Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models

Joint work with

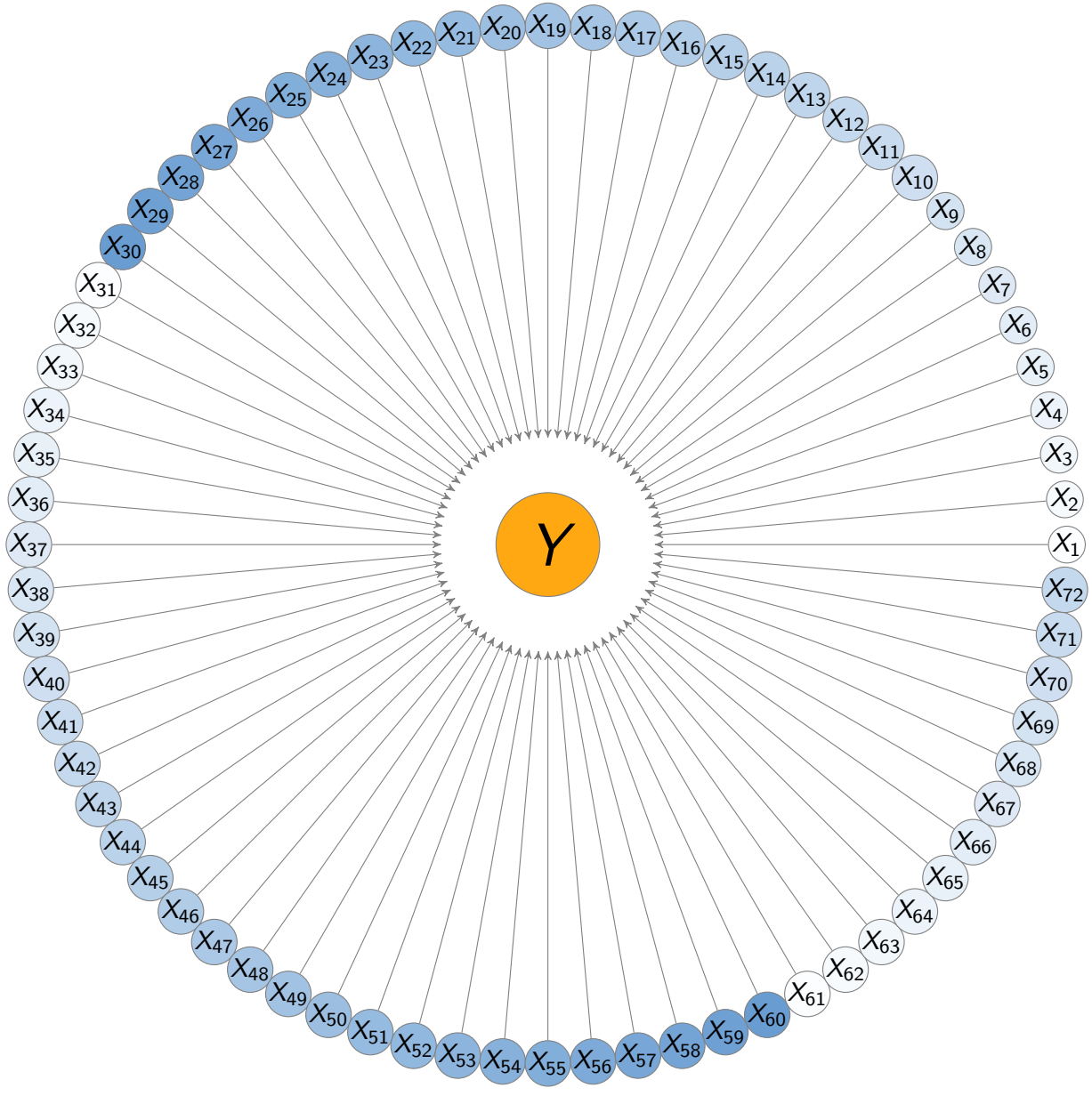
Yi Yang, Tianyuan Lu, Erwin Schurr, Celia Greenwood (McGill),
Marie Forest (ÉTS), JC Loredo-Osti (Memorial),
Karim Oualkacha (UQÀM)

CMStatistics, London 2019

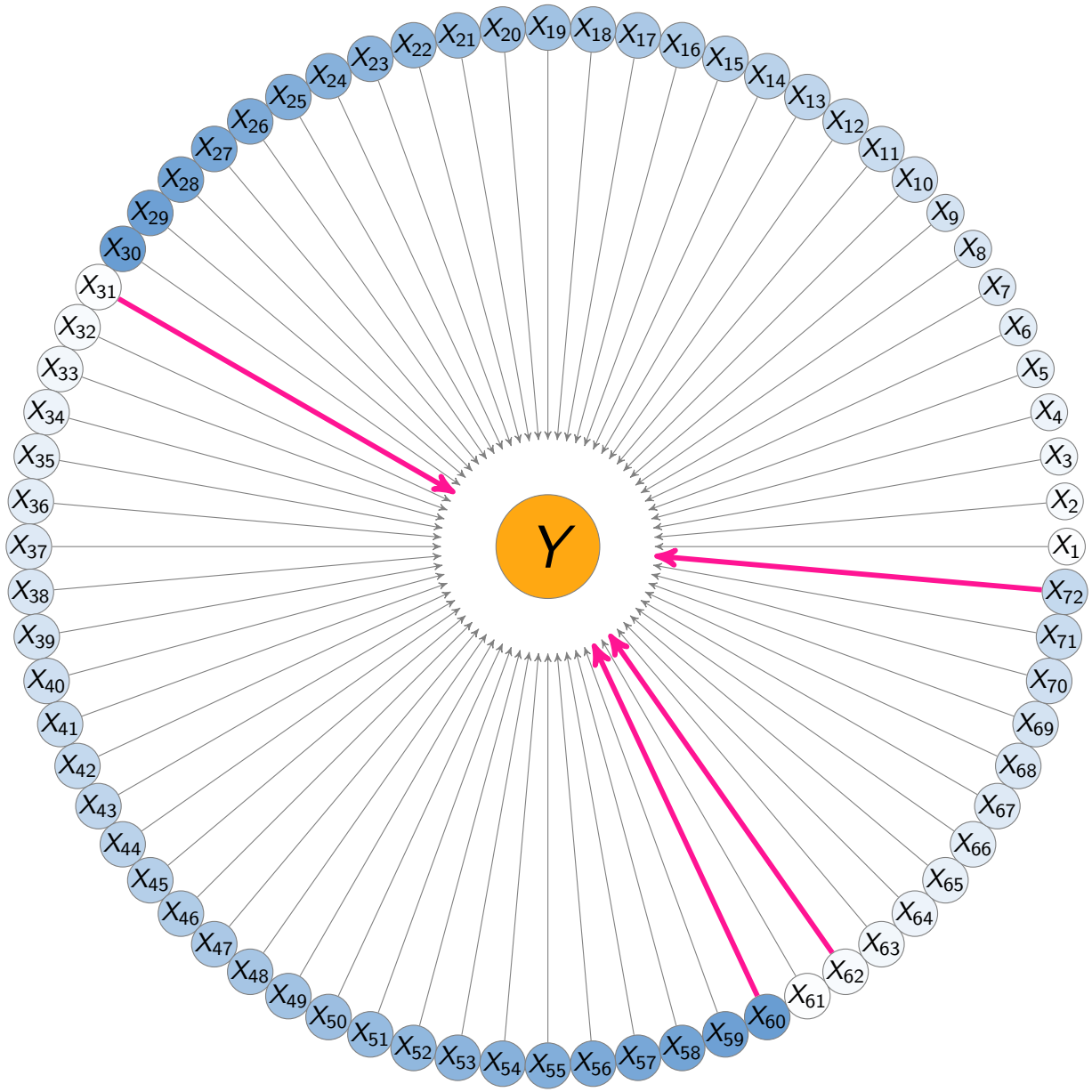
sahirbhatnagar.com

Betting on Sparsity

Bet on Sparsity Principle



Bet on Sparsity Principle



Bet on Sparsity Principle

Use a procedure that does well in sparse problems, since no procedure does well in dense problems.[1]

[1] The elements of statistical learning. Springer series in statistics, 2001.

Bet on Sparsity Principle

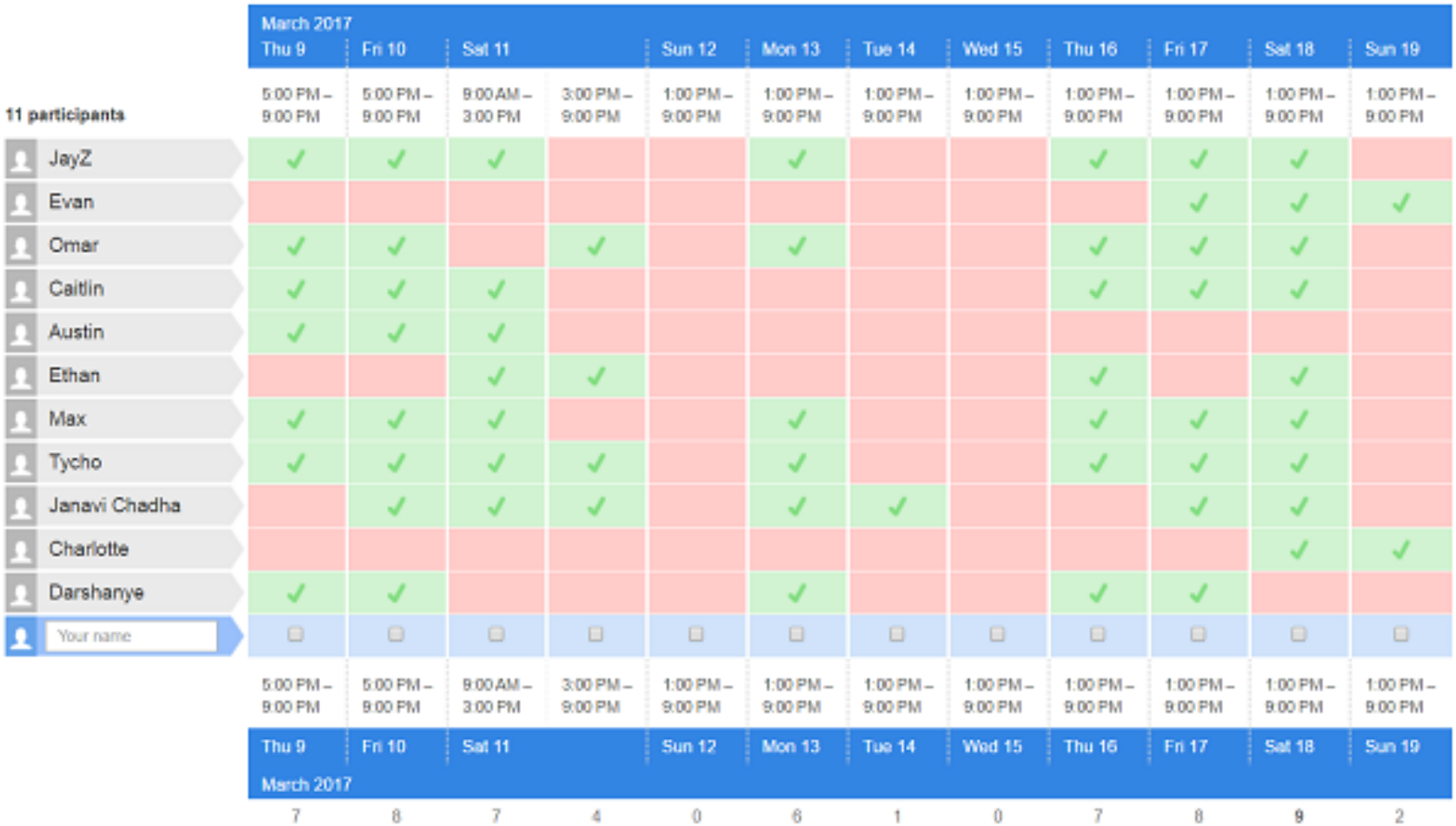
Use a procedure that does well in sparse problems, since no procedure does well in dense problems.[1]

- We often don't have enough data to estimate so many parameters
- Even when we do, we might want to identify a **relatively small number of predictors** ($k < N$) that play an important role
- Faster computation, easier to understand, and stable predictions on new datasets.

[1] The elements of statistical learning. Springer series in statistics, 2001.

How would you schedule a meeting of 20 people?

How would you schedule a meeting of 20 people?



Doctors Bet on Sparsity Also

Motivation

Motivating Dataset: Two Problems

	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	−1.255	1	2	0	0	0	1
2	4114347	−0.339	1	2	0	2	0	1
3	4399930	−0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	−0.421	2	2	0	1	0	1
7	4870063	−0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	−2.29	1	2	0	0	0	1
10	5805218	2.289	1	2	0	1	1	1

Problem 1: Which Predictors Affect the Response

	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	−1.255	1	2	0	0	0	1
2	4114347	−0.339	1	2	0	2	0	1
3	4399930	−0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	−0.421	2	2	0	1	0	1
7	4870063	−0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	−2.29	1	2	0	0	0	1
10	5805218	2.289	1	2	0	1	1	1

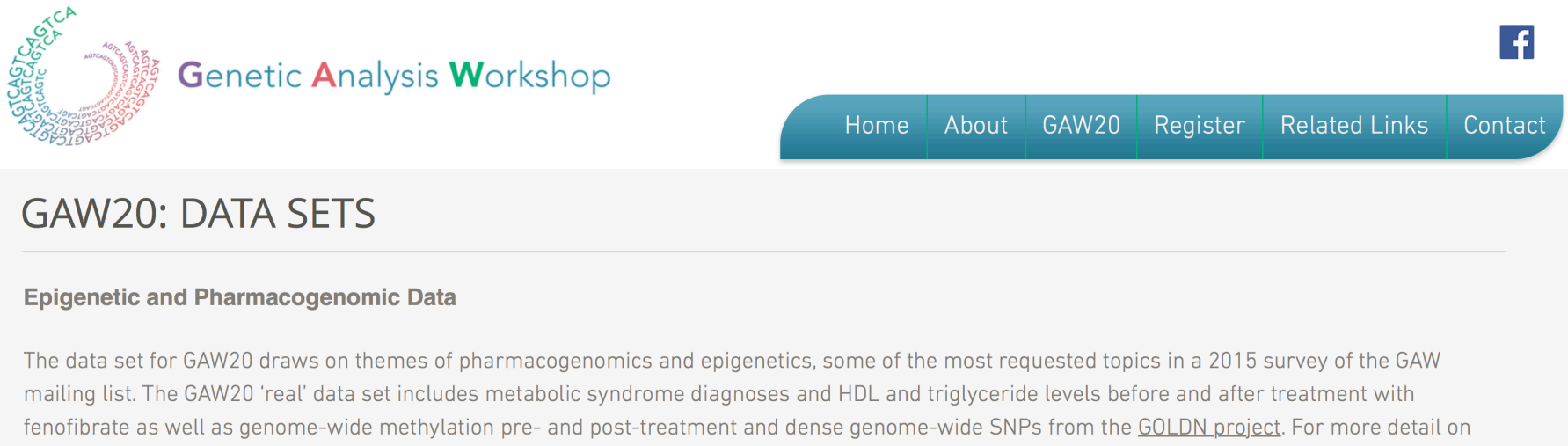
Problem 2: Observations are not Independent

- Observations are **correlated**, but this information is **unknown**
- However it can be **estimated** from the data

	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	−1.255	1	2	0	0	0	1
2	4114347	−0.339	1	2	0	2	0	1
3	4399930	−0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	−0.421	2	2	0	1	0	1
7	4870063	−0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	−2.29	1	2	0	0	0	1
10	5805218	2.289	1	2	0	1	1	1

GAW20 Dataset

Genetic Analysis Workshop (GAW20, March 4-7, 2017, San Diego, US)



[1] GOLDEN project: Genetics of Lipid Lowering Drugs and Diet Network Study

GAW20 Dataset

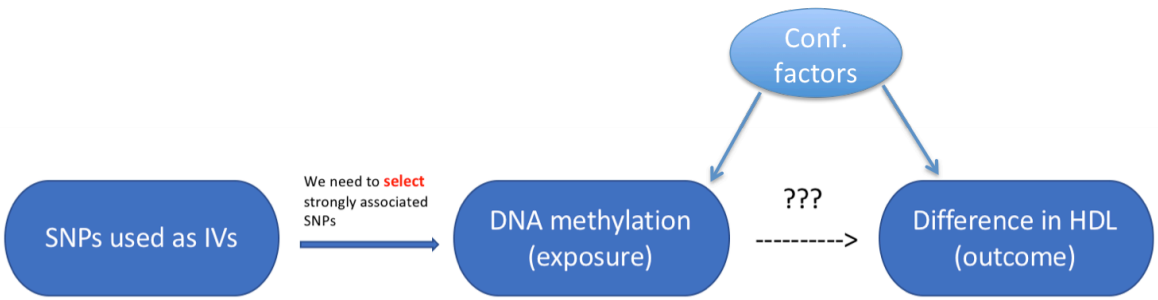
- Our contribution in GAW20

Investigating potential causal relationships between SNPs, DNA methylation and HDL

Lai Jiang[1,2], Kaiqiong Zhao[1,2], Kathleen Klein[2], Angelo J Canty[5],
Karim Oualkacha[3], Celia MT Greenwood*[1,2,4]

Motivation

- Our contribution in GAW20 consisted of investigation of causal relationship between DNA methylation (exposure) within some genes and ΔHDL (outcome)
- DNA methylation in these genes has been shown association with HDL
- We used Mendelian randomization to explore causal relationship
- We used SNPs around the analyzed genes as Instrumental Variables (IVs) to interrogate the causal relationship



Challenges in GAW20 Data Sets

- GAW20 SNPs data was high-dimensional
- There was a need for data regularization in order to select SNPs strongly associated with the exposure
- Penalized LS regression can be used (Lasso, SCAD, MCP or Elastic net)

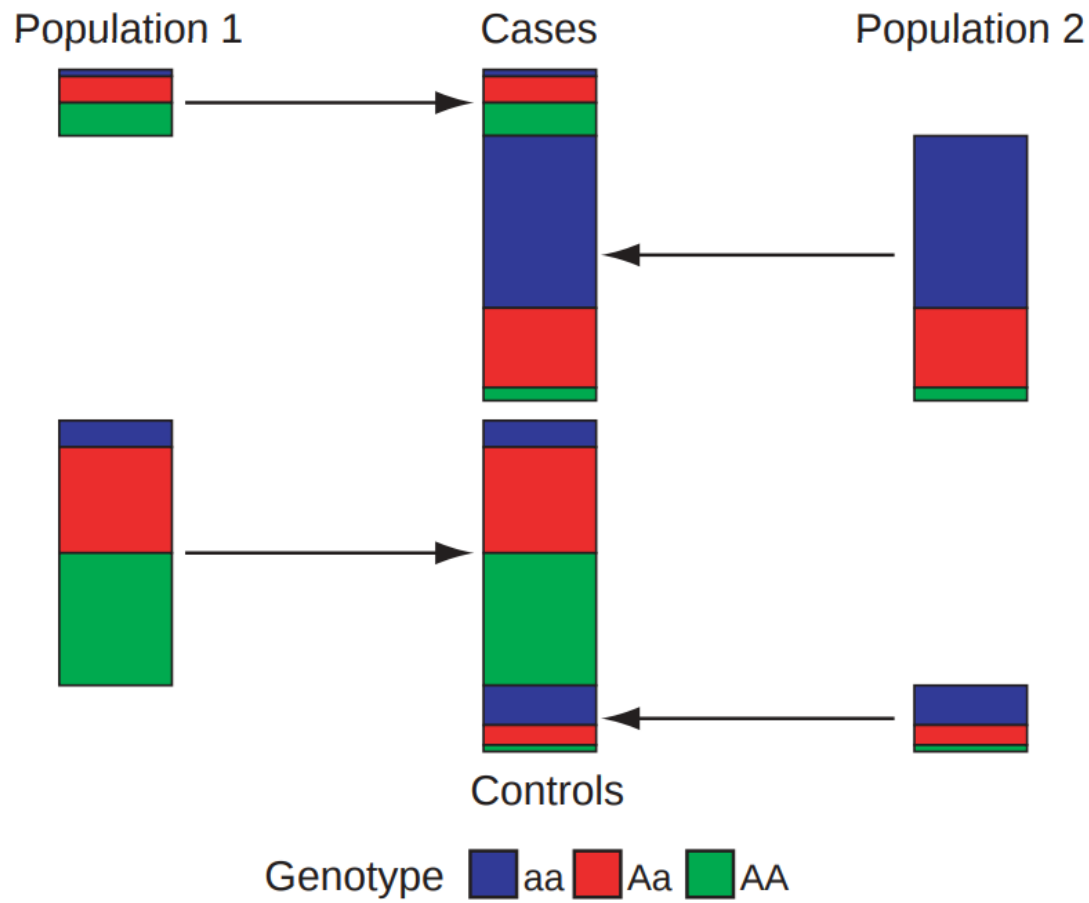
Challenges in GAW20 Data Sets

- GAW20 SNPs data was high-dimensional
- There was a need for data regularization in order to select SNPs strongly associated with the exposure
- Penalized LS regression can be used (Lasso, SCAD, MCP or Elastic net)
- But, data consists of families !
- In the GAW20, all penalized regression methods
  - either did not control for the family structure

Challenges in GAW20 Data Sets

- GAW20 SNPs data was high-dimensional
- There was a need for data regularization in order to select SNPs strongly associated with the exposure
- Penalized LS regression can be used (Lasso, SCAD, MCP or Elastic net)
- But, data consists of families !
- In the GAW20, all penalized regression methods
  - either did not control for the family structure
  - or used **two-stage adjustment** for the family structure (including our group)

Population structure in genetic association studies



[1] Marchini et al. Nature genetics (2004)

Kinship Matrix: Measuring Genetic Similarity

- Let *kinship* be a list of SNPs used to estimate the kinship matrix
- Let $X_{kinship}$ be a standardized $n \times q$ genotype matrix.
- A kinship matrix ($\mathbf{\Phi}$) can be computed as

$$\mathbf{\Phi} = \frac{1}{q-1} X_{kinship} X_{kinship}^{\top} \tag{1}$$

Two Stage Procedure

- Step 1: Fit a null LMM with a single random effect

$$\mathbf{Y} = \mathbf{P} + \boldsymbol{\varepsilon}$$

$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$$

- σ^2 is the phenotype total variance
- $\eta \in [0, 1]$ is the phenotype heritability (narrow sens)
- $\mathbf{Y}|(\eta, \sigma^2) \sim \mathcal{N}(\mathbf{0}, \eta\sigma^2\boldsymbol{\Phi} + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$

Two Stage Procedure

- Step 1: Fit a null LMM with a single random effect

$$\mathbf{Y} = \mathbf{P} + \boldsymbol{\varepsilon}$$

$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$$

- σ^2 is the phenotype total variance
- $\eta \in [0, 1]$ is the phenotype heritability (narrow sens)
- $\mathbf{Y}|(\eta, \sigma^2) \sim \mathcal{N}(\mathbf{0}, \eta\sigma^2\boldsymbol{\Phi} + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$
- Step 2: Use residuals from Step 1 as new *independent* response

Two step procedure

$\mathbf{X}_{\text{kinship}}$

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

Two step procedure

$X\_kinship$

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

$X\_kinship\ X\_kinship^{T}$

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

Two step procedure

X_kinship

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

X_kinship X_kinship^T

Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289

Y

$\sim$

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

P

$+$ **E**

Two step procedure

Step 1:

Y

Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289

$\sim$

P

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

$+ \; E_1$

Step 2:

Residuals from Step 1

$\sim$

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

$+ \; E_2$

Two step procedure

Step 1:

Y

Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289

$\sim$

P

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

$+ \; E_1$

Step 2:

Residuals from Step 1

$\sim$

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

$+ \; E_2$

- In association testing, it is known to suffer from huge power loss (Oualkacha et al. Gene. Epi. (2013))

Our proposal

Proposal

Aim:

We believe that performing variable selection and controlling for familial and/or hidden relationships simultaneously in high-dimensional settings, are likely to be of great interest to the genetics community

Proposal

Aim:

We believe that performing variable selection and controlling for familial and/or hidden relationships simultaneously in high-dimensional settings, are likely to be of great interest to the genetics community

Proposal:

We propose, `ggmix`, a **one stage** procedure which simultaneously controls for structured populations and performs variable selection in Linear Mixed Models (LMMs)

ggmix: One step procedure

Y

Response
−1.255
−0.339
−0.6
0.809
0.279
−0.421
−0.454
1.383
−2.29
2.289

$\sim$

X

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
ID1	2	2	2	2	2	2
ID2	0	2	2	2	2	2
ID3	0	2	2	2	2	2
ID4	1	2	2	2	2	2
ID5	0	2	2	2	2	2
ID6	1	2	2	2	1	2
ID7	2	2	2	2	1	2
ID8	1	2	2	2	2	2
ID9	0	2	2	2	1	2
ID10	1	2	2	1	2	2

+

P

	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10
ID1	0.97	0	0	0	−0.02	0.03	0.02	−0.01	−0.02	0.03
ID2	0	1	0	−0.01	0	−0.01	−0.01	0	0	0
ID3	0	0	0.98	0.01	0.01	0.01	0	0.03	−0.01	−0.01
ID4	0	−0.01	0.01	1.03	0.04	0.01	−0.01	0.01	0.01	−0.01
ID5	−0.02	0	0.01	0.04	0.97	−0.01	−0.01	0.01	0.03	0.03
ID6	0.03	−0.01	0.01	0.01	−0.01	1.02	0	0	0	0.01
ID7	0.02	−0.01	0	−0.01	−0.01	0	1	0.02	0.02	0
ID8	−0.01	0	0.03	0.01	0.01	0	0.02	1.01	0.01	0
ID9	−0.02	0	−0.01	0.01	0.03	0	0.02	0.01	1.04	0.01
ID10	0.03	0	−0.01	−0.01	0.03	0.01	0	0	0.01	0.95

+ E

[2] Bhatnagar et al. Revision submitted (2019+)

[3] R package: `sahirbhatnagar.com/ggmix`

Data and Model

- Phenotype: $\mathbf{Y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- SNPs: $\mathbf{X} = (\mathbf{X}_1; \ldots, \mathbf{X}_n)^T \in \mathbb{R}^{n \times p}$, where $p \gg n$
- Twice the Kinship matrix or Realized Relationship matrix: $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$
- Polygenic random effect: $\mathbf{P} = (P_1, \ldots, P_n) \in \mathbb{R}^n$
- Error: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$

Data and Model

- Phenotype: $\mathbf{Y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- SNPs: $\mathbf{X} = (\mathbf{X}_1; \ldots, \mathbf{X}_n)^T \in \mathbb{R}^{n \times p}$, where $p \gg n$
- Twice the Kinship matrix or Realized Relationship matrix: $\mathbf{\Phi} \in \mathbb{R}^{n \times n}$
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$
- Polygenic random effect: $\mathbf{P} = (P_1, \ldots, P_n) \in \mathbb{R}^n$
- Error: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$
- We consider the following LMM with a single random effect:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P} + \boldsymbol{\varepsilon}$$

$$\mathbf{P} \sim \mathcal{N}(0, \eta\sigma^2\mathbf{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$$

- σ^2 is the phenotype total variance
- $\eta \in [0, 1]$ is the phenotype heritability (narrow sens)
- $\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\mathbf{\Phi} + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$

Likelihood

- The negative log-likelihood is given by

$$-\ell(\mathbf{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(\mathbf{V})\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y} - \mathbf{X}\beta\right)^T \mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\beta\right)$$

$$\mathbf{V} = \eta\mathbf{\Phi} + (1-\eta)\mathcal{I}$$

Likelihood

- The negative log-likelihood is given by

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(\mathbf{V})\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y} - \mathbf{X}\beta\right)^T \mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\beta\right)$$

$$\mathbf{V} = \eta\boldsymbol{\Phi} + (1-\eta)\boldsymbol{\mathcal{I}}$$

- Assume the spectral decomposition of $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

- $\mathbf{U}$ is an $n \times n$ orthogonal matrix and $\mathbf{D}$ is an $n \times n$ diagonal matrix

Likelihood

- The negative log-likelihood is given by

$$-\ell(\mathbf{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(\mathbf{V})\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y} - \mathbf{X}\beta\right)^T \mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\beta\right)$$

$$\mathbf{V} = \eta\mathbf{\Phi} + (1-\eta)\mathcal{I}$$

- Assume the spectral decomposition of $\mathbf{\Phi}$

$$\mathbf{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

- $\mathbf{U}$ is an $n \times n$ orthogonal matrix and $\mathbf{D}$ is an $n \times n$ diagonal matrix
- One can write

$$\mathbf{V} = \mathbf{U}(\eta\mathbf{D} + (1-\eta)\mathcal{I})\mathbf{U}^\top = \mathbf{U}\mathbf{W}\mathbf{U}^\top$$

Likelihood

- The negative log-likelihood is given by

$$-\ell(\boldsymbol{\Theta}) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\log\left(\det(\mathbf{V})\right) + \frac{1}{2\sigma^2}\left(\mathbf{Y} - \mathbf{X}\beta\right)^T \mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\beta\right)$$

$$\mathbf{V} = \eta\boldsymbol{\Phi} + (1-\eta)\boldsymbol{\mathcal{I}}$$

- Assume the spectral decomposition of $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

- $\mathbf{U}$ is an $n \times n$ orthogonal matrix and $\mathbf{D}$ is an $n \times n$ diagonal matrix
- One can write

$$\mathbf{V} = \mathbf{U}(\eta\mathbf{D} + (1-\eta)\boldsymbol{\mathcal{I}})\mathbf{U}^\top = \mathbf{U}\mathbf{W}\mathbf{U}^\top$$

with $\mathbf{W} = \text{diag}\left(w_i\right)_{i=1}^n$, $w_i = \eta\mathbf{D}_{ii} + (1-\eta)$

Likelihood

- Projection of $\mathbf{Y}$ (and columns of $\mathbf{X}$) into Span($\mathbf{U}$) leads to a simplified correlation structure for the transformed data: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\beta, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\beta, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$

Likelihood

- Projection of $\mathbf{Y}$ (and columns of $\mathbf{X}$) into Span($\mathbf{U}$) leads to a simplified correlation structure for the transformed data: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\beta, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\beta, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$
- The negative log-likelihood can then be expressed as

$$-\ell(\mathbf{\Theta}) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \log(w_i) + \frac{1}{2\sigma^2} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)^T \mathbf{W}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)$$

Likelihood

- Projection of $\mathbf{Y}$ (and columns of $\mathbf{X}$) into Span($\mathbf{U}$) leads to a simplified correlation structure for the transformed data: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\beta, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\beta, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$
- The negative log-likelihood can then be expressed as

$$-\ell(\mathbf{\Theta}) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \log(w_i) + \frac{1}{2\sigma^2} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)^T \mathbf{W}^{-1} (\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)$$

- For fixed σ^2 and η, solving for β is a weighted least squares problem

Penalized Maximum Likelihood Estimator

- Define the objective function:

$$Q_\lambda(\boldsymbol{\Theta}) = -\ell(\boldsymbol{\Theta}) + \lambda \sum_j p_j(\beta_j)$$

- $p_j(\cdot)$ is a penalty term on $\beta_1, \ldots, \beta_p$
- An estimate of the model parameters $\widehat{\boldsymbol{\Theta}}_\lambda$ is obtained by

$$\widehat{\boldsymbol{\Theta}}_\lambda = \arg\min_{\boldsymbol{\Theta}} Q_\lambda(\boldsymbol{\Theta})$$

Block Relaxation (De Leeuw, 1994)

To solve for the optimization problem we use a block relaxation technique

Set $k \leftarrow 0$, initial values for the parameter vector $\boldsymbol{\Theta}^{(0)}$ and ϵ;
for $\lambda \in \{\lambda_{max}, \ldots, \lambda_{min}\}$ **do**
  repeat
    For $j = 1, \ldots, p$,

$$\beta_j^{(k+1)} \leftarrow \arg\min_{\beta_j} Q_\lambda \left(\boldsymbol{\beta}_{-j}^{(k)}, \eta^{(k)}, \sigma^{2\ (k)} \right)$$

$$\eta^{(k+1)} \leftarrow \arg\min_{\eta} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta, \sigma^{2\ (k)} \right)$$

$$\sigma^{2\ (k+1)} \leftarrow \arg\min_{\sigma^2} Q_\lambda \left(\boldsymbol{\beta}^{(k+1)}, \eta^{(k+1)}, \sigma^2 \right)$$

    $k \leftarrow k + 1$
  until *convergence criterion is satisfied:*
  $||\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Theta}^{(k)}||_2 < \epsilon$;
end

Algorithm 1: Block Relaxation Algorithm

Coordinate Gradient Descent Method

- We take advantage of smoothness of $\ell(\mathbf{\Theta})$
- We approximate $Q_\lambda(\mathbf{\Theta})$ by a strictly convex quadratic function (using gradient)
- We use CGD to calculate a descent direction
- To achieve the descent property for the objective function, we employ further line search

[1] Tseng P& Yun S. Math. Program., Ser. B, (2009)

Coordinate Gradient Descent Method

- We take advantage of smoothness of $\ell(\mathbf{\Theta})$
- We approximate $Q_\lambda(\mathbf{\Theta})$ by a strictly convex quadratic function (using gradient)
- We use CGD to calculate a descent direction
- To achieve the descent property for the objective function, we employ further line search

Theorem [Convergence] [1]**:**
If $\{\mathbf{\Theta}^{(k)}, k = 0, 1, 2, \ldots\}$ is a sequence of iterates generated by the iteration map of Algorithm 1, then each cluster point (i.e. limit point) of $\{\mathbf{\Theta}^{(k)}, k = 0, 1, 2, \ldots\}$ is a stationary point of $Q_\lambda(\mathbf{\Theta})$

[1] Tseng P& Yun S. Math. Program., Ser. B, (2009)

Choice of the tuning parameter

- We use the BIC:

$$BIC_\lambda = -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \widehat{\eta}) + c \cdot \widehat{df}_\lambda$$

- $\widehat{df}_\lambda$ is the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_\lambda$ plus two [1]
- Several authors [2] have used this criterion for variable selection in mixed models with $c = \log n$
- Other authors [3] have proposed $c = \log(\log(n)) * \log(n)$

[1] Zou et al. The Annals of Statistics, (2007)
[2] Bondell et al. Biometrics (2010)
[3] Wang et al. JRSS(Ser. B), (2009)

Results

Simulation Results

		1% Causal SNPs			
		No overlap		All causal SNPs in kinship	
Metric	Method	10%	30%	10%	30%
TPR at FPR=5%	twostep	0.84 (0.05)	0.84 (0.05)	0.76 (0.09)	0.77 (0.08)
	lasso	0.86 (0.05)	0.85 (0.05)	0.86 (0.05)	0.86 (0.05)
	ggmix	0.86 (0.05)	0.86 (0.05)	0.85 (0.05)	0.86 (0.05)
Model Size	twostep	338 (71)	339 (68)	289 (62)	285 (55)
	lasso	282 (51)	281 (52)	285 (50)	284 (54)
	ggmix	43 (7)	43 (8)	44 (8)	43 (9)
RMSE	twostep	1.42 (0.10)	1.41 (0.10)	1.44 (0.33)	1.40 (0.22)
	lasso	1.39 (0.09)	1.38 (0.09)	1.40 (0.08)	1.38 (0.08)
	ggmix	1.22 (0.10)	1.20 (0.10)	1.23 (0.11)	1.23 (0.12)
Estimation Error	twostep	2.97 (0.60)	2.92 (0.60)	3.60 (5.41)	3.21 (3.46)
	lasso	2.76 (0.46)	2.69 (0.47)	2.82 (0.48)	2.75 (0.48)
	ggmix	2.11 (1.28)	2.04 (1.22)	2.21 (1.24)	2.28 (1.34)

Real data applications

1. **UK Biobank**
   - 10,000 LD-pruned SNPs (Essentially un-correlated variables) to predict standing height in 18k related individuals
   - Standing height is highly polygenic (many variables associated with response)

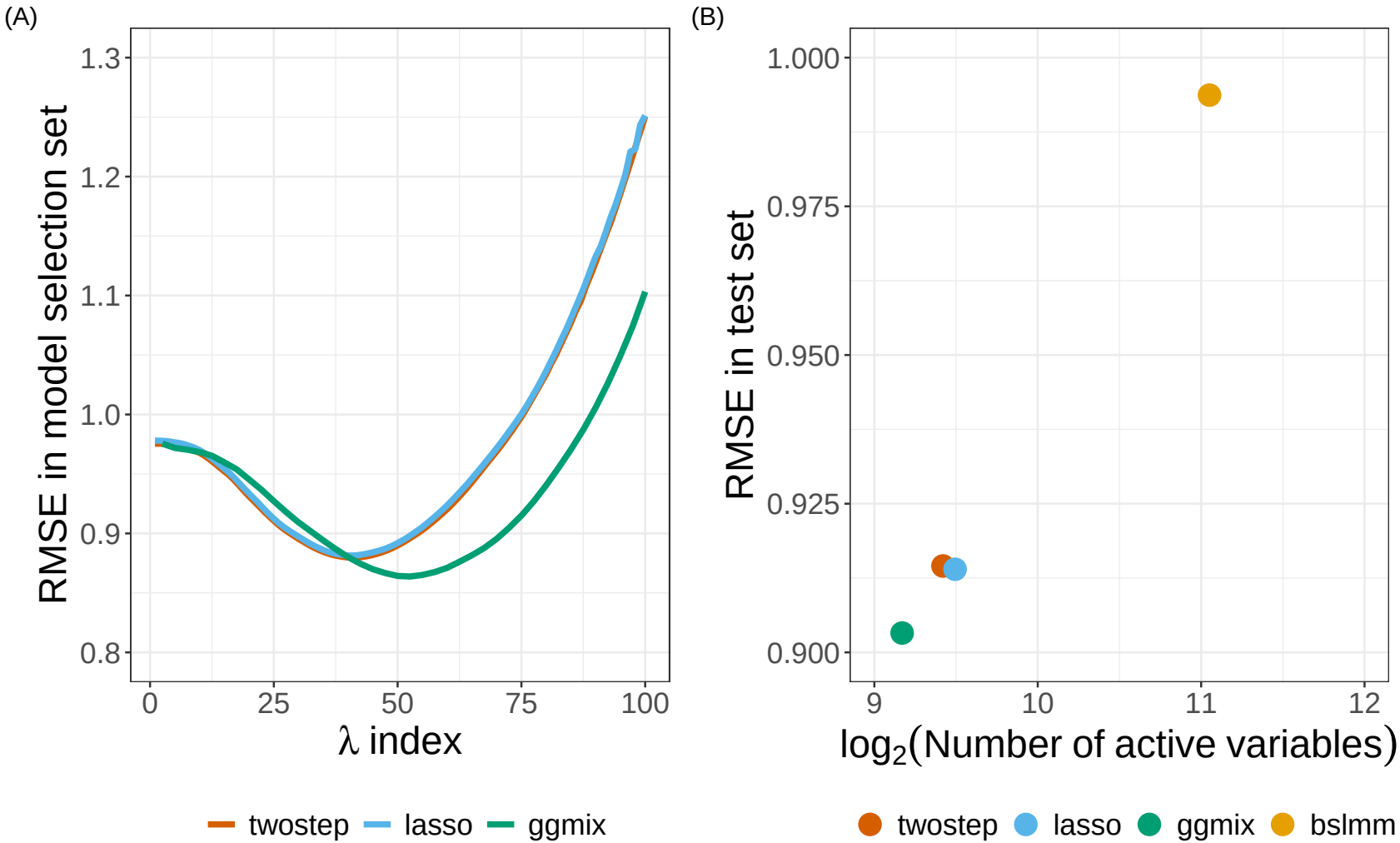
Real data applications

1. **UK Biobank**
   - 10,000 LD-pruned SNPs (Essentially un-correlated variables) to predict standing height in 18k related individuals
   - Standing height is highly polygenic (many variables associated with response)
2. **GAW20 Simulated dataset**
   - 50,000 SNPs (all on chromosome 1) to predict high-density lipoproteins in 679 related individuals
   - Not much correlation between causal SNP and others
   - Very sparse signals (only 1 causal variant)

Real data applications

1. **UK Biobank**
   - 10,000 LD-pruned SNPs (Essentially un-correlated variables) to predict standing height in 18k related individuals
   - Standing height is highly polygenic (many variables associated with response)
2. **GAW20 Simulated dataset**
   - 50,000 SNPs (all on chromosome 1) to predict high-density lipoproteins in 679 related individuals
   - Not much correlation between causal SNP and others
   - Very sparse signals (only 1 causal variant)
3. **Mouse Crosses**
   - Find loci associated with mouse sensitivity to mycobacterial infection
   - 189 samples, and 625 microsatellite markers
   - Highly correlated variables

Results: UK Biobank

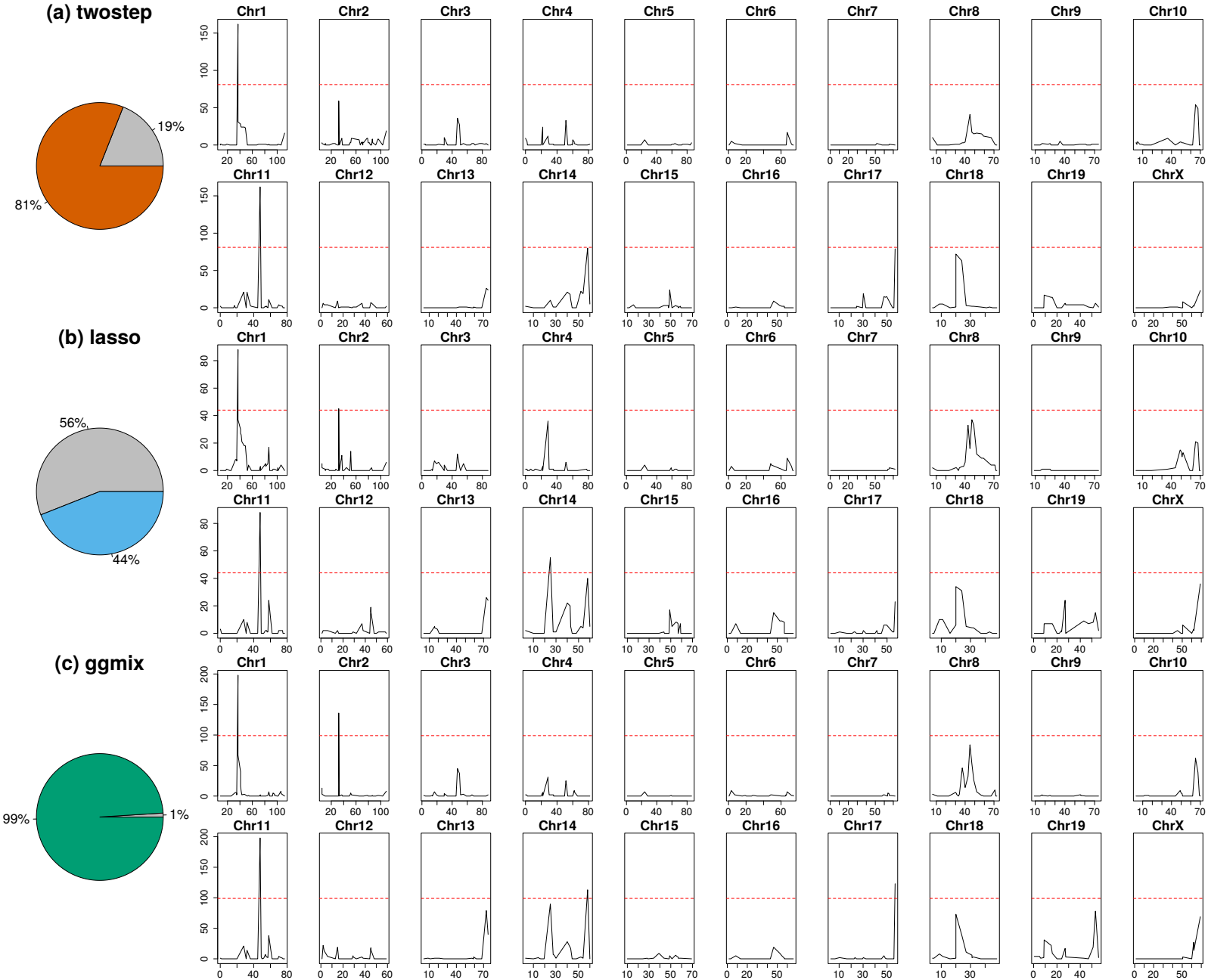


Results: GAW20

Method	**Median number of active variables (Inter-quartile range)**	**RMSE (SD)**
`twostep`	1 (1 - 11)	0.3604 (0.0242)
`lasso`	1 (1 - 15)	0.3105 (0.0199)
`ggmix`	1 (1 - 12)	0.3146 (0.0210)
`BSLMM`	40,737 (39,901 - 41,539)	0.2503 (0.0099)

Table 1: Summary of model performance based on 200 GAW20 simulations. Five-fold cross-validation root-mean-square error was reported for each simulation replicate.

Results: Mouse crosses



Discussion and Future Work

Discussion

- Two-step procedure leads to a large number of false positives and false negatives

Discussion

- Two-step procedure leads to a large number of false positives and false negatives
- Principal component adjustment in `lasso` may not be sufficient to control for confounding, particularly when there is a lot of correlation between observations

Discussion

- Two-step procedure leads to a large number of false positives and false negatives
- Principal component adjustment in `lasso` may not be sufficient to control for confounding, particularly when there is a lot of correlation between observations
- `ggmix` performs well even when the causal variables are used in the calculation of the kinship matrix

Discussion

- Two-step procedure leads to a large number of false positives and false negatives
- Principal component adjustment in `lasso` may not be sufficient to control for confounding, particularly when there is a lot of correlation between observations
- `ggmix` performs well even when the causal variables are used in the calculation of the kinship matrix
- `ggmix` showed the biggest improvement over `twostep` and `lasso` when there were highly correlated variables with lots of structure (e.g. mouse crosses example)

Future work

- `ggmix` is limited by the number of individuals (not applicable to enitre UK Biobank cohort of 500k) $\rightarrow$ low-rank approximations to kinship matrix

Future work

- `ggmix` is limited by the number of individuals (not applicable to enitre UK Biobank cohort of 500k) $\rightarrow$ low-rank approximations to kinship matrix
- Run into memory issues when the number of covariates in the model exceeds 50k $\rightarrow$ memory mapping strategies (e.g. `biglasso` by Zeng and Breheny (2017))

Future work

- `ggmix` is limited by the number of individuals (not applicable to enitre UK Biobank cohort of 500k) $\rightarrow$ low-rank approximations to kinship matrix
- Run into memory issues when the number of covariates in the model exceeds 50k $\rightarrow$ memory mapping strategies (e.g. `biglasso` by Zeng and Breheny (2017))
- Extension to other (non-convex) penalties $\rightarrow$ more consistent variable selection

Future work

- `ggmix` is limited by the number of individuals (not applicable to enitre UK Biobank cohort of 500k) $\rightarrow$ low-rank approximations to kinship matrix
- Run into memory issues when the number of covariates in the model exceeds 50k $\rightarrow$ memory mapping strategies (e.g. `biglasso` by Zeng and Breheny (2017))
- Extension to other (non-convex) penalties $\rightarrow$ more consistent variable selection
- Model selection. Is HDBIC appropriate? $\rightarrow$ `cAIC4` (Greven et al.) ?

References

1. Sahir R Bhatnagar, Yi Yang, Tianyuan Lu, Erwin Schurr, JC Loredo-Osti, Marie Forest, Karim Oualkacha, and Celia MT Greenwood. Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *Revision submitted*. https://doi.org/10.1101/408484
2. Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. Nature methods, 8(10):833–835, 2011.
3. Matti Pirinen, Peter Donnelly, Chris CA Spencer, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics, 7(1):369–390, 2013.
4. Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming, 117(1):387–423, 2009.
5. Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for gen- eralized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010.

`sahirbhatnagar.com`