IOWA

Best Subset Selection

Devin Spolsdoff Jacob Seedorff

Methods

Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms

Hussein Hazimeh,[a] Rahul Mazumder[a,b]

[a] Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [b] MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso

Following "Best Subset Selection from a Modern Optimization Lens" by Bertsimas, King, and Mazumder (2016)

Trevor Hastie   Robert Tibshirani   Ryan J. Tibshirani

Outline

- Introduction to best subset selection
- Recent developments in best subset selection
  - Best subset selection
  - Classes of Minima
  - Algorithms
- Extended comparisons of best subset selection
  - Relaxed lasso
  - Simulation
  - Discussion

Introduction to Best Subset Selection

Best Subset Selection

- $\min_{\beta} \frac{1}{2} ||y - X\beta||_2^2 + \lambda||\beta||_0$
- Where $||\beta||_0 = \sum_j 1_{(\beta_j \neq 0)}$ is the $l0$-pseudo norm
- This function is discontinuous at 0 which causes difficulty with optimization
  – The traditional method of solving this problem is to fit a model for each combination of covariates and then select the model that resulted in the smallest value of the objective function
- However, there have been many recent advances that have allowed for much faster computation for this problem

IOWA

Best Subset Selection Cont.

- The best subset selection estimator can instead be found by solving the following problem
- $\min_{\beta} \frac{1}{2} ||y - X\beta||_2^2 \; s.t. ||\beta||_0 \leq k$
- For $k = 0, 1, \ldots, p$ and the best subset selection estimator for a given λ will be one of these $p + 1$ models

IOWA

Recent developments in best subset selection

Best Subset Selection

Methods

Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms

Hussein Hazimeh,[a] **Rahul Mazumder**[a,b]

[a] Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [b] MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

Notation

- $F(\beta) = \frac{1}{2}||y - X\beta||_2^2 + \lambda_0||\beta||_0 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$
- $\bar{\beta_i} = \left(y - \sum_{j\neq i} X_j\beta_j\right)^T X_i = y^T X_i - \sum_{j\neq i} X_j^T X_i \beta_j$
- The support (S) is the set of nonzero coefficients
- U^S denotes a $p \times p$ matrix with the following properties
  - $(U^S\beta)_i = \beta_i$ if $i \in S$
  - $(U^S\beta)_i = 0$ if $i \notin S$.
- The set $\{1, 2, \ldots, p\}$ is denoted by $[p]$

IOWA

Recent developments in best subset selection

Classes of Minima

CW Minima

- A vector β^* is a coordinate-wise (CW) minima if for every $i \in [p]$, β_i^* is a minimizer of $F(\beta^*)$ with respect to the i^{th} coordinate with all the other coordinates held fixed
- The single coordinate solution is given by the following thresholding operator

$$\bar{T}\left(\bar{\beta}_i^*, \lambda_0, \lambda_1, \lambda_2\right) = \begin{cases} \left\{sign\left(\bar{\beta}_i^*\right)\dfrac{\left|\bar{\beta}_i^*\right| - \lambda_1}{1 + 2\,\lambda_2}\right\} & if\ \dfrac{\left|\bar{\beta}_i^*\right| - \lambda_1}{1 + 2\,\lambda_2} > \sqrt{\dfrac{2\lambda_0}{1 + 2\lambda_2}} \\ \{0\} & if\ \dfrac{\left|\bar{\beta}_i^*\right| - \lambda_1}{1 + 2\,\lambda_2} < \sqrt{\dfrac{2\lambda_0}{1 + 2\lambda_2}} \\ \left\{0, sign\left(\bar{\beta}_i^*\right)\dfrac{\left|\bar{\beta}_i^*\right| - \lambda_1}{1 + 2\,\lambda_2}\right\} & if\ \dfrac{\left|\bar{\beta}_i^*\right| - \lambda_1}{1 + 2\,\lambda_2} = \sqrt{\dfrac{2\lambda_0}{1 + 2\lambda_2}} \end{cases}$$

IOWA

PSI(k) and FSI(k) Minima

- A vector β^* with support S is a partial swap-inescapable minima of order k (PSI(k) minima) if for every $S_1 \subseteq S$, $S_2 \subseteq S^C$, with $|S_1| \leq k$, $|S_2| \leq k$, the following holds

$$F(\beta^*) \leq \min_{\beta_{S_2}} F(\beta^* - U^{S_1}\beta^* + U^{S_2}\beta)$$

- A vector β^* with support S is a full swap-inescapable minima of order k (FSI(k) minima) if for every $S_1 \subseteq S$, $S_2 \subseteq S^C$, with $|S_1| \leq k$, $|S_2| \leq k$, the following holds

$$F(\beta^*) \leq \min_{\beta_{(S/S_1)\cup S_2}} F(\beta^* - U^{S_1}\beta^* + U^{(S/S_1)\cup S_2}\beta)$$

Ordering of Minima

- $FSI(k) \subset PSI(k) \subset CW$
- This means that $FSI(k)$ minima are the strongest, followed by $PSI(k)$ minima and CW minima are the weakest
- When k is sufficiently large $FSI(k)$ and $PSI(k)$ minima coincide with the class of global minimizers, however as we increase k we are also increasing the difficulty of the problem and thus it will take longer to find the solution

IOWA

Recent developments in best subset selection

Algorithms

CW Minima

- Since the CW minima is based on doing updates one coordinate at a time, it would make sense to use coordinate descent to perform the updates
- The authors propose the use of coordinate descent (CDSS) with the following modified thresholding operator

$$T(\bar{\beta}_i^*, \lambda_0, \lambda_1, \lambda_2) = \begin{cases} \left\{ sign(\bar{\beta}_i^*) \dfrac{|\bar{\beta}_i^*| - \lambda_1}{1 + 2\,\lambda_2} \right\} & if\ \dfrac{|\bar{\beta}_i^*| - \lambda_1}{1 + 2\,\lambda_2} \geq \sqrt{\dfrac{2\lambda_0}{1 + 2\lambda_2}} \\ \{0\} & if\ \dfrac{|\bar{\beta}_i^*| - \lambda_1}{1 + 2\,\lambda_2} < \sqrt{\dfrac{2\lambda_0}{1 + 2\lambda_2}} \end{cases}$$

Spacer Steps

- In the coordinate descent algorithm, they also introduce the use of spacer steps
- Spacer steps entail the following process for some fixed number C
  - When a support S has been encountered Cp-many times, then a spacer step is performed
  - Reoptimize over each coordinate in S with the following thresholding operator $T(\bar{\beta}_i, 0, \lambda_1, \lambda_2)$
  - Reset the counter for this support
- These are necessary for their proof of convergence to a CW minima

IOWA

PSI(k) Minima

- In order to converge to a PSI(k) minima, we again perform CDSS and after it converges, we check if there is a feasible solution to the following problem

$$\min_{\beta, S_1, S_2} F\left(\beta^l - U^{S_1}\beta^l + U^{S_2}\beta\right) \; s.t. \, S_1 \subseteq S, S_2 \subseteq S^c, |S_1| \leq k, |S_2| \leq k$$

- If there is a solution to this problem, then we update the support and β
- If there is no feasible solution, then stop and declare convergence

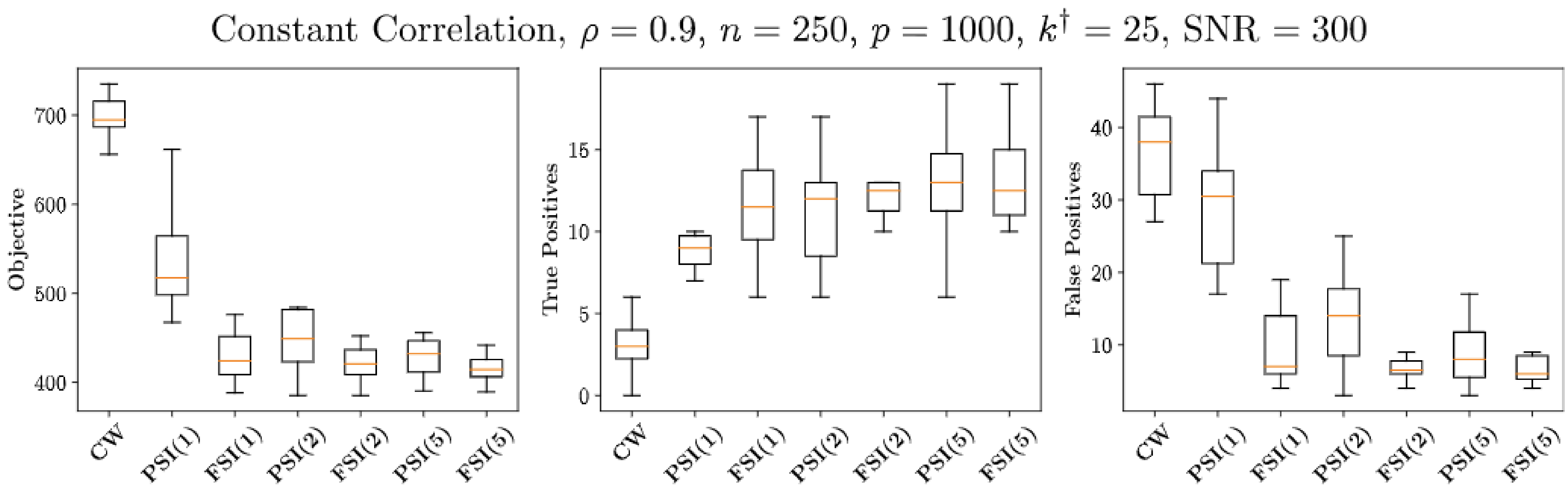
FSI(k) Minima

- In order to converge to an FSI(k) minima, we again perform CDSS and after it converges, we check if there is a feasible solution to the following problem

$$\min_{\beta, S_1, S_2} F\left(\beta^l - U^{S_1}\beta^l + U^{(S/S_1)\cup S_2}\beta\right) \; s.t. S_1 \subseteq S, S_2 \subseteq S^c, |S_1| \leq k, |S_2| \leq k$$

- If there is a solution to this problem, then we update the support and β and repeat both steps again
- If there is no feasible solution, then stop and declare convergence

Comparison of Methods



Extended comparisons of best subset selection

Relaxed lasso

Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso

Following "Best Subset Selection from a Modern Optimization Lens" by Bertsimas, King, and Mazumder (2016)

Trevor Hastie   Robert Tibshirani   Ryan J. Tibshirani

A (Simplified) Relaxed Lasso

- X_{A_λ}: is a submatrix of X that only contains the columns of the nonzero coefficients for the lasso solution for the given λ
- The authors use the following relaxed lasso estimator
  - $\hat{\beta}^{relax}(\lambda, \gamma) = \gamma\hat{\beta}^{lasso}(\lambda) + (1 - \gamma)\hat{\beta}^{LS}(\lambda)$
  - $\hat{\beta}^{LS}(\lambda) = \left(X_{A_\lambda}^T X_{A_\lambda}\right)^{-1} X_{A_\lambda}^T Y$
- $\gamma \in [0, 1]$

IOWA

Extended comparisons of best subset selection

Simulations

Variable Definitions

- n, p: problem dimensions
- s: sparsity level (number of nonzero coefficients)
- Beta-type: pattern of sparsity
- ρ: predictor autocorrelation level
- ν: Signal to noise ratio (SNR) level

Four Beta-Type Settings

Beta-type 1

- s components equal to 1 at roughly equally-spaced indices between 1 and p
- The rest are zero

Beta-type 2

- The first s components equal to 1
- The rest are zero

Beta-type 3

- The first s components are nonzero values equally-spaced between 10 and 0.5
- The rest equal to 0

Beta-type 5

- The first s components equal to 1
- The rest decaying exponentially to 0, $0.5^{i-s}, for\ i = s + 1, \dots, p$

Steps

I. Defined coefficients according to s and beta-type

II. Drew the rows of the predictor matrix $X \in \mathbb{R}^{n \times p}$ i.i.d. from $N_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ has entry (i, j) equal to $\rho^{|i-j|}$

III. Drew the response vector $Y \in \mathbb{R}^n$ from $N_n(X\beta_0, \sigma^2 I)$, with σ^2 defined to meet the desired SNR level, i.e. $\sigma^2 = \frac{\beta_0^T \Sigma \beta_0}{\nu}$

IOWA

Steps

IV. Ran the lasso, relaxed lasso, forward stepwise selection, and best subset selection on the data each over a wide range of tuning parameter values

V. Record metrics of interest
   - Relative risk, relative test error, proportion of variance explained (PVE), number of nonzeros

VI. Repeat steps ii-v a total of 10 times, and average the results

IOWA

Configuration

- Considered following problem settings
  - Low: $n = 100, p = 10, s = 5$
  - Medium: $n = 500, p = 100, s = 5$
  - High-5: $n = 50, p = 1000, s = 5$
  - High-10: $n = 100, p = 1000, s = 10$
- Predictor autocorrelation considered $\rho = 0, \mathbf{0.35}, 0.7$
- The following values for the SNR and corresponding PVE were considered

SNR	0.05	0.09	0.14	0.25	0.42	0.71	1.22	2.07	3.52	6.00
PVE	0.05	0.08	0.12	0.20	0.30	0.42	0.55	0.67	0.78	0.86

Tuning

Setting	Lasso	Relaxed Lasso	Forward Selection and Best Subset Selection
Low Setting	Tuned over 50 values of λ	Same 50 values as lasso Tuned over 10 values of γ equally spaced from 0 to 1	Tuned over subsets of size $k = 0, \ldots, 10$
All other Settings	Tuned over 100 values of λ	Same 100 values as lasso Same 10 values of γ	Tuned over subsets of size $k = 0, \ldots, 50$
Tuning performed by minimizing prediction error on an external validation set of size n which was independently and identically generated			

IOWA

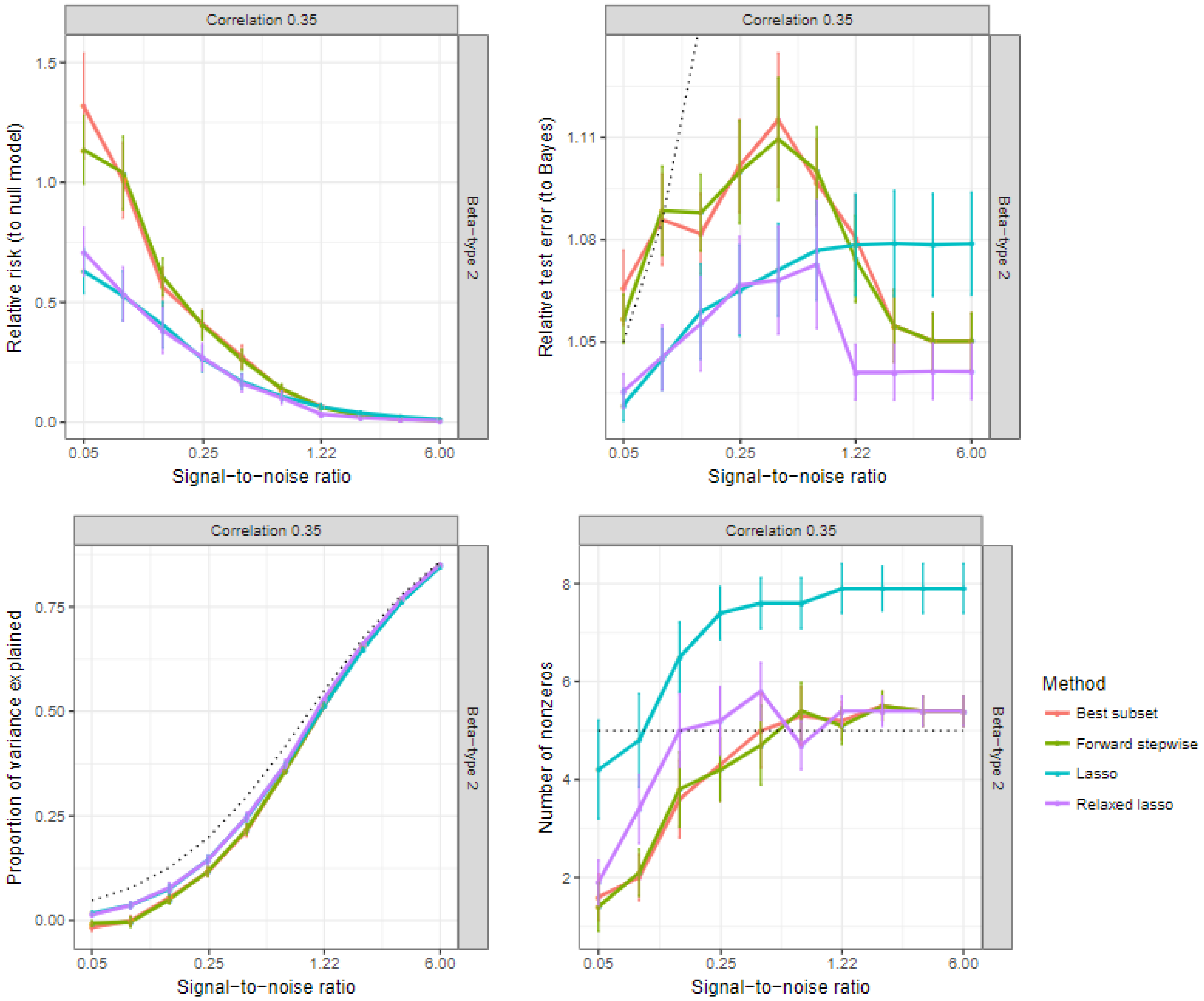
Setting	BS	FS	Lasso	RLasso
low ($n = 100,\ p = 10,\ s = 5$)	3.43	0.006	0.002	0.002
medium ($n = 500,\ p = 100,\ s = 5$)	$\approx 120\ min$	0.818	0.009	0.009
high-5 ($n = 50,\ p = 1000,\ s = 5$)	$\approx 126\ min$	0.137	0.011	0.011
high-10 ($n = 100,\ p = 1000,\ s = 10$)	$\approx 144\ min$	0.277	0.019	0.021

Computational Costs

- Rather than approximating the exact solution as seen in previous paper, these authors try to find the exact solution for best subset selection
- Despite recent advances in best subset selection, it can still be very slow, so a time limit of 3 minutes was set for each subset size

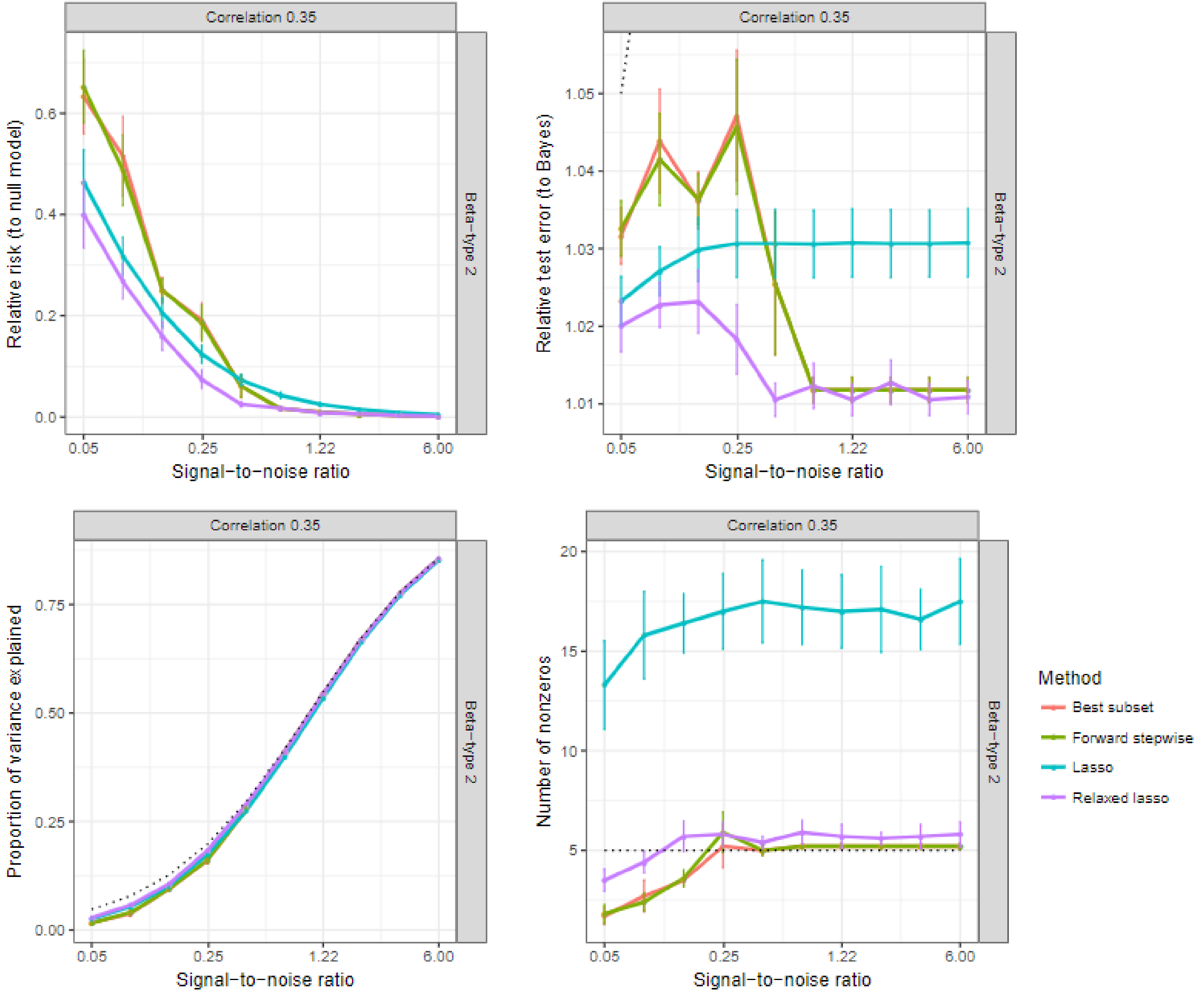
Low Setting
Beta-type 2

Low setting: $n = 100,\ p = 10,\ s = 5$
Correlation $\rho = 0.35$**, beta-type 2**



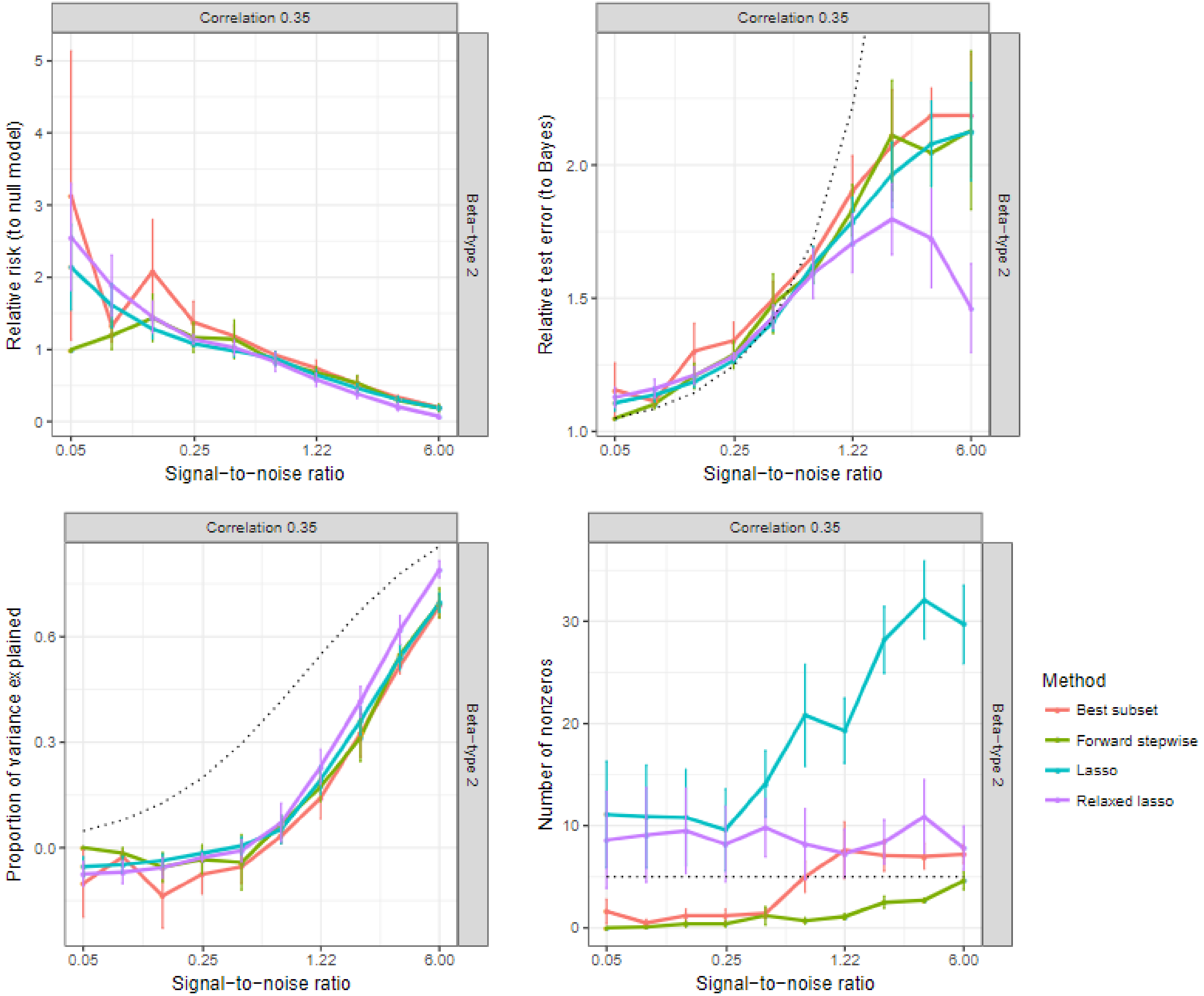
Medium Setting
Beta-type 2

Medium setting: $n = 500,\ p = 100,\ s = 5$
Correlation $\rho = 0.35$**, beta-type 2**



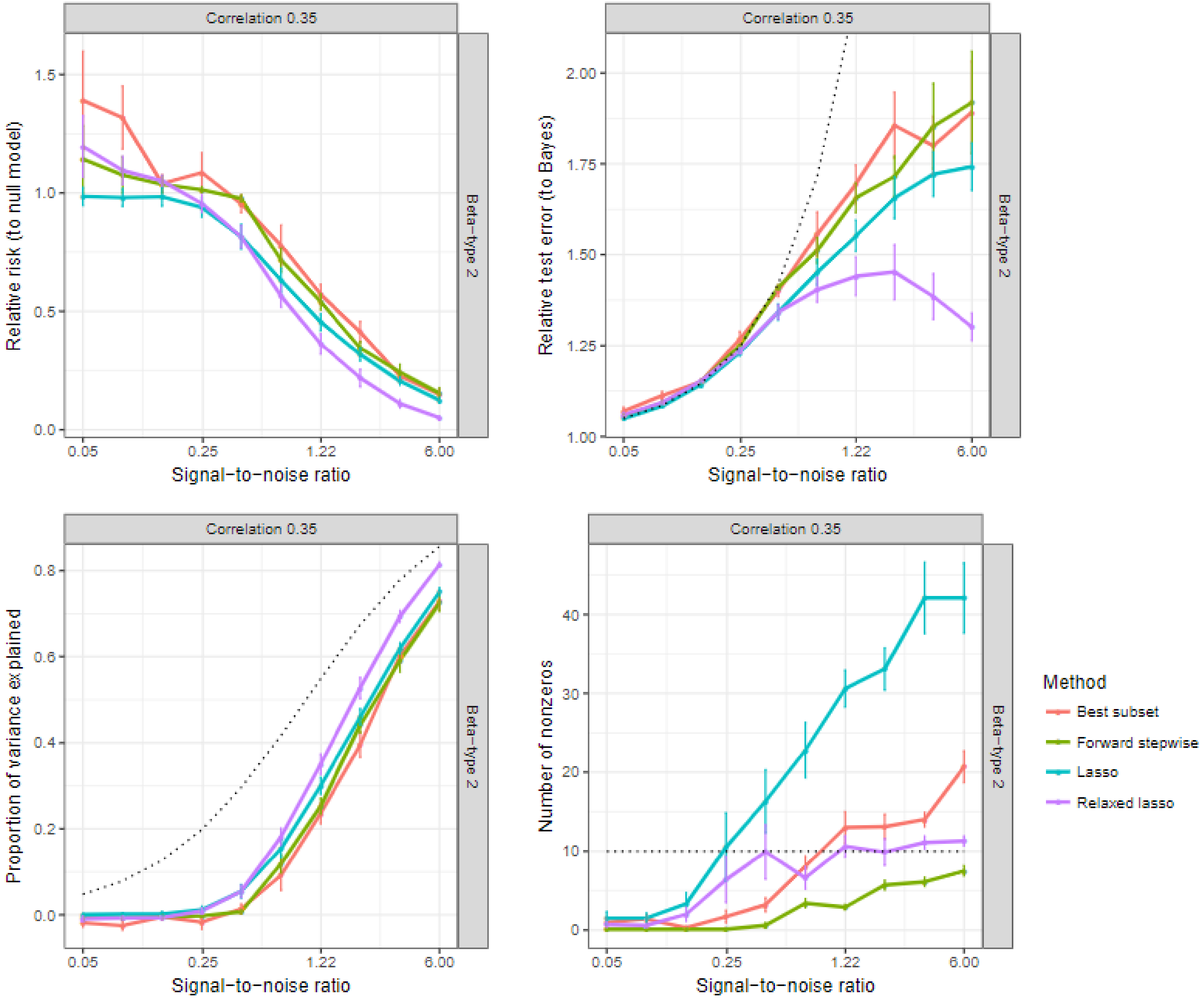
High-5 Setting
Beta-type 2

High-5 setting: $n = 50$, $p = 1000$, $s = 5$
Correlation $\rho = 0.35$, beta-type 2



High-10 Setting
Beta-type 2

High-10 setting: $n = 100,\ p = 1000,\ s = 10$
Correlation $\rho = 0.35$, **beta-type 2**



Extended comparisons of best subset selection

Discussion

Conclusions

- Best subset selection may have underperformed due to 3-minute per problem instance per subset size restriction
  - Particularly at high SNR levels in the high settings
- Lasso gives better results than best subset selection in the low SNR range and worse in the high SNR range
  - The transition point between specific SNR depends on problem dimensions
- Relaxed lasso performs as well as or better than all other methods
  - Utilized γ to get heavy shrinkage from lasso when useful and reverses it when it is not useful
- Comparable PVE results suggest that best practice is to favor the method that is easiest to compute

References

- Hazimeh, H., & Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, *68*(5), 1517–1537. https://doi.org/10.1287/opre.2019.1919
- Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

IOWA

Questions?

Devin Spolsdoff
Jacob Seedorff
University of Iowa
Department of Biostatistics

uiowa.edu

Appendix

Classes of Minima

- Coordinate-wise (CW) minima
  - Solutions where optimizing with respect to one coordinate at a time cannot improve the objective function
- Partial swap-inescapable minima of order k (PSI(k) minima)
  - Solutions where removing any subset of size at most k of the support, adding a subset of size at most k to the support, and optimizing over the newly added subset cannot improve the objective function
- Full swap-inescapable minima of order k (FSI(k) minima)
  - Solutions where removing any subset of size at most k of the support, adding a subset of size at most k to the support, and optimizing over the new support cannot improve the objective function

A (Simplified) Relaxed Lasso

- A_λ: set that contains all the indices of where the nonzero variables are for the lasso solution of a given λ
- X_{A_λ}: is a submatrix of X that only contains the columns of the nonzero coefficients for the lasso solution for the given λ
- The authors use the following relaxed lasso estimator
  - $\hat{\beta}^{relax}(\lambda, \gamma) = \gamma \hat{\beta}^{lasso}(\lambda) + (1 - \gamma)\hat{\beta}^{LS}(\lambda)$
  - $\hat{\beta}^{LS}(\lambda) = \left(X_{A_\lambda}^T X_{A_\lambda}\right)^{-1} X_{A_\lambda}^T Y$
- $\gamma \in [0, 1]$

IOWA

Solution to PSI(k) Subproblem

$$
\begin{aligned}
\min_{\theta,\beta,z} \quad & f(\theta) + \lambda_0 \sum_{i\in[p]} z_i \\
\text{s.t.} \quad & \theta = \beta^{\ell} - \sum_{i\in S} e_i \beta_i^{\ell}(1 - z_i) + \sum_{i\in S^C} e_i \beta_i \\
& -\mathcal{M} z_i \le \beta_i \le \mathcal{M} z_i, \quad \forall i \in S^c \\
& \sum_{i\in S^c} z_i \le k \\
& \sum_{i\in S} z_i \ge |S| - k \\
& \beta_i \in \mathbb{R}, \quad \forall i \in S^c \\
& z_i \in \{0,1\}, \quad \forall i \in [p],
\end{aligned}
$$

IOWA

Solution to FSI(k) Subproblem

$$
\begin{aligned}
\min_{\theta, w, z} \quad & f(\theta) + \lambda_0 \sum_{i \in [p]} z_i \\
& -\mathcal{M} z_i \le \theta_i \le \mathcal{M} z_i, \quad \forall i \in [p] \\
& z_i \le w_i, \quad \forall i \in S \\
& \sum_{i \in S^c} z_i \le k \\
& \sum_{i \in S} w_i \ge |S| - k \\
& \theta_i \in \mathbb{R}, \quad \forall i \in [p] \\
& z_i \in \{0, 1\}, \quad \forall i \in [p] \\
& w_i \in \{0, 1\} \quad \forall i \in S.
\end{aligned}
$$

IOWA

IOWA