Selection of causal SNPs in Twin Studies data from multi-SNP mixed models

Subho Majumdar

April 20, 2017

The model

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

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i \boldsymbol{\beta} + \mathbf{C}_i \boldsymbol{\beta}_c + \boldsymbol{\epsilon}_i \tag{1}$$

$$\mathbf{V}_i = \sigma_a^2 \boldsymbol{\Phi}_i + \sigma_c^2 \mathbf{1}\mathbf{1}^T + \sigma_e^2 \mathbf{I}_{n_i} \tag{2}$$

where $\boldsymbol{\Phi}_i$ is the known relationship matrix = twice the kinship matrix, and $\sigma_a^2, \sigma_c^2, \sigma_e^2$ are the variances corresponding to polygenic effext outside the group of SNPs modeled, shared environment and random error, respectively.

Objective

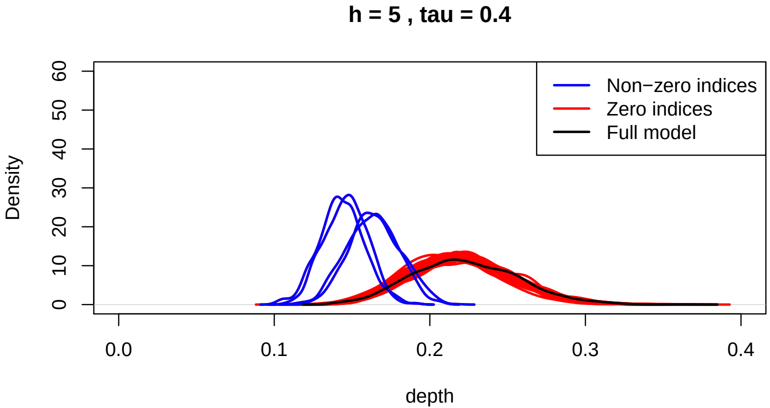
Want to detect the non-zero entries of $\boldsymbol{\beta}_g$ in the above model. State-of-the-art is to perform single-SNP analysis (e.g. using RFGLS) and then correct for multiple correlation. This loses power. We want to use a new bootstrap-based method of *e*-**values** to improve that.

The *e*-values method

1. Estimate the full model coefficient, say $\hat{\boldsymbol{\beta}}_g$ (by `regress` etc.)
2. Obtain its bootstrap distribution: $[\hat{\boldsymbol{\beta}}]$;
3. Replace the *j*-th coefficient with 0, name it $\hat{\boldsymbol{\beta}}_{-j}$. Do the same for its bootstrap distribution, say $[\hat{\boldsymbol{\beta}}_{-j}]$. Repeat for all j;
4. *e*-value of *j*-th covariate = tail probability of the *q*-th quantile of $E([\hat{\boldsymbol{\beta}}_{-j}])$ with respect to $E([\hat{\boldsymbol{\beta}}])$, where $E(.)$ is an *evaluation function*;
5. Select *j*-th covariate if *e*-value is less than *qt*-th quantile of $E([\hat{\boldsymbol{\beta}}])$.

Simulation setup

- 250 pedigrees, each of size 4: consisting of parents and MZ twins;
- $\alpha = 0$, no environmental covariates;
- 50 SNPs in correlated blocks of 6,4,6,4 and 30: MAF of SNPs in the blocks 0.2, 0.4, 0.4, 0.25 and 0.25;
- $\sigma_a^2 = 4, \sigma_c^2 = 1, \sigma_e^2 = 1$;
- First SNP of first 4 blocks are causal: each having heritability *h*%
- Full setup replicated 100 times.


h = 5 , tau = 0.4
Density
0
10
20
30
40
50
60
0.0
0.1
0.2
0.3
0.4
depth
Non−zero indices
Zero indices
Full model

Simulation results

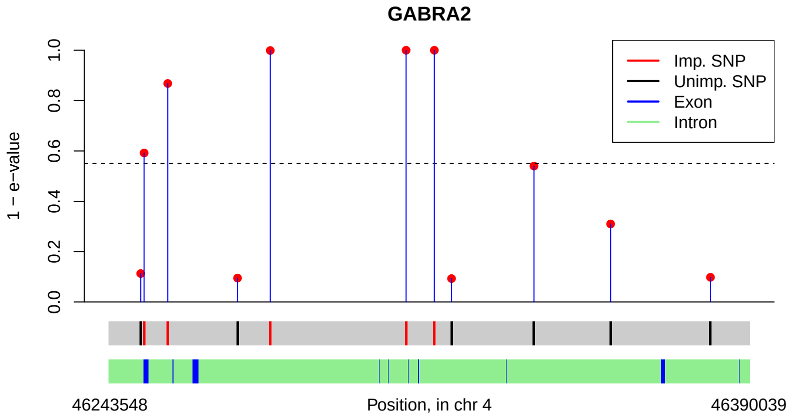
Case	**BIC**	**RFGLS**	*e*-**values**
$h = 10$	0.82/0.96	0.62/0.97	0.94/0.93
$h = 5$	0.36/0.98	0.36/0.97	0.73/0.90
$h = 2$	0.08/0.99	0.16/0.97	0.30/0.93
$h = 1$	0.02/1.00	0.10/0.97	0.14/0.96
$h = 0$	0.00/1.00	0.01/0.98	0.02/0.98

Table: True Poisitive/True negative proportions over 100 replications for 3 methods ($q = 0.5, t = 0.8$)

Analyzing the MCTFR data

- Analyze data on families with MZ twins: 682 families;
- Look at gene specific models: GABRA2, ADH1B, ADH1C, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT. Group together ADH genes. Also do SLC6A4+DRD2.

GABRA2



Detects rs1808851 and rs279856, which are at perfect LD with the well-known rs279858. This is missed by a previous analysis (Irons 2012).