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

Subhabrata Majumdar

Abstract: Genome Wide Association Studies (GWAS) on samples obtained from families instead of unrelated individuals are an established way of assessing gene-environment interactions and effects of Single Nucleotide Polymorphisms (SNP) on the development of behavioral disorders. Due to multiple reasons like weak signals of causal SNPs or multiple correlated SNPs, traditional methods of SNP detection based on single-marker analysis suffer from loss of power in such studies. Although statistical model selection methods are able to remedy this, owing to the dependent structure of the familial data they are computationally demanding. Here we propose a fast variable selection algorithm to address this problem. Working on a linear mixed model with three variance components and fixed effects for multiple SNPs, we obtain a quantity called the e-value for each SNP by only training the model with all covariates, and select SNPs having e-values below a threshold. To compute the e-values, we utilize a fast and scalable bootstrap procedure that relies on Monte-Carlo sampling to obtain bootstrapped copies of estimated fixed effect vectors. Numerical studies reveal our method to be more effective in detecting causal SNPs than either single-marker analysis on mixed models or model selection methods that ignore the familial dependency structure. We also use the e-values to perform gene-level analysis in a familial GWAS dataset and detect several SNPs that have potential effect on alcohol consumption in individuals.

Keywords: Model selection, bootstrap, data depth, family data, twin studies, ACE model, alcoholism

1 Model

We have data $\mathcal{Z} = \{\mathcal{Z}^1, \dots, \mathcal{Z}^K\}; \mathcal{Z}^k = (\mathbf{Y}^k, \mathbf{X}^k)$ where $\mathbf{Y}^k \in \mathbb{R}^{n \times q}, \mathbf{X}^k \in \mathbb{R}^{n \times p}$ for $1 \le k \le K$.

$$\mathbf{X}^k = (\mathbf{X}_1^k, \dots, \mathbf{X}_p^k)^T \sim \mathcal{N}(0, \Sigma_x^k) \tag{1.1}$$

$$\mathbf{Y}^k = \mathbf{X}^k \mathbf{B}^k + \mathbf{E}^k; \quad \mathbf{E}^k = (\mathbf{E}_1^k, \dots, \mathbf{E}_p^k)^T \sim \mathcal{N}(0, \Sigma_y^k) \tag{1.2}$$

$$\Omega_x^k = (\Sigma_x^k)^{-1}; \quad \Omega_y^k = (\Sigma_y^k)^{-1} \tag{1.3}$$

Want to estimate $\{(\Omega_x^k, \Omega_y^k, \mathbf{B}^k); 1 \le k \le K\}$ in presence of known grouping structures $\mathcal{G}_x, \mathcal{G}_y, \mathcal{H}$ respectively. Estimation of $\{\Omega_x^k\}$ done using JSEM. For the other part, we use the following two-step procedure:

1. Run neighborhood selection on y-network incorporating effects of x-data:

$$\min \left\{ \sum_{i=1}^p \frac{1}{n_k} \left[\sum_{k=1}^K \|\mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_i^k - \mathbf{X}^k \mathbf{B}_i^k\|^2 + 2 \sum_{j \ne i} \sum_{g \in \mathcal{G}_y^{ij}} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\| \right] + 2 \sum_{h \in \mathcal{H}} \eta^h \|\mathbf{B}^{[h]}\| \right\} \tag{1.4}$$

$$= \min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}, \Theta) + P(\Theta) + Q(\mathcal{B}) \right\} \tag{1.5}$$

   where $\Theta = \{\Theta_i\}, \mathcal{B} = \{\mathbf{B}^k\}, \mathcal{Y} = \{\mathbf{Y}^k\}, \mathcal{X} = \{\mathbf{X}^k\}, \mathcal{E} = \{\mathbf{E}^k\}$.

   This estimates $\mathcal{B}$ (maybe refit and/or within-group threshold?) .

2. Step I part 2 and step II of JSEM (see 15-656 pg 6) follows to estimate $\{\Omega_y^k\}$.

The objective function is bi-convex, so we are going to do the following in step 1-

- Start with initial estimates of $\mathcal{B}$ and Θ, say $\mathcal{B}^{(0)}, \Theta^{(0)}$.

- Iterate:

$$\Theta^{(k+1)} = \arg\min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}^{(k)}, \Theta^{(k)}) + P(\Theta^{(k)}) \right\} \tag{1.6}$$

$$\mathcal{B}^{(k+1)} = \arg\min \left\{ f(\mathcal{Y}, \mathcal{X}, \mathcal{B}^{(k)}, \Theta^{(k+1)}) + Q(\mathcal{B}^{(k)}) \right\} \tag{1.7}$$

- Continue till convergence.

2 Conditions

Conditions A1, A2, A3 from JSEM paper.

3 Results

Define

$$\hat{\boldsymbol{\beta}}^k = \tag{3.1}$$

$$\hat{\Theta}^i = \arg\min_{\Theta_i} \left\{ \frac{1}{n_k} \sum_{k=1}^K \|\mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_i^k - \mathbf{X}^k \hat{\mathbf{B}}_i^k\|^2 + 2 \sum_{j \ne i} \sum_{g \in \mathcal{G}_y^{ij}} \lambda_{ij}^g \|\boldsymbol{\theta}_{ij}^{[g]}\| \right\} \tag{3.2}$$

Theorem 3.1. *Assume fixed $\mathcal{X}, \mathcal{E}$ and deterministic $\hat{\mathcal{B}} = \{\mathbf{B}^k\}$. Also*
(T1) $\|\hat{\mathbf{B}}_i^k - \mathbf{B}_i^k\| \leq v_\beta$;
(T2) $\|\mathbf{X}^k(\hat{\mathbf{B}}_i^k - \mathbf{B}_i^k)\| \leq c(v_\beta)$ *for some non-negative function $c(.)$;*
Group uniform IC.
Then
(I) Estimation consistency
(II) Direction consistency

Proof of Theorem 3.1. Part I. Follows proof of thm 1 in 15-656. The proof has 3 parts: consistency of neighborhood regression, selection of edge sets, and finally the refitting step.

For any $g \in \mathcal{G}^{ij}, k \in g$, and $j \neq i$, let

$$\hat{\boldsymbol{\epsilon}}_i^k = \mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_{0,i}^k - \mathbf{X}^k \hat{\mathbf{B}}_i^k; \quad \hat{\zeta}_{ij}^k = \frac{(\hat{\boldsymbol{\epsilon}}_i^k)^T \mathbf{Y}_j^k}{n}; \quad \hat{\boldsymbol{\zeta}}_{ij}^{[g]} = (\hat{\zeta}_{ij}^k)_{k \in g}$$

Consider the random event $\mathcal{A} = \bigcap_{i,j \neq i,g} \mathcal{A}_{ij}^g$ with $\mathcal{A}_{ij}^g = \{2\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\| \leq \lambda_{ij}^g\}$.

Proposition 3.2. *Given that λ_{ij}^g are chosen as*

$$\lambda_{ij}^g \geq \max_{k \in g} \frac{2}{\sqrt{n\omega_{ii}^k}} \left(\sqrt{|g|} + \frac{\pi}{\sqrt{2}} \sqrt{q \log G_0} + \sqrt{c(v_\beta)} \right)$$

we shall have $\mathbb{P}(\mathcal{A}) \geq 1 - 2pG_0^{1-q}$ for some $q > 1$.

Proof of Proposition 3.2. We follow the proof of Lemma E.2 in 15-656, with $\mathbf{Y}_j^k, \hat{\boldsymbol{\epsilon}}_i^k, \hat{\zeta}_{ij}^k, \hat{\boldsymbol{\zeta}}_{ij}^{[g]}$ in place of $\mathbf{X}_j^k, \boldsymbol{\epsilon}_i^k, \zeta_{ij}^k, \boldsymbol{\zeta}_{ij}^{[g]}$ respectively. Proceeding in a similar fashion we get

$$\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\|^2 = \frac{1}{n}(\|\mathbf{Z}^{[g]}\|^2 + 2 \sum_{k \in g} Z^k (\mathbf{Q}_j^k)^T \boldsymbol{\delta}_i^k + \|(\mathbf{Q}_j^k)^T \boldsymbol{\delta}_i^k\|^2)$$

where $\mathbf{Z}^{[g]} = (Z^k)_{k \in g}; Z^k = (\mathbf{Q}_j^k)^T \boldsymbol{\epsilon}_i^k$ with $\boldsymbol{\epsilon}_i^k := \mathbf{Y}_i^k - \mathbf{Y}_{-i}^k \boldsymbol{\theta}_{0,i}^k - \mathbf{X}^k \mathbf{B}_{0,i}^k$, $\mathbf{Q}_j^k$ is the first eigenvector of $\mathbf{Y}_j^k (\mathbf{Y}_j^k)^T / n$, and $\boldsymbol{\delta}_i^k := \mathbf{X}^k(\mathbf{B}_{0,i}^k - \hat{\mathbf{B}}_i^k)$. Applying Cauchy-schwarz inequality to right side and by assumption (T2),

$$\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\| \leq \frac{1}{\sqrt{n}}(\|\mathbf{Z}^{[g]}\| + \sqrt{c(v_\beta)})$$

thus

$$\mathbb{P}(\{\mathcal{A}_{ij}^g\}^c) = \mathbb{P}\left(\|\hat{\boldsymbol{\zeta}}_{ij}^{[g]}\| > \frac{\lambda_{ij}^g}{2} \right) \leq \mathbb{P}\left(\|\mathbf{Z}^{[g]}\| > \frac{\sqrt{n}\lambda_{ij}^g}{2} - \sqrt{c(v_\beta)} \right)$$

We now proceed through the proof of Lemma E.2 in 15-656 to end up with the choice of λ_{ij}^g. $\qquad\square$

All subsequent derivations in the theorem go through with the new choice of λ_{ij}^g.

Part II. Proof of Thm 2 in 15-656 follows. We only need a new bound for $Var(\mathbf{Y}_i^k | \mathbf{Y}_{-i}^k, \mathbf{X}^k, \hat{\mathbf{B}}_i^k)$. For this we have

$$Var(\mathbf{Y}_i^k | \mathbf{Y}_{-i}^k, \mathbf{X}^k, \hat{\mathbf{B}}_i^k) = \mathbb{E}(\hat{\boldsymbol{\epsilon}}_i^k)^2 = \mathbb{E}(\boldsymbol{\epsilon}_i^k + \boldsymbol{\delta}_i^k)^2 \leq \left(\frac{1}{d_0} + \frac{c(v_\beta)}{n} \right)^2$$

applying cauchy-schwarz inequality followed by assumption (A2). Now Replace $1/\sqrt{nd_0}$ in choice of λ, α_n in Thm 2 statement with $1/\sqrt{n}(\sqrt{1/d_0} + \sqrt{c(v_\beta)/n})$.

$\qquad\square$

Proposition 3.3. *Given fixed $\hat{\mathcal{B}}$, prediction errors follow bound in T2 with high enough probability.*