Sparse Additive Interaction Learning

Sahir Bhatnagar

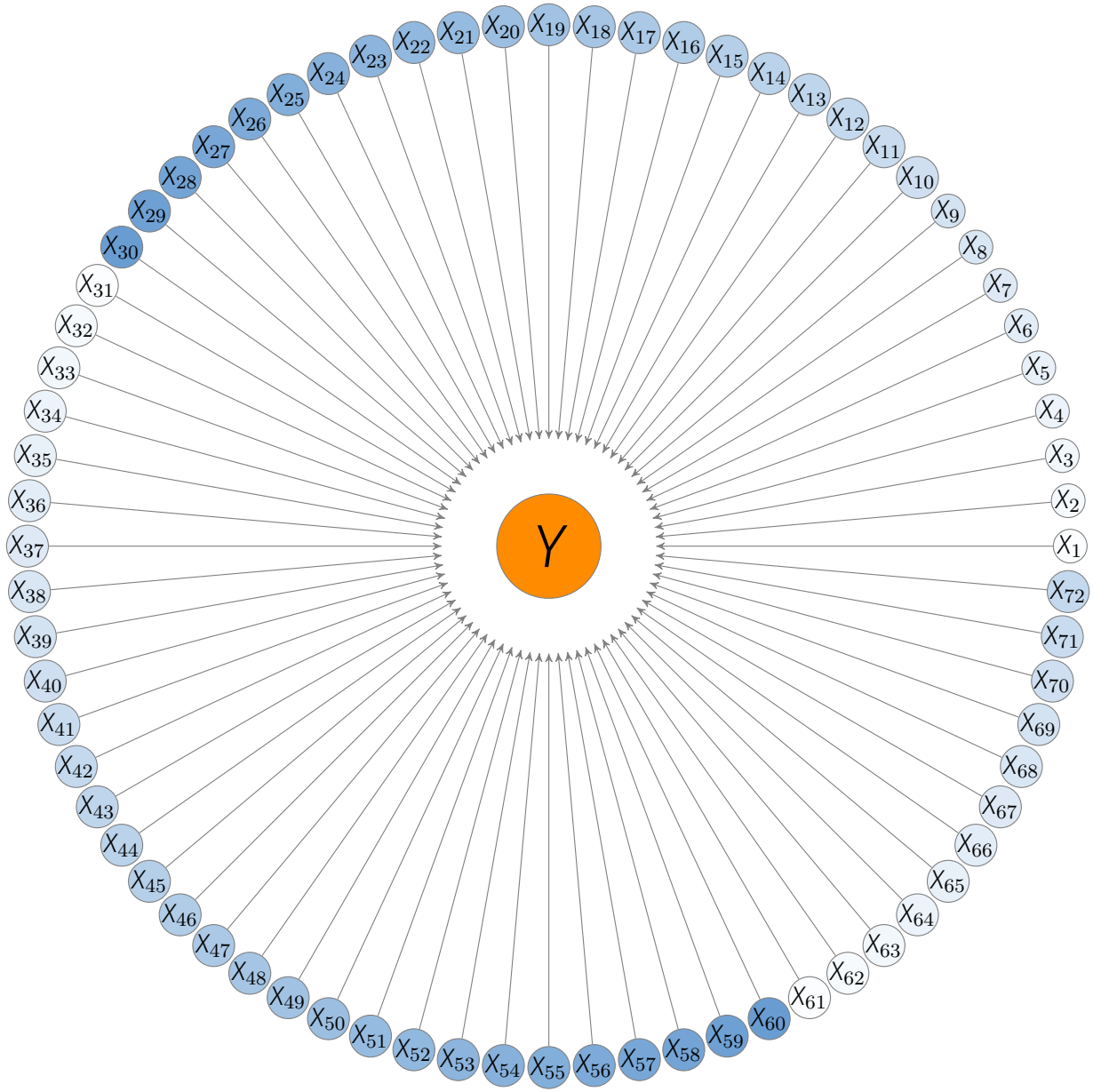
Department of Epidemiology, Biostatistics and Occupational Health
Department of Diagnostic Radiology

Joint work with Yi Yang and Celia Greenwood (McGill)

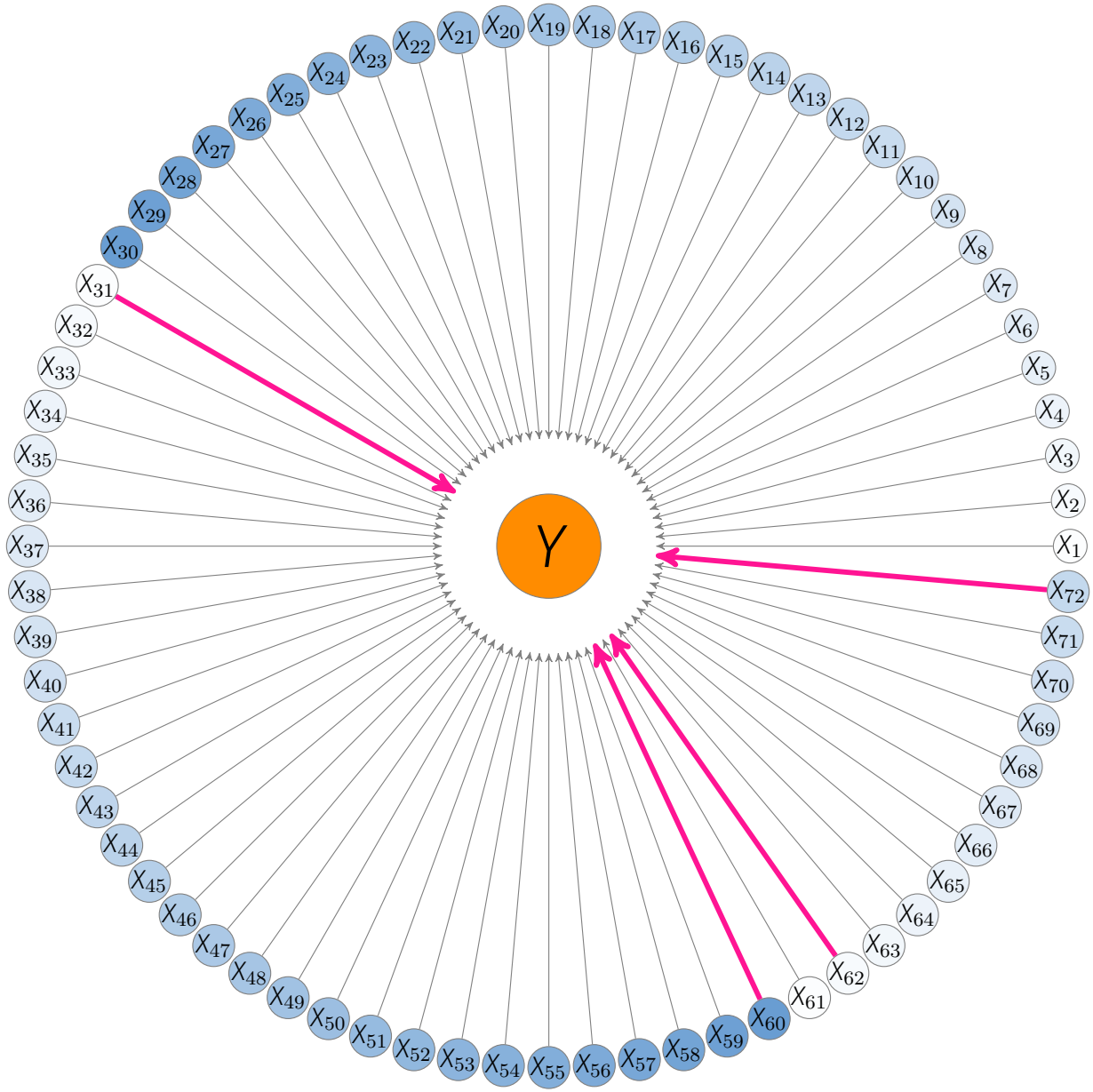
HDDA Conference
June 24, 2019
https://sahirbhatnagar.com/sail

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]

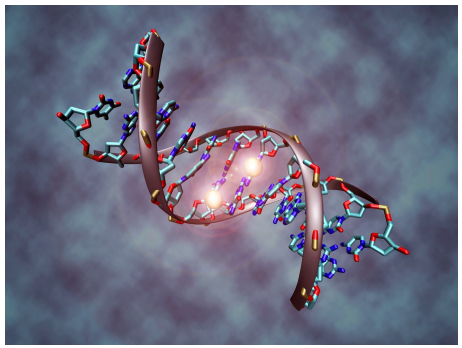
- We often don't have enough data to estimate so many parameters
- Even when we do, we might want to identify a **relatively small number of predictors** ($k < N$) that play an important role
- Faster computation, easier to understand, and stable predictions on new datasets.

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

`sail`: Strong Additive Interaction Learning

Motivation 1: Non-linear Interactions

$\sim$



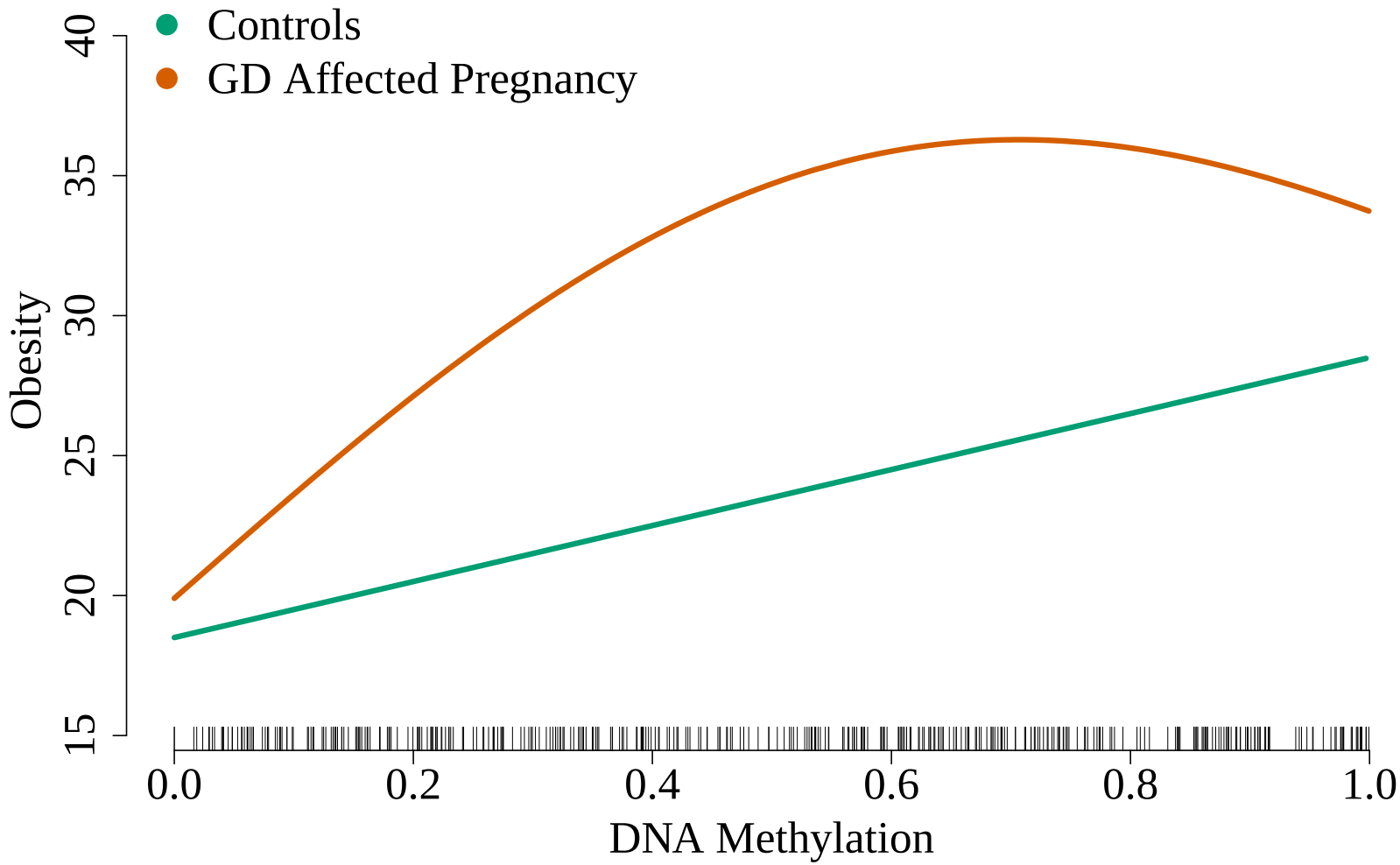
$\times$

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

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

[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

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

Strong Heredity[1]

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

- 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)

Lasso interaction model

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

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

$$\underset{\boldsymbol{\Theta}:=(\beta_0,\boldsymbol{\beta},\boldsymbol{\tau})}{\operatorname{argmin}} \quad \mathcal{L}(\boldsymbol{\Theta}) + \lambda(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\tau}\|_1)$$

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. 2019+, bioRxiv)	`sail(x, e, y, basis)`

Our Extension to Nonlinear Effects

Consider the basis expansion

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

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

`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

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

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

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

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

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

    $k \leftarrow k + 1$

  until *convergence criterion is satisfied*;

end

Implementation

Objective Function

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

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

Implementation

Objective Function

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

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://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

Theory

Sparsity

Theorem 1

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

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

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

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

Sparsity

Theorem 1

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

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

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

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

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

Asymptotic normality

Theorem 2

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

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

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

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

Asymptotic normality

Theorem 2

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

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

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

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

Simulations

Simulation Scenarios

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

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

Simulation Scenarios

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

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

2. Truth obeys weak hierarchy

Simulation Scenarios

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

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

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

Simulation Scenarios

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

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

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

Simulation Scenarios

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

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

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

Simulation Scenarios

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

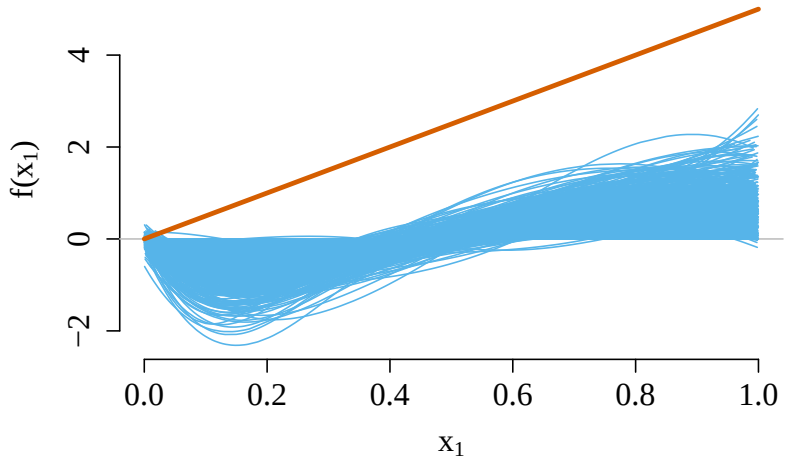
$$Y = \sum_{j=1}^{4} f_j(X_j) + \beta_E \cdot X_E + X_E \times (f_3(X_3) + f_4(X_4)) + \varepsilon$$

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

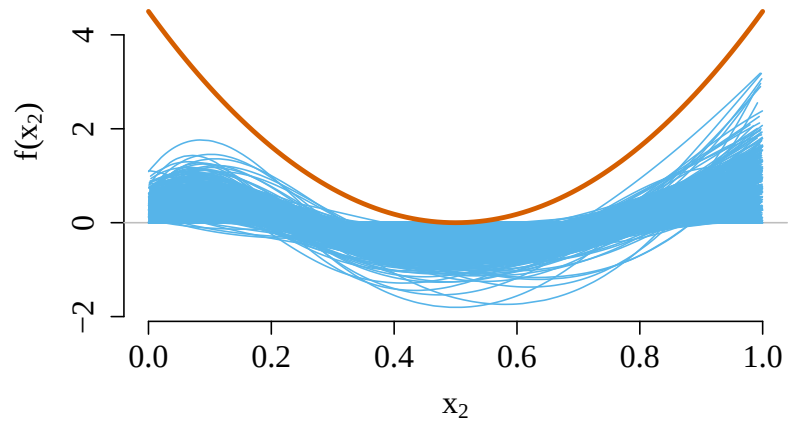
- $n_{train} = n_{tuning} = 200$, $n_{test} = 800$, $p = 1000$, $\beta_E = 1$, $SNR = 2$
- $X_j \sim$ `truncnorm(0,1)`, $j = 1, \ldots, 1000$,
  $E \sim$ `truncnorm(-1,1)`
- `sail` needs to estimate $1000 \times 5 \times 2 = 10$k parameters

Scenario 1: Main Effects for 500 Simulations

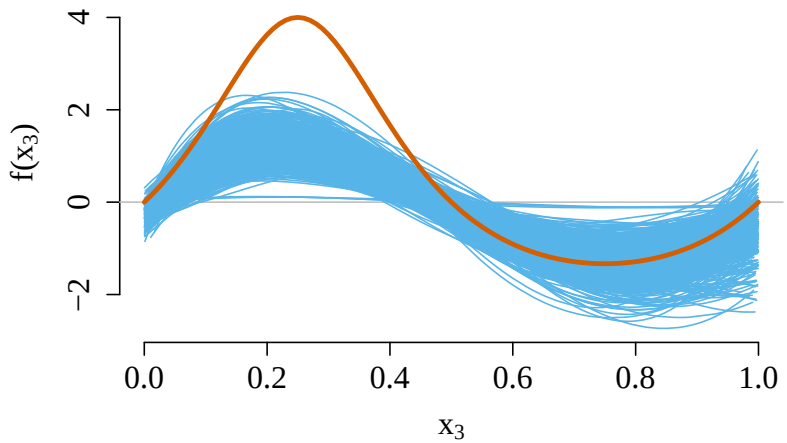
$f(x_1) = 5x_1$



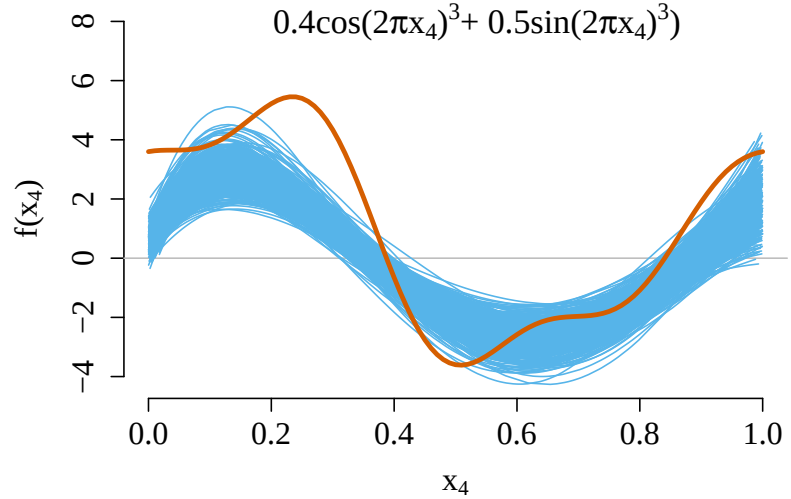
$f(x_2) = 4.5(2x_2 - 1)^2$



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

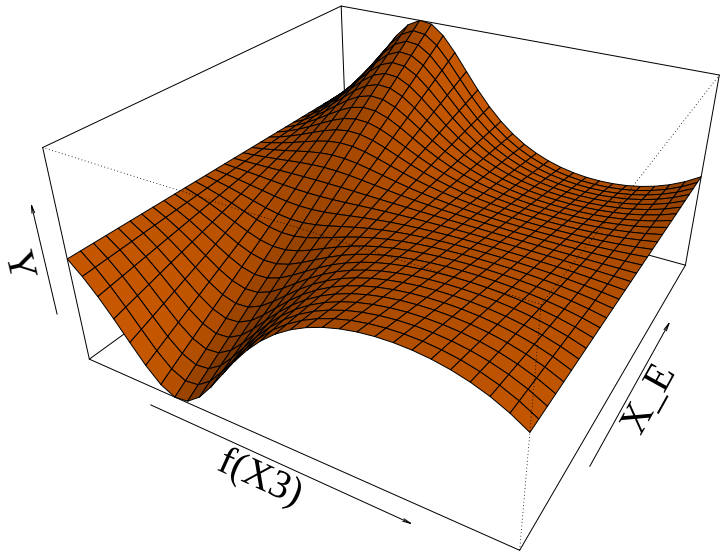


$f(x_4) = 6(0.1\sin(2\pi x_4) + 0.2\cos(2\pi x_4) + 0.3\sin(2\pi x_4)^2 + 0.4\cos(2\pi x_4)^3 + 0.5\sin(2\pi x_4)^3)$

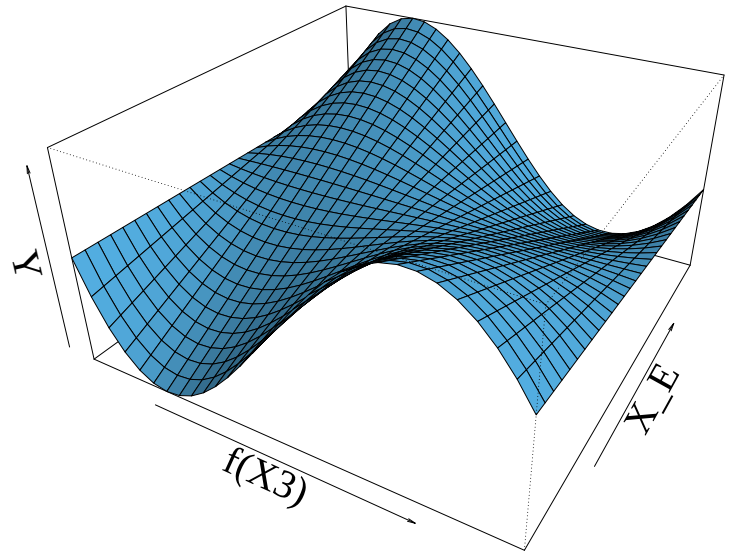


Scenario 1: Estimated Interaction Effects for $E \cdot f(X_3)$

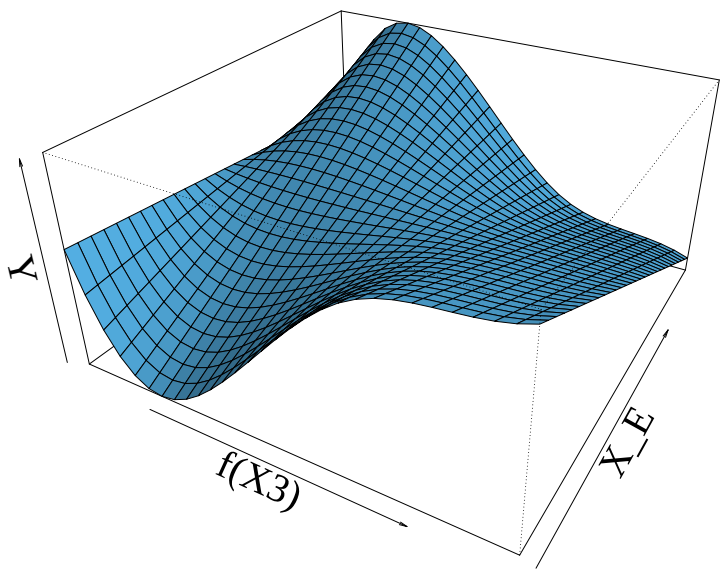
Truth



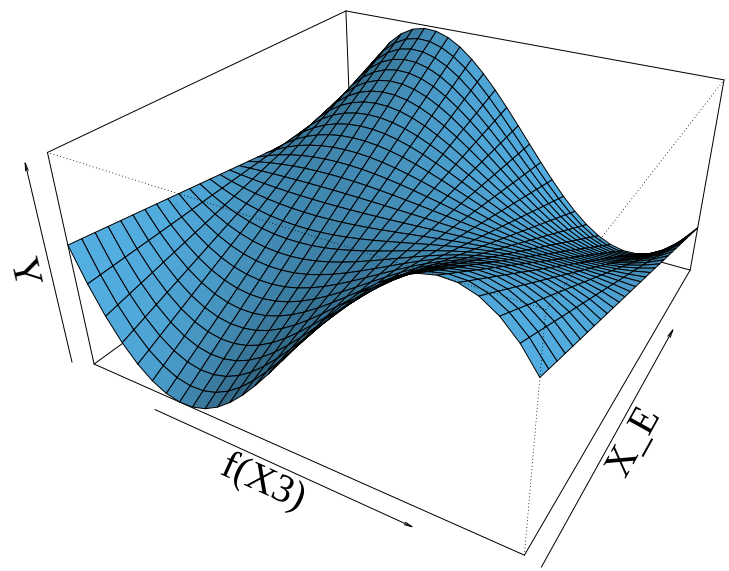
Estimated: 25th Percentile



Estimated: 50th Percentile

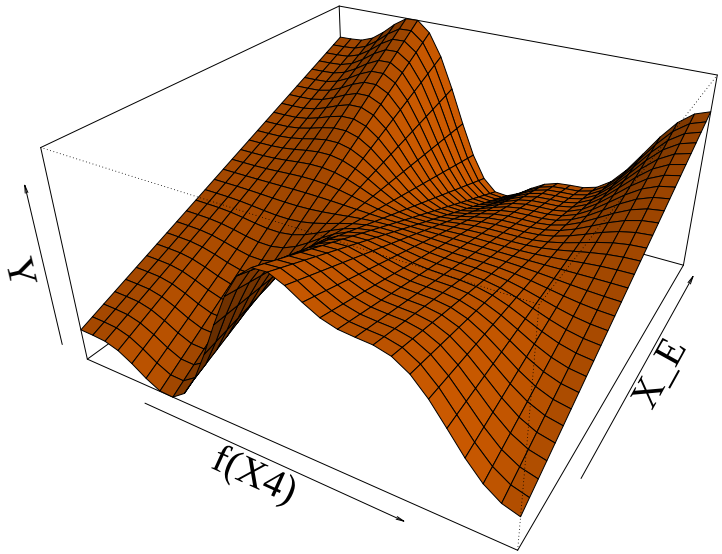


Estimated: 75th Percentile

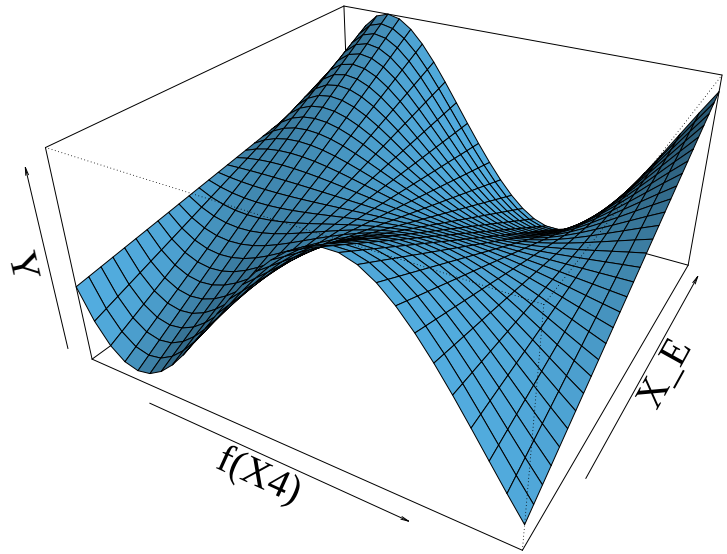


Scenario 1: Estimated Interaction Effects for $E \cdot f(X_4)$

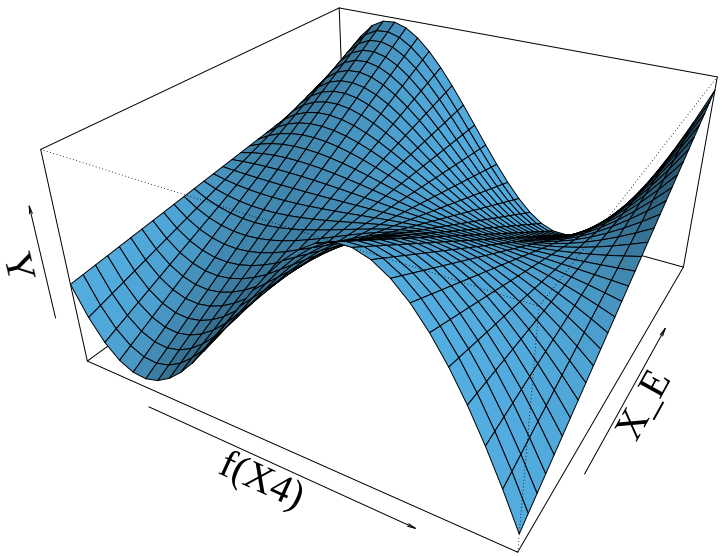
Truth



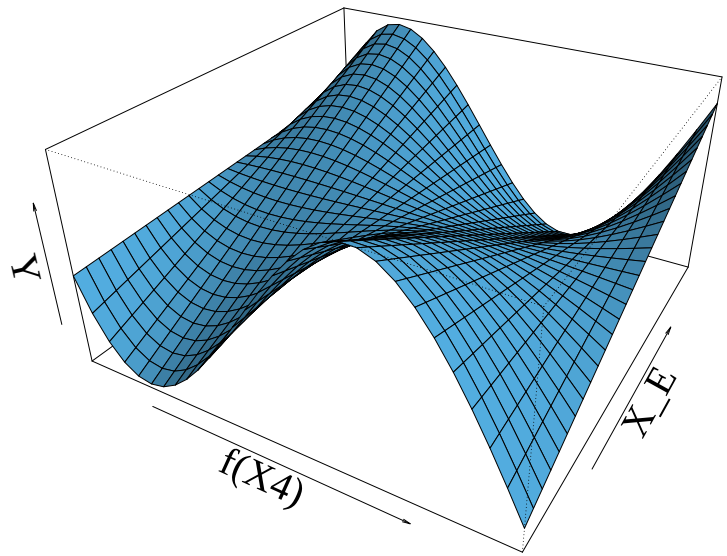
Estimated: 25th Percentile



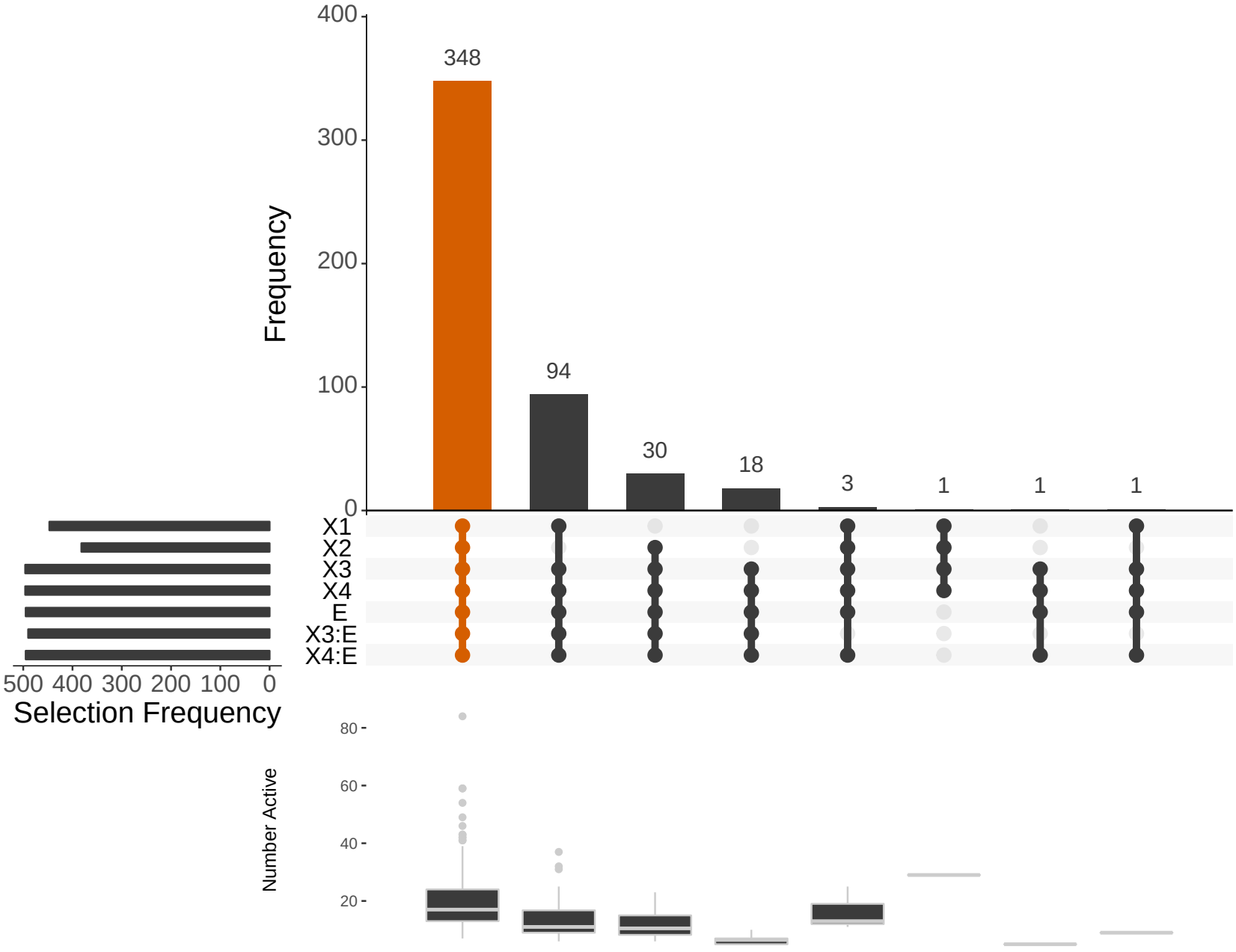
Estimated: 50th Percentile



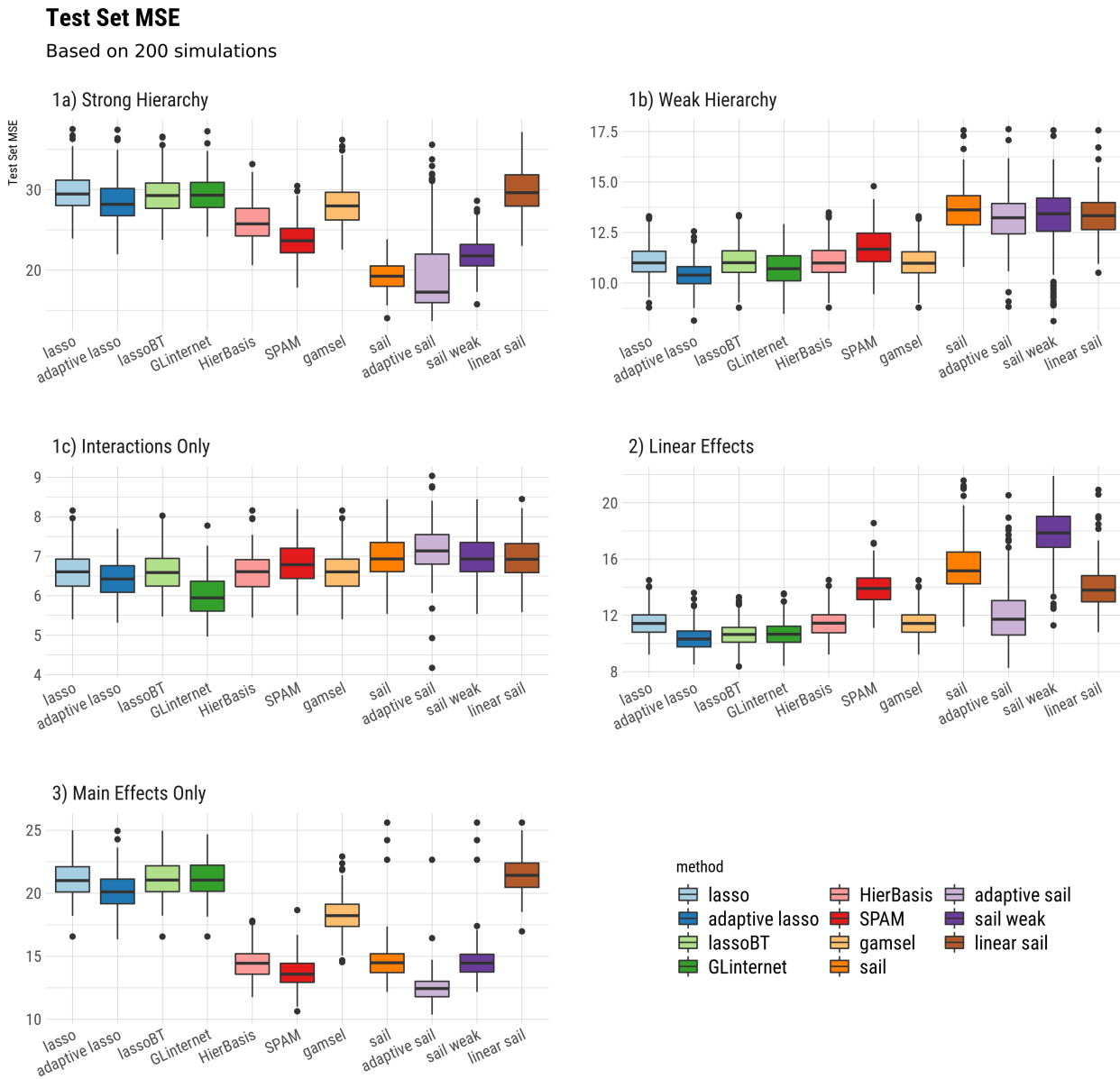
Estimated: 75th Percentile



Right in Our Wheel House Simulation Results



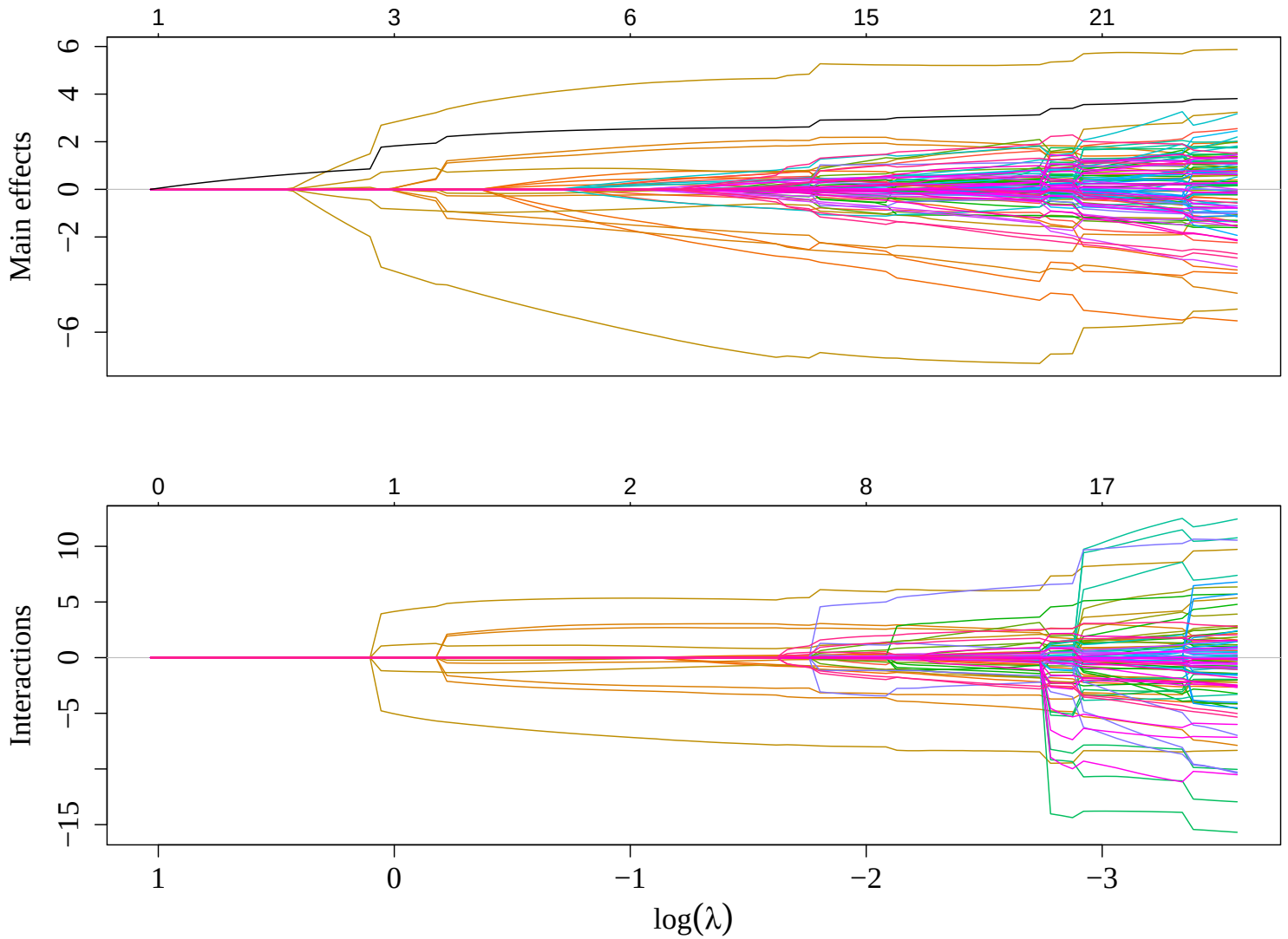
All Scenarios MSE



`sail` R package

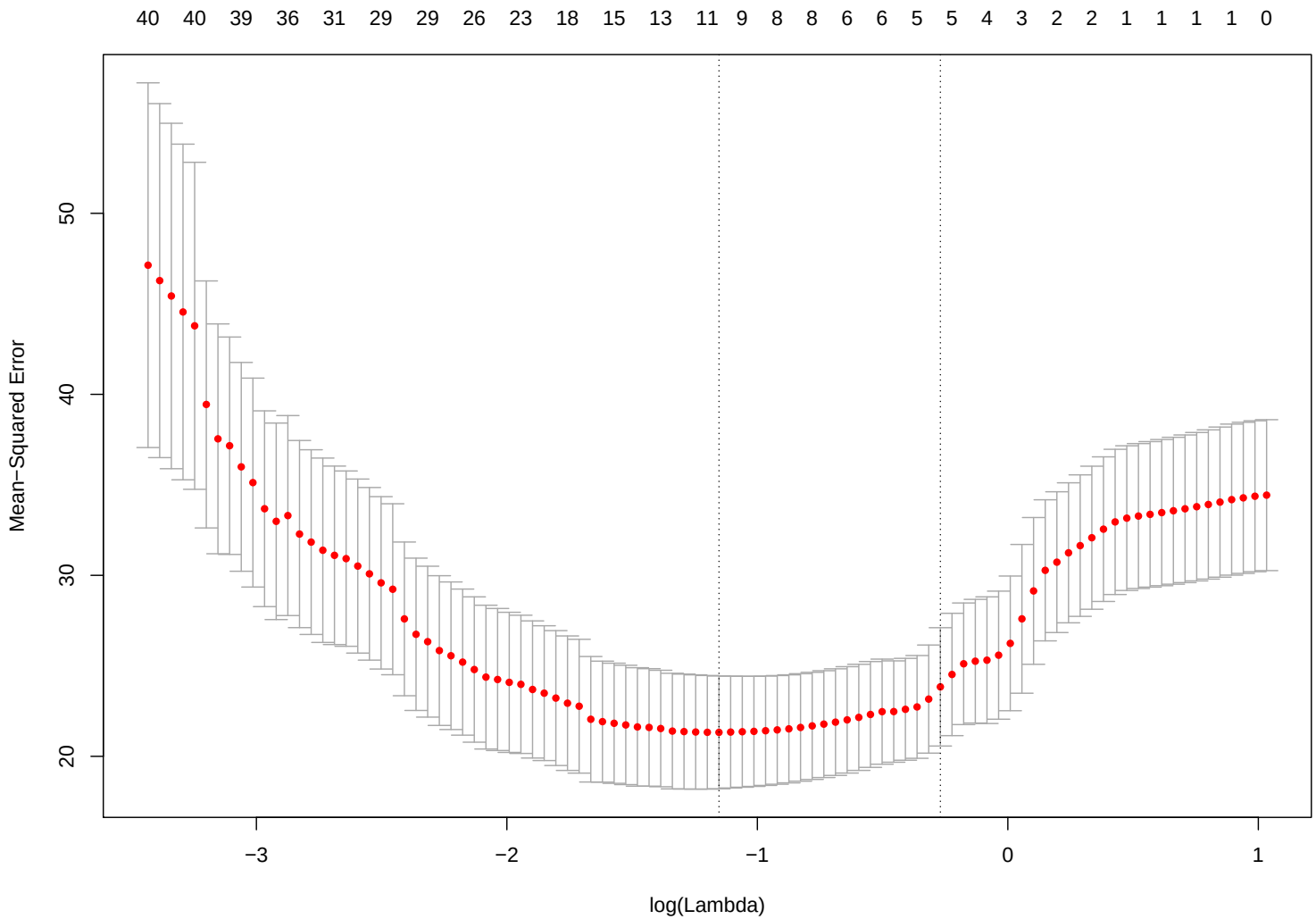
`sail` R package: Solution Path results

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



`sail` R package: Cross-validation results

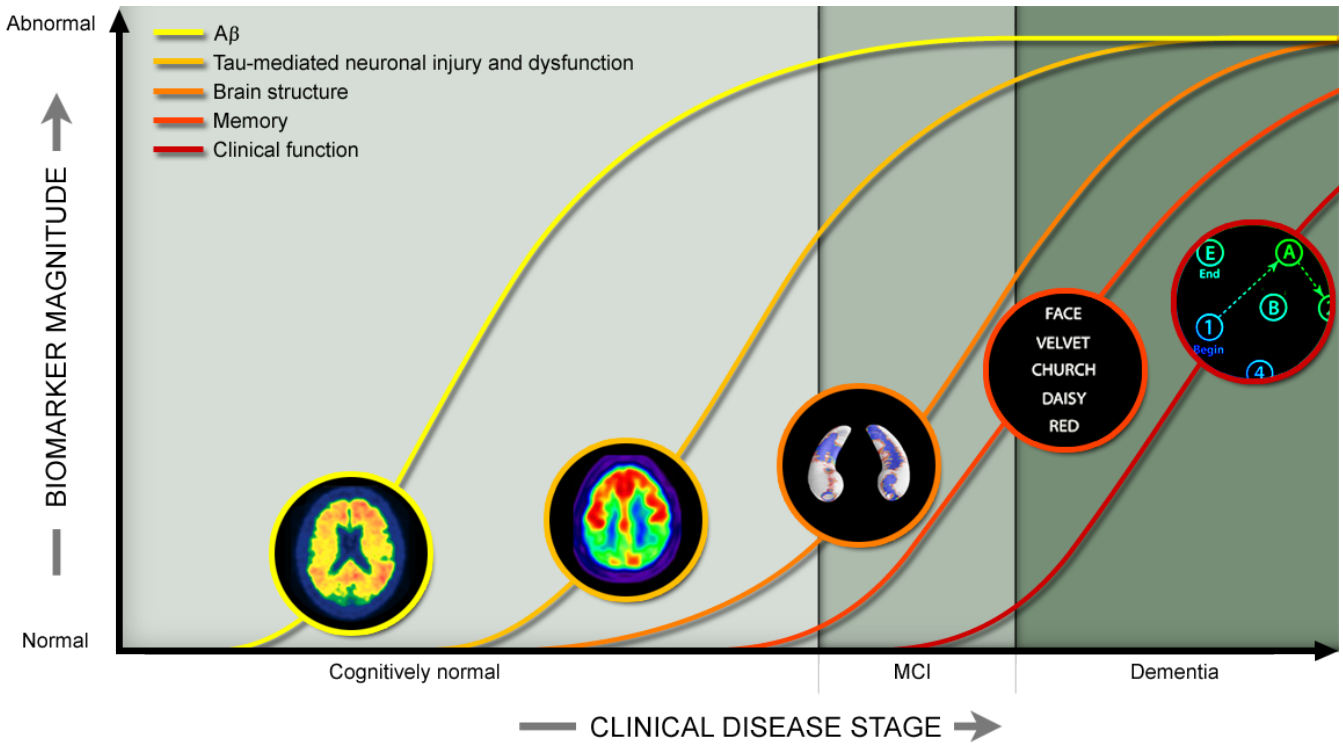
```
sail::plot(cvfit)
```



Real Data Application

Alzheimer's Disease Neuroimaging Initiative (ADNI)

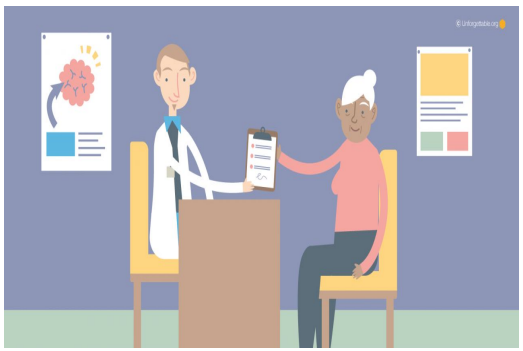
- Alzheimer's is an irreversible neurodegenerative disease that results in a loss of mental function due to the **deterioration of brain tissue**.
- The overall goal of ADNI is to **validate biomarkers** for use in Alzheimer's disease clinical treatment trials



Interaction between Aβ Protein and APOE gene

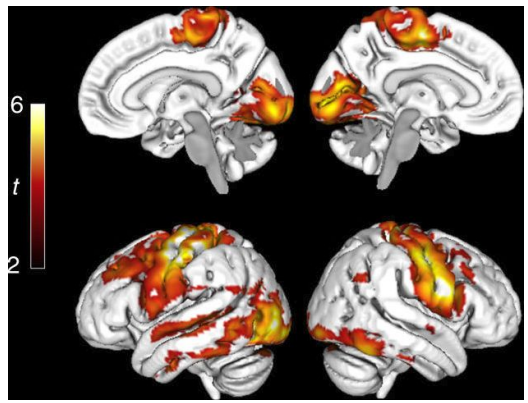
- $\mathbf{E}$: APOE4 allele increases the risk for Alzheimer's and lowers the age of onset
- $\mathbf{X}$: PET amyloid imaging to assess Aβ protein load in 96 brain regions
- $\mathbf{Y}$: General cognitive decline measured by mini-mental state examination
- $3 \times 96 \times 2 + 1 = 577$ parameters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

Mini–Mental State Examination



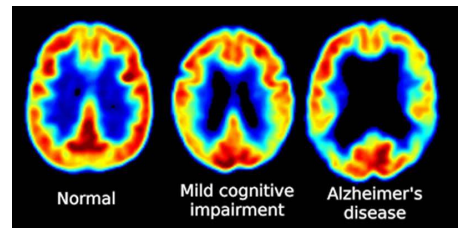
$\sim$

Amyloid Beta Protein



$\times$

Cognitive Status



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

MSE vs. number of active variables

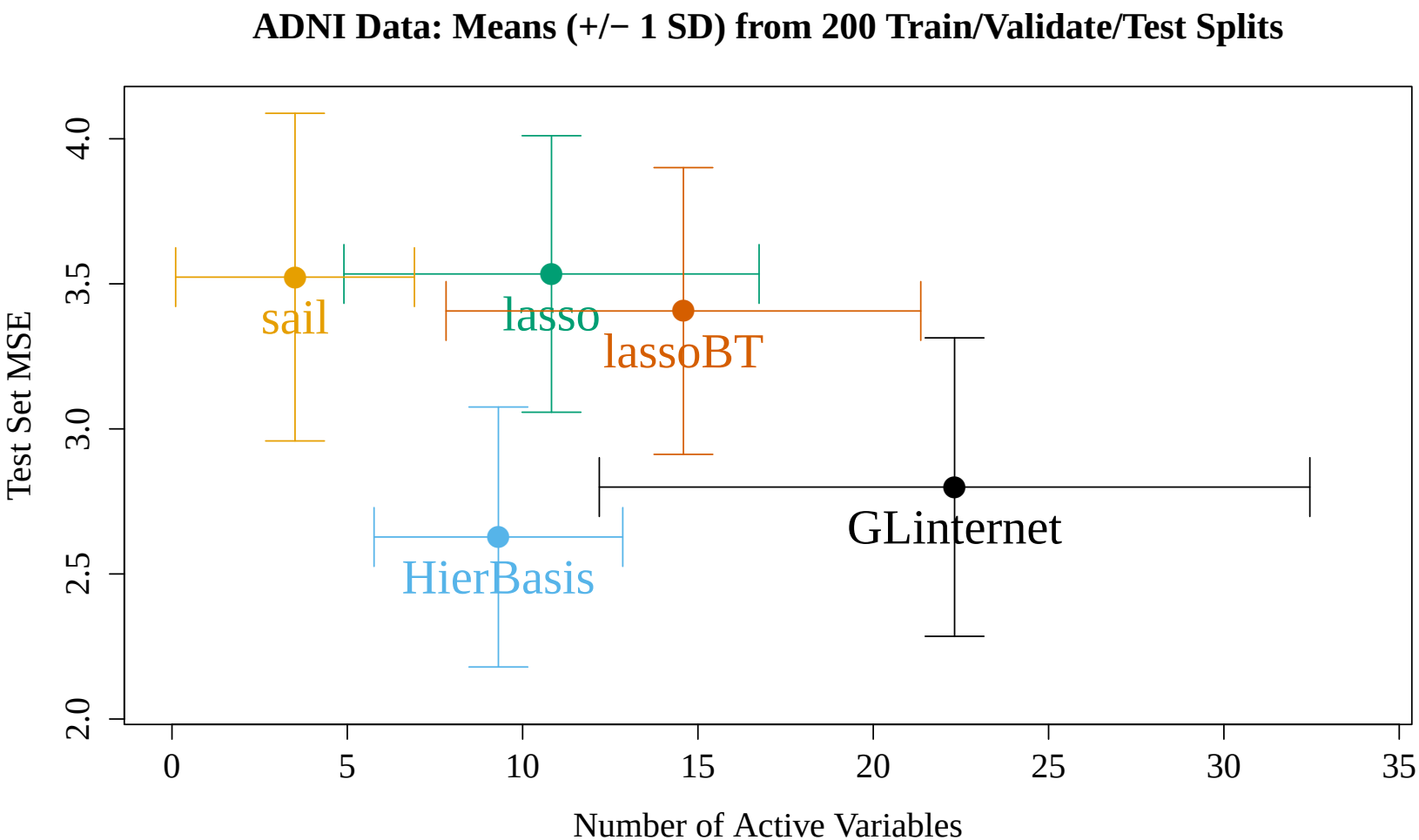
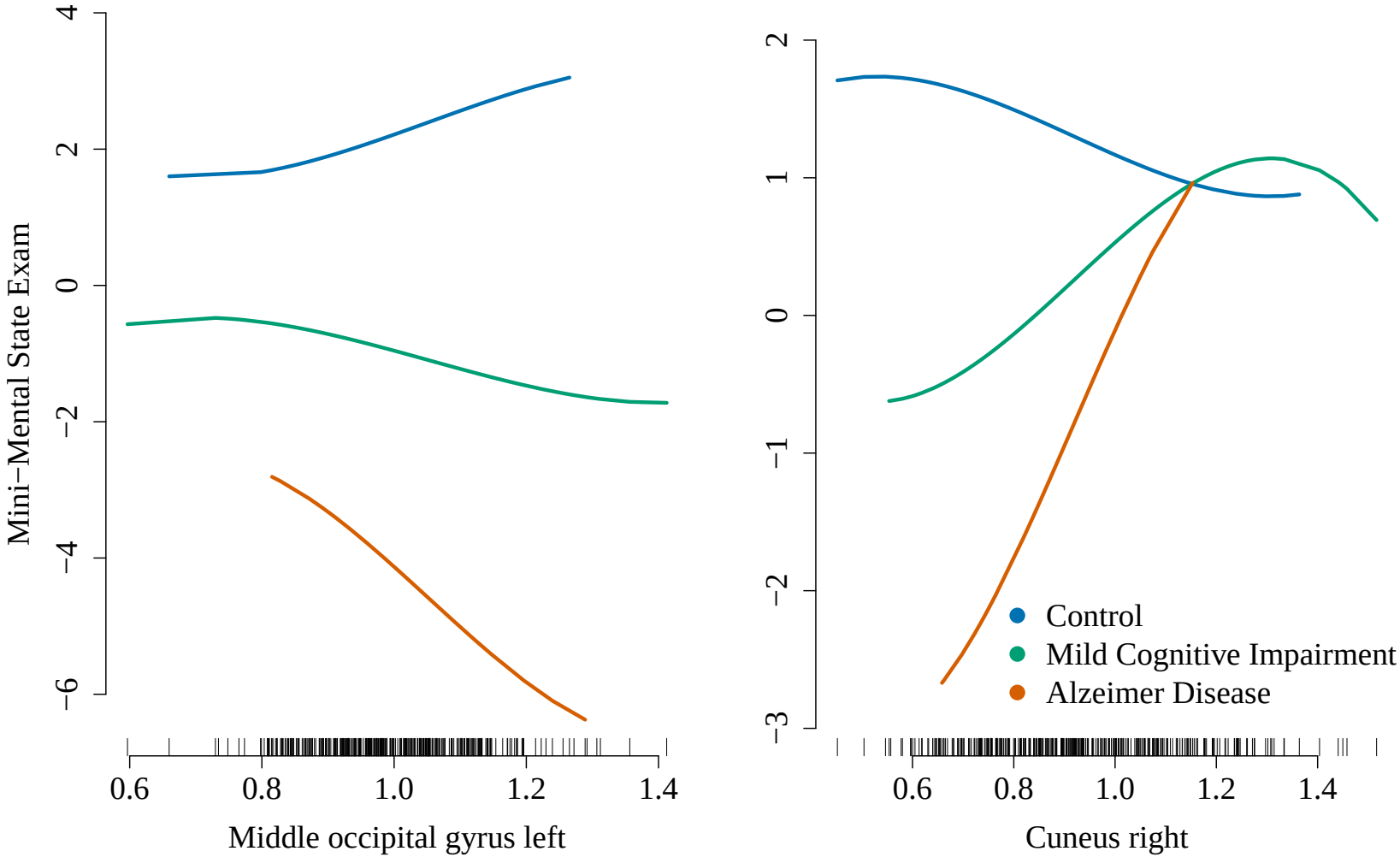


Fig.: Mean test set MSE vs. mean number of active variables (± 1 SD) for ADNI data based on 200 train/validation/test splits.

sail: Interactions



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

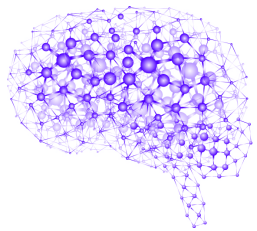
- `sail` can currently only handle $E \cdot f(X)$ or $f(E) \cdot X$
- Does not allow for $f(X_1, E)$ or $f(X_1, X_2)$
- Memory footprint is an issue

Future Directions

- Implement ADMM algorithm for scalability. Distributed computing (GPU)
- Binary Outcomes
- bi-level selection:

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

Acknowledgements



CENTRE 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.

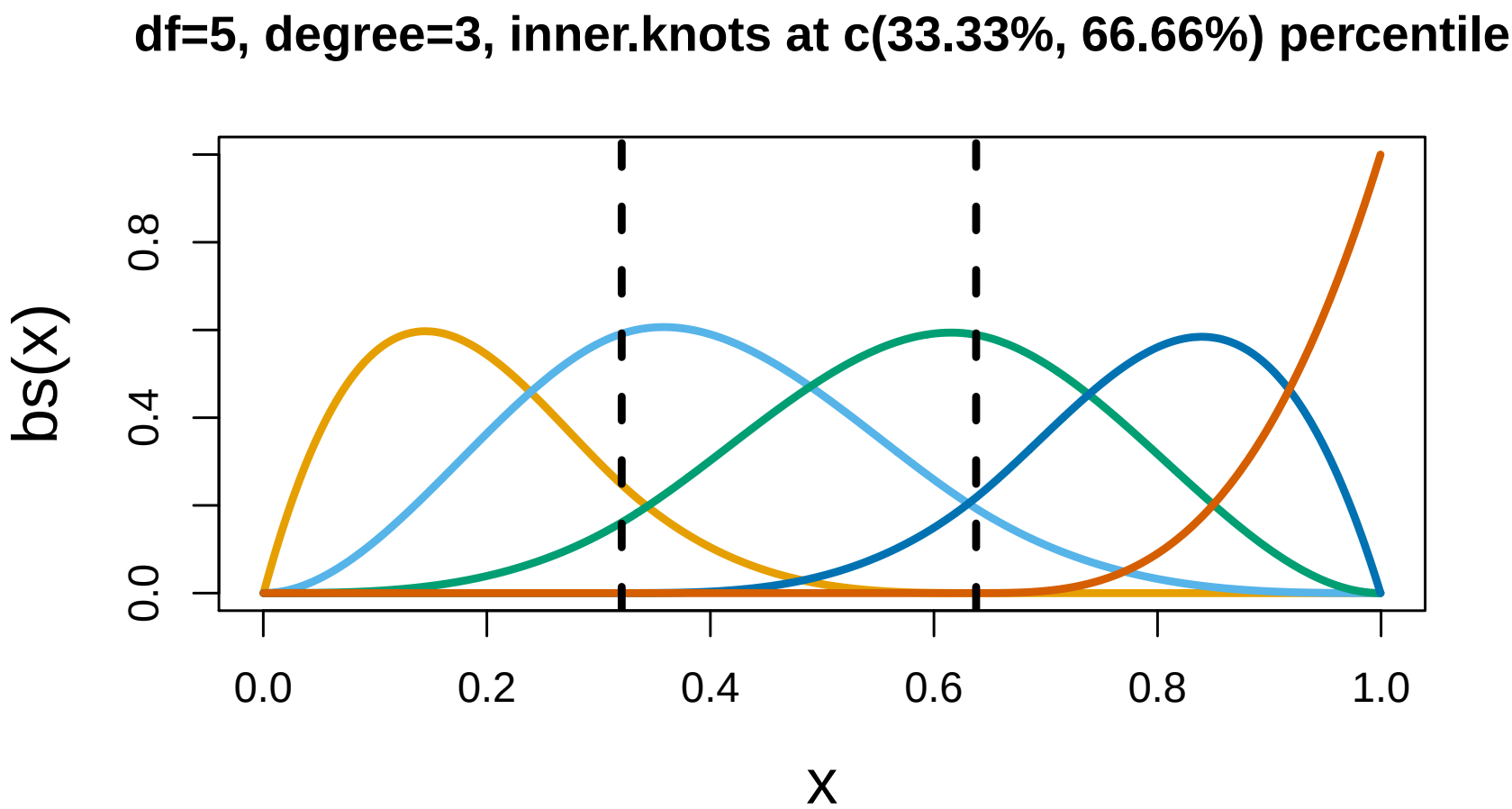
sahirbhatnagar.com

Session Info

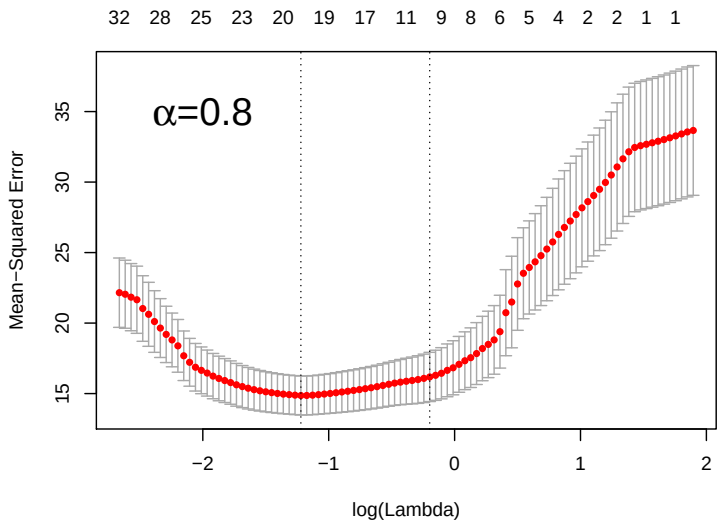
```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 18.10

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.8.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.8.0

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

other attached packages:
[1] xtable_1.8-4      rpart.plot_3.0.6  rpart_4.1-15      data.table_1
[5] ISLR_1.2          ggplot2_3.1.0     knitr_1.23

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.1          magrittr_1.5        tidyselect_0.2.5
 [4] munsell_0.5.0       colorspace_1.4-0    R6_2.4.0
 [7] rlang_0.3.4         highr_0.8           stringr_1.4.0
[10] plyr_1.8.4          dplyr_0.8.0.1       tools_3.6.0
[13] grid_3.6.0          gtable_0.2.0        xfun_0.7
[16] pacman_0.5.0        withr_2.1.2         digest_0.6.19
[19] lazyeval_0.2.1      assertthat_0.2.1    tibble_2.1.1
[22] crayon_1.3.4        RSkittleBrewer_1.1  purrr_0.3.2
[25] glue_1.3.1          evaluate_0.14       stringi_1.4.3
[28] compiler_3.6.0      pillar_1.4.0        scales_1.0.0
```

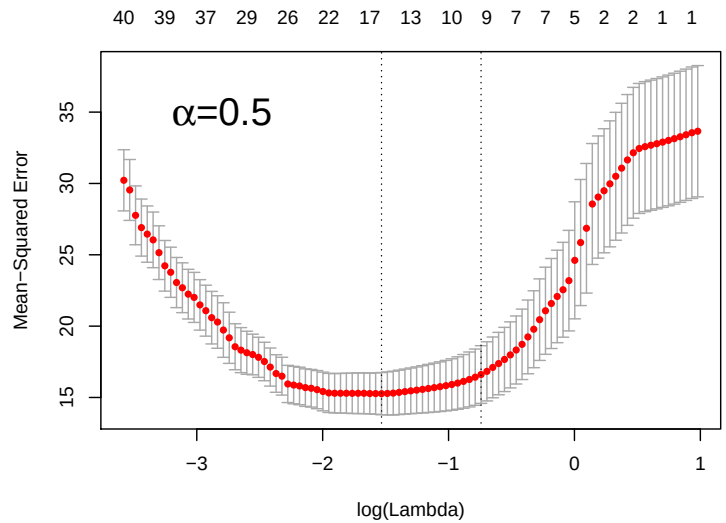
B-Spline Expansion

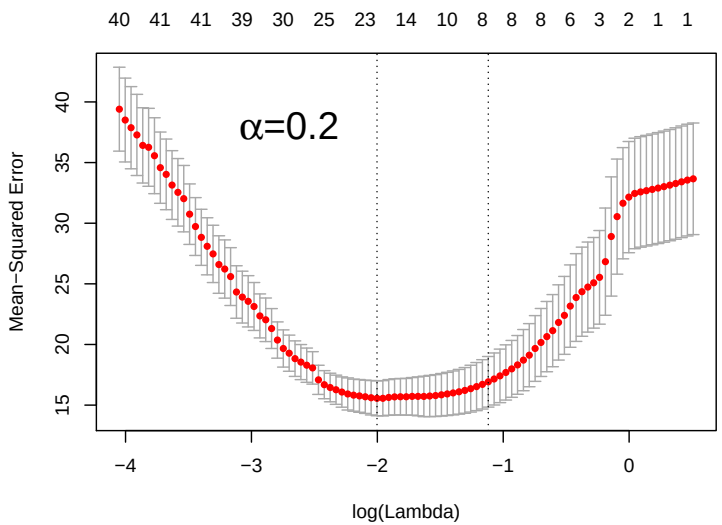
```r
x <- truncnorm::rtruncnorm(1000, a = 0, b = 1)
B <- splines::bs(x, df = 5, degree=3, intercept = FALSE)
```

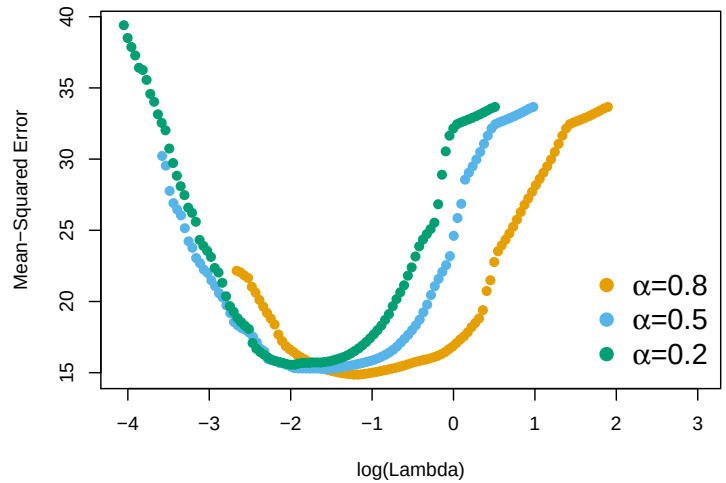


`sail` A Note on the Second Tuning Parameter results









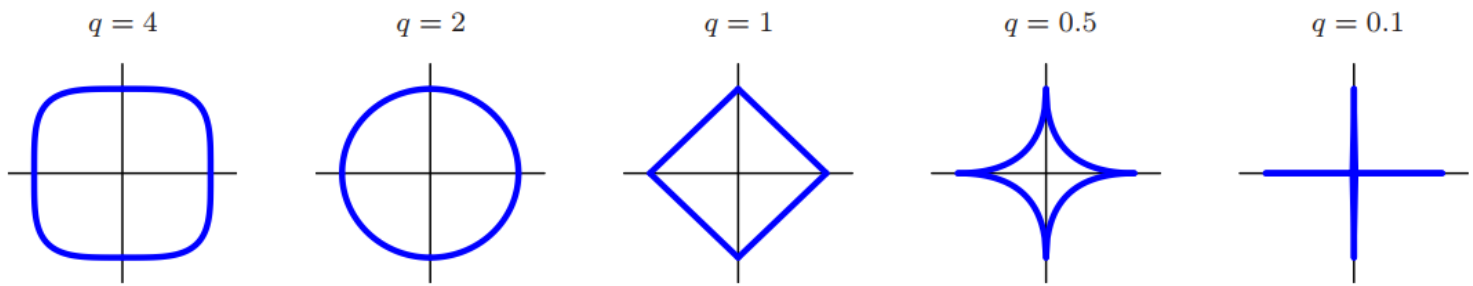
Appendix

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