

# Shrinkage models for variable selection in high-dimensional problems

Alex Lewin<sup>1</sup>, Leonardo Bottolo<sup>2</sup> and Sylvia Richardson<sup>1</sup>

<sup>1</sup>Department of Epidemiology, <sup>2</sup>Institute of Mathematical Sciences; Imperial College  
a.m.lewin@imperial.ac.uk

## Motivation

Spike and slab models have been proposed as a solution to the problem of variable selection (Ishwaran and Rao 2005).

These models are attractive because they have a simple structure and can be fit in a fully Bayesian manner using Gibbs sampling.

Jia and Xu recently proposed an extension of such a model for analysing large eQTL (expression Quantitative Trait Loci) data sets with multiple outcomes.

In this work we investigate this model using another eQTL data set and compare results found by a Bayesian variable selection model (Bottolo and Richardson 2008, see poster).

## Summary

- A spike and slab model is tested on a data set of gene expression and SNPs, to search for associations between transcripts and markers.
- The Gibbs sampler for the spike and slab model encounters serious mixing problems: it is not able to estimate posterior probabilities of association, though regression coefficients appear stable.
- Results (eg. regression coefficients) are highly sensitive to parameter choice.
- Spike and slab model does not produce enough sparsity for variable selection.
- By comparison, a variable selection model using evolutionary MCMC is able to find a reasonable number of associations in this data set.

## Conclusion

The spike and slab model appears unsuitable for variable selection problems involving large data sets.

## Spike and slab model

$y_{it}$  gene expression (individual  $i$ , transcript  $t$ )  
 $X_{ij}$  SNP (individual  $i$ , marker  $j$ )

$y_{it} \sim N(\sum_j X_{ij}\beta_{ij}, \sigma^2)$   
 $\beta_{ij} \sim (1-z_{ij})N(0, 0.001) + z_{ij}N(0, \lambda_j^2)$

$z_{ij} \sim \text{Bern}(\rho_j)$ ,  $\rho_j \sim \text{Beta}(1, 1)$

EITHER  $\lambda_j^2 \sim \text{Inv Gam}(3, \tau)$ ,  $\tau$  fixed  
OR  $\lambda_j^2 = \lambda^2$  fixed

$\sigma^2 \sim \text{InvGam}(0.01, 0.01)$

This model is essentially that proposed in Jia and Xu (2007):

- Spike and slab (mixture) prior on regression coefficients.
- Shares information across transcripts for each marker using a hierarchical model.

### Parameters of interest:

- $z$  (posterior mean is posterior probability of association between transcript  $t$  and marker  $j$ )
- $\beta$  (regression coefficient linking transcript  $t$  and marker  $j$ )
- $\rho$  (probability of marker  $j$  being associated with any transcripts), i.e. probability of marker  $j$  being a master-regulator

We have implemented a Monte Carlo Markov Chain (MCMC) Gibbs sampler in Matlab. 10,000 iterations take 8 hours on dual processor 2.4 GHz running Windows.

## References:

Hübner et al. 2005, Nature Genetics 37: 243-253.

Ishwaran and Rao 2005, JASA 33: 730-773; Jia and Xu 2007, Genetics 176: 611-623.

## What can be estimated in this model?

### 4 models:

- 1)  $\lambda^2$  fixed = 10 (non-hierarchical)
- 2)  $\lambda^2$  fixed = 3
- 3)  $\tau$  fixed = 1 (hierarchical)
- 4)  $\tau$  fixed = 5

Model (1) run 500,000 iterations, other models 15,000 to 25,000 iterations.

Run on data below (1000 transcripts, 770 markers).

### MCMC performance:

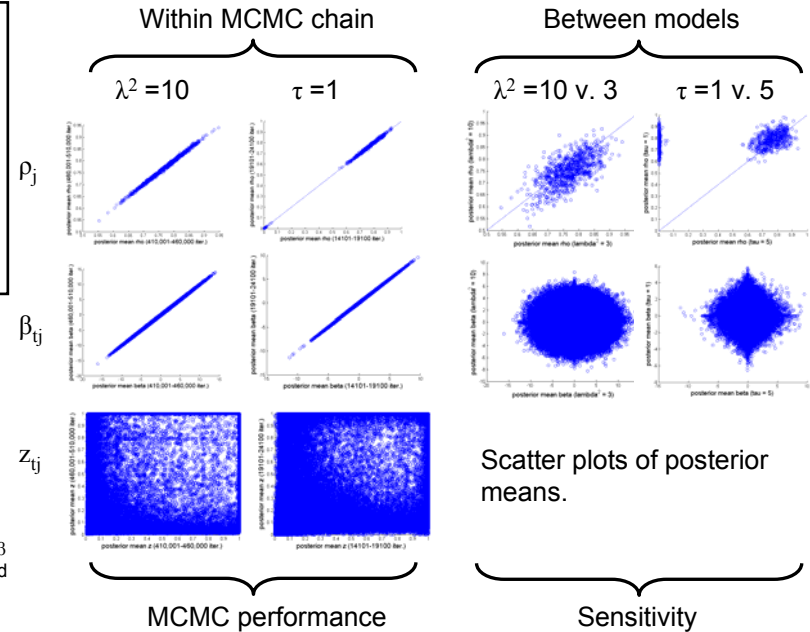
Compare posterior means of  $\rho$ ,  $\beta$  and  $z$  from two different parts of the same MCMC chain.

Marker probabilities  $\rho$  and regression coefficients  $\beta$  can be estimated well, but posterior probabilities of association  $z$  cannot.

### Sensitivity:

Compare posterior means from different models.

Cannot even estimate regression coefficients  $\beta$  with any reliability (eg. between models (1) and (2)) the correlation between regression coefficients is 0.001.



## Results on inbred rats data

### eQTL (expression Quantitative Trait Loci)

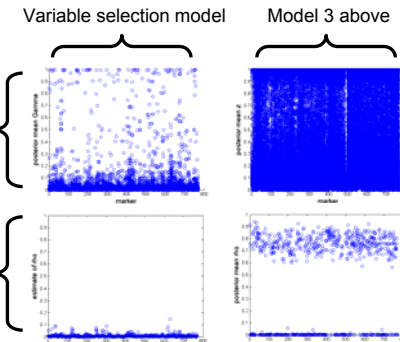
Fat tissue from 29 inbred rats (Hübner et al. 2005).

For each rat: gene expression of 1000 transcripts (those most varying between rats), and genotype at 770 SNPs (single nucleotide polymorphisms).

Compare spike and slab with variable selection model (Bottolo and Richardson 2008: see poster).

$P(\text{association of marker } j \text{ with transcript } t \mid \text{data}) \text{ v. marker } j$   
(probs. for all 1000 transcripts plotted against each marker)

$P(\text{association of marker } j \text{ with any transcript } \mid \text{data}) \text{ v. marker } j$   
(probability of being a master regulator)



Variable selection model succeeds in estimating  $z$  (results not shown).

Variable selection model induces sparsity on  $z$  and finds very few markers to be master-regulators (as expected biologically).

Spike and slab model does not induce sparsity on  $z$ .

Spike and slab model claims 62% of markers to be master-regulators ( $\rho \sim 0.8$ ), with only 38% having  $\rho < 0.1$ .