Selection of causal SNPs in Twin Studies data

Subho Majumdar

November 17, 2016

Objective

- Most GWAS using family-based designs focus on single-SNP analysis to do association analysis;
- The reason is the difficulty to do association tests for individual SNPs in a multiple-SNP linear mixed model;
- Our objective is to take a variable selection approach while remaining within the mixed model sructure;
- We shall utilize a frugal model selection method to identify SNPs with possible association with the quantitative trait in question.

The model

- m pedigrees, i^{th} pedigree has n_i indivuduals.
- $y_{ij} =$ measured phenotype in j-th individual of i-th pedigree.

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i\beta_g + \mathbf{C}_i\beta_c + \epsilon_i$$

for i-th pedigree. The matrices $\mathbf{G}_i$ and $\mathbf{C}_i$ contain genotype scores for a number of SNPs and environmental covariate values respectively, for all members of the pedigree. Also

$$\epsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{V}_i); \quad \mathbf{V}_i = \Phi\sigma_a^2 + \mathbf{I}_{n_i}\sigma_e^2$$

where Φ is the known kinship matrix.

Simulation setup

- 500 pedigrees, each of size 4: consisting of parents and HZ twins;
- $\alpha = 0$, no environmental covariates;
- 1000 independent SNPs, with probabilities of dominant alleles chosen from Unif$(0.1, 0.3)$;
- $\sigma_a^2 = 3, \sigma_e^2 = 4$;
- First 10 SNPs are causal:
  Case 1- $\beta_{g,1}, ..., \beta_{g,10} \sim$ Unif$(0.1, 0.2)$ iid;
  Case 2- $\beta_{g,1}, ..., \beta_{g,10} \sim$ Unif$(0.5, 1)$ iid.
- Full setup replicated 100 times.

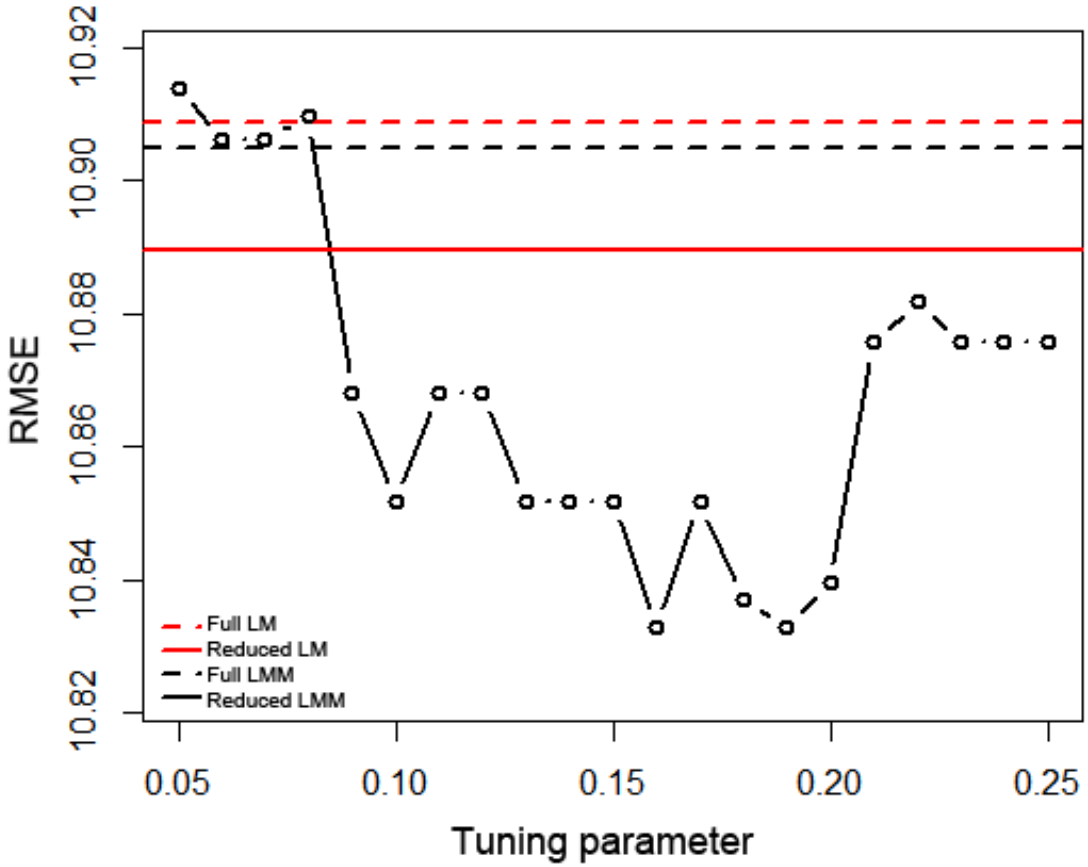
Results

Case	True positive	True negative
1	0.84 (0.13)	0.21 (0.06)
2	0.997 (0.02)	0.20 (0.05)

Runtime:

- 1 minute for 100 SNPs, 20 minutes for 1000 SNPs;
- Faster than backward deletion on linear model.

Prediction performance



(On smaller data)