Advances in mixed effects models for prediction and variable selection in high-dimensional data

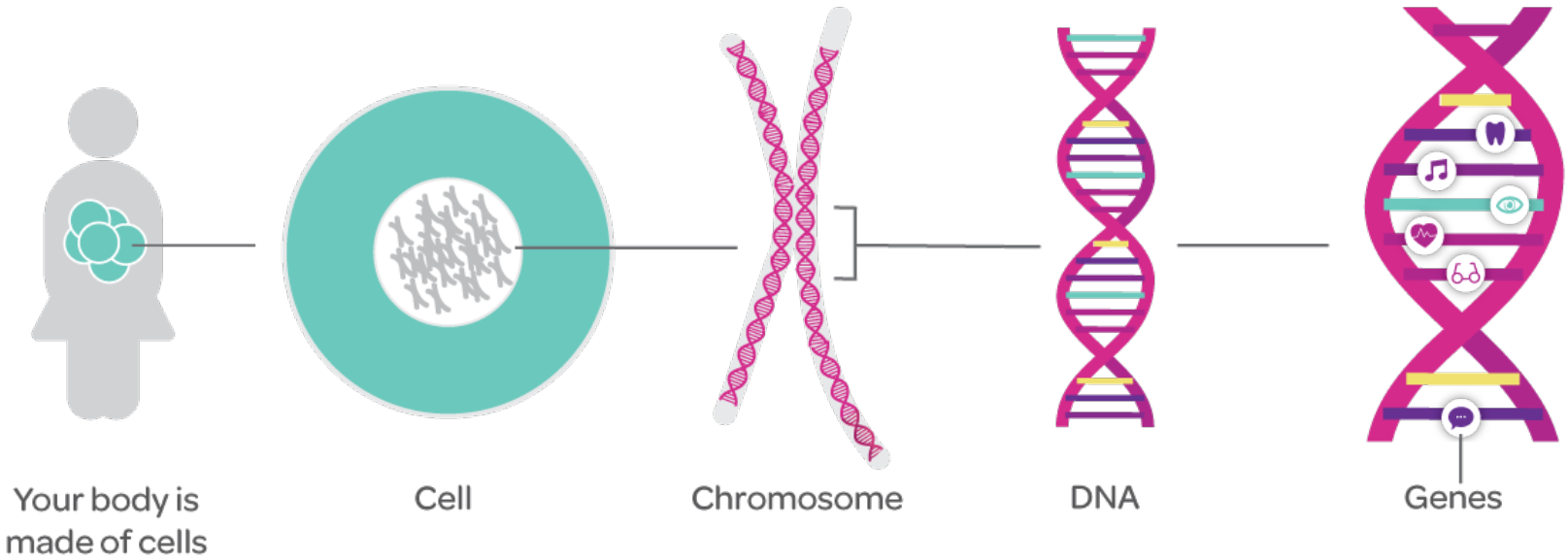
Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
Department of Diagnostic Radiology
McGill University

sahirbhatnagar.com

April 14, 2021

Chromosomes, DNA and Genes

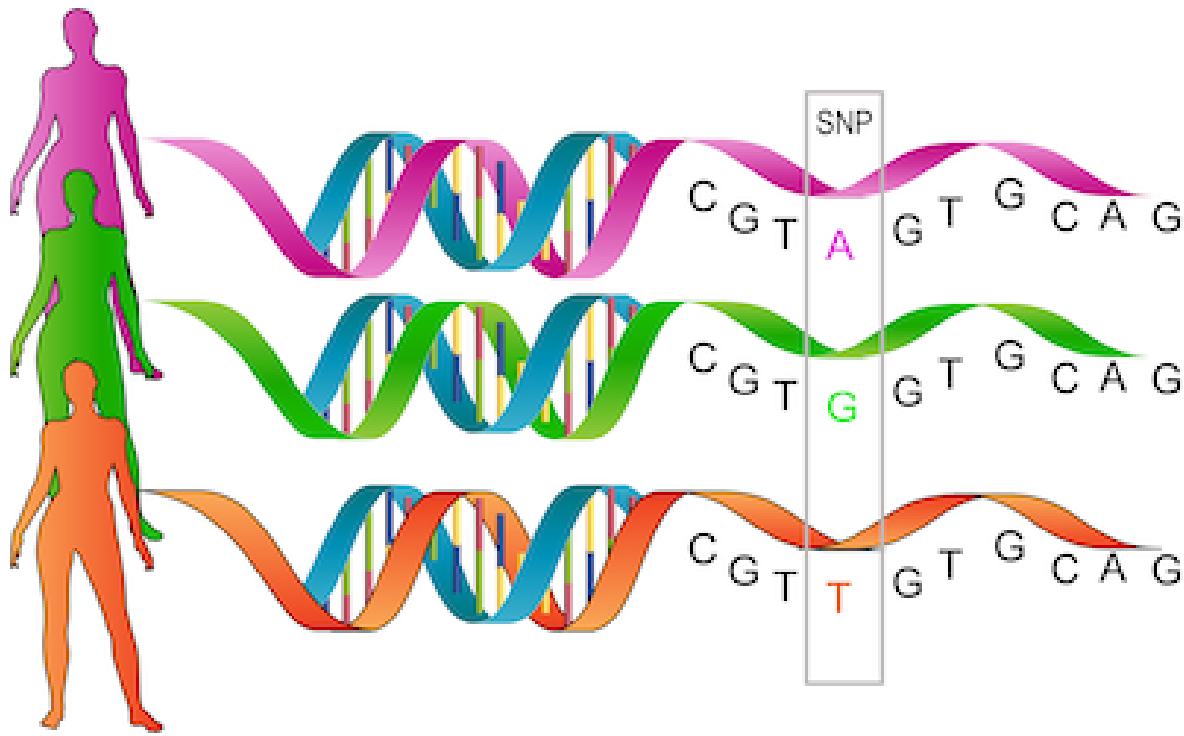
- DNA is a molecule containing genetic information written in a language whose words consist of 4 letters: A, T, C, G.
- A gene is a piece of DNA formed by a precise sequence of several of these letters; this sequence of letters forms the sequence of the gene.
- There are more than 25,000 genes in the human genome that code for various physical characteristics and control the functioning of the body and contribute to the state of health of the individual at all stages of their life.



[1] https://mirakind.org/genetics-101

Single-nucleotide polymorphism (SNP)

- A SNP is a variation at a single position in a DNA sequence among individuals
- If more than 1% of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP.



[1] https://www.nature.com/scitable/definition/snp-295/

UKBiobank: High-dimensional data ($n << p$)

- UK Biobank has collected and continues to collect extensive environmental, lifestyle, and genetic data on half a million participants
- Genotyping: 800,000 genome-wide variants and imputation to 90 million variants
- Lots of phenotypes (response variables) available including bone mineral density, height, mental health, ect.
- Objectives
  - Prediction: determine the optimal weighted combination of SNPs to predict the response e.g. polygenic risk scores (PRS)
  - Discovery: which SNPs cause the phenotype

[1] https://www.ukbiobank.ac.uk/enable-your-research/about-our-data

Example 1: Prediction study with UK Biobank

- Osteoporosis screening identifies only a small proportion of the screened population to be eligible for intervention to prevent osteoporosis-related fractures
- Much of the screening expenditure is spent on individuals who will not qualify for intervention

PLOS MEDICINE

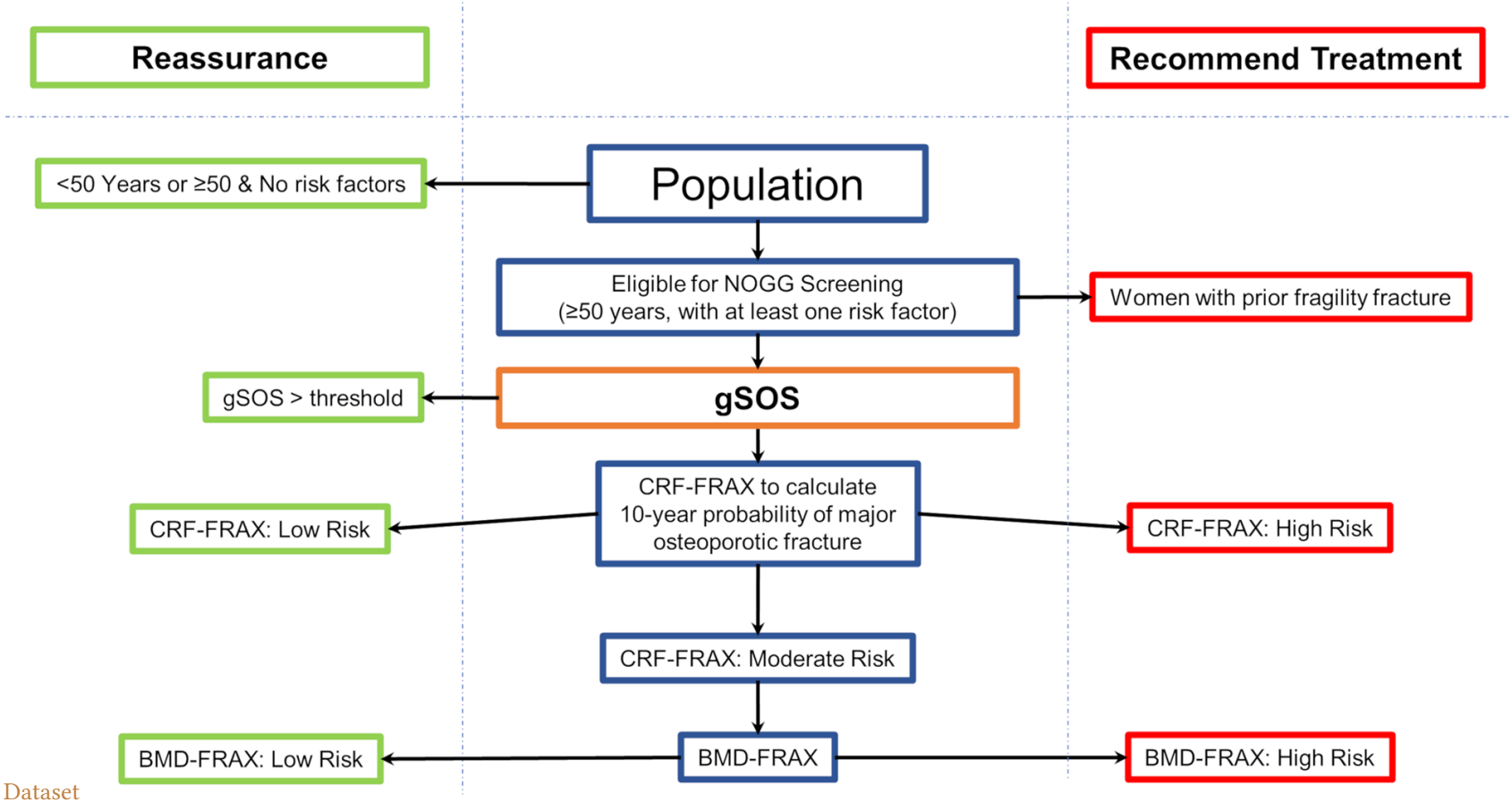
RESEARCH ARTICLE

Development of a polygenic risk score to improve screening for fracture risk: A genetic risk prediction study

Vincenzo Forgetta[1☯], Julyan Keller-Baruch[2☯], Marie Forest[1], Audrey Durand[3], Sahir Bhatnagar[1], John P. Kemp[4,5], Maria Nethander[6,7], Daniel Evans[8], John A. Morris[1], Douglas P. Kiel[9], Fernando Rivadeneira[10], Helena Johansson[11,12], Nicholas C. Harvey[13,14], Dan Mellström[7], Magnus Karlsson[15], Cyrus Cooper[13,14,16], David M. Evans[4,5], Robert Clarke[17], John A. Kanis[11,12], Eric Orwoll[18,19], Eugene V. McCloskey[20], Claes Ohlsson[7], Joelle Pineau[3], William D. Leslie[21], Celia M. T. Greenwood[1,2,22,23], J. Brent Richards[1,2,24]*

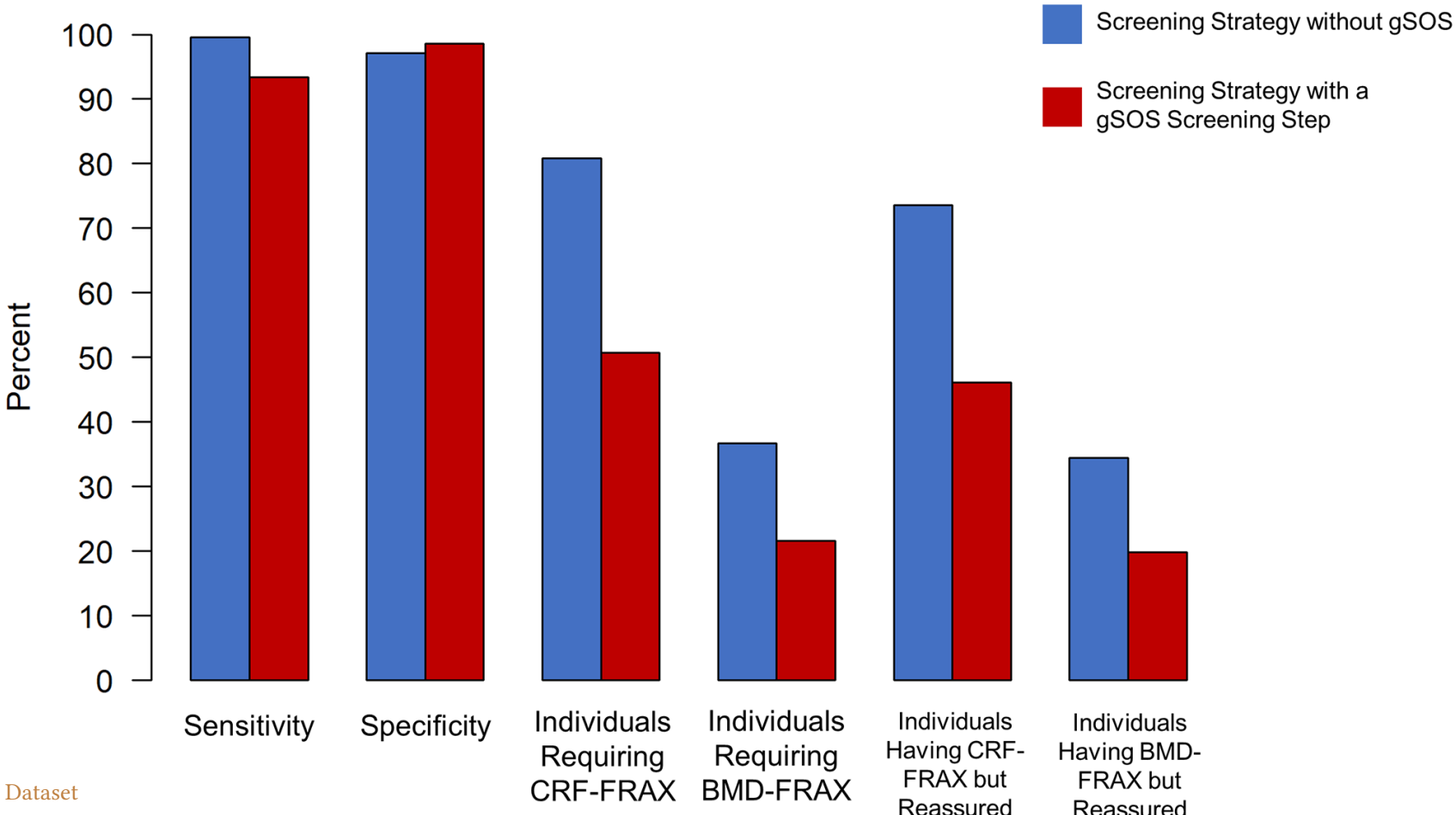
gSOS: a PRS for heel quantitative ultrasound speed of sound (SOS)

- SOS is a heritable risk factor for osteoporotic fracture
- The prediction of SOS using PRS could decrease the number of screened individuals by reassuring those with low genetic risk (*negative screening*)



gSOS: a cheap *negative screening* tool

- 81% of the population required expensive testing to achieve 99.6% sensitivity and 97.1% specificity
- Our polygenic risk score (gSOS) consisting of 21,717 genetic variants, only required 51% of the population while maintaining similar sensitivity and specificity.



Example 2: Identify individuals with rare variants

- An LDL-C PRS could be used to identify individuals with a higher probability of harboring FH variants
- We find that those with a low LDL-C PRS had a 21-fold higher probability of carrying an FH variant compared with those with a high LDL-C PRS

Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated With Risk of Ischemic Heart Disease and Enriches for Individuals With Familial Hypercholesterolemia

Haoyu Wu, MSc; Vincenzo Forgetta, PhD; Sirui Zhou, PhD; Sahir R. Bhatnagar, PhD; Guillaume Paré, MD; J. Brent Richards, MD

My former MSc student Haoyu Wu

Classical Methods

- A nice and powerful toolbox for analyzing the more traditional datasets where the sample size (N) is far **greater than** the number of covariates (p):
  - linear regression, logistic regression, LDA, QDA, glm,
  - regression spline, smoothing spline, kernel smoothing, local smoothing, GAM,
  - Neural Network, SVM, Boosting, Random Forest, ...

$$\mathbf{X}_{n\times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{12} & \cdots & x_{1p} \\ x_{31} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & x_{np} \end{bmatrix}$$

High-dimensional data ($n << p$)

- Our data are *wide*
- e.g. UK Biobank – $\mathbf{X} \in \mathbb{R}^{500{,}000 \times 90{,}000{,}000}$

$$\mathbf{X}_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x_{np} \end{bmatrix}$$

New challenges arise from how such data is *used*

A	
y	x_1
0.0	0
2.1	1
2.7	0
5.9	3
7.3	3
0.0	0
2.0	1

B								
y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	2	0	0	1	0	1	0
2.1	1	0	2	3	2	0	0	3
2.7	0	0	0	2	2	1	1	1
5.9	3	0	1	0	0	0	2	0
7.3	3	4	0	1	1	1	0	0
0.0	0	2	0	0	3	0	0	0
2.0	1	0	2	1	0	0	0	1

Estimated model	R^2_{adj}
$y = 0.66 + 1.92x_1$	0.83
$y = 0.22 + 1.78x_1 + 0x_2 + 0x_3 + 0x_4 + 0x_5 + 2.11x_6 + 0x_7 + 0x_8$	0.98

Overarching reaserch focus: including prior information

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \{ \text{DataFitting } [\mathbf{X}, y, \beta] + \lambda \text{ Prior } [\beta] \}$$

Overarching reaserch focus: including prior information

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \{ \text{DataFitting } [\mathbf{X}, y, \beta] + \lambda \text{ Prior } [\beta] \}$$

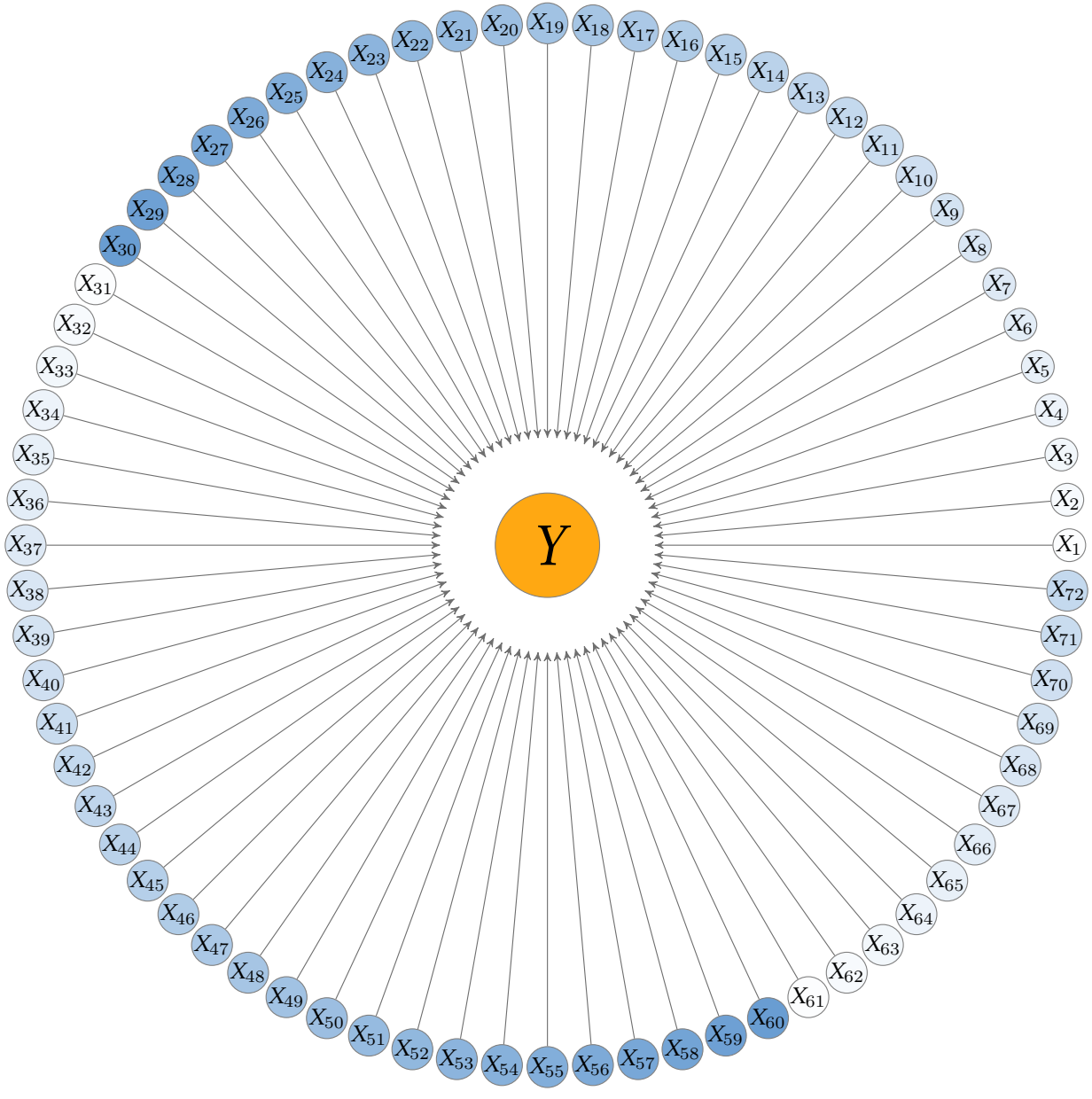
Examples:

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_0 \qquad \text{(Best subset selection)}$$

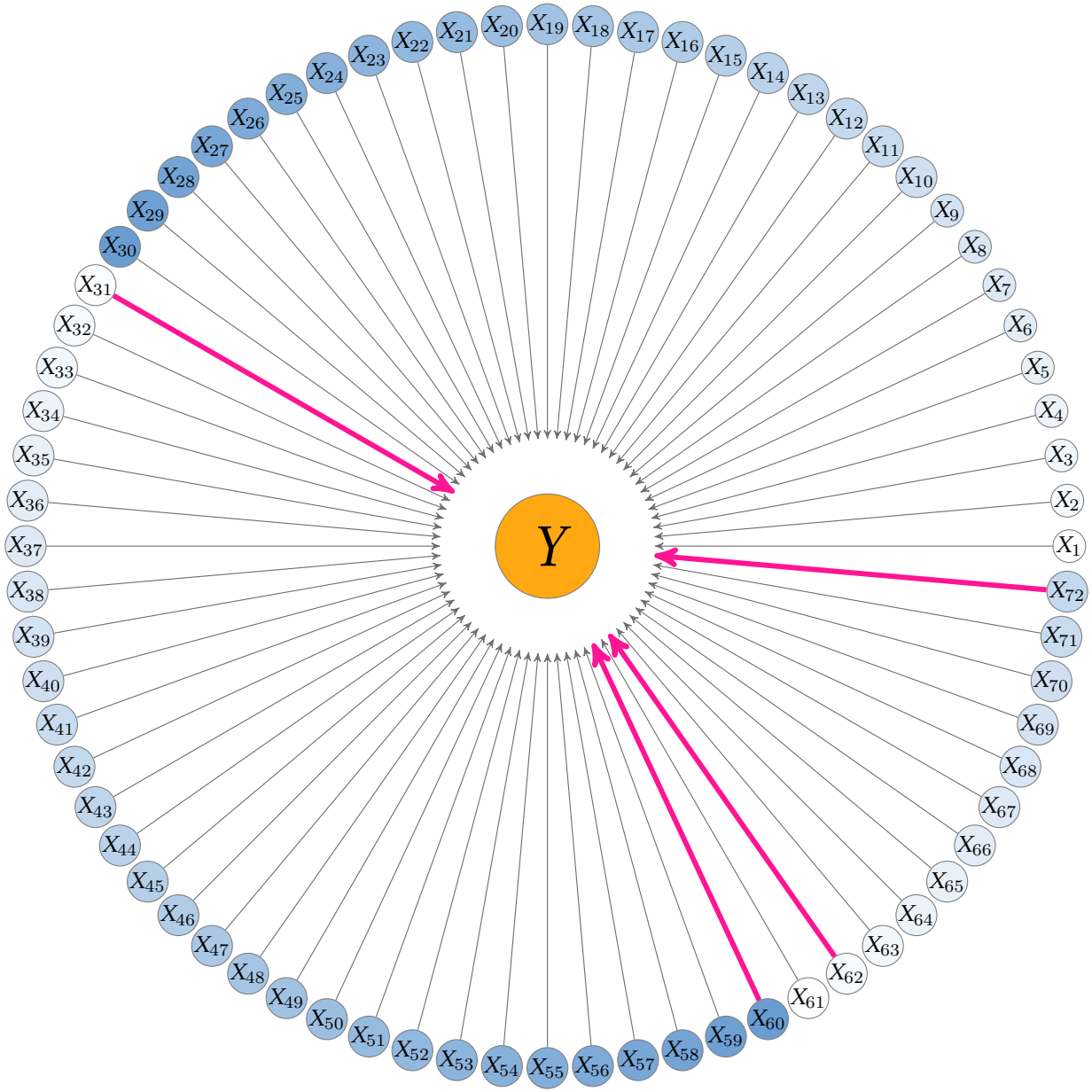
$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \qquad \text{(Lasso regression)}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \qquad \text{(Ridge regession)}$$

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

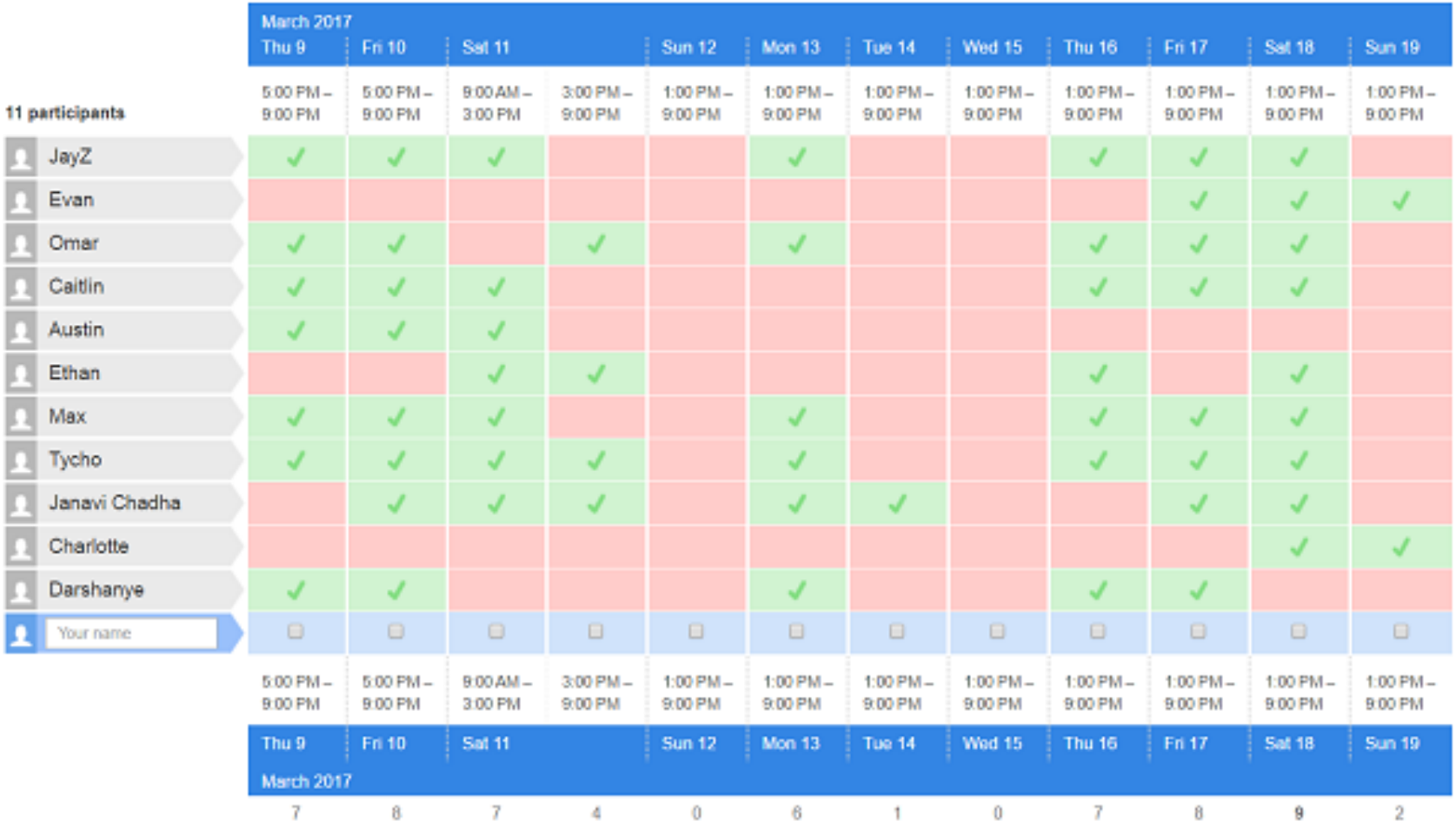
Use a procedure that does well in sparse problems, since no procedure does well in dense problems.[1]

- 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.

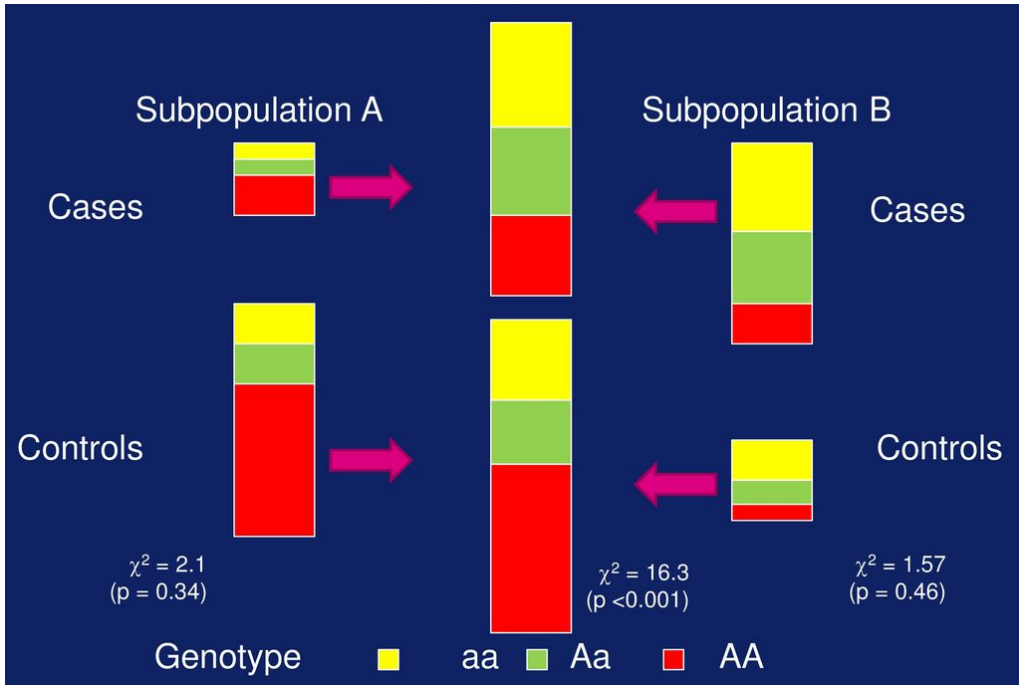
How would you schedule a meeting of 20 people?

How would you schedule a meeting of 20 people?



Doctors also bet on sparsity

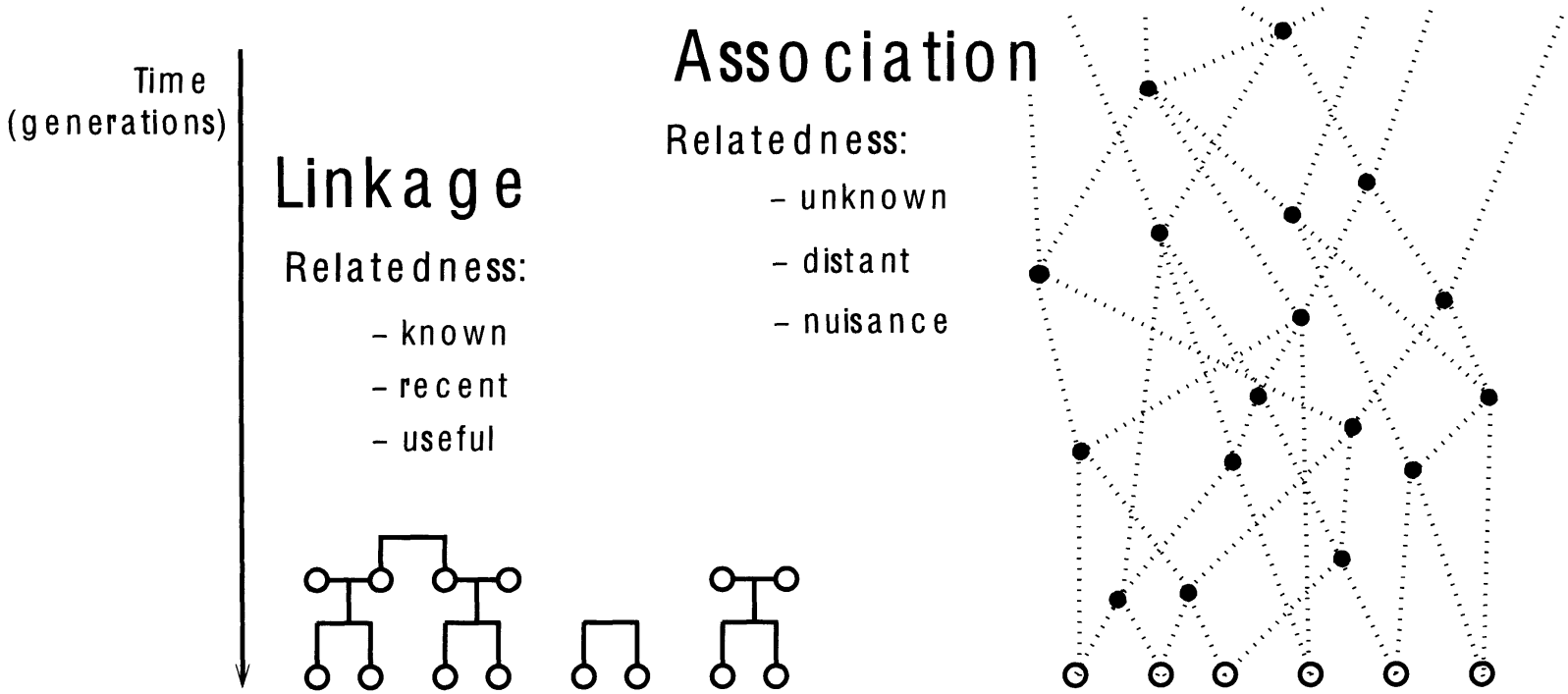
Additional challenges in genetic data – confounding by population structure



[1] Tam V. et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet (2019)

Population structure

- GWAS compares unrelated individuals, but *unrelated* signifies that the relationships are typically unknown and assumed to be distant



[1] Astle and Balding. Population structure and cryptic relatedness in genetic association studies. Statistical Science (2009)

Observations are not independant

- Observations are **correlated**, but this relationship is often **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

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

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

Multivariable Penalized Linear mixed effects models (LMM)

$$\mathbf{Y} = \sum_{j=1}^{p} \beta_j \cdot \text{SNP}_j + \mathbf{P} + \boldsymbol{\varepsilon} \tag{2}$$

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

- σ^2 is the phenotype total variance
- $\eta \in [0, 1]$ is the phenotype heritability
- $\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\sum_{j=1}^{p} \beta_j \cdot \text{SNP}_j, \eta\sigma^2\boldsymbol{\Phi} + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$
- In our applications, $n << p$

Multivariable Penalized Linear mixed effects models (LMM)

$$\mathbf{Y} = \sum_{j=1}^{p} \beta_j \cdot \text{SNP}_j + \mathbf{P} + \boldsymbol{\varepsilon} \tag{2}$$

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

- σ^2 is the phenotype total variance
- $\eta \in [0, 1]$ is the phenotype heritability
- $\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\sum_{j=1}^{p} \beta_j \cdot \text{SNP}_j, \eta\sigma^2\boldsymbol{\Phi} + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$
- In our applications, $n << p$

Lasso, ridge, ect. are not directly applicable to LMM

Current solution: 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)
- $\mathbf{Y}|(\eta, \sigma^2) \sim \mathcal{N}(\mathbf{0}, \eta\sigma^2\boldsymbol{\Phi} + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$

Current solution: 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

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

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

	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: ggmix

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

PLOS GENETICS

RESEARCH ARTICLE

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

Sahir R. Bhatnagar[1,2]*, **Yi Yang**[3], **Tianyuan Lu**[4,5], **Erwin Schurr**[6], **JC Loredo-Osti**[7], **Marie Forest**[8], **Karim Oualkacha**[9], **Celia M. T. Greenwood**[1,4,5,10,11]

1 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada, **2** Department of Diagnostic Radiology, McGill University, Montréal, Québec, Canada, **3** Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada, **4** Quantitative Life Sciences, McGill University, Montréal, Québec, Canada, **5** Lady Davis Institute, Jewish General Hospital, Montréal, Québec, Canada, **6** Department of Medicine, McGill University, Montréal, Québec, Canada, **7** Department of Mathematics and Statistics, Memorial University, St. John's, Newfoundland and Labrador, Canada, **8** École de Technologie Supérieure, Montréal, Québec, Canada, **9** Département de Mathématiques, Université du Québec à Montréal, Montréal, Québec, Canada, **10** Gerald Bronfman Department of Oncology, McGill University, Montréal, Québec, Canada, **11** Department of Human Genetics, McGill University, Montréal, Québec, Canada

* sahir.bhatnagar@mcgill.ca

Check for updates

[1] R package: `sahirbhatnagar.com/ggmix`, `https://cran.r-project.org/package=ggmix`

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

[1] R package: sahirbhatnagar.com/ggmix, https://cran.r-project.org/package=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: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$
- 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

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

where

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

Likelihood

- The negative log-likelihood is given by

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

  where

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

- 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

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

  where

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

- 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: $\tilde{\mathbf{Y}} = \mathbf{U}^\top \mathbf{Y}$
- $\tilde{\mathbf{Y}} | (\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 \mathbf{W})$, with $\tilde{\mathbf{X}} = \mathbf{U}^\top \mathbf{X}$
- The negative log-likelihood can then be expressed as

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

Likelihood

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

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

- For fixed σ^2 and η, solving for $\boldsymbol{\beta}$ 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

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

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

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

    $k \leftarrow k + 1$
  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 $\ell(\boldsymbol{\Theta})$
- We approximate $Q_\lambda(\boldsymbol{\Theta})$ 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

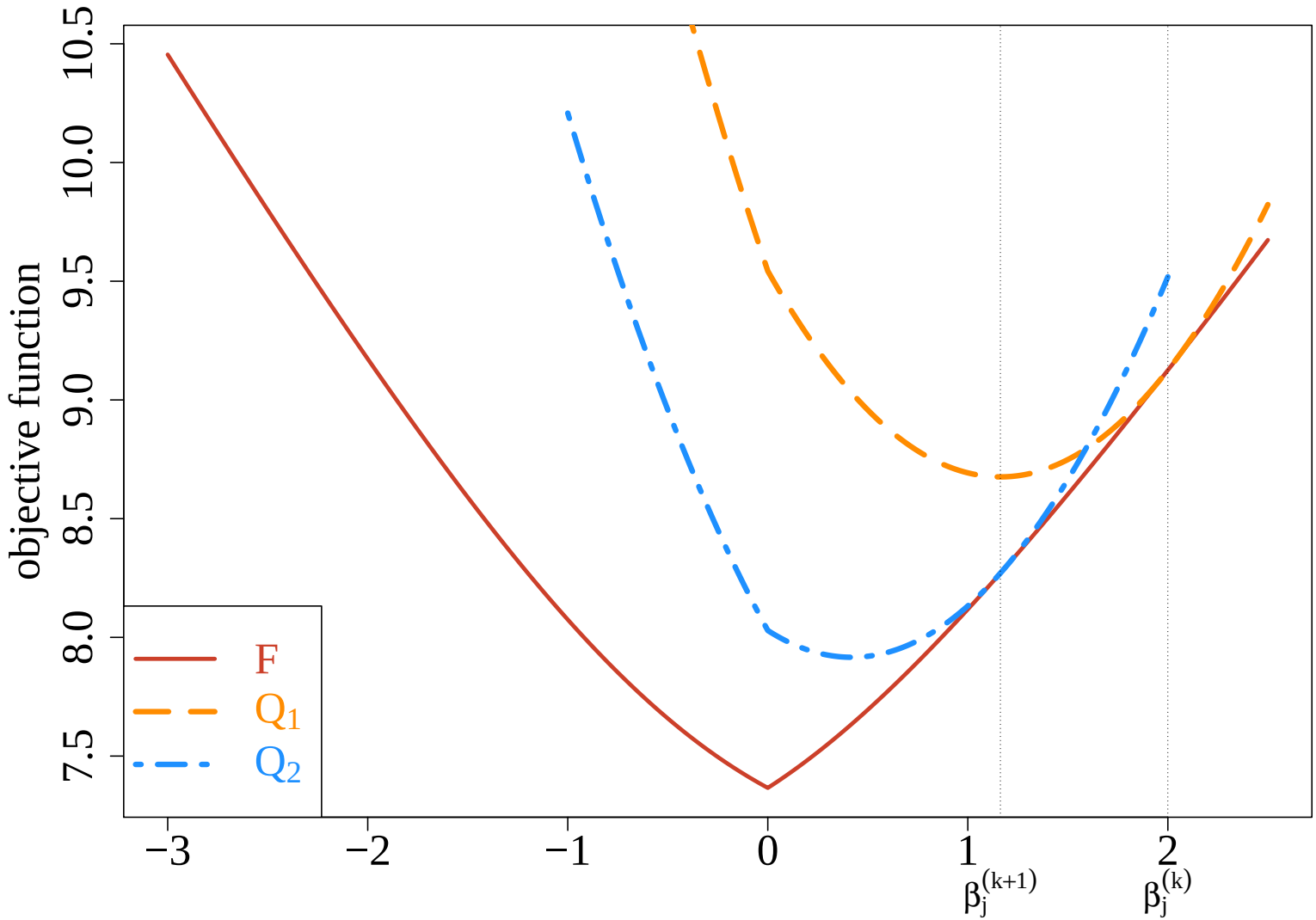
[1] Tseng P& Yun S. Math. Program., Ser. B, (2009)

Coordinate Gradient Descent Method

- We take advantage of smoothness of $\ell(\boldsymbol{\Theta})$
- We approximate $Q_\lambda(\boldsymbol{\Theta})$ 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 $\{\boldsymbol{\Theta}^{(k)}, k = 0, 1, 2, \ldots\}$ is a sequence of iterates generated by the iteration map of Algorithm 1, then each cluster point (i.e. limit point) of $\{\boldsymbol{\Theta}^{(k)}, k = 0, 1, 2, \ldots\}$ is a stationary point of $Q_\lambda(\boldsymbol{\Theta})$

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


objective function
10.5
10.0
9.5
9.0
8.5
8.0
7.5
F
Q1
Q2
−3
−2
−1
0
1
2
β_j^(k+1)
β_j^(k)

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)

Simulation study

- We simulated data from the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{P} + \boldsymbol{\varepsilon}$
- We used heritability $\eta = \{0.1, 0.3\}$, number of covariates $p = 5,000$, number of *kinship* SNPs $k = 10,000$, percentage of *causal* SNPs $c = \{0\%, 1\%\}$ and $\sigma^2 = 1$.
- In addition to these parameters, we also varied the amount of overlap between the *causal* list and the *kinship* list:
  1. None of the *causal* SNPs are included in *kinship* set.
  2. All of the *causal* SNPs are included in the *kinship* set.
- These were meant to contrast the model behavior when causal SNPs are included in both the main effects and random effects vs. when the causal SNPs are only included in the main effects.
- These scenarios are motivated by the current standard of practice in GWAS where the candidate marker is excluded from the calculation of the kinship matrix.
- This approach becomes much more difficult to apply in large-scale multivariable models where there is likely to be overlap between the variables in the design matrix and kinship matrix.

Simulation study results

- Both the `lasso+PC` and `twostep` selected more false positives compared to `ggmix`
- Overall, we observed that variable selection results and RMSE for `ggmix` were similar regardless of whether the causal SNPs were in the kinship matrix or not.
- This result is encouraging since in practice the kinship matrix is constructed from a random sample of SNPs across the genome, some of which are likely to be causal, particularly in polygenic traits.
- In particular, our simulation results show that the principal component adjustment method may not be the best approach to control for confounding by population structure, particularly when variable selection is of interest.

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)

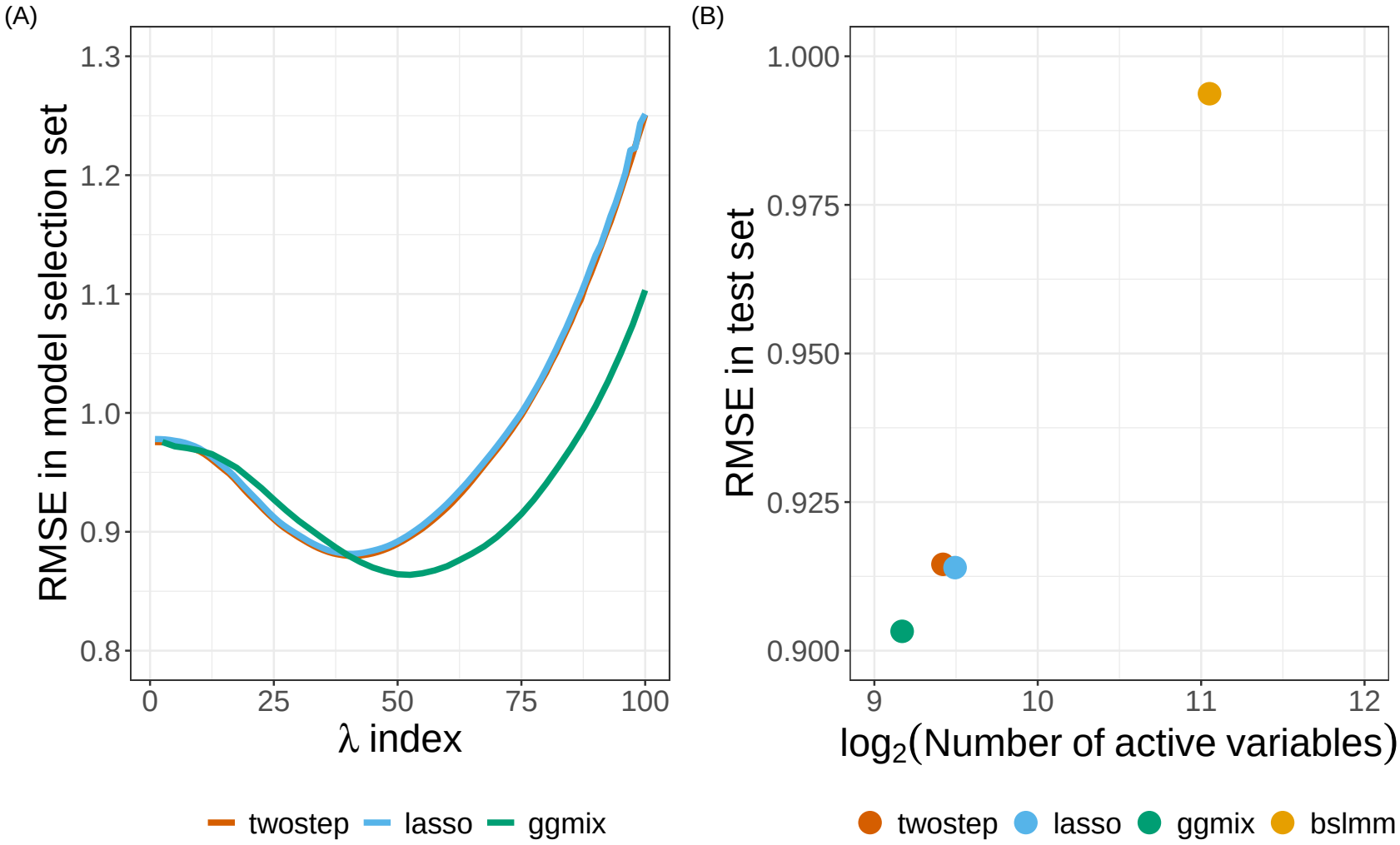
Real data applications

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

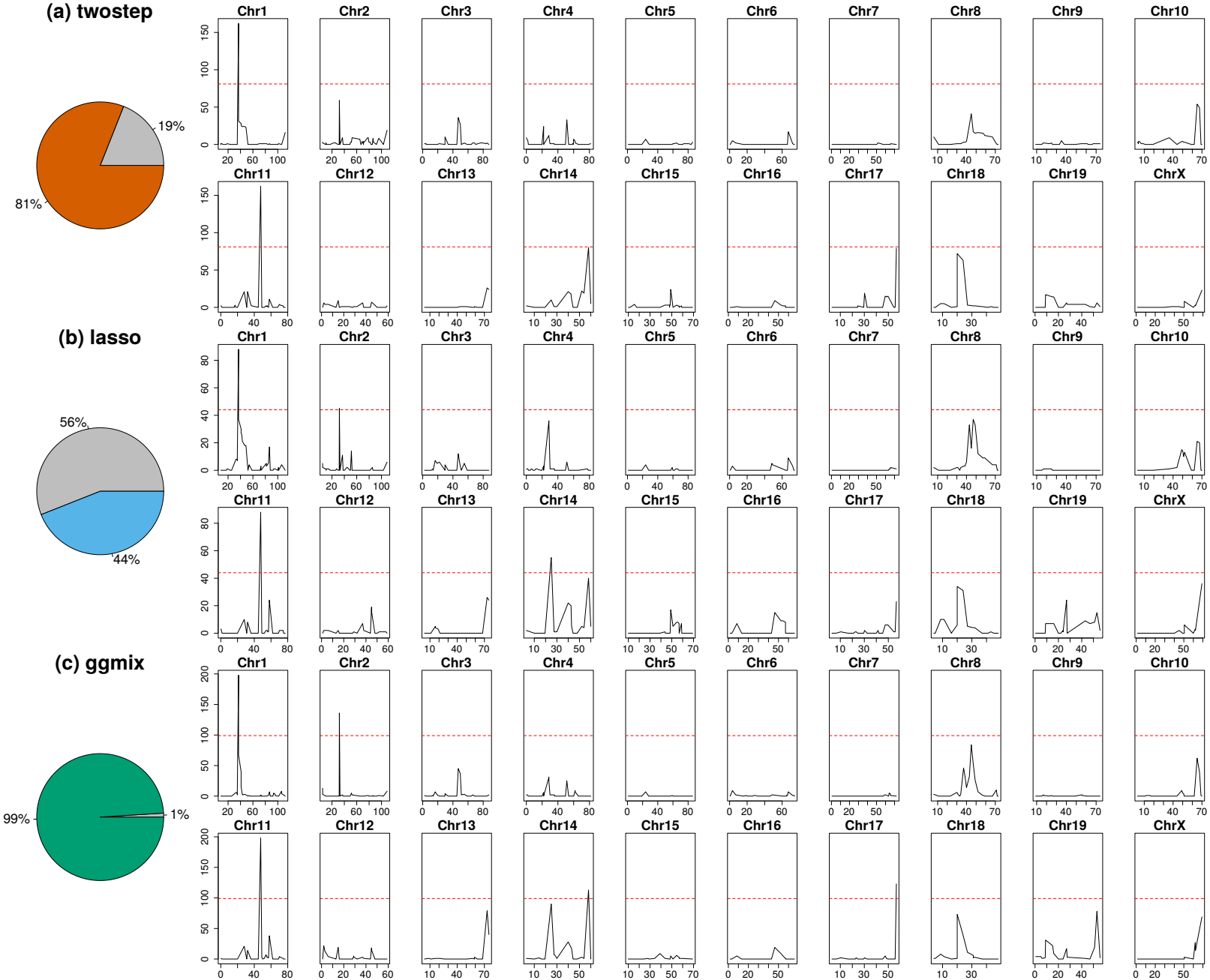


Results: GAW20

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)

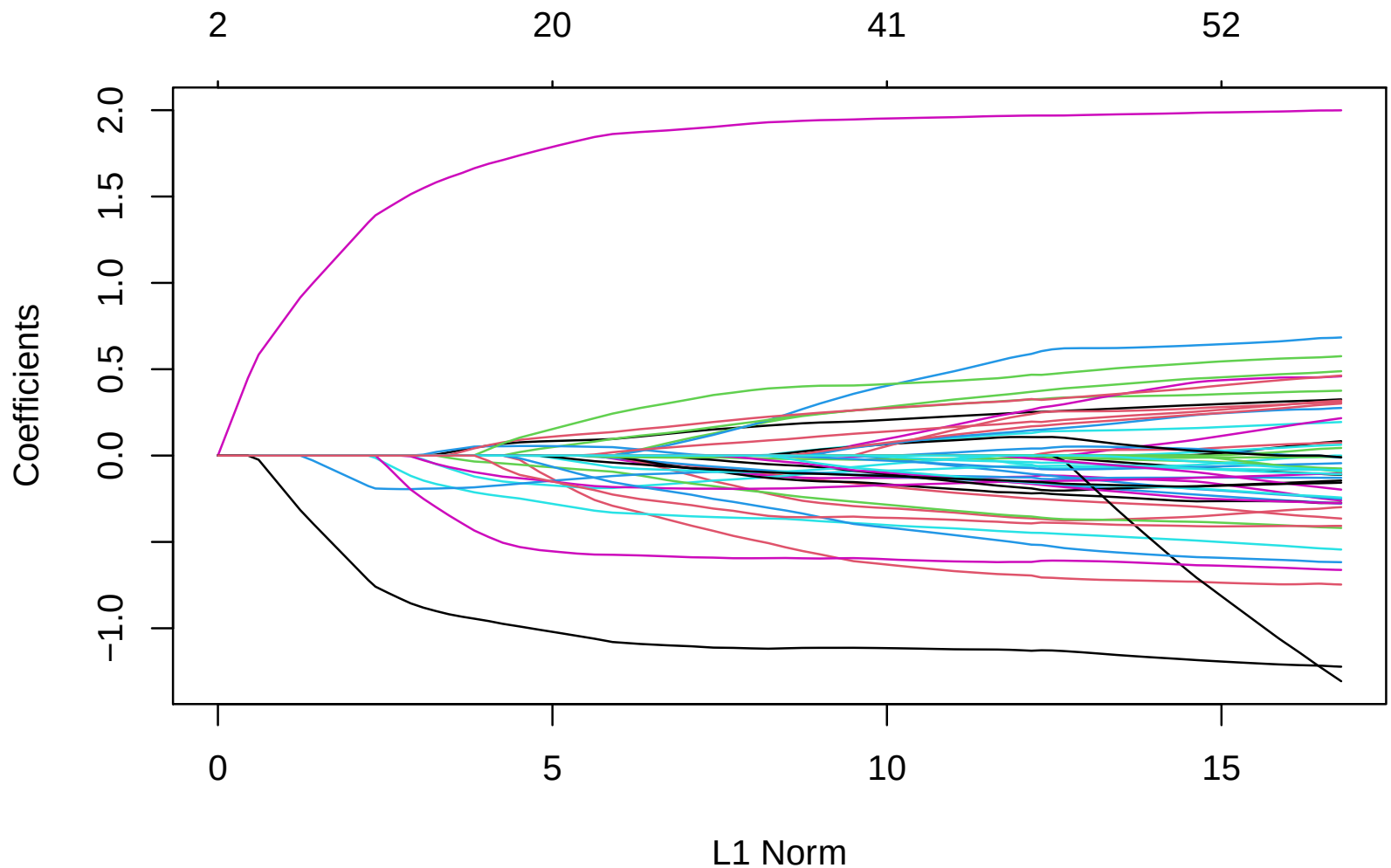
Table: Summary of model performance based on 200 GAW20 simulations. Five-fold cross-validation root-mean-square error was reported for each simulation replicate.

Results: Mouse crosses



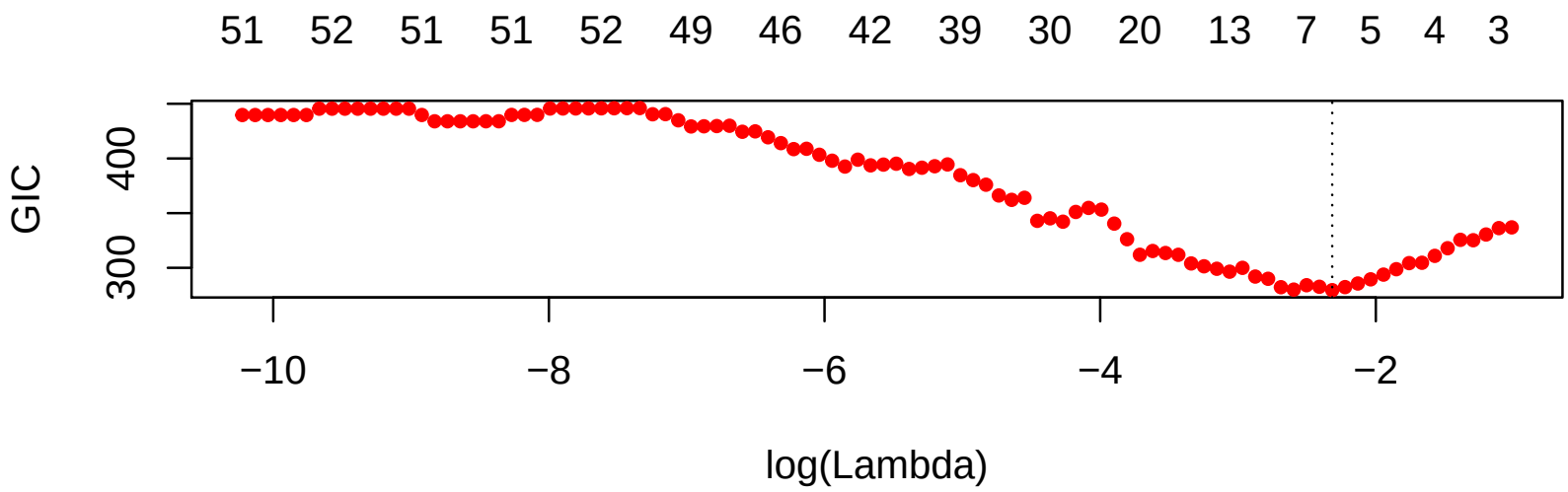
ggmix R package

```
library(ggmix)
data("admixed")
fit <- ggmix(x = admixed$xtrain,
             y = admixed$ytrain,
             kinship = admixed$kin_train)
plot(fit)
```



ggmix R package

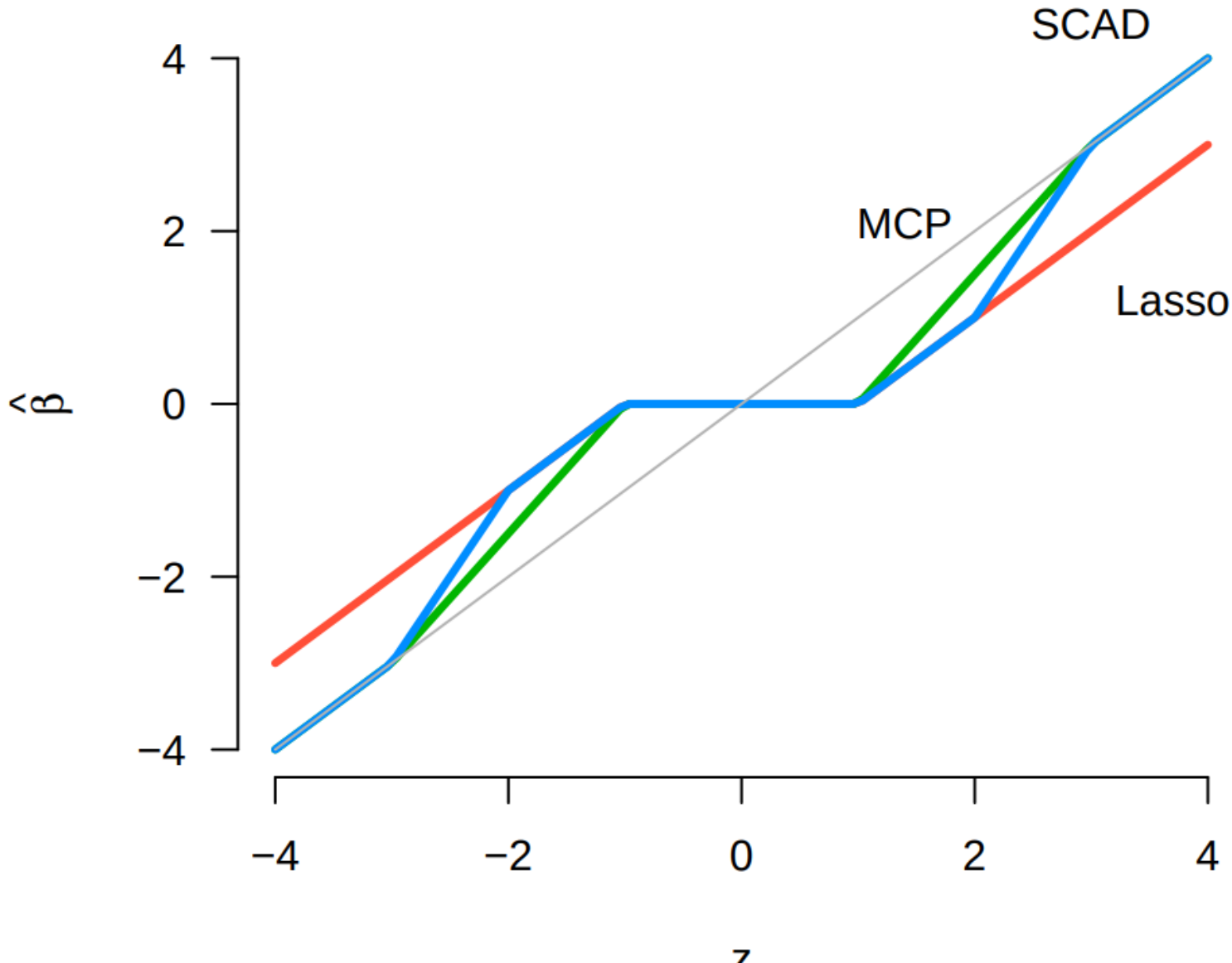
```
hdbic <- gic(fit)
plot(hdbic)
```




```
coef(hdbic, type = "nonzero")
```

```
##                       1
## (Intercept) -0.03598164
## X302        -0.17617815
## X524         1.34917874
## X538        -0.72073279
## eta          0.99000000
## sigma2       1.60477653
```

SCAD (Fan et Li, JASA, 2001), MCP (Zhang, Ann. Stat., 2010)



Computational challenges

- Past approaches for optimization for SCAD/MCP relies upon descent method, first- or second- order
- e.g., `sparsenet` (Mazumder et al. 2011) uses coordinate descent with full step size, whose coordinate update cycles through $\tilde{\beta}_j = S_{\gamma_k}\left(\sum_{i=1}^{n}\left(y_i - \tilde{y}_i^j\right)x_{ij}, \lambda_\ell\right)$, where $\tilde{y}_i^j = \sum_{k\neq j} x_{ik}\tilde{\beta}_k$
- However, coordinate descent is difficult to vectorize, and rate of convergence is difficult of establish – though past literature suggests $O(1/k)$ rate of convergence for ISTA

Our proposal: Accelerated gradient (AG) method

Improving Convergence for Nonconvex Composite Programming

Kai Yang · **Masoud Asgharian** · **Sahir Bhatnagar**

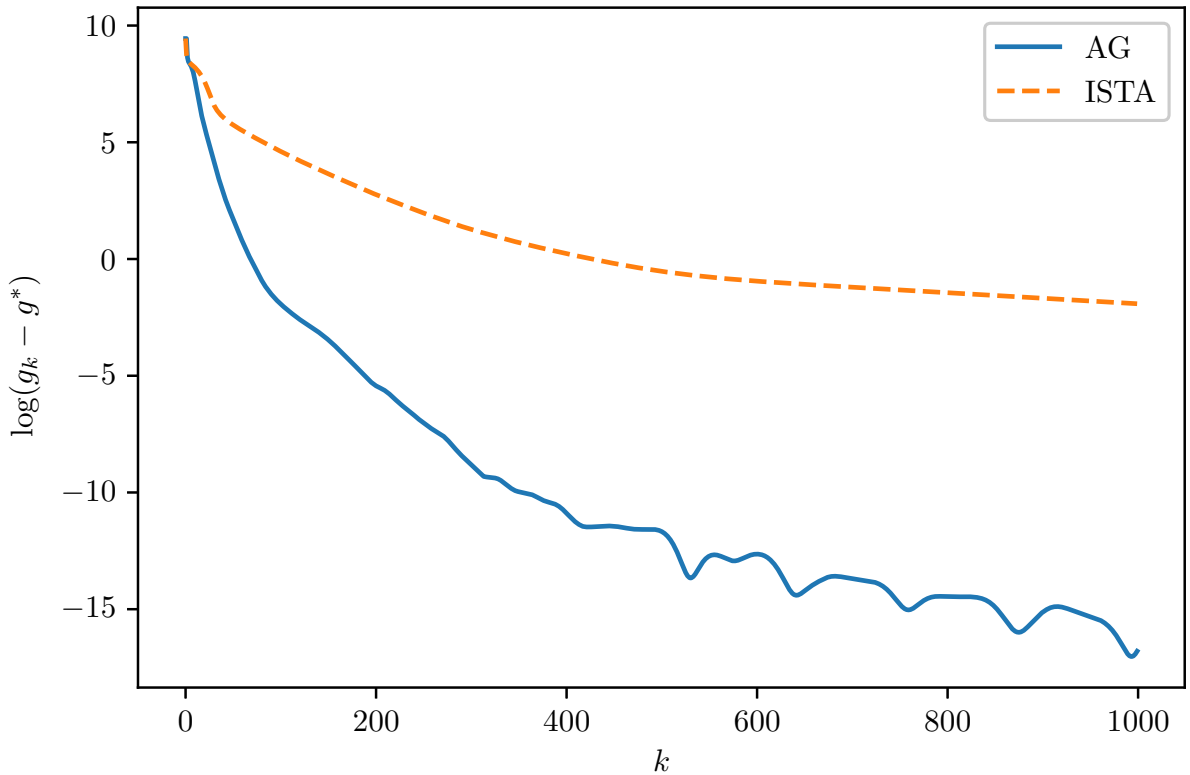
Received: date / Accepted: date

Abstract High-dimensional nonconvex composite problems are popular in today's machine learning and statistical genetics research. Recently, Ghadimi and Lan [1] proposed an algorithm to optimize nonconvex high-dimensional problems. There are several parameters in their algorithm that are to be set before running the algorithm. It is not trivial how to choose these parameters nor there is, to the best of our knowledge, an explicit rule how to select the parameters to make the algorithm converges faster. We analyze Ghadimi and Lan's algorithm to gain an interpretation based on the inequality constraints for convergence and the upper bound for the norm of the gradient analogue. Our interpretation of their algorithm suggests this to be a damped accelerated gradient scheme. Based on this, we propose an approach how to select the parameters to improve convergence of the algorithm. Our numerical studies using high-dimensional nonconvex sparse learning problems, motivated by image denoising and statistical genetics applications, show that convergence can be made, on average, considerably faster than that of the conventional ISTA algorithm for such optimization problems with over 10000 variables should the parameters be chosen using our proposed approach.

Keywords Accelerated Gradient · Composite Optimization · Nonconvex Optimization

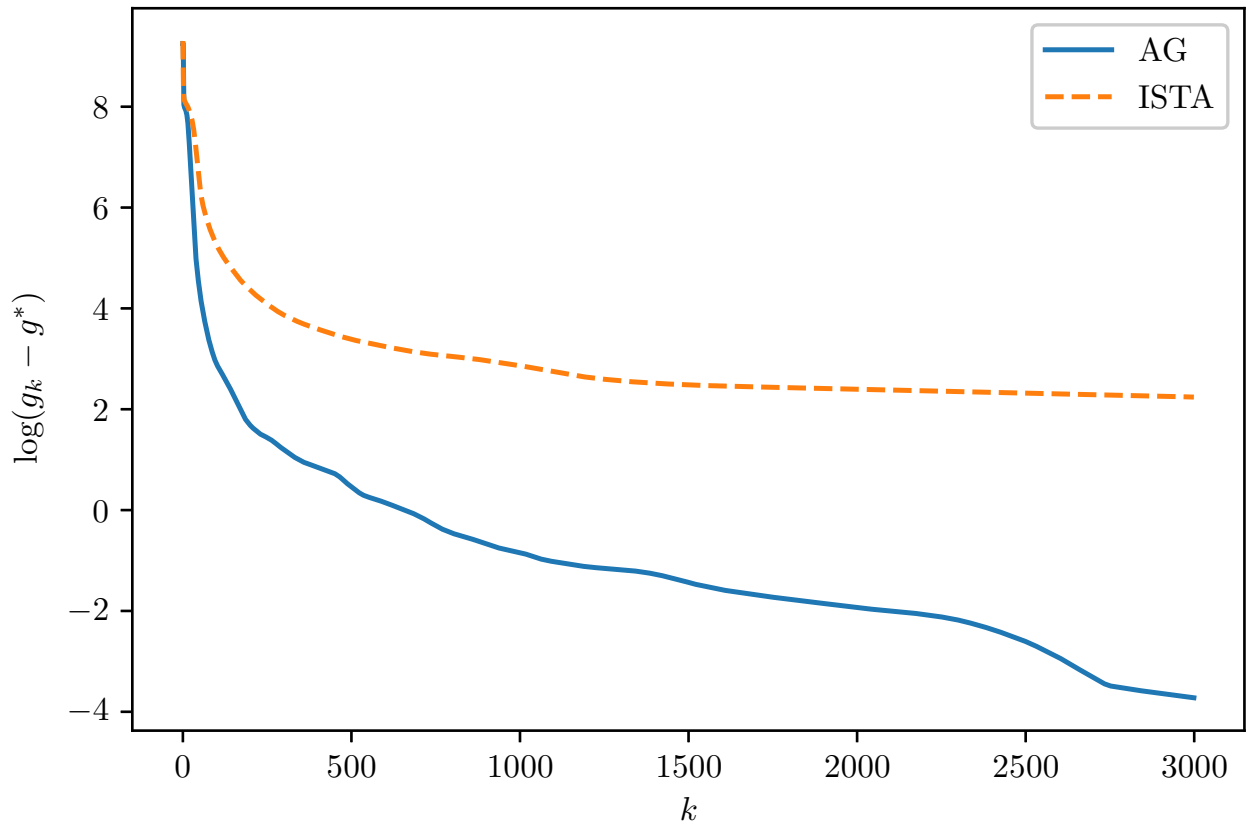
[1] https://arxiv.org/abs/2009.10629

Numerical Study for SCAD

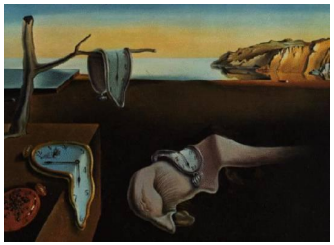


$\mathbf{x}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{I})$, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $\mathbf{y} = \mathbf{X}\boldsymbol{\tau}_{\text{generate}} + \boldsymbol{\varepsilon}$, $\sigma^2 = \frac{\left\|\boldsymbol{\tau}_{\text{generate}}\right\|^2}{3}$, $\boldsymbol{\tau}_{\text{generate}} \in \mathbb{R}^{10006}$ is a sparse constant vector with 6 values of $1.23(\text{intercept}), 3, 4, 5, 6, 59$ as true effect coefficients and 10000 values of 0. Start point: $\boldsymbol{\tau}_0 = \mathbf{1}_{10006}$, $a = 3.7$, $\lambda = 0.6$.

Numerical Study for MCP



Simulation settings here is same as before in SCAD, $\gamma = 2.5,\ \lambda = 0.6$.



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates

Sahir Rai Bhatnagar*
McGill University

Maxime Turgeon*
University of Manitoba

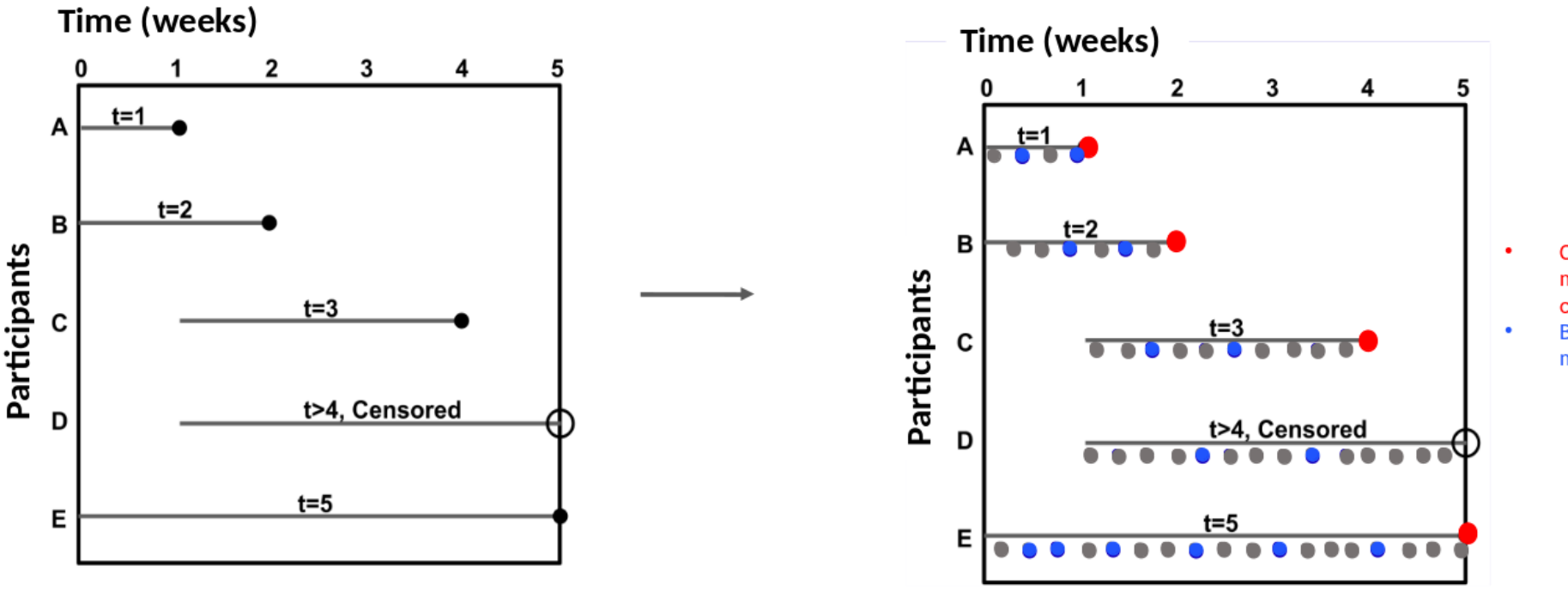
Jesse Islam
McGill University

James A. Hanley
McGill University

Olli Saarela
University of Toronto

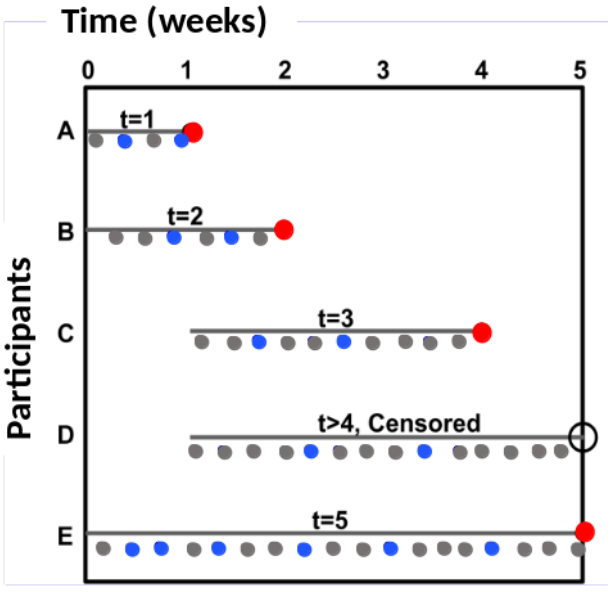
[1] https://arxiv.org/abs/2009.10264, https://cran.r-project.org/package=casebase

Case-base sampling – Toy example



- Case series: all the person-moments where an event occured
- Base series: sample of all moments

Case-base sampling with logistic regression



$$e^{\beta(x,t)} = \frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)}$$

...

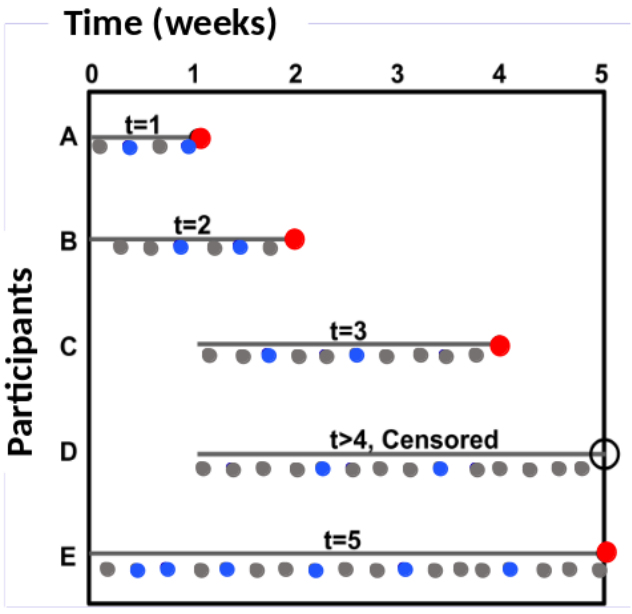
$$ln(h(\hat{x},t)) = \hat{\beta}(x,t) + ln\left(\frac{b}{B}\right)$$

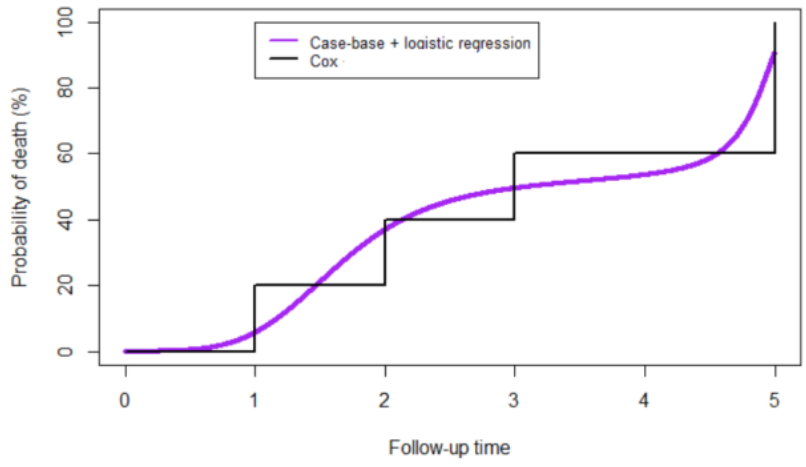
b = # Blue
B = # all moments

To have a flexible baseline hazard:

$$ln(\hat{h}(x,t)) = \hat{\beta}_{t_1}t + \hat{\beta}_{t_2}t^2 + \hat{\beta}_{t_3}t^3 + \hat{\beta}x + ln\left(\frac{b}{B}\right)$$

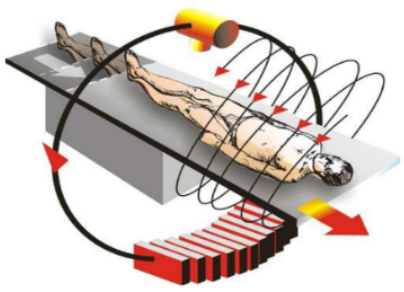
Smooth-in-time cumulative incidence curves

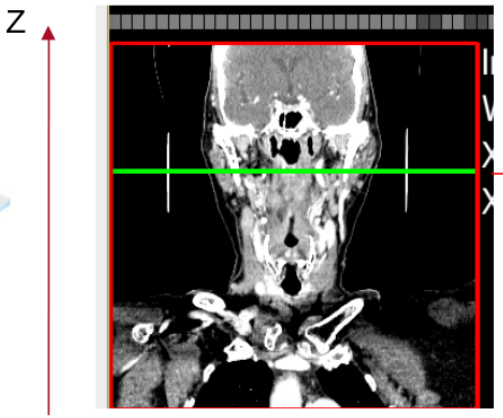




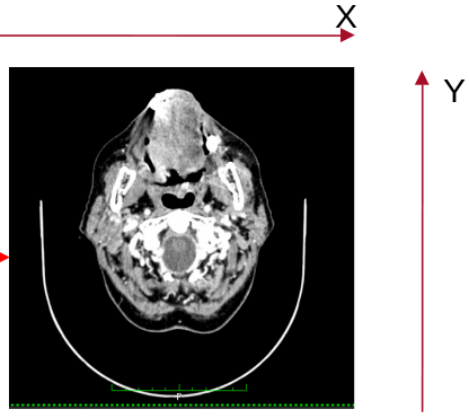
Single Energy CT – 3-D Tensor

Single Energy CT Scans





One patient

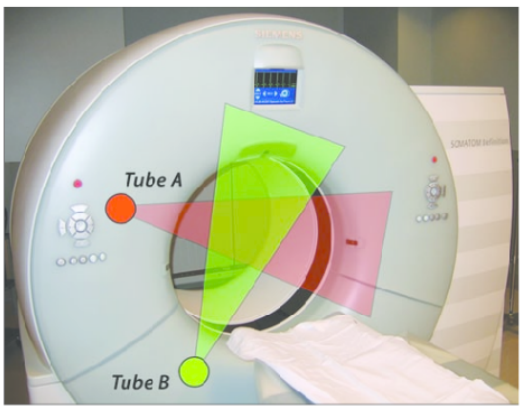


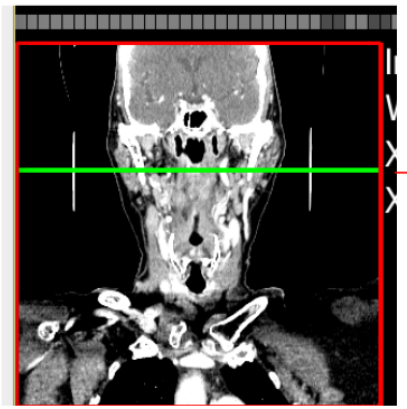
One Slice (fix Z value)
Grey area: low intensity, high in H and C: fat, tissues
White area: high intensity, high in calcium: bones
Black: air

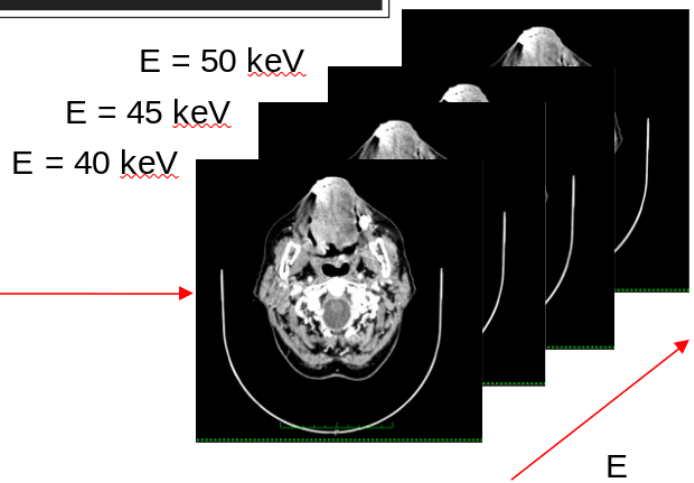
5

Dual Energy CT – 4-D Tensor

Dual Energy CT Scans







Two tubes incite x-rays at different energies: Xlow and Xhigh
DECT can reconstruct the signal received to scans as if they are taken at different incoming X-ray energy in Conventional CT scans

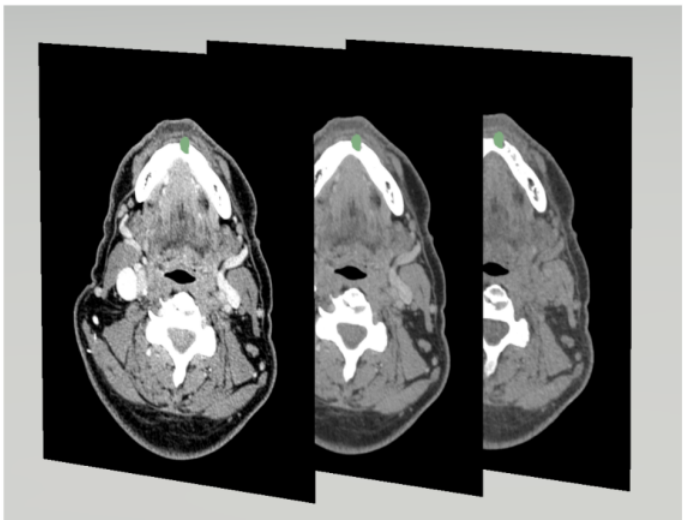
For a higher energy level, intensity typically decreases
Different materials have different decrease in intensity

8

HNSCC

Head and Neck Squamous Cell Carcinoma

DECT slice at 3 energy levels: 40/70/140kev
and ground truth tumor (segmented in green)



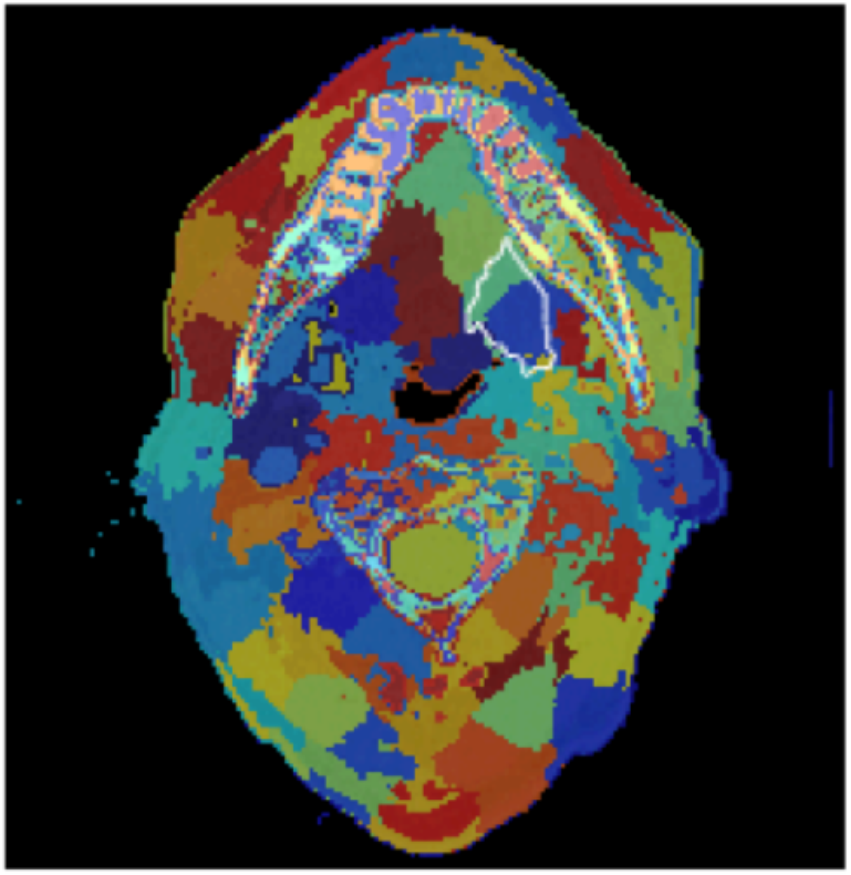
Segolene Brivet | Tumor extraction on DECT | Feb 24, 2021 | 2 / 18

One Approach – Clustering

Goal:
Cut image in different areas that fit anatomical regions

Focus on tumor regions.

Method: Statistical approach to differentiate energy decay curves of voxels



Acknowledgements

CRSNG RGPIN-2020-05133

- Kai Yang: Non-convex optimization
- Jesse Islam: High-dimensional survival analysis

Kai Yang, PhD (c)

Jesse Islam, PhD (c)

NSERC CRSNG

Acknowledgements

MiCM

- Julien St-Pierre: LMM with multiple random effects, longitudinal data, combining multiple cohorts

Julien St-Pierre, PhD (c)

McGill initiative in Computational Medicine

Acknowledgements

CIHR Project Grant, CANSSI CRT

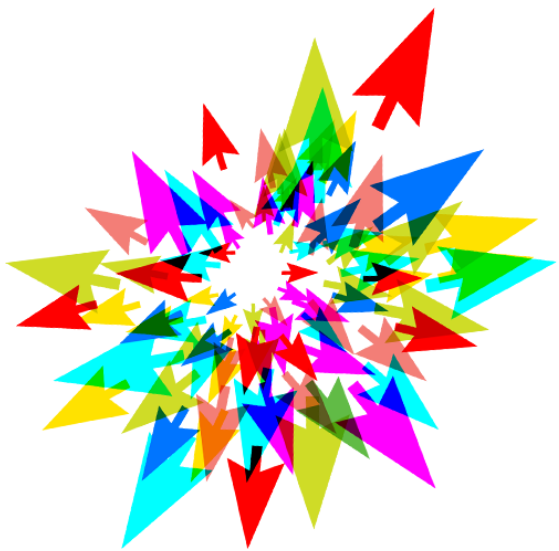
- Zeyu Bian: Low-rank approximations, memory mapping
- Mohan Zhao: Multivariate outcomes and matrix covariates

Zeyu Bian, PhD (c)

Mohan Zhao, BSc (c)

Acknowledgements

- Masoud Asgharian (McGill)
- Tianyuan Lu (McGill)
- Yi Yang (McGill)
- Karim Oualkacha (UQÀM)
- Celia Greenwood (Lady Davis Institute)
- Erica Moodie (McGill)
- James Hanley (McGill)
- Maxime Turgeon (UManitoba)
- Olli Saarela (UofT)
- Luda Diatchenko (McGill)
- UK Biobank Resource under project number 27449. We appreciate the generosity of UK Biobank volunteers



References

1. **Yang K**, Asgharian M, Bhatnagar SR (2020+). Improving Rate of Convergence for Nonconvex Composite Programming. *Submitted to Optimization Letters*. https://arxiv.org/abs/2009.10629.
2. Bhatnagar SR, Turgeon M, **Islam J**, Hanley JA, Saarela O (2020+). casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates. *Submitted to Journal of Statistical Software*. https://arxiv.org/abs/2009.10264.
3. Bhatnagar SR, Yang Y, Lu T, Schurr E, Loredo-Osti JC, Forest M, Oualkacha K, Greenwood CMT (2020). Simultaneous SNP selection and adjustment for population structure in high dimensional prediction models. *PLoS Genetics* 16(5): e1008766. DOI 10.1371/journal.pgen.1008766.

sahirbhatnagar.com

Session Info

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 20.10

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.10.so

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggmix_0.0.1 knitr_1.31

loaded via a namespace (and not attached):
 [1] lattice_0.20-41  codetools_0.2-16 glmnet_4.1-1     foreach_1.5.1
 [5] grid_4.0.2       magrittr_2.0.1   evaluate_0.14    highr_0.8
 [9] stringi_1.5.3    Matrix_1.2-18    splines_4.0.2    iterators_1.0.13
[13] tools_4.0.2      stringr_1.4.0    survival_3.2-3   xfun_0.22
[17] compiler_4.0.2   shape_1.4.5
```