Betting on Sparsity

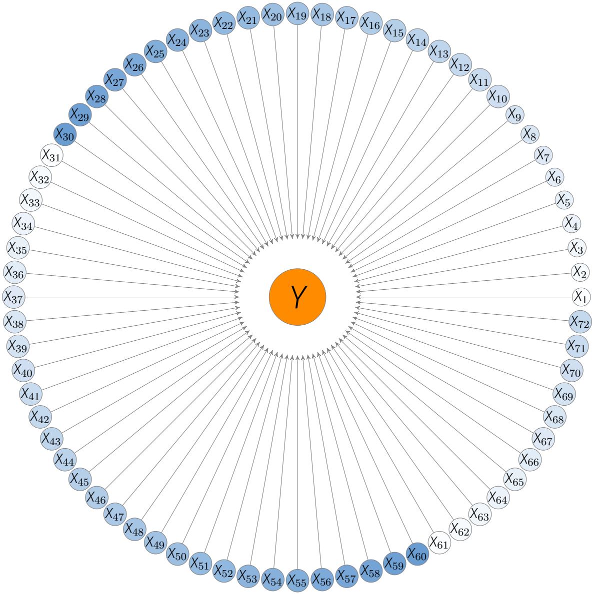
Sahir Bhatnagar, PhD Candidate, McGill Biostatistics

Joint work with Yi Yang and Celia Greenwood (McGill)

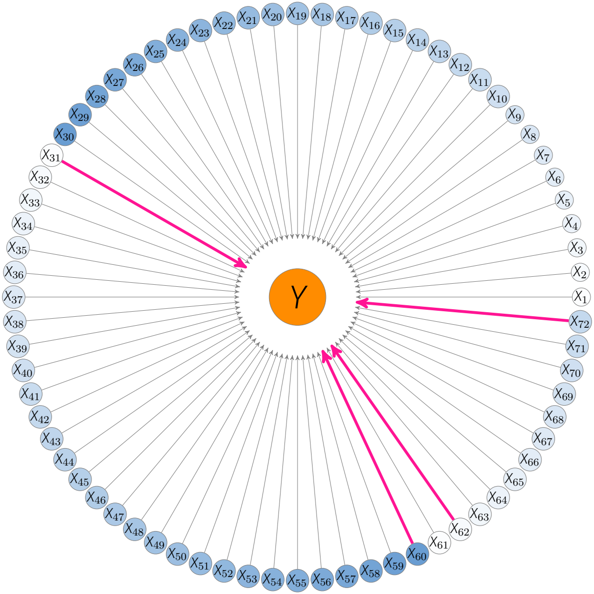
March 1, 2018

Betting on Sparsity

Bet on Sparsity Principle



Bet on Sparsity Principle



Bet on Sparsity Principle

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

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

Bet on Sparsity Principle

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

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

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

A Thought Experiment

How would you schedule a meeting of 20 people?

How would you schedule a meeting of 20 people?

11 participants	March 2017 Thu 9 5:00 PM – 9:00 PM	Fri 10 5:00 PM – 9:00 PM	Sat 11 9:00 AM – 3:00 PM	3:00 PM – 9:00 PM	Sun 12 1:00 PM – 9:00 PM	Mon 13 1:00 PM – 9:00 PM	Tue 14 1:00 PM – 9:00 PM	Wed 15 1:00 PM – 9:00 PM	Thu 16 1:00 PM – 9:00 PM	Fri 17 1:00 PM – 9:00 PM	Sat 18 1:00 PM – 9:00 PM	Sun 19 1:00 PM – 9:00 PM
JayZ	✓	✓	✓			✓			✓	✓	✓	
Evan										✓	✓	✓
Omar	✓	✓		✓		✓			✓	✓	✓	
Caitlin	✓	✓	✓						✓	✓	✓	
Austin	✓	✓	✓									
Ethan			✓	✓					✓		✓	
Max	✓	✓	✓			✓			✓	✓	✓	
Tycho	✓	✓	✓	✓		✓			✓	✓	✓	
Janavi Chadha		✓	✓	✓		✓	✓			✓	✓	
Charlotte											✓	✓
Darshanye	✓	✓				✓			✓	✓		
Your name	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	5:00 PM – 9:00 PM Thu 9 March 2017	5:00 PM – 9:00 PM Fri 10	9:00 AM – 3:00 PM Sat 11	3:00 PM – 9:00 PM	1:00 PM – 9:00 PM Sun 12	1:00 PM – 9:00 PM Mon 13	1:00 PM – 9:00 PM Tue 14	1:00 PM – 9:00 PM Wed 15	1:00 PM – 9:00 PM Thu 16	1:00 PM – 9:00 PM Fri 17	1:00 PM – 9:00 PM Sat 18	1:00 PM – 9:00 PM Sun 19
	7	8	7	4	0	6	1	0	7	8	9	2

Doctors Bet on Sparsity Also

Doctors Bet on Sparsity Also

Motivating Example

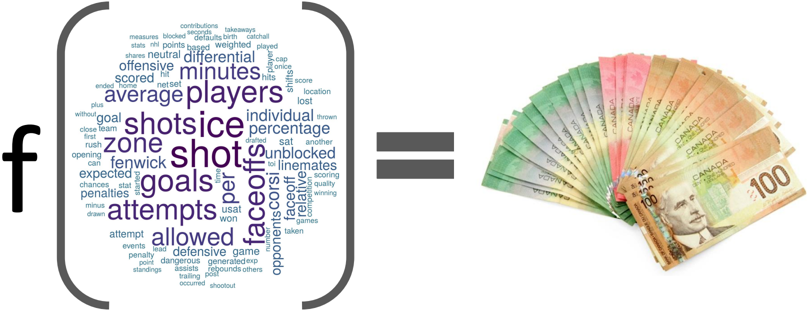
Predictors of NHL Salary[2]

---

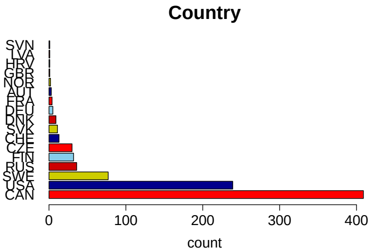
[2] https://www.kaggle.com/camnugent/nhl-salary-data-prediction-cleaning-and-modelling

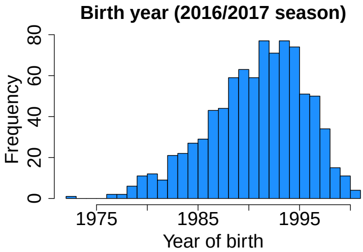
Supervised Learning

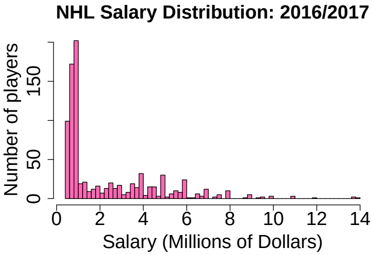
- Learn the function f

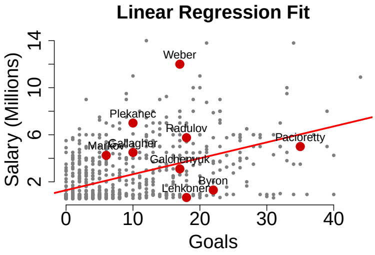


Predictors of NHL Salary

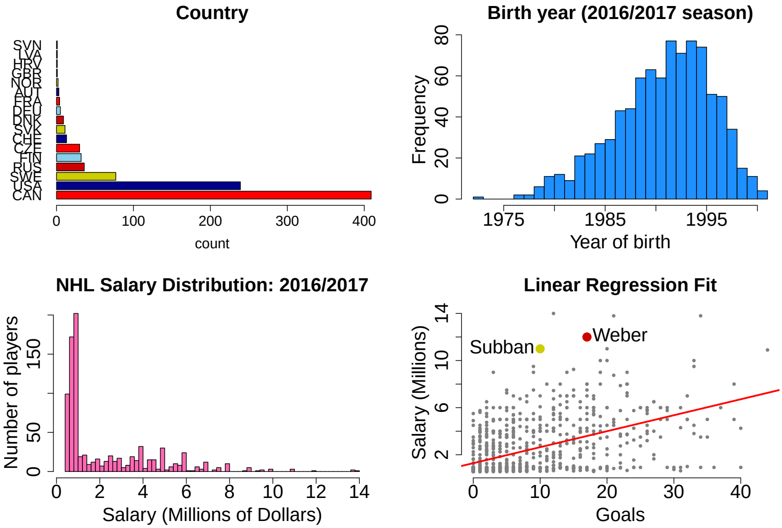




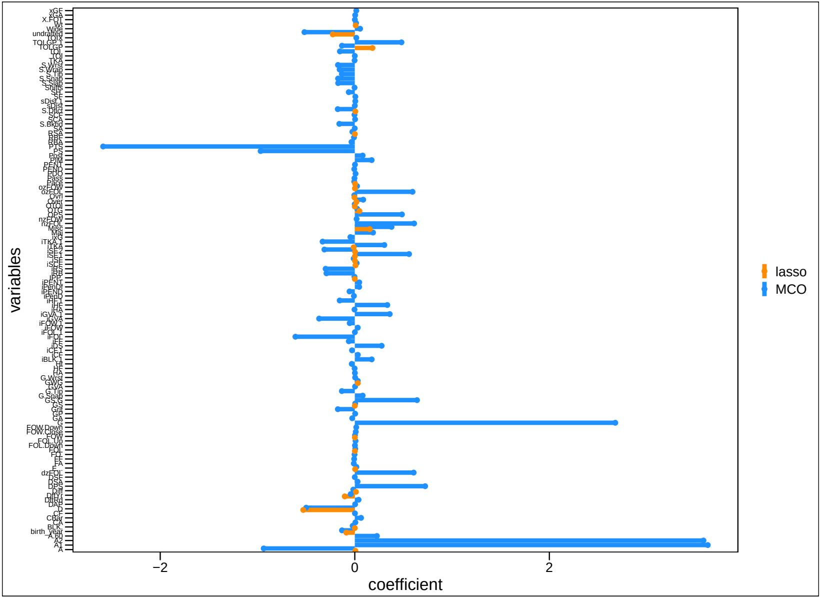




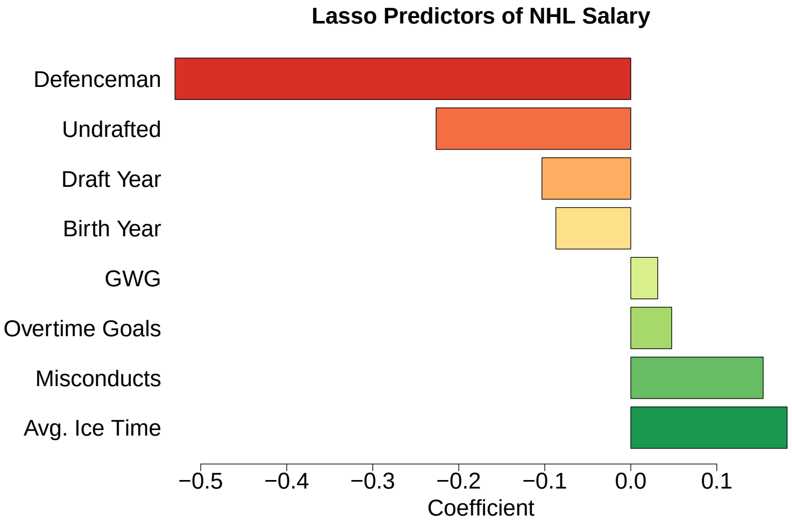
Predictors of NHL Salary



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

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

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

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

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

[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

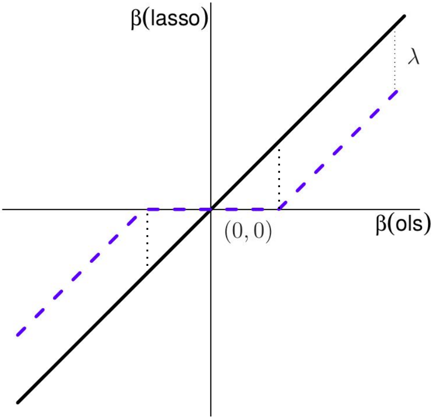
$$\widehat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{1}$$

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

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

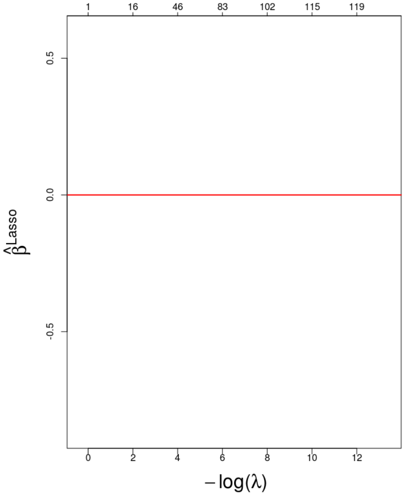
Inspection of the Lasso Solution

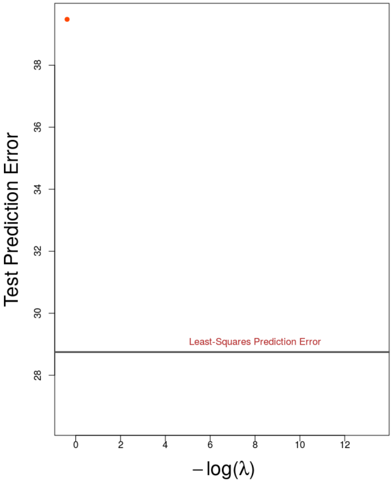
- When the data are standardized, the lasso solution **shrinks the LS estimate toward zero** by the amount λ



[1]Hastie et al. Statistical learning with sparsity: the lasso and generalizations

Choosing the Model Complexity





Group Lasso Illustration

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

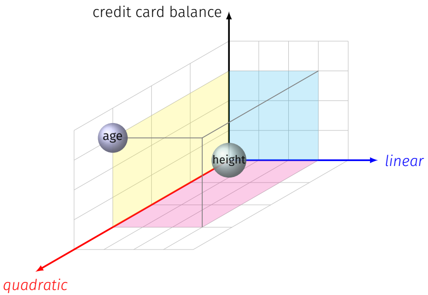
$$\min_{(\beta_0,\boldsymbol{\beta})} \frac{1}{2}\|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k}\|\boldsymbol{\beta}^{(k)}\|_2 \qquad p_k - \text{group size}$$

Group Lasso Illustration

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

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

$$\text{Credit card balance} \sim \text{age} + \text{age}^2 + \text{height} + \text{height}^2$$



(c) Lasso (d) 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 (2018+, preprint)
  https://github.com/sahirbhatnagar/sail
- **ggmix** – Bhatnagar, Oualkacha, Yang, Greenwood (2018+, preprint)
  https://github.com/sahirbhatnagar/ggmix
- **casebase** – Bhatnagar[1], Turgeon[1], Yang, Hanley and Saarela (2018+, preprint)
  https://cran.r-project.org/package=casebase

[1] joint co-authors

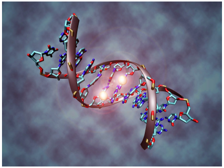
Overview of Our Software Packages

	`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)	$(x, y, \mathbf{\Psi})$	(x, t, δ)

`sail`: Strong Additive Interaction Learning

Motivation 1: Non-linear Interactions

$\sim$



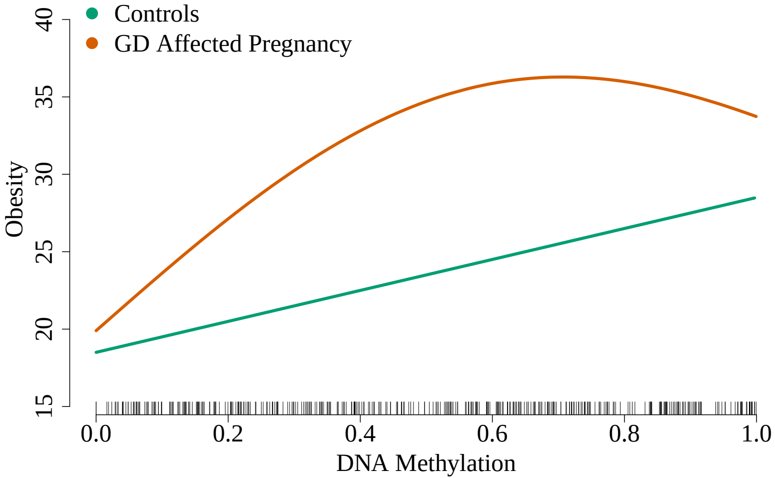
$\times$

Phenotype
Obesity measures

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

Environment
Gestational
Diabetes

Motivation 1: Non-linear Interactions



Motivation 2: Heredity Property

$$Y = \beta_0 \cdot \mathbf{1} + \underbrace{\sum_{j=1}^{p} \beta_j X_j + \beta_E X_E}_{\text{main effects}} + \underbrace{\sum_{j=1}^{p} \alpha_j X_E X_j}_{\text{interactions}} + \varepsilon$$

[1] Chipman. Canadian Journal of Statistics (1996)
[2] McCullagh and Nelder. Generalized Linear Models (1983)
[3] Cox. International Statistical Review (1984)

Motivation 2: Heredity Property

$$Y = \beta_0 \cdot \mathbf{1} + \underbrace{\sum_{j=1}^{p} \beta_j X_j + \beta_E X_E}_{\text{main effects}} + \underbrace{\sum_{j=1}^{p} \alpha_j X_E X_j}_{\text{interactions}} + \varepsilon$$

Strong Heredity[1]

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

- 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)

Lasso interaction model

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

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

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

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

Our Extension to Nonlinear Effects

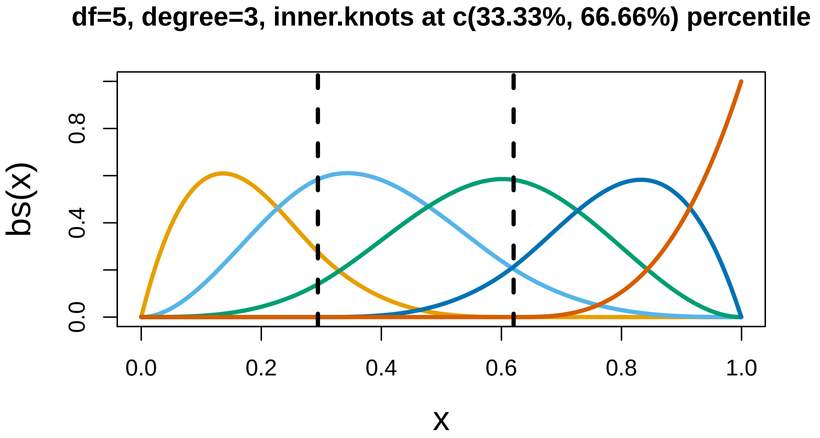
Consider the basis expansion

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

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

B-Spline Expansion

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



`sail`: Additive Interactions

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

Model

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

sail: Strong Heredity

Reparametrization[1]

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

Model

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

Objective Function

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

[1] Choi et al. JASA (2010)

Algorithm

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

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

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

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

    $k \leftarrow k + 1$

  until *convergence criterion is satisfied*;

end

Implementation

Objective Function

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

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

Implementation

Objective Function

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

Lasso problem

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

---

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

Implementation

Objective Function

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

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

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 only has main effects:

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

Simulation Scenarios

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

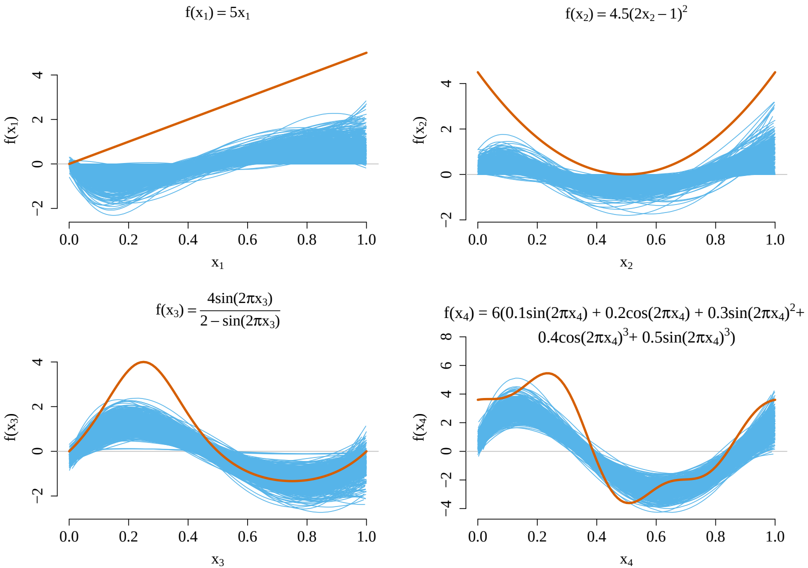
$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

2. Truth only has main effects:

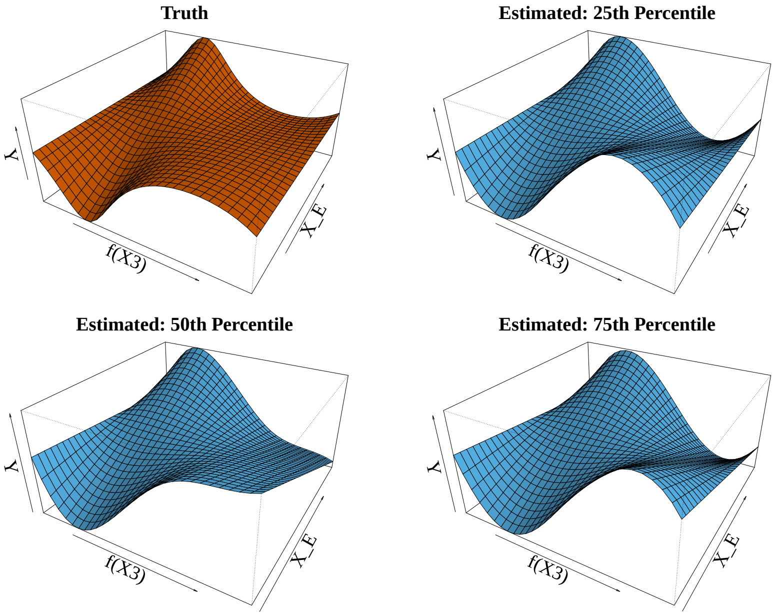
$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + \varepsilon$$

- $n = 200$, $p = 1000$, $\beta_E = 1$, $SNR = 2$
- $X_j \sim$ `truncnorm(0,1)`, $j = 1, \ldots, 1000$,
  $E \sim$ `truncnorm(-1,1)`
- `sail` needs to estimate $1000 \times 5 \times 2 = 10$k parameters

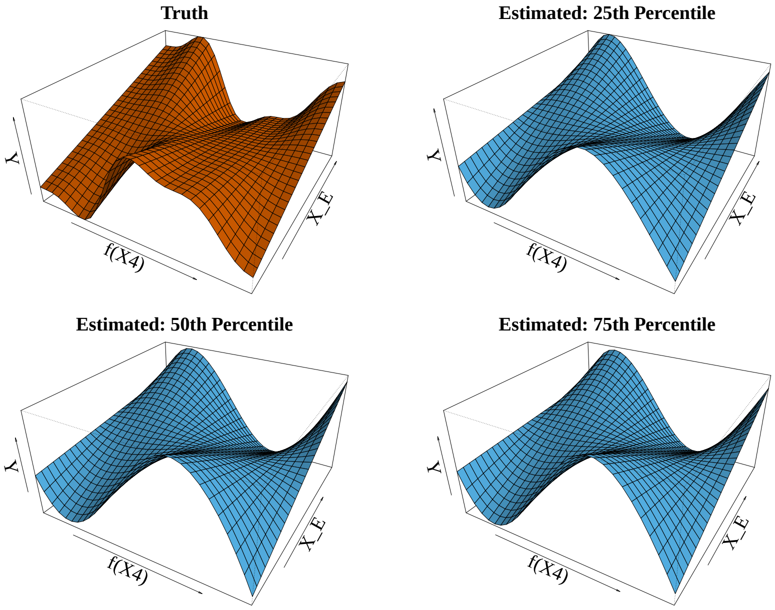
Scenario 1: Main Effects for 500 Simulations



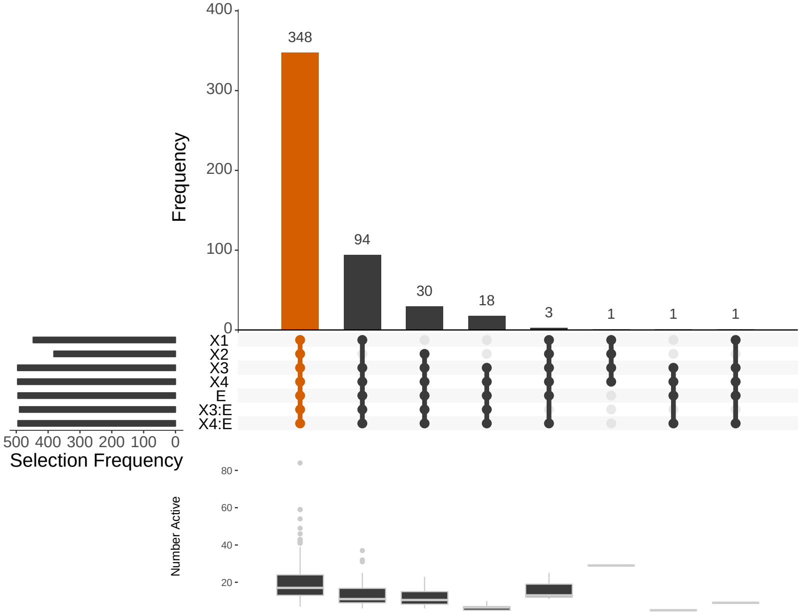
Scenario 1: Estimated Interaction Effects for $E \cdot f(X_3)$



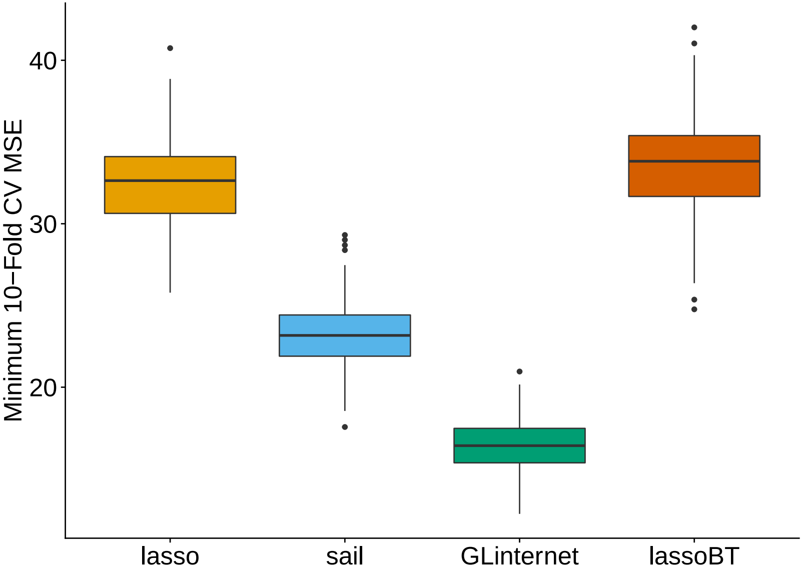
Scenario 1: Estimated Interaction Effects for $E \cdot f(X_4)$



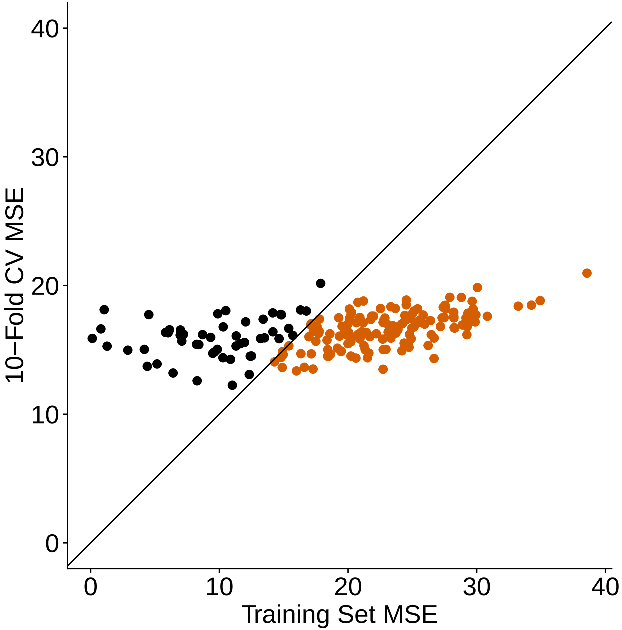
Right in Our Wheel House Simulation Results



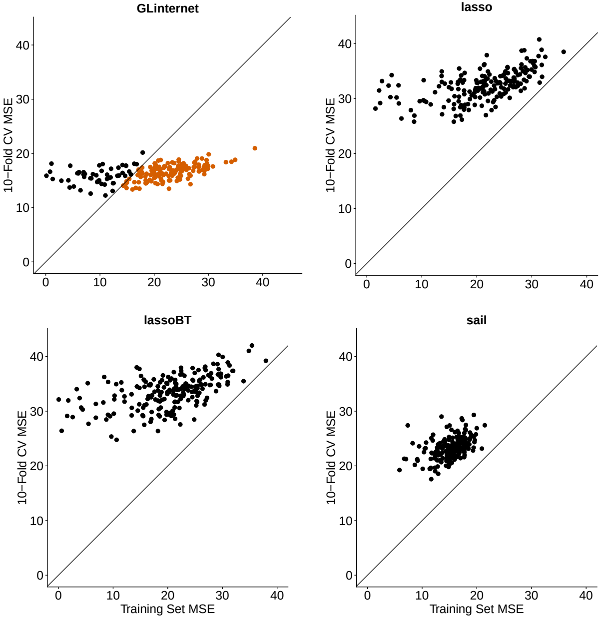
Right in Our Wheel House Simulation - Comparison



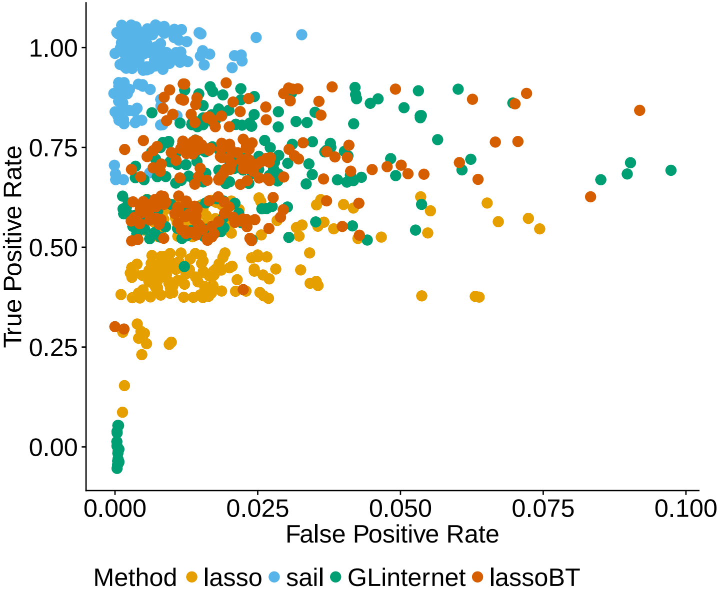
GLinternet: 70% of points below the line



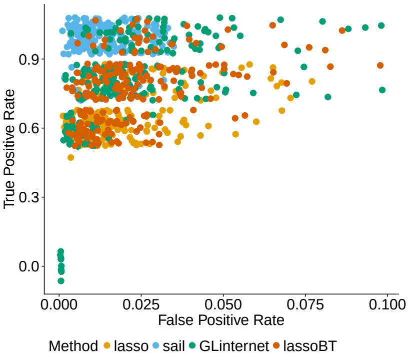
10-Fold CV MSE vs. Training MSE Comparison

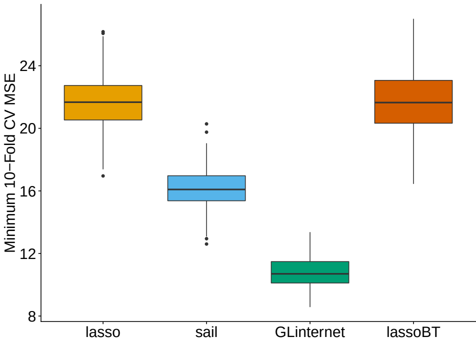


Right in Our Wheel House Simulation - Comparison

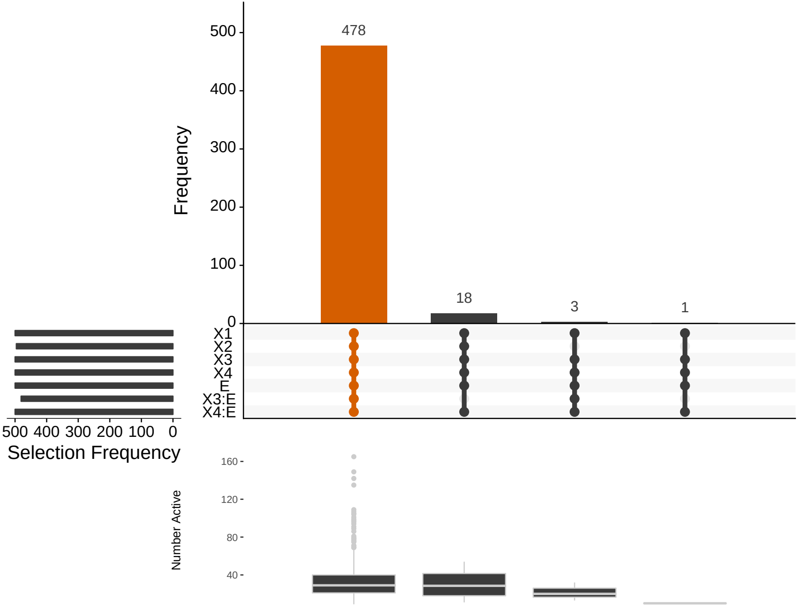


No Interactions Simulation - Comparison

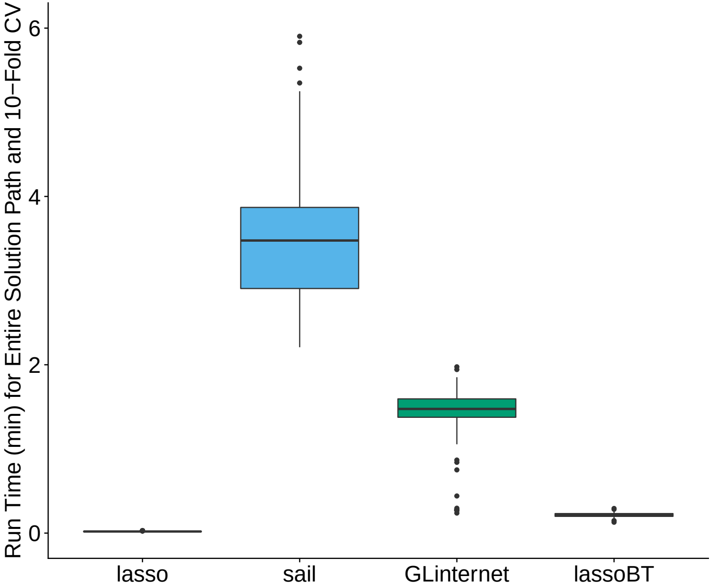




`sail` with `degree=1` when Truth is Linear



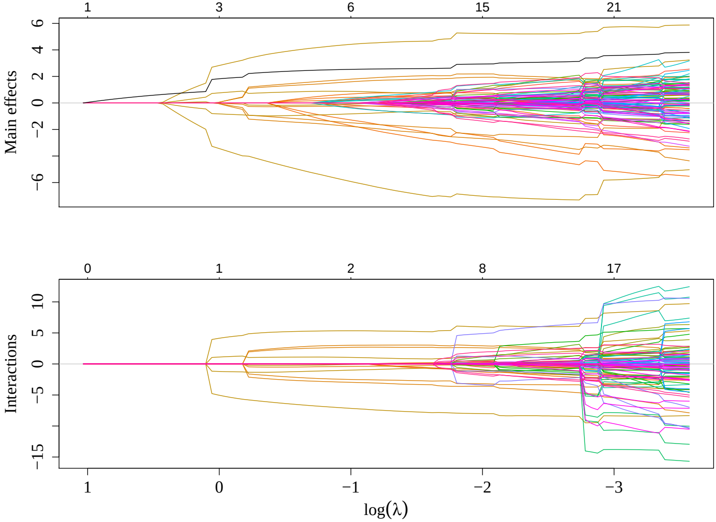
Computing time



`sail` R package

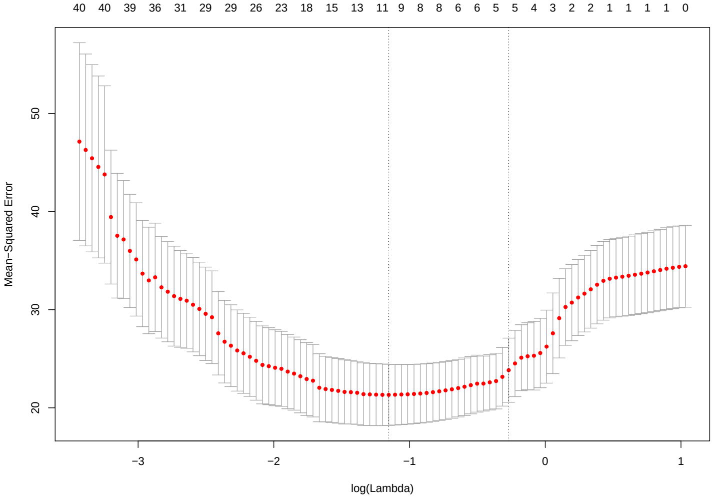
sail R package: Solution Path results

```r
sail::plot(fit)
```

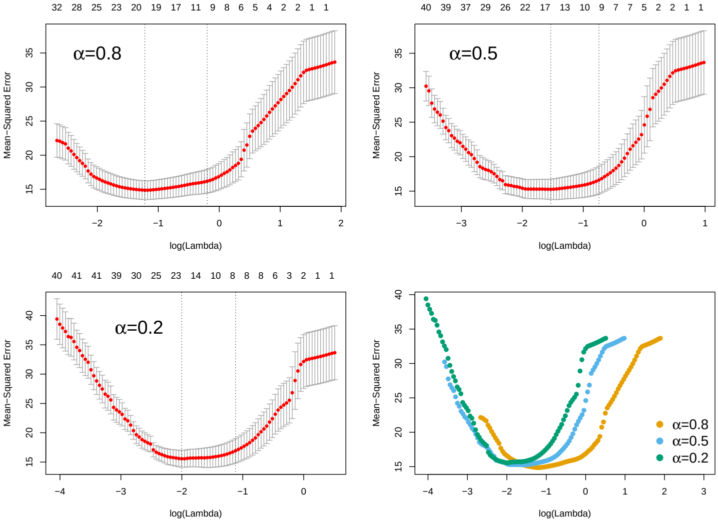


sail R package: Cross-validation results

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



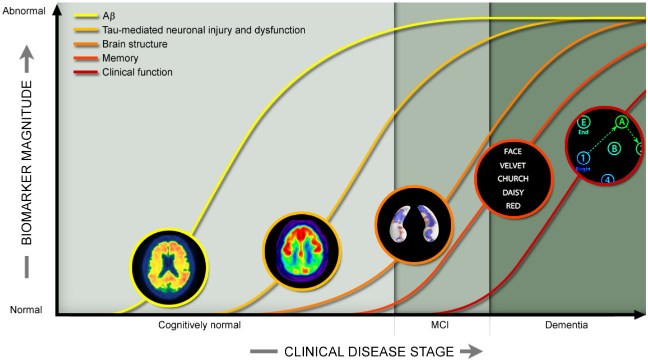
`sail` A Note on the Second Tuning Parameter results



Real Data Application

Alzheimer's Disease Neuroimaging Initiative (ADNI)

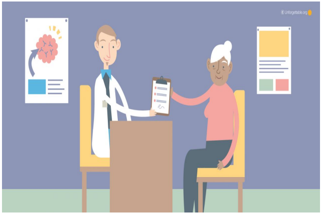
- Alzheimer's is an irreversible neurodegenerative disease that results in a loss of mental function due to the **deterioration of brain tissue**.
- The overall goal of ADNI is to **validate biomarkers** for use in Alzheimer's disease clinical treatment trials



Interaction between Aβ Protein and APOE gene

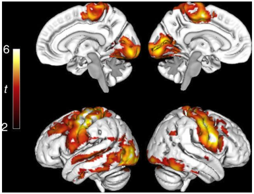
- $\mathbf{E}$: APOE4 allele increases the risk for Alzheimer's and lowers the age of onset
- $\mathbf{X}$: PET amyloid imaging to assess Aβ protein load in 96 brain regions
- $\mathbf{Y}$: General cognitive decline measured by mini-mental state examination
- $3 \times 96 \times 2 + 1 = 577$ parameters to estimate

Mini–Mental State Examination



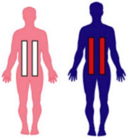
$\sim$

Amyloid Beta acide aminé



$\times$

Gène APOE



$Y_{343 \times 1}$ $\quad$ $X_{343 \times 96}$ $\quad$ $E_{343 \times 1}$

Variable Selection Results: `sail` vs. `lasso`

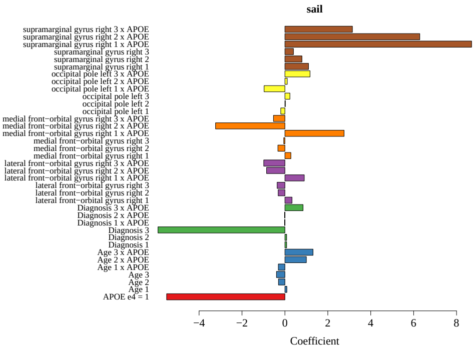


Fig.: `sail`: 7 variables

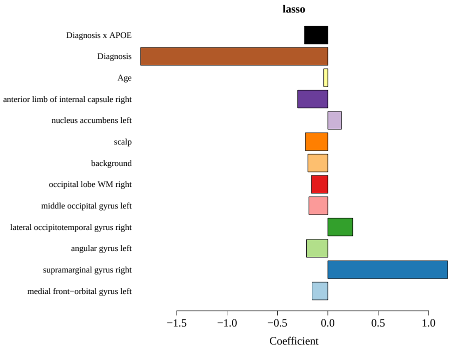
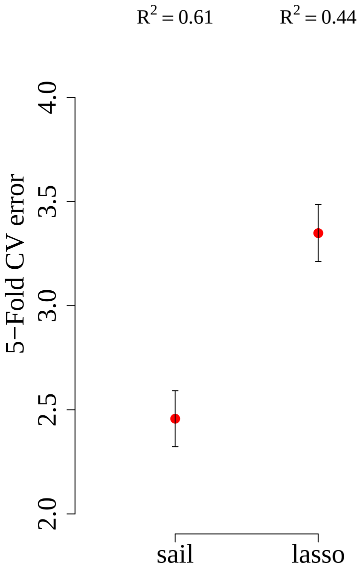
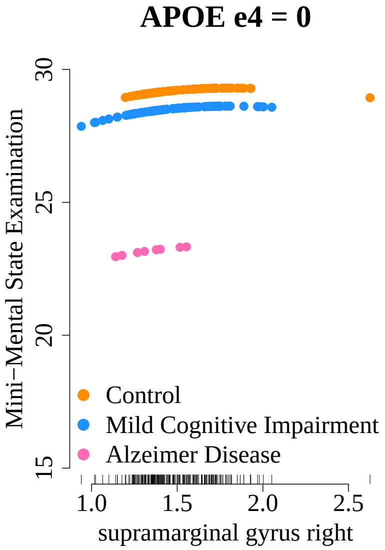


Fig.: `lasso`: 13 variables

5-Fold Cross-Validated MSE



`sail`: Interactions with the `supramarginal gyrus` region



Discussion

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

Future Directions

- Weak heredity property $\rightarrow \alpha_j = \gamma_j(|\beta_j| + |\beta_E|)$
- Implement ADMM algorithm for scalability. Distributed computing (GPU)
- Binary Outcomes
- 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

CENTRE LUDMER CENTER
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

Session Info

```
R version 3.4.1 (2017-06-30)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.3 LTS

Matrix products: default
BLAS: /usr/lib/openblas-base/libblas.so.3
LAPACK: /usr/lib/libopenblasp-r0.2.18.so

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

other attached packages:
[1] xtable_1.8-2        rpart.plot_2.1.2    rpart_4.1-11
[4] data.table_1.10.4-3 ISLR_1.2            ggplot2_2.2.1.9000
[7] knitr_1.19

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.15        magrittr_1.5        splines_3.4.1
 [4] munsell_0.4.3       colorspace_1.3-2    rlang_0.1.6
 [7] stringr_1.2.0       highr_0.6           plyr_1.8.4
[10] tools_3.4.1         grid_3.4.1          gtable_0.2.0
[13] pacman_0.4.6        lazyeval_0.2.1      digest_0.6.15
[16] tibble_1.4.2        RSkittleBrewer_1.1 codetools_0.2-15
[19] evaluate_0.10.1     stringi_1.1.5       compiler_3.4.1
[22] pillar_1.1.0        methods_3.4.1       scales_0.5.0.9000
[25] truncnorm_1.0-7
```

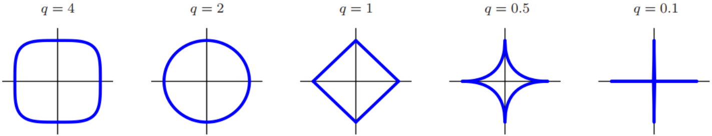
Appendix

Why the L1 norm ?

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

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

- 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