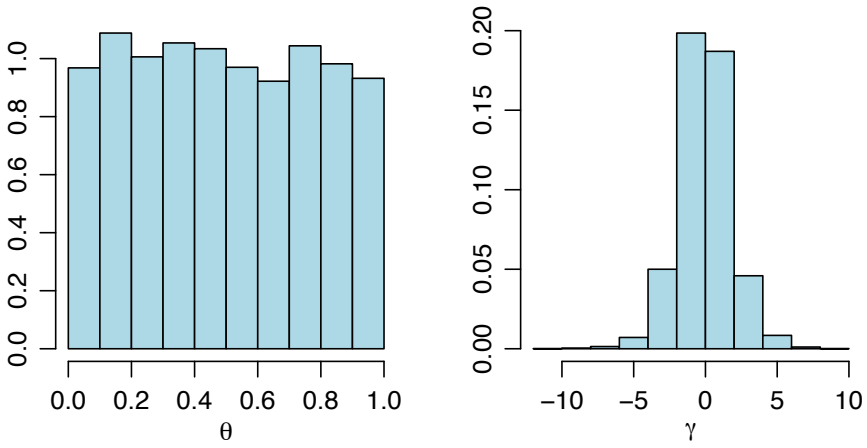
Note that a prior $\pi_A(\theta)$ over θ induces a prior $\pi_A(\gamma)$ over $\gamma = \log \frac{\theta}{1-\theta}$.

This *induced prior* can be obtained via

- calculus;
- simulation.

Induced priors

```
theta<-rbeta(5000,1,1)
gamma<-log(theta/(1-theta))
```



Internally consistent procedures

This fact creates a small conundrum:

We could generate a prior for γ via the induced prior on θ:

$$\text{Procedure}(M_A) \rightarrow \pi_A(\theta) \rightarrow \pi_A(\gamma)$$

Alternatively, a prior for γ could be obtained directly from M_B:

$$\text{Procedure}(M_B) \rightarrow \pi_B(\gamma)$$

Both $\pi_A(\gamma)$ and $\pi_B(\gamma)$ are obtained from the Procedure.

Which one should we use?

Jeffreys' principle

Jeffreys (1949) says that any default Procedure should be *internally consistent* in the sense that the two priors on γ should be the same.

More generally, his principle states if M_B is a reparameterization of M_A, then

$$\pi_A(\gamma) = \pi_B(\gamma).$$

Of course, all of this logic applies to the model in terms of θ:

$$\begin{aligned} \text{Procedure}(M_A) &\rightarrow \pi_A(\theta) \\ \text{Procedure}(M_B) &\rightarrow \pi_B(\gamma) \rightarrow \pi_B(\theta) \\ \pi_A(\theta) &= \pi_B(\theta) \end{aligned}$$

Jeffreys' prior

It turns out that Jeffreys' principle leads to a unique Procedure:

$$\pi_J(\theta) = \sqrt{|\mathrm{E}[(\tfrac{d}{d\theta} \log p(y|\theta))^2]|}$$

Example: Binomial/binary model

$$\begin{aligned} y_1, \ldots, y_n &\sim \text{i.i.d. binary}(\theta) \\ \pi_J(\theta) &\propto \theta^{-1/2}(1-\theta)^{-1/2} \end{aligned}$$

We recognize this prior as a beta(1/2,1/2) distribution:

$$\theta \sim \text{beta}(1/2, 1/2)$$

Default Bayesian inference is then based on the following posterior:

$$\theta|y_1, \ldots, y_n \sim \text{beta}(1/2 + \sum y_i, 1/2 + \sum(1-y_i)).$$

Jeffreys' prior

Example: Poisson model

$$y_1, \ldots, y_n \sim \text{i.i.d. Poisson}(\theta)$$
$$\pi_J(\theta) \propto 1/\sqrt{\theta}$$

Recall our conjugate prior for θ in this case was a gamma(a, b) density:

$$\theta(\theta|a, b) \propto \theta^{a-1} e^{-\theta/b}$$

For the Poisson model and gamma prior,

$$\theta \sim \text{gamma}(a, b) \rightarrow \theta|y_1, \ldots, y_n \sim \text{gamma}(a + \sum y_i, b + n)$$

What about under the Jeffreys prior?

$\pi_J(\theta)$ "looks like" a gamma distribution with $(a, b) == (1/2, 0)$. It follows that

$$\theta \sim \pi_J \rightarrow \theta|y_1, \ldots, y_n \sim \text{gamma}(1/2 + \sum y_i, n).$$

(Note: π_J is not an actual gamma density - it is not a probability density at all!)

Jeffreys' prior

Example: Normal model

$$y_1, \ldots, y_n \sim \text{i.i.d. Normal}(\mu, \sigma^2)$$
$$\pi_J(\mu, \sigma^2) = 1/\sigma^2$$

(this is a particular version of Jeffreys' prior for multiparameter problems)

It is very interesting to note that the resulting posterior for μ is

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}$$

This means that a 95% objective Bayesian confidence interval for μ is

$$\mu \in \bar{y} \pm t_{.975, n-1} s/\sqrt{n}$$

This is exactly the same as the usual t-confidence interval for a normal mean.

Notes on Jeffreys' prior

1. Jeffreys' principle leads to Jeffreys' prior.
2. Jeffreys' prior isn't always a "proper" prior distribution.
3. Improper priors can lead to proper posteriors.
   - These often lead to Bayesian interpretations of "frequentist" procedures.

Data-based priors

Recall from the binary/beta analysis:

$$\begin{aligned} \theta &\sim \text{beta}(a, b) \\ y_1, \ldots, y_n &\sim \text{binary}(\theta) \\ \theta | y_1, \ldots, y_n &\sim \text{beta}(a + \sum y_i, b + \sum (1 - y_i)) \end{aligned}$$

Under this posterior,

$$\begin{aligned} \text{E}[\theta | y_1, \ldots, y_n] &= \frac{a + \sum y_i}{a + b + n} \\ &= \left(\frac{a + b}{a + b + n} \right) \frac{a}{a + b} + \left(\frac{n}{a + b + n} \right) \bar{y} \end{aligned}$$

- $\frac{a}{a+b} \approx$ guess at what θ is
- $a + b \approx$ confidence in guess.

Data-based priors

We may be reluctant to guess at what θ is. Wouldn't $\bar{y}$ be better than a guess?

Idea: Set $\frac{a}{a+b} = \bar{y}$.

Problem: This is cheating! Using $\bar{y}$ for your prior misrepresents the amount of information you have.

Solution: Cheat as little as possible:

- Set $\frac{a}{a+b} = \bar{y}$.
- Set $a + b = 1$.
- This implies $a = \bar{y}$, $b = 1 - \bar{y}$.

The amount of cheating has the information content of only one observation.

Unit information principle

If you don't have prior information about θ, then

1. Obtain an MLE/OLS estimator $\hat{\theta}$ of θ;
2. Make the prior $\pi(\theta)$
   - weakly centered around $\hat{\theta}$,
   - have the information equivalent of one observation.

Again, such a prior leads to double-use of the information in your sample.

However, the amount of "cheating" is small, and decreases with n.

Poisson example:

$$y_1, \ldots, y_n \sim \text{i.i.d. Poisson}(\theta)$$

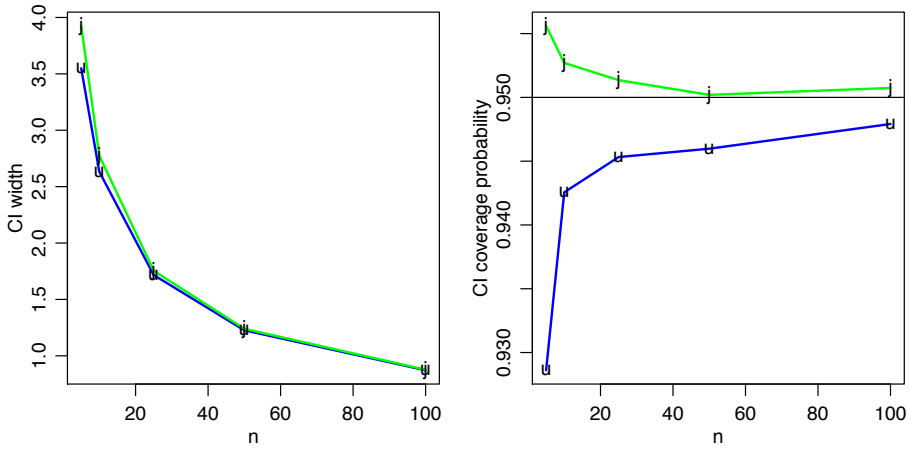
Under the gamma(a, b) prior,

$$\begin{aligned} \mathrm{E}[\theta | y_1, \ldots, y_n] &= \frac{a + \sum y_i}{b + n} \\ &= \left(\frac{b}{b+n}\right)\frac{a}{b} + \left(\frac{n}{b+n}\right)\bar{y} \end{aligned}$$

Unit information prior:

$$a/b = \bar{y}, b = 1 \Rightarrow (a, b) = (\bar{y}, 1)$$

Comparison to Jeffreys' prior



Notes on UI priors

1. UI priors weakly concentrate around a data-based estimator.
2. Inference under UI priors is anti-conservative, but this bias decreases with n.
3. Can be used in multiparameter settings, and is related to BIC.

Normal means problem

$$y_j = \theta_j + \epsilon_j, \quad \epsilon_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, 1)$$

Task: Estimate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

An odd problem:

- What does estimation of θ_j have to do with estimation of θ_k?
- There is only one observation y_j per parameter θ_j - how well can we do?

Where the problem comes from:

- Comparison of two groups A and B on p variables (e.g. expression levels)
- For each variable j, construct a two-sample t-statistic

$$y_j = \frac{\bar{x}_{A,j} - \bar{x}_{B,j}}{s_j/\sqrt{n}}$$

- For each j, y_j is approximately normal with
  - mean $\theta_j = \sqrt{n}(\mu_{A,j} - \mu_{B,j})/\sigma_j$
  - variance 1.

Normal means problem

$$y_j = \theta_j + \epsilon_j, \quad \epsilon_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, 1)$$

One obvious estimator of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is $\mathbf{y} = (y_1, \ldots, y_p)$.

- $\mathbf{y}$ is the MLE;
- $\mathbf{y}$ is unbiased and the UMVUE.

However, it turns out that $\mathbf{y}$ is not so great in terms of risk:

$$R(\mathbf{y}, \boldsymbol{\theta}) = \mathrm{E}[\sum_{j=1}^{p} (y_j - \theta_j)^2]$$

When $p > 2$ we can find an estimator that beats $\mathbf{y}$ for every value of $\boldsymbol{\theta}$, and is much better when p is large. This estimator has been referred to as an *empirical Bayes estimator*.

Bayesian normal means problem

$$y_j = \theta_j + \epsilon_j, \quad \epsilon_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, 1)$$

Consider the following prior on $\boldsymbol{\theta}$:

$$\theta_1, \ldots, \epsilon_p \sim \text{i.i.d. normal}(0, \tau^2)$$

Under this prior,

$$\hat{\theta}_j = \mathrm{E}[\theta_j | y_1, \ldots, y_n] = \frac{\tau^2}{\tau^2 + 1} y_j$$

This is a type of "shrinkage" prior:

- It "shrinks" the estimates towards zero, away from y_j;
- It is particularly good if many of the true θ_j's are very small or zero.

Empirical Bayes

$$\hat{\theta}_j = \frac{\tau^2}{\tau^2+1} y_j$$

We might know we want to shrink towards zero.

We might not know the appropriate amount of shrinkage.

Solution: Estimate τ^2 from the data!

$$\left.\begin{array}{ll} y_j & = \theta_j + \epsilon_j \\ \epsilon_j & \sim N(0,1) \\ \theta_j & \sim N(0,\tau^2) \end{array}\right\} \quad \Rightarrow \quad y_j \sim N(0, \tau^2+1)$$

We should have

$$\sum y_j^2 \approx p(\tau^2+1)$$
$$\sum y_j^2/p - 1 \approx \tau^2$$

Idea: Use $\hat{\tau}^2 = \sum y_j^2/p - 1$ for the shrinkage estimator.

Modification Use $\hat{\tau}^2 = \sum y_j^2/(p-2) - 1$ for the shrinkage estimator.

James-Stein estimation

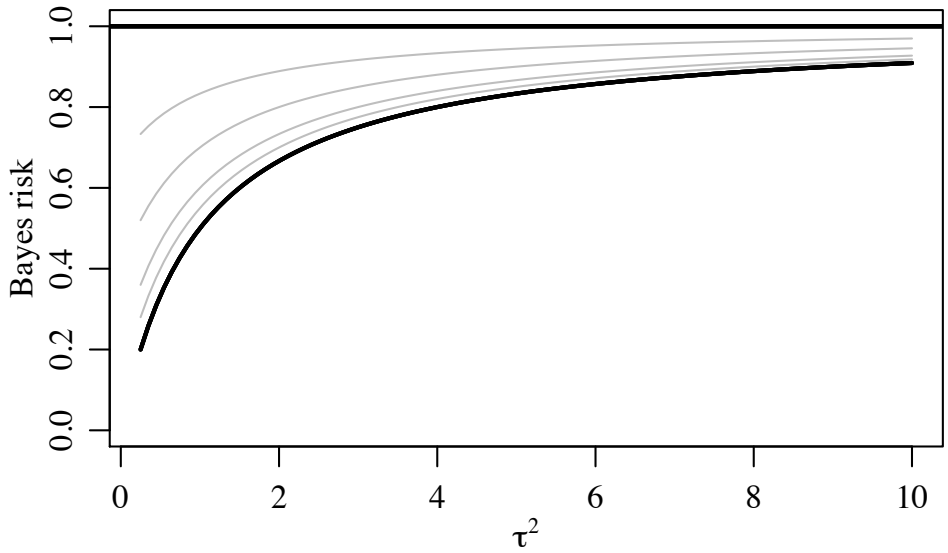
$$\hat{\theta}_j = \frac{\hat{\tau}^2}{\hat{\tau}^2+1} y_j \qquad \hat{\tau}^2 = \sum y_j^2/(p-2) - 1$$

It has been shown theoretically that from a non-Bayesian perspective, this estimator beats $\mathbf{y}$ in terms of risk for all $\boldsymbol{\theta}$.

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) < R(\mathbf{y}, \boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta}$$

Also, from a Bayesian perspective, this estimator is almost as good as the optimal Bayes estimator, under a known τ^2.

Comparison of risks



The Bayes risk of the JSE is between that of $\mathbf{X}$ and the Bayes estimator.

Bayes risk functions are plotted for $p \in \{3, 5, 10, 20\}$.

Empirical Bayes in general

Model: $p(y|\theta), \theta \in \Theta$

Prior class: $\pi(\theta|\psi), \psi \in \Psi$

What value of ψ to choose?

Empirical Bayes:

1. Obtain the "marginal likelihood" $p(y|\psi) = \int p(y|\theta)\pi(\theta|\psi)d\theta$;
2. Find an estimator $\hat{\psi}$ based on $p(y|\psi)$;
3. Use the prior $\pi(\theta|\hat{\psi})$.

Notes on empirical Bayes

1. Empirical Bayes procedures are obtained by "estimating" hyperparameters from the data.
2. Often these procedures behave well from both Bayes and frequentist procedures.
3. They work best when the number of parameters is large and hyperparameters are distinguishable.

Module 4: Bayesian Methods
Lecture 9 B: QTL interval mapping

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

QTLs

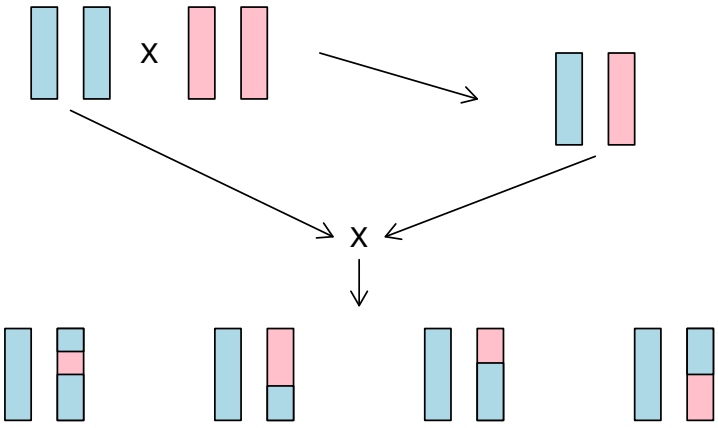
Genetic variation $\Rightarrow$ quantitative phenotypic variation

QTLs have been associated with many health-related phenotypes:

- cancer
- obesity
- heritable disease

QTL interval mapping: A statistical approach to the identification of QTLs from marker and phenotype data.

F1 Backcross



At any given locus, an animal could be AA or AB.

Two-component mixture model

Suppose there is a single QTL affecting a continuous trait. Let

- x be the location of the QTL;
- $g(x)$ be the genotype at x
  - $g(x) = 0$ if AA at x
  - $g(x) = 1$ if AB at x
- y be a continuous quantitative trait.

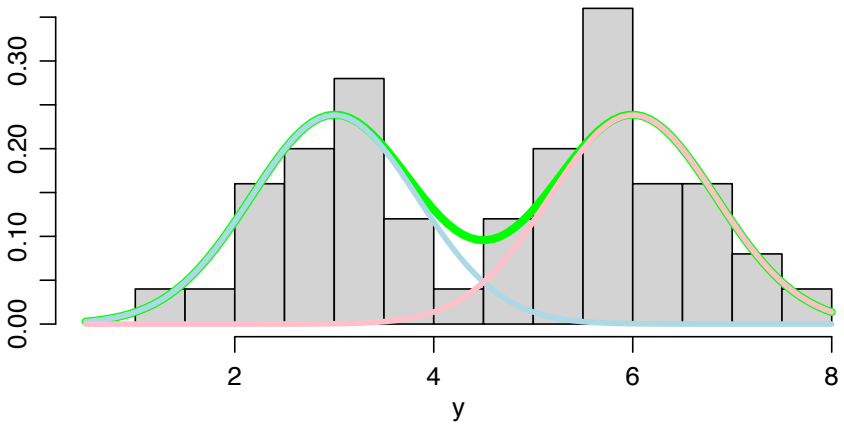
Two-component mixture model:

$$y \sim \begin{cases} \text{normal}(\mu_{AA}, \sigma^2) & \text{if } g(x) = 0 \\ \text{normal}(\mu_{AB}, \sigma^2) & \text{if } g(x) = 1 \end{cases}$$

About half of the animals are $g(x) = 0$ and half are $g(x) = 1$, but we don't know which are which.

Two-component mixture model

Data from 50 animals:



Marker data

If the location of x of the QTL were known, we could genotype it:

- $\mathbf{y}_0 = \{y_i : g_i(x) = 0\}$
- $\mathbf{y}_1 = \{y_i : g_i(x) = 1\}$
- evaluate effect size with a two sample t-test.

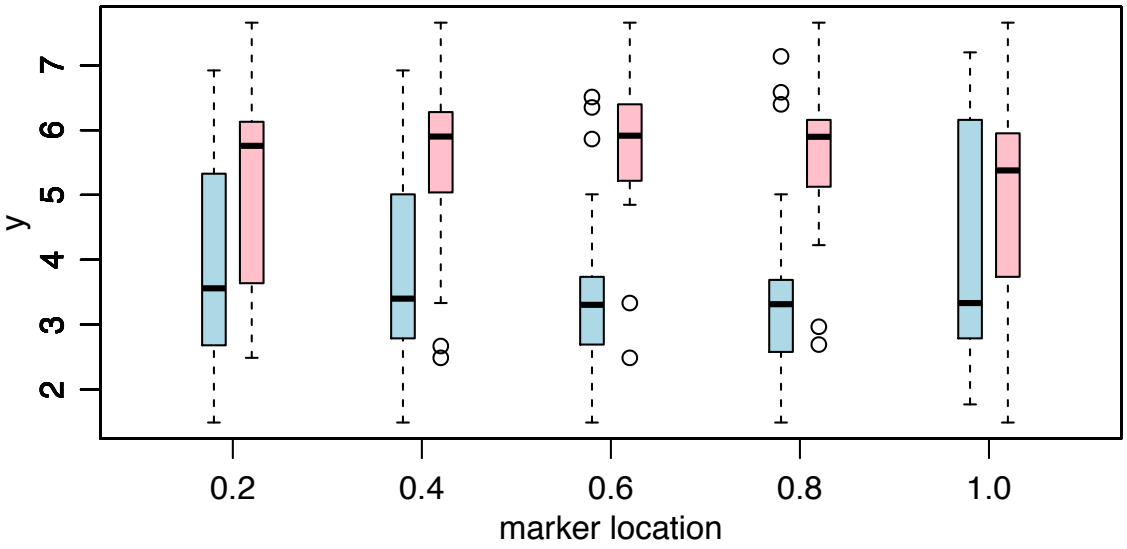
Instead of $g(x)$ we have genotype information at a set of markers:

Genotype information at evenly spaced markers $m_1, \ldots, m_K$

$$g_i(m_k) = \begin{cases} 0 & \text{if animal } i \text{ is homozygous at } m_k \\ 1 & \text{if animal } i \text{ is heterozygous at } m_k \end{cases}$$

Comparisons at marker locations

$n = 50$ animals at $K = 6$ equally spaced marker locations:



Comparisons across the genome

Procedure: Move along each chromosome, making comparisons of heterozygotes to homozygotes at each possible QTL location x.

Problem: Genotypes at non-marker locations x are not known.

However, they are known probabilistically: Let

- r = recombination rate between left and right flanking markers;
- r_l = recombination rate between left flanking marker m_l and x;
- r_r = recombination rate between right flanking marker m_r and x.

$$\Pr(g(x) = 1 | g(m_l) = 1, g(m_r) = 1) = \frac{(1 - r_l) \times (1 - r_r)}{1 - r}$$

$$\Pr(g(x) = 1 | g(m_l) = 0, g(m_r) = 1) = \frac{r_l \times (1 - r_r)}{r}$$

etc.

Knowns and unknowns quantities

Unknown quantities in the system include

- QTL location x
- genotypes $G(x) = \{g_1(x), \ldots, g_n(x)\}$
- parameters of the QTL distributions: $\theta = \{\mu_{AA}, \mu_{AB}, \sigma^2\}$

Known quantities include

- quantitative trait data $\mathbf{y} = y_1, \ldots, y_n$
- marker data $\mathbf{M} = \{g_i(m_k), i = 1, \ldots, n, k = 1, \ldots K\}$

Bayesian analysis: Obtain Pr(unknowns|knowns)

$$\Pr(x, G(x), \theta | \mathbf{y}, \mathbf{M})$$

Gibbs sampler

We can approximate $\Pr(x, G(x), \theta|\mathbf{y}, \mathbf{M})$ with a Gibbs sampler:

1. simulate $x \sim p(x|\theta, \mathbf{y}, \mathbf{M})$
2. simulate $G(x) \sim p(G(x)|\theta, x, \mathbf{y}, \mathbf{M})$
3. simulate $\theta \sim p(\theta|x, G(x), \mathbf{y}, \mathbf{M})$

For example, based on marker data alone,

$$\begin{aligned}\Pr(g_i(x) = 1|\mathbf{M}) &= \frac{\Pr(g_i(x) = 1|\mathbf{M})}{\Pr(g_i(x) = 1|\mathbf{M}) + \Pr(g_i(x) = 0|\mathbf{M})} \\ &= \frac{p_{i1}}{p_{i1} + p_{i0}}.\end{aligned}$$

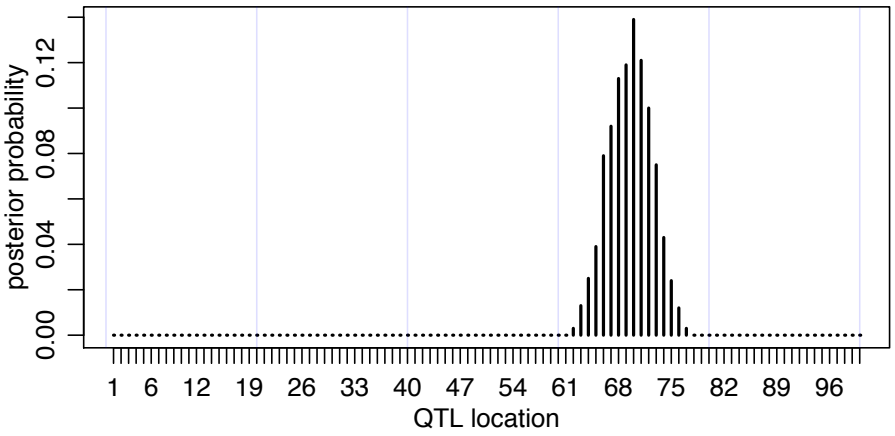
Given phenotype data,

$$\begin{aligned}\Pr(g_i(x) = 1|x, \theta, \mathbf{y}, \mathbf{M}) &= \frac{p_{i1} \times p(y_i|g_i(x) = 1, \theta)}{p_{i1} \times p(y_i|g_i(x) = 1, \theta) + p_{i0} \times p(y_i|g_i(x) = 0, \theta)} \\ &= \frac{p_{i1} \times \text{dnorm}(y_i, \mu_{AB}, \sigma)}{p_{i1} \times \text{dnorm}(y_i, \mu_{AB}, \sigma) + p_{i0} \times \text{dnorm}(y_i, \mu_{AA}, \sigma)}.\end{aligned}$$

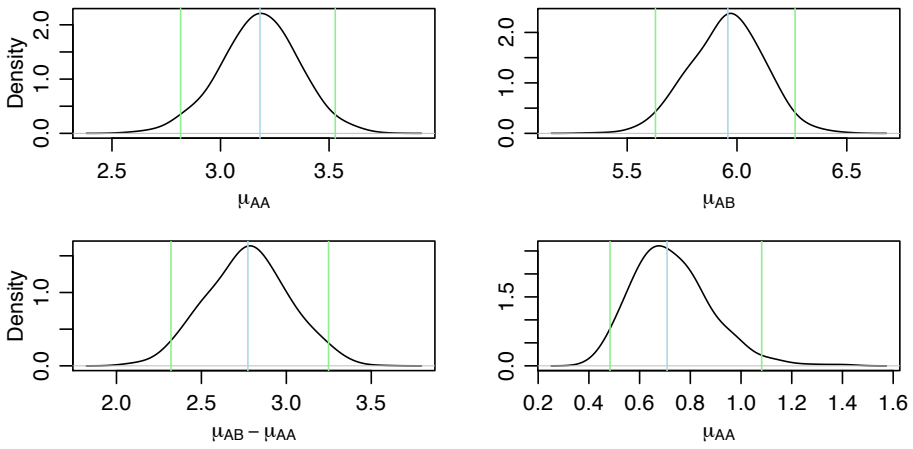
R-code for Gibbs sampler

```
for(s in 1:25000)
{

  ## update x
  lpy.x<-NULL;for(x in 1:100){lpy.x<-c(lpy.x,lpy.theta(y,G,x,mu,s2))}
  x<-sample(1:100,1,prob=exp(lpy.x-max(lpy.x)))

  ## update gx
  pg1.x<-prhet.sG(x,G,mpos)
  py.g1<-dnorm(y,mu[2],sqrt(s2))
  py.g0<-dnorm(y,mu[1],sqrt(s2))
  pg1.yx<- py.g1*pg1.x/( py.g1*pg1.x + py.g0*(1-pg1.x))
  gx<-rbinom(n,1,pg1.yx)

  ## update s2
  s2<-1/rgamma( 1,(nu0+n)/2,(nu0*s20+sum( (y-mu[gx+1])^2) )/2   )

  ## update mu
  mu<-rnorm(2,(mu0*k0+tapply(y,gx,sum))/(k0+table(gx)),
                sqrt(s2/(k0+ table(gx))))
}
```

QTL location



Parameter estimates



Some references

- "Review of statistical methods for QTL mapping in experimental crosses" (Broman, 2001).
- "QTLBIM" - QTL Bayesian interval mapping: R-package.