Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons

Myungjun Kim, Seungyeop Hyun and Taeyoung Chang

Seoul National University

December 15, 2021

Contents

Introduction

- In recent work, Bertsimas, King and Mazumder (2016) suggested a *Mixed Integer Optimization* (MIO) approach to solve the best subset selection problem,

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \|Y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_0 \leq k.$$

- Using recent advances in MIO algorithms, they demonstrated that best subset selection can now be solved at much larger problem sizes than what was thought possible.

Is Best Subset the Holy Grail?

- Hastie, Tibshirani and Tibshirani (2020) pointed out that neither best subset nor the lasso uniformly dominate the other over the wide range of signal-to-noise ratio (SNR).
- When there is an observational noise like real world dataset, whether the best subset gives a better estimator than others is a subtle question.
- Different procedures have different operating characteristics, that is, give rise to different bias-variance tradeoffs as tuning parameters vary.

What Is a Realistic Signal-to-Noise Ratio?

Let $y_0 = f(x_0) + \epsilon_0$ where x_0 and ϵ_0 are independent. The SNR and the proportion of variance explained (PVE) are defined as

$$SNR = \frac{Var(f(x_0))}{Var(\epsilon_0)} \quad \text{and} \quad PVE(f) = 1 - \frac{Var(\epsilon_0)}{Var(y_0)} = \frac{SNR}{1 + SNR}.$$

- A PVE of 0.5 (SNR = 1) is rare for noisy observational data, and 0.2 (SNR = 0.25) may be more typical. A PVE of 0.86 (SNR = 6) seems unrealistic.
- Bertsimas, King and Mazumder (2016) considered SNRs in the range of about 2 to 8 in low-dimensional cases, and about 3 to 10 in high-dimensional cases.

Goal

This paper is *not* about:

- What is the best prediction algorithm?
- What is the best variable selector?
- Empirically validating theory for ℓ_0 and ℓ_1 penalties.

Rather, this paper is about:

- The relative merits of the three most canonical forms for sparse estimation in a linear model: ℓ_0, ℓ_1 and forward stepwise selection.

Contents

The Best Subset Problem

The best subset problem is written by

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \|Y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_0 \leq k. \end{aligned}$$

- Best subset finds the k predictors that produces the best fit in terms of squared error.
- It is nonconvex problem and is known to be NP-hard.
- A mixed integer optimization (MIO) formulation for the best subset problem is suggested.

MIO Formulations for the Best Subset Problem

- It can be structured as the following MIO formulation

$$
\begin{aligned}
\underset{\beta,\mathbf{z}}{\text{minimize}} \quad & \|Y - X\beta\|_2^2 \\
\text{subject to} \quad & \beta_i(1 - z_i) = 0, \quad \forall\, i = 1, \cdots, p \\
& z_i \in \{0, 1\}, \quad \forall\, i = 1, \cdots, p \\
& \sum_{i=1}^{p} z_i \leq k.
\end{aligned}
$$

MIO Formulations for the Best Subset Problem

- Adding problem-dependent constants M_U and M_ℓ, a more structured rerpresentation can be given as

$$\begin{aligned} \underset{\beta,\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\beta^T(X^TX)\beta - \langle X^Ty, \beta\rangle + \frac{1}{2}\|y\|_2^2 \\ \text{subject to} \quad & \beta_i(1-z_i) = 0, \quad \forall\, i = 1, \cdots, p \\ & z_i \in \{0,1\}, \quad \forall\, i = 1, \cdots, p \\ & \sum_{i=1}^p z_i \le k, \\ & \|\beta\|_\infty \le M_U \quad \text{and} \quad \|\beta\|_1 \le M_\ell. \end{aligned}$$

- Utilizing these bounds typically leads to improved performance of MIO.
- For $n < p$ case, we add another optimization variable $\xi \in \mathbb{R}^n$ with constraints $\xi = X\beta$.

Obtaining Warmstart for the Optimization

- Our situation can be viewed as

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & g(\beta) \\ \text{subject to} \quad & \|\beta\|_0 \leq k, \end{aligned}$$

  where $g(\beta) = \frac{1}{2}\|y - X\beta\|_2^2$.
- Note that g is convex and has Lipshcitz continuous gradient with Lipshcitz constant $\ell = \lambda_{\max}(X^T X)$.
- For such convex function $g(\beta)$, with any $L \geq \ell$ we have

$$g(\eta) \leq Q_L(\eta, \beta) = g(\beta) + \frac{L}{2}\|\eta - \beta\|_2^2 + \langle \nabla(\beta), \eta - \beta \rangle.$$

- We want to find $\text{argmin}_{\|\eta\|_0 \leq k} Q_L(\eta, \beta)$ with given β for getting close to the minimizer of $g(\beta)$.

Projected Gradient Method

- $\operatorname{argmin}_{\|\eta\|_0 \le k} Q_L(\eta, \beta)$ has a closed form solution which is

$$H_k\left(\beta - \frac{1}{L}\nabla g(\beta)\right),$$

  where $H_k(\mathbf{c})$ denotes the projection to the coordinates having k largest (in absolute value) elements of $\mathbf{c}$.
- By updating

$$\beta_{m+1} \in H_k\left(\beta_m - \frac{1}{L}\nabla g(\beta_m)\right),$$

  we can find a stationary point of the main problem.
- We exploit this value as a warmstart for the optimization of MIO problem using solver.

Contents

Forward Stepwise Selection

Forward stepwise is less ambitious version of best subset.

- It starts with the empty model and iteratively adds the variable that best improves the fit.
- Formally, the procedure starts with an empty active set A_0 and for each step $k = 1, \cdots, \min\{n, p\}$, we select the variable indexed by

$$j_k = \text{argmin}_{j \notin A_{k-1}} \|y - P_{A_{k-1} \cup \{j_k\}} y\|_2^2.$$

- It means that it chooses the variable that leads to the lowest squared error when added to A_{k-1}.

Forward Stepwise Selection

- Equivalently, it adds the variable which achieves the maximum absolute correlation with y after we project out the contributions from $X_{A_{k-1}}$.

$$
\begin{aligned}
&\text{minimize} \quad \|y - P_{A_{k-1}\cup\{j_k\}} y\|_2^2 \\
&\Leftrightarrow \text{maximize} \quad \|P_{A_{k-1}\cup\{j_k\}} y\|_2^2 \qquad \because \text{ Pythagorean Law} \\
&\Leftrightarrow \text{maximize} \quad \|P_{(I-P_{A_{k-1}})x_{j_k}} y\|_2^2 \qquad \because \ \|P_{A_{k-1}\cup\{j_k\}} y\|_2^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad = \|P_{A_{k-1}} y\|_2^2 + \|P_{(I-P_{A_{k-1}})x_{j_k}} y\|_2^2 \\
&\Leftrightarrow \text{maximize} \quad \frac{|\langle\, (I - P_{A_{k-1}})X_{j_k}\ ,\ y\,\rangle|}{\|(I - P_{A_{k-1}})X_{j_k}\|_2}
\end{aligned}
$$

Algorithm for Forward Selection

- The forward stepwise selection is highly structured and this greatly aids its computation.
- Suppose that we have maintained a QR decomposition of active submatrix $X_{A_{k-1}}$ of predictors and the orthogonalized remaining predictors with respect to $X_{A_{k-1}}$.
- Then we find one of remaining predictor which has maximum absolute correlation with y.
- To update

$$X_{A_{k-1}} = Q_{k-1}R_{k-1} \quad \text{to} \quad X_{A_k} = Q_k R_k$$

with selected variable X_{j_k}, we shall take advantage of modified Gram-Schimidt algorithm.

Algorithm for Forward Selection

- Using MGS, we can derive k-th column of Q_k and k-th column of R_k

$$
\begin{aligned}
\mathbf{v}_k &= \mathbf{x}_{j_k} - P_{\mathrm{span}(\{\mathbf{q}_1,\cdots,\mathbf{q}_{k-1}\})}(\mathbf{x}_{j_k}) = \mathbf{x}_{j_k} - \sum_{j=1}^{k-1} \langle \mathbf{q}_j, \mathbf{x}_k \rangle \cdot \mathbf{q}_j \\
&= \mathbf{x}_k - \sum_{j=1}^{k-1} \left\langle \mathbf{q}_j, \mathbf{x}_{j_k} - \sum_{i=1}^{j-1} \langle \mathbf{q}_i, \mathbf{x}_{j_k} \rangle \mathbf{q}_i \right\rangle \cdot \mathbf{q}_j \\
\mathbf{q}_k &= \mathbf{v}_k / \|\mathbf{v}_k\|_2
\end{aligned}
$$

- Orthogonalizing the remaing predictors with respect to the one just included can be done using Q_k since $I - P_{A_k} = I - Q_k Q_k^T$.

Contents

The Lasso

The lasso problem is written by

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1, \quad \lambda \geq 0.$$

- It solves a convex relaxtion of best subset problem where we replace the ℓ_0 norm by the ℓ_1 norm.

The Lasso with Pathwise Coordinate Descent

- R package `glmnet` solves the lasso problem using *pathwise coordinate descent*.
- Compute the solutions for a decreasing sequence of λ,

$$\lambda_{\text{max}} = \|X^T Y\|_\infty > \cdots > \lambda_{\text{min}} = \epsilon\lambda_{\text{max}},$$

  typically with $\epsilon = 0.001$ and $K = 100$ values of λ on the log scale.
- Starting at λ_{max}, where all coefficients of the solution $\hat{\beta}$ are zero, we use *warm starts* in computing the solutions at the sequence of λ, i.e., $\hat{\beta}(\lambda_k)$ is used as an initial value for λ_{k+1}.

Active Set Strategy

- After one or several cycles through p variables, store the nonzero coefficient in the active set $\mathcal{A}$.
- Iterates coordinate descent restricting further iterations to $\mathcal{A}$ till convergence.
- One more cycle through all variables to check KKT optimality conditions:

$$\begin{aligned} \left|\langle\, x_j\,,\, y - X\hat{\beta}(\lambda)\,\rangle\right| &= \lambda \quad \text{for all members of the active set,} \\ \left|\langle\, x_j\,,\, y - X\hat{\beta}(\lambda)\,\rangle\right| &\leq \lambda \quad \text{for all variables not in the active set.} \end{aligned}$$

If there were a variable violating the conditions, then add it in $\mathcal{A}$ and go back to the previous step.

Screening Rule

- For some problems, screening rules can be used in combination with coordinate descent to further wittle down the active set.
- For the lasso, Tibshirani (2012) suggested the *sequential strong rules* which discards the jth predictor from the optimization problem at λ_k if
$$\left|\langle\, x_j \,,\, y - X\hat{\beta}(\lambda_{k-1})\,\rangle\right| < 2\lambda_k - \lambda_{k-1}.$$

Screening Rule

- Motivation for the strong rules comes with KKT conditions.
- If we assume that we can bound the amount that $c_j(\lambda) = \langle\, x_j\,,\, y - X\hat{\beta}(\lambda)\,\rangle$ changes as we move from λ to another $\tilde{\lambda}$, i.e.,

$$|c_j(\lambda) - c_j(\tilde{\lambda})| \leq |\lambda - \tilde{\lambda}| \quad \forall\, \lambda\,,\, \tilde{\lambda}, \quad \text{and} \quad \forall j = 1, \cdots, p$$

  then $|c_j(\lambda_{k-1})| < 2\lambda_k - \lambda_{k-1}$ (which satisfying strong rule) implies

$$\begin{aligned} |c_j(\lambda_k)| &\leq |c_j(\lambda_k) - c_j(\lambda_{k-1})| + |c_j(\lambda_{k-1})| \\ &< (\lambda_{k-1} - \lambda_k) + (2\lambda_k - \lambda_{k-1}) = \lambda_k \end{aligned}$$

  so that $\hat{\beta}_j(\lambda_k) = 0$ by the KKT conditions.
- The sequential strong rule can mistakenly discard active predictors, so it must be combined with a check of the KKT conditions.

Algorithm for Lasso implemented by `glmnet`

- Using both *ever-active* set of predictors $\mathcal{A}(\lambda)$ and the strong set $S(\lambda)$ which is the set of the indices of the predictors that survive the screening rule can be advantageous.
  1. Set $\mathcal{E} = \mathcal{A}(\lambda)$.
  2. Solve the problem at value λ by using only the predictors in $\mathcal{E}$.
  3. Check the KKT conditions at this solution for all predictors in $S(\lambda)$. If violated, then add these violating predictors into $\mathcal{E}$ and go back to previous step using the current solution as a warm start.
  4. Check the KKT conditions at all predictors. No violations means we are done. Otherwise, add these violators into $\mathcal{E}$, recompute $S(\lambda)$ and go back to the first step using the current solution as a warm start.
- Note that violations in the third step are faily common whereas those in the fourth step are rare. Hence the fact that the size of $S(\lambda)$ is very much less than p makes this an effective strategy.

A (Simplified) Relaxed Lasso

A simplified version of the relaxed lasso estimator is

$$\hat{\beta}^{\text{relax}}(\lambda, \gamma) = \gamma \hat{\beta}^{\text{lasso}}(\lambda) + (1-\gamma)\hat{\beta}^{\text{LS}}(\lambda),$$

where $\lambda \geq 0$ and $\gamma \in [0, 1]$.

- $\mathcal{A}_\lambda$: the active set of $\hat{\beta}^{\text{lasso}}(\lambda)$
- $\hat{\beta}^{\text{LS}}_{\mathcal{A}_\lambda} = (X^T_{\mathcal{A}_\lambda} X_{\mathcal{A}_\lambda})^{-1} X^T_{\mathcal{A}_\lambda} y$, i.e., it denotes the least squares solution obtained by regressing of y on $X_{\mathcal{A}_\lambda}$.
- $\hat{\beta}^{\text{LS}}(\lambda)$: the full-sized version of $\hat{\beta}^{\text{LS}}_{\mathcal{A}_\lambda}$, padded with zeros.
- The relaxed lasso tries to undo the shrinkage inherent in the lasso estimator to a varying degree depending on γ.

Contents

Setup

- Define coefficients $\beta_0 \in \mathbb{R}^p$ according to s (sparsity level) and the beta-type.
- The predictor matrix $X \in \mathbb{R}^{n \times p}$ i.i.d. from $N_p(0, \Sigma)$ where $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho \in \{0, 0.35, 0.70\}$.
- 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., $\sigma^2 = \beta_0^T \Sigma \beta_0 / \nu$.
- Run the lasso, relaxed lasso, forward stepwise, and best subset on the data over a wide range of parameters, and choose the parameter by minimizing prediction error on a validation set.
- Record several metrics of interest and repeat total of 10 times, and average the results.

Coefficients

- beta-type 1: β_0 has s components equal to 1, occuring at equally-spaced indices between 1 and p, and the rest equal to 0.
- beta-type 2: β_0 has its first s components equal to 1, and the rest equal to 0.
- beta-type 3: β_0 has its first s components taking nonzero values equally-spaced between 10 and 0.5, and the rest equal to 0
- beta-type 5: β_0 has its first s components equal to 1, and the rest decaying exponentially to 0, specifically, $\beta_{0i} = 0.5^{i-s}$, for $i = s+1, \ldots, p$.

Configurations

We considered the following four problem settings:

Setting	n	p	s
Low	100	10	5
Medium	500	100	5
High-5	50	1000	5
High-10	100	1000	10

In each setting, we considered ten values for the SNR ranging from 0.05 to 6 on a log scale:

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

Evaluation Metrics

- Relative risk:

$$RR(\hat{\beta}) = \frac{\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0)^2}{\mathbb{E}(x_0^T \beta_0)^2} = \frac{(\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0)}{\beta_0^T \Sigma \beta_0}$$

- Relative test error:

$$RTE(\hat{\beta}) = \frac{\mathbb{E}(y_0 - x_0^T \hat{\beta})^2}{\sigma^2} = \frac{(\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0) + \sigma^2}{\sigma^2}$$

- Proportion of variance explained:

$$PVE(\hat{\beta}) = 1 - \frac{\mathbb{E}(y_0 - x_0^T \hat{\beta})^2}{Var(y_0)} = 1 - \frac{(\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0) + \sigma^2}{\beta_0^T \Sigma \beta_0 + \sigma^2}$$

- Number of nonzeros: $\|\hat{\beta}\|_0 = \sum_{i=1}^p 1\{\hat{\beta}_i \neq 0\}$
- F-Score: the harmonic mean of recall and precision

Results: Computation Time

Setting	n	p	s	BS	FS	Lasso	RLasso
Low	100	10	5	0.313	0.003	0.002	0.002
Medium	500	100	5	76.8 hr	0.890	0.013	0.154
High-5	50	1000	5	44.2 hr	0.123	0.014	0.159
High-10	100	1000	10	61.7 hr	0.254	0.024	0.158

Table 1: Time in seconds for one path of solutions for each method

Setting	n	p	s	BS	FS	Lasso	RLasso
Low	100	10	5	2.20	0.026	0.0006	0.0009
Medium	500	100	5	4634	1.801	0.004	0.056
High-5	50	1000	5	4896	0.127	0.003	0.018
High-10	100	1000	10	4905	0.454	0.010	0.038

Table 2: Reproduced time in seconds for one path of solutions for each method

Results: Effective Degrees of Freedom

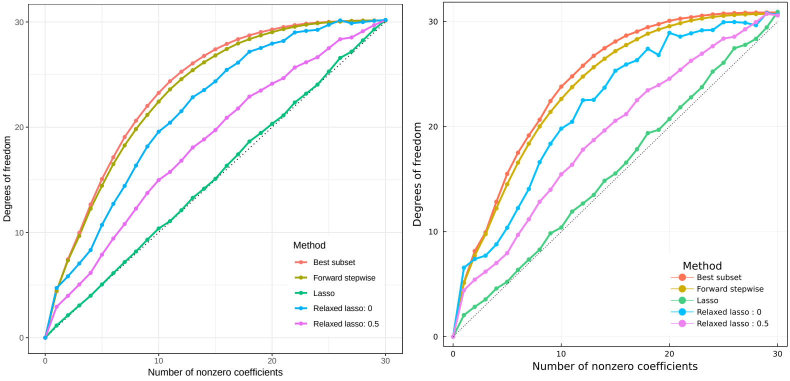


Figure 1: Degrees of freedom $\sum_{i=1}^{n} Cov(Y_i, \hat{Y}_i)/\sigma^2$ for the lasso, forward stepwise, best subset and the relaxed lasso with $\gamma = 0.5$ and $\gamma = 0$.

Results: Accuracy Metrics (Low)

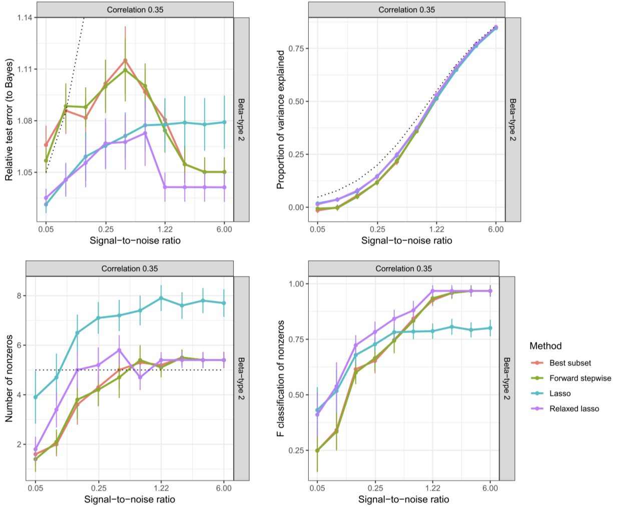
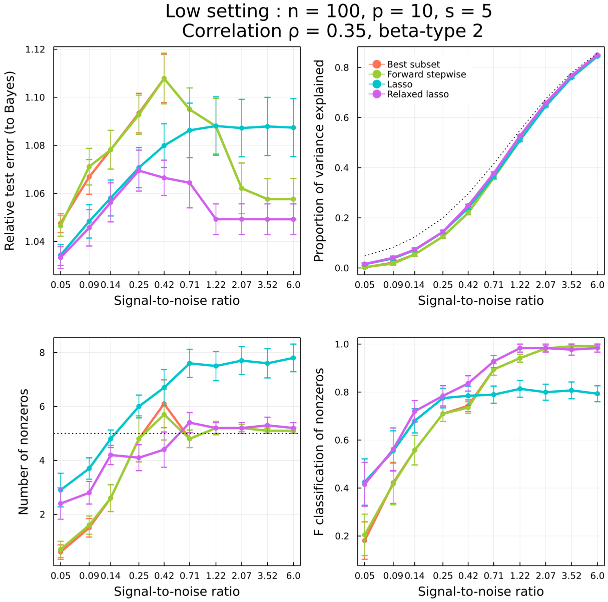


Figure 2: RTE, PVE, number of nonzero coefficients and F-score as functions of SNR with $n = 100, p = 10, s = 5, \rho = 0.35$ and beta-type 2.

Results: Accuracy Metrics (Low)



Results: Accuracy Metrics (Medium)

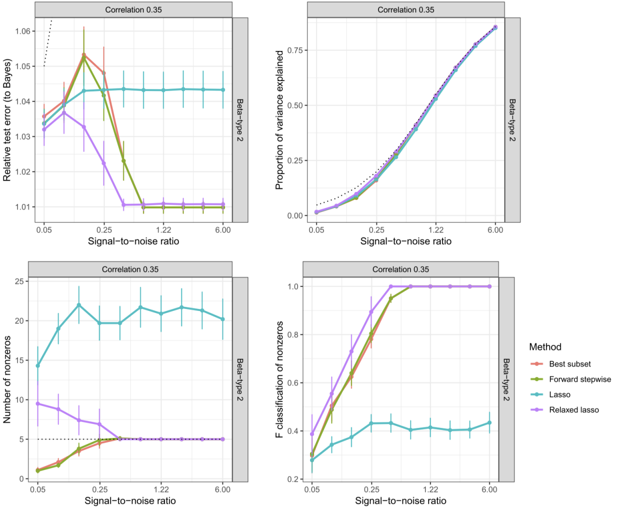
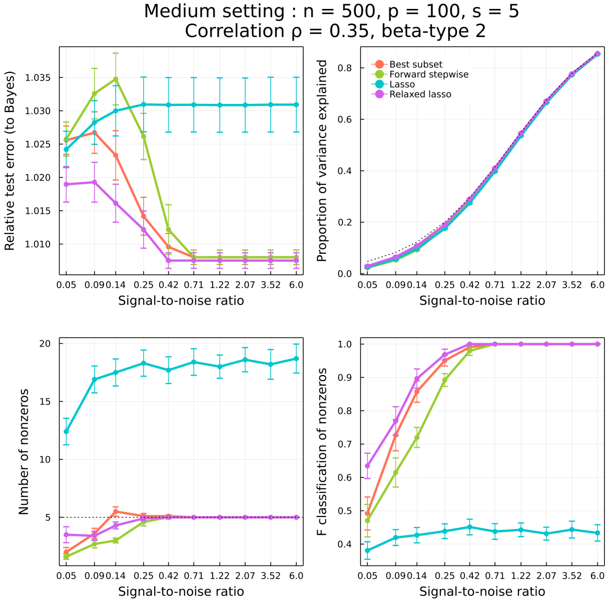


Figure 3: RTE, PVE, number of nonzero coefficients and F-score as functions of SNR with $n = 500, p = 100, s = 5, \rho = 0.35$ and beta-type 2.

Results: Accuracy Metrics (Medium)



Results: Accuracy Metrics (High-5)

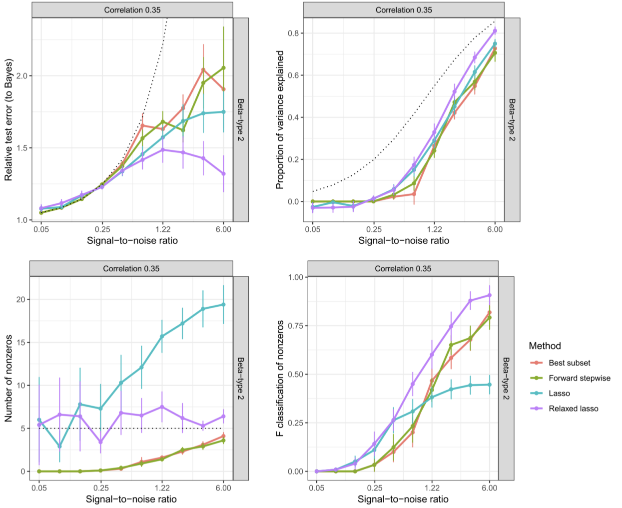
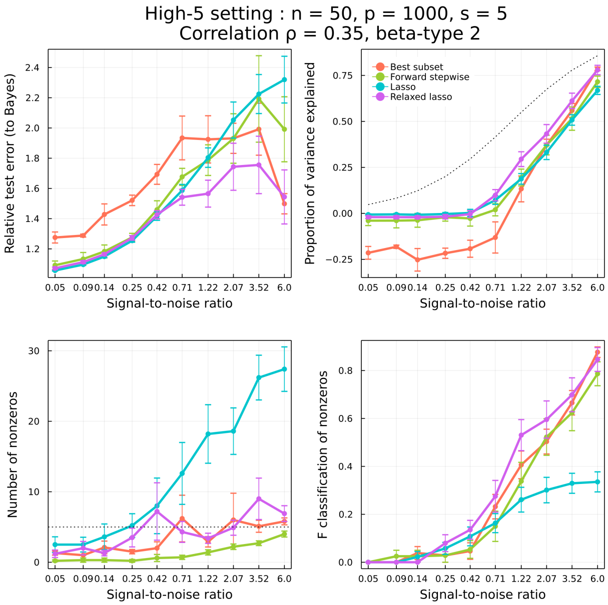


Figure 4: RTE, PVE, number of nonzero coefficients and F-score as functions of SNR with $n = 50, p = 1000, s = 5, \rho = 0.35$ and beta-type 2.

Results: Accuracy Metrics (High-5)



Results: Accuracy Metrics (High-5)

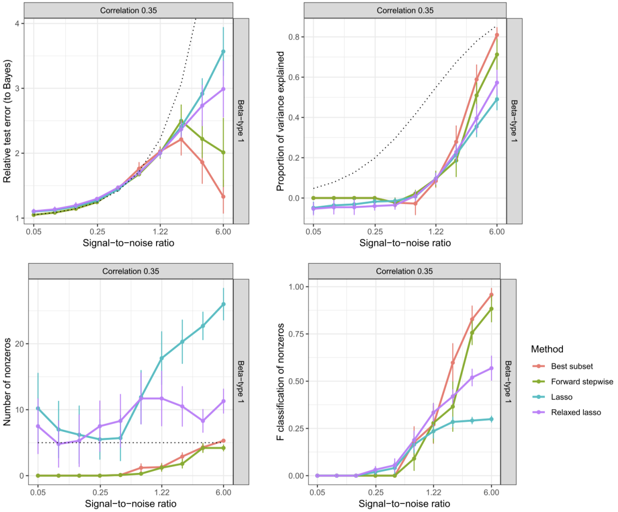
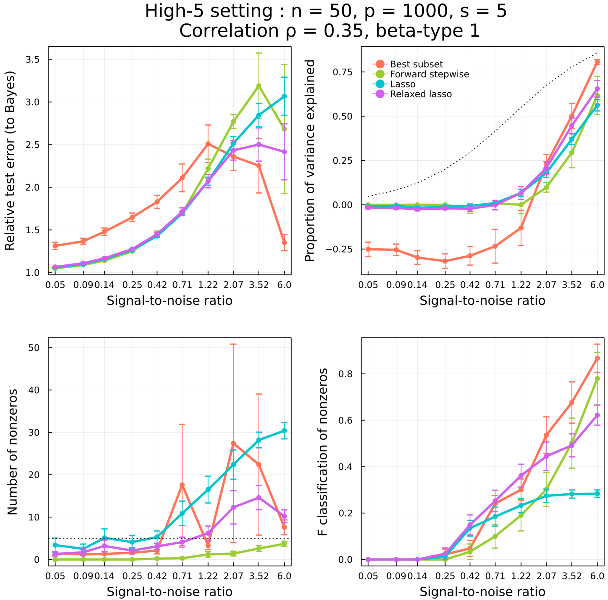


Figure 5: RTE, PVE, number of nonzero coefficients and F-score as functions of SNR with $n = 50, p = 1000, s = 5, \rho = 0.35$ and beta-type 1.

Results: Accuracy Metrics (High-5)



Summary of Results

- The lasso and relaxed lasso are very fast and forward stepwise is also fast, though not quite as fast as the lasso.
- In high-5 and high-10 setting, best subset selection gives very poor accuracy because of time-limit.
- Forward stepwise selection and best subset selection perform quite similarly over all settings, but the former one is much faster.
- In the low SNR range, the lasso outperforms the best subset selection while it has worse accuracy than best subset selection in the high SNR range.
- The relaxed lasso performs better than all other methods over all settings.

References

- Hastie, T., Tibshirani, R. and Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*. **35** 579-592.
- Bertsimas, D., King, A. and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*. **44** 813-852.
- Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J and Tibshirani, R. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistics Society*. **74** 245-266.
- Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *The Annals of Statistics*. **1** 302-332.