Simultaneous Selection of Multiple Important Single Nucleotide Polymorphisms in Familial Genome Wide Association Studies Data

Subhabrata Majumdar, AT&T Data Science and AI Research
Saonli Basu, Matt McGue and Snigdhansu Chatterjee
University of Minnesota Twin Cities

May 21, 2021

Outline

- Motivation
- Statistical model
- The *e*-values framework
- Simulation study
- The Minnesota Twin Studies data example

Simultaneous Selection of Multiple Important Single Nucleotide Polymorphisms in **Familial** Genome Wide Association Studies Data

Motivation

Genome-Wide Association Studies (GWAS) based on families are used in behavioral genetics to control for environmental variation, thus requiring smaller sample size to detect Single Nucleotide Polymorphisms (SNP) responsible behind traits like alcoholism and drug addiction, and also to quantify gene-environment interaction.

Two challenges:

1. SNPs highly correlated, weak signals of individual SNPs;
2. Need to use mixed models to account for within-family dependence.

Objective

We propose a computationally efficient approach for SNP detection in families while utilizing information on multiple SNPs simultaneously.

There are two state-of-the-art approaches:

- Perform single-SNP analysis and then correct for multiple testing. This loses power.
- Group-based association test (SKAT etc.) that test for whether a group of SNPs is associated with the phenotype. They do not generally prioritize within the group and are unable to detect individual SNPs associated with the trait.

We use our recently proposed framework of *e*-values to improve upon that.

Statistical model

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i\beta_g + \mathbf{C}_i\beta_c + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{V}_i); \quad \mathbf{V}_i = \sigma_a^2\boldsymbol{\Phi}_i + \sigma_c^2\mathbf{1}\mathbf{1}^T + \sigma_e^2\mathbf{I}_{n_i}$$

- Total m families, with the i-th pedigree containing n_i individuals;
- $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$ are the quantitative trait values for individuals in i-th pedigree, $\mathbf{G}_i \in \mathbb{R}^{n_i \times p_s}$ containing their genotypes for a bunch of SNPs, $\mathbf{C}_i \in \mathbb{R}^{n_i \times p}$ contain the data on individual-specific covariates;
- Three variance components correspond to polygenic effect due to other SNPs, shared environment effect and individual-specific effects. This is called the **ACE model**.

The relationship matrix Φ_i

$$\mathbf{\Phi}_{MZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1 \\ 1/2 & 1/2 & 1 & 1 \end{bmatrix},$$

$$\mathbf{\Phi}_{DZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{bmatrix},$$

$$\mathbf{\Phi}_{Adopted} = \mathbf{I}_4$$

$\mathbf{\Phi}_i$ depends on the type of the i-th family:

MZ = family with identical or monozygous twins,
DZ = family with identical or dizygous twins.

The *e*-values framework

- *e*-values are a fast and general method for best subset variable selection (Majumdar and Chatterjee, 2017).
- The *e*-values are able to compare sampling distributions of coefficient vectors coresponding to two statistical models.
- They are based on point-to-distribution distance measures, which we call *evaluation functions* $E(\mathbf{x}, [\mathbf{X}])$, $[\mathbf{X}]$ denoting distribution of a random variable $\mathbf{X}$.

ACE models and *e*-values

1. Estimate the full model coefficient, say $\hat{\beta}_g \equiv \hat{\beta}$ (by R package `regress`)
2. Obtain its bootstrap distribution: $[\hat{\beta}]$;
3. Replace the j-th coefficient with 0, name it $\hat{\beta}_{-j}$. Do the same for its bootstrap distribution, say $[\hat{\beta}_{-j}]$. Repeat for all j;
4. e-value of j-th SNP = tail probability of the q-th quantile of $[E(\hat{\beta}_{-j}, [\hat{\beta}])]$ with respect to $[E(\hat{\beta}, [\hat{\beta}])]$;
5. Select j-th SNP if its e-value is less than tq, for some $0 < t < 1$.

Advantages

- Need to train one single model– saves a lot of computation time for mixed models;
- SNPs are selected taking the effects of *all* other SNPs into account;
- We use a fast generalized bootstrap (Chatterjee and Bose, 2005) for the calculation steps, which is based on monte carlo random sampling.

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 (a measure of magnitude of non-zero effect) $h/6\%$;
- Full setup replicated 1000 times.

- Methods compared:
  mBIC2 - Variant of BIC that control false discovery rate at 0.05;
  RFGLS - Fast method of fitting single-SNP ACE models. Do Benjamini-Hochberg correction on *p*-values to control FDR at 0.05.

Simulation results

Method		$h = 10$	$h = 7$	$h = 5$	$h = 3$	$h = 2$	$h = 1$	$h = 0$
mBIC2		0.79/0.99	0.59/0.99	0.41/0.99	0.2/0.99	0.11/0.99	0.05/0.99	-/0.99
RFGLS+BH		0.95/0.92	0.82/0.95	0.62/0.97	0.29/0.98	0.14/0.99	0.04/1	-/1
E_2	$t = 0.8$	0.97/0.98	0.9/0.97	0.79/0.96	0.54/0.96	0.34/0.97	0.15/0.98	-/0.99
	$t = 0.74$	0.96/0.98	0.88/0.97	0.75/0.97	0.48/0.97	0.29/0.98	0.12/0.98	-/0.99
	$t = 0.68$	**0.95/0.99**	0.87/0.98	0.72/0.98	0.45/0.98	0.26/0.98	0.1/0.99	-/0.99
	$t = 0.62$	**0.95/0.99**	0.84/0.98	**0.68/0.98**	**0.4/0.99**	**0.22/0.99**	0.09/0.99	-/0.99
	$t = 0.56$	0.94/0.99	**0.82/0.99**	**0.65/0.99**	**0.36/0.99**	**0.19/0.99**	**0.07/1**	**-/1**
	$t = 0.5$	0.92/0.99	0.79/0.99	0.6/0.99	**0.31/0.99**	**0.16/1**	**0.05/1**	**-/1**

Table: Average True Positive (TP)/ True Negative (TN) rates for mBIC2, RFGLS+BH and the e-values method with E_2 as evaluation maps and different values of t over 1000 replications

$$E_2(\mathbf{x}, [\mathbf{X}]) = \exp\left[-\left\|\frac{\mathbf{x} - \mathbb{E}\mathbf{X}}{\sqrt{\mathrm{diag}(\mathbb{V}\mathbf{X})}}\right\|\right]$$

Analyzing the Minnesota Twin Studies data

- Analyze data on families with MZ and DZ twins: 682 families;
- Response variable: amount of alcohol consumption;
- Look at models specific to well-studied genes for alcoholism: GABRA2, ADH1B, ADH1C, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT;
- Group together ADH genes as individual genes have very small number of SNPs.

Number of detected SNPs

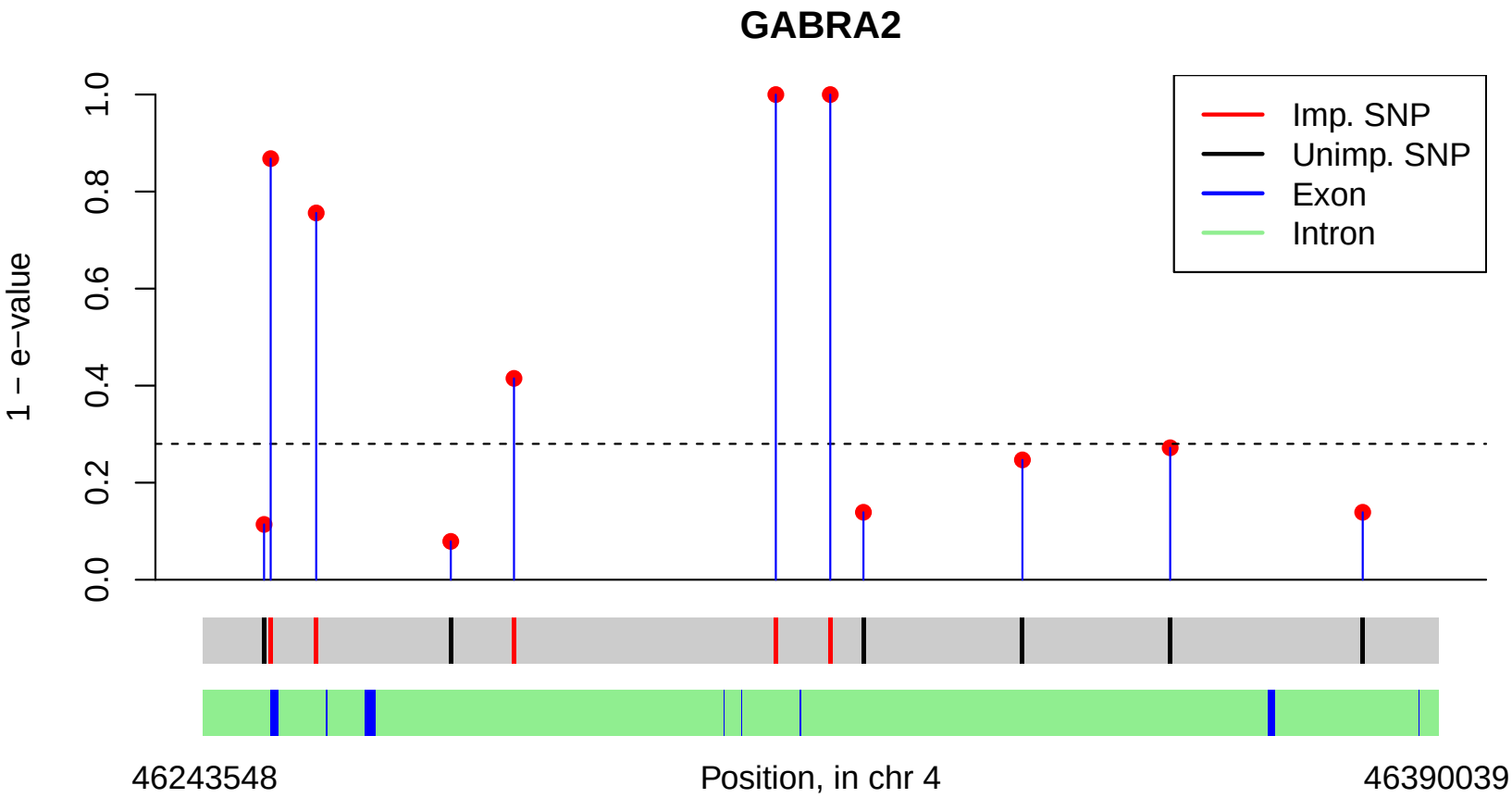
Gene	Total no. of SNPs	No. of SNPs detected by		
		e-value	RFGLS+BH	mBIC2
GABRA2	11	5	0	0
ADH	44	3	1	0
OPRM1	47	25	1	0
CYP2E1	9	5	0	0
ALDH2	6	5	0	1
COMT	15	14	0	0
SLC6A3	18	4	0	0
SLC6A4	5	0	0	0
DRD2	17	0	0	1

Table: Table of analyzed genes and number of detected SNPs in them by the three methods

Summary of detected SNPs

Gene	Detected SNPs with known associations	Reference for associated SNP
GABRA2	rs1808851, rs279856: close to rs279858	Cui et al. (2012)
ADH genes	rs17027523: 20kb upstream of rs1229984	Multiple studies (https://www.snpedia.com/index.php/Rs1229984)
OPRM1	rs12662873: 1 kb upstream of rs1799971	Multiple studies (https://www.snpedia.com/index.php/Rs1799971)
CYP2E1	rs9419624: 600b downstream of rs4646976; rs9419702: 10kb upstream of rs4838767	Lind et al. (2012)
ALDH2	rs16941437: 10kb upstream of rs671	Multiple studies (https://www.snpedia.com/index.php/Rs671)
COMT	rs4680, rs165774	Voisey et al. (2011)
SLC6A3	rs464049	Huang et al. (2017)

Table: Table of detected SNPs with known references



Detects rs1808851 and rs279856, which have very high correlation with the well-known rs279858. This was missed by a previous analysis (Irons, 2012).

References

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.*, 33:414–436.

Cui, W. Y., Seneviratne, C., Gu, J., and Li, M. D. (2012). Genetics of GABAergic signaling in nicotine and alcohol dependence. *Hum. Genet.*, 131:843–855.

Huang, C.-C., Kuo, S.-C., Weh, Y.-W., et al. (2017). The SLC6A3 gene possibly affects susceptibility to late-onset alcohol dependence but not specific personality traits in a Han Chinese population. *PLoS ONE*, 12(2):e0171170.

Irons, D. E. (2012). *Characterizing specific genetic and environmental influences on alcohol use*. PhD thesis, University of Minnesota.

Lind, P. A., Macgregor, S., Heath, A. C., and Madden, P. A. F. (2012). Association between *in vivo* alcohol metabolism and genetic variation in pathways that metabolize the carbon skeleton of ethanol and NADH reoxidation in the Alcohol Challenge Twin Study. *Alcohol Clin. Exp. Res.*, 36:2074–2085.

Majumdar, S. and Chatterjee, S. (2017). Fast and General Model Selection using Data Depth and Resampling. https://arxiv.org/abs/1706.02429.

Voisey, J., Swagell, C. D., Hughes, I. P., et al. (2011). A novel SNP in COMT is associated with alcohol dependence but not opiate or nicotine dependence: a case control study. *Behav. Brain Funct.*, 7(51).

Future work

- Extending to group SNP detection;
- Extending to analyzing SNPs from multiple genes- detect SNPs inside a gene, then detect significant genes among a group of genes;
- High-dimensional situations.

THANK YOU!

Acknowledgements: NSF grant IIS-1029711, University of Minnesota Interdisciplinary Doctoral Fellowship