HEC MONTRÉAL

Betting on Sparsity

Sahir Bhatnagar, PhD Candidate, McGill Biostatistics

Joint work with Karim Oualkacha, Yi Yang and Celia Greenwood

November 16, 2017

Introduction

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)
- 2017: Chargé de cours théorie des probabilités et l'inférence statistique (McGill)

Mon cheminement

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)
- 2017: Chargé de cours théorie des probabilités et l'inférence statistique (McGill)
- sahirbhatnagar.com

Outline

1. A motivating example

Outline

1. A motivating example
2. Background on penalization methods lasso and group lasso

Outline

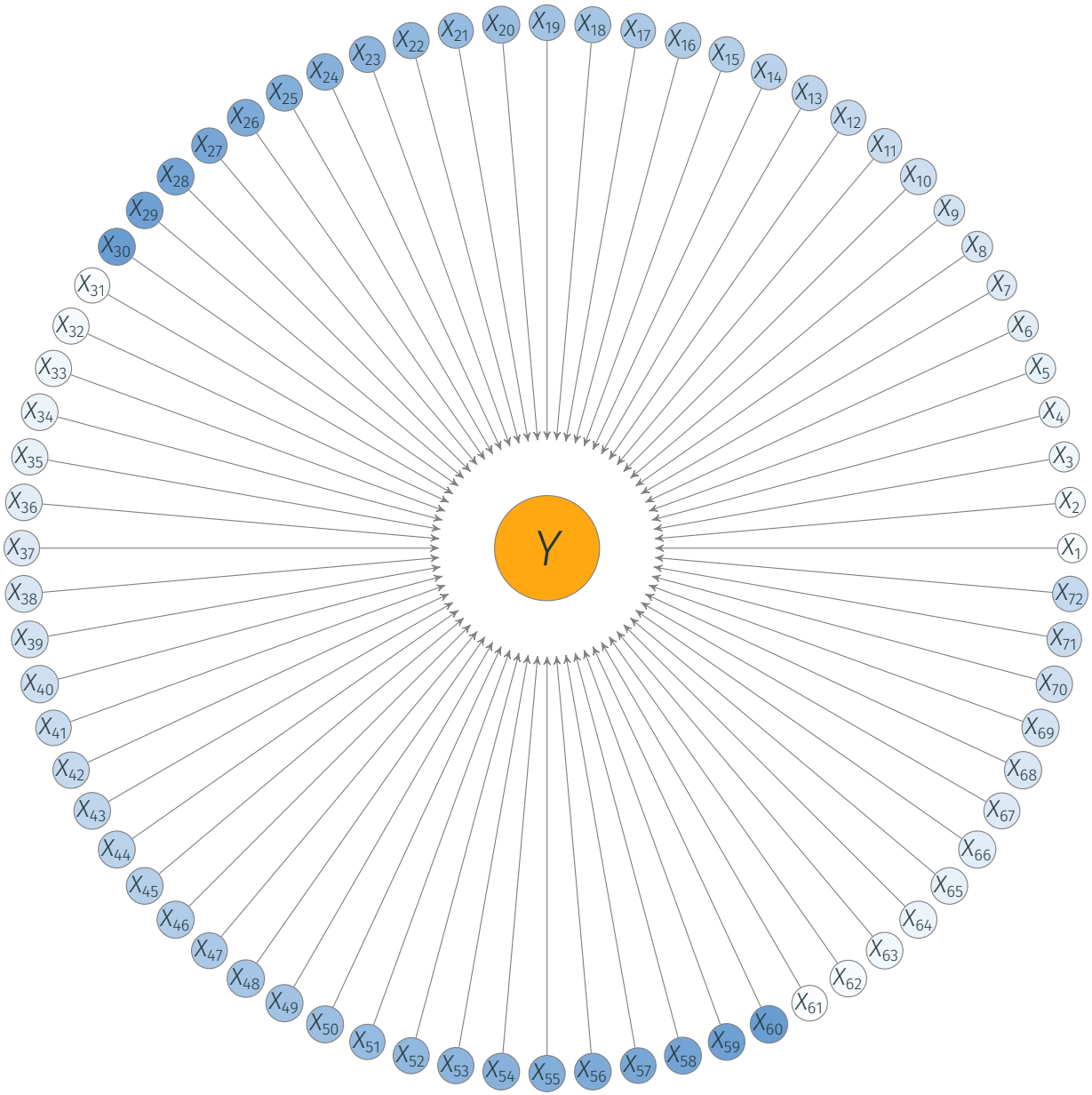
1. A motivating example
2. Background on penalization methods lasso and group lasso
3. Overview of our software packages

Outline

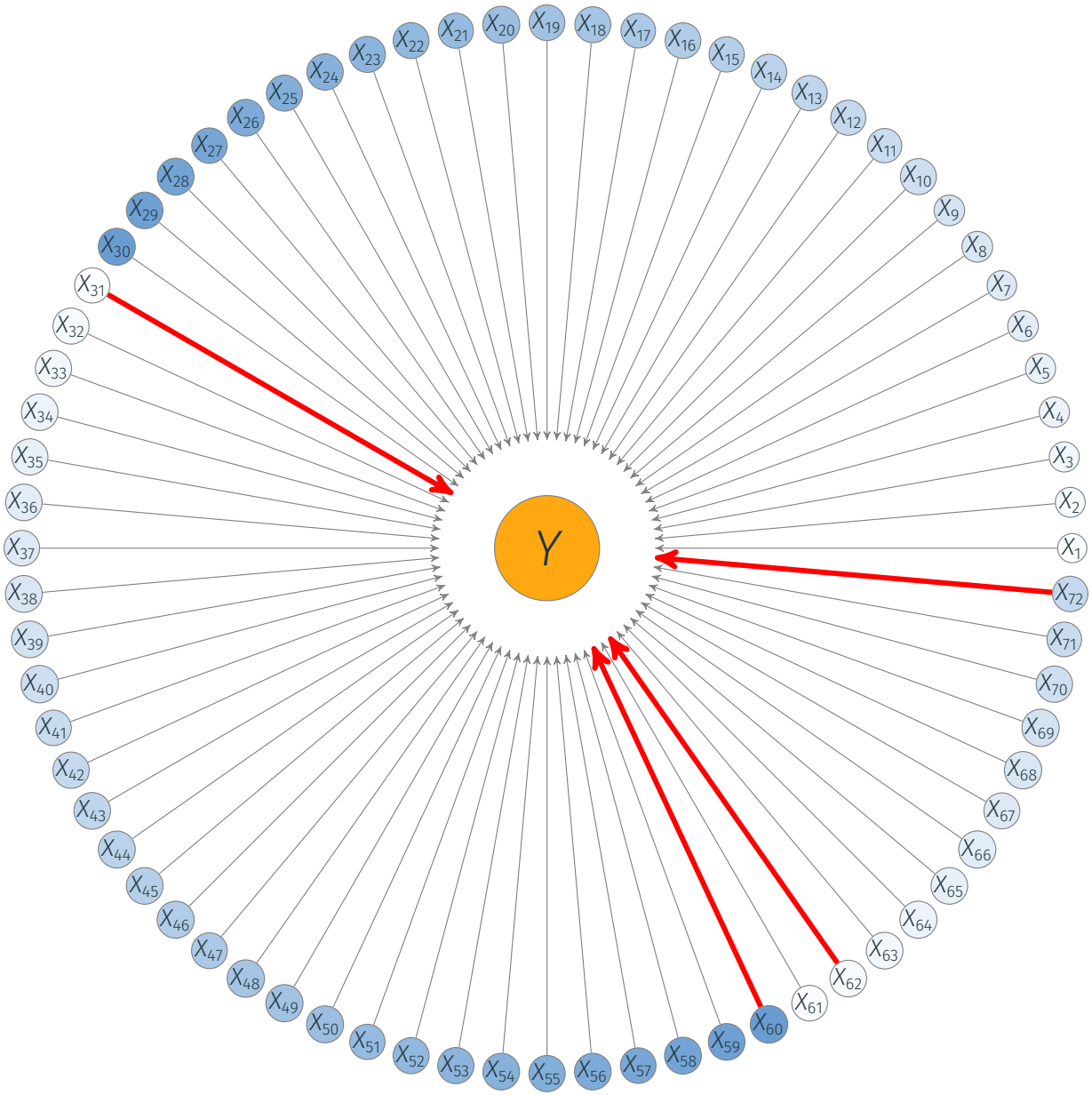
1. A motivating example
2. Background on penalization methods lasso and group lasso
3. Overview of our software packages
4. Present two of our penalization methods: `sail` and `ggmix`

Bet on Sparsity

Bet on Sparsity Principle



Bet on Sparsity Principle



Bet on Sparsity Principle

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

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

Bet on Sparsity Principle

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

- An underlying assumption of **simplicity** in high-dimensional data ($N << p$)
- A sparse statistical model is one in which only a **relatively small number of predictors** $k < N$ play an important role
- Sparse models can be faster to compute, easier to understand, and yield more stable predictions.

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

Motivating Example

Predictors of NHL Salary[2]

[2] https://www.kaggle.com/camnugent/nhl-salary-data-prediction-cleaning-and-modelling

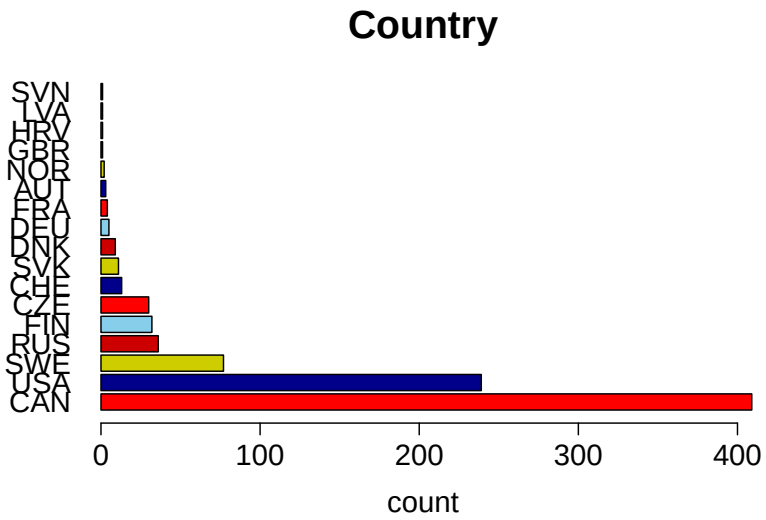
Supervised Learning

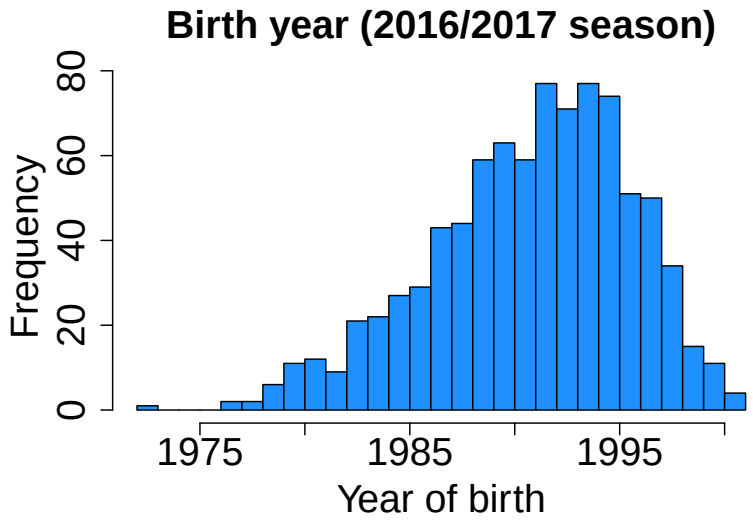
- Learn the function f

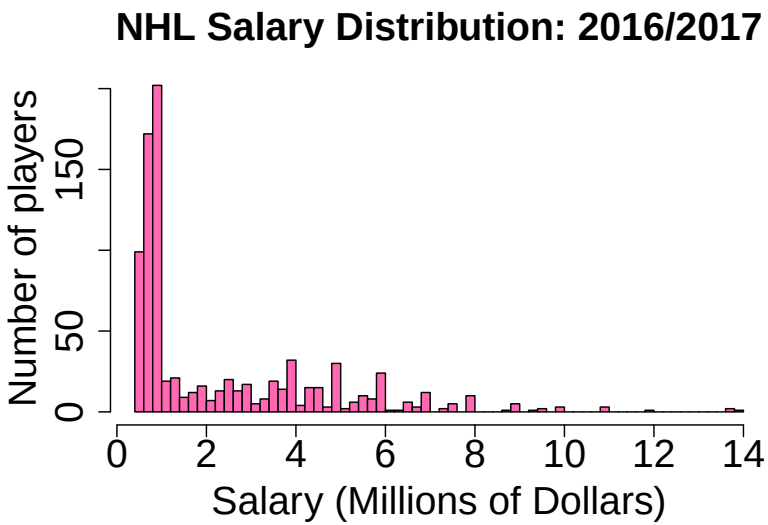
f 

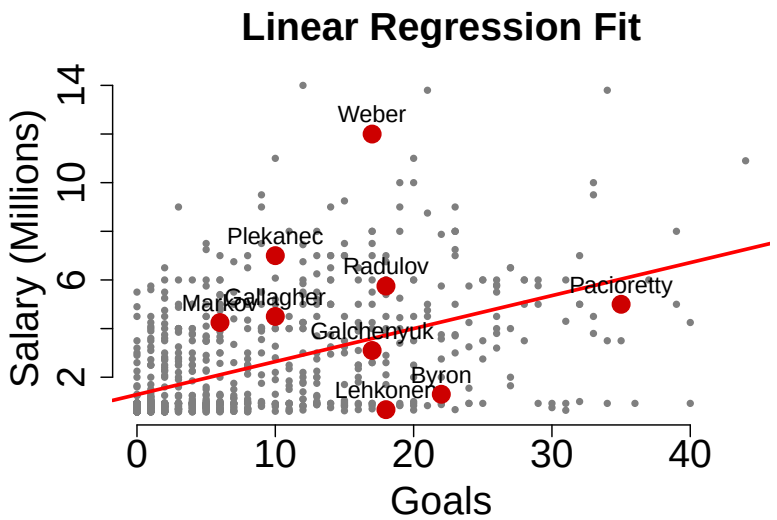
=


Predictors of NHL Salary

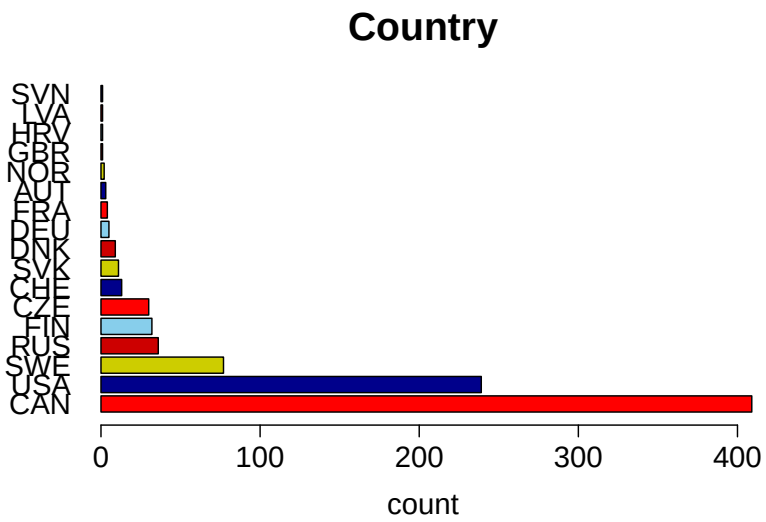


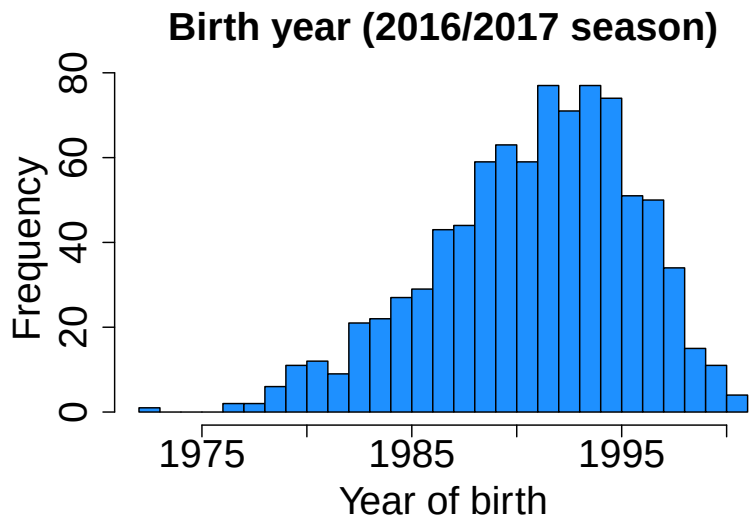


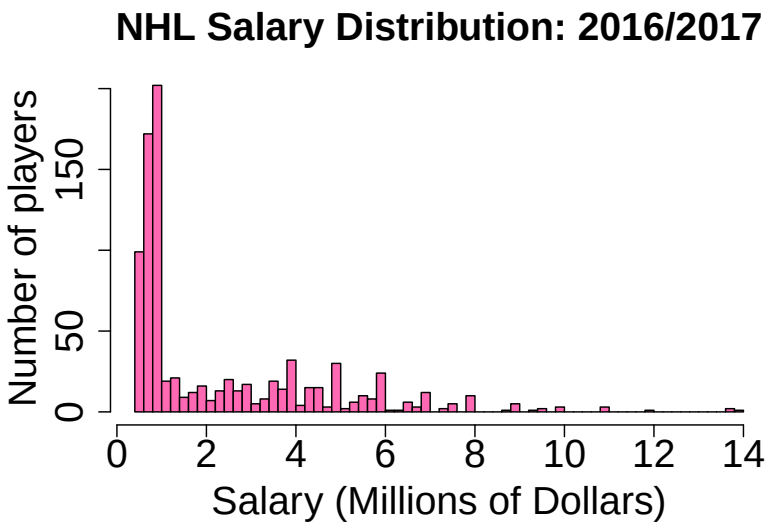


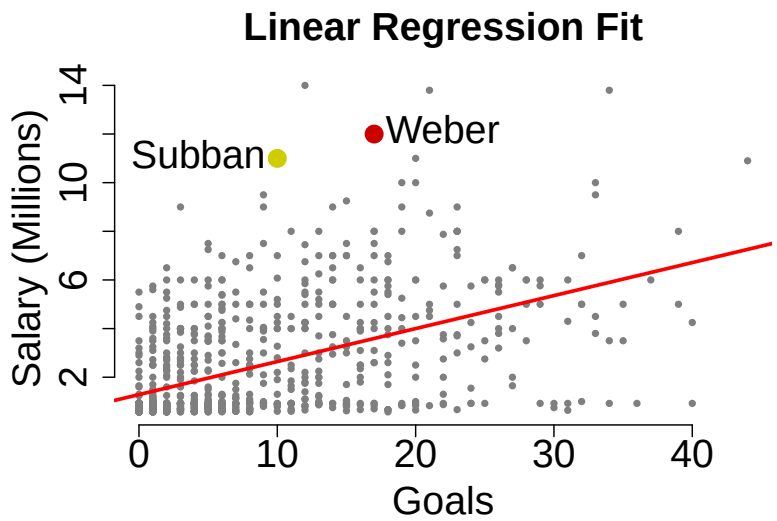


Predictors of NHL Salary

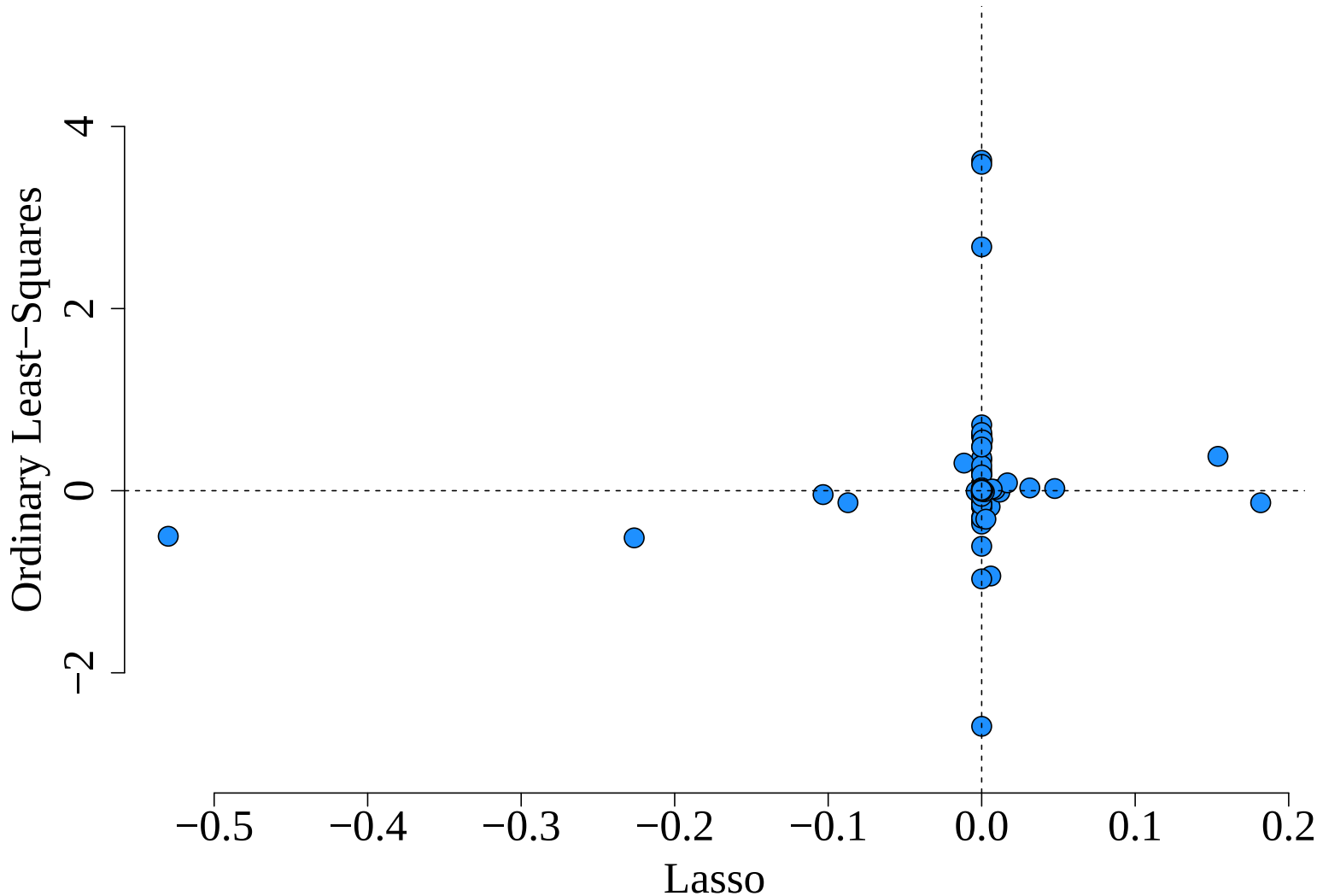




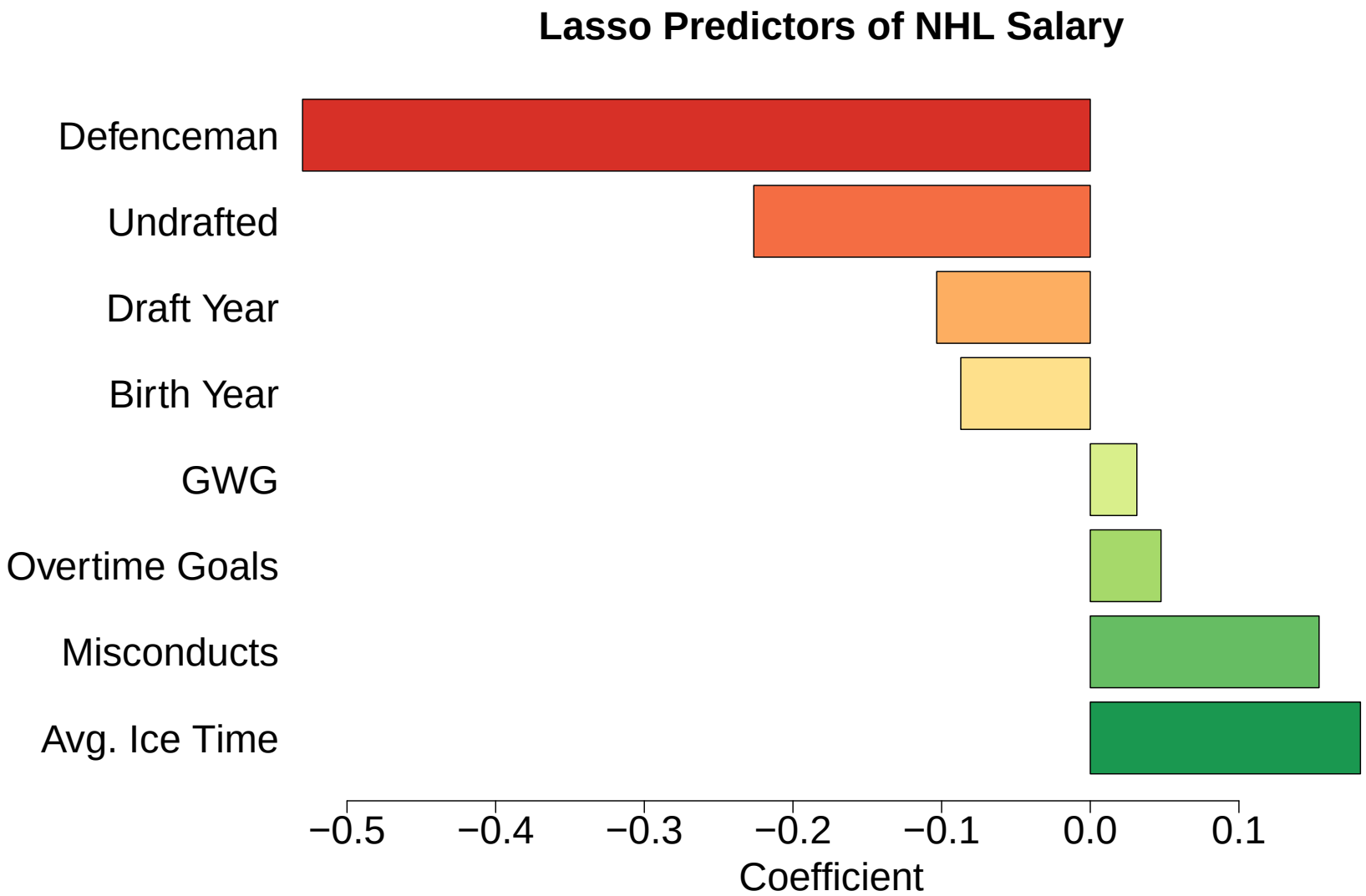




OLS vs. Lasso Coefficients



Lasso Selected Predictors



Background on the Lasso

- Predictors x_{ij}, $j = 1, \ldots, p$ and outcome values y_i for the ith observation, $i = 1, \ldots, n$
- Assume x_{ij} are standardized so that $\sum_i x_{ij}/n = 0$ and $\sum_i x_{ij}^2 = 1$.

[1] Tibshirani. JRSSB (1996)

Background on the Lasso

- Predictors x_{ij}, $j = 1, \ldots, p$ and outcome values y_i for the ith observation, $i = 1, \ldots, n$
- Assume x_{ij} are standardized so that $\sum_i x_{ij}/n = 0$ and $\sum_i x_{ij}^2 = 1$. The lasso[1] solves

$$\widehat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s, \qquad s > 0$$

[1] Tibshirani. JRSSB (1996)

Background on the Lasso

- Predictors x_{ij}, $j = 1, \ldots, p$ and outcome values y_i for the ith observation, $i = 1, \ldots, n$
- Assume x_{ij} are standardized so that $\sum_i x_{ij}/n = 0$ and $\sum_i x_{ij}^2 = 1$. The lasso[1] solves

$$\widehat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s, \qquad s > 0$$

- Equivalently, the Lagrange version of the problem, for $\lambda > 0$

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

[1] Tibshirani. JRSSB (1996)

Inspection of the Lasso Solution

- Consider a single predictor setting based on the observed data $\{(x_i, y_i)\}_{i=1}^n$. The problem then is to solve

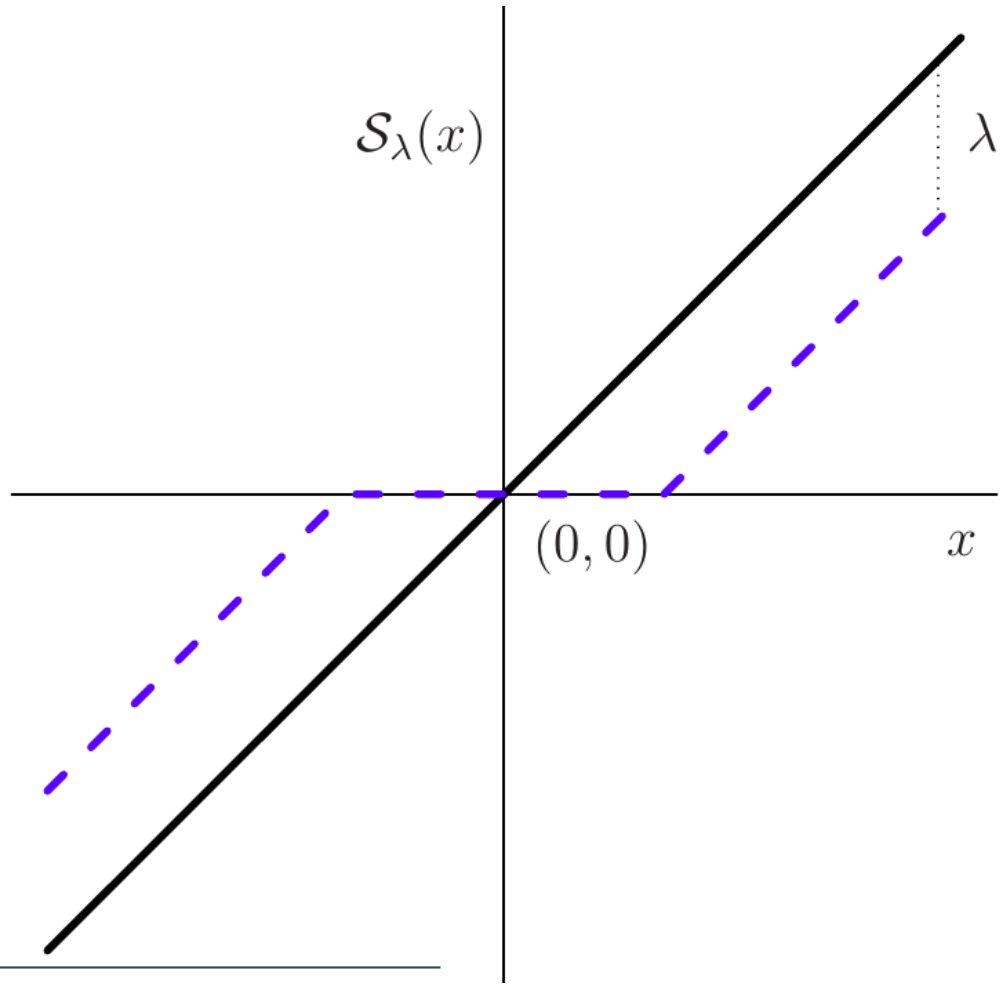
$$\widehat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{1}$$

- With a standardized predictor, the lasso solution (1) is a **soft-thresholded** version of the least-squares (LS) estimate $\widehat{\beta}^{LS}$

$$\begin{aligned}
\widehat{\beta}^{lasso} &= S_\lambda\left(\widehat{\beta}^{LS}\right) = \text{sign}\left(\widehat{\beta}^{LS}\right)\left(|\widehat{\beta}^{LS}| - \lambda\right)_+ \\
&= \begin{cases} \widehat{\beta}^{LS} - \lambda, & \widehat{\beta}^{LS} > \lambda \\ 0 & |\widehat{\beta}^{LS}| \leq \lambda \\ \widehat{\beta}^{LS} + \lambda & \widehat{\beta}^{LS} \leq -\lambda \end{cases}
\end{aligned}$$

Inspection of the Lasso Solution

- When the data are standardized, the lasso solution **shrinks the LS estimate toward zero** by the amount λ



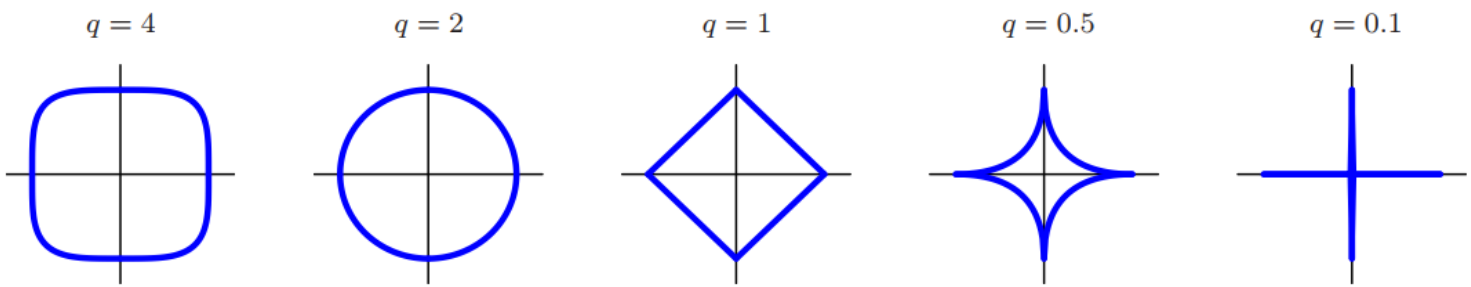
[1] Hastie et al. Statistical learning with sparsity: the lasso and generalizations. CRC press, (2015).

Why the L1 norm ?

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

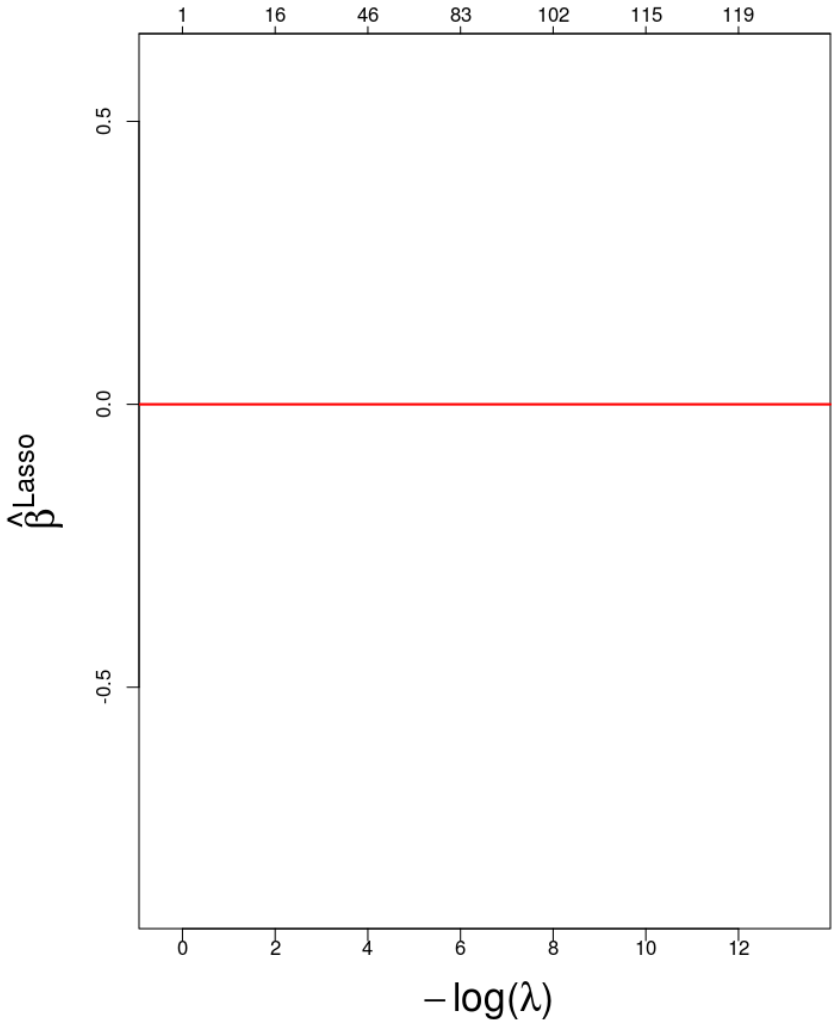
$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

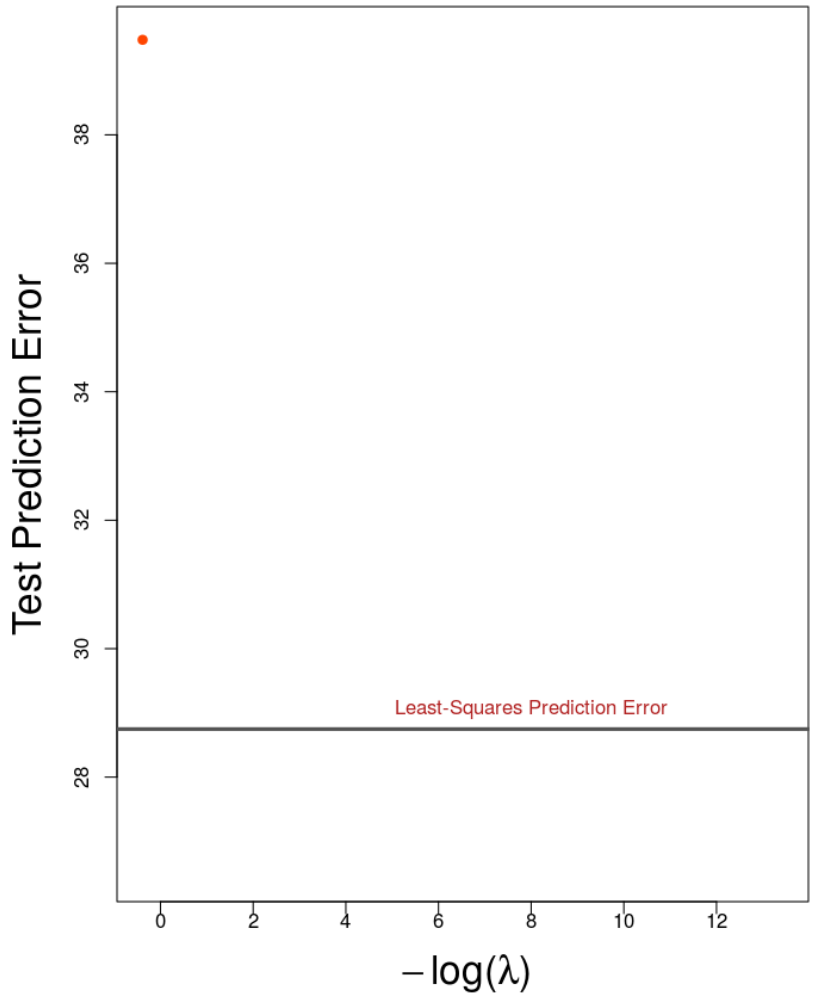
- Why do we use the ℓ_1 norm? Why not use the $q = 2$ (Ridge) or any ℓ_q norm?



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

Choosing the Model Complexity

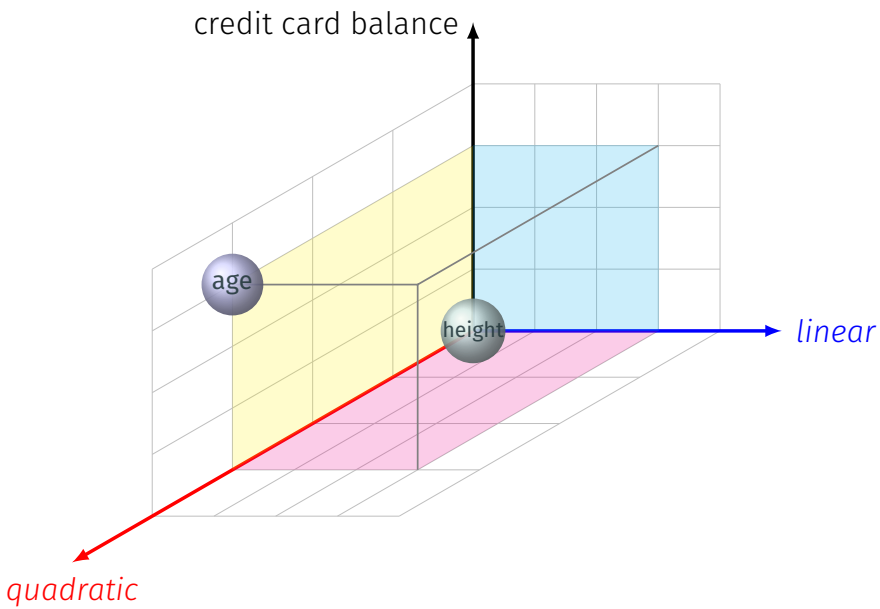




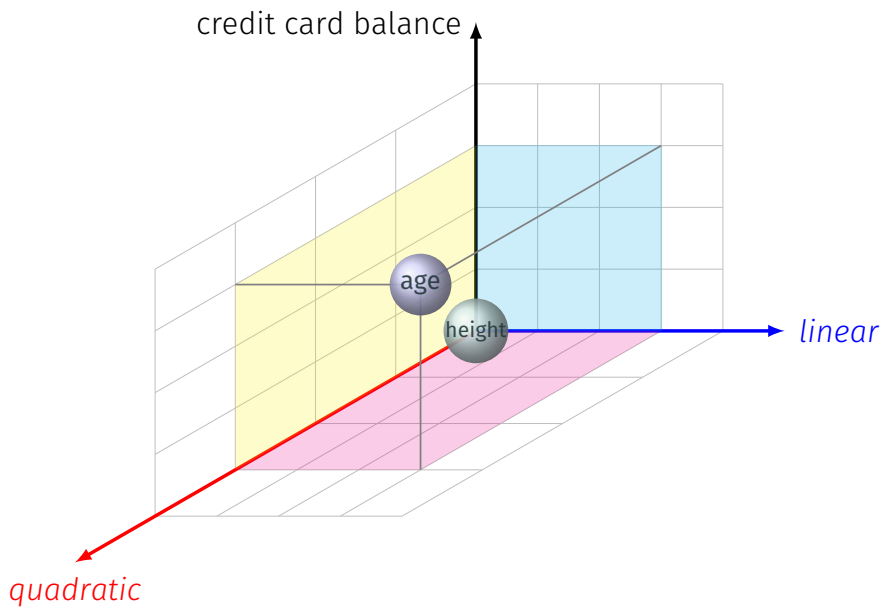
Group Lasso

Extended from the lasso penalty, the **group lasso** estimator is:

$$\min_{(\beta_0,\boldsymbol{\beta})} \frac{1}{2} \|\mathbf{y} - \beta_0 - \mathbf{x}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} \|\boldsymbol{\beta}^{(k)}\|_2 \qquad p_k - \text{group size}$$



(a) Lasso



(b) Group Lasso

Our Software

Overview of Our Software Packages

- **eclust** – Bhatnagar et al. (2017, Genetic Epidemiology)
  https://cran.r-project.org/package=eclust
- **sail** – Bhatnagar, Yang and Greenwood (2017+, preprint)
  https://github.com/sahirbhatnagar/sail
- **ggmix** – Bhatnagar, Oualkacha, Yang, Greenwood (2017+, preprint)
  https://github.com/sahirbhatnagar/ggmix
- **casebase** – Bhatnagar[1], Turgeon[1], Yang, Hanley and Saarela (2017+, preprint)
  https://cran.r-project.org/package=casebase

[1] joint co-authors

Overview of Our Software Packages

	`eclust`	`sail`	`ggmix`	`casebase`
Model				
Least-Squares	✓	✓	✓	
Binary Classification	✓			
Survival Analysis				✓
Penalty				
Ridge	✓		✓	✓
Lasso	✓	✓	✓	✓
Elastic Net	✓		✓	✓
Group Lasso		✓	✓	
Feature				
Interactions	✓	✓		✓
Flexible Modeling	✓	✓		✓
Random Effects			✓	
Data	(x, y, e)	(x, y, e)	(x, y, Ψ)	(x, t, δ)

sail: Strong Additive Interaction Learning

Motivation 1: Heredity Property

$$Y = \beta_0 \cdot 1 + \sum_{j=1}^{p} \beta_j X_j + \beta_E X_E + \sum_{j=1}^{p} \alpha_j X_E X_j + \varepsilon$$

[1] Chipman. Canadian Journal of Statistics (1996)

[2] McCullagh and Nelder. Generalized Linear Models (1983)

[3] Cox. International Statistical Review (1984)

Motivation 1: Heredity Property

$$Y = \beta_0 \cdot 1 + \sum_{j=1}^{p} \beta_j X_j + \beta_E X_E + \sum_{j=1}^{p} \alpha_j X_E X_j + \varepsilon$$

Strong Heredity[1]

$$\hat{\alpha}_j \neq 0 \quad \Rightarrow \quad \hat{\beta}_j \neq 0 \quad \text{and} \quad \hat{\beta}_E \neq 0$$

Weak Heredity[1]

$$\hat{\alpha}_j \neq 0 \quad \Rightarrow \quad \hat{\beta}_j \neq 0 \quad \text{or} \quad \hat{\beta}_E \neq 0$$

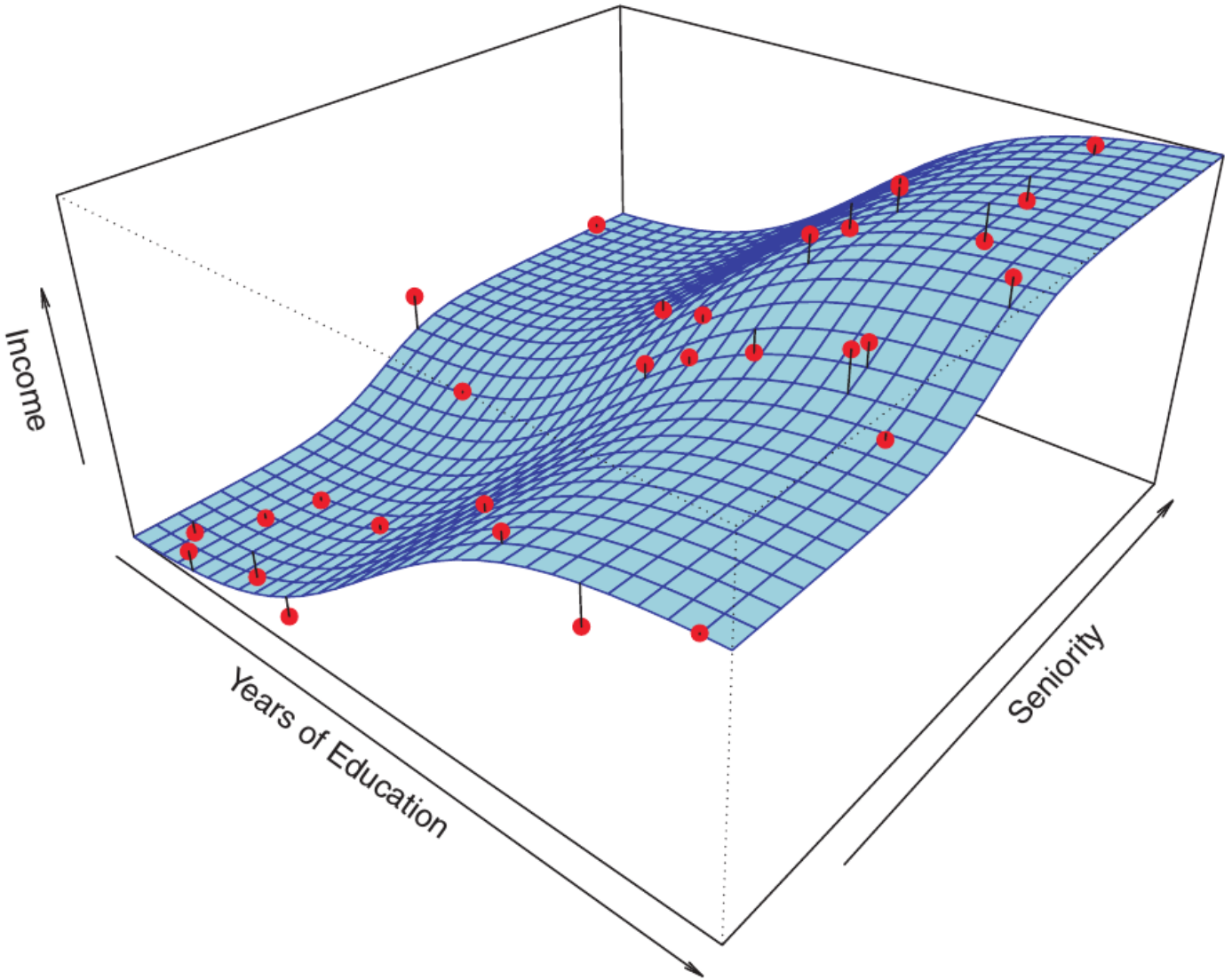
- Heredity property is desired for the purposes of **interpretability**[2]
- Large main effects are more likely to lead to appreciable interactions[3]

[1] Chipman. Canadian Journal of Statistics (1996)

[2] McCullagh and Nelder. Generalized Linear Models (1983)

[3] Cox. International Statistical Review (1984)

Motivation 2: Non-linear Interactions



[1] James et al. Introduction to Statistical Learning (2013)

Lasso interaction model

- $Y \longrightarrow$ response
- $X_E \longrightarrow$ environment
- $X_j \longrightarrow$ predictors, $j = 1, \ldots, p$

$$Y = \beta_0 \cdot 1 + \sum_{j=1}^{p} \beta_j X_j + \beta_E X_E + \sum_{j=1}^{p} \alpha_j X_E X_j + \varepsilon$$

$$\underset{\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}}{\operatorname{argmin}} \quad \mathcal{L}(Y; \Theta) + \lambda(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\alpha}\|_1)$$

Strong Heredity Interactions: Current State of the Art

Type	Model	Software
Linear	`CAP` (Zhao et al. 2009, *Ann. Stat*)	✗
	`SHIM` (Choi et al. 2009, *JASA*)	✗
	`hiernet` (Bien et al. 2013, *Ann. Stat*)	`hierNet(x, y)`
	`GRESH` (She and Jiang 2014, *JASA*)	✗
	`FAMILY` (Haris et al. 2014, *JCGS*)	`FAMILY(x, z, y)`
	`glinternet` (Lim and Hastie 2015, *JCGS*)	`glinternet(x, y)`
	`RAMP` (Hao et al. 2016, *JASA*)	`RAMP(x, y)`
Non-linear	`VANISH` (Radchenko and James 2010, *JASA*)	✗
	`sail` (Bhatnagar et al. 2017+)	`sail(x, e, y)`

Our Extension to Nonlinear Effects

Consider the basis expansion

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

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

sail: Additive Interactions

- $\theta_j = (\beta_{j1}, \ldots, \beta_{jp_j}) \in \mathbb{R}^{p_j}$
- $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jp_j}) \in \mathbb{R}^{p_j}$
- $\Psi_j \to n \times p_j$ matrix of evaluations of the $\psi_{j\ell}$
- In our implementation, we use `bsplines` with 5 degrees of freedom

Model

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \Psi_j \theta_j + \beta_E X_E + \sum_{j=1}^{p} X_E \Psi_j \alpha_j + \varepsilon$$

sail: Strong Heredity

Reparametrization[1]

$$\alpha_j = \gamma_j \beta_E \theta_j$$

Model

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^{p} \Psi_j \theta_j + \beta_E X_E + \sum_{j=1}^{p} \gamma_j \beta_E X_E \Psi_j \theta_j + \varepsilon$$

Objective Function

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

[1] Choi et al. JASA (2010)

Algorithm

Block Relaxation (De Leeuw, 1994)

Algorithm 1: Block Relaxation Algorithm

Set the iteration counter $k \leftarrow 0$, initial values for the parameter vector $\Theta^{(0)}$;

for *each pair* $(\lambda_\beta, \lambda_\gamma)$ **do**

  repeat

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

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

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

    $k \leftarrow k + 1$

  until *convergence criterion is satisfied*;

end

Implementation

Objective Function

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

[1] https://github.com/sahirbhatnagar/sail

Implementation

Objective Function

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

Lasso problem

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

[1] https://github.com/sahirbhatnagar/sail

Implementation

Objective Function

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

[1] https://github.com/sahirbhatnagar/sail

Implementation

Objective Function

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

Group Lasso problem

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

[1] https://github.com/sahirbhatnagar/sail

Simulations

Simulations Scenarios

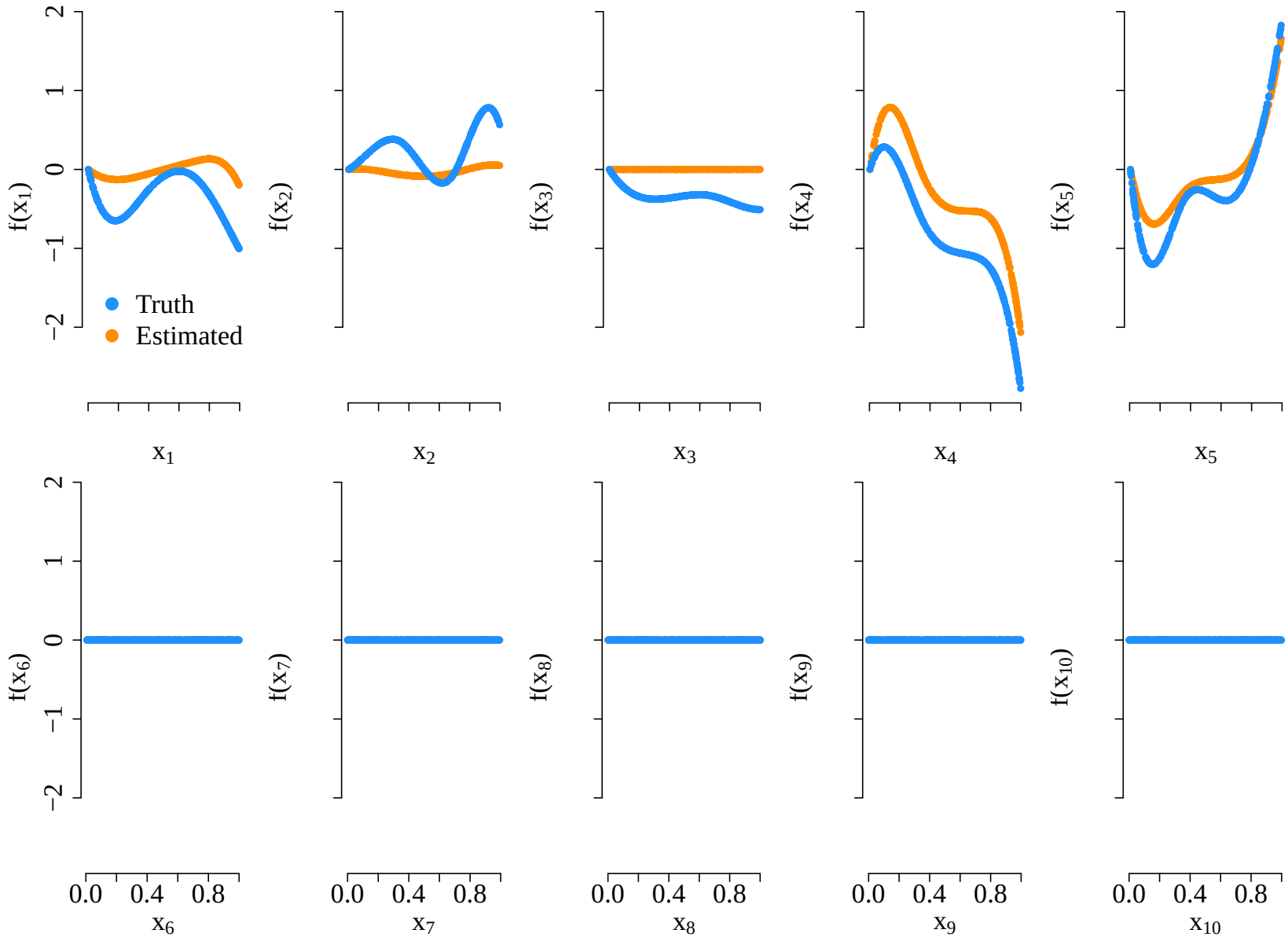
Scenario 1: Easy

- $Y = \sum_{j=1}^{5} f(X_j) + X_E + E \times (f(X_1) + f(X_2))$
- $f(\cdot) \rightarrow$ B-splines with 5 df
- $\theta_j \sim \mathcal{N}(0, 1)$
- $N = 400, p = 50$
- $50 \times 5 \times 2 + 1 = 501$ parameters to estimate

Scenario 2: Hard

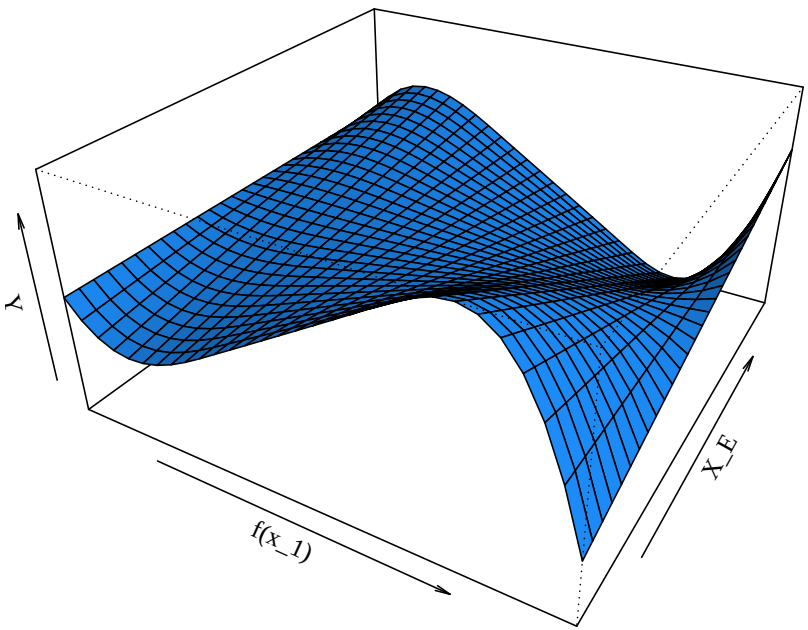
- $Y = \sum_{j=1}^{4} f(X_j) + X_E + E \times (f(X_3) + f(X_4))$
- $f(X_1) \rightarrow$ linear
- $f(X_2) \rightarrow$ quadratic
- $f(X_3) \rightarrow$ sinusoidal
- $f(X_4) \rightarrow$ complicated sinusoidal

Scenario 1: Main Effects

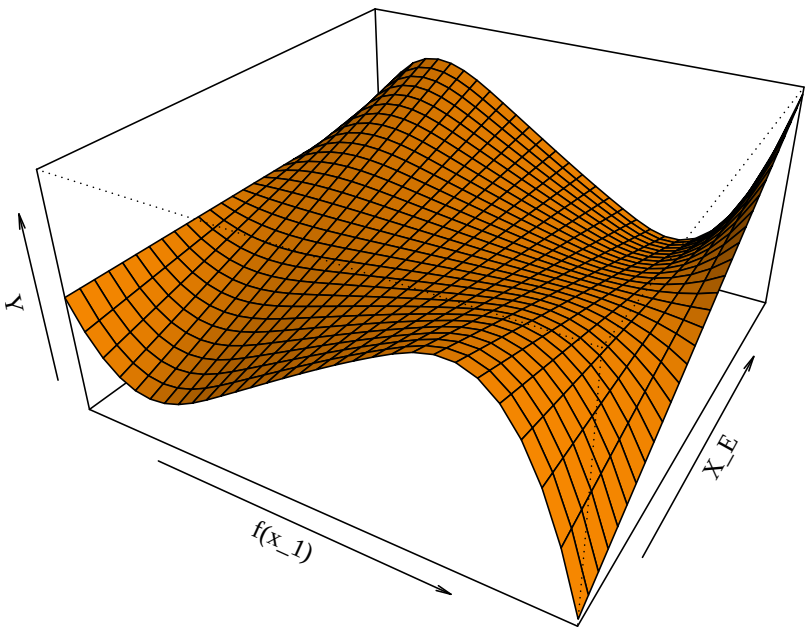


Scenario 1: Interaction Effects

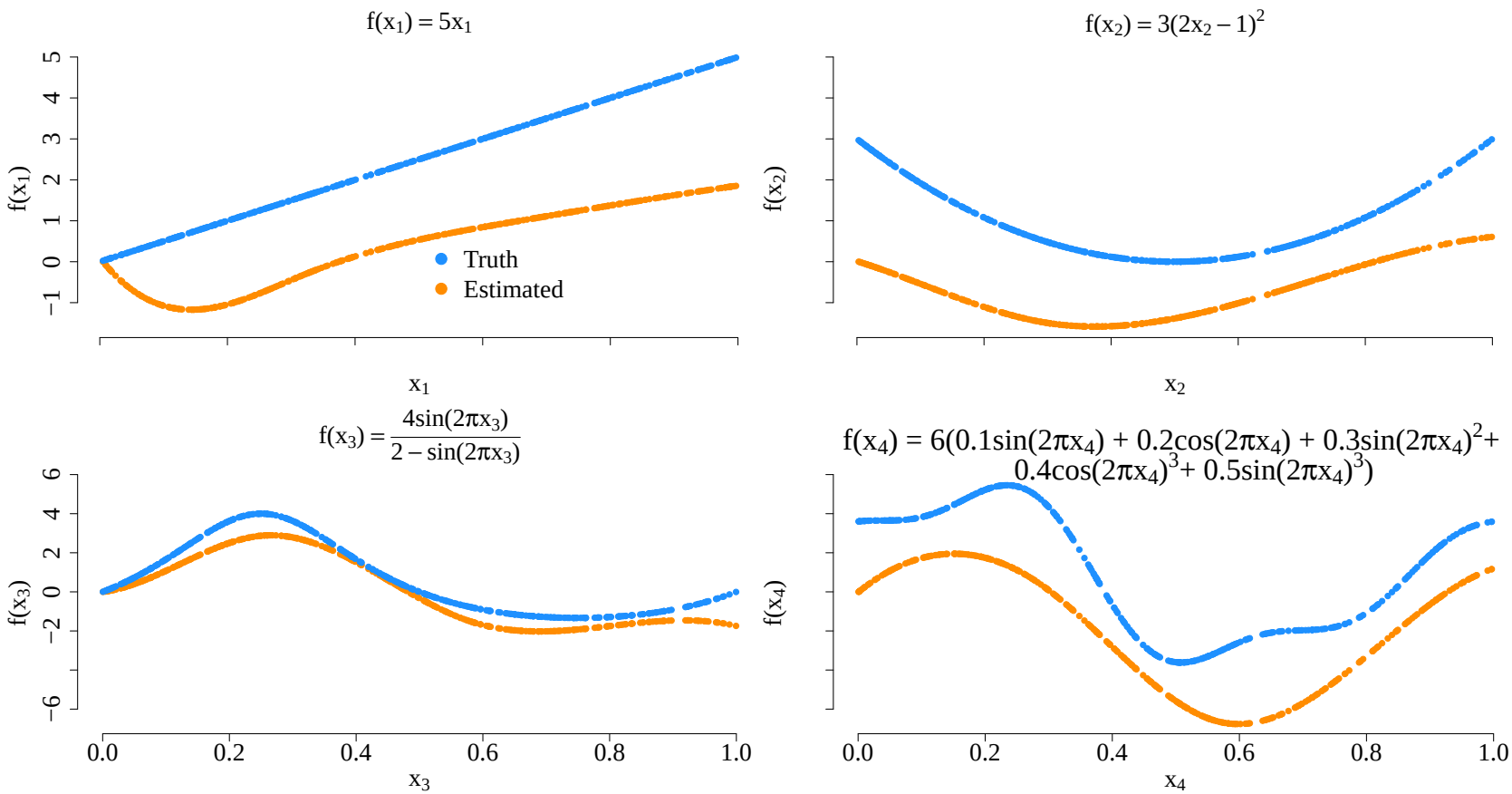
Truth



Estimated X_E*f(X_1)

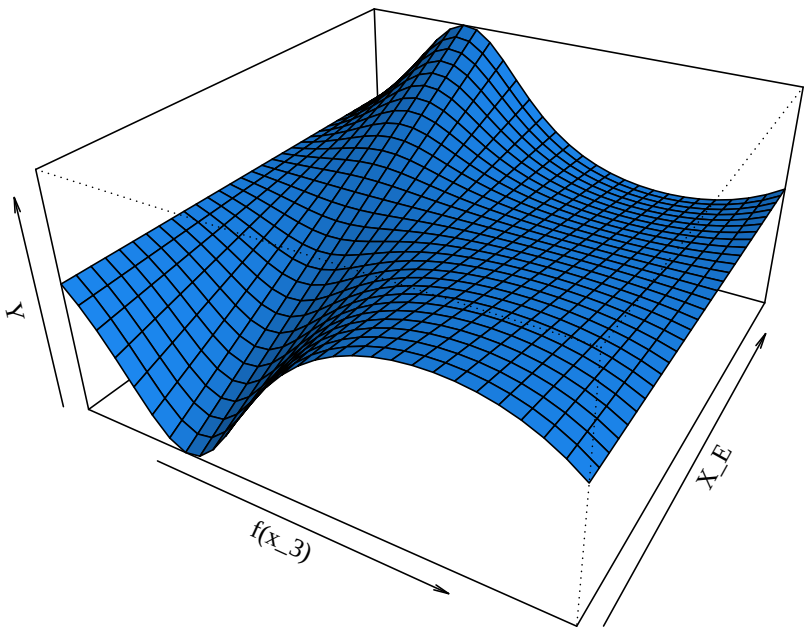


Scenario 2: Main Effects

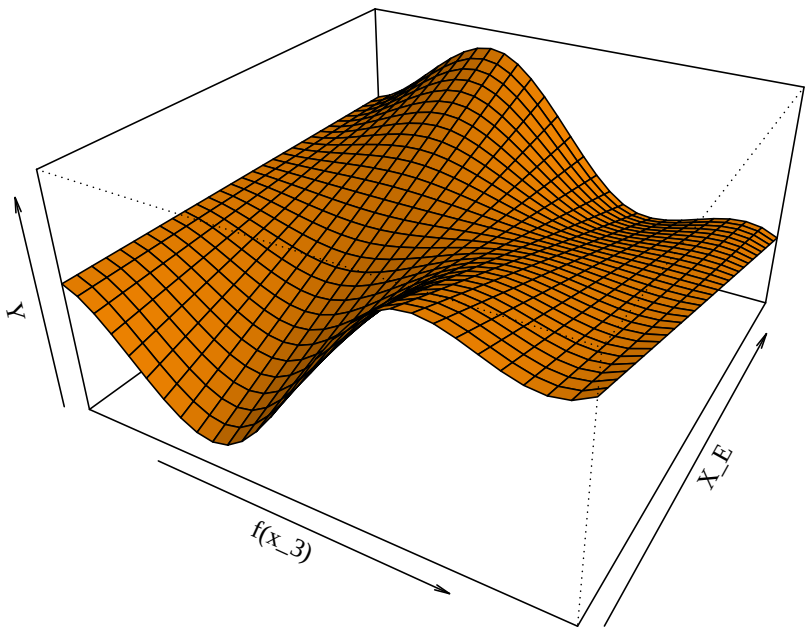


Scenario 2: Interaction Effects

Truth

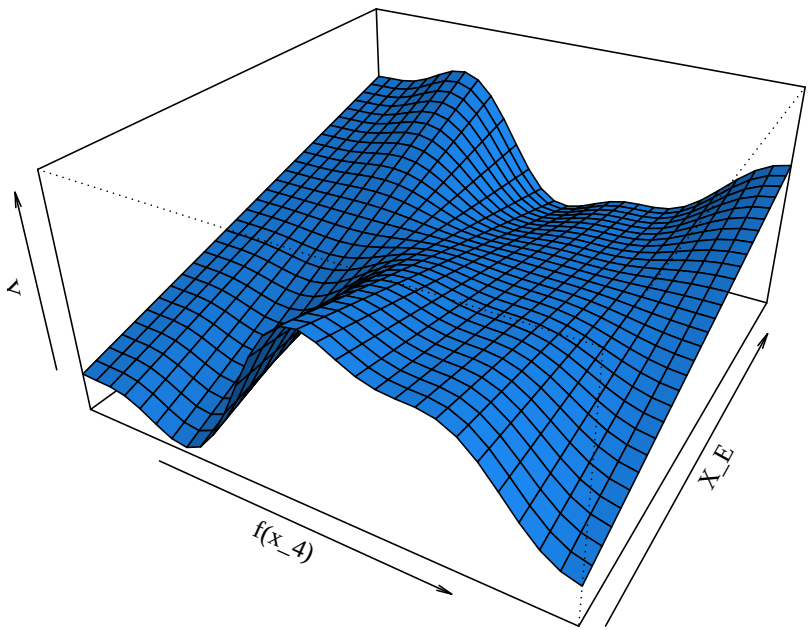


Estimated X_E*f(X_3)

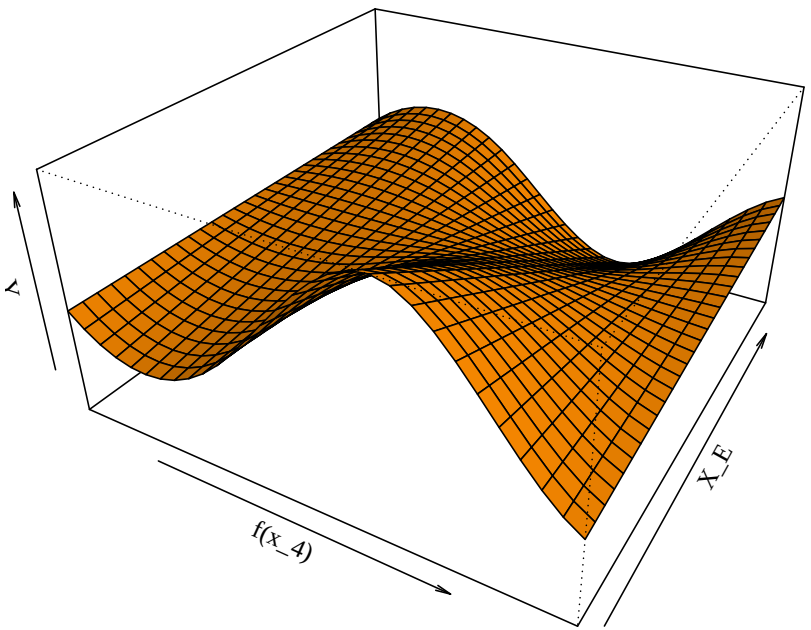


Scenario 2: Interaction Effects

Truth

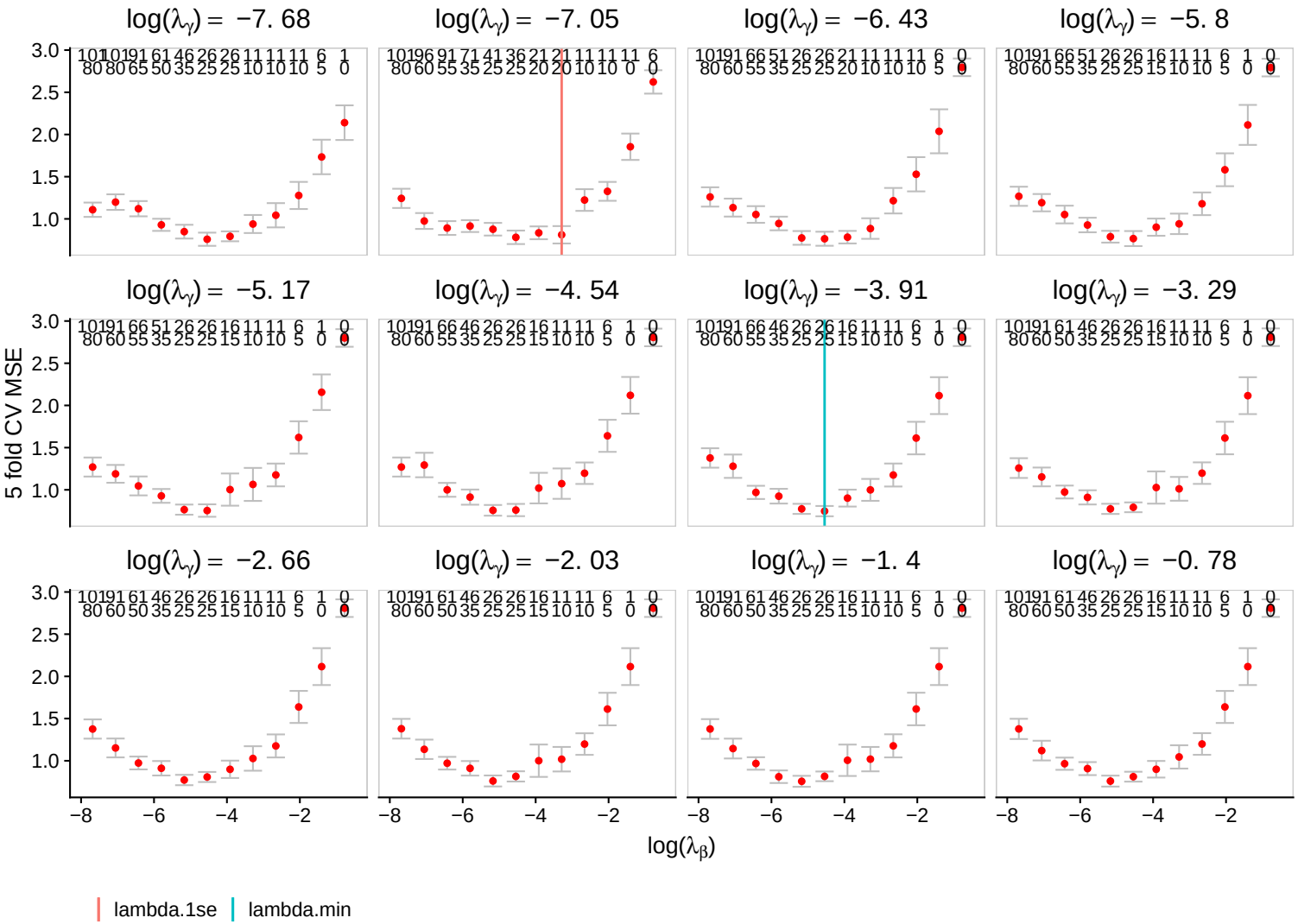


Estimated X_E*f(X_4)



sail R package: Cross-validation results

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



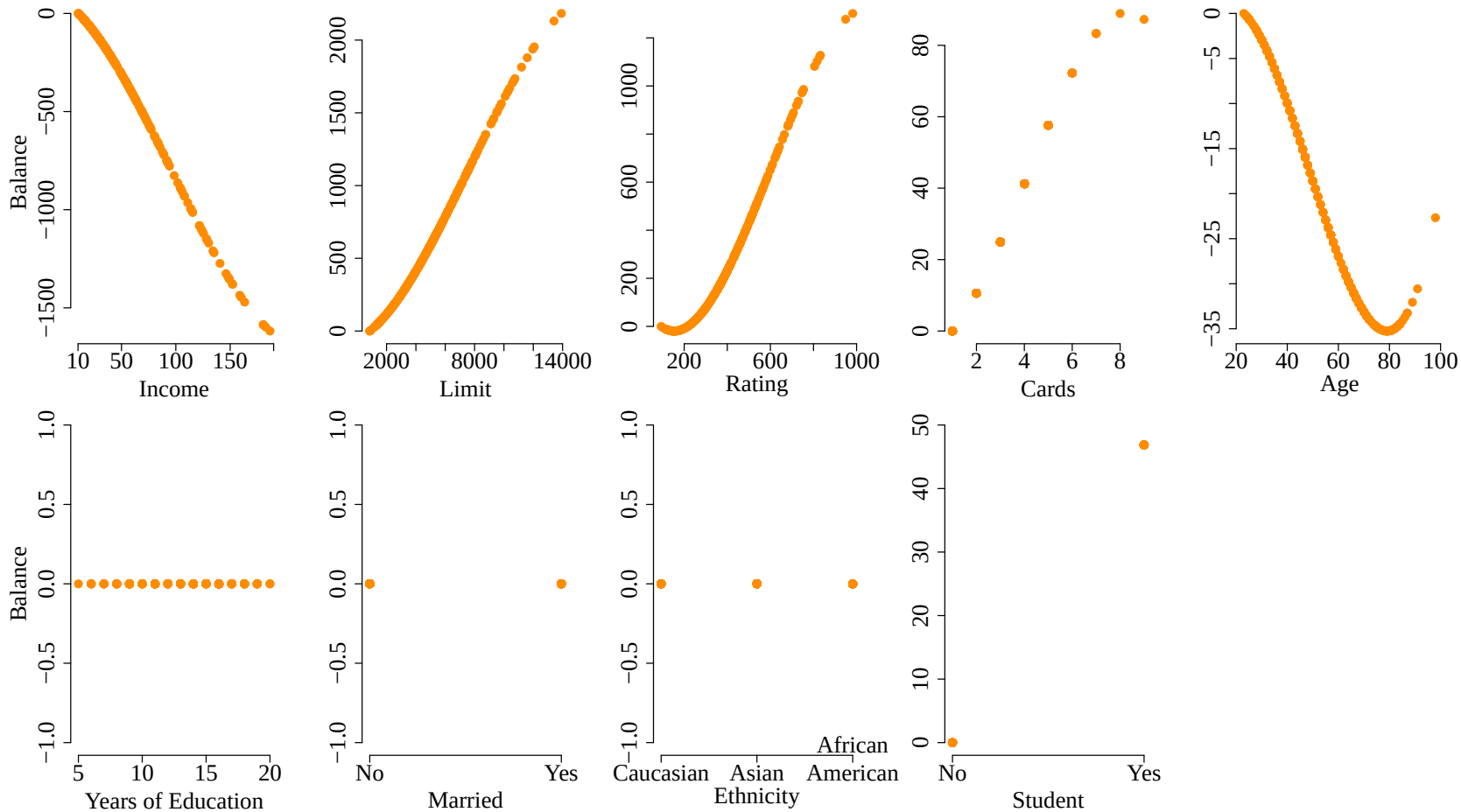
`sail`: Results on A Real Data Set

- `Credit` dataset[1]: $\mathbf{Y}_{400\times 1} \rightarrow$Credit card balance, $\mathbf{X}_{400\times 9} \rightarrow$ predictors

[1] `ISLR` package on CRAN

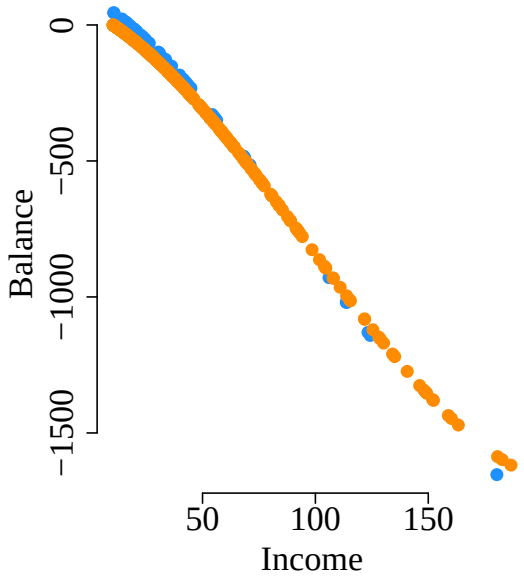
sail: Results on A Real Data Set

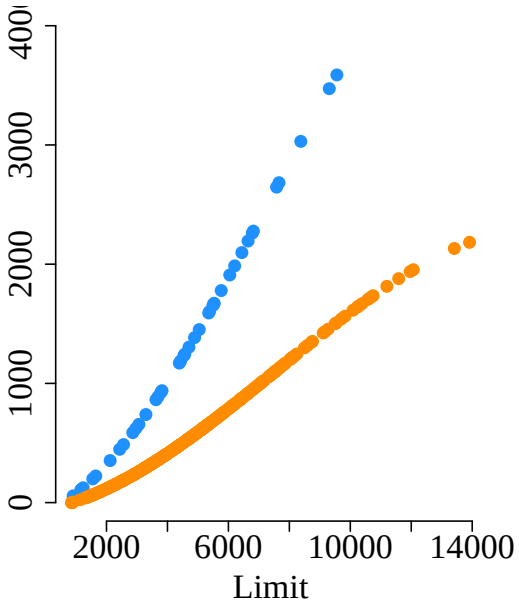
- `Credit` dataset[1]: $\mathbf{Y}_{400\times 1} \rightarrow$Credit card balance, $\mathbf{X}_{400\times 9} \rightarrow$ predictors

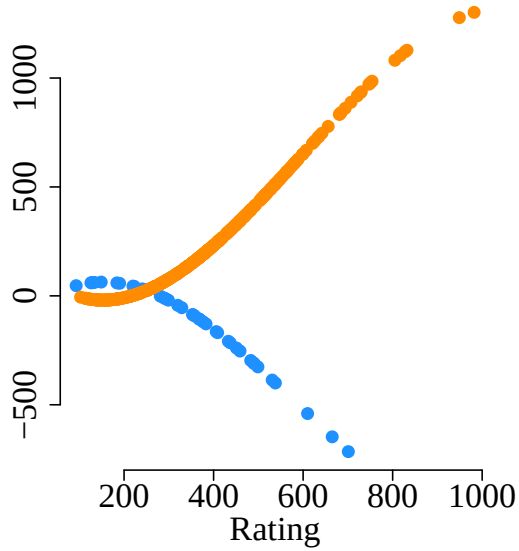


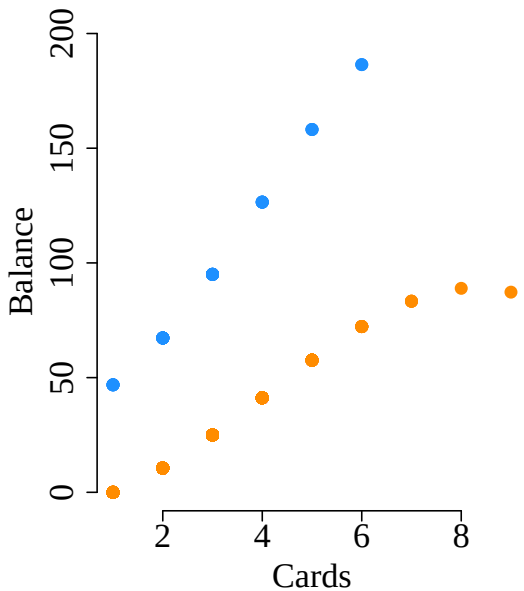
[1] `ISLR` package on CRAN

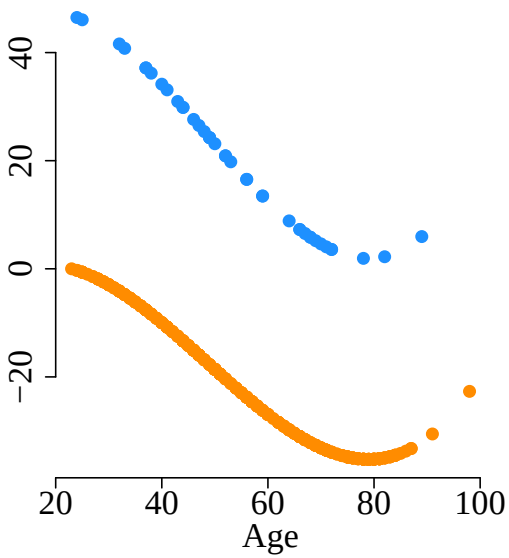
sail: Interactions with Student





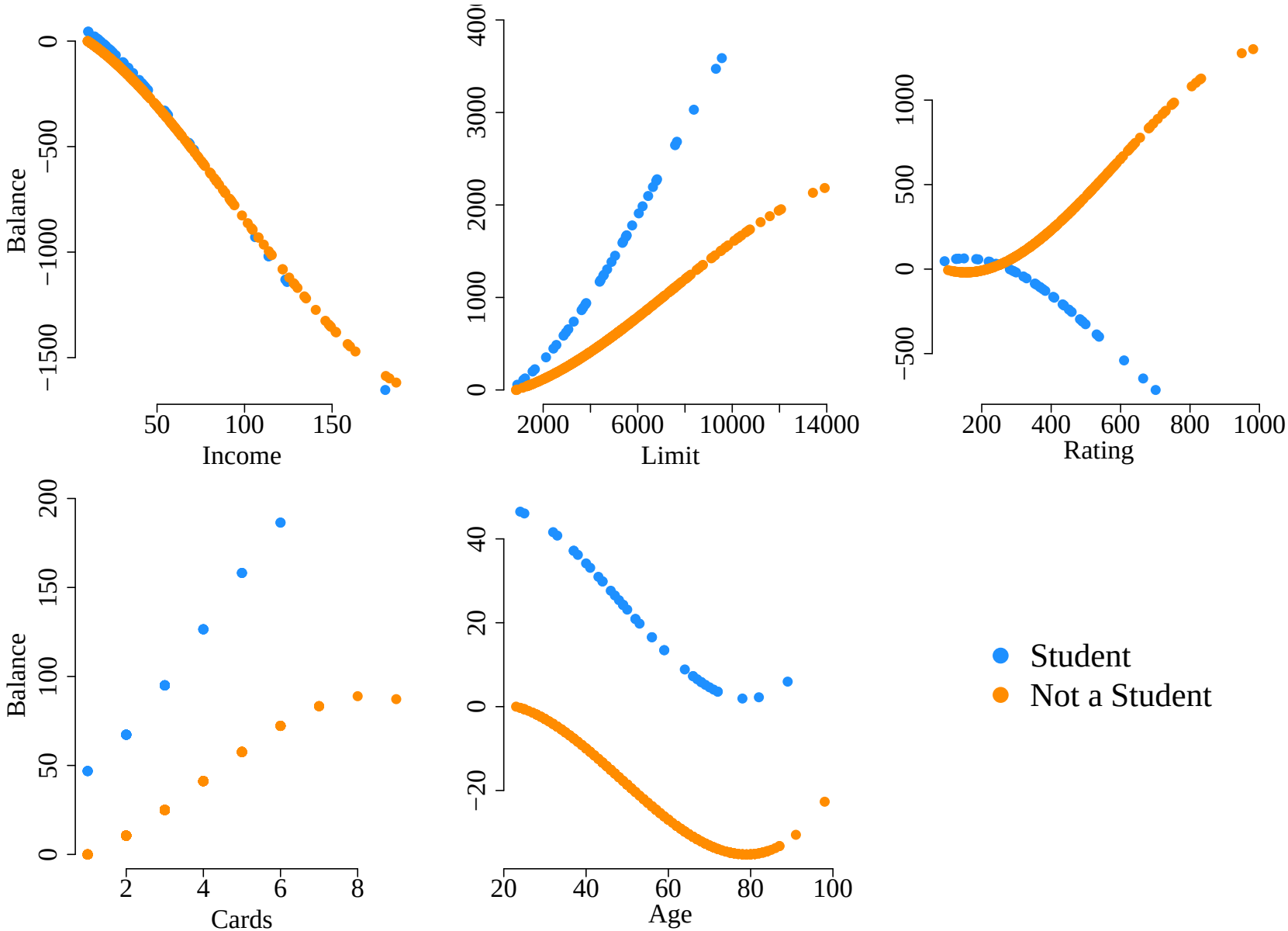






Which color represents a Student?

sail: Interactions with Student



ggmix: Mixed Models with the Group Lasso

Motivating Dataset: UK Biobank

- **500,000 individuals** with over **40 million predictors**
- 1000's of responses (e.g. disease status, bone mineral density)
- Problem: Which predictors are associated with the response?

Motivating Dataset: Two Problems

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

Problem 1: Groups of Predictors Affect the Response

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

Problem 2: Observations are not Independent

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

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

ggmix: Simulation Results

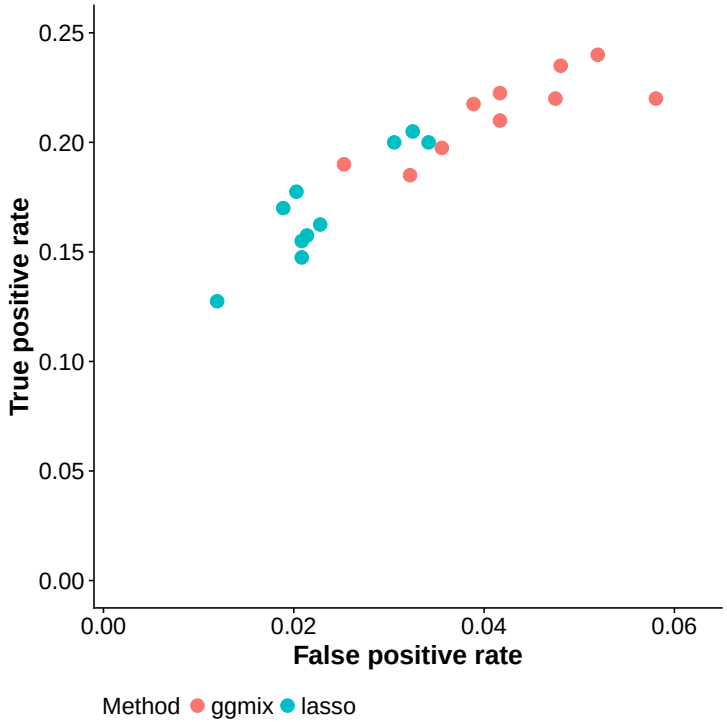


Figure 1: True Positive vs. False Positive (10 simulations). $N = 1000, p = 5000, p_{active} = 500$

Lasso	ggmix
32.8 (0.87)	26.7 (1.06)

Table 1: Mean root mean squared error (sd)

Data

- Response: $\mathsf{Y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- Predictors: $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_n)^T \in \mathbb{R}^{n \times p}$, where $p \gg n$
- Similarity Matrix: $\Phi \in \mathbb{R}^{n \times n}$
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$
- Random effect: $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$
- Error: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$

Data

- Response: $\mathsf{Y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- Predictors: $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_n)^T \in \mathbb{R}^{n \times p}$, where $p \gg n$
- Similarity Matrix: $\Phi \in \mathbb{R}^{n \times n}$
- Regression Coefficients: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$
- Random effect: $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$
- Error: $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$
- We consider the following LMM with a single random effect:

$$\mathsf{Y} = \mathsf{X}\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\varepsilon} \tag{2}$$

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

- σ^2 and $\eta \in [0, 1]$ determine how the variance is divided between $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$
- $\mathsf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathsf{X}\boldsymbol{\beta}, \eta\sigma^2\Phi + (1-\eta)\sigma^2\boldsymbol{\mathcal{I}})$

Likelihood

- The negative log-likelihood is given by

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

where $\mathsf{V} = \eta\Phi + (1-\eta)\boldsymbol{\mathcal{I}}$ [1]

[1] Pirinen et al. Annals of Applied Statistics (2013)

Likelihood

- The negative log-likelihood is given by

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

  where $\mathsf{V} = \eta\Phi + (1-\eta)\boldsymbol{\mathcal{I}}$ [1]

- Assume we have a low-rank matrix $\mathsf{K} \in \mathbb{R}^{n\times k}$ ($k < n$), to compute the factored similarity matrix $\Phi = \mathsf{K}\mathsf{K}^T$

[1] Pirinen et al. Annals of Applied Statistics (2013)

Likelihood

- The negative log-likelihood is given by

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

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

- Assume we have a low-rank matrix $\mathbf{K} \in \mathbb{R}^{n\times k}$ ($k < n$), to compute the factored similarity matrix $\Phi = \mathbf{K}\mathbf{K}^T$
- Let $\mathbf{K} = \mathbf{U}\Lambda\mathbf{V}^T$ be the SVD of $\mathbf{K}$, then

$$\Phi = \mathbf{U}_1\Lambda\Lambda\mathbf{U}_1^T = \mathbf{U}_1\Sigma\mathbf{U}_1^T$$

  where $\mathbf{U}_1 \in \mathbb{R}^{n\times k}$ is the matrix of singular vectors corresponding to the k non-zero eigenvalues.

[1] Pirinen et al. Annals of Applied Statistics (2013)

Penalized Maximum Likelihood Estimator

- The negative log-likelihood can then be expressed as

$$-\ell(\Theta) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\left(\sum_{i=1}^{k}\log(1+\eta(\Sigma_i-1)) + (n-k)\log(1-\eta)\right) + \frac{1}{2}\left\{(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta})^T\left[\frac{1}{\sigma^2(1-\eta)}\left(\boldsymbol{\mathcal{I}}_n - \mathsf{U}_1\left(\frac{1-\eta}{\eta}\Sigma_1^{-1} + \boldsymbol{\mathcal{I}}_k\right)^{-1}\mathsf{U}_1^T\right)\right](\mathsf{Y}-\mathsf{X}\boldsymbol{\beta})\right\}$$

Penalized Maximum Likelihood Estimator

- The negative log-likelihood can then be expressed as

$$-\ell(\Theta) \propto \frac{n}{2}\log(\sigma^2) + \frac{1}{2}\left(\sum_{i=1}^{k}\log(1+\eta(\Sigma_i-1)) + (n-k)\log(1-\eta)\right) + \frac{1}{2}\left\{(\mathsf{Y}-\mathsf{X}\boldsymbol{\beta})^T\left[\frac{1}{\sigma^2(1-\eta)}\left(\boldsymbol{\mathcal{I}}_n - \mathsf{U}_1\left(\frac{1-\eta}{\eta}\Sigma_1^{-1} + \boldsymbol{\mathcal{I}}_k\right)^{-1}\mathsf{U}_1^T\right)\right](\mathsf{Y}-\mathsf{X}\boldsymbol{\beta})\right\}$$

- Define the objective function:

$$Q_\lambda(\Theta) = -\ell(\Theta) + \lambda\sum_j P_j(\beta_j)$$

- $P_j(\cdot)$ is a penalty term on $\beta_1, \ldots, \beta_{p+1}$
- An estimate of the regression parameters $\widehat{\Theta}_\lambda$ is obtained by

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

Group Lasso for Mixed Models

- Assume the predictors in $\mathbf{X} \in \mathbb{R}^{n \times p}$ belong to K **non-overlapping groups** with **pre-defined** group membership and cardinality p_k
- Let $\boldsymbol{\beta}_{(k)}$ to denote the segment of $\boldsymbol{\beta}$ corresponding to group k

Group Lasso for Mixed Models

- Assume the predictors in $\mathbf{X} \in \mathbb{R}^{n \times p}$ belong to K **non-overlapping groups** with **pre-defined** group membership and cardinality p_k
- Let $\boldsymbol{\beta}_{(k)}$ to denote the segment of $\boldsymbol{\beta}$ corresponding to group k
- We consider the group lasso penalized estimator

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|\mathbf{D}) + \lambda \sum_{k=1}^{K} w_k \|\boldsymbol{\beta}_{(k)}\|_2, \tag{3}$$

- where

$$L(\boldsymbol{\beta} \mid \mathbf{D}) = \frac{1}{2} \left[\mathbf{Y} - \widehat{\mathbf{Y}}\right]^\top \mathbf{W} \left[\mathbf{Y} - \widehat{\mathbf{Y}}\right] \tag{4}$$

$\widehat{\mathbf{Y}} = \sum_{j=1}^{p} \beta_j X_j$, $\mathbf{D}$ is the working data $\{\mathbf{Y}, \mathbf{X}\}$, and

$$\mathbf{W}_{n \times n} = \frac{1}{\sigma^2(1-\eta)} \left(\boldsymbol{\mathcal{I}}_n - \mathbf{U}_1 \left(\frac{1-\eta}{\eta} \Sigma_1^{-1} + \boldsymbol{\mathcal{I}}_k \right)^{-1} \mathbf{U}_1^T \right) \tag{5}$$

Groupwise Descent: Exploiting Sparsity Structure

Minimize the objective function

$$\frac{1}{2}\left[\mathsf{Y}-\widehat{\mathsf{Y}}\right]^{\top}\mathsf{W}\left[\mathsf{Y}-\widehat{\mathsf{Y}}\right]+\lambda\sum_{k=1}^{K}w_k\|\boldsymbol{\beta}^{(k)}\|_2$$

During each sub-iteration only optimize $\boldsymbol{\beta}^{(k)}$. Set $\boldsymbol{\beta}^{(k')}=\widetilde{\boldsymbol{\beta}}^{(k')}$ for $k'\neq k$ at their current value.

1. Initialization: $\widetilde{\boldsymbol{\beta}}$
2. Cyclic groupwise descent: for $k=1,2,\ldots,K$, update $\boldsymbol{\beta}^{(k)}$ by minimizing the objective function

$$\widetilde{\boldsymbol{\beta}}^{(k)}(\mathrm{new})\leftarrow\arg\min_{\boldsymbol{\beta}^{(k)}}L(\boldsymbol{\beta}\mid\mathsf{D})+\lambda w_k\|\boldsymbol{\beta}^{(k)}\|_2$$

3. Repeat (2) till convergence.

Quadratic Majorization Condition

$$\arg\min_{\boldsymbol{\beta}^{(k)}} \frac{1}{2}\left[\mathbf{Y}-\widehat{\mathbf{Y}}\right]^{\top}\mathbf{W}\left[\mathbf{Y}-\widehat{\mathbf{Y}}\right]+\lambda\sum_{k=1}^{K}w_k\left\|\boldsymbol{\beta}^{(k)}\right\|_2 \tag{6}$$

- Unfortunately, there is no closed form solution to (6)

[1] Yang and Zou. Statistical Computing (2014)

Quadratic Majorization Condition

$$\arg\min_{\boldsymbol{\beta}^{(k)}} \frac{1}{2}\left[\mathbf{Y}-\widehat{\mathbf{Y}}\right]^{\top}\mathbf{W}\left[\mathbf{Y}-\widehat{\mathbf{Y}}\right]+\lambda\sum_{k=1}^{K}w_k\left\|\boldsymbol{\beta}^{(k)}\right\|_2 \tag{6}$$

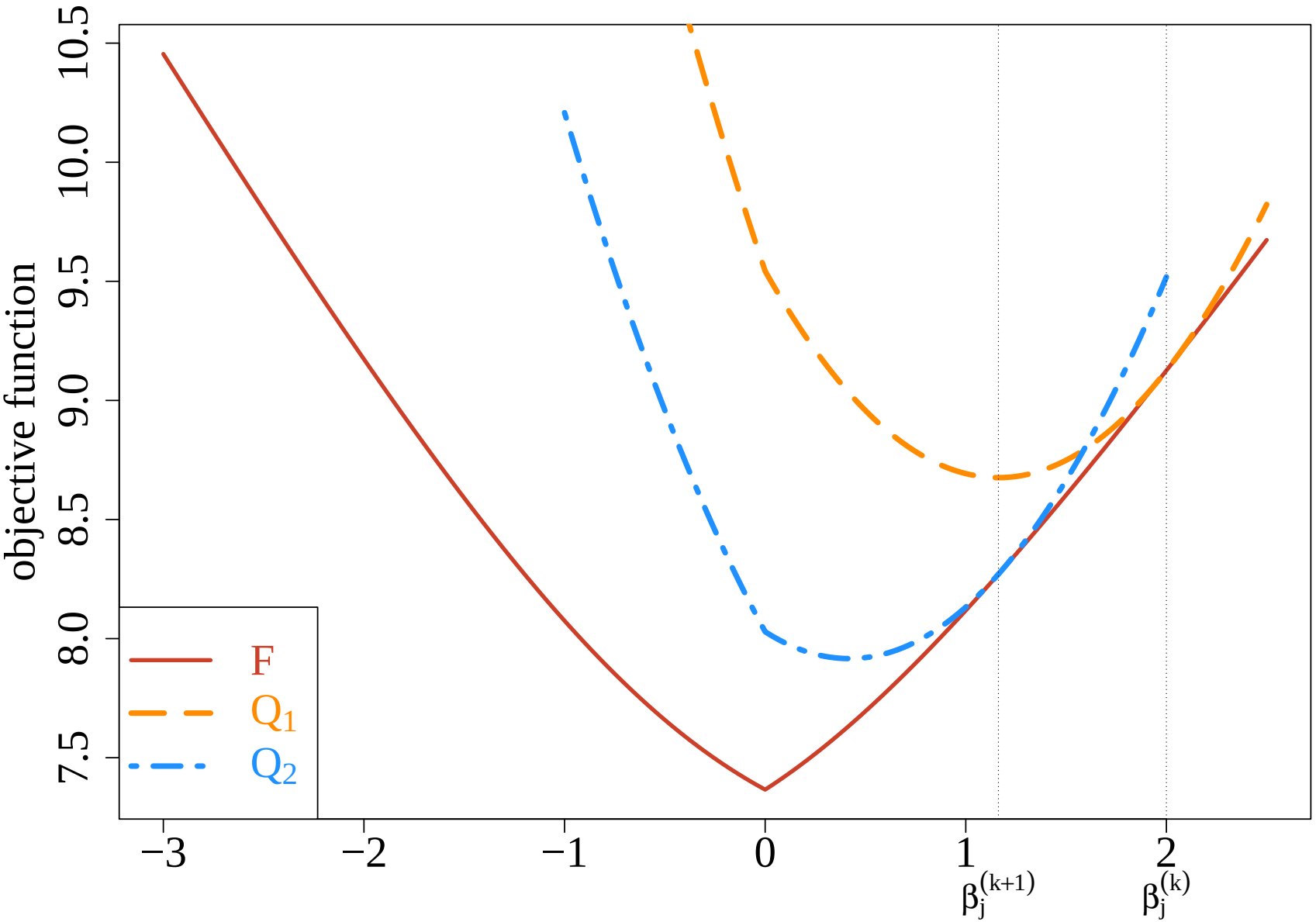
- Unfortunately, there is no closed form solution to (6)
- However, the loss function $L(\boldsymbol{\beta}|\mathbf{D})$ satisfies the quadratic majorization (QM) condition[1], since there exists
  - a $p \times p$ matrix $\mathbf{H} = \mathbf{X}^{\top}\mathbf{W}\mathbf{X}$, and
  - $\nabla L(\boldsymbol{\beta}|\mathbf{D}) = -\left(Y-\hat{Y}\right)^{\top}\mathbf{W}\mathbf{X}$

  which may only depend on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}, \boldsymbol{\beta}^*$,

$$L(\boldsymbol{\beta}\mid\mathbf{D}) \leq L(\boldsymbol{\beta}^*\mid\mathbf{D})+(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\top}\nabla L(\boldsymbol{\beta}^*|\mathbf{D})+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\top}\mathbf{H}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)$$

[1] Yang and Zou. Statistical Computing (2014)

Generalized Coordinate Descent (GCD)



Groupwise Majorization Descent

- Update β in a groupwise fashion

$$\beta - \widetilde{\beta} = (\underbrace{0, \ldots, 0}_{k-1}, \beta^{(k)} - \widetilde{\beta}^{(k)}, \underbrace{0, \ldots, 0}_{K-k})$$

Groupwise Majorization Descent

- Update β in a groupwise fashion

$$\beta - \widetilde{\beta} = (\underbrace{0, \ldots, 0}_{k-1}, \beta^{(k)} - \widetilde{\beta}^{(k)}, \underbrace{0, \ldots, 0}_{K-k})$$

- Only need to compute the majorization function on group level

$$L(\beta \mid \mathsf{D}) \leq L(\widetilde{\beta} \mid \mathsf{D}) - (\beta^{(k)} - \widetilde{\beta}^{(k)})^\top U^{(k)} + \frac{1}{2}\gamma_k (\beta^{(k)} - \widetilde{\beta}^{(k)})^\top (\beta^{(k)} - \widetilde{\beta}^{(k)})$$

$$U^{(k)} = \frac{\partial}{\partial \beta_{(k)}} L(\beta \mid \mathsf{D}) = -\left(Y - \hat{Y}\right)^\top \mathsf{W}\mathsf{X}_{(k)}$$

$$\mathsf{H}^{(k)} = \frac{\partial^2}{\partial \beta_{(k)} \partial \beta_{(k)}^\top} L(\beta \mid \mathsf{D}) = \mathsf{X}_{(k)}^\top \mathsf{W}\mathsf{X}_{(k)}$$

- $\gamma_k = \mathrm{eigen}_{\max}(\mathsf{H}^{(k)})$

Groupwise Majorization Descent

- Update β in a groupwise fashion

$$\beta - \widetilde{\beta} = (\underbrace{0, \ldots, 0}_{k-1}, \beta^{(k)} - \widetilde{\beta}^{(k)}, \underbrace{0, \ldots, 0}_{K-k})$$

- Only need to compute the majorization function on group level

$$L(\beta \mid \mathsf{D}) \leq L(\widetilde{\beta} \mid \mathsf{D}) - (\beta^{(k)} - \widetilde{\beta}^{(k)})^\top U^{(k)} + \frac{1}{2}\gamma_k (\beta^{(k)} - \widetilde{\beta}^{(k)})^\top (\beta^{(k)} - \widetilde{\beta}^{(k)})$$

$$U^{(k)} = \frac{\partial}{\partial \beta_{(k)}} L(\beta \mid \mathsf{D}) = -\left(Y - \hat{Y}\right)^\top \mathsf{W}\mathsf{X}_{(k)}$$

$$\mathsf{H}^{(k)} = \frac{\partial^2}{\partial \beta_{(k)} \partial \beta_{(k)}^\top} L(\beta \mid \mathsf{D}) = \mathsf{X}_{(k)}^\top \mathsf{W}\mathsf{X}_{(k)}$$

- $\gamma_k = \mathrm{eigen}_{\max}(\mathsf{H}^{(k)})$
- Update $\widetilde{\beta}^{(k)}$ with a fast operation:

$$\widetilde{\beta}^{(k)}(\mathrm{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\beta}^{(k)}\right)\left(1 - \frac{\lambda w_k}{\|U^{(k)} + \gamma_k \widetilde{\beta}^{(k)}\|_2}\right)_+$$

Discussion

Strengths and Limitations

Strengths

- Environment interactions with strong heredity property in $p >> N$
- `sail` allows for flexible modeling of input variables
- `ggmix` package provides first implementation of group lasso in mixed models

Strengths and Limitations

Strengths

- Environment interactions with strong heredity property in $p >> N$
- `sail` allows for flexible modeling of input variables
- `ggmix` package provides first implementation of group lasso in mixed models

Limitations

- `sail` can currently only handle $E \cdot f(X)$ or $f(E) \cdot X$
- Does not allow for $f(X_1, E)$ or $f(X_1, X_2)$
- Current implementation of `sail` is slow due to cross validation for 2 tuning parameters
- Memory footprint is an issue for both `sail` and `ggmix`

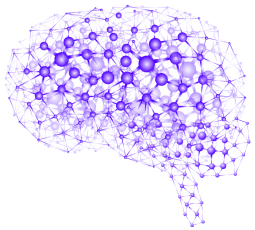
Future Directions

- Are two tuning parameters really necessary ?

$$\lambda \left\{ (1-\alpha) \left[w_E |\beta_E| + \sum_{j=1}^{p} w_j \|\theta_j\|_2 \right] + \alpha \sum_{j=1}^{p} w_{jE} |\gamma_j| \right\}$$

- Weak heredity property $\rightarrow \alpha_j = \gamma_j(|\beta_j| + |\beta_E|)$
- Implement ADMM algorithm for scalability. Distributed computing (GPU)
- Extension to nonconvex penalties (SCAD, MCP)
- Binary Outcomes

Acknowledgements



CENTRE CENTER
LUDMER
NEUROINFORMATIQUE & SANTÉ MENTALE | NEUROINFORMATICS & MENTAL HEALTH

Jewish General Hospital
Lady Davis Institute for Medical Research

McGill

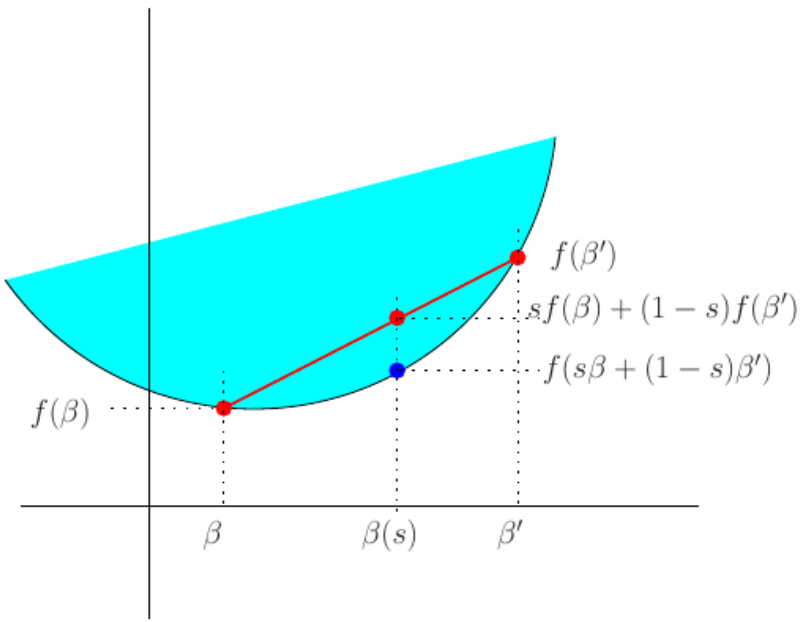
References

- Radchenko, P., & James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. Journal of the American Statistical Association, 105(492), 1541-1553.
- Choi, N. H., Li, W., & Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. Journal of the American Statistical Association, 105(489), 354-364.
- Chipman, H. (1996). Bayesian variable selection with related predictors. Canadian Journal of Statistics, 24(1), 17-36.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1)
- Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. Statistics and Computing, 25(6), 1129-1141
- De Leeuw, J. (1994). Block-relaxation algorithms in statistics. In Information systems and data analysis (pp. 308-324). Springer Berlin Heidelberg.

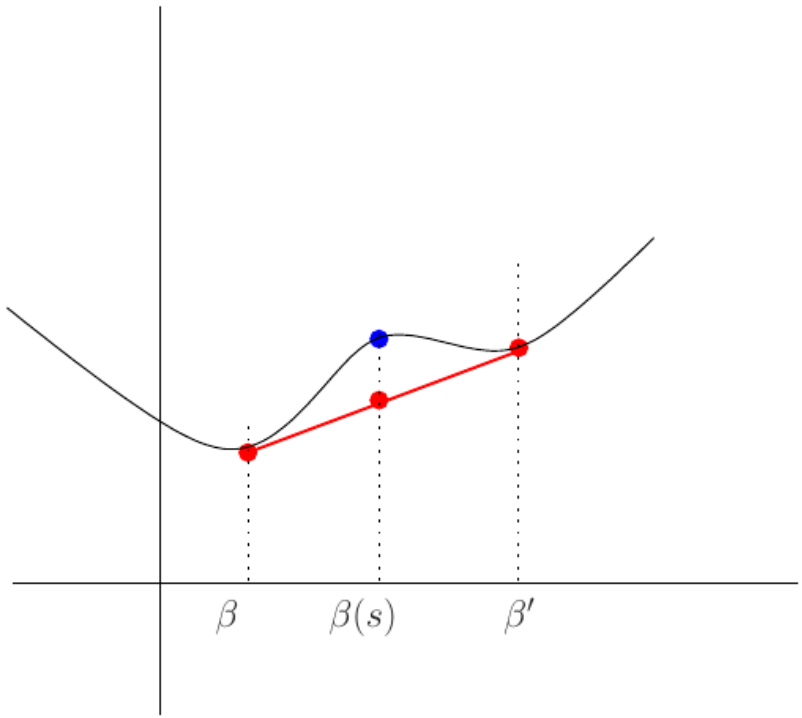
sahirbhatnagar.com

Appendix

Convex Function



(a)

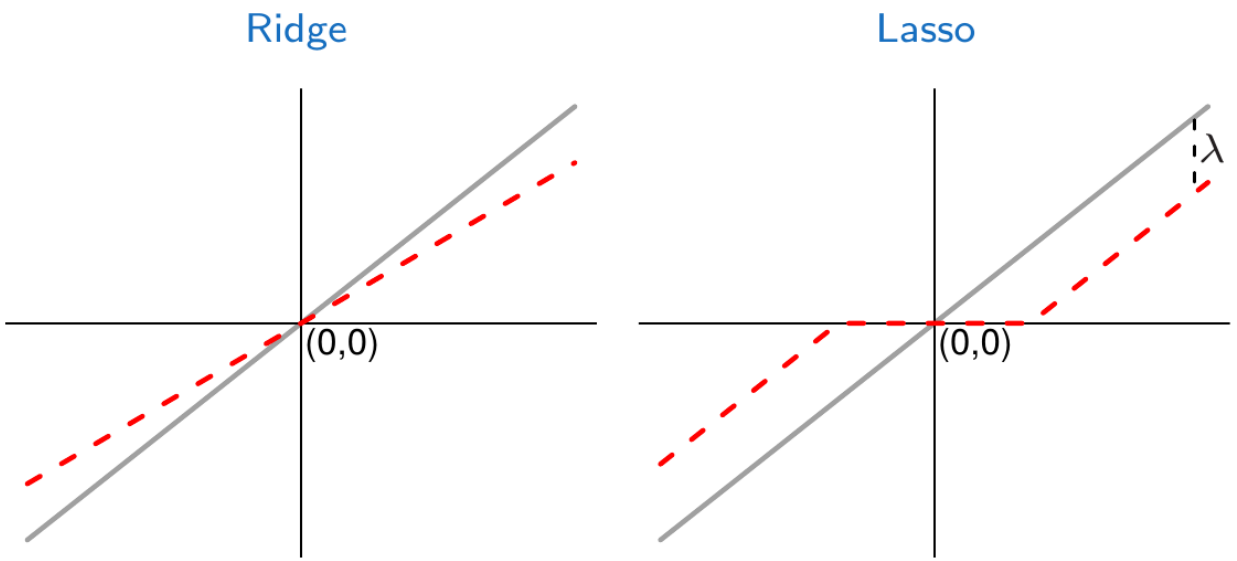


(b)

Ridge vs. Lasso

- Estimators of β_j in terms of the least-squares estimate $\widehat{\beta}_j^{LS}$ for an orthonormal model matrix $\mathbf{X}$

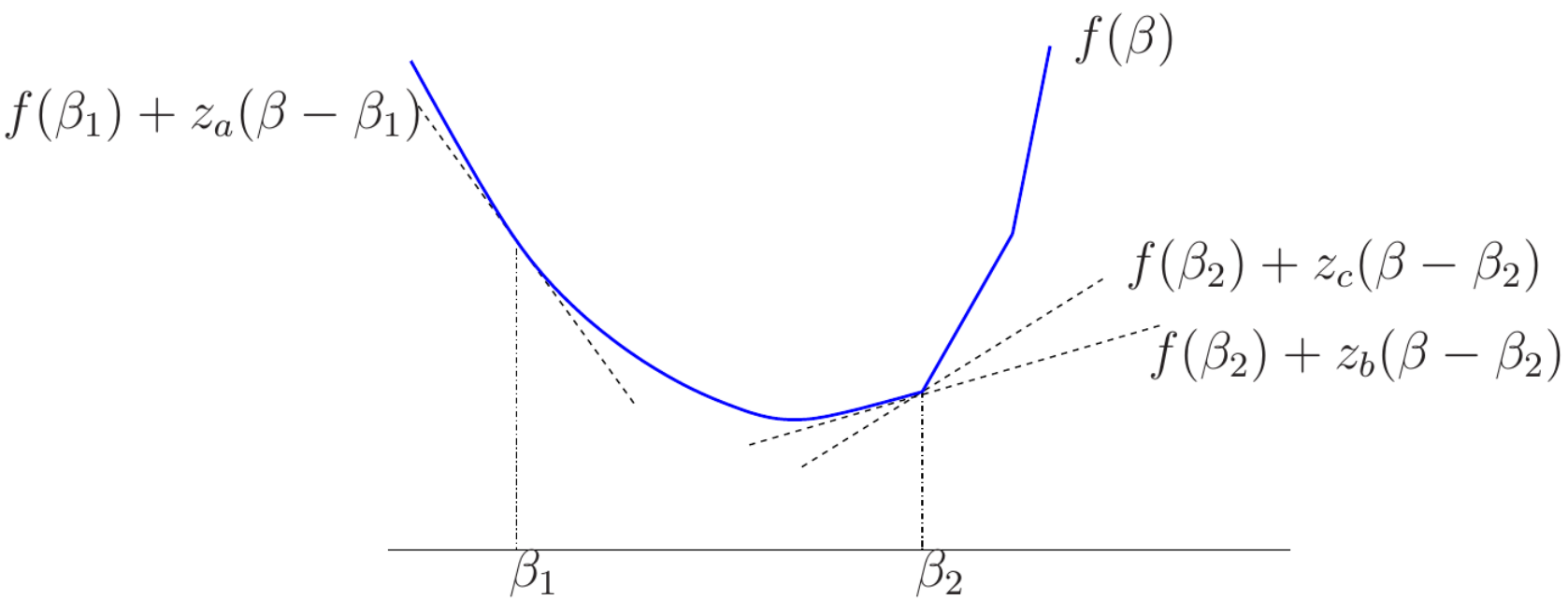
q	Estimator	Formula
1	Lasso	$\text{sign}(\widehat{\beta}_j^{LS})(\lvert\widehat{\beta}_j^{LS}\rvert - \lambda)_+$
2	Ridge	$\widehat{\beta}_j^{LS}/(1+\lambda)$



Subgradient

- A basic property of differentiable convex functions is that the first-order tangent approximation always provides a lower bound.
- The notion of subgradient is based on a natural generalization of this idea. In particular, given a convex function $f : \mathbb{R}^p \to \mathbb{R}$, a vector $z \in \mathbb{R}^p$ is said to be a *subgradient* of f at β if

$$f(\beta') \geq f(\beta) + \langle z, \beta' - \beta \rangle, \quad \text{for all } \beta' \in \mathbb{R}^p$$



Algorithm for ggmix

Algorithm 2: Block Relaxation Algorithm

Set the iteration counter $k \leftarrow 0$, initial values for the parameter vector $\Theta^{(0)}$ and convergence threshold ϵ;

for $\lambda \in \{\lambda_{max}, \ldots, \lambda_{min}\}$ **do**

  repeat

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

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

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

    $k \leftarrow k + 1$

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

end

Coordinate Descent

- Minimize the objective function

$$\min_{\boldsymbol{\beta}} F(\boldsymbol{\beta}) \equiv L(\beta_1, \ldots, \beta_p) + p_\lambda(\boldsymbol{\beta})$$

- $L : \mathbb{R}^p \to \mathbb{R}$ is differentiable convex, $p_\lambda : \mathbb{R}^p \to \mathbb{R}$ bounded from below but not necessarily smooth. e.g. $p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$

Coordinate Descent (Fu (1998), Friedman et al. (2007), Wu and Lange (2008))

1. Initialization: $\tilde{\boldsymbol{\beta}}$
2. Cyclic coordinate descent: for $j = 1, 2, \ldots, p$, update β_j by minimizing the objective function

$$\tilde{\beta}_j^{new} \leftarrow \arg\min_{\beta_j} F(\beta_j | \beta_k = \tilde{\beta}_k, k \neq j)$$

3. Repeat (2) till convergence.

Quadratic Majorization Condition

Empirical loss

$$L(\beta \mid \mathbf{D}) = \frac{1}{n}\sum_{i=1}^{n} \Phi(y_i, \beta^\top x_i)$$

Φ satisfies the QM condition, if and only if:

1. $L(\beta \mid \mathbf{D})$ is differentiable as a function of β.
2. Can find $\mathbf{H}$ (p by p), may only depend on the data $\mathbf{D}$, such that for all β, β^*

$$L(\beta \mid \mathbf{D}) \leq L(\beta^* \mid \mathbf{D}) + (\beta - \beta^*)^\top \nabla L(\beta^* | \mathbf{D}) + \frac{1}{2}(\beta - \beta^*)^\top \mathbf{H}(\beta - \beta^*)$$

Loss	$-\nabla L(\beta \mid \mathbf{D})$	$\mathbf{H}$
Least squares	$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)x_i$	$\mathbf{X}^\top\mathbf{X}/n$
Logistic regression	$\frac{1}{n}\sum_{i=1}^{n} y_i x_i \frac{1}{1+\exp(y_i x_i^\top \beta)}$	$\frac{1}{4}\mathbf{X}^\top\mathbf{X}/n$
Squared hinge loss	$\frac{1}{n}\sum_{i=1}^{n} 2y_i x_i(1 - y_i x_i^\top \beta)_+$	$4\mathbf{X}^\top\mathbf{X}/n$
Huberized hinge loss	$\frac{1}{n}\sum_{i=1}^{n} y_i x_i \mathrm{hsvm}'(y_i x_i^\top \beta)$	$\frac{2}{\delta}\mathbf{X}^\top\mathbf{X}/n$

Verifying Quadratic Majorization Condition

Lemma 1 (Yang and Zou, 2014)

Assume $\Phi(y, f)$ is differentiable with respect to f and write $\Phi'_f = \frac{\partial \Phi(y,f)}{\partial f}$, $\nabla L(\boldsymbol{\beta}|\boldsymbol{D}) = \frac{1}{n}\sum_{i=1}^{n} \Phi'_f(y_i, \boldsymbol{x}_i^\top \boldsymbol{\beta})\boldsymbol{x}_i$.

1. *If Φ'_f is* **Lipschitz continuous** *with constant C such that*

$$|\Phi'_f(y, f_1) - \Phi'_f(y, f_2)| \leq C|f_1 - f_2| \quad \forall\, y, f_1, f_2,$$

   then the QM condition holds for Φ and $\boldsymbol{H} = \frac{2C}{n}\boldsymbol{X}^\top\boldsymbol{X}$.

2. *If $\Phi''_f = \frac{\partial \Phi^2(y,f)}{\partial f^2}$* **exists** *and*

$$\Phi''_f \leq C_2 \quad \forall\, y, f,$$

   then the QM condition holds for Φ and $\boldsymbol{H} = \frac{C_2}{n}\boldsymbol{X}^\top\boldsymbol{X}$.

Strict Descent Property of GMD

Proposition (Yang and Zou, 2014)

- If $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \neq \widetilde{\boldsymbol{\beta}}^{(k)}$ then

$$L(\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \mid \mathbf{D}) + \lambda w_k \|\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})\|_2 < L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) + \lambda w_k \|\widetilde{\boldsymbol{\beta}}^{(k)}\|_2$$

  the objective function is **strictly decreased** after updating all k in a cycle.

- If $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \widetilde{\boldsymbol{\beta}}^{(k)}$ for all k, then the solution must satisfy the KKT conditions:

$$\begin{aligned} -U^{(k)} + \lambda w_k \cdot \frac{\widetilde{\boldsymbol{\beta}}^{(k)}}{\|\widetilde{\boldsymbol{\beta}}^{(k)}\|_2} = 0 \qquad & \text{if } \widetilde{\boldsymbol{\beta}}^{(k)} \neq 0, \\ \left\|U^{(k)}\right\|_2 \leq \lambda w_k \qquad & \text{if } \widetilde{\boldsymbol{\beta}}^{(k)} = 0, \end{aligned}$$

  which means that the algorithm converges and finds **the right answer**.

Separability and Coordinate Descent

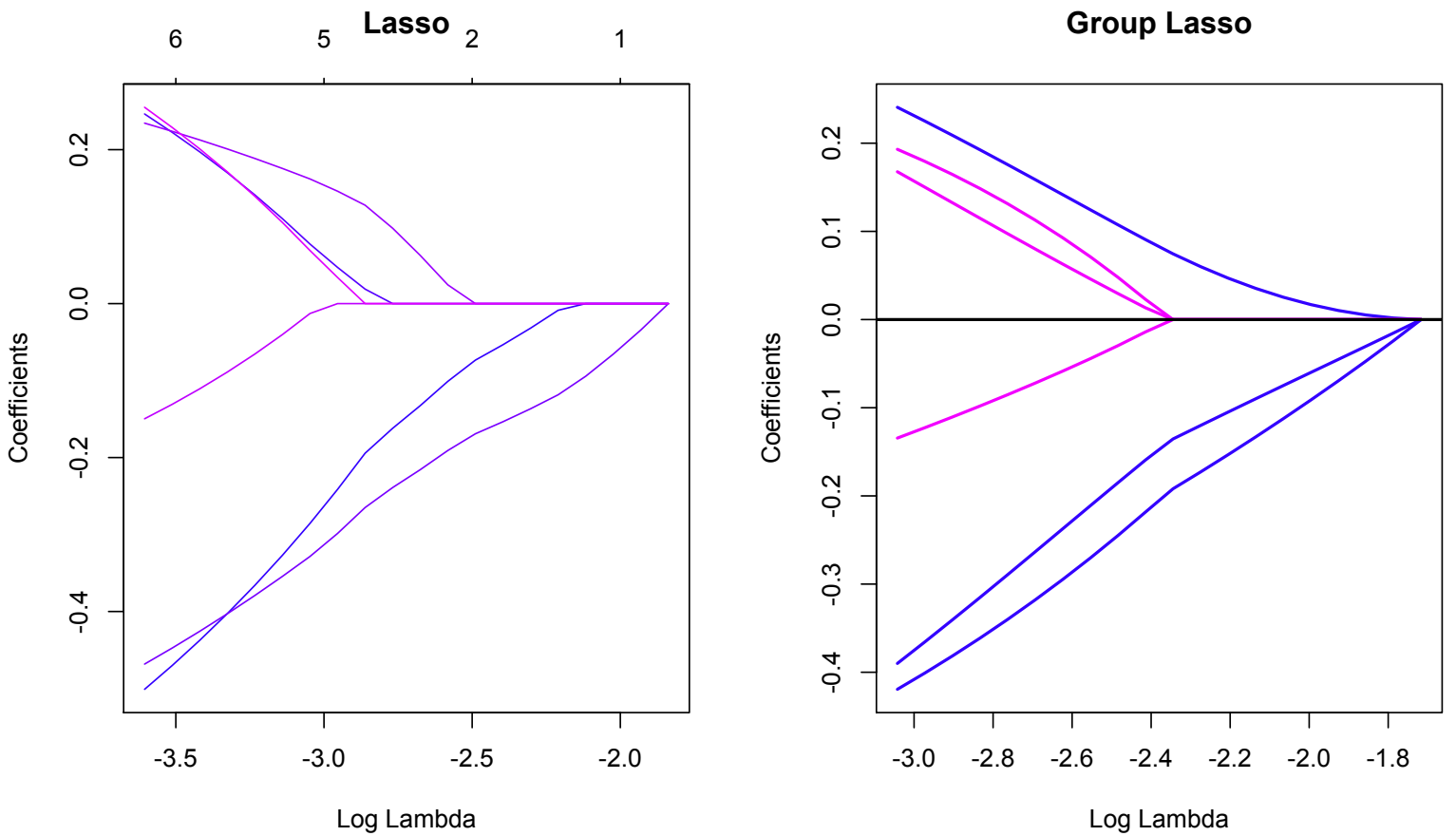
$$f(\beta_1, \ldots, \beta_p) = g(\beta_1, \ldots, \beta_p) + \sum_{j=1}^{p} h_j(\beta_j)$$

- $g : \mathbb{R}^p \to \mathbb{R}$ is differentiable and convex, and the univariate functions $h_j : \mathbb{R} \to \mathbb{R}$ are convex **but not necessarily differentiable**.
- Tseng (2001) shows that for any convex cost function f with separable structure, the coordinate descent algorithm is guaranteed to converge to the global minimizer
- The key property underlying this result is the separability of the nondifferentiable component $h(\beta) = \sum_{j=1}^{p} h_j(\beta_j)$, as a sum of functions of each individual parameter.

[1] Tseng. Journal of optimization theory and applications (2001)

Lasso vs. Group Lasso

- Logistic regression with group lasso: $n = 50$, $p = 6$.
- Group lasso: specify $(\beta_1, \beta_2, \beta_3)$, $(\beta_4, \beta_5, \beta_6)$. Variable selection at the group level.
- Solution path: view β as function of λ.



Group Lasso Motivation

- **Categorical predictors (factors):** dummy variables
- **Additive Model:** $\sum_{k=1}^{K} f_k(X^{(k)}) \approx \sum_{k=1}^{K} \sum_{m=1}^{M} \beta_{km} h_m(X^{(k)})$
  - ex. **birth weight** predicted by the mother's **age** and **weight**, Age, Age^2, Age^3 and $Weight$, $Weight^2$, $Weight^3$

Group lasso partitions the variable coefficients into K groups

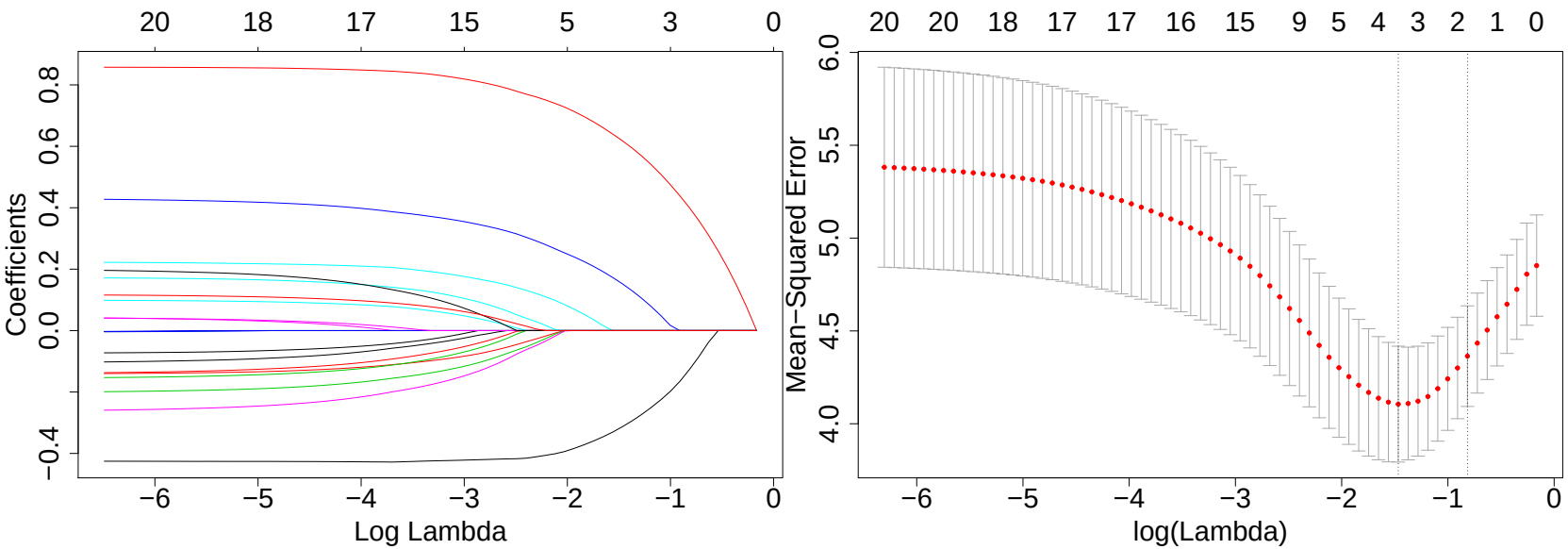
$$\boldsymbol{\beta} = ([\boldsymbol{\beta}^{(1)}]^\mathsf{T}, [\boldsymbol{\beta}^{(2)}]^\mathsf{T}, \cdots, [\boldsymbol{\beta}^{(K)}]^\mathsf{T})^\mathsf{T}$$

Extended from the lasso penalty, the group lasso estimator is:

$$\min_{(\beta_0, \boldsymbol{\beta})} \frac{1}{2} \|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} \|\boldsymbol{\beta}^{(k)}\|_2 \qquad p_k - \text{group size}$$

Choosing Model Complexity

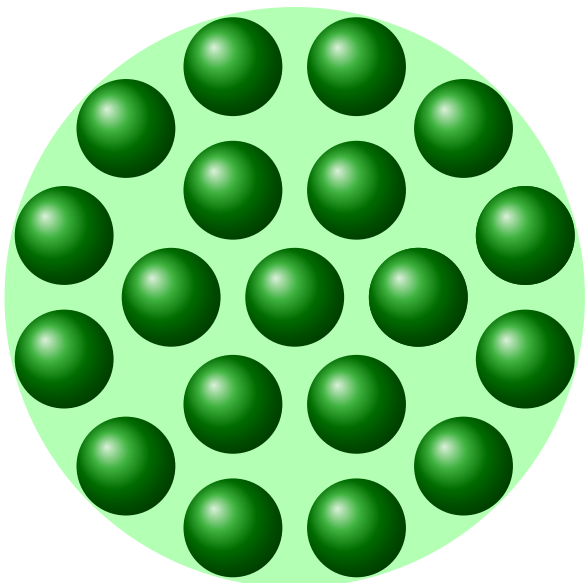
- The tuning parameter λ controls the complexity of the model
- **Generalization ability of the model**: we select the λ that gives the most accurate model for predicting independent test data from the same population



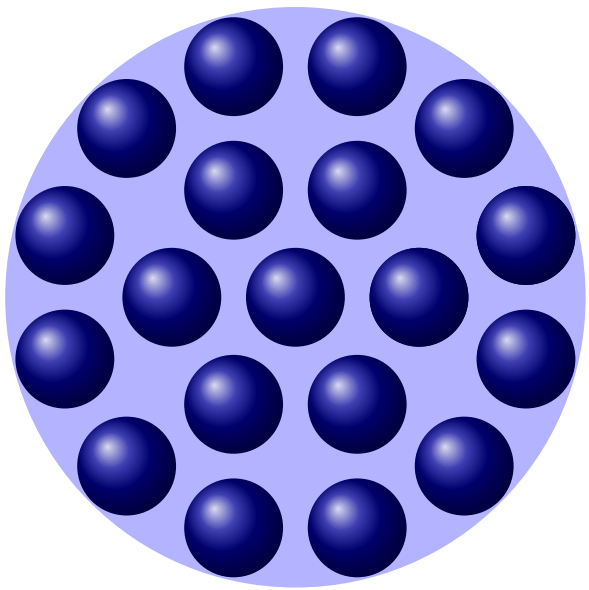
Simulation Study

- $\mathbf{X}^{(test)}$: 4k SNPs from UK Biobank genotyped data from 1k randomly sampled individuals
- $\mathbf{X}^{(causal)}$: random sample of 400 SNPs from $\mathbf{X}^{(test)}$
- $\mathbf{X}^{(kinship)}$: random sample of 4k SNPs used to construct the kinship matrix (Φ) including the causal SNPs
- $\beta_j \sim Unif(0.1, 0.3)$ for $j = 1, \ldots, 400$
- $Y = \beta_0 + \sum_{j=1}^{400} \beta_j \mathbf{X}_j^{(causal)} + P + E$
- $P \sim \mathcal{N}(0, 0.6 \cdot \Phi_{n \times n})$, $E \sim \mathcal{N}(0, 0.4 \cdot \mathbf{I}_{n \times n})$

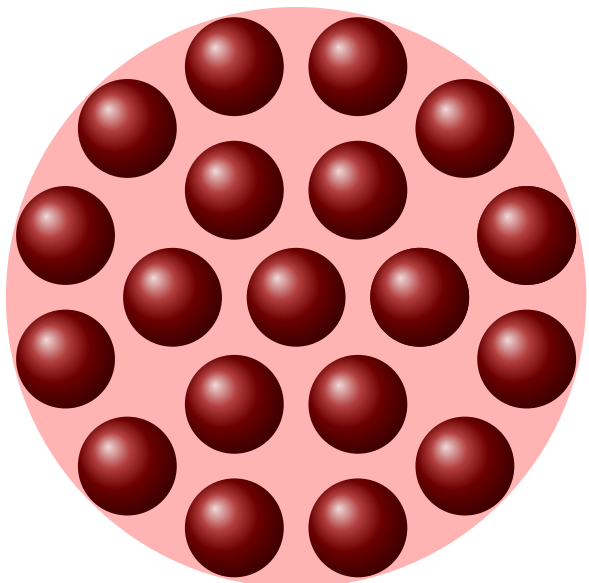
Motivation



(a) Retired

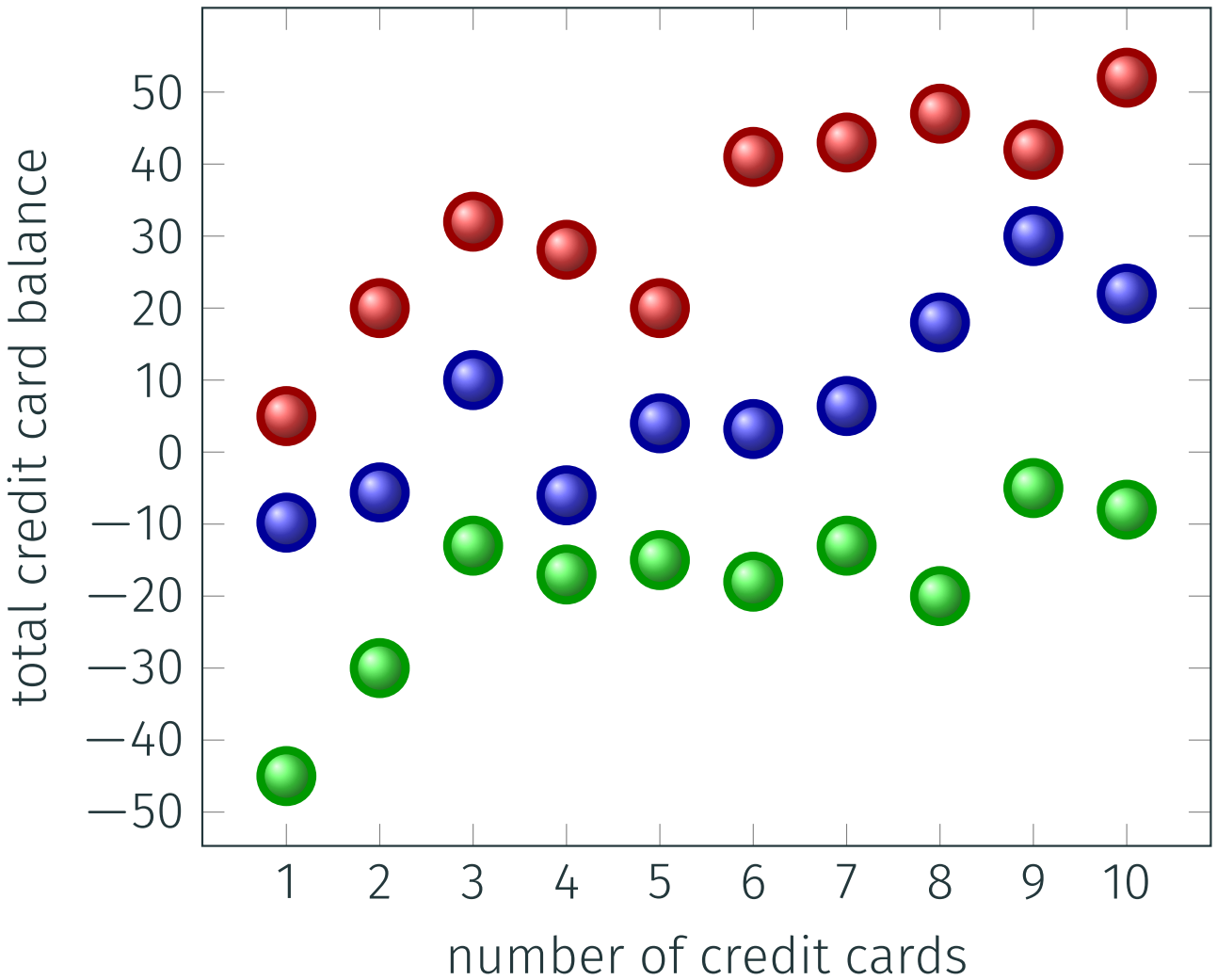


(b) Employed

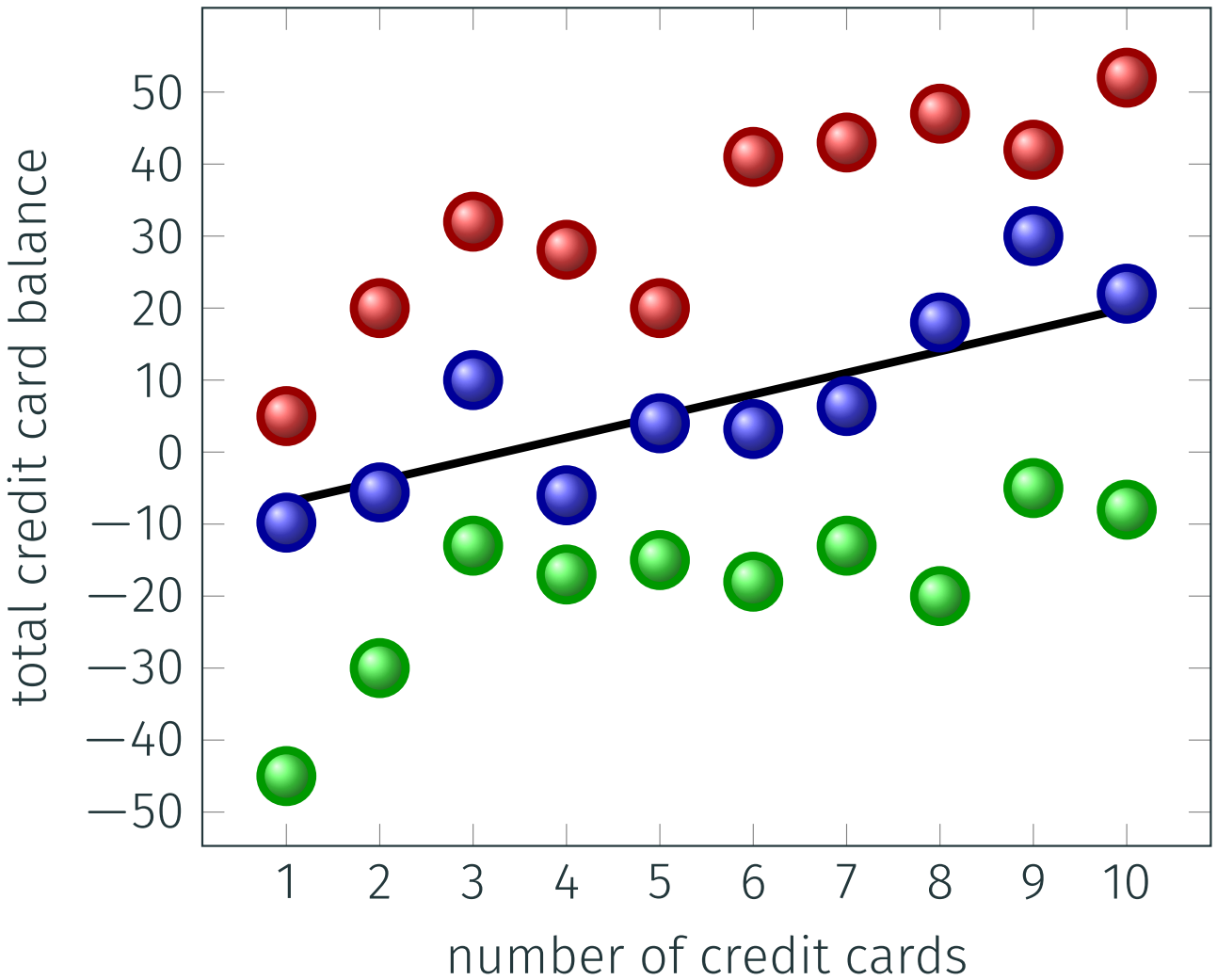


(c) Students

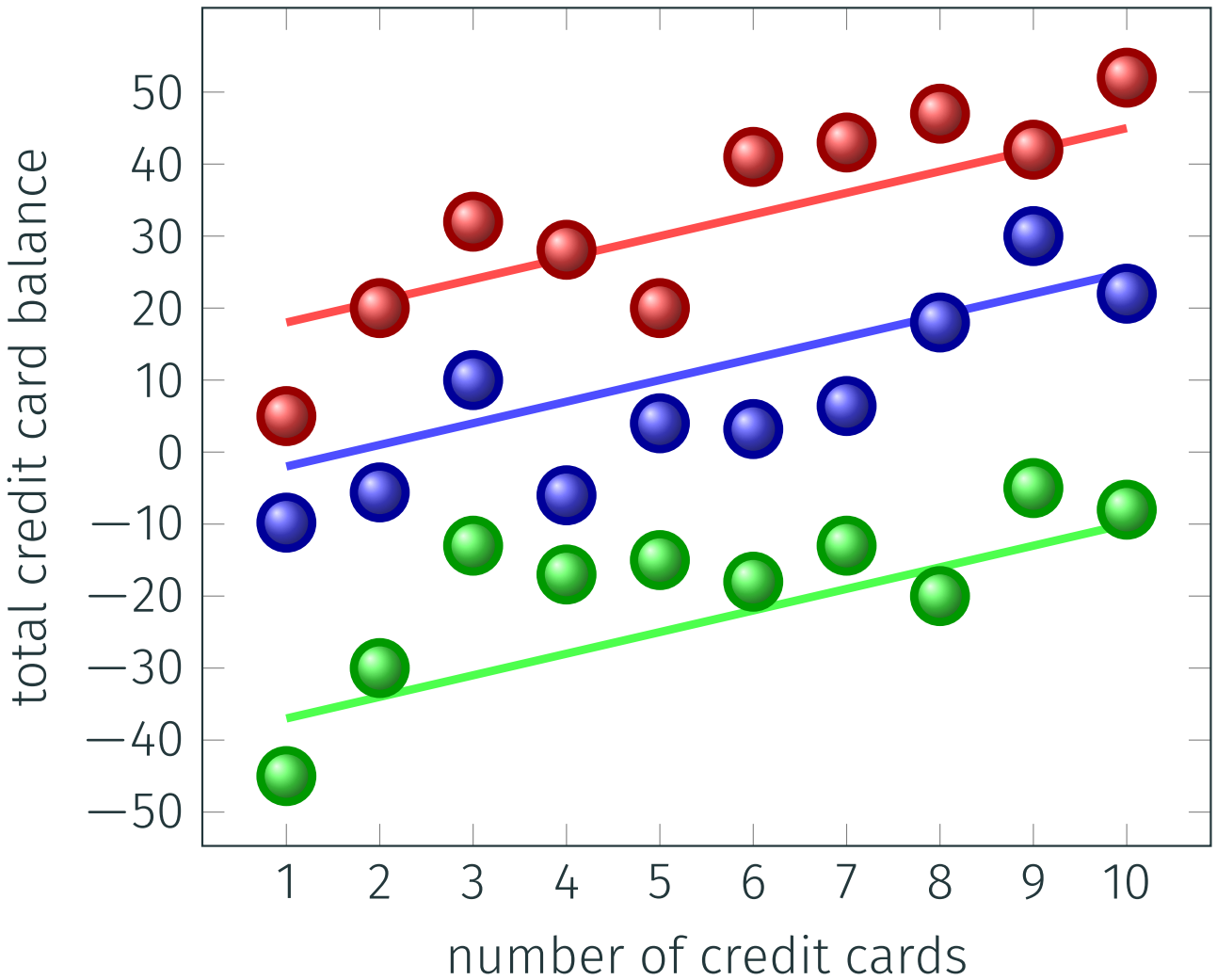
Linear Model



Linear Model

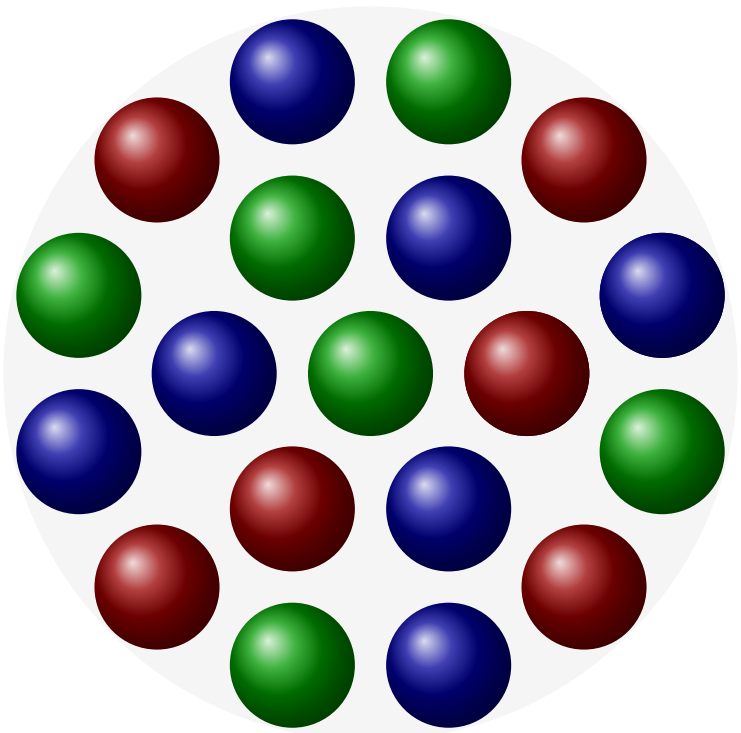


Random Intercept

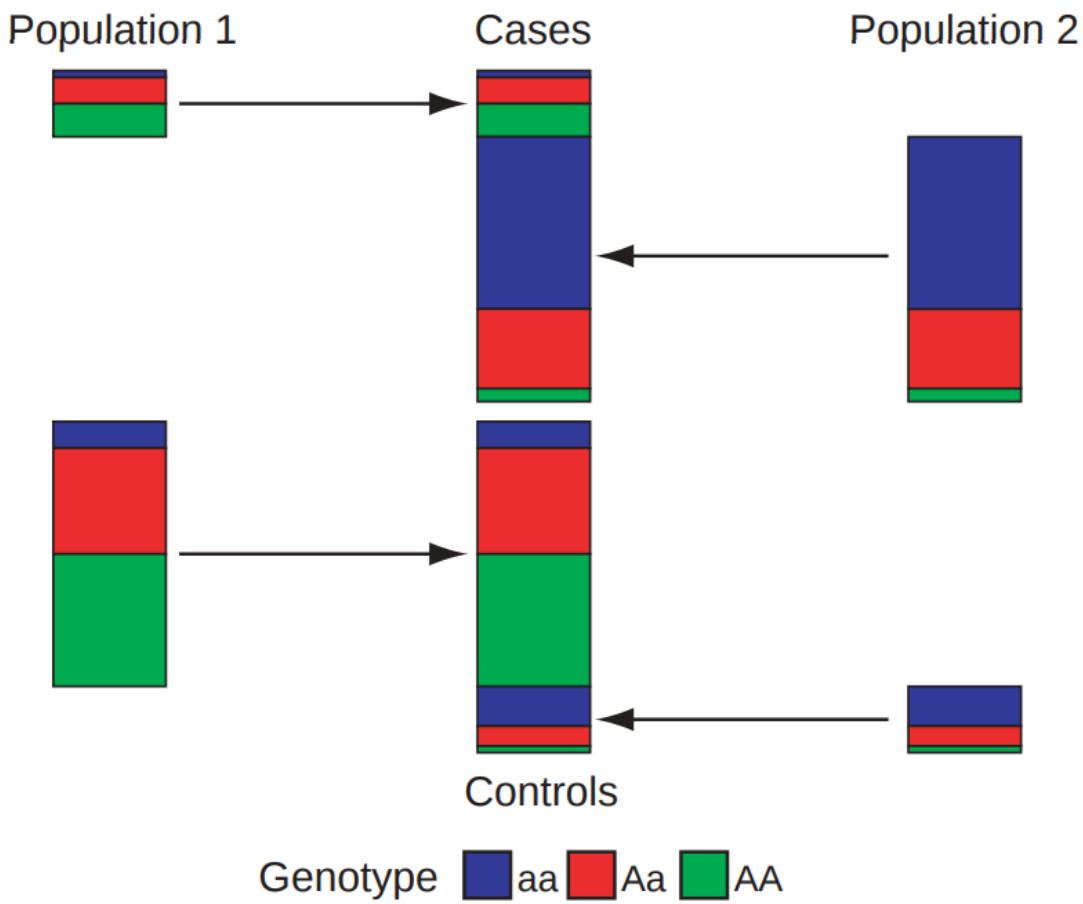


When Grouping Information is Unknown

- In our applications, the grouping information is unknown
- It must be estimated from the data



Motivation



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