A Sparse Additive Model for High-Dimensional Interactions with an Exposure Variable

Sahir R Bhatnagar[1,2], Tianyuan Lu[3,4], Amanda Lovato[5], David L Olds[6], Michael S Kobor[7], Michael J Meaney[8], Kieran O'Donnell[9], Yi Yang[10], and Celia MT Greenwood[1,3,5]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University

[2]Department of Diagnostic Radiology, McGill University

[3]Quantitative Life Sciences, McGill University

[4]Lady Davis Institute, Jewish General Hospital, Montréal, QC

[5]Statistics Canada, Ottawa, ON

[6]Department of Pediatrics, University of Colorado School of Medicine, Denver

[7]Department of Medical Genetics, University of British Columbia, BC

[8]Singapore Institute for Clinical Sciences, Singapore; McGill University

[9]Department of Psychiatry, McGill University

[10]Department of Mathematics and Statistics, McGill University

[11]Departments of Oncology and Human Genetics, McGill University

September 5, 2021

```
## /scratch/bhatnagar-lab/sbhatnagar/git_repositories/sail/manuscript/bin
```

```
## +-- ADNI.R
```

```
## +-- PRS_bootstrap.R
```

```
## +-- PRS_eval_functions.R
```

```
## +-- PRS_method_functions.R
```

```
## +-- PRS_model_functions.R
```

```
## +-- PRS_plot_functions.R
```

```
## +-- PRS_plots.R
```

```
## +-- README_for_CSDA_Figures.Rnw
```

```
## +-- README_for_CSDA_Figures.pdf
## +-- README_for_CSDA_Figures.tex
## +-- cache
## |   +-- __packages
## |   +-- globals_6d1d04419b10b1a06759621e8a2eb8a6.RData
## |   +-- globals_6d1d04419b10b1a06759621e8a2eb8a6.rdb
## |   +-- globals_6d1d04419b10b1a06759621e8a2eb8a6.rdx
## |   +-- packages_9258a6f76ff684c0e2fd1b9c0ebacc65.RData
## |   +-- packages_9258a6f76ff684c0e2fd1b9c0ebacc65.rdb
## |   +-- packages_9258a6f76ff684c0e2fd1b9c0ebacc65.rdx
## |   +-- plot-mse-sim_9c0c9a1ddefc6e0131421d7b889a78cc.RData
## |   +-- plot-mse-sim_9c0c9a1ddefc6e0131421d7b889a78cc.rdb
## |   +-- plot-mse-sim_9c0c9a1ddefc6e0131421d7b889a78cc.rdx
## |   +-- setup2_900cd6b26f4a65ef22443e3819f71205.RData
## |   +-- setup2_900cd6b26f4a65ef22443e3819f71205.rdb
## |   +-- setup2_900cd6b26f4a65ef22443e3819f71205.rdx
## |   +-- simulation-results_05ea566347202b9db8e06fd8152aac10.RData
## |   +-- simulation-results_05ea566347202b9db8e06fd8152aac10.rdb
## |   +-- simulation-results_05ea566347202b9db8e06fd8152aac10.rdx
## |   +-- toy-effects_601ac4af2c99af20848947802c3de27d.RData
## |   +-- toy-effects_601ac4af2c99af20848947802c3de27d.rdb
## |   +-- toy-effects_601ac4af2c99af20848947802c3de27d.rdx
## |   +-- toy-example_a90f8c38d0355dd05704afeb06d0c2c6.RData
## |   +-- toy-example_a90f8c38d0355dd05704afeb06d0c2c6.rdb
## |   +-- toy-example_a90f8c38d0355dd05704afeb06d0c2c6.rdx
## |   +-- toy-solution-path_991c1813a5d6d78869c629849579836d.RData
## |   +-- toy-solution-path_991c1813a5d6d78869c629849579836d.rdb
## |   +-- toy-solution-path_991c1813a5d6d78869c629849579836d.rdx
## |   +-- unnamed-chunk-1_9950a44ce550dc548513d2b840367e35.RData
## |   +-- unnamed-chunk-1_9950a44ce550dc548513d2b840367e35.rdb
## |   \-- unnamed-chunk-1_9950a44ce550dc548513d2b840367e35.rdx
## +-- figure
## |   +-- plot-mse-sim-1.pdf
## |   +-- toy-effects-1.pdf
## |   \-- toy-solution-path-1.pdf
## +-- intro.R
## +-- setup.R
## +-- simulation.R
## +-- support_bootstrap.R
## \-- support_plots.R
```

1 Figure 1 - Toy example solution path and effects

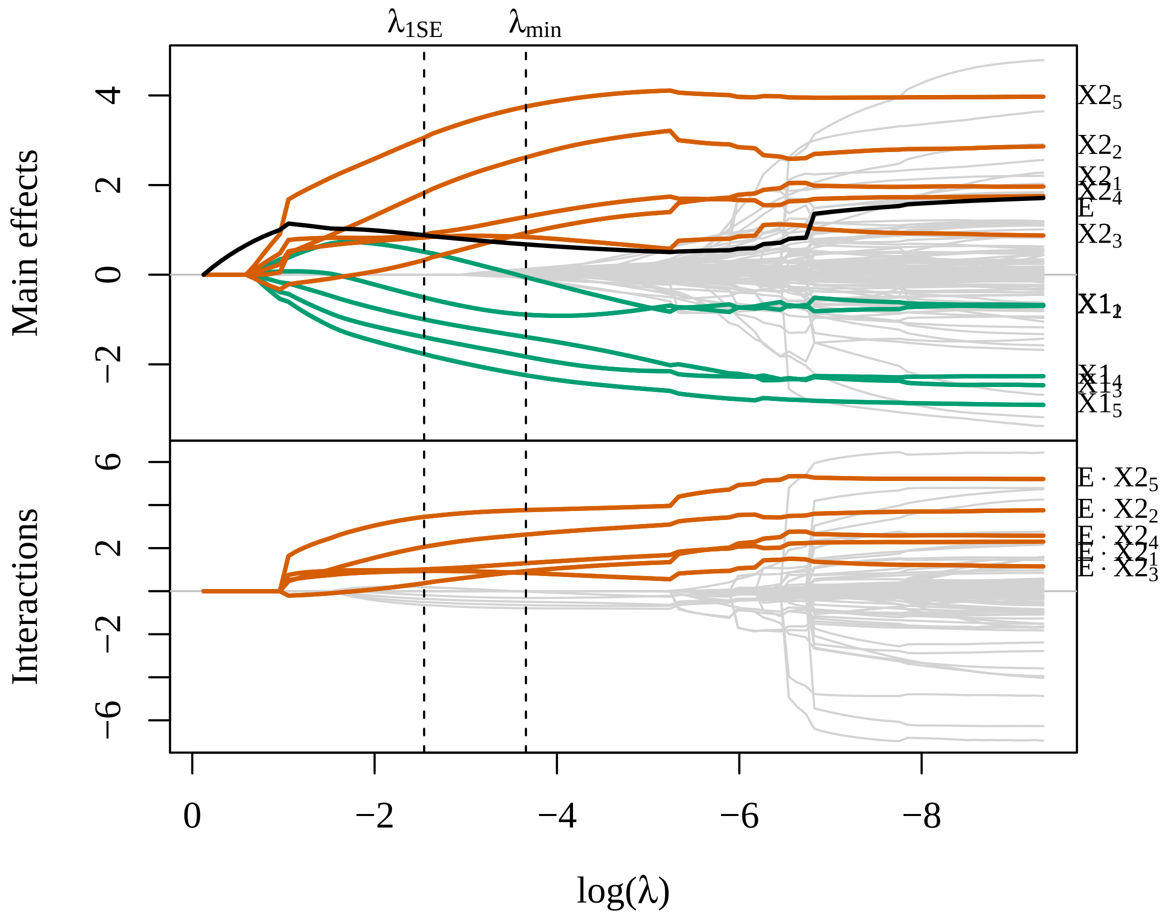

Figure 1: Toy example solution path for main effects (top) and interactions (bottom). $\{X1_1, X1_2, X1_3\}$ and $\{X2_1, X2_2, X2_3\}$ are the three basis coefficients for X_1 and X_2, respectively. λ_{1SE} is the largest value of penalization for which the CV error is within one standard error of the minimizing value λ_{min}.

In Figure 2, we plot the true and estimated component functions $\hat{f}_1(X_1)$ and $E \cdot \hat{f}_2(X_2)$, and their estimates from this analysis with `sail`. We are able to capture the shape of the correct functional form, but the means are not well aligned with the data. Lack-of-fit for $f_1(X_1)$ can be partially explained by acknowledging that `sail` is trying to fit a cubic spline to a linear function. Nevertheless, this example demonstrates that `sail` can still identify trends reasonably well.

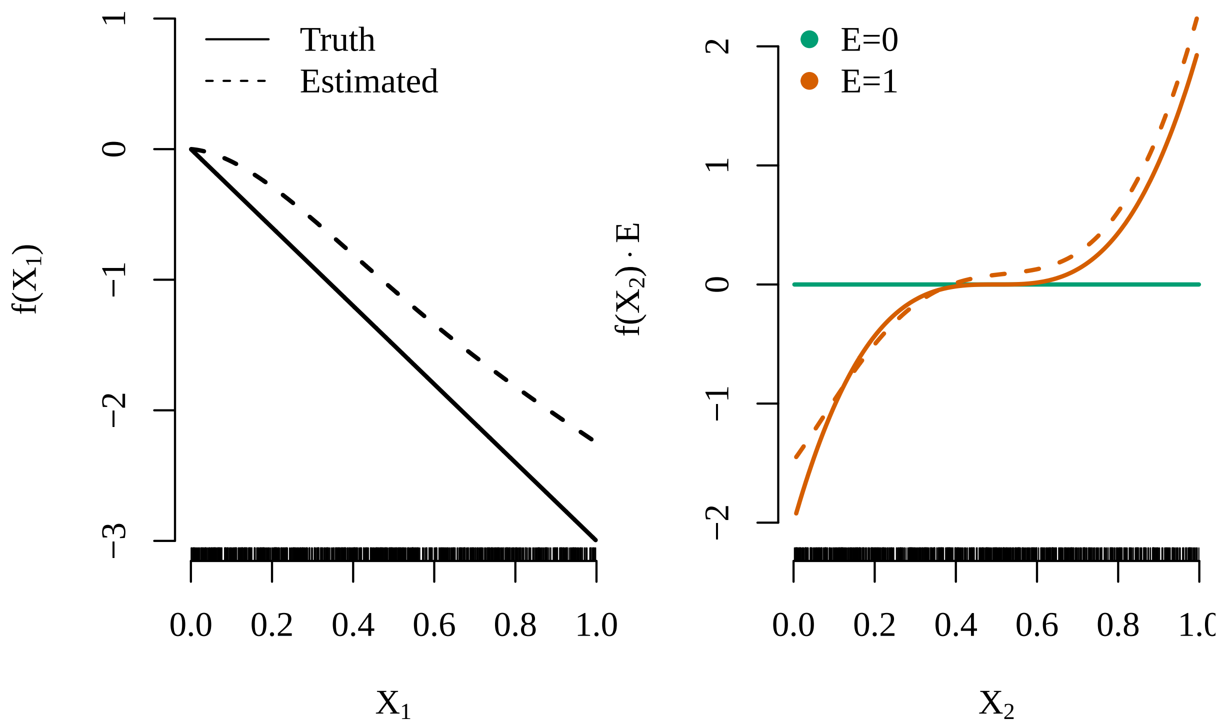

Figure 2: Estimated smooth functions for X_1 and the $X_2 \cdot E$ interaction by the `sail` method based on λ_{min}.

2 Figure 2 - Test set MSE

```
## Error in `:=`(scenario, as.numeric(as.character(stringr::str_extract_all(parameterIndex, :  Check that is.data.table(DT)
== TRUE. Otherwise, := and `:=`(...)  are defined for use in j, once only and in particular ways.  See help(":=").

## Error in `:=`(scen, ifelse(scenario == 1, "Strong Hierarchy", ifelse(scenario == :  Check that is.data.table(DT)
== TRUE. Otherwise, := and `:=`(...)  are defined for use in j, once only and in particular ways.  See help(":=").

## Error in `:=`(scen, factor(scen, levels = c("Strong Hierarchy", "Weak Hierarchy", :  Check that is.data.table(DT)
== TRUE. Otherwise, := and `:=`(...)  are defined for use in j, once only and in particular ways.  See help(":=").
```

3 Table 1

Table 1: Mean (standard deviation) of the number of selected variables ($|\widehat{\mathcal{J}}|$), true positive rate (TPR) and false positive rate (FPR) as a percentage from 200 simulations for each of the five scenarios. $|\mathcal{J}|$ is the number of truly associated variables.

	Linear Main Effects		Linear Interactions		Non-linear Main Effects			Non-linear Interactions		
	lasso	adaptive lasso	lassoBT	GLinternet	HierBasis	SPAM	gamsel	sail	adaptive sail	sail weak
1a) Strong heredity ($|\mathcal{J}| = 7$)										
$|\widehat{\mathcal{J}}|$	28 (15)	8 (4)	35 (18)	40 (20)	133 (48)	42 (19)	46 (21)	21 (3)	8 (3)	21 (3)
TPR	53.9 (8.4)	49.3 (10.1)	61.7 (11.5)	66.4 (14.0)	65.2 (8.1)	60.9 (8.5)	56.9 (7.7)	86.8 (8.0)	81.4 (13.0)	82.1 (10.9)
FPR	1.2 (0.7)	0.2 (0.2)	1.5 (0.9)	1.8 (1.0)	6.5 (2.4)	1.9 (0.9)	2.1 (1.1)	0.8 (0.1)	0.1 (0.1)	0.8 (0.1)
1b) Weak heredity ($|\mathcal{J}| = 5$)										
$|\widehat{\mathcal{J}}|$	19 (12)	4 (2)	20 (13)	38 (23)	24 (23)	28 (16)	21 (15)	16 (7)	5 (3)	14 (10)
TPR	40.7 (3.6)	40.1 (1.4)	40.8 (3.8)	64.1 (14.9)	42.2 (6.3)	53.9 (9.4)	42.7 (6.8)	50.5 (10.4)	46.4 (10.1)	55.0 (13.7)
FPR	0.9 (0.6)	0.1 (0.1)	0.9 (0.7)	1.7 (1.1)	1.1 (1.1)	1.2 (0.8)	1.0 (0.7)	0.7 (0.3)	0.2 (0.1)	0.6 (0.5)
1c) Interactions Only ($|\mathcal{J}| = 2$)										
$|\widehat{\mathcal{J}}|$	12 (12)	3 (2)	14 (13)	38 (21)	12 (13)	13 (12)	12 (12)	7 (7)	2 (2)	26 (30)
TPR	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	81.4 (27.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	22.9 (36.9)
FPR	0.6 (0.6)	0.6 (6.9)	0.7 (0.7)	1.8 (1.0)	0.6 (0.7)	0.7 (0.6)	0.6 (0.6)	0.4 (0.3)	0.1 (0.1)	1.3 (1.5)
2) Linear Effects ($|\mathcal{J}| = 7$)										
$|\widehat{\mathcal{J}}|$	37 (17)	8 (3)	48 (19)	51 (23)	37 (19)	42 (19)	37 (16)	20 (4)	11 (4)	20 (4)
TPR	70.4 (3.7)	67.2 (6.7)	72.3 (6.3)	93.4 (8.5)	70.3 (3.8)	65.0 (8.1)	70.4 (3.7)	91.8 (10.5)	86.0 (18.5)	68.1 (14.9)
FPR	1.6 (0.8)	0.2 (0.2)	2.2 (1.0)	2.2 (1.2)	1.6 (0.9)	1.9 (0.9)	1.6 (0.8)	0.7 (0.2)	0.2 (0.2)	0.7 (0.2)
3) Main Effects Only ($|\mathcal{J}| = 5$)										
$|\widehat{\mathcal{J}}|$	29 (14)	7 (4)	31 (15)	34 (18)	154 (17)	46 (21)	56 (20)	22 (2)	9 (2)	22 (2)
TPR	75.9 (10.9)	66.5 (15.3)	76.0 (10.9)	77.0 (9.5)	97.5 (6.6)	93.1 (10.7)	81.3 (9.5)	88.3 (10.3)	84.1 (9.2)	85.2 (12.1)
FPR	1.3 (0.7)	0.2 (0.2)	1.3 (0.8)	1.5 (0.9)	7.5 (0.9)	2.1 (1.0)	2.6 (1.0)	0.9 (0.1)	0.2 (0.1)	0.9 (0.1)

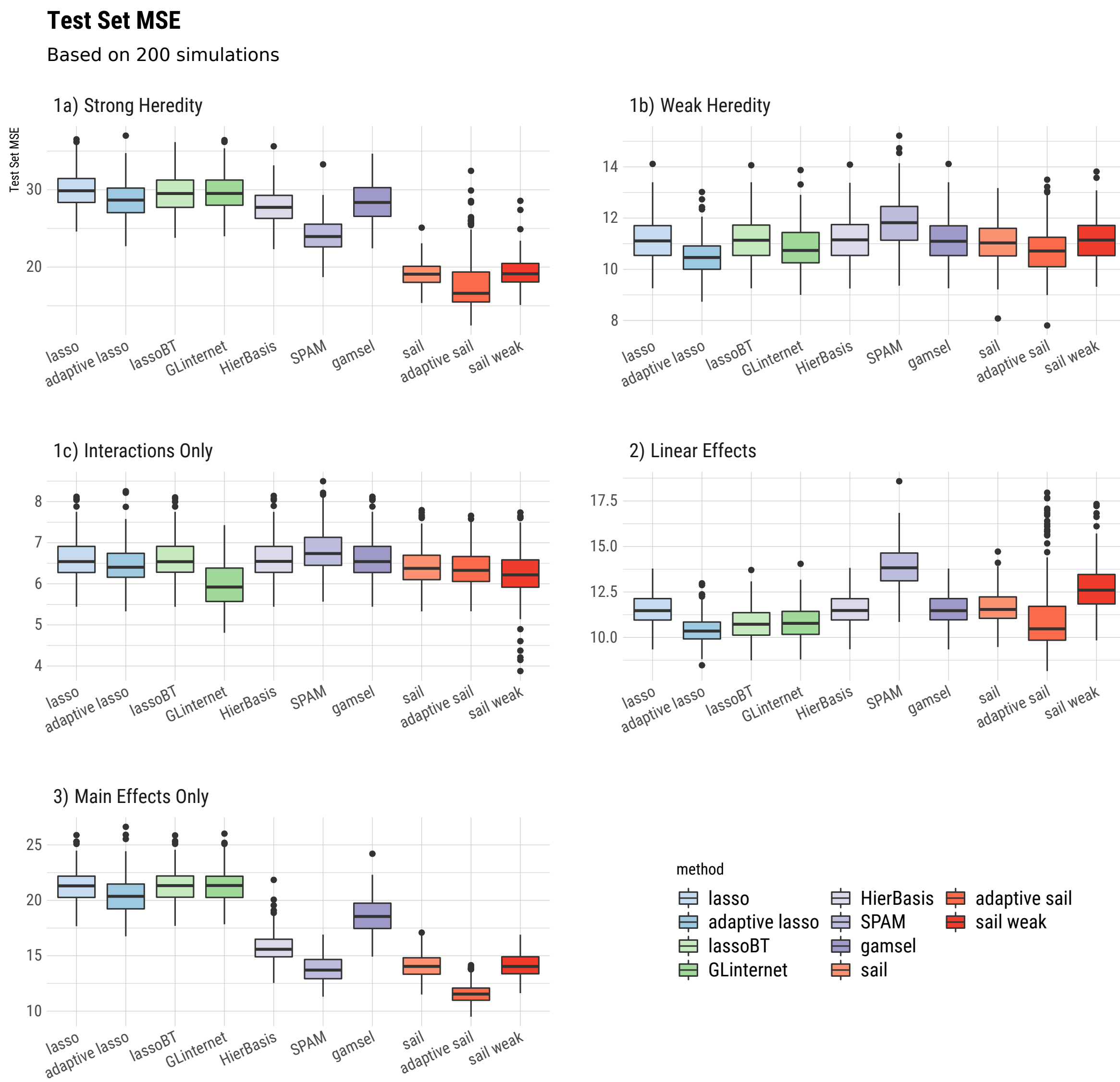

Figure 3: Boxplots of the test set mean squared error from 200 simulations for each of the five simulation scenarios.