QTL Mapping II: Model Selection

Karl W Broman
Department of Biostatistics
Johns Hopkins University

kbroman@jhsph.edu

www.biostat.jhsph.edu/~kbroman

Outline

• The model selection problem:
  – Class of models
  – Compare models
  – Search model space
  – Assess the performance of a procedure

• A simulation study
Hypothesis testing?

- In the past, QTL mapping has been regarded as a task of hypothesis testing.

  Is this a QTL?

  Much of the focus has been on adjusting for test multiplicity.

- It is better to view the problem as one of model selection.

  What set of QTLs are well-supported?

  Is there evidence for QTL-QTL interactions?

  Model = a defined set of QTLs and QTL-QTL interactions (and possibly covariates and QTL-covariate interactions).

Perfect data situation

To ease discussion, we’ll focus on a simple special case:

- Complete marker genotype data
- Markers are only putative QTLs
- Normally distributed residuals

Example model (in a backcross):

\[ y_i = \mu + \Delta_1 q_{i1} + \Delta_2 q_{i2} + \Delta_3 q_{i3} + \epsilon_i \]

\( \epsilon_i \) are iid \( N(0, \sigma^2) \)

\( q_j \) are 0/1 variables (QTL genotypes)

\( \mu, \Delta \)'s are parameters, estimated by least squares

Fitted values: \( \hat{y}_i = \hat{\mu} + \hat{\Delta}_1 q_{i1} + \hat{\Delta}_2 q_{i2} + \hat{\Delta}_3 q_{i3} \)

\( \text{RSS} = \sum_i (y_i - \hat{y}_i)^2 \) indicates model fit.
Model selection

1. Class of models
2. Compare models
3. Search model space
4. Assess performance of a procedure

Note:

2 and 4 are much the same.

There might be a 0th item: Method for model fitting. (e.g., imputation, EM, etc.)

Class of models

- Additive models
  \[ y = \mu + \sum_j \Delta_j q_j + \epsilon \]

- Also pairwise interactions
  Preserve hierarchy: if include interaction, also include both main effects.

- Also higher-order interactions
  Again, preserve hierarchy.

- Regression trees
Example regression tree

### Intercrosses: class of models

- Always bring in both degrees of freedom with a QTL or
  Try to distinguish additivity/dominance/recessiveness?

\[
\begin{array}{ccc}
A & H & B \\
A & 100 & 80 & 80 \\
H & 20 & 20 & 20 \\
B & 20 & 20 & 20 \\
\end{array}
\]

- Always bring in all four d.f. with a QTL:QTL interaction or
  Try to distinguish ways to subdivide the \(3 \times 3\) table?

\[
\begin{array}{ccc}
A & H & B \\
A \\
H \\
B \\
\end{array}
\]
Compare models

- Imagine you could fit all possible models; which would you like best?

- This issue is like LOD thresholds, but more complicated.

- For models with the same number of parameters (QTLs and interactions), we prefer that with the “best fit” (smallest RSS or largest likelihood).

- If you fit more parameters, you’ll get a “better fit.”
  - How much better, before including additional terms?
  - I like a form of penalized likelihood.

- Note: Bayesians must also confront this issue (through the prior on models).

The additive QTL case

n backcross mice; M markers

\[ x_{ij} = \text{genotype (1/0) of mouse } i \text{ at marker } j \]

\[ y_i = \text{phenotype (trait value) of mouse } i \]

\[ y_i = \mu + \sum_{j=1}^{M} \Delta_j x_{ij} + \epsilon_i \]

Which \( \Delta_j \neq 0 \)?

\[ \text{BIC}_\delta = \log \text{RSS} + \text{no. markers} \times \left( \delta \times \frac{\log n}{n} \right) \]
Why BIC$_{\delta}$?

- For a fixed no. markers, letting $n \rightarrow \infty$, BIC$_{\delta}$ is consistent.
- There exists a prior (on models + coefficients) for which BIC$_{\delta}$ is the $-\log$ posterior.
- BIC$_{\delta}$ is essentially equivalent to use of a threshold on the conditional LOD score.
- It performs well.

BIC$_{\delta} \leftrightarrow$ conditional LOD

Conditional LOD score:

$$LOD(x^*_k | x^*_1, \ldots, x^*_{k-1}) = \frac{n}{2} \log_{10} \left\{ \frac{\text{RSS}(x^*_1, \ldots, x^*_{k-1})}{\text{RSS}(x^*_1, \ldots, x^*_k)} \right\}$$

Minimizing BIC$_{\delta}$ is approximately equivalent to choosing the largest $k$ such that

$$LOD(x^*_k | x^*_1, \ldots, x^*_{k-1}) \geq \frac{\delta}{2} \log_{10} n$$
Choice of $\delta$

Smaller $\delta$: include more loci; higher false positive rate

Larger $\delta$: include fewer loci; lower false positive rate

Let $L = 95\%$ genome-wide LOD threshold

(compare single-QTL models to the null model)

Choose $\delta = 2 \frac{L}{\log_{10} n}$

With this choice of $\delta$, in the absence of QTLs, we'll include at least one extraneous locus, 5% of the time.

Search model space

- Consider the case of additive QTL models, with 100 putative QTLs.
- There are $2^{100} \approx 10^{30}$ possible models, far more than can be inspected individually.
- Need a way to search through this space, to find the good ones.
- This is really a matter of “grunt work.” (More is better; the tradeoff is with computational time.)
Methods of model search

- **Forward selection**
  - Find the best single-QTL model: \( q_1^* \).
  - Find the best two-QTL model that includes \( q_1^* \) : \( (q_1^*, q_2^*) \).
  - Find the best three-QTL model that includes \( q_1^*, q_2^* \) : \( (q_1^*, q_2^*, q_3^*) \).
  - Etc.

- **Backward elimination**

- **Forward selection followed by backward elimination**

- **Stepwise selection**

- **Randomized algorithms** (e.g., MCMC, genetic algorithms, etc.)

Assess performance

- Selection of a model includes two types of errors:
  - Missing important terms (QTLs or interactions)
  - Include extraneous terms

- Unlike in hypothesis testing, we can make **both errors** at the same time!

- **Identify as many correct terms as possible, while controlling the rate of inclusion of extraneous terms.**

- You **can’t know** the performance of your procedure with your data—you need to know the truth.

- You **can know**:
  - How a particular procedure performs in simulated cases
  - How a procedure performs in simulated data close to what you’ve inferred
Science isn’t really model selection

I’ve said here: pick a good model

Really:

• Want to guide future experiments
• Want some understanding of the uncertainty in different aspects of the chosen model

A simulation study

• Backcross with n=250
• No crossover interference
• 9 chr, each 100 cM
• Markers at 10 cM spacing; complete genotype data
• 7 QTLs
  – One pair in coupling
  – One pair in repulsion
  – Three unlinked QTLs
• Heritability = 50%
• 2000 simulation replicates
Methods

- ANOVA at marker loci
- Composite interval mapping (CIM)
- Forward selection with permutation tests
- Forward selection with BIC$_\delta$
- Backward elimination with BIC$_\delta$
- FS followed by BE with BIC$_\delta$
- MCMC with BIC$_\delta$

→ A selected marker is deemed correct if it is within 10 cM of a QTL (i.e., correct or adjacent)

A simplified version of CIM

Select a set of markers, $S$
  (e.g., by FS to a fixed number)

For each marker, $x$, in the genome:
  (a) If $x \not\in S$, calculate $\text{LOD}(x \mid S)$
  (b) If $x \in S$, calculate $\text{LOD}(x \mid S \setminus \{x\})$

Compare to a genome-wide threshold.
  (Take into account the choice of $S$.)
QTLs linked in coupling

Ave no. chosen

ANOVA 3 5 7 9 11

CIM fs, perm fs be fs/be mcmc

QTLs linked in repulsion

Ave no. chosen

ANOVA 3 5 7 9 11

CIM fs, perm fs be fs/be mcmc

ANOVA

CIM

fs, perm

fs

be

fs/be

mcmc

BIC
Summary

- QTL mapping is a **model selection problem** (rather than hypothesis testing).

- **Model selection** =
  - Select a class of models
  - Select a criterion for comparing models
  - Select a method of searching model space
  - Figure out how your procedure performs

- **Key issue**: the comparison of models.

- Large-scale computer simulations are necessary for assessing the performance of procedures.
References

  A paper, soon to appear, containing the simulation study described above.

  Another paper on the model selection aspects of QTL mapping.

  A good book on model selection in regression; this new edition has just been printed.