Sparse Additive Interaction Learning

Sahir Bhatnagar

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

November 16, 2021

https://sahirbhatnagar.com/sail

https://sahirbhatnagar.com

McGill
UNIVERSITY

Outline

Introduction

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

Overarching reaserch focus: including prior information

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \{ \text{DataFitting}\,[\mathbf{X}, y, \beta] + \color{red}{\lambda\ \text{Prior}\,[\beta]}\color{black}{\}}$$

Overarching reaserch focus: including prior information

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

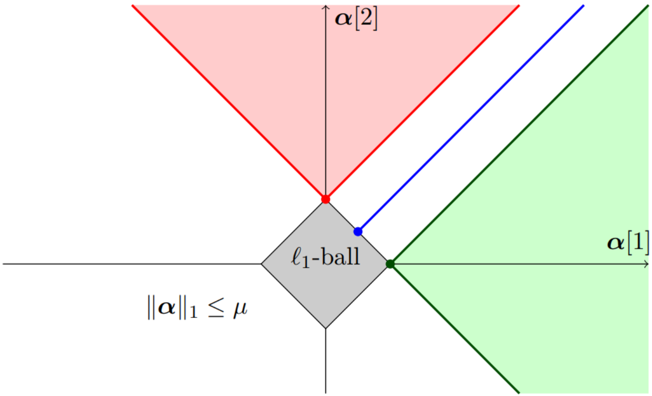
Examples:

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

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

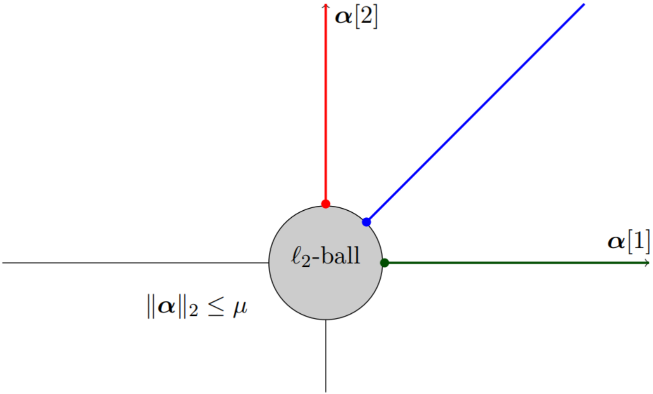
$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \qquad \text{(Ridge regession)}$$

Effect of the Euclidean projection onto the ℓ_1-ball



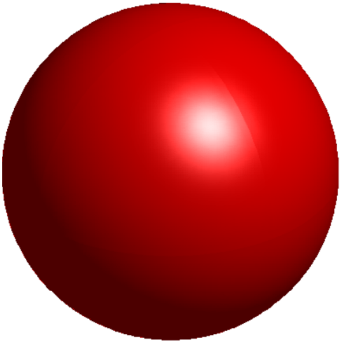
[1] Mairal, Bach and Ponce (2012). Sparse Modeling for Image and Vision Processing.

Effect of the Euclidean projection onto the ℓ_2-ball

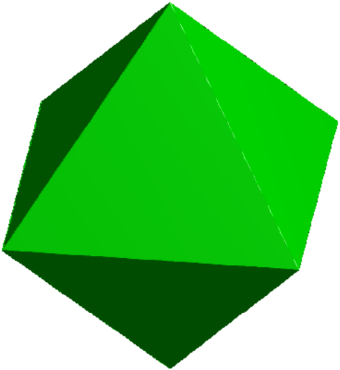


[1] Mairal, Bach and Ponce (2012). Sparse Modeling for Image and Vision Processing.

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.


High-dimensional data:
• Genomics
• Radiomics
• Neuroimaging
Penalized Regression
GLMM
Accelerated gradient for nonconvex penalties JCGS (2021+)
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 DTR 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

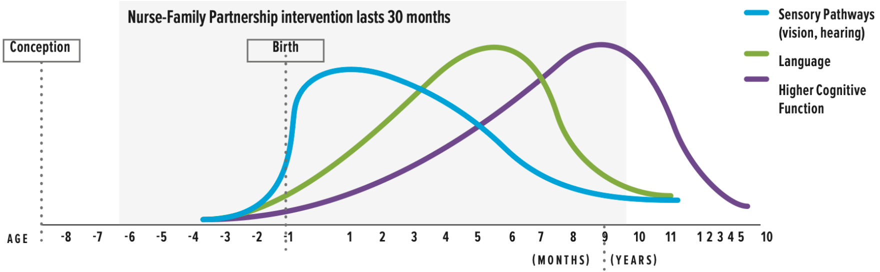
Nurse-Family Partnership

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).

ALL CAUSE MORTALITY OVER 20 YEAR FOLLOW-UP

3x Mothers who did not receive nurse home visits were nearly **3 times more likely to die** from all causes of death than nurse visited mothers (3.7% versus 1.3%)[1]

8x Mothers that did not receive nurse home visits were **8 times more likely to die** from external causes – including unintentional injuries, suicide, drug overdose and homicide – than nurse visited mothers (1.7% versus 0.2%)[1]

PREVENTABLE CHILD MORTALITY OVER 20 YEAR FOLLOW-UP

- Among Nurse-Family Partnership participants, there were **lower rates of preventable child mortality** from birth until age 20.[1]
- 1.6% of the children not receiving nurse home visits died from preventable causes – including sudden infant death syndrome, unintentional injuries and homicide – while none of the nurse visited children died from these causes.[1]

Additional Maternal and Child Health Outcomes

Maternal Health Outcomes

35% fewer cases of pregnancy-induced hypertension[5]

18% fewer preterm births[6]

79% reduction in preterm delivery among women who smoke cigarettes[7]

31% reduction in very closely spaced (<6 months) subsequent pregnancies[8]

Child Health Outcomes

48% reduction in child abuse and neglect[9]

39% fewer health care encounters for injuries or ingestions in the first 2 years of life among children born to mothers with low psychological resources[10]

67% less behavioral and intellectual problems in children at age 6[11]

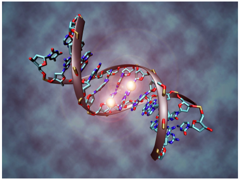
56% fewer emergency room visits for accidents and poisonings through age 2[12]

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

$\sim$



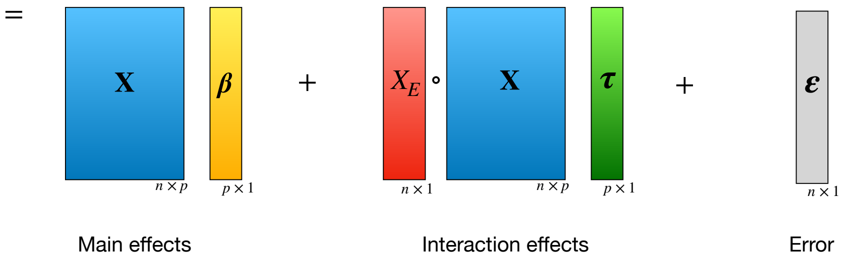
$\times$

Phenotype
IQ Score

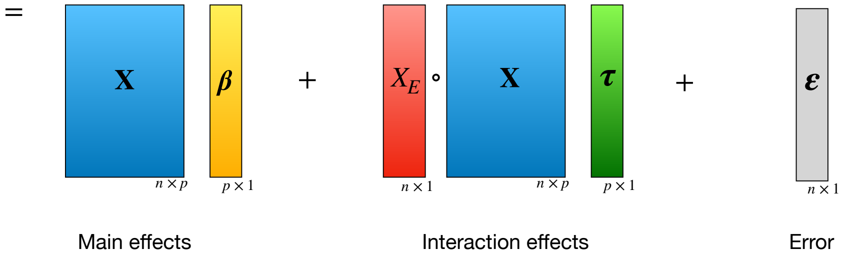
Large Data
Genetic Markers

Environment
NFP Intervention

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



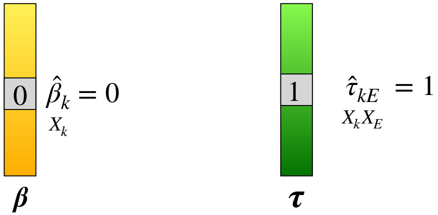
$$Y = \sum_{j=1}^{p} X_j \beta_j + \sum_{j=1}^{p} X_j X_E \tau_j + \varepsilon$$



Let $\quad Z_{jE} = X_E X_j$

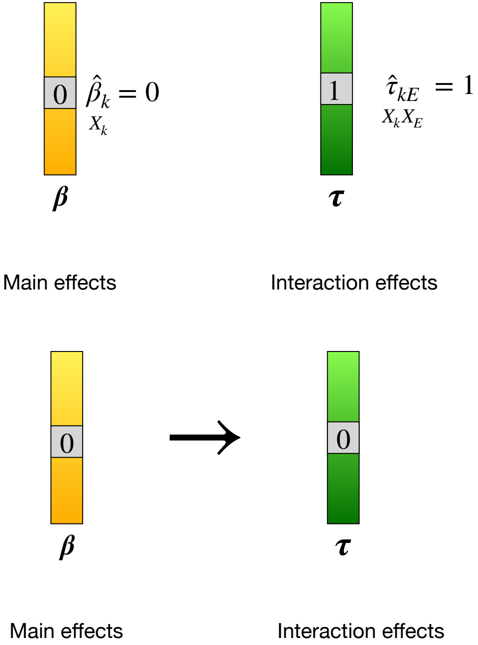
$$Y = \sum_{j=1}^{p} X_j \beta_j + \sum_{j=1}^{p} Z_{jE} \tau_j + \varepsilon$$

$$= \underbrace{\begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix}}_{n \times 2p} \underbrace{\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\tau} \end{bmatrix}}_{2p \times 1} + \underbrace{\boldsymbol{\varepsilon}}_{n \times 1}$$

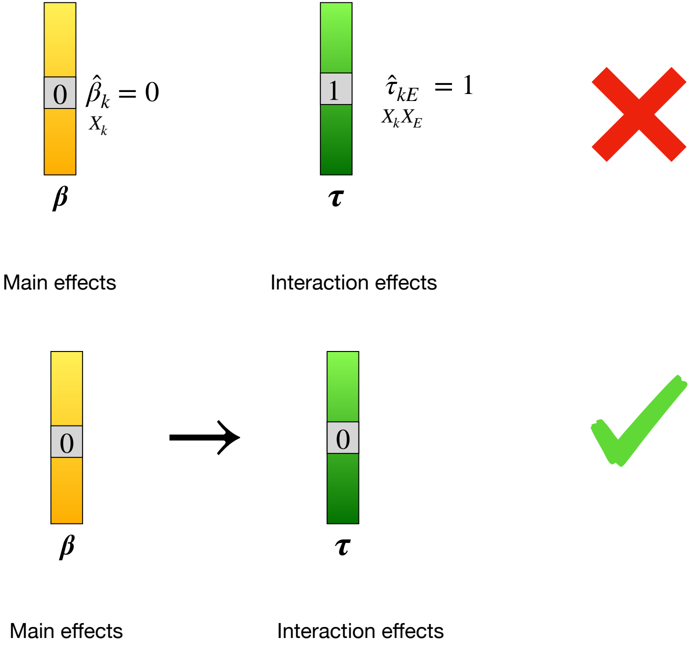


Main effects

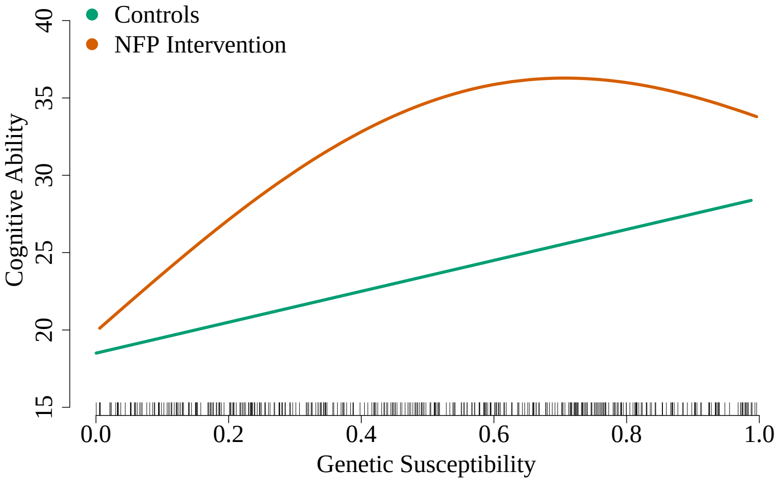
Interaction effects



0
$\hat{\beta}_k = 0$
X_k
$\boldsymbol{\beta}$
1
$\hat{\tau}_{kE} = 1$
$X_k X_E$
$\boldsymbol{\tau}$
Main effects
Interaction effects
0
$\boldsymbol{\beta}$
0
$\boldsymbol{\tau}$
Main effects
Interaction effects


0
$\hat{\beta}_k = 0$
X_k
$\boldsymbol{\beta}$
1
$\hat{\tau}_{kE} = 1$
$X_k X_E$
$\boldsymbol{\tau}$
Main effects
Interaction effects
0
$\boldsymbol{\beta}$
0
$\boldsymbol{\tau}$
Main effects
Interaction effects

Non-linear Interactions



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)`
	`LassoBacktracking` (Shah 2018, JMLR)	`LassoBT(x, y)`
Non-linear	`VANISH` (Radchenko and James 2010, JASA)	✗
	`sail` (Bhatnagar et al. 2020+, in revision CSDA)	`sail(x, e, y, basis)`

Our Extension to Nonlinear Effects

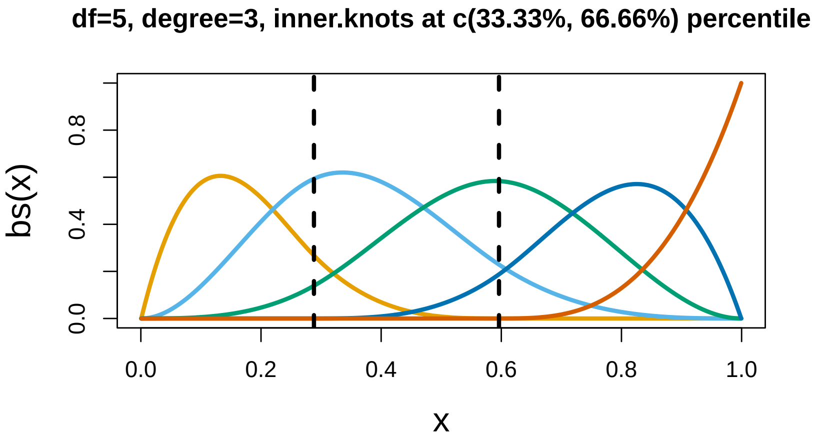
Consider the basis expansion

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

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

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

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

sail: Strong Heredity

Reparametrization[1]

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

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] Choi et al. JASA (2010)

`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

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

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

Implementation

Objective Function

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

[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

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

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

Implementation

Objective Function

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

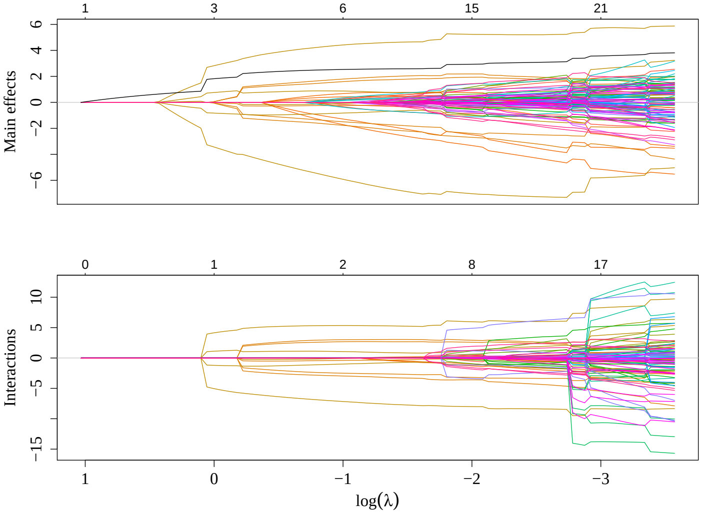
Group Lasso problem

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

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

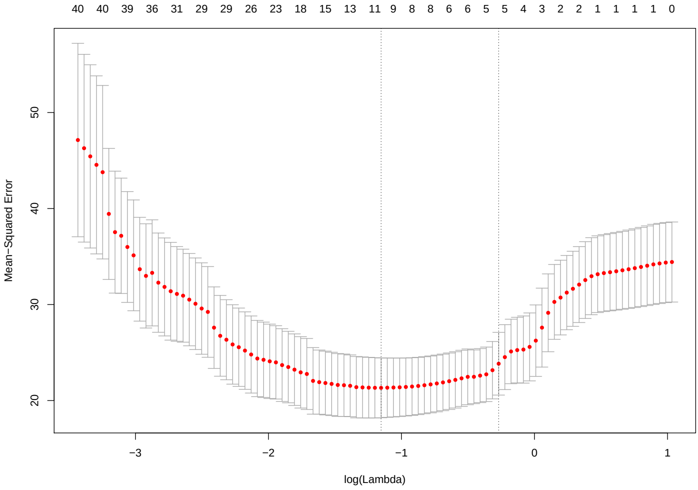
sail R package: Solution Path results

```r
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

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



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

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

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

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

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

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)

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 ($n = 89$).

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.

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

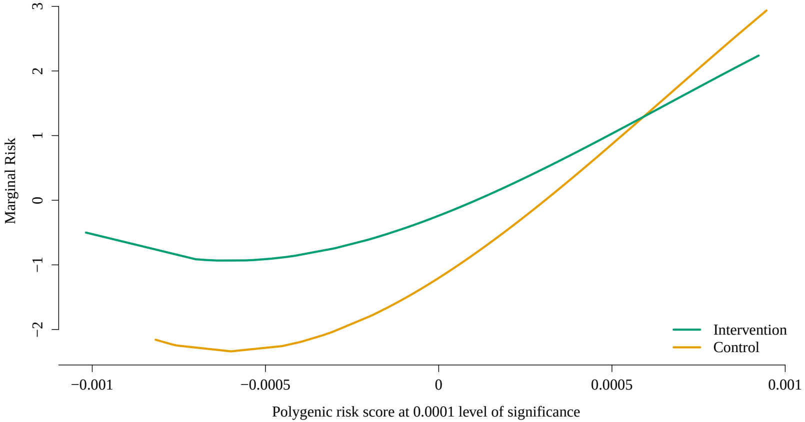


Fig.: The selected model, chosen via 10-fold cross-validation, contained three variables: the main effects for the intervention and the PRS for educational attainment using genetic variants significant at the 0.0001 level, as well as their interaction.

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

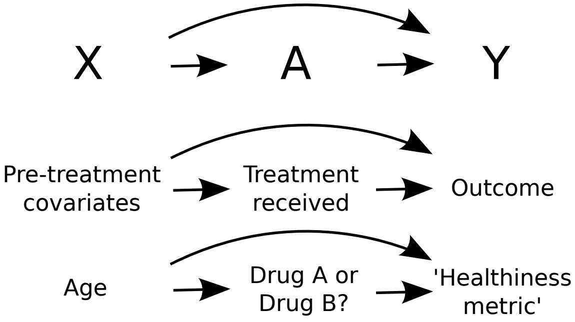
Strengths

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

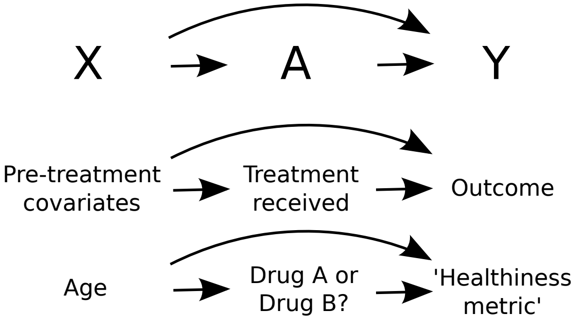
Limitations

- `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

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

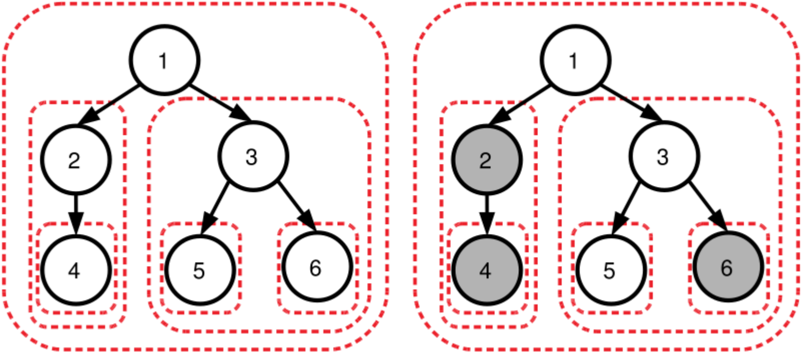
Zeyu Bian, Erica EM Moodie, Susan M Shortreed, Sahir Bhatnagar

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 revision at Biometrics.* https://arxiv.org/abs/2101.07359

Hierarchical Penalty Structure



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

Bi-level selection

- Bi-level selection:

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

Acknowledgements


Zeyu Bian, PhD (c)

CANSSI INCASS

NSERC CRSNG

CIHR IRSC

MiCM 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 revision at Biometrics.*
3. De Leeuw, J. (1994). Block-relaxation algorithms in statistics. In Information systems and data analysis (pp. 308-324). Springer Berlin Heidelberg.
4. 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.
5. Chipman, H. (1996). Bayesian variable selection with related predictors. Canadian Journal of Statistics, 24(1), 17-36.

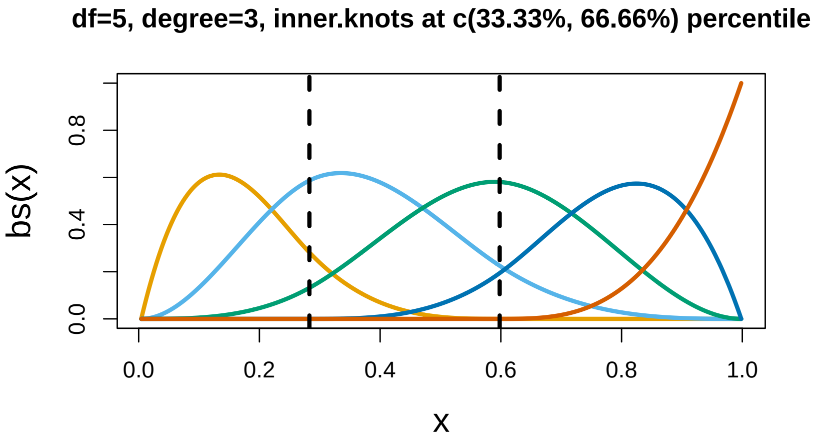
`sahirbhatnagar.com`

Session Info

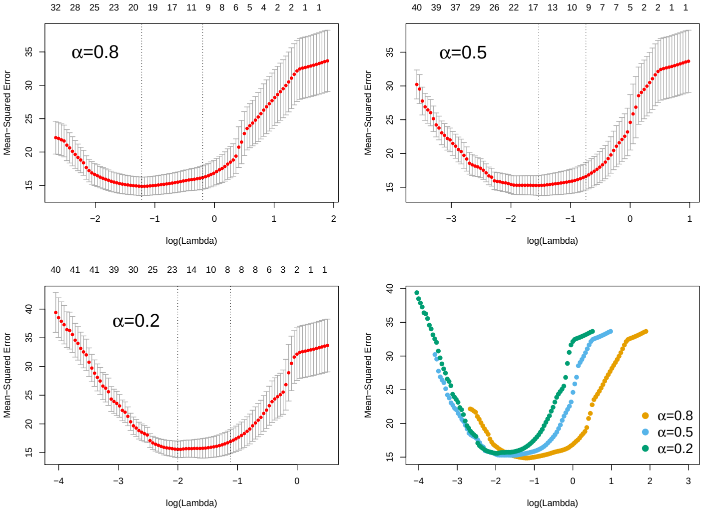
```
R version 4.1.1 (2021-08-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

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] codetools_0.2-18   splines_4.1.1      scales_1.1.1       ellipsis_0.3.2
[33] assertthat_0.2.1   colorspace_2.0-2   utf8_1.2.2         stringi_1.7.5
[37] munsell_0.5.0      truncnorm_1.0-8    crayon_1.4.1
```

B-Spline Expansion

```r
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

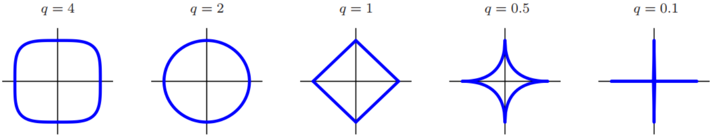


Why the L1 norm ?

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

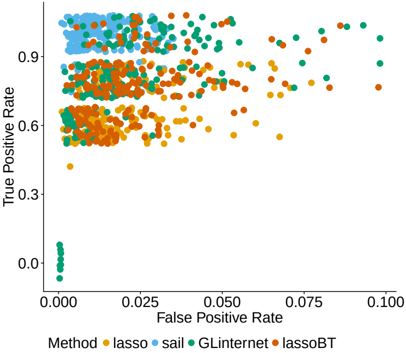
$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

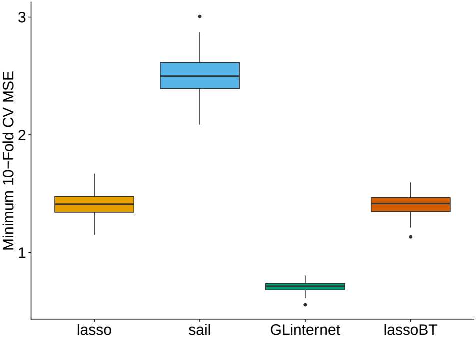
- Why do we use the ℓ_1 norm? Why not use the $q = 2$ (Ridge) or any ℓ_q norm?



- $q = 1$ 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

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

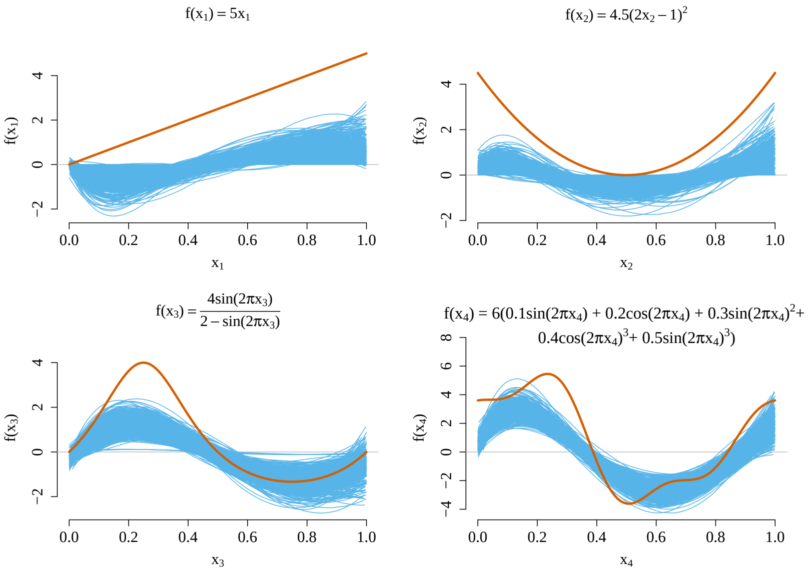
1. **Truth obeys strong hierarchy (right in our wheel house):**

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

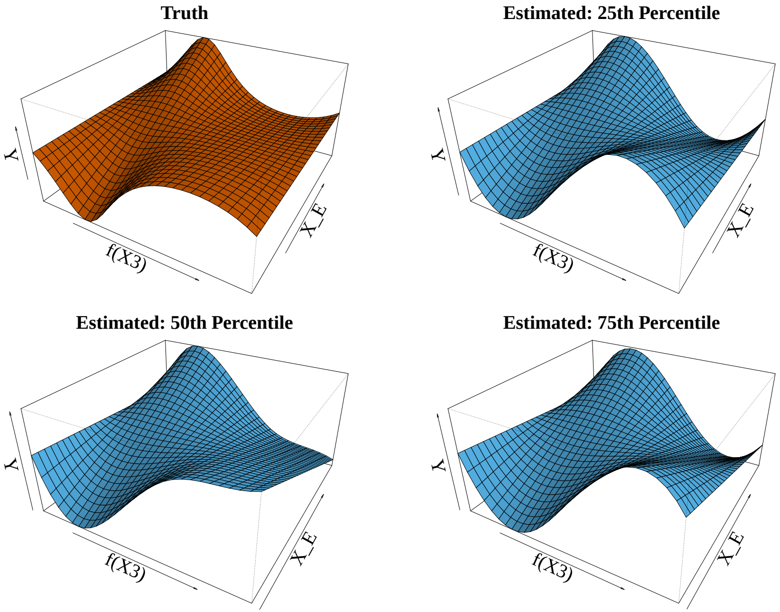
2. **Truth obeys weak hierarchy**
3. **Truth only has interactions**
4. **Truth is linear**
5. **Truth only has main effects**

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

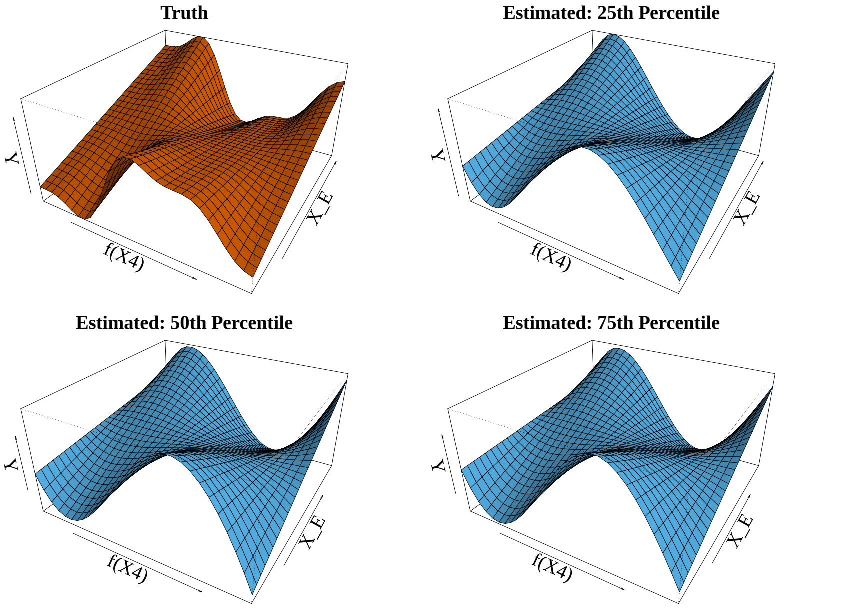
Scenario 1: Main Effects for 500 Simulations



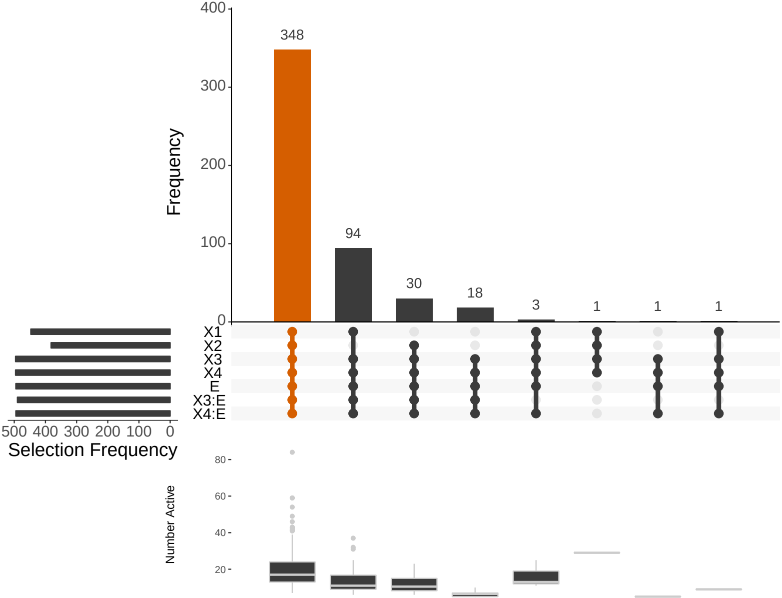
Scenario 1: Estimated Interaction Effects for $E \cdot f(X_3)$



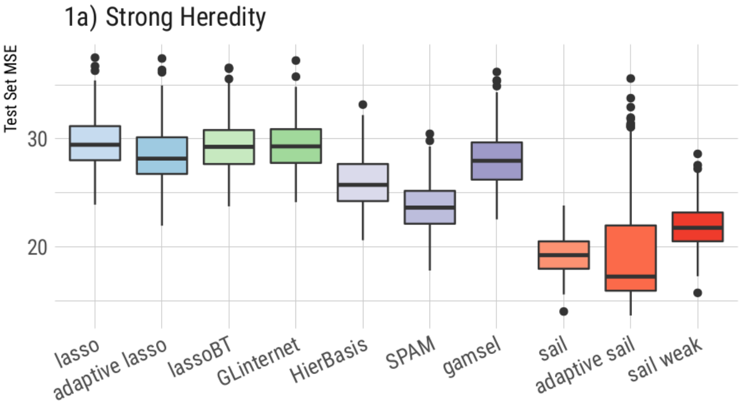
Scenario 1: Estimated Interaction Effects for $E \cdot f(X_4)$



Right in Our Wheel House Simulation Results



Strong Heredity



Main Effects Only