Variable Selection and Prediction for High-Dimensional Genetic Data with Complex Structures

Sahir Bhatnagar, PhD

Assistant Professor
Department of Epidemiology, Biostatistics and Occupational Health
Department of Diagnostic Radiology

December 8, 2021

https://sahirbhatnagar.com

Outline

High Dimensional (HD) Data Analysis

Classical

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

McGill Summer School in Health Data Analytics. https://sahirbhatnagar.com/assets/pdf/mcgillHDA_2021.pdf

High Dimensional (HD) Data Analysis

Classical

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

HD data

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

McGill Summer School in Health Data Analytics. https://sahirbhatnagar.com/assets/pdf/mcgillHDA_2021.pdf

New challenges arise from how such data is *used*

A

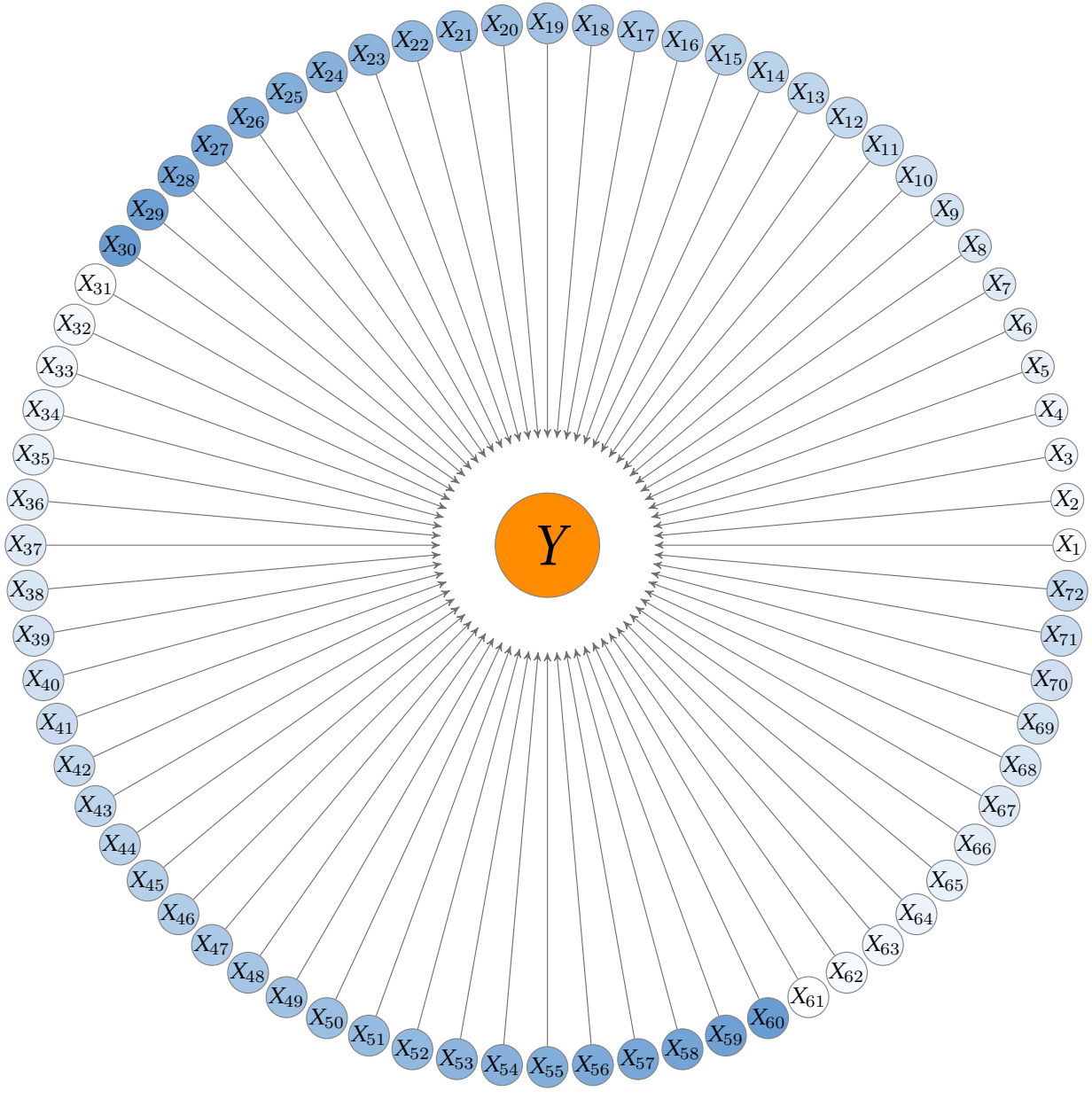
y	x_1
0.0	0
2.1	1
2.7	0
5.9	3
7.3	3
0.0	0
2.0	1

B

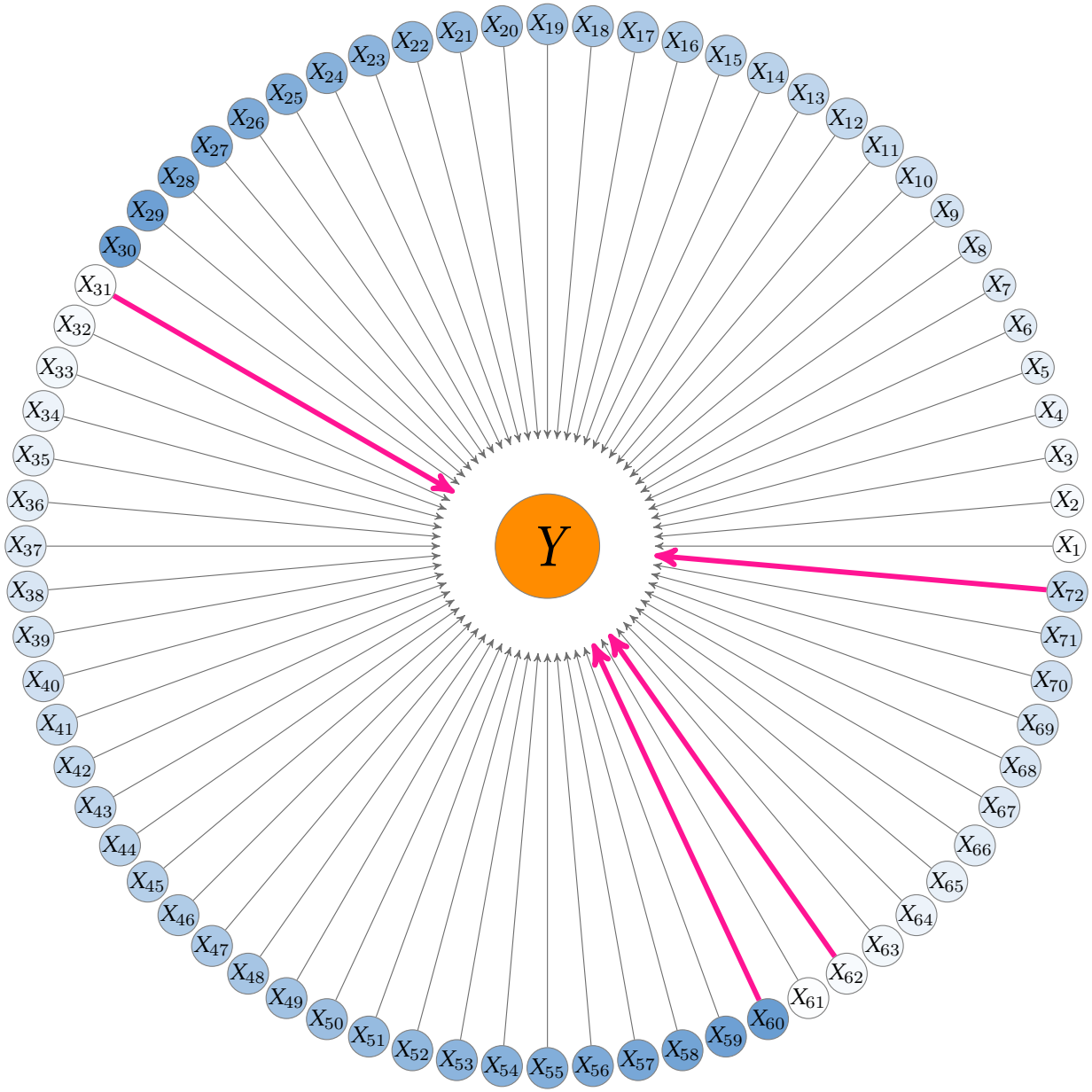
y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
0	0	2	0	0	1	0	1	0
2.1	1	0	2	3	2	0	0	3
2.7	0	0	0	2	2	1	1	1
5.9	3	0	1	0	0	0	2	0
7.3	3	4	0	1	1	1	0	0
0.0	0	2	0	0	3	0	0	0
2.0	1	0	2	1	0	0	0	1

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

Bet on Sparsity Principle



Bet on Sparsity Principle



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:

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

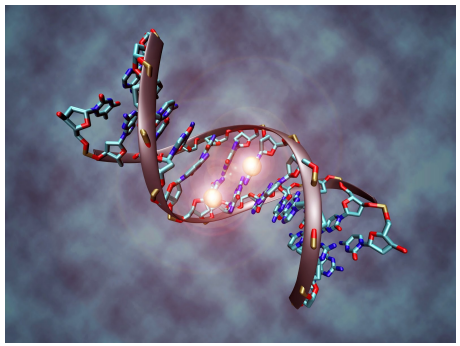
$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \qquad \text{(Lasso regression)}$$

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


High-dimensional data:
Genomics
Radiomics
Neuroimaging
Penalized Regression
Accelerated gradient for nonconvex penalties JCGS (2021+)
GLMM
Tuning free parameter selection
Error-in-variables Gen Epi (2021)
LMM PloS Genetics (2020)
Interaction Selection
Environment-based clustering Gen Epi (2018)
Sparse additive model CSDA (2021+)
Variable selection in Dynamic Treatment Regimes Biometrics (2021)
Regularized estimating equations
Tree based penalties
Imaging Analysis
Mean curvature of isophotes Radiology: Artificial Intelligence (2021)
Sparse Bayesian model Stat (2021+)
Sparse tensor regression
Dual energy CT
Smooth Hazards
casebase R Journal (2021+)
Person-moment neural networks Stat Med (2021+)
Polygenic hazard scores
Penalized competing risks

Gestational diabetes, DNA methylation and obesity

$\sim$



$\times$

Phenotype
Obesity measures

Large Data
Child's epigenome
$(p \approx 450\text{k})$

Environment
Gestational Diabetes

Differential Correlation between environments

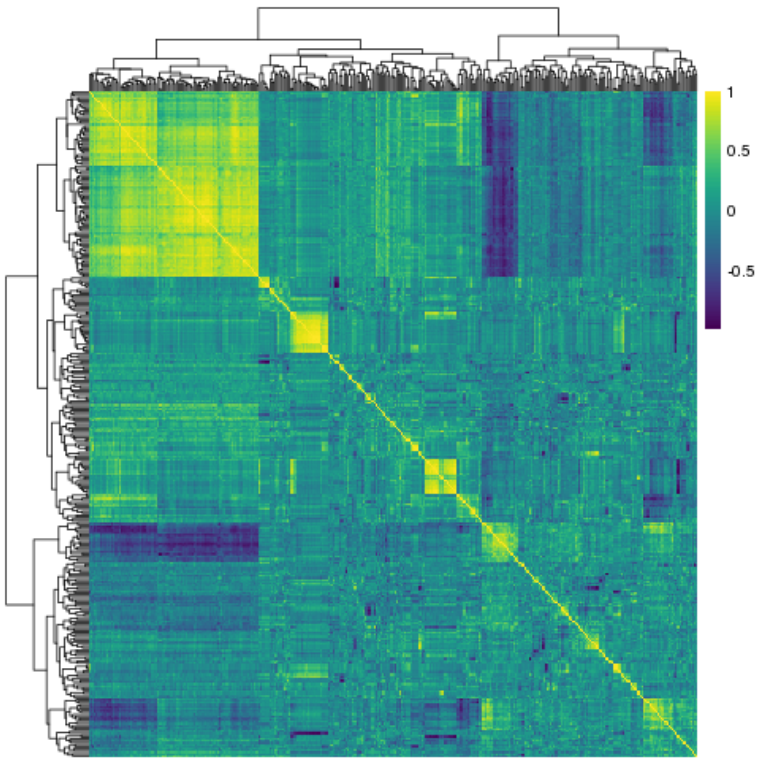


Fig.: Gestational diabetes

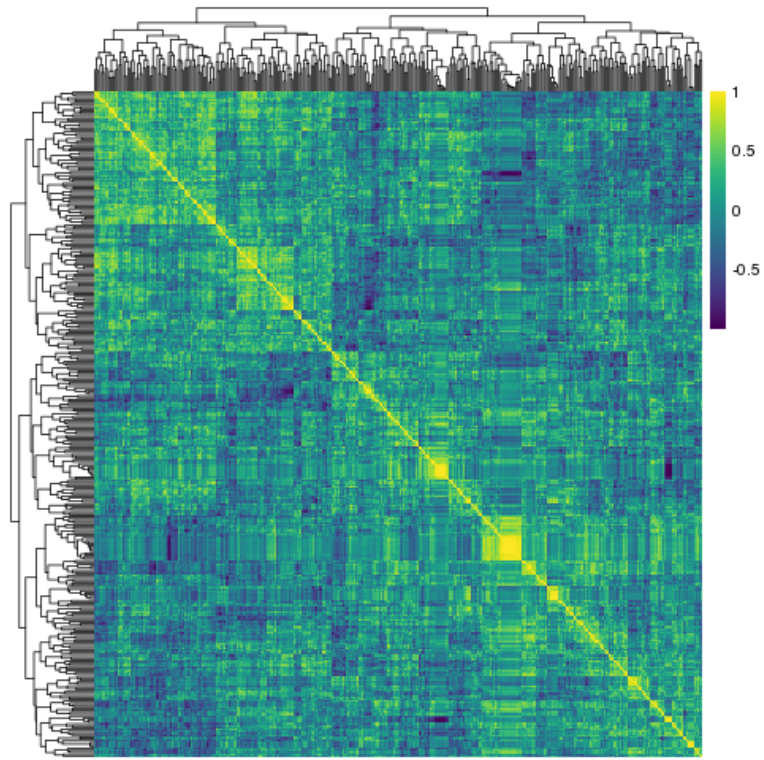
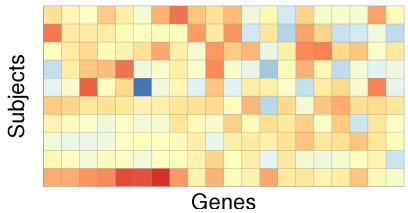


Fig.: Controls

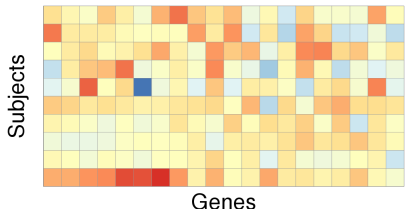
`eclust`: our proposed 2 step method

Original Data



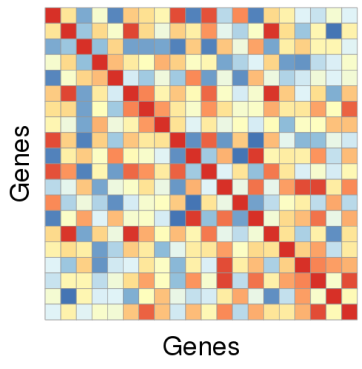
`eclust`: our proposed 2 step method

Original Data

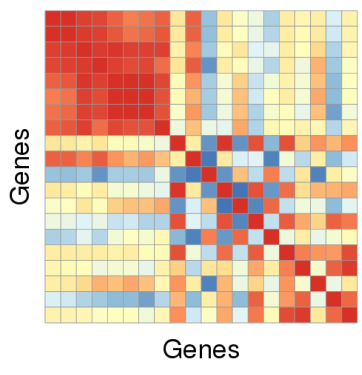


1a) Gene Similarity

$E = 0$

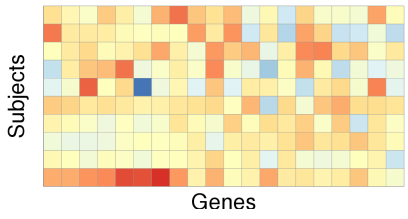


$E = 1$



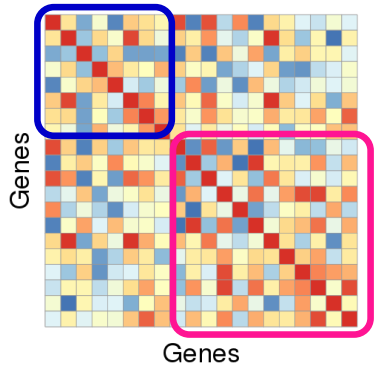
`eclust`: our proposed 2 step method

Original Data

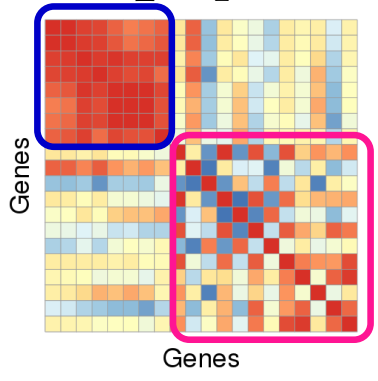


1a) Gene Similarity

$E = 0$

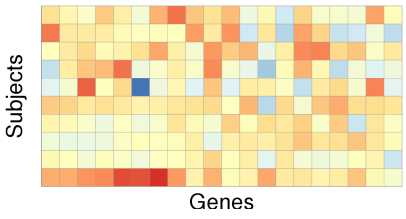


$E = 1$



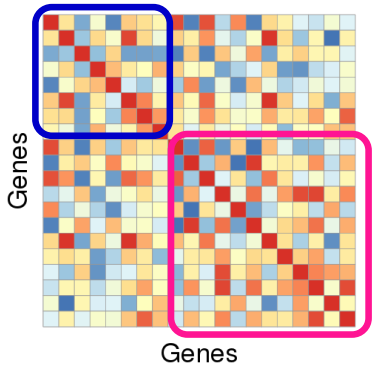
`eclust`: our proposed 2 step method

Original Data

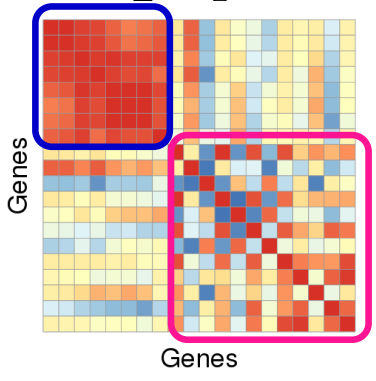


1a) Gene Similarity

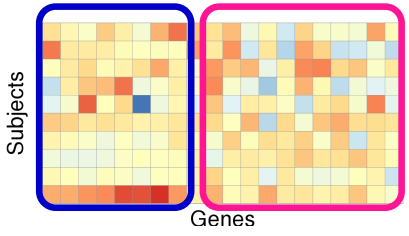
$E = 0$



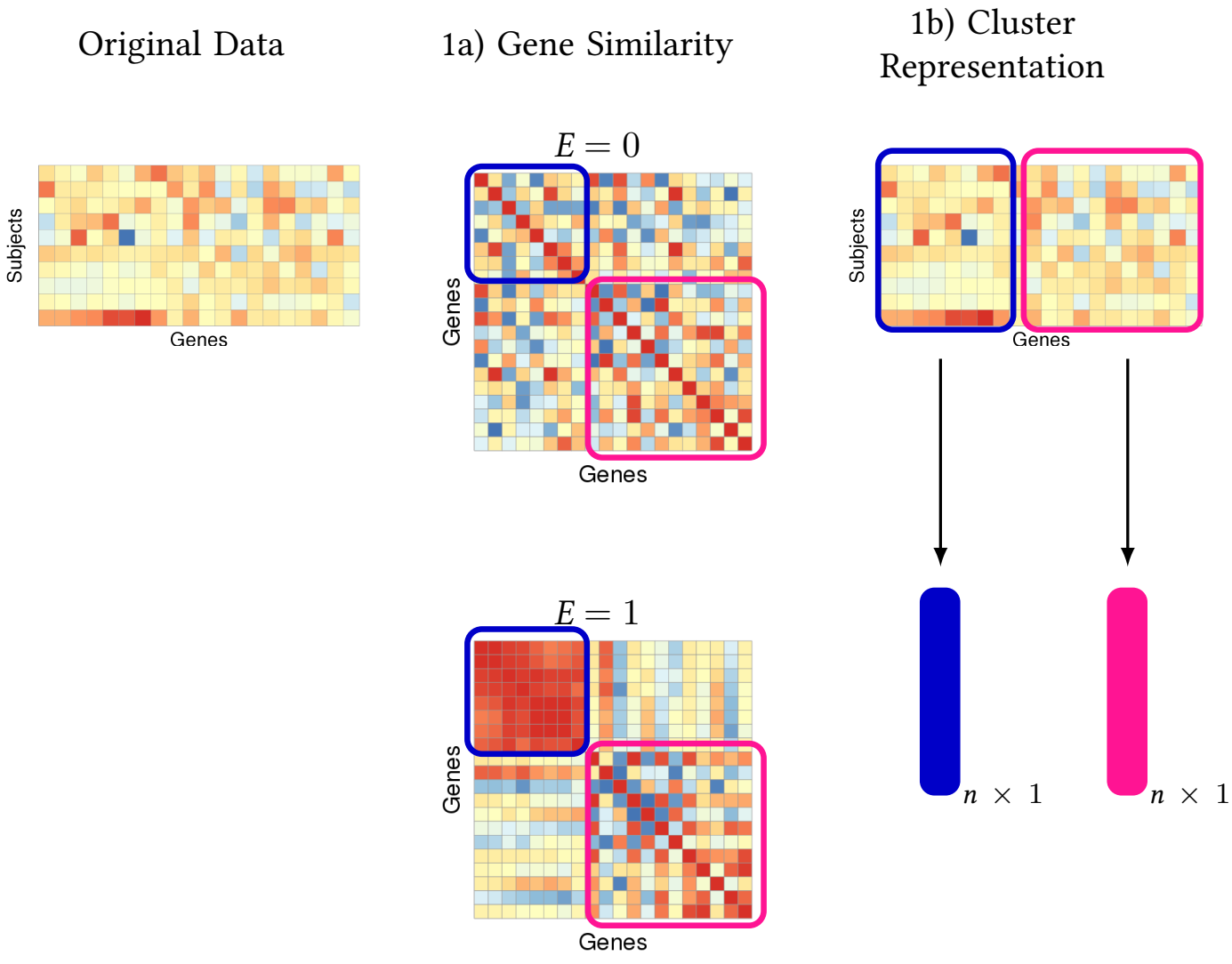
$E = 1$



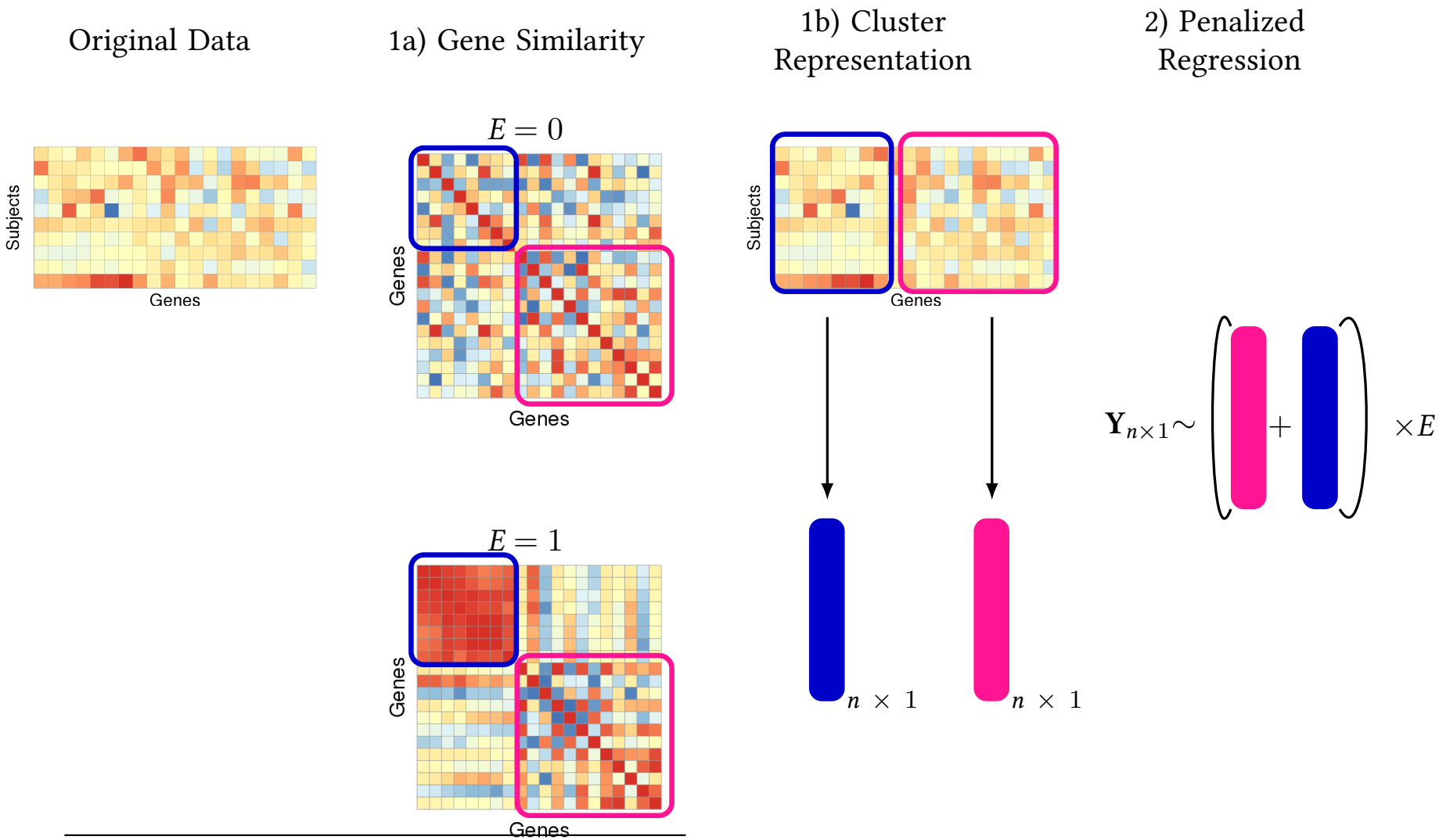
1b) Cluster Representation



`eclust`: our proposed 2 step method



`eclust`: our proposed 2 step method



Bhatnagar et al. An analytic approach for interpretable predictive models in high dimensional data, in the presence of interactions with exposures. *Genetic Epidemiology (2018)*. https://cran.r-project.org/package=eclust

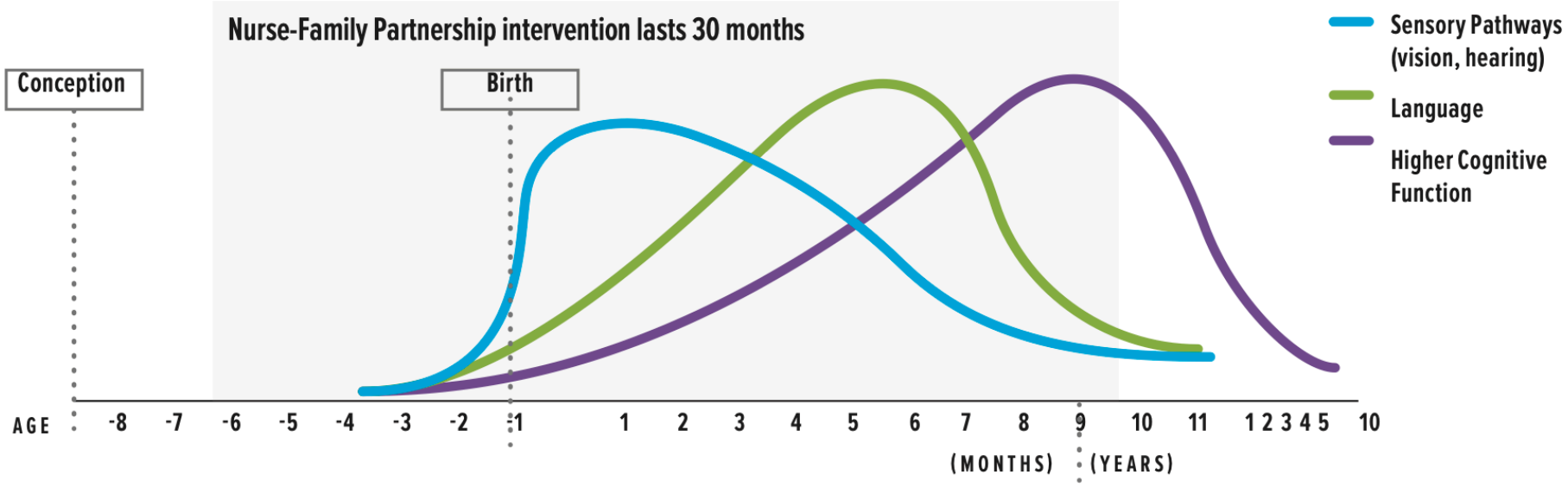
Nurse-Family Partnership

Helping First-Time Parents Succeed®

Nurse-Family Partnership is an evidence-based, community health program with over 40 years of evidence showing significant improvements in the health and lives of first-time moms and their children living in poverty.

Human Brain Development

Synapse formation dependent on early experiences



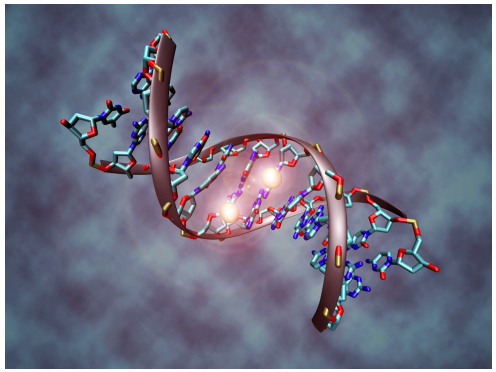
Source: Nelson, C.A., In Neurons to Neighborhoods (2000).

Interactions between Intervention and Genetics

Stanford–Binet Fifth Edition (SB5) classification[4]

IQ Range ("deviation IQ")	IQ Classification
145–160	Very gifted or highly advanced
130–144	Gifted or very advanced
120–129	Superior
110–119	High average
90–109	Average
80–89	Low average
70–79	Borderline impaired or delayed
55–69	Mildly impaired or delayed
40–54	Moderately impaired or delayed

$\sim$



$\times$

Phenotype
IQ Score

Large Data
Genetic Markers

Environment
NFP Intervention

$$Y = \sum_{j=1}^{p} X_j \beta_j + \sum_{j=1}^{p} X_j X_E \tau_j + \varepsilon$$

$$= \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + X_{E\,n \times 1} \circ \mathbf{X}_{n \times p} \boldsymbol{\tau}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

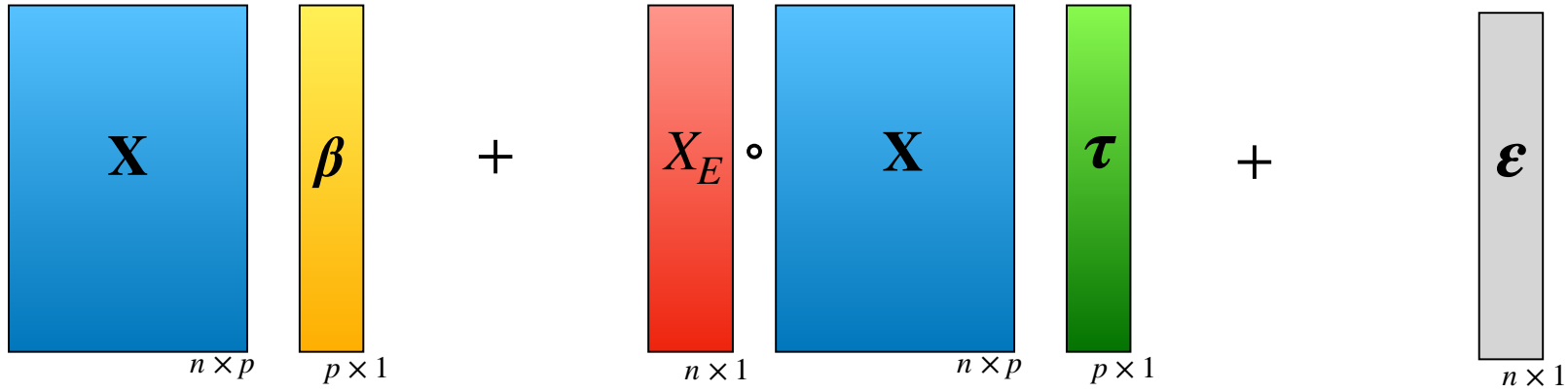
Main effects

Interaction effects

Error

$$Y = \sum_{j=1}^{p} X_j \beta_j + \sum_{j=1}^{p} X_j X_E \tau_j + \varepsilon$$

$$= \mathbf{X}_{n \times p} \boldsymbol{\beta}_{p \times 1} + X_{E\,n \times 1} \circ \mathbf{X}_{n \times p} \boldsymbol{\tau}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$



Main effects

Interaction effects

Error

Let $Z_{jE} = X_E X_j$

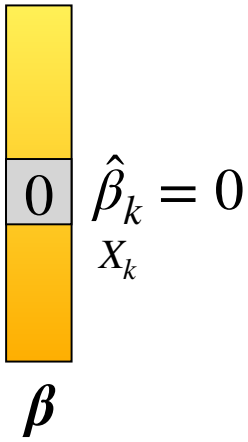
$$Y = \sum_{j=1}^{p} X_j \beta_j + \sum_{j=1}^{p} Z_{jE} \tau_j + \varepsilon$$

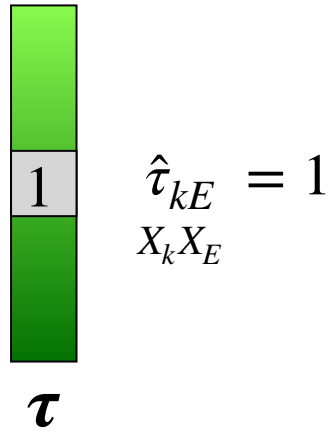
$$= \mathbf{X} \quad \mathbf{Z} \quad \boldsymbol{\beta} \quad \boldsymbol{\tau} + \boldsymbol{\varepsilon}$$

$n \times 2p$

$2p \times 1$

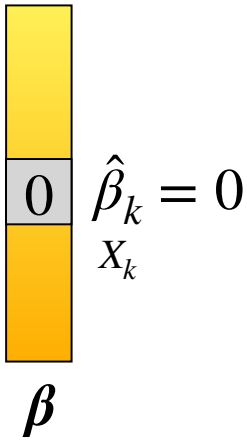
$n \times 1$

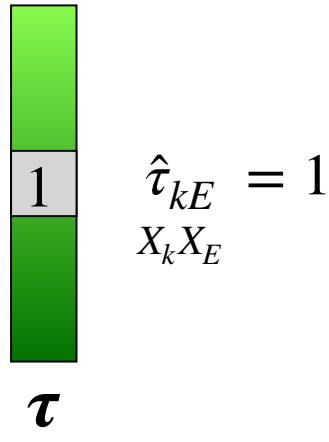




Main effects

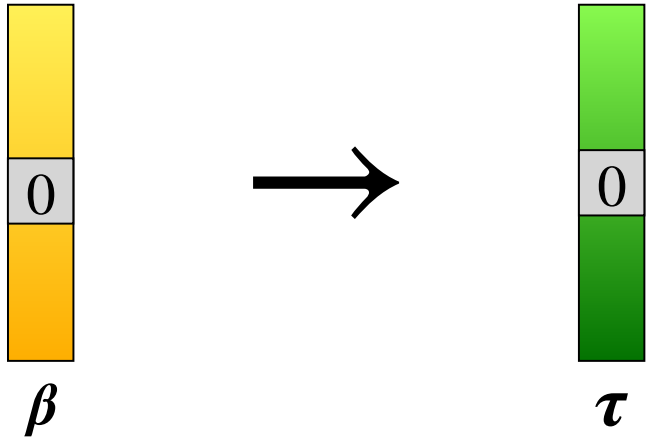
Interaction effects





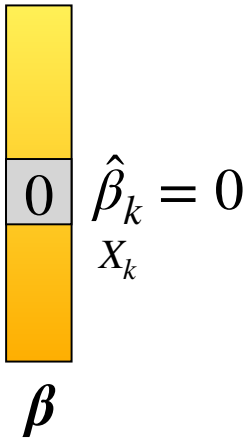
Main effects

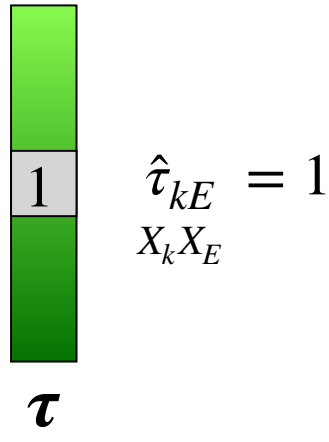
Interaction effects



Main effects

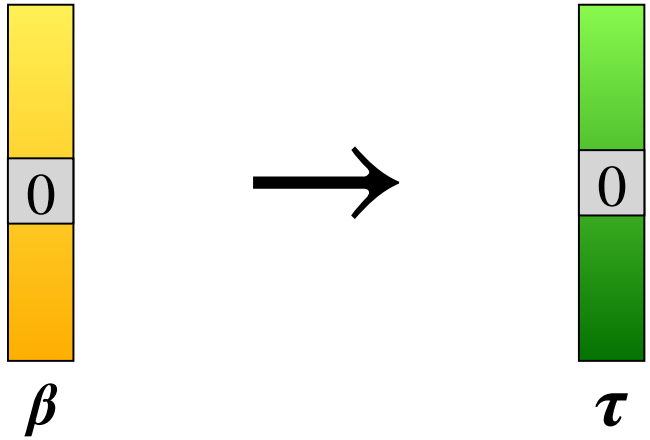
Interaction effects





Main effects

Interaction effects



Main effects

Interaction effects

Our Extension to Nonlinear Effects

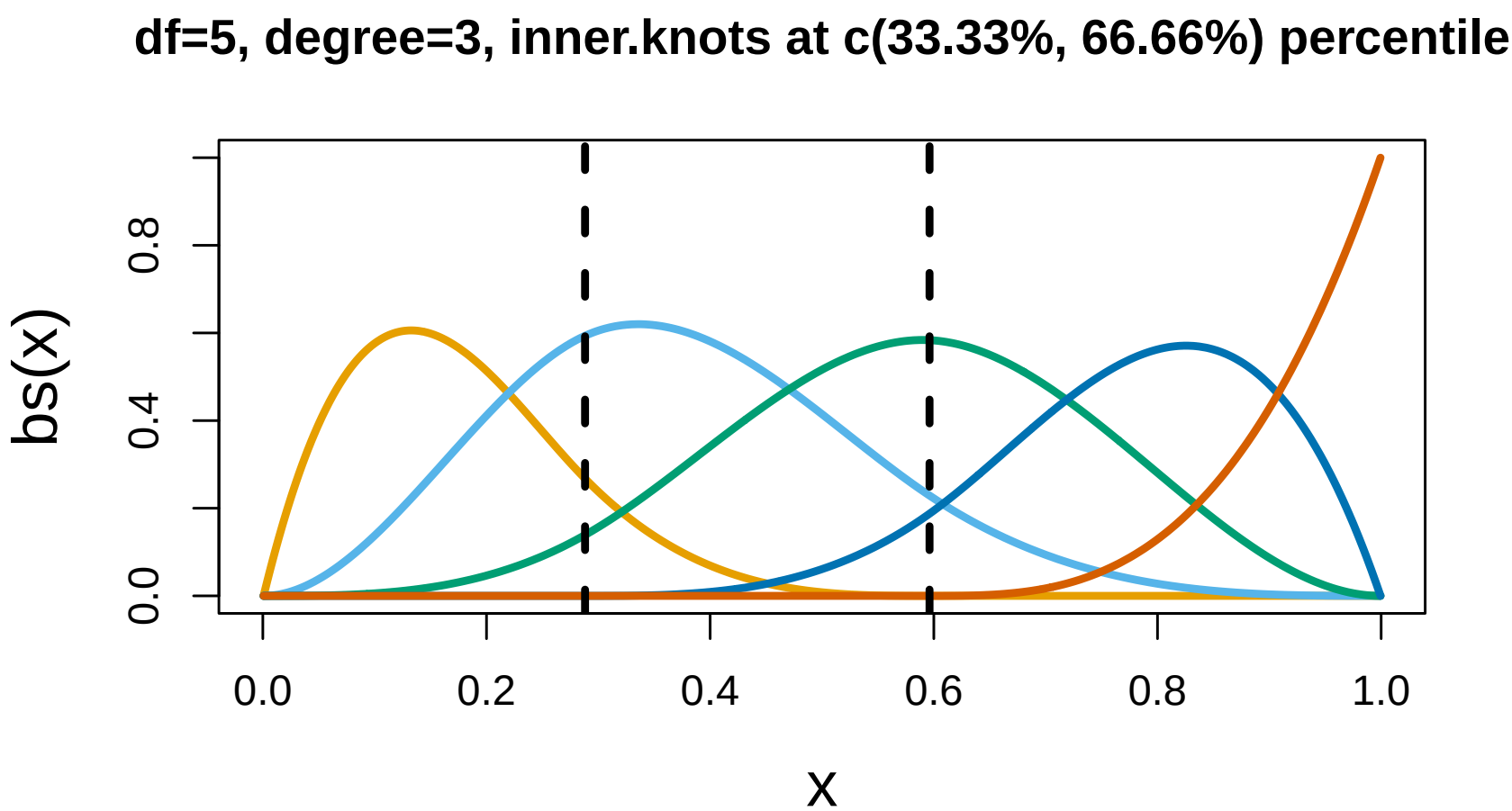
Consider the basis expansion

$$f_j(X_j) = \sum_{\ell=1}^{m_j} \psi_{j\ell}(X_j)\beta_{j\ell}$$

$$f(X_1) = \underbrace{\begin{bmatrix} \psi_{11}(X_{11}) & \psi_{12}(X_{12}) & \cdots & \psi_{11}(X_{15}) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{11}(X_{i1}) & \psi_{12}(X_{i2}) & \cdots & \psi_{11}(X_{i5}) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{11}(X_{N1}) & \psi_{12}(X_{N2}) & \cdots & \psi_{11}(X_{N5}) \end{bmatrix}_{N\times 5}}_{\mathbf{\Psi_1}} \times \underbrace{\begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{14} \\ \beta_{15} \end{bmatrix}_{5\times 1}}_{\theta_1}$$

B-Spline Expansion

```
x <- truncnorm::rtruncnorm(1000, a = 0, b = 1)
B <- splines::bs(x, df = 5, degree=3, intercept = FALSE)
```



`sail`: Additive Interactions

- $\boldsymbol{\theta}_j = (\beta_{j1}, \ldots, \beta_{jm_j}) \in \mathbb{R}^{m_j}$
- $\boldsymbol{\tau}_j = (\tau_{j1}, \ldots, \tau_{jm_j}) \in \mathbb{R}^{m_j}$
- $\boldsymbol{\Psi}_j \to n \times m_j$ matrix of evaluations of the $\psi_{j\ell}$
- In our implementation, we use cubic `bsplines` with 5 degrees of freedom

Model

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\tau}_j + \varepsilon$$

Bhatnagar et al. *In revision at Computational Statistics and Data Analysis* (2021+)

`sail`: Strong Heredity

Reparametrization

$$\boldsymbol{\tau}_j = \gamma_j \beta_E \boldsymbol{\theta}_j$$

Model

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \boldsymbol{\theta}_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j \beta_E (X_E \circ \boldsymbol{\Psi}_j) \boldsymbol{\theta}_j + \varepsilon$$

Objective Function

$$\underset{\boldsymbol{\Theta} := (\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma})}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(1-\alpha) \left(w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\boldsymbol{\theta}_j\|_2 \right) + \lambda\alpha \sum_{j=1}^{p} w_{jE} |\gamma_j|$$

[1] https://cran.r-project.org/package=sail

Nurse Family Partnership Program

- The Stanford Binet IQ scores at 4 years of age were collected for 189 subjects born to women randomly assigned to control (n = 100) or nurse-visited intervention groups (n = 89).
- For each subject, we calculated a polygenic risk score (PRS) for educational attainment at different p-value thresholds using weights from a previous GWAS.

Nurse Family Partnership Program

- The Stanford Binet IQ scores at 4 years of age were collected for 189 subjects born to women randomly assigned to control (n = 100) or nurse-visited intervention groups (n = 89).
- For each subject, we calculated a polygenic risk score (PRS) for educational attainment at different p-value thresholds using weights from a previous GWAS.
- In this context, individuals with a higher PRS have a propensity for higher educational attainment.
- The goal of this analysis was to determine if there was an interaction between genetic predisposition to educational attainment (X) and maternal participation in the NFP program (E) on child IQ at 4 years of age (Y).

Application of `sail` to NFP data

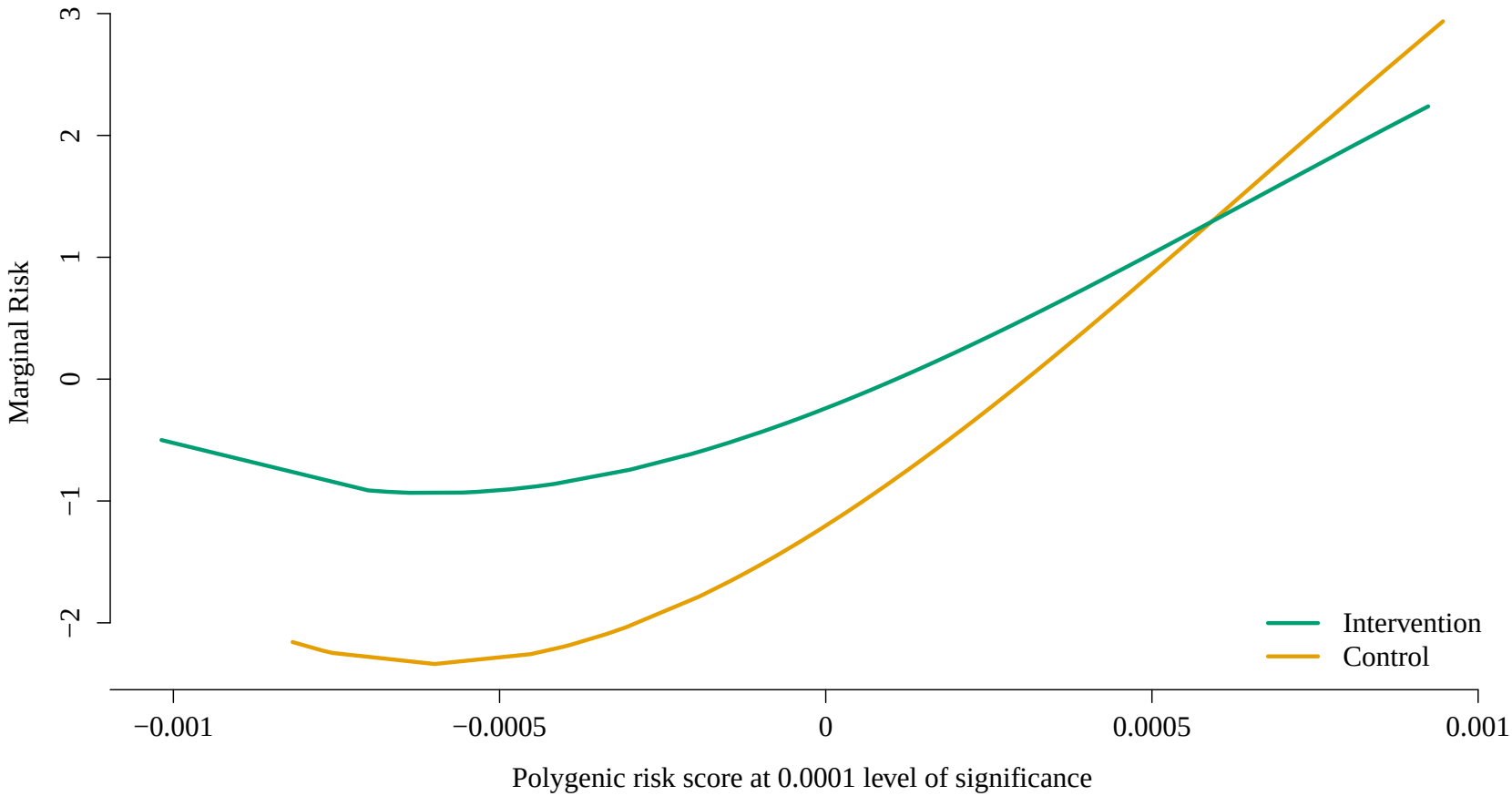
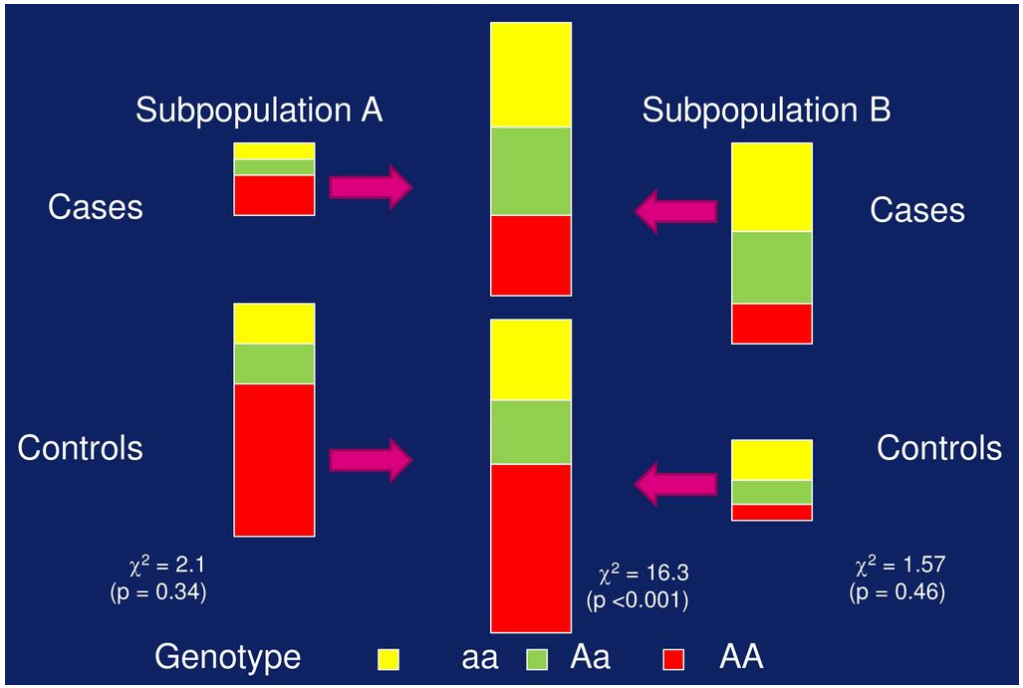


Fig.: The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational attainment using genetic variants significant at the 0.0001 level, as well as their interaction.

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)

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

$X\_kinship$

	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

$\downarrow$ $X\_kinship\, X\_kinship^T$

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

Current solution: Two Stage 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

$+ E_2$

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

[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:

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

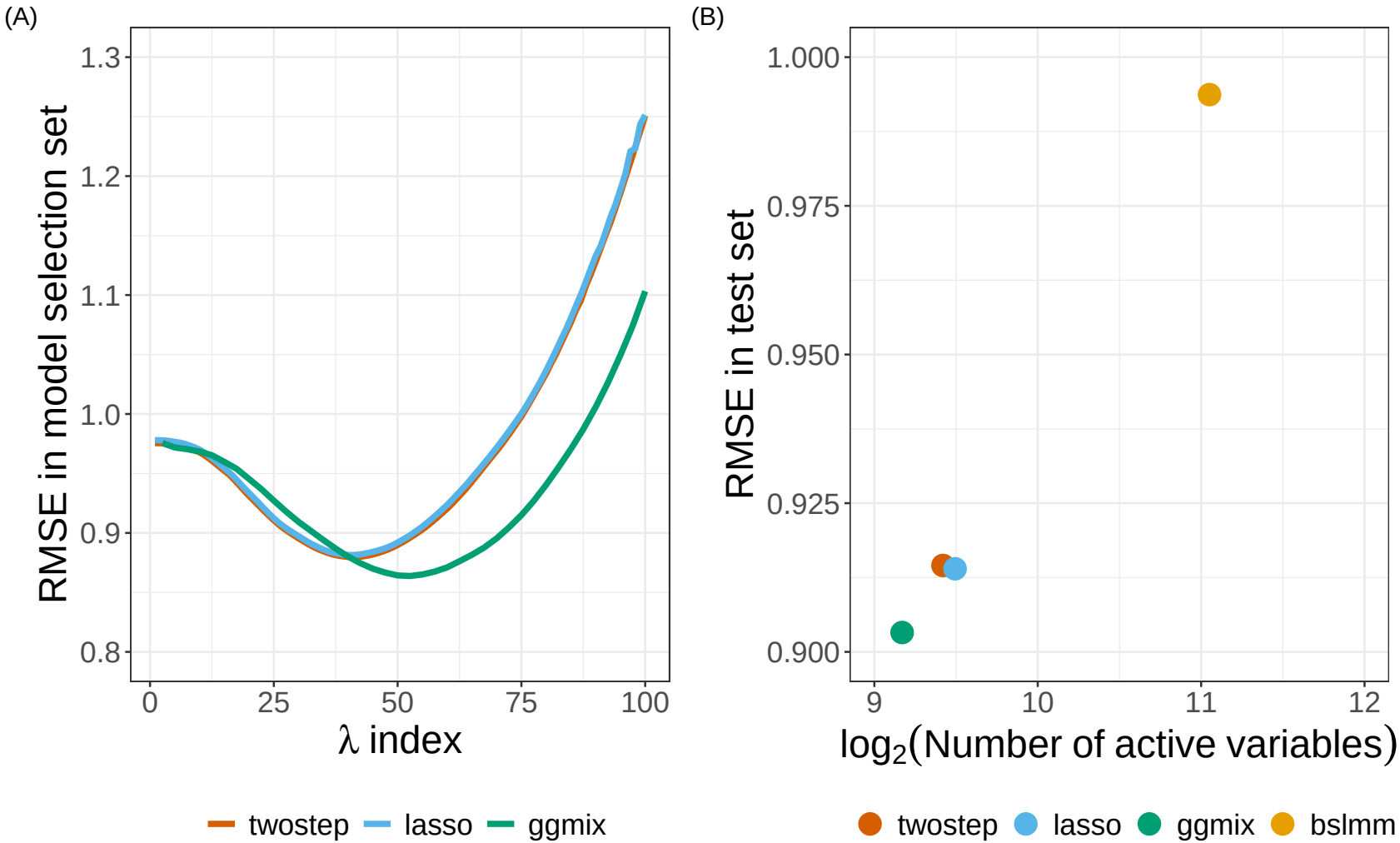
- $p_j(\cdot)$ is a penalty term on $\beta_1, \ldots, \beta_p$
- An estimate of the model parameters $\widehat{\boldsymbol{\Theta}}_\lambda$ is obtained by

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

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



Neural network survival analysis

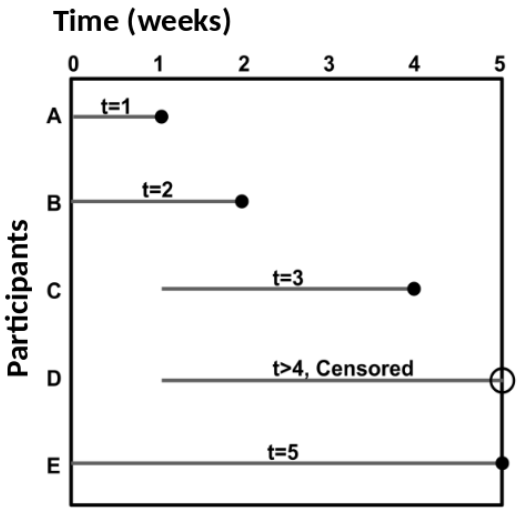
- DeepSurv – Cox neural networks.
  - Cox regression extended using neural networks.
  - Only uses proportional hazards (PH).
- DeepHit – First Hitting Time neural networks.
  - Inverse Gaussian distribution used as baseline hazard.
  - Does not let model determine baseline hazard.
- DeepSurvivalMachines (DSM) – Mixture model used for baseline hazard.
  - User specifies a set of distributions to be used as the baseline hazard.
  - Does not permit time-varying interactions.

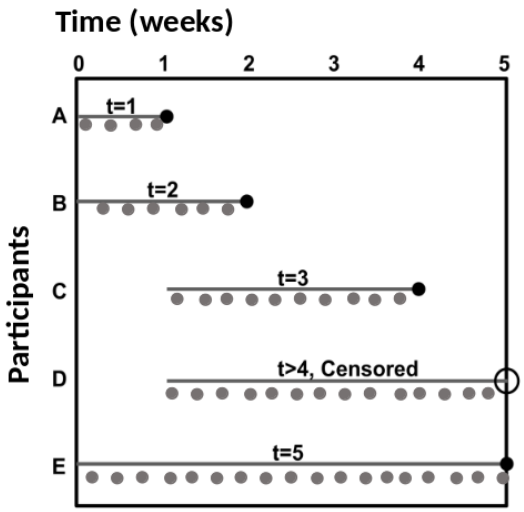
Neural network survival analysis

- DeepSurv – Cox neural networks.
  - Cox regression extended using neural networks.
  - Only uses proportional hazards (PH).
- DeepHit – First Hitting Time neural networks.
  - Inverse Gaussian distribution used as baseline hazard.
  - Does not let model determine baseline hazard.
- DeepSurvivalMachines (DSM) – Mixture model used for baseline hazard.
  - User specifies a set of distributions to be used as the baseline hazard.
  - Does not permit time-varying interactions.

- **Person-moment neural networks (PMNN)**
  - Provides a flexible baseline hazard.
  - Permits time-varying interactions of covariates.
  - Applicable to high-dimensional datasets

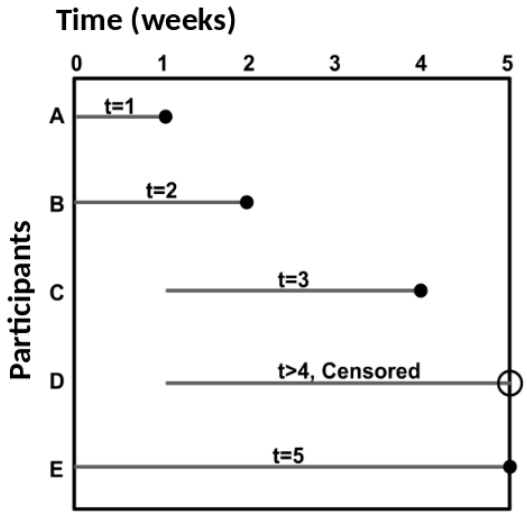
Case-base sampling

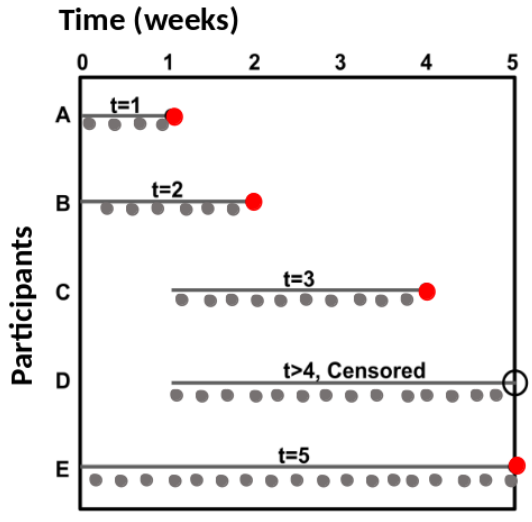




- Base: All the person-moments experienced in the study.

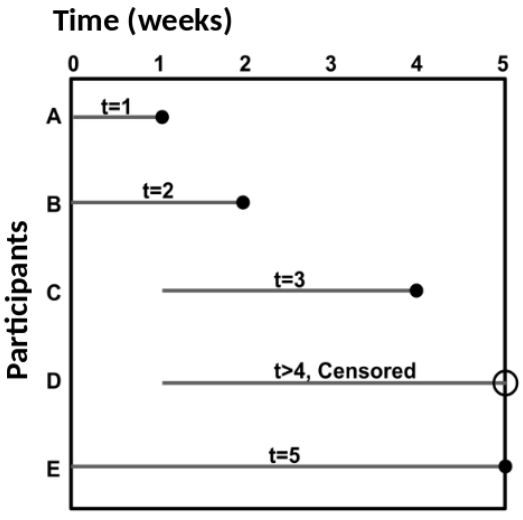
Case-base sampling

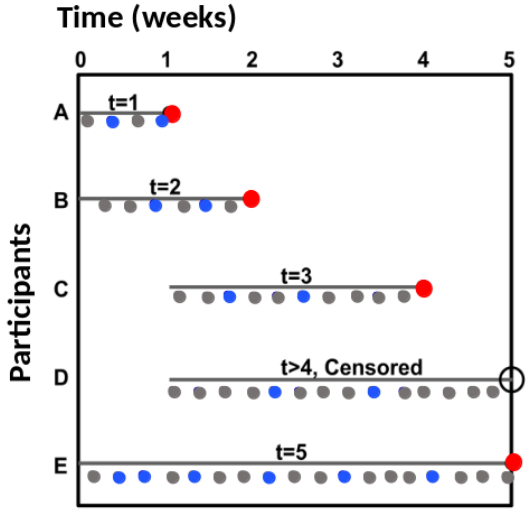




- Base: All the person-moments experienced in the study.
- Case series: all the person-moments where an event occurred.

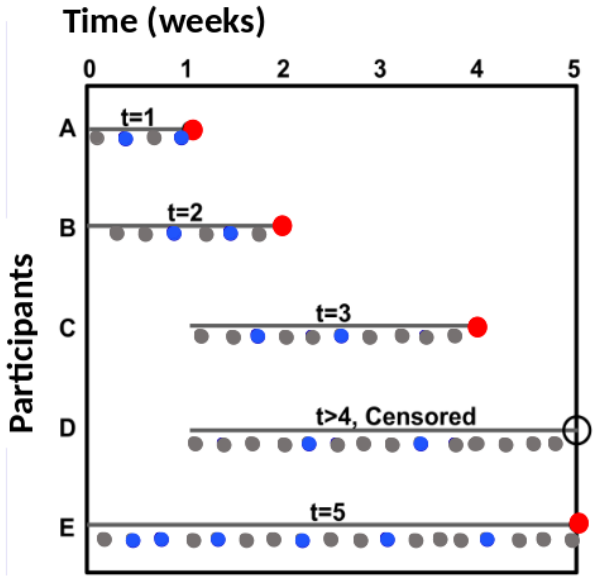
Case-base sampling





- Base: All the person-moments experienced in the study.
- Case series: all the person-moments where an event occurred.
- Base series: sample of the base.

Case-base sampling and logistic regression



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

$$\frac{Pr(Y=1|x,t)}{Pr(Y=0|x,t)} = \frac{h(x,t) * B(x,t)}{b[B(x,t)/B]}$$

$$\frac{h(x,t) * B(x,t)}{b[B(x,t)/B]} = \frac{h(x,t) * B}{b}$$

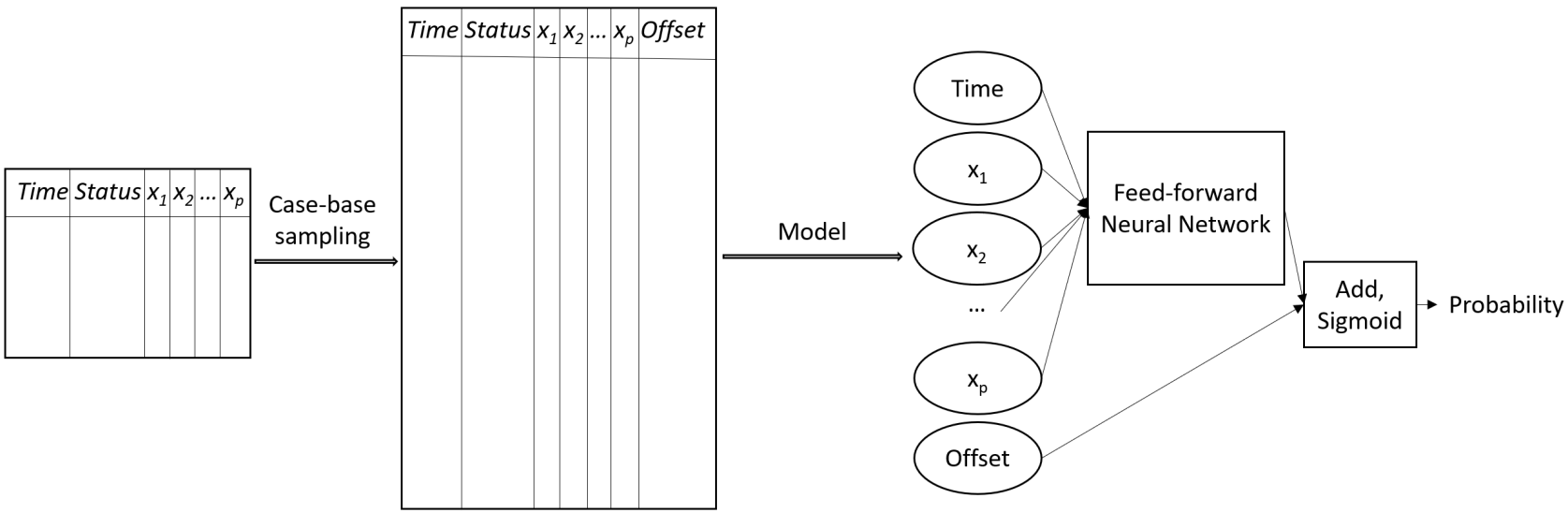
$$h(x,t) = e^{\beta(x,t)} \frac{b}{B}$$

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

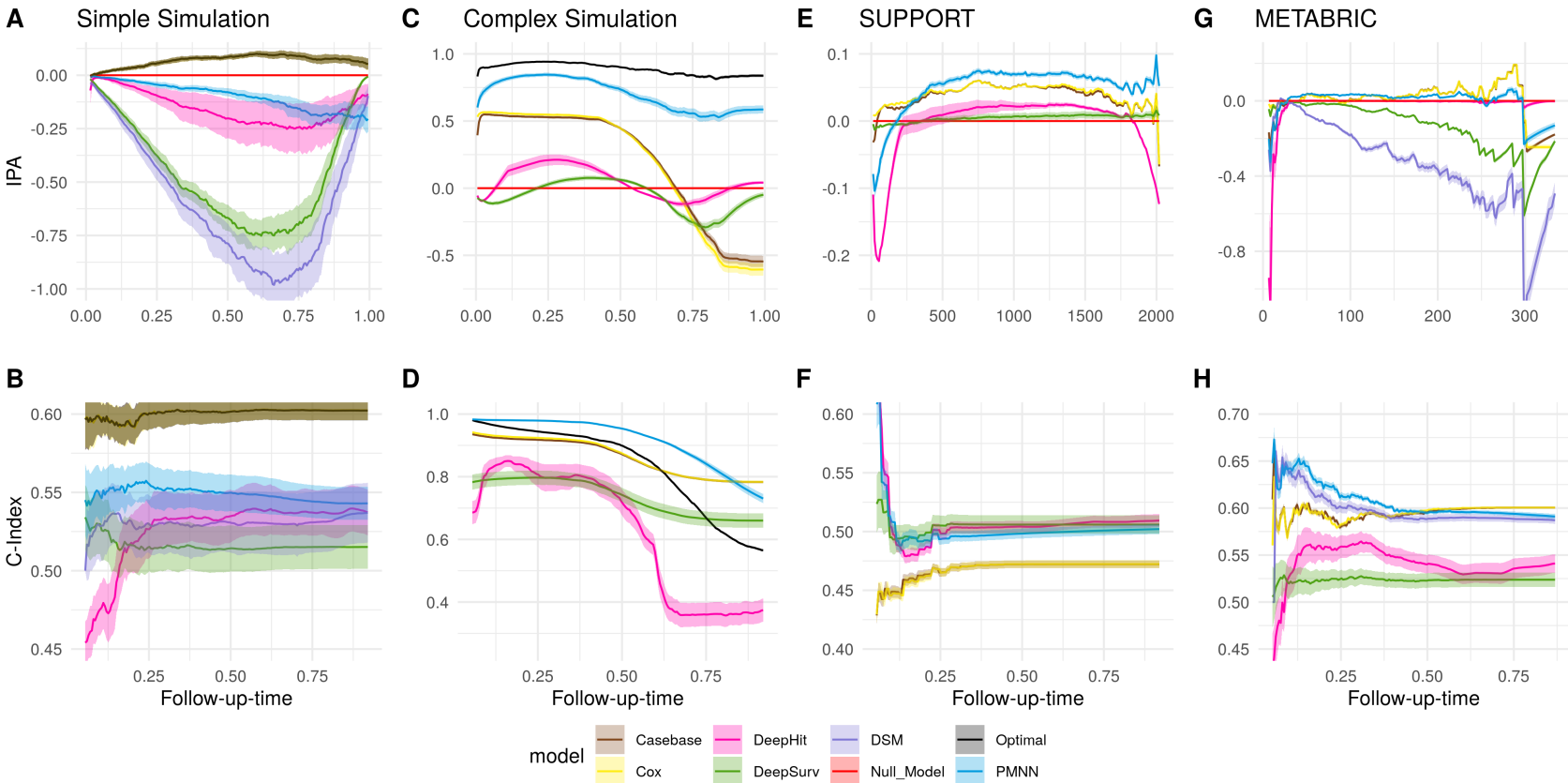
b = # Blue
B = # Moments

Bhatnagar et al. *In revision at R Journal* (2021+).
https://cran.r-project.org/package=casebase

Overview of our method



Results


A
Simple Simulation
C
Complex Simulation
E
SUPPORT
G
METABRIC
IPA
B
D
F
H
C-Index
Follow-up-time
model
Casebase
Cox
DeepHit
DeepSurv
DSM
Null_Model
Optimal
PMNN

Acknowledgements

Zeyu Bian, PhD (c)

CANSSI INCASS

NSERC CRSNG

CIHR IRSC

McGill initiative in Computational Medicine

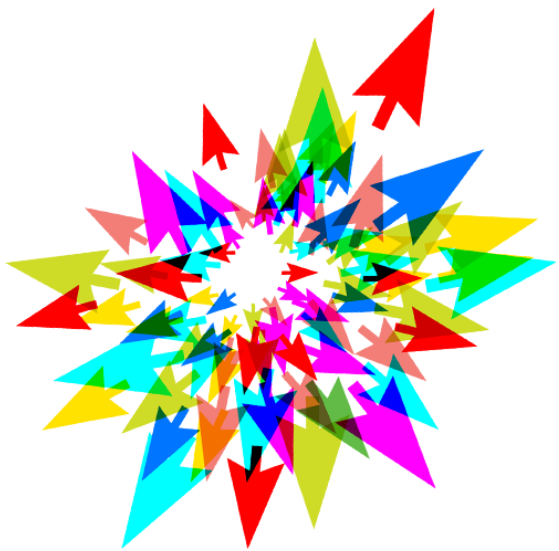
Mitacs

Acknowledgements

- Tianyuan Lu (McGill)
- Yi Yang (McGill)
- Celia Greenwood (Lady Davis Institute)
- Erica Moodie (McGill)
- Kieran O'Donnell (Yale)

Calcul Québec

compute canada | calcul canada



References

1. Bhatnagar, SR, Lu, T, Lovato, A, Olds, DL, Kobor, MS, Meaney, MJ, O'Donnell, K, Yang, Y, and Greenwood, CMT (2021+). A Sparse Additive Model for High-Dimensional Interactions with an Exposure Variable. bioRxiv. DOI 10.1101/445304. *In revision at Computational Statistics and Data Analysis.*

2. **Bian Z**, Moodie EEM, Shortreed S, Bhatnagar SR (2021). Variable Selection in Regression-based Estimation of Dynamic Treatment Regimes. https://arxiv.org/abs/2101.07359. *In press at Biometrics.*

3. Bhatnagar SR, Turgeon M, **Islam J**, Hanley JA, Saarela O (2021+). casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates. https://arxiv.org/abs/2009.10264. *Revision submitted at R Journal.*

4. 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.

5. Bhatnagar SR, Yang Y, Khundrakpam B, Evans A, Blanchette M, Bouchard L, Greenwood CMT (2017). An analytic approach for interpretable predictive models in high dimensional data, in the presence of interactions with exposures. *Genetic Epidemiology.* Apr 1;42(3):233-49. DOI 10.1101/102475.

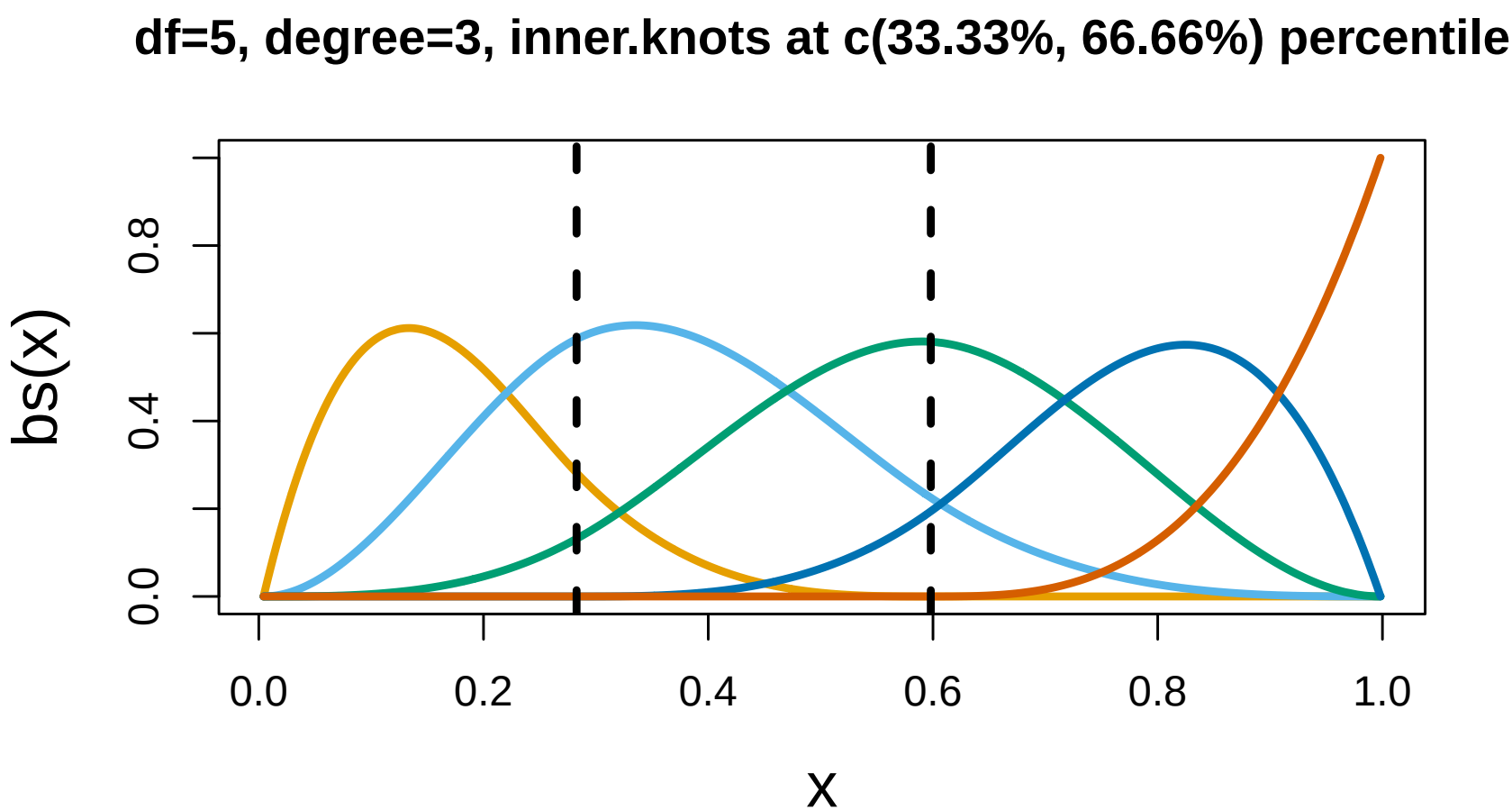
sahirbhatnagar.com

Session Info

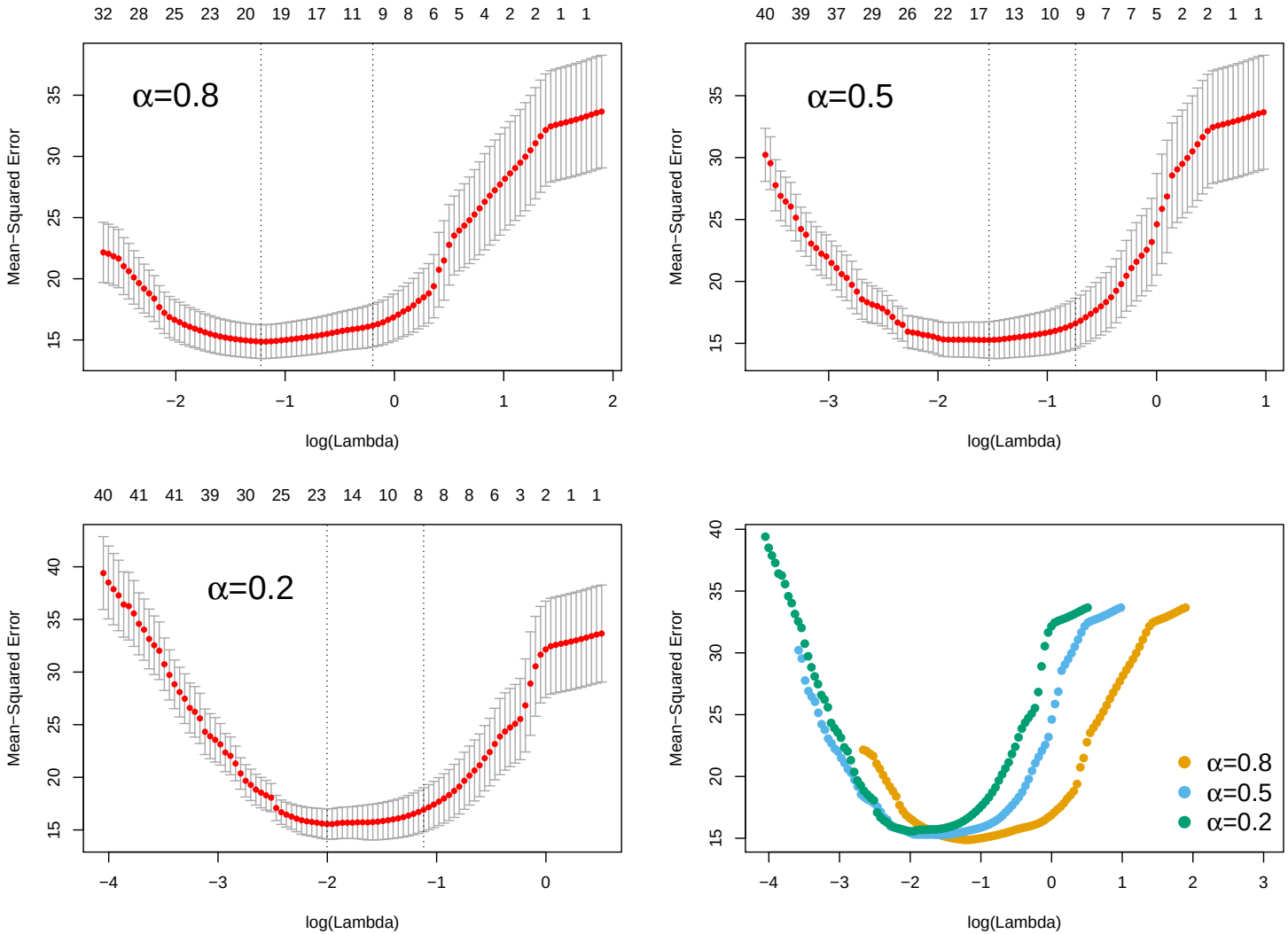
```
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

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-r0.3.13.so

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

other attached packages:
[1] xtable_1.8-4       rpart.plot_3.1.0   rpart_4.1-15       data.table_1.14.2
[5] ISLR_1.2           ggplot2_3.3.5.9000 knitr_1.36

loaded via a namespace (and not attached):
 [1] pillar_1.6.4       compiler_4.1.1     highr_0.9          tools_4.1.1
 [5] digest_0.6.28      evaluate_0.14      lifecycle_1.0.1    tibble_3.1.5
 [9] gtable_0.3.0       pkgconfig_2.0.3    rlang_0.4.12       DBI_1.1.1
[13] xfun_0.26          withr_2.4.2        dplyr_1.0.7        stringr_1.4.0
[17] generics_0.1.0     vctrs_0.3.8        grid_4.1.1         tidyselect_1.1.1
[21] glue_1.4.2         R6_2.5.1           fansi_0.5.0        pacman_0.5.1
[25] purrr_0.3.4        RSkittleBrewer_1.1 blob_1.2.1         magrittr_2.0.1
[29] scales_1.1.1       ellipsis_0.3.2     assertthat_0.2.1   colorspace_2.0-2
[33] utf8_1.2.2         stringi_1.7.5      munsell_0.5.0      crayon_1.4.1
```

B-Spline Expansion

```
x <- truncnorm::rtruncnorm(1000, a = 0, b = 1)
B <- splines::bs(x, df = 5, degree=3, intercept = FALSE)
```



`sail` A Note on the Second Tuning Parameter results

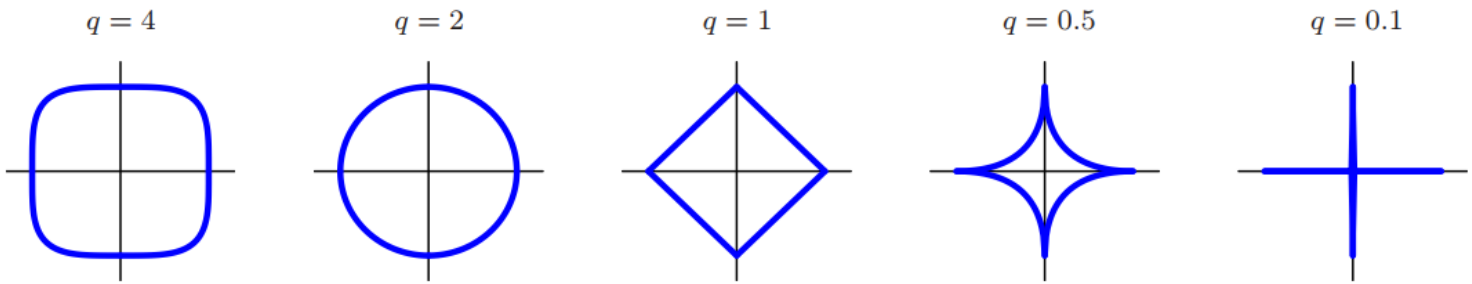


Why the L1 norm ?

- For a fixed real number $q \geq 0$ consider the criterion

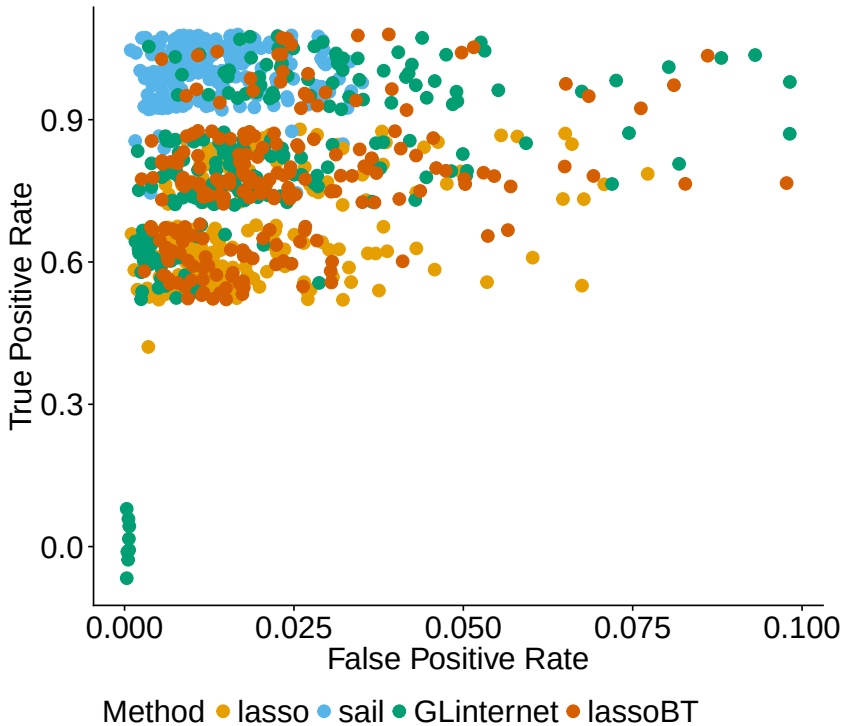
$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

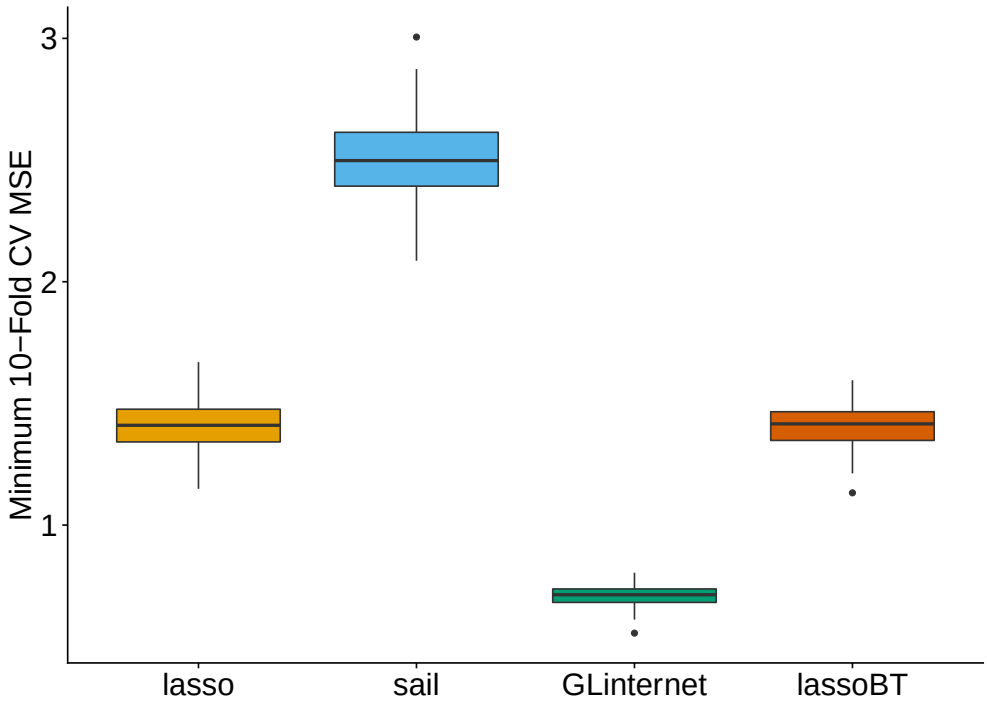
- Why do we use the ℓ_1 norm? Why not use the $q = 2$ (Ridge) or any ℓ_q norm?



- $q = 1$ is the smallest value that yields a sparse solution and yields a **convex** problem $\rightarrow$ scalable to high-dimensional data
- For $q < 1$ the constrained region is **nonconvex**

Linear Effects Simulation - Comparison





Simulations

Theory

Algorithm

Discussion

ggmix appendix

DTR

Dynamic Treatment Regimes

Simulation Scenarios

1. **Truth obeys strong hierarchy** (**right in our wheel house**):

$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

Simulation Scenarios

1. **Truth obeys strong hierarchy** (**right in our wheel house**):

$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

2. **Truth obeys weak hierarchy**

Simulation Scenarios

1. **Truth obeys strong hierarchy (right in our wheel house):**

$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

2. **Truth obeys weak hierarchy**
3. **Truth only has interactions**

Simulation Scenarios

1. **Truth obeys strong hierarchy (right in our wheel house):**

$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

2. **Truth obeys weak hierarchy**
3. **Truth only has interactions**
4. **Truth is linear**

Simulation Scenarios

1. **Truth obeys strong hierarchy (right in our wheel house):**

$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

2. **Truth obeys weak hierarchy**
3. **Truth only has interactions**
4. **Truth is linear**
5. **Truth only has main effects**

Simulation Scenarios

1. **Truth obeys strong hierarchy** (**right in our wheel house**):

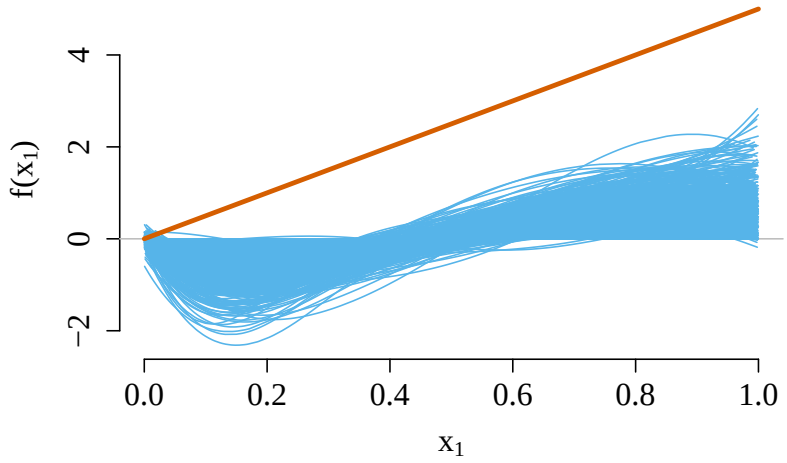
$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

2. **Truth obeys weak hierarchy**
3. **Truth only has interactions**
4. **Truth is linear**
5. **Truth only has main effects**

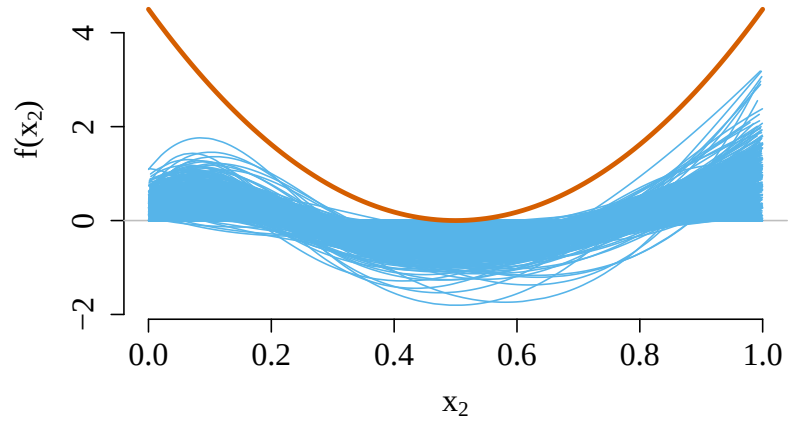
- $n_{train} = n_{tuning} = 200$, $n_{test} = 800$, $p = 1000$, $\beta_E = 1$, $SNR = 2$
- $X_j \sim$ `truncnorm(0,1)`, $j = 1, \ldots, 1000$, $E \sim$ `truncnorm(-1,1)`
- `sail` needs to estimate $1000 \times 5 \times 2 = 10$k parameters

Scenario 1: Main Effects for 500 Simulations

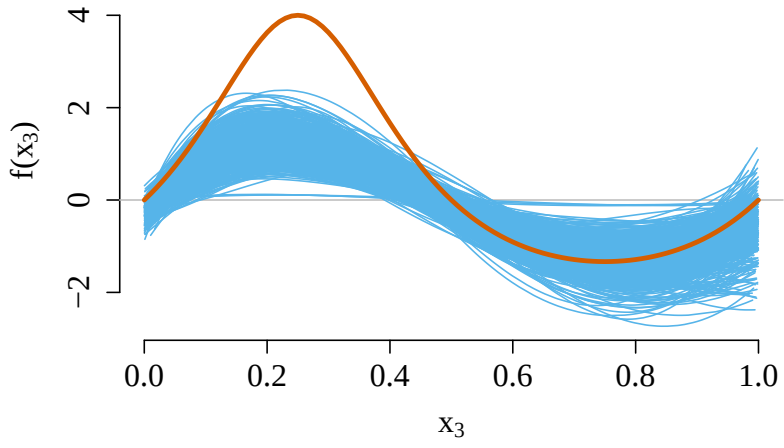
$f(x_1) = 5x_1$



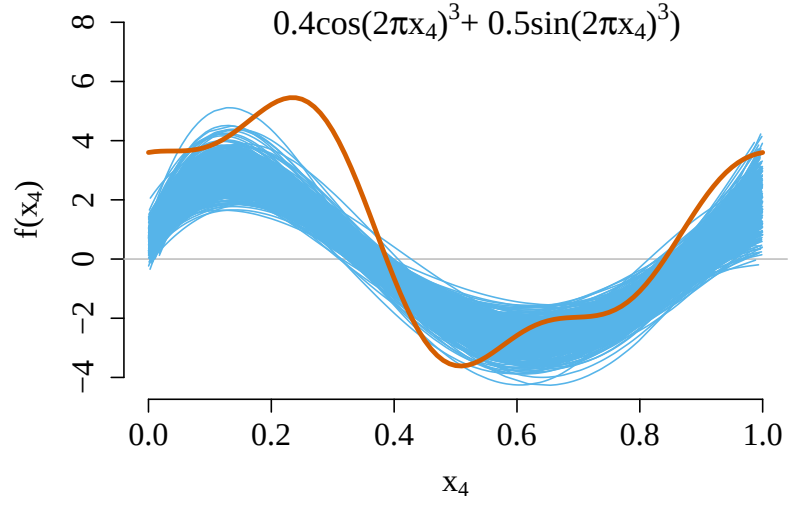
$f(x_2) = 4.5(2x_2 - 1)^2$



$f(x_3) = \frac{4\sin(2\pi x_3)}{2 - \sin(2\pi x_3)}$

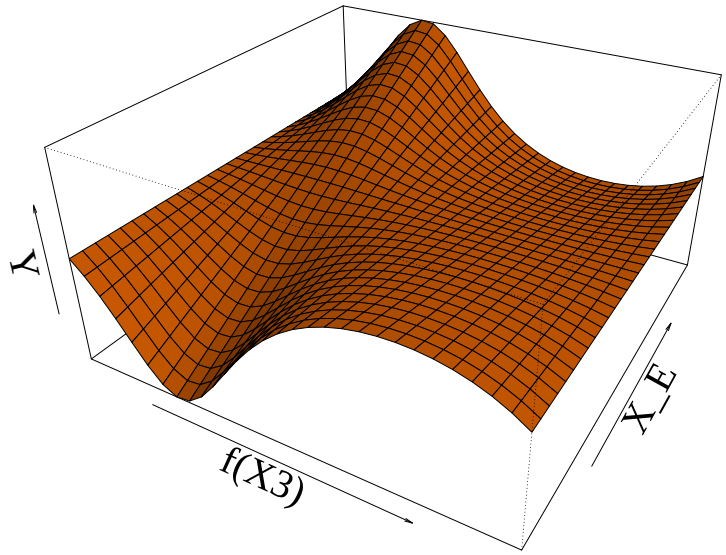


$f(x_4) = 6(0.1\sin(2\pi x_4) + 0.2\cos(2\pi x_4) + 0.3\sin(2\pi x_4)^2 + 0.4\cos(2\pi x_4)^3 + 0.5\sin(2\pi x_4)^3)$

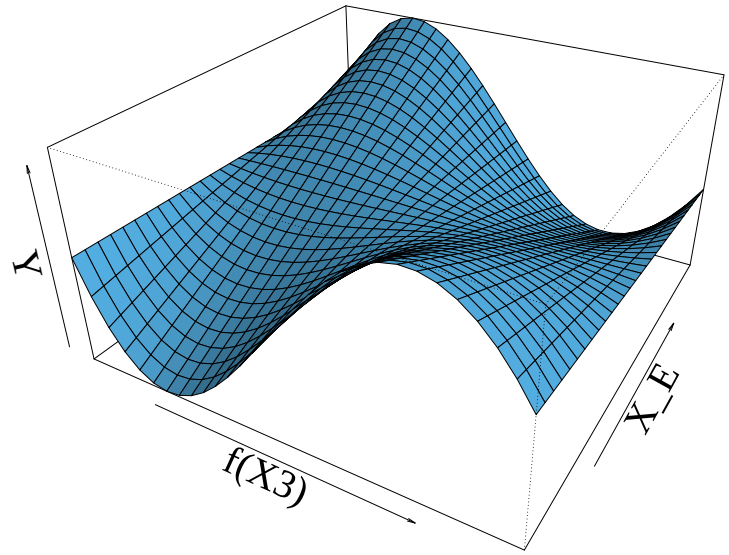


Scenario 1: Estimated Interaction Effects for $E \cdot f(X_3)$

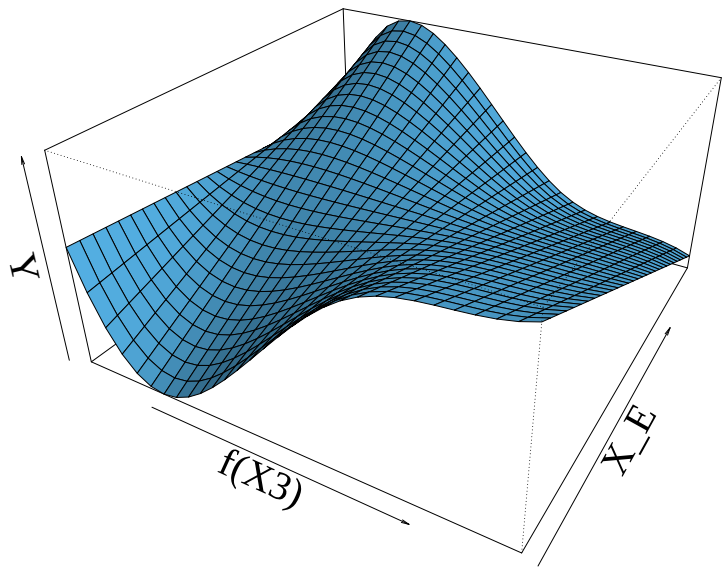
Truth



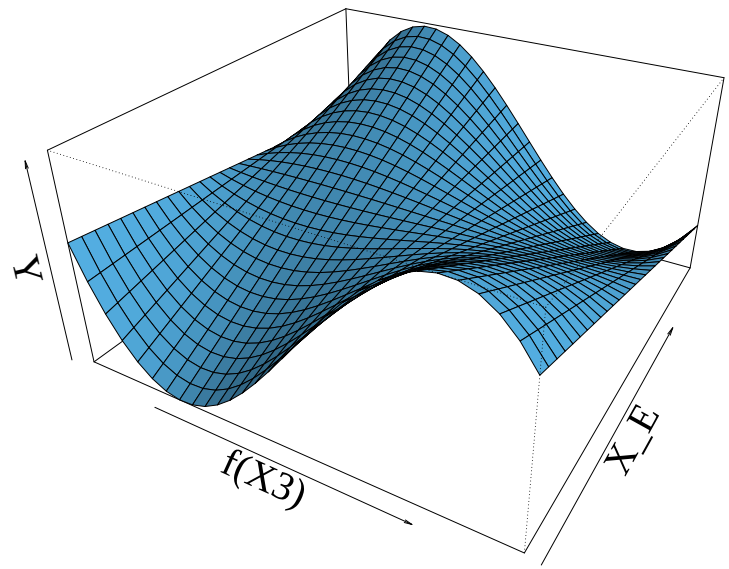
Estimated: 25th Percentile



Estimated: 50th Percentile

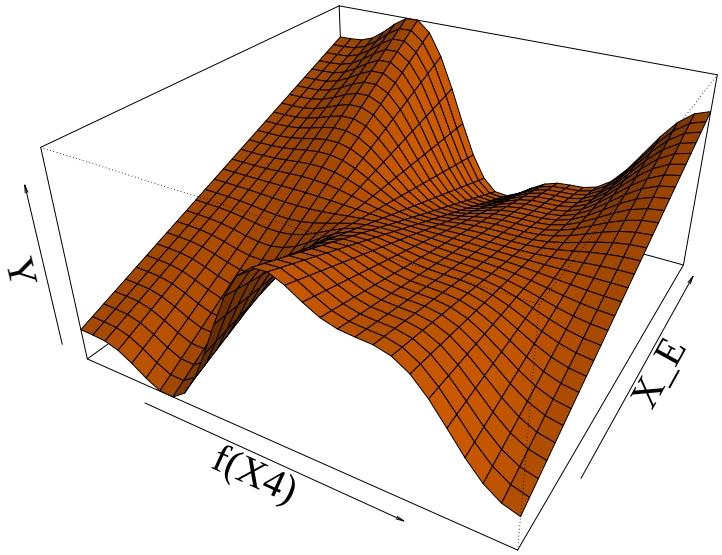


Estimated: 75th Percentile

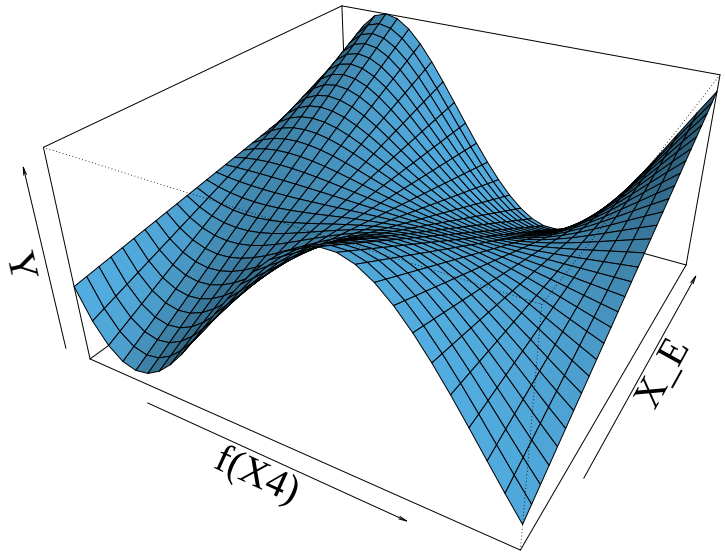


Scenario 1: Estimated Interaction Effects for $E \cdot f(X_4)$

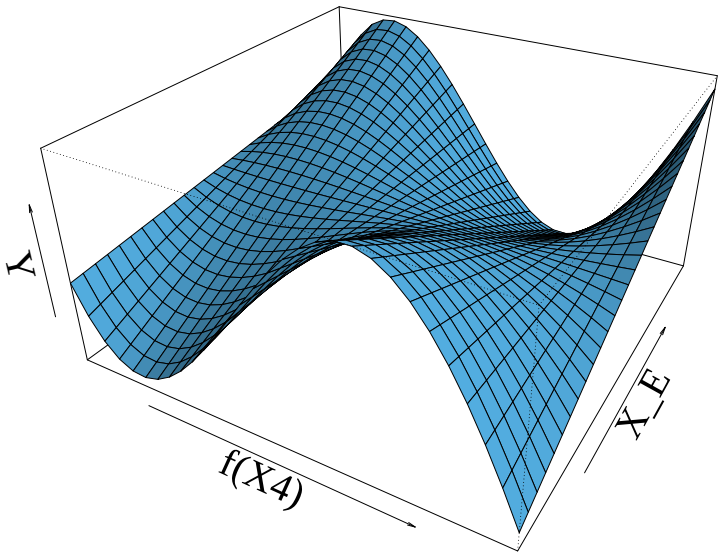
Truth



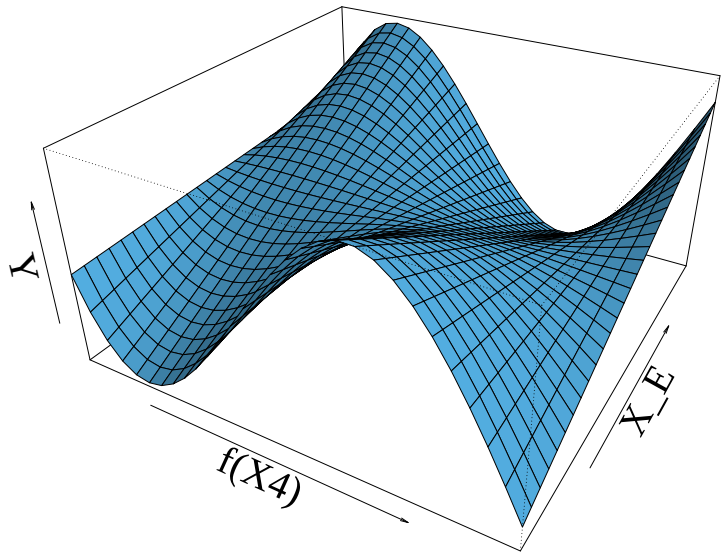
Estimated: 25th Percentile



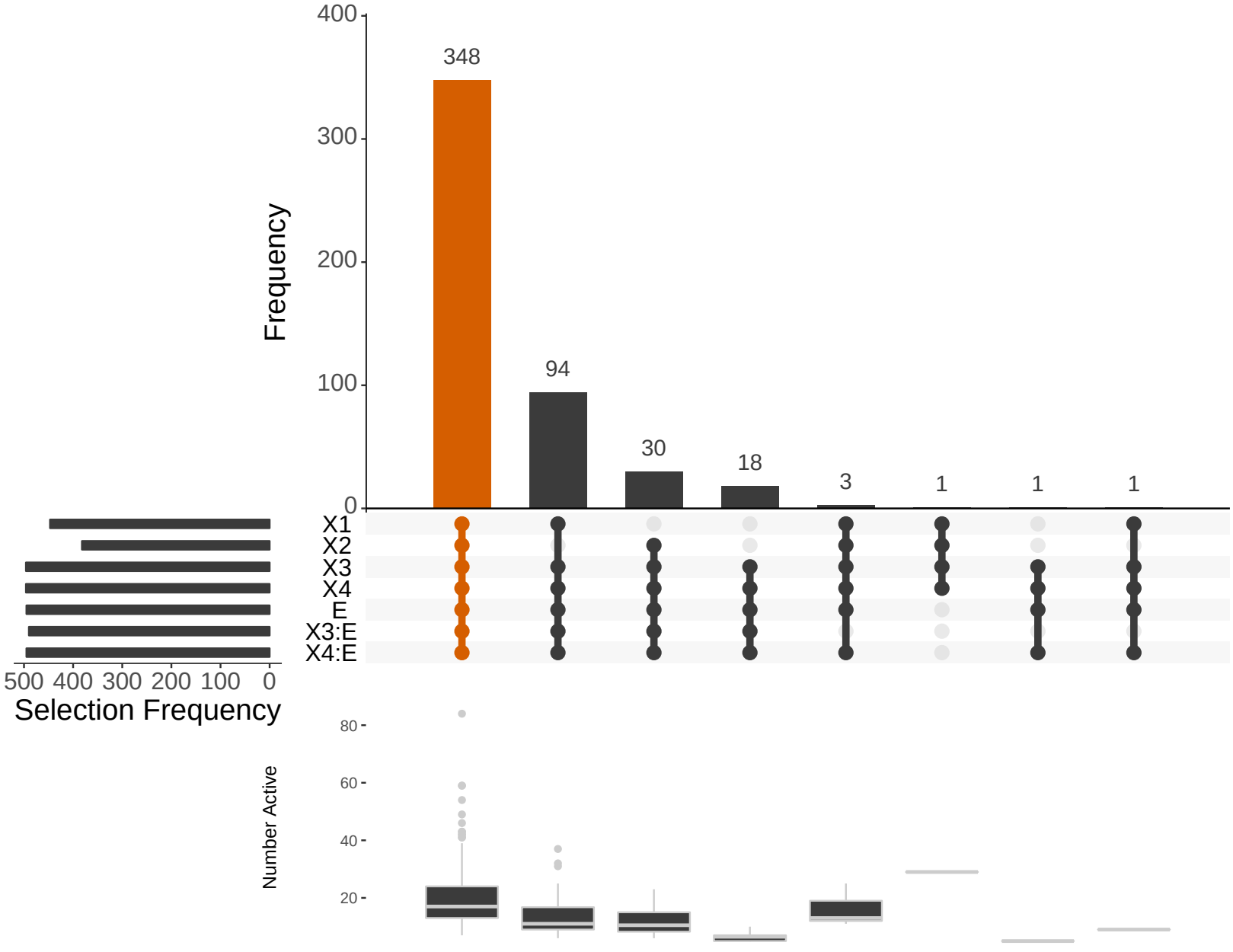
Estimated: 50th Percentile



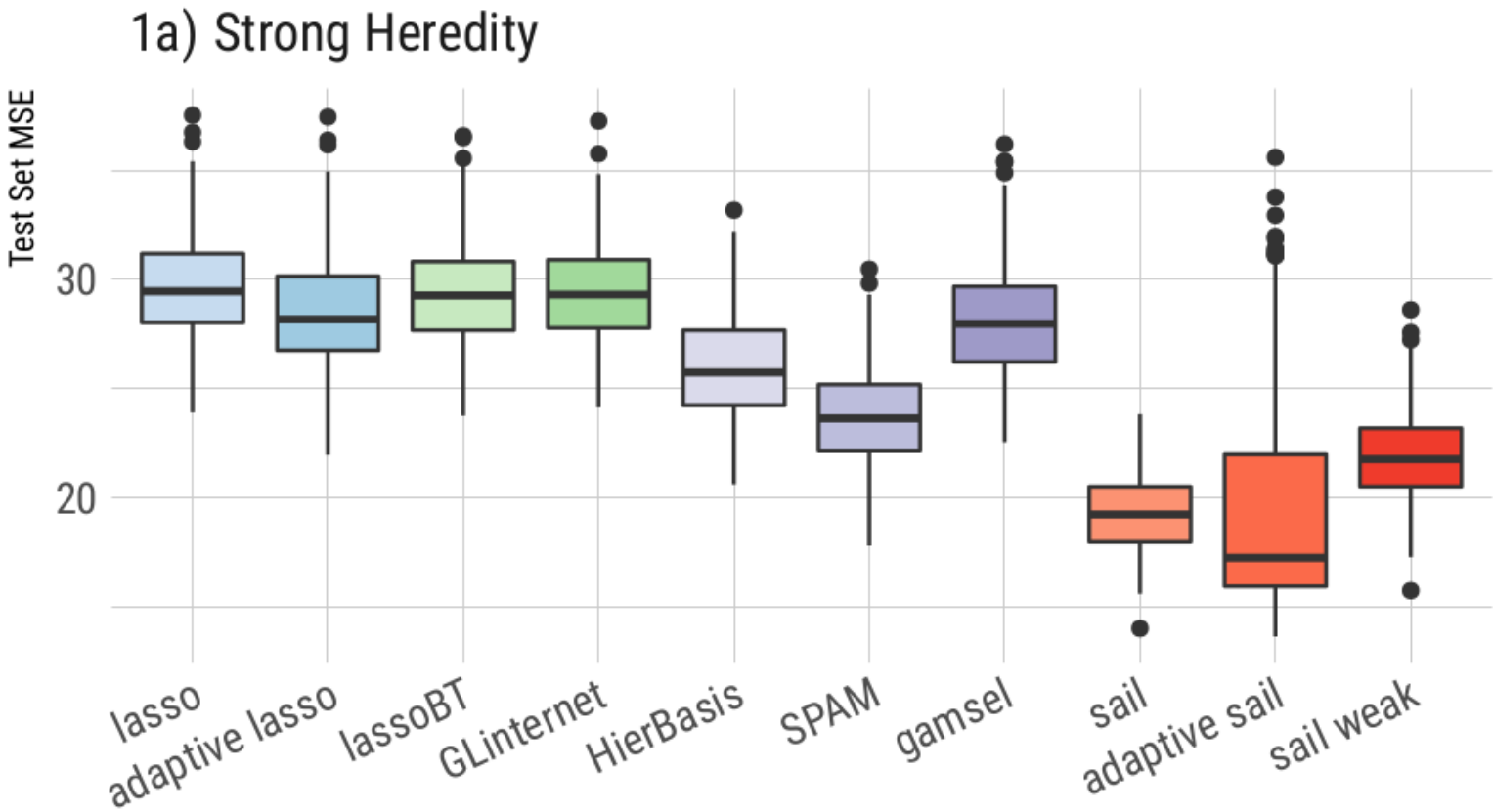
Estimated: 75th Percentile



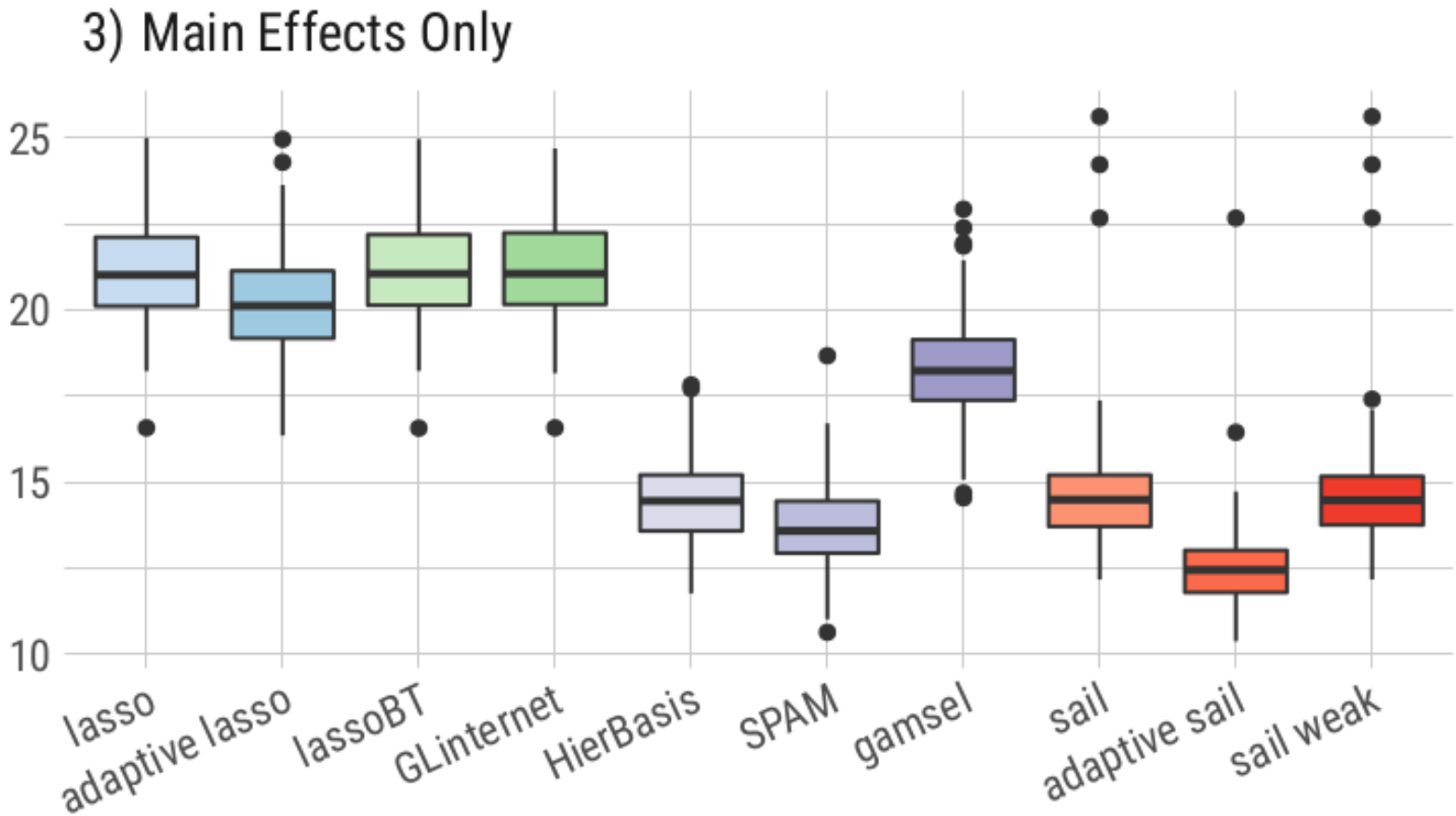
Right in Our Wheel House Simulation Results



Strong Heredity



Main Effects Only



Sparsity

Theorem 1

$$\widehat{\boldsymbol{\Theta}}_n = \underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

$$\mathcal{A}_1 = \{j : \theta_j \neq 0, \beta_j \neq 0\}$$
$$\mathcal{A}_2 = \{k : \gamma_k \neq 0\}, \qquad \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$$

Under certain regularity conditions and the existence of a local minimizer $\widehat{\boldsymbol{\Theta}}_n$ that is $\sqrt{n}$-consistent

$$P\left(\widehat{\boldsymbol{\Theta}}_{\mathcal{A}^c} = 0\right) \to 1$$

Sparsity

Theorem 1

$$\widehat{\boldsymbol{\Theta}}_n = \underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

$$\mathcal{A}_1 = \{j : \theta_j \neq 0, \beta_j \neq 0\}$$
$$\mathcal{A}_2 = \{k : \gamma_k \neq 0\}, \qquad \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$$

Under certain regularity conditions and the existence of a local minimizer $\widehat{\boldsymbol{\Theta}}_n$ that is $\sqrt{n}$-consistent

$$P\left(\widehat{\boldsymbol{\Theta}}_{\mathcal{A}^c} = 0\right) \to 1$$

Theorem 1 shows that when the tuning parameters for the nonzero coefficients converge to 0 faster than $n^{-1/2}$ `sail` can consistently remove the noise terms with probability tending to 1.

Asymptotic normality

Theorem 2

$$\widehat{\boldsymbol{\Theta}}_n = \underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

Under certain regularity conditions, the component $\widehat{\boldsymbol{\Theta}}_{\mathcal{A}}$ of the local minimizer $\widehat{\boldsymbol{\Theta}}_n$ satisfies

$$\sqrt{n}\left(\widehat{\boldsymbol{\Theta}}_{\mathcal{A}} - \boldsymbol{\Theta}_{\mathcal{A}}\right) \to_d \mathcal{N}\left(0, \mathbf{I}^{-1}\left(\boldsymbol{\Theta}_{\mathcal{A}}\right)\right)$$

Theorem 2 shows that the `sail` estimates for nonzero coefficients in the true model have the same asymptotic distribution as they would have if the zero coefficients were known in advance.

Asymptotic normality

Theorem 2

$$\widehat{\boldsymbol{\Theta}}_n = \underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

Under certain regularity conditions, the component $\widehat{\boldsymbol{\Theta}}_{\mathcal{A}}$ of the local minimizer $\widehat{\boldsymbol{\Theta}}_n$ satisfies

$$\sqrt{n}\left(\widehat{\boldsymbol{\Theta}}_{\mathcal{A}} - \boldsymbol{\Theta}_{\mathcal{A}}\right) \to_d \mathcal{N}\left(0, \mathbf{I}^{-1}\left(\boldsymbol{\Theta}_{\mathcal{A}}\right)\right)$$

Theorem 2 shows that the `sail` estimates for nonzero coefficients in the true model have the same asymptotic distribution as they would have if the zero coefficients were known in advance.
Theorem 1 + 2 –> Oracle property (Fan and Li, 2001)

Simulations

Theory

Algorithm

Discussion

ggmix appendix

DTR
    Dynamic Treatment Regimes

Block Relaxation (De Leeuw, 1994)

Algorithm 1: Block Relaxation Algorithm

Set the iteration counter $k \leftarrow 0$ and fix $\alpha \in (0, 1)$;

for each λ **do**

  repeat

$$\boldsymbol{\gamma}^{(k+1)} \leftarrow \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \quad Q_\lambda\left(\boldsymbol{\gamma}, \beta_E^{(k)}, \boldsymbol{\theta}^{(k)}\right)$$

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \quad Q_\lambda\left(\boldsymbol{\theta}, \beta_E^{(k)}, \boldsymbol{\gamma}^{(k+1)}\right)$$

$$\beta_E^{(k+1)} \leftarrow \underset{\boldsymbol{\beta_E}}{\operatorname{argmin}} \quad Q_\lambda\left(\boldsymbol{\theta}^{(k+1)}, \beta_E, \boldsymbol{\gamma}^{(k+1)}\right)$$

    $k \leftarrow k + 1$

  until convergence criterion is satisfied;

end

sail: Weak Heredity

Reparametrization

$$\boldsymbol{\tau}_j = \gamma_j(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j)$$

Model

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \boldsymbol{\Psi}_j \theta_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j (X_E \circ \boldsymbol{\Psi}_j)(\beta_E \cdot \mathbf{1}_{m_j} + \boldsymbol{\theta}_j) + \varepsilon$$

Objective Function

$$\underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

Implementation

Objective Function

$$\underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\text{argmin}} \quad \mathcal{L}(Y; \boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

[1] https://cran.r-project.org/package=sail

Implementation

Objective Function

$$\underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(Y; \boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

Lasso problem

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(Y; \boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

[1] https://cran.r-project.org/package=sail

Implementation

Objective Function

$$\underset{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}}{\text{argmin}} \quad \mathcal{L}(Y; \boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2\right) + \lambda\alpha \sum_{j=1}^{p} w_{jE}|\gamma_j|$$

[1] https://cran.r-project.org/package=sail

Implementation

Objective Function

$$\underset{\beta_E,\boldsymbol{\theta},\boldsymbol{\gamma}}{\operatorname{argmin}} \quad \mathcal{L}(Y;\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j\|\theta_j\|_2\right) + \lambda\alpha\sum_{j=1}^{p} w_{jE}|\gamma_j|$$

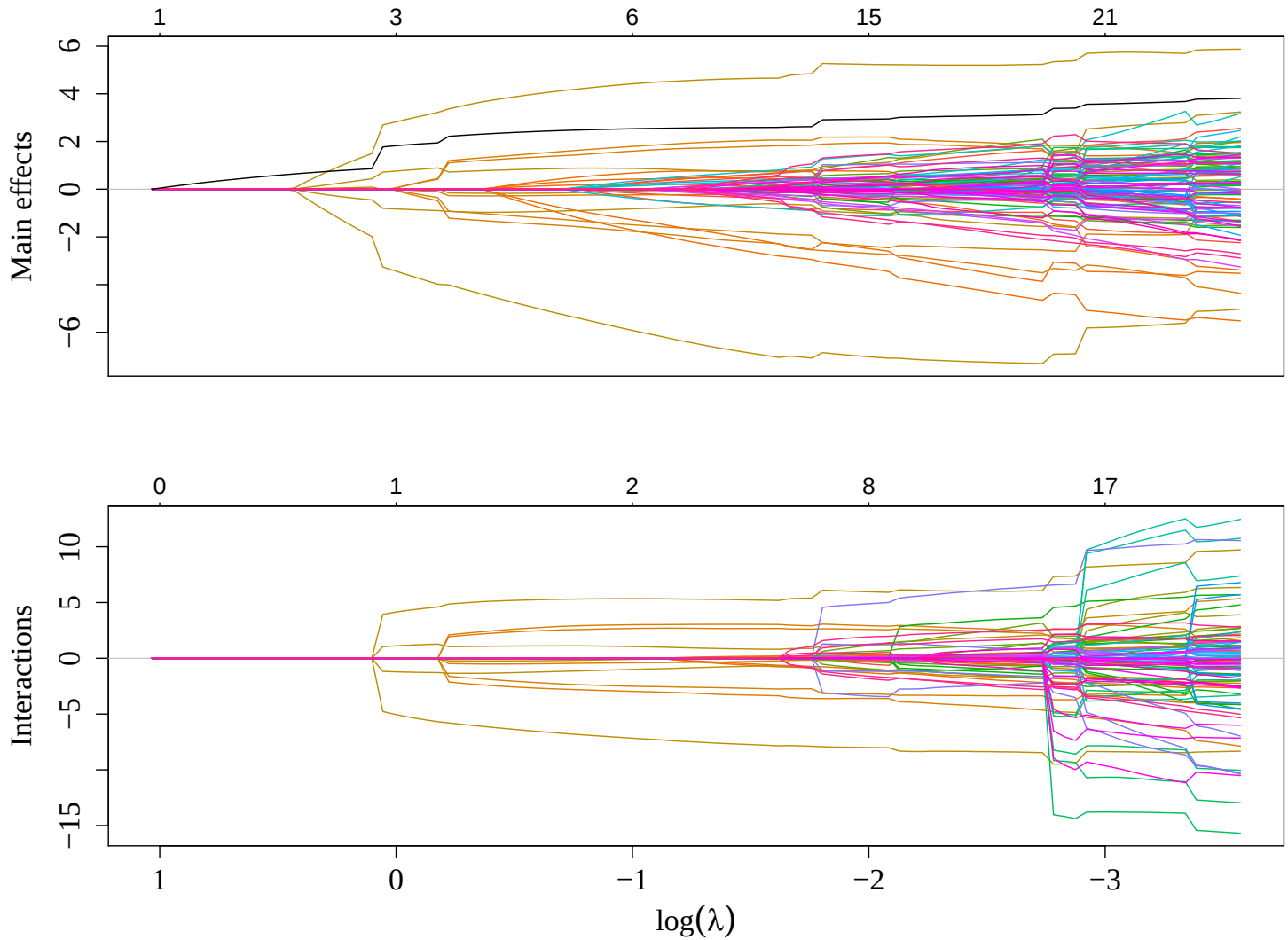
Group Lasso problem

$$\underset{\beta_E,\boldsymbol{\theta}}{\operatorname{argmin}} \quad \mathcal{L}(Y;\boldsymbol{\Theta}) + \lambda(1-\alpha)\left(w_E|\beta_E| + \sum_{j=1}^{p} w_j\|\theta_j\|_2\right) + \lambda\alpha\sum_{j=1}^{p} w_{jE}|\gamma_j|$$

[1] https://cran.r-project.org/package=sail

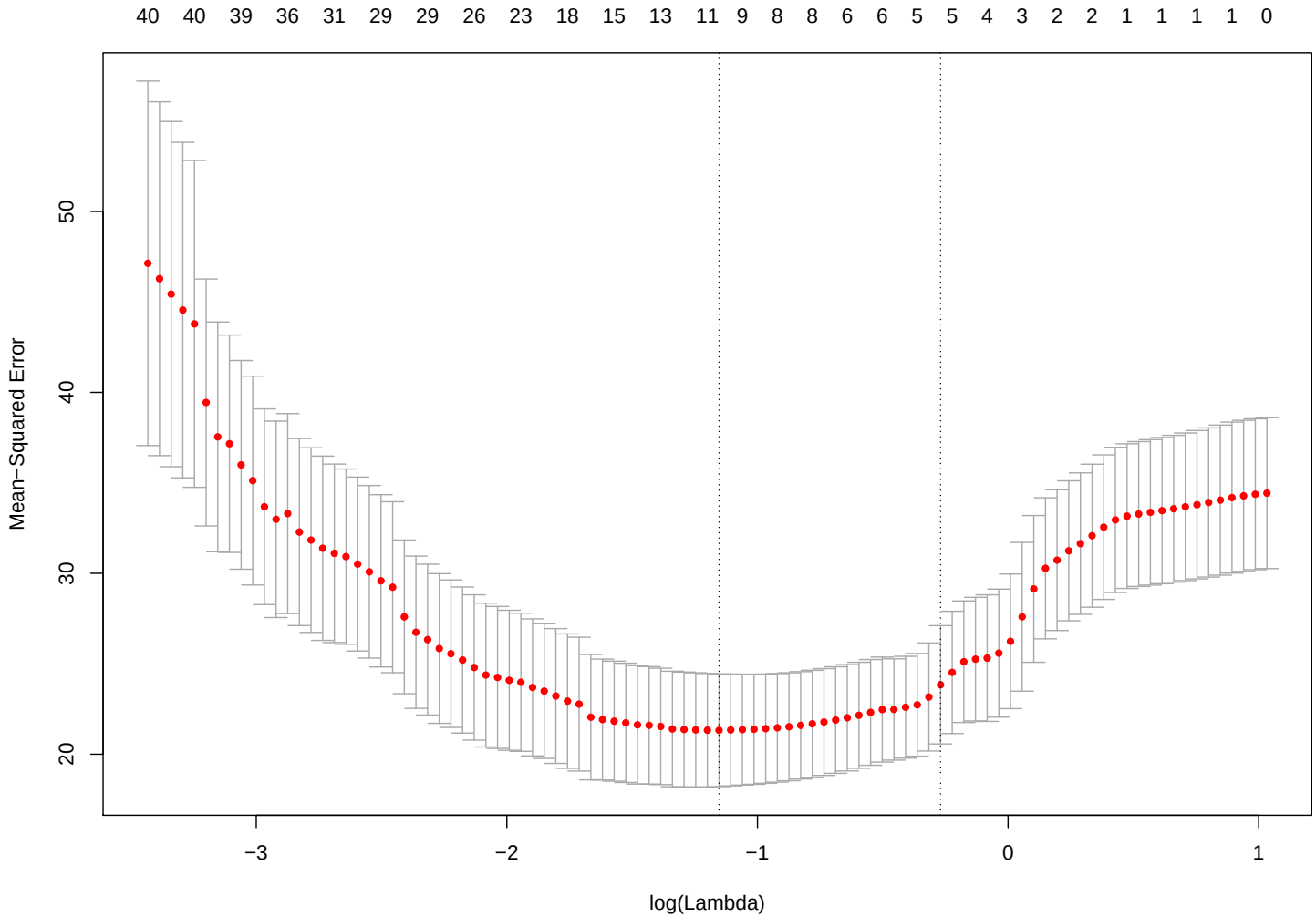
`sail` R package: Solution Path results

```
f.basis <- function(x) splines::bs(x, degree = 5)
fit <- sail(x, y, e, basis = f.basis)
plot(fit)
```



`sail` R package: Cross-validation results

```
sail::plot(cvfit)
```



Strengths and Limitations

Strengths

- Non-linear environment interactions with strong heredity property in $p >> N$
- `sail` allows for flexible modeling of input variables

Strengths and Limitations

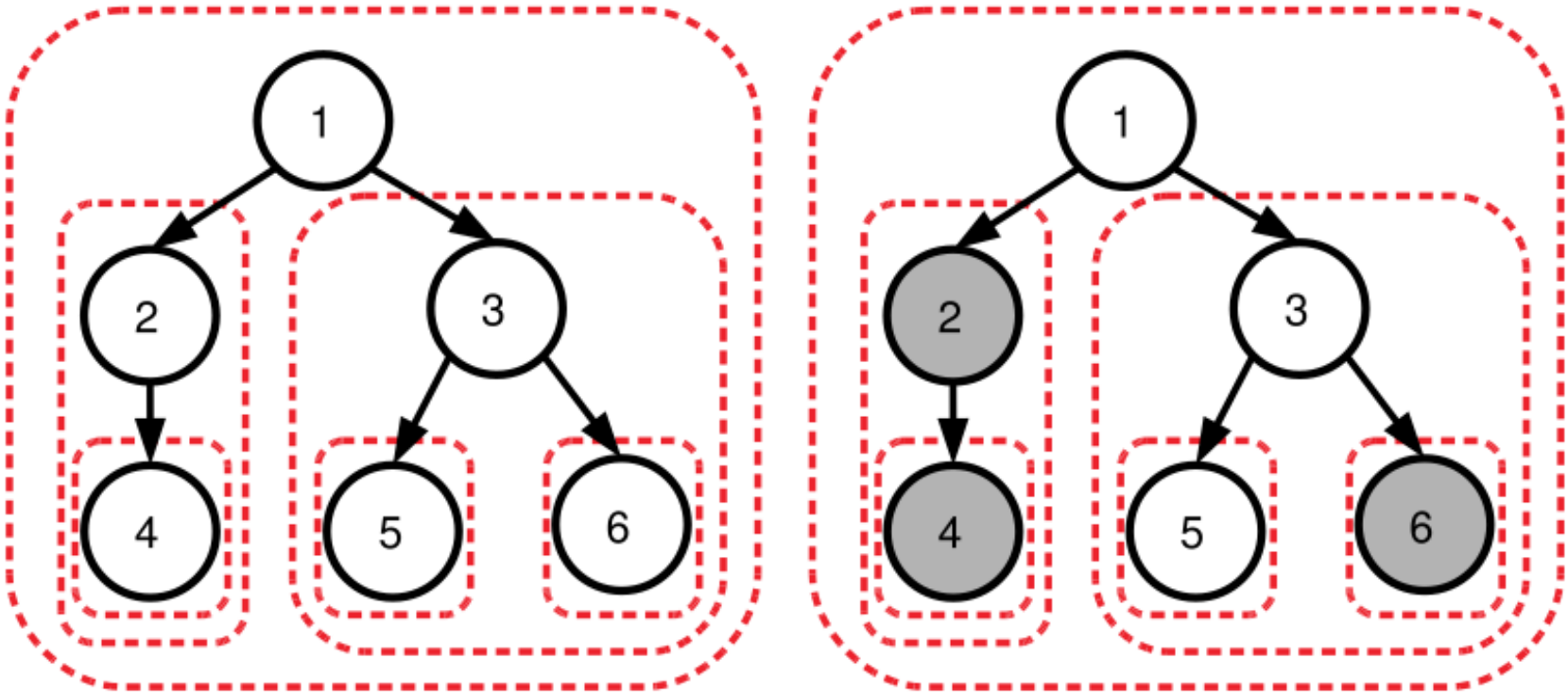
Strengths

- Non-linear environment interactions with strong heredity property in $p >> N$
- `sail` allows for flexible modeling of input variables

Limitations

- `sail` can currently only handle $E \cdot f(X)$ or $f(E) \cdot X$
- Does not allow for $f(X_1, E)$ or $f(X_1, X_2)$
- Memory footprint is an issue

Hierarchical Penalty Structure



[1] Bach, Jenatton, Mairal and Obozinski (2011). Optimization with Sparsity-Inducing Penalties.

Bi-level selection

- Bi-level selection:

$$f(X_1) = \underbrace{\begin{bmatrix} X_{11} & \psi_{11}(X_{11}) & \psi_{12}(X_{12}) & \cdots & \psi_{11}(X_{15}) \\ & \vdots & \vdots & \cdots & \vdots \\ & \vdots & \vdots & \cdots & \vdots \\ X_{i1} & \psi_{11}(X_{i1}) & \psi_{12}(X_{i2}) & \cdots & \psi_{11}(X_{i5}) \\ & \vdots & \vdots & \cdots & \vdots \\ & \vdots & \vdots & \cdots & \vdots \\ X_{N1} & \psi_{11}(X_{N1}) & \psi_{12}(X_{N2}) & \cdots & \psi_{11}(X_{N5}) \end{bmatrix}_{N\times 5}}_{\mathbf{\Psi_1}} \times \underbrace{\begin{bmatrix} \beta_{\text{linear}} \\ \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{14} \\ \beta_{15} \end{bmatrix}_{6\times 1}}_{\theta_1}$$

Simulations

Theory

Algorithm

Discussion

ggmix appendix

DTR
    Dynamic Treatment Regimes

Block Relaxation (De Leeuw, 1994)

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

Algorithm 2: Block Relaxation Algorithm

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

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

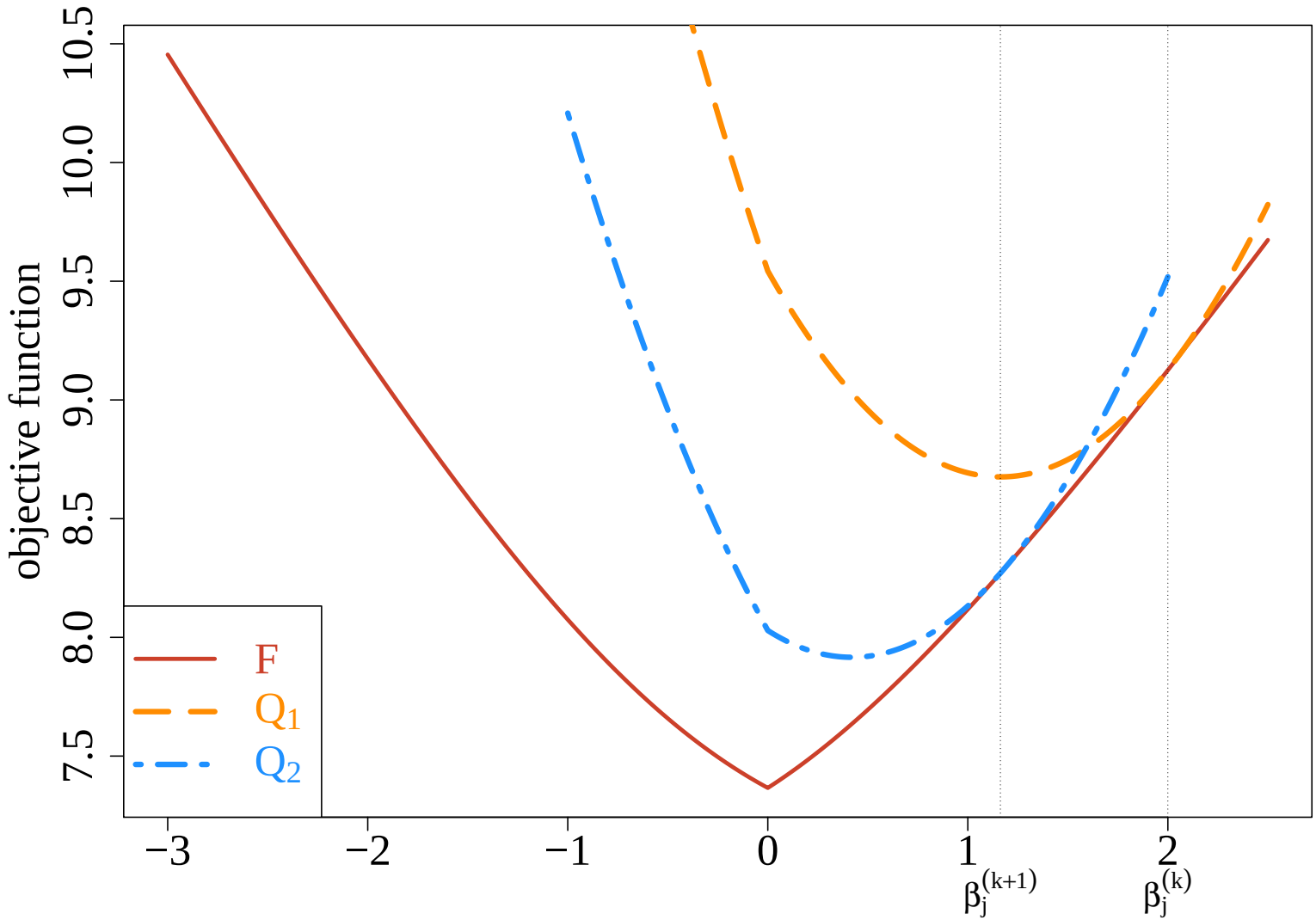
[1] Tseng P& Yun S. Math. Program., Ser. B, (2009)

Coordinate Gradient Descent Method

- We take advantage of smoothness of $\ell(\mathbf{\Theta})$
- 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)


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

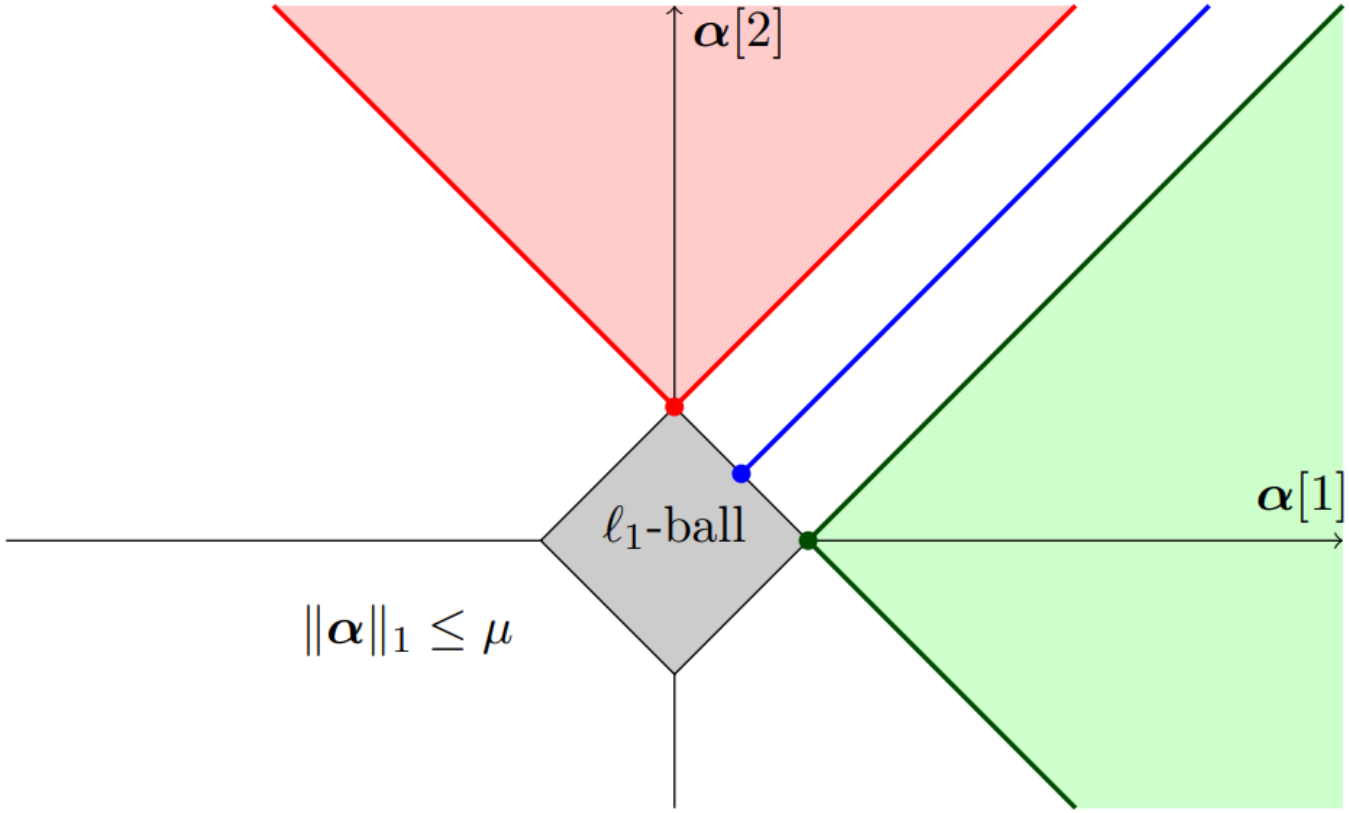
- We use the BIC:

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

- $\widehat{df}_\lambda$ 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 $c = \log(\log(n)) * \log(n)$

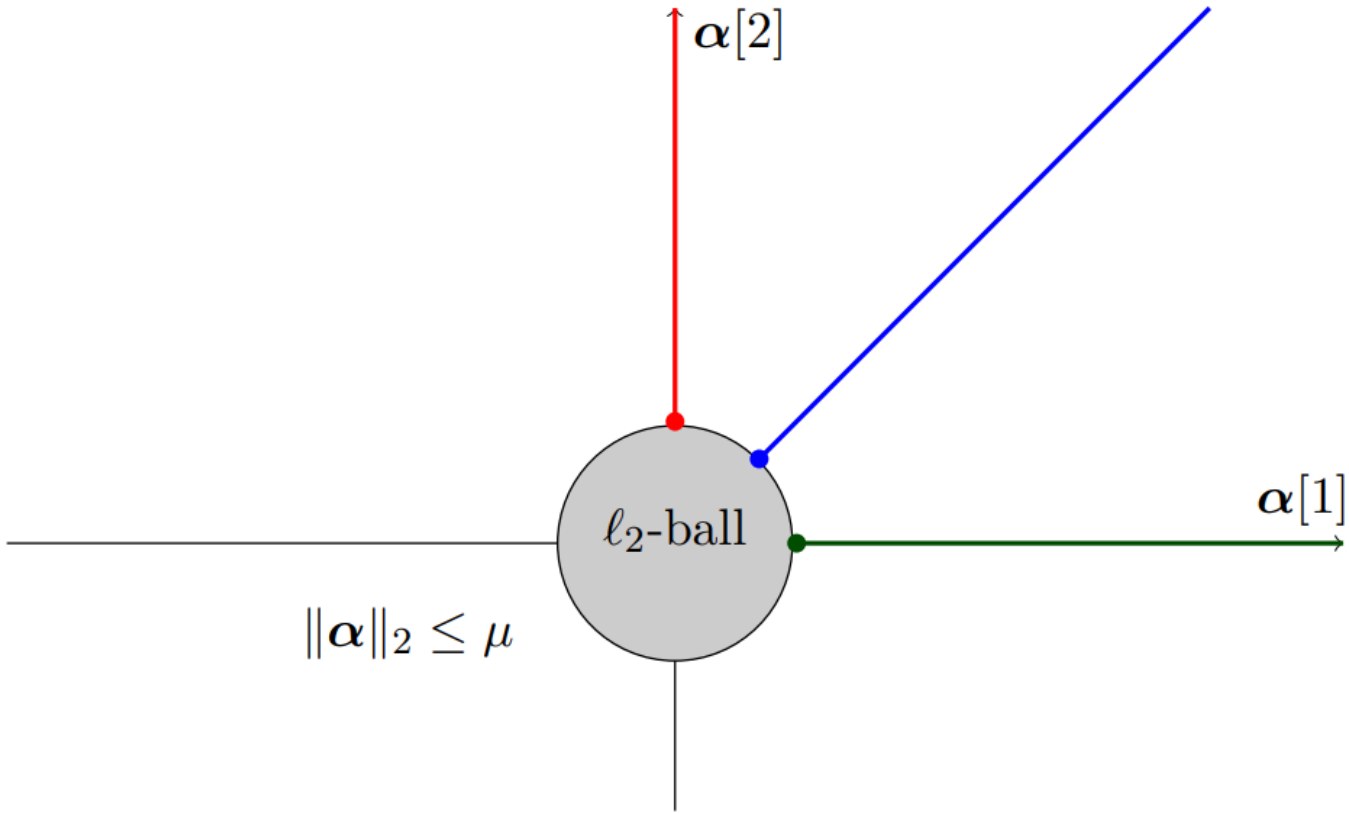
[1] Zou et al. The Annals of Statistics, (2007)
[2] Bondell et al. Biometrics (2010)
[3] Wang et al. JRSS(Ser. B), (2009)

Effect of the Euclidean projection onto the ℓ_1-ball



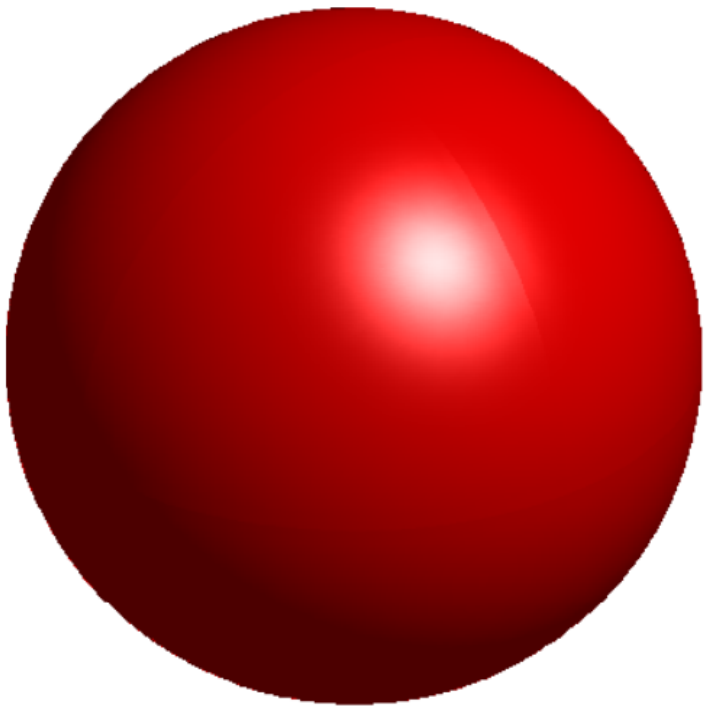
[1] Mairal, Bach and Ponce (2012). Sparse Modeling for Image and Vision Processing.

Effect of the Euclidean projection onto the ℓ_2-ball

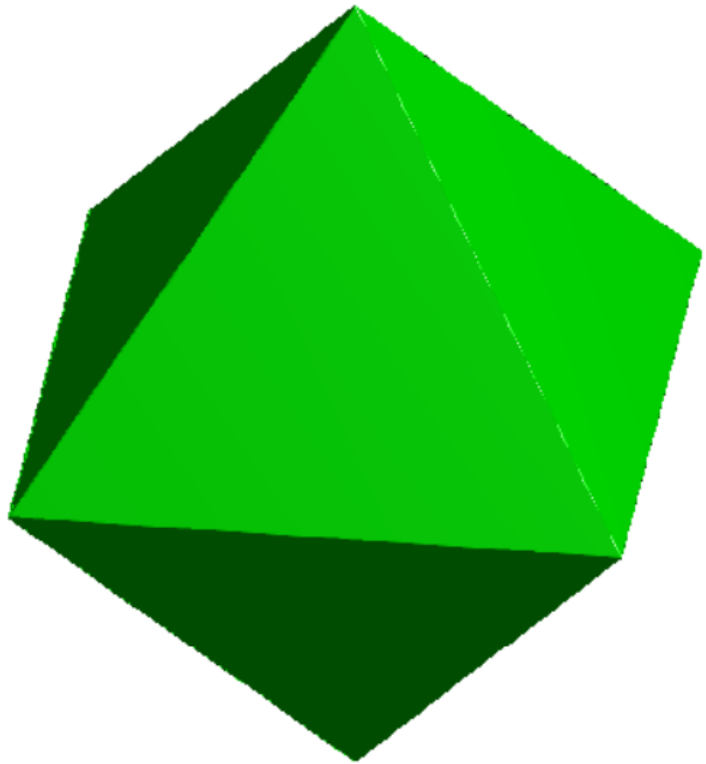


[1] Mairal, Bach and Ponce (2012). Sparse Modeling for Image and Vision Processing.

Representation in three dimensions of the ℓ_1- and ℓ_2-balls



(a) ℓ_2-ball in 3D



(b) ℓ_1-ball in 3D

[1] Mairal, Bach and Ponce (2012). Sparse Modeling for Image and Vision Processing.

Simulations

Theory

Algorithm

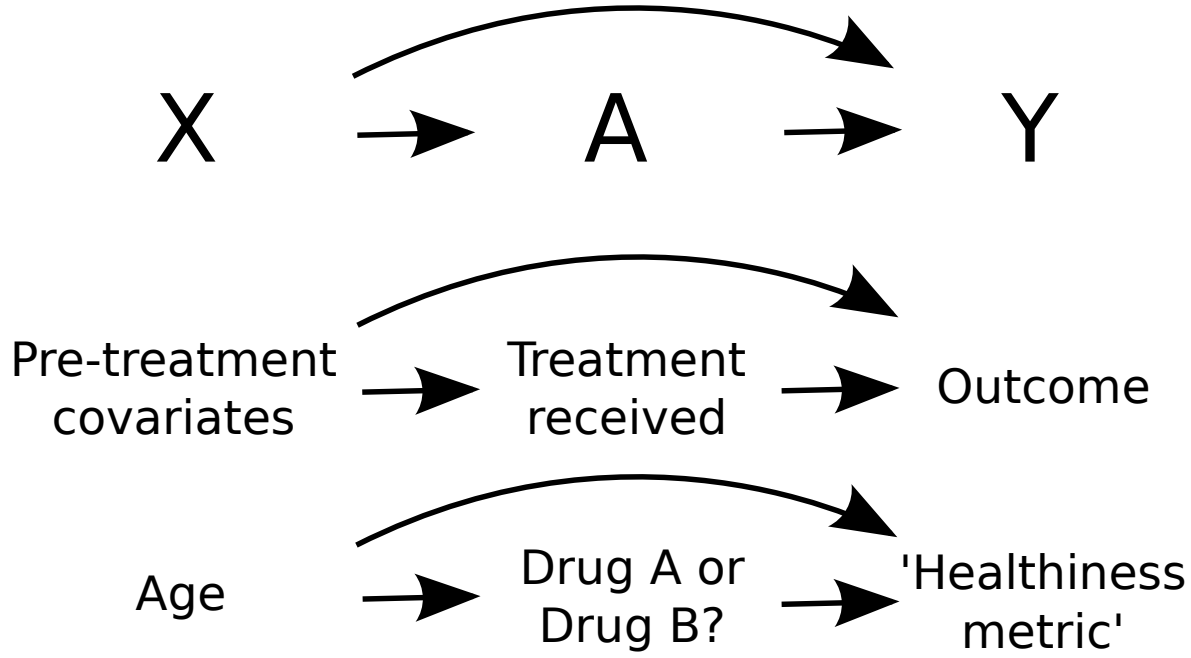
Discussion

ggmix appendix

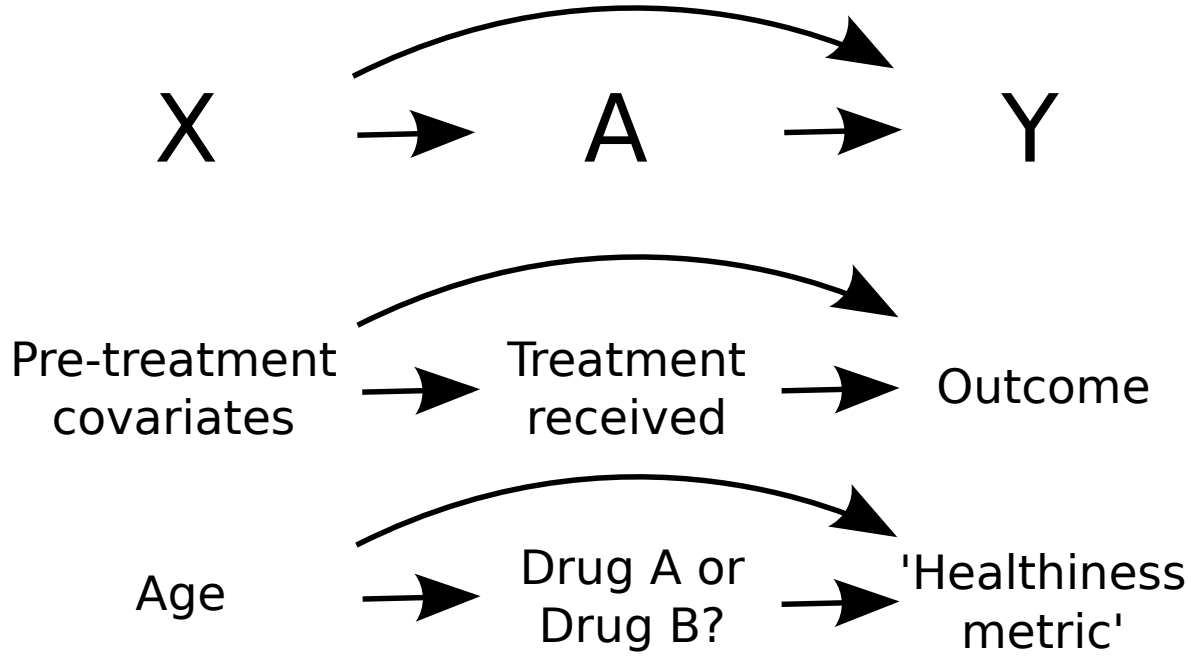
DTR

Dynamic Treatment Regimes

Dynamic Treatment Regimes (DTRs)

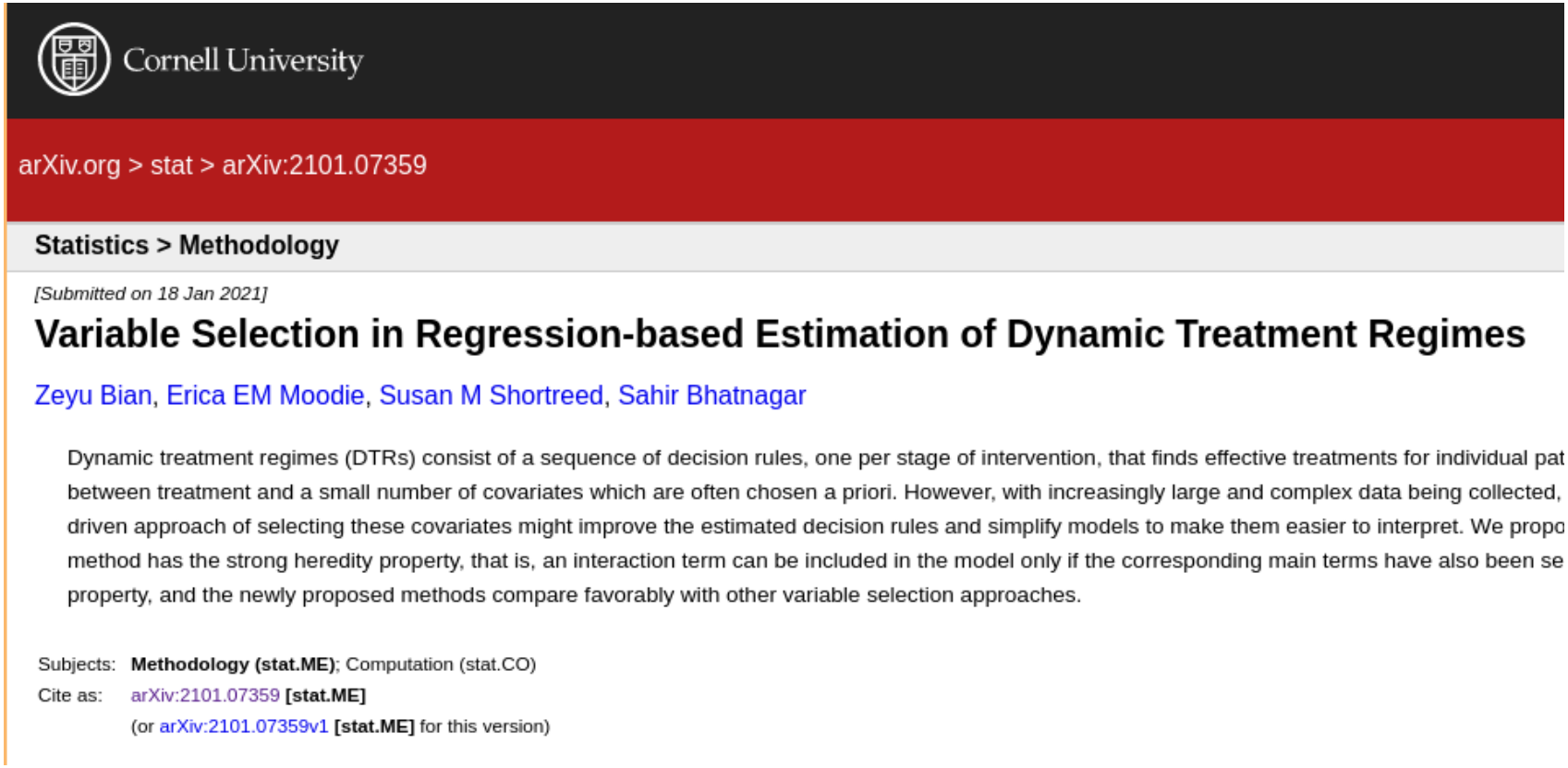


Dynamic Treatment Regimes (DTRs)



$$\mathbb{E}[Y \mid \mathbf{X}, A; \boldsymbol{\psi}, \boldsymbol{\beta}] = \underbrace{\mathbf{X}\boldsymbol{\beta}}_{\substack{\text{Impact of patient history} \\ \text{in the absence of treatment}}} + \underbrace{\psi_0 A + \boldsymbol{\psi} A \mathbf{X}}_{\substack{\text{Impact of treatment} \\ \text{on outcome}}}$$

Extension of `sail` to DTRs



Cornell University

arXiv.org > stat > arXiv:2101.07359

Statistics > Methodology

[Submitted on 18 Jan 2021]

Variable Selection in Regression-based Estimation of Dynamic Treatment Regimes

Zeyu Bian, Erica EM Moodie, Susan M Shortreed, Sahir Bhatnagar

Dynamic treatment regimes (DTRs) consist of a sequence of decision rules, one per stage of intervention, that finds effective treatments for individual pat between treatment and a small number of covariates which are often chosen a priori. However, with increasingly large and complex data being collected, driven approach of selecting these covariates might improve the estimated decision rules and simplify models to make them easier to interpret. We propo method has the strong heredity property, that is, an interaction term can be included in the model only if the corresponding main terms have also been se property, and the newly proposed methods compare favorably with other variable selection approaches.

Subjects: **Methodology (stat.ME)**; Computation (stat.CO)
Cite as: arXiv:2101.07359 **[stat.ME]**
(or arXiv:2101.07359v1 **[stat.ME]** for this version)

[1] *In press at Biometrics.* https://arxiv.org/abs/2101.07359