Variable Selection at Scale

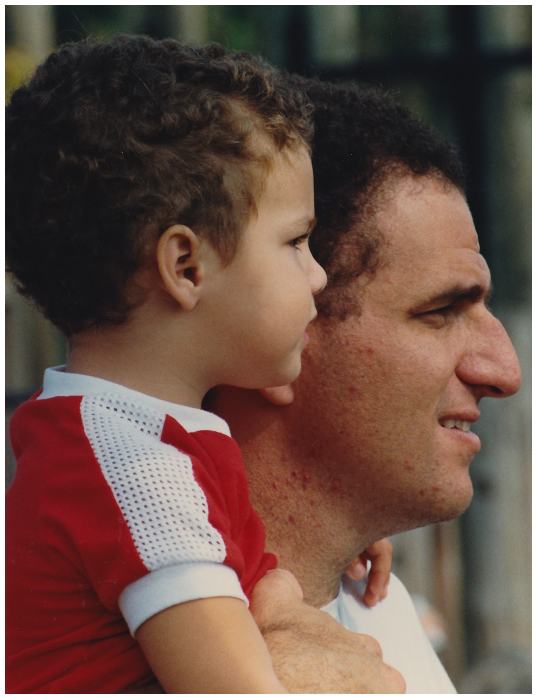
Trevor Hastie, Stanford University

with Ryan Tibshirani and Rob Tibshirani

Variable Selection at Scale

Trevor Hastie, Stanford University

with Ryan Tibshirani and Rob Tibshirani



Outline and Summary

We consider linear regression models $\eta(X) = X^T\beta$ with potentially very large numbers of variables, and methods for selecting an informative subset.

- Revisit two baby boomers (best-subset selection and forward-stepwise selection), one millennial (lasso) and a newborn (relaxed lasso).
- Simulation study to evaluate them all over a wide range of settings.

Conclusions:

- forward stepwise very close to best subset, but much faster.
- relaxed lasso overall winner, and fastest by far.
- In wide-data settings, and low SNR, lasso can beat best subset and forward stepwise.

Outline and Summary

We consider linear regression models $\eta(X) = X^T\beta$ with potentially very large numbers of variables, and methods for selecting an informative subset.

- Revisit two baby boomers (best-subset selection and forward-stepwise selection), one millennial (lasso) and a newborn (relaxed lasso).
- Simulation study to evaluate them all over a wide range of settings.

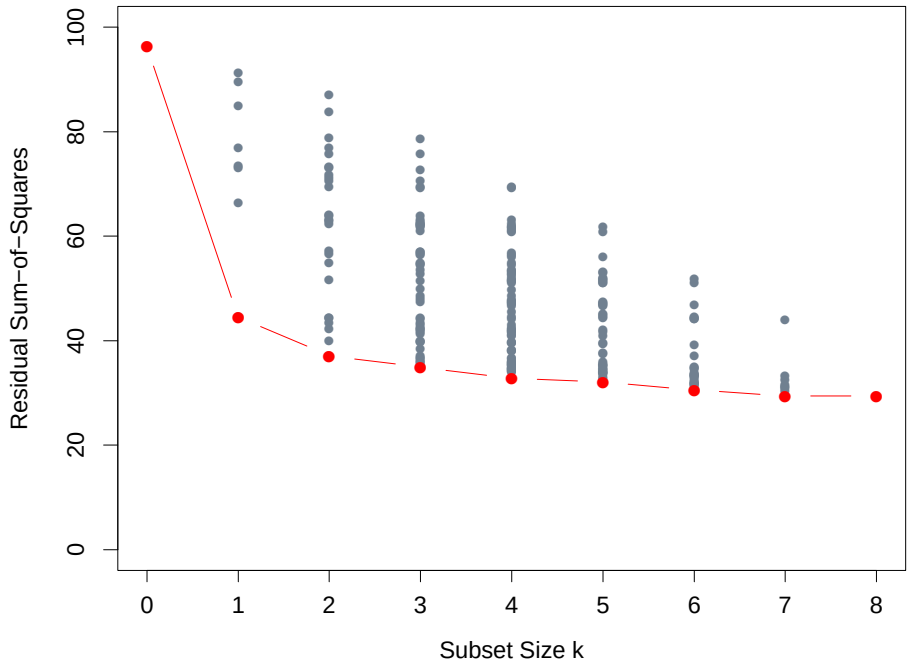
Conclusions:

- forward stepwise very close to best subset, but much faster.
- relaxed lasso overall winner, and fastest by far.
- In wide-data settings, and low SNR, lasso can beat best subset and forward stepwise.

Paper: https://arxiv.org/abs/1707.08692

R package: https://github.com/ryantibs/best-subset/

Best Subset Selection



1. For each subset of size k of the p variables, evaluate the fitting objective (e.g. RSS) via linear regression on the training data.
2. Candidate models $\hat{\beta}_{(k)}$ are at the lower frontier — the best for each k on the training data.
3. Pick $\hat{k}$ using a validation dataset (or CV), and deliver $\hat{\beta}_{(\hat{k})}$

Properties of Best Subset Selection

- ✔ Well-defined goal — the obvious gold standard for variable selection.
- ✔ Feasible for least squares regression with $p \approx 35$ using clever algorithms (Furnival and Wilson, 1974, "Leaps and Bounds").

Properties of Best Subset Selection

- ✔ Well-defined goal — the obvious gold standard for variable selection.
- ✔ Feasible for least squares regression with $p \approx 35$ using clever algorithms (Furnival and Wilson, 1974, "Leaps and Bounds").
- ✖ Combinatorially hard for large p.

Properties of Best Subset Selection

- ✔ Well-defined goal — the obvious gold standard for variable selection.
- ✔ Feasible for least squares regression with $p \approx 35$ using clever algorithms (Furnival and Wilson, 1974, "Leaps and Bounds").
- ✖ Combinatorially hard for large p.
- ? Obvious gold standard — really?

Best Subset Selection Breakthrough

Rahul Mazumder, with Bertsimas and King (AoS 2016) crack the forty year old best-subset selection bottleneck! They use *mixed-integer programming (MIO)* along with the GUROBI solver.

Best Subset Selection Breakthrough

Rahul Mazumder, with Bertsimas and King (AoS 2016) crack the forty year old best-subset selection bottleneck! They use *mixed-integer programming (MIO)* along with the GUROBI solver.

$$\begin{aligned} \text{minimize}_{z,\beta} \quad & \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ \text{subject to} \quad & -Mz_j \le \beta_j \le Mz_j,\ z_j \in \{0,1\},\ j = 1,\ldots,p \\ & \sum_{j=1}^{p} z_j \le k. \end{aligned}$$

Their procedure iteratively narrows the optimality gap — if the gap hits zero, they have found the solution.

Forward Stepwise Selection

Greedy forward algorithm, traditionally thought of as a sub-optimal but feasible alternative to best-subset regression.

1. Start with null model (response mean).
2. Choose among the p variables to find the best single-variable model in terms of fitting objective.
3. Choose among the remaining $p-1$ variables to find the one, when included with the previously chosen variable, best improves the fitting objective.
4. Choose among the remaining $p-2$ ..., and so on.

Forward stepwise produces a nested sequence of models $\hat{\beta}_{(k)},\ k = 1, 2, \ldots$.
Pick k using a validation dataset.

Forward Stepwise Selection Properties

- ✔ Computationally feasible with big data, and also works with $n \ll p$.
- ✔ Efficient computations with squared-error loss. Computations can be arranged as a guided QR decomposition of the X matrix, and hence costs the same as a full least-squares fit $O(np \min(n, p))$.
- ✔ Performance very similar to best subset selection, although difficult counter examples can be constructed.

Forward Stepwise Selection Properties

- ✔ Computationally feasible with big data, and also works with $n \ll p$.
- ✔ Efficient computations with squared-error loss. Computations can be arranged as a guided QR decomposition of the X matrix, and hence costs the same as a full least-squares fit $O(np\min(n,p))$.
- ✔ Performance very similar to best subset selection, although difficult counter examples can be constructed.
- ✖ Efficiency not available for GLMs, although score approximations can be used.
- ✖ Tedious with very large p and n, since terms augmented one at a time.

Lasso

The lasso (Tibshirani, 1996) solves

$$\underset{\beta}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{s.t. } \|\beta\|_1 \leq t$$

Generally, the smaller t, the *sparser* the solutions, and approximate nesting occurs.
We compute many solutions over a range of values of t, and select t using validation data.

Lasso

The lasso (Tibshirani, 1996) solves

$$\underset{\beta}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{s.t. } \|\beta\|_1 \leq t$$

Generally, the smaller t, the *sparser* the solutions, and approximate nesting occurs.
We compute many solutions over a range of values of t, and select t using validation data.

Often thought of as a *convex relaxation* for the best-subset problem

$$\underset{\beta}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{s.t. } \|\beta\|_0 \leq k$$

Lasso properties

We typically solve lasso in *Lagrange* form

$$\underset{\beta}{\text{minimize}}\, \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\|\beta\|_1$$

- ✔ Extremely fast algorithms for solving lasso problems (with many loss functions). Pathwise coordinate descent via GLMNET (Friedman, H, Tibshirani, 2010) exploits sparsity, active-set convergence, strong rules, and more, to rapidly compute entire solution path on a grid of values of λ.
- ✔ With large p provides convenient subset selection, taking *leaps* rather than single steps.

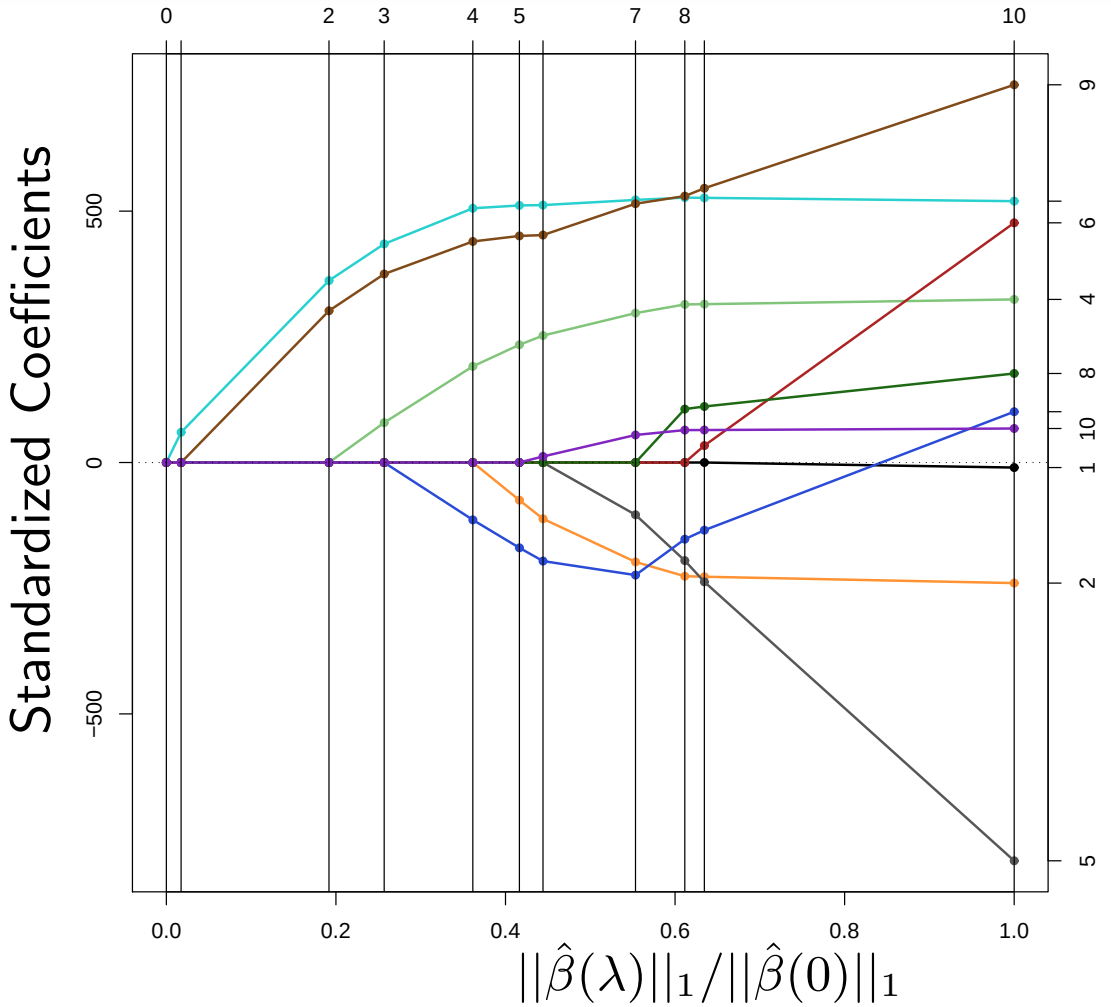
Lasso properties

We typically solve lasso in *Lagrange* form

$$\underset{\beta}{\text{minimize}}\, \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\|\beta\|_1$$

- ✔ Extremely fast algorithms for solving lasso problems (with many loss functions). Pathwise coordinate descent via GLMNET (Friedman, H, Tibshirani, 2010) exploits sparsity, active-set convergence, strong rules, and more, to rapidly compute entire solution path on a grid of values of λ.
- ✔ With large p provides convenient subset selection, taking *leaps* rather than single steps.
- ✖ Since coefficients are both *selected* and *regularized*, can suffer from shrinkage bias.

Lasso Coefficient Path



Lasso: $\hat{\beta}(\lambda) = \text{argmin}_\beta \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda ||\beta||_1$

fit using LARS package in R (Efron, H, Johnstone, Tibshirani 2002)

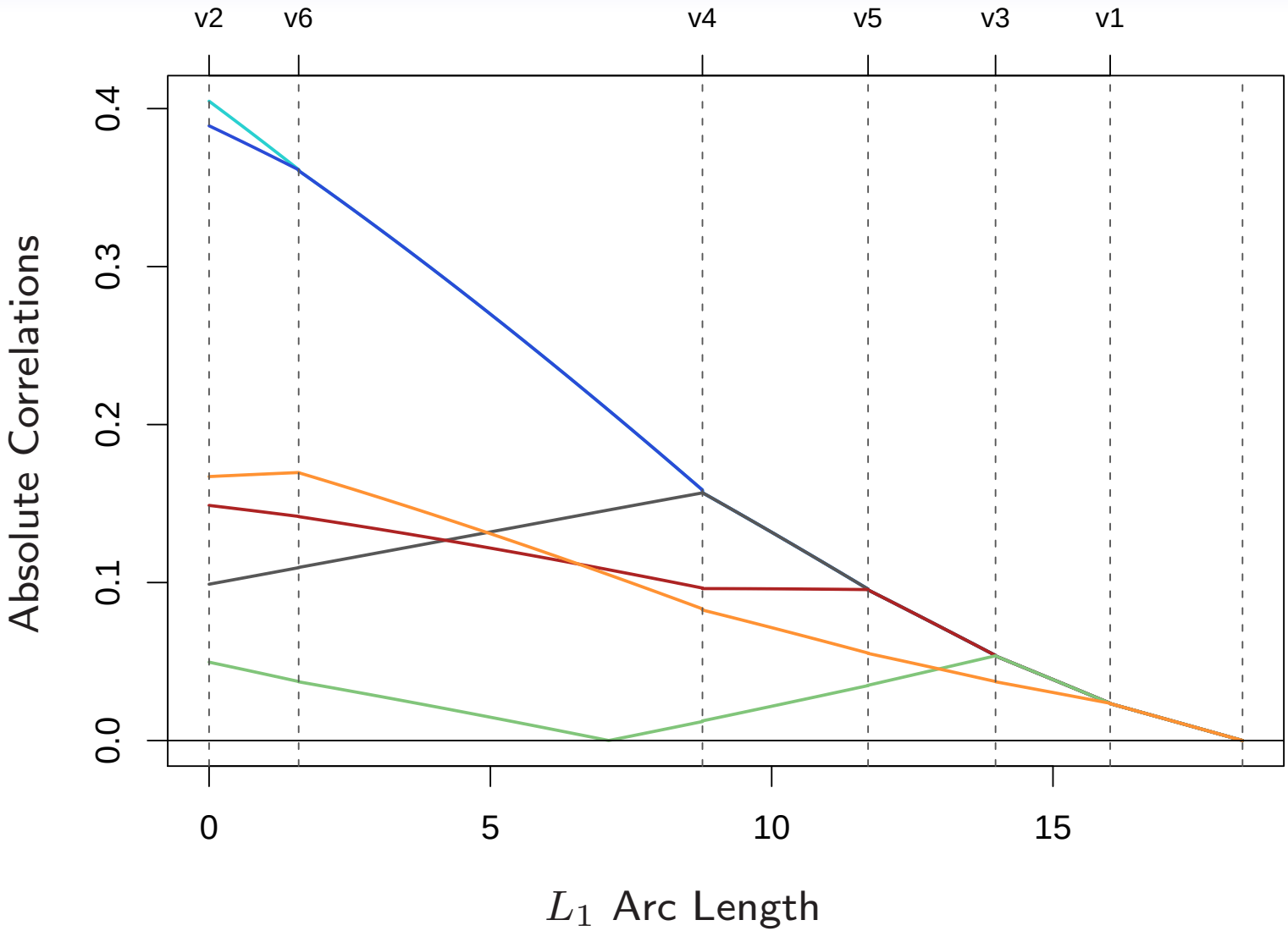
Lasso and Least-Angle Regression (LAR)

Interesting connection between Lasso and Forward Stepwise.

LAR algorithm: Democratic Forward Stepwise

1. Find variable $X_{(1)}$ most correlated with the response.
2. While moving towards the least-squares fit on $X_{(1)}$, keep track of correlations of other variables with the evolving residual.
3. When $X_{(2)}$ catches up in correlation, include it in model, and move the pair toward least squares fit (correlations stay tied!)
4. And so on.

LAR path = Lasso path (almost always).
Forward Stepwise goes all the way with each variable, while LAR lets others in when they catch up. This *slow learning* was inspired by the forward stagewise approach of boosting.


v2
v6
v4
v5
v3
v1
Absolute Correlations
0.0
0.1
0.2
0.3
0.4
0
5
10
15
L_1 Arc Length

Relaxed Lasso

Originally proposed by Meinshausen (2006). We present a simplified version.

- Suppose $\hat{\beta}_\lambda$ is the lasso solution at λ, and let A_λ be the *active set* of indices with nonzero coefficients in $\hat{\beta}_\lambda$.
- Let $\hat{\beta}^{LS}_{A_\lambda}$ be the coefficients in the least squares fit, using only the variables in A_λ. Let $\hat{\beta}^{LS}_\lambda$ be the full-sized version of this coefficient vector, padded with zeros.

  $\hat{\beta}^{LS}_\lambda$ debiases the lasso, while maintaining its sparsity.
- Define the *Relaxed Lasso*

$$\hat{\beta}^{RELAX}_\lambda(\gamma) = \gamma \cdot \hat{\beta}_\lambda + (1-\gamma) \cdot \hat{\beta}^{LS}_\lambda$$

Relaxed Lasso

Originally proposed by Meinshausen (2006). We present a simplified version.

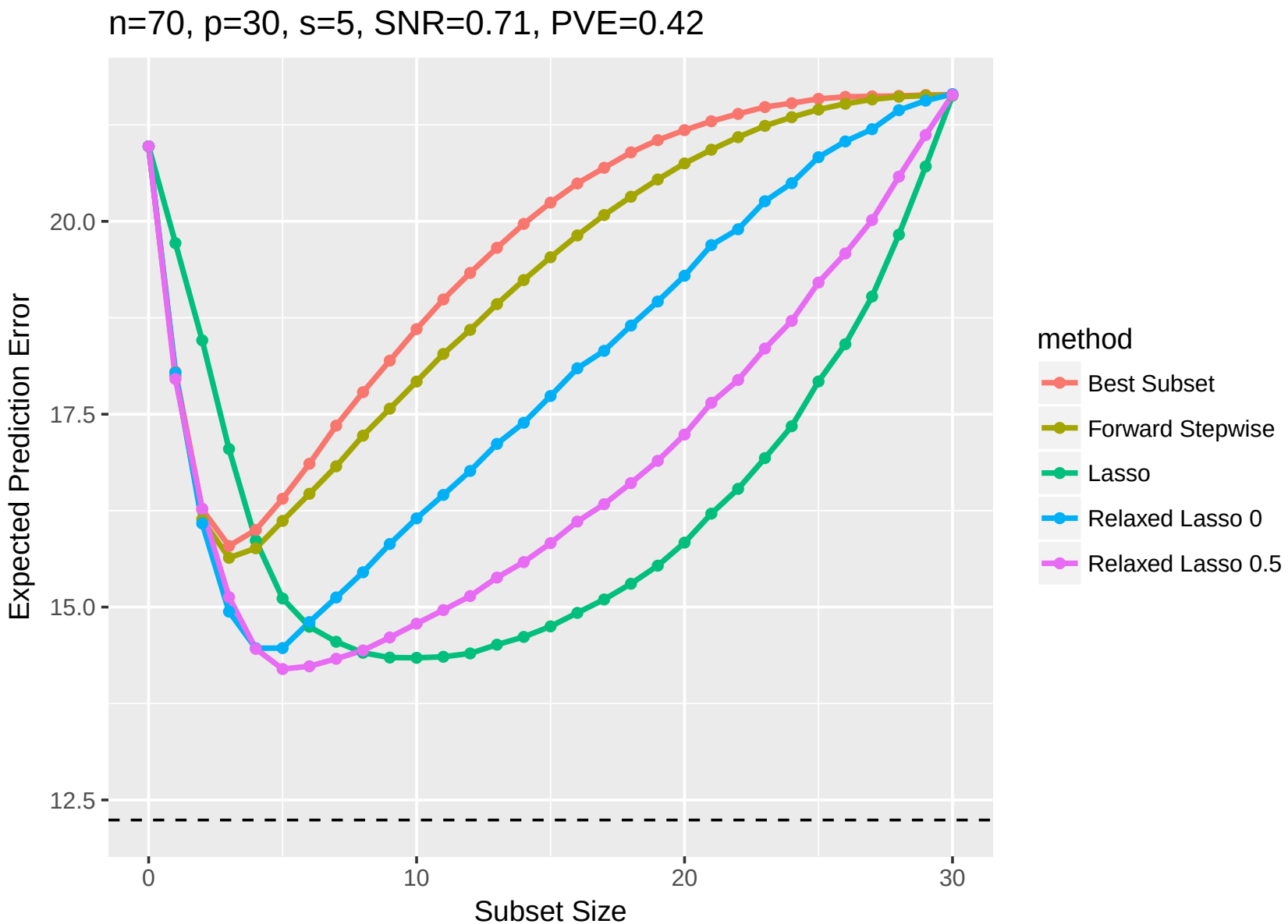
- Suppose $\hat{\beta}_\lambda$ is the lasso solution at λ, and let A_λ be the *active set* of indices with nonzero coefficients in $\hat{\beta}_\lambda$.
- Let $\hat{\beta}^{LS}_{A_\lambda}$ be the coefficients in the least squares fit, using only the variables in A_λ. Let $\hat{\beta}^{LS}_\lambda$ be the full-sized version of this coefficient vector, padded with zeros.

  $\hat{\beta}^{LS}_\lambda$ debiases the lasso, while maintaining its sparsity.
- Define the *Relaxed Lasso*

$$\hat{\beta}^{RELAX}_\lambda(\gamma) = \gamma \cdot \hat{\beta}_\lambda + (1-\gamma) \cdot \hat{\beta}^{LS}_\lambda$$

Once $\hat{\beta}^{LS}_\lambda$ is computed at desired values of λ, the whole family $\hat{\beta}^{RELAX}_\lambda(\gamma)$ comes free of charge!

Simulation


n=70, p=30, s=5, SNR=0.71, PVE=0.42
Expected Prediction Error
20.0
17.5
15.0
12.5
0
10
20
30
Subset Size
method
Best Subset
Forward Stepwise
Lasso
Relaxed Lasso 0
Relaxed Lasso 0.5

Simulation Setup

$$\begin{aligned} Y &= \sum_{j=1}^{p} X_j \beta_j + \epsilon \\ X &\sim N_p(0, \Sigma) \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

- $p = 30$, sample size $n = 70$, and first $s = 5$ values of β are 1, the rest are zero.
- Σ is correlation matrix, with $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$, and $\rho = 0.35$
- σ^2 is chosen here to achieve desired SNR $= \text{Var}(X\beta)/\sigma^2$ of 0.71.

Simulation Setup

$$\begin{aligned} Y &= \sum_{j=1}^{p} X_j \beta_j + \epsilon \\ X &\sim N_p(0, \Sigma) \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

- $p = 30$, sample size $n = 70$, and first $s = 5$ values of β are 1, the rest are zero.
- Σ is correlation matrix, with $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$, and $\rho = 0.35$
- σ^2 is chosen here to achieve desired SNR $= \text{Var}(X\beta)/\sigma^2$ of 0.71.
- This is equivalent to a *percentage variance explained* (R^2) of 42%, since population PVE $= \text{SNR}/(1 + \text{SNR})$.

Simulation Setup

$$
\begin{aligned}
Y &= \sum_{j=1}^{p} X_j \beta_j + \epsilon \\
X &\sim N_p(0, \Sigma) \\
\epsilon &\sim N(0, \sigma^2)
\end{aligned}
$$

- $p = 30$, sample size $n = 70$, and first $s = 5$ values of β are 1, the rest are zero.
- Σ is correlation matrix, with $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$, and $\rho = 0.35$
- σ^2 is chosen here to achieve desired SNR $= \text{Var}(X\beta)/\sigma^2$ of 0.71.
- This is equivalent to a *percentage variance explained* (R^2) of 42%, since population PVE $= \text{SNR}/(1 + \text{SNR})$.
- Where appropriate we have a separate validation set of size n, and an infinite test set.

Degrees of Freedom

We can get some insight into the aggressiveness of the procedures by looking at their *degrees of freedom*.

Suppose $y_i = f(x_i) + \epsilon_i$, $i = 1, \ldots, n$, and assume $\mathrm{Var}(\epsilon_i) = \sigma^2$. Let $\hat{y}_i$ be the fitted value for observation i, after applying some regression method to the n pairs (x_i, y_i) (e.g. best-subset linear regression of size k, lasso with parameter λ)

$$\mathrm{df} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i)$$

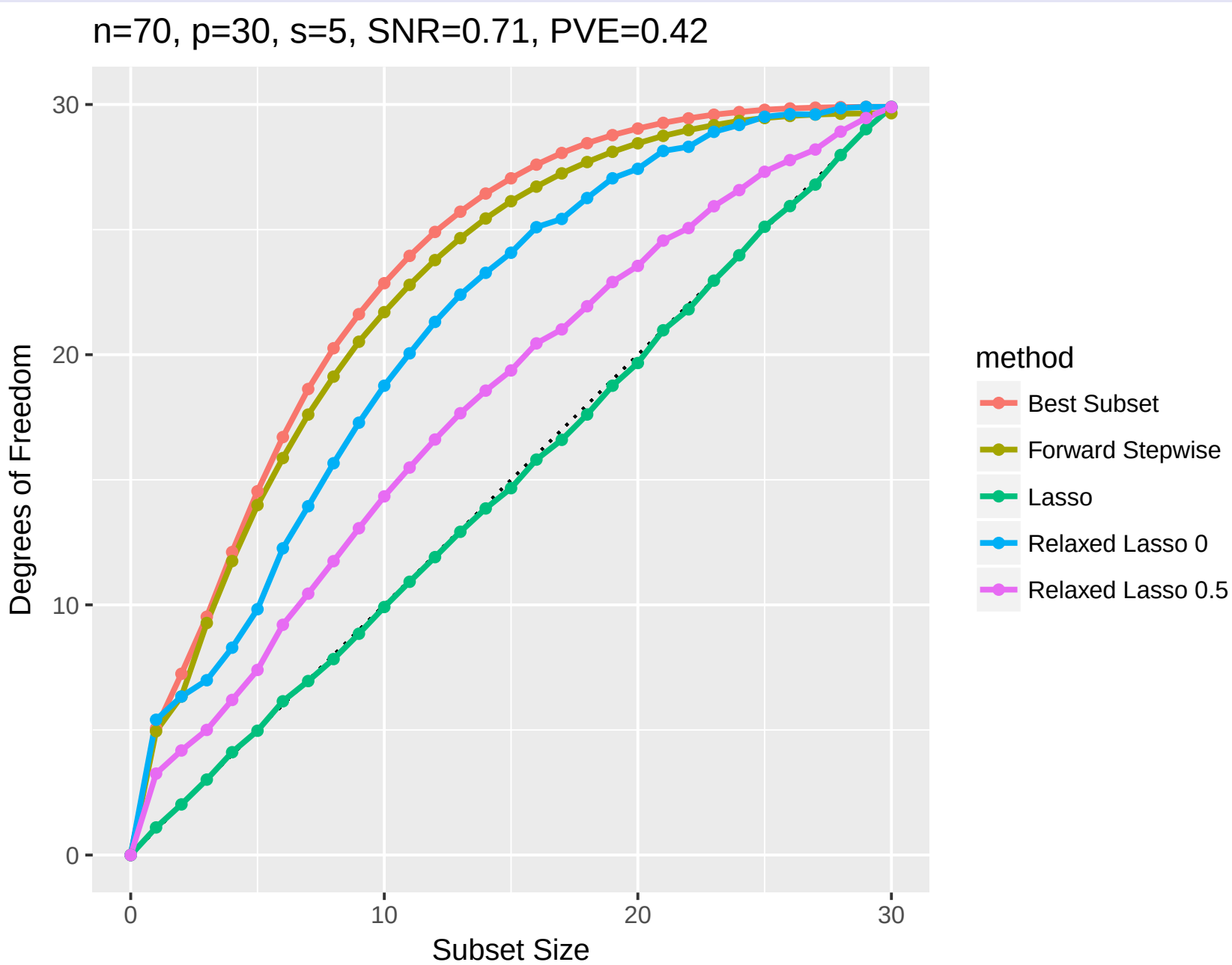
Degrees of Freedom

We can get some insight into the aggressiveness of the procedures by looking at their *degrees of freedom*.

Suppose $y_i = f(x_i) + \epsilon_i$, $i = 1, \ldots, n$, and assume $\text{Var}(\epsilon_i) = \sigma^2$. Let $\hat{y}_i$ be the fitted value for observation i, after applying some regression method to the n pairs (x_i, y_i) (e.g. best-subset linear regression of size k, lasso with parameter λ)

$$\text{df} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Cov}(y_i, \hat{y}_i)$$

These covariances are wrt the sampling distribution of the y_i. The more aggressive the procedure, the more it will overfit the training responses, and hence the higher the covariances and df.

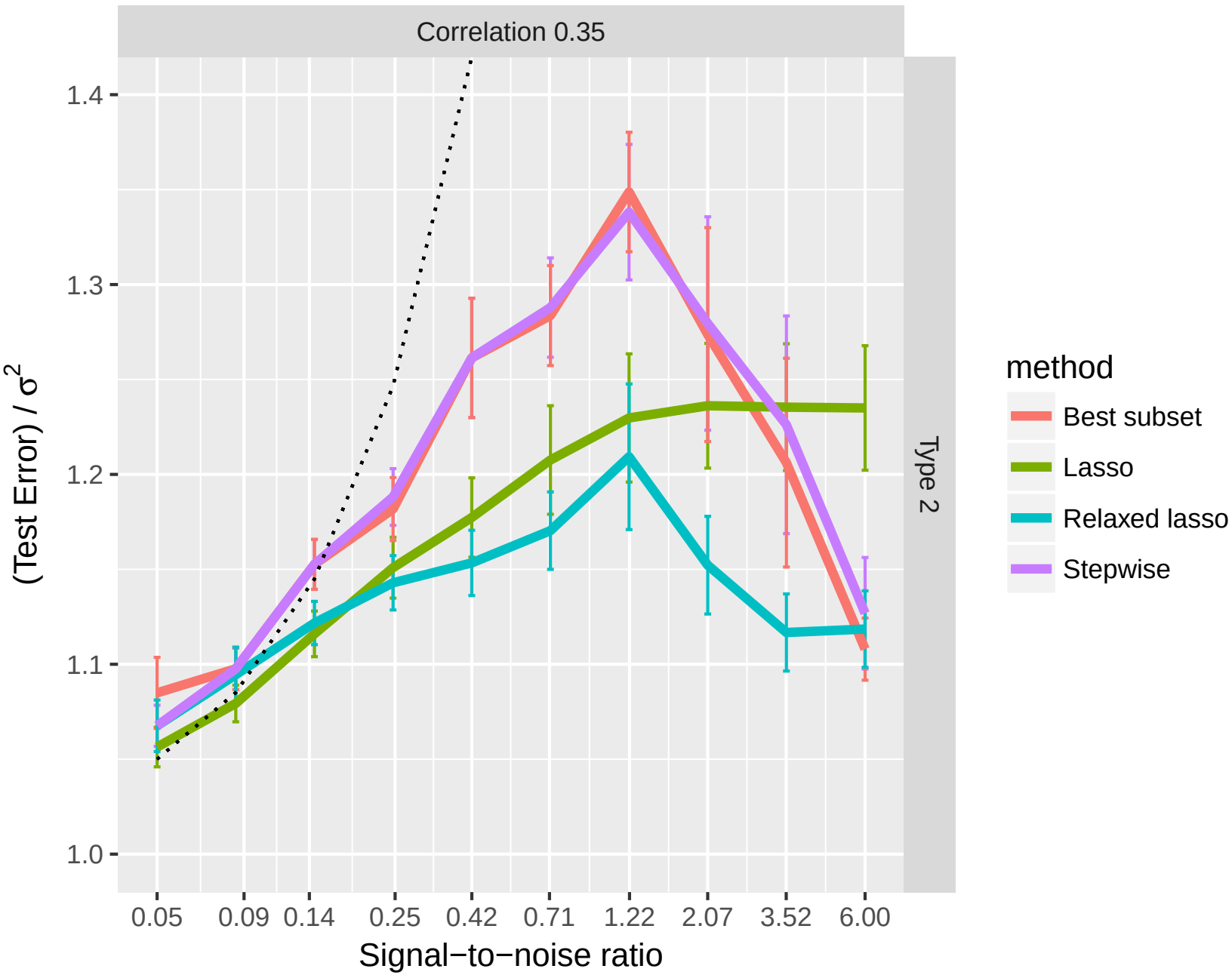

n=70, p=30, s=5, SNR=0.71, PVE=0.42
Degrees of Freedom
Subset Size
0
10
20
30
method
Best Subset
Forward Stepwise
Lasso
Relaxed Lasso 0
Relaxed Lasso 0.5

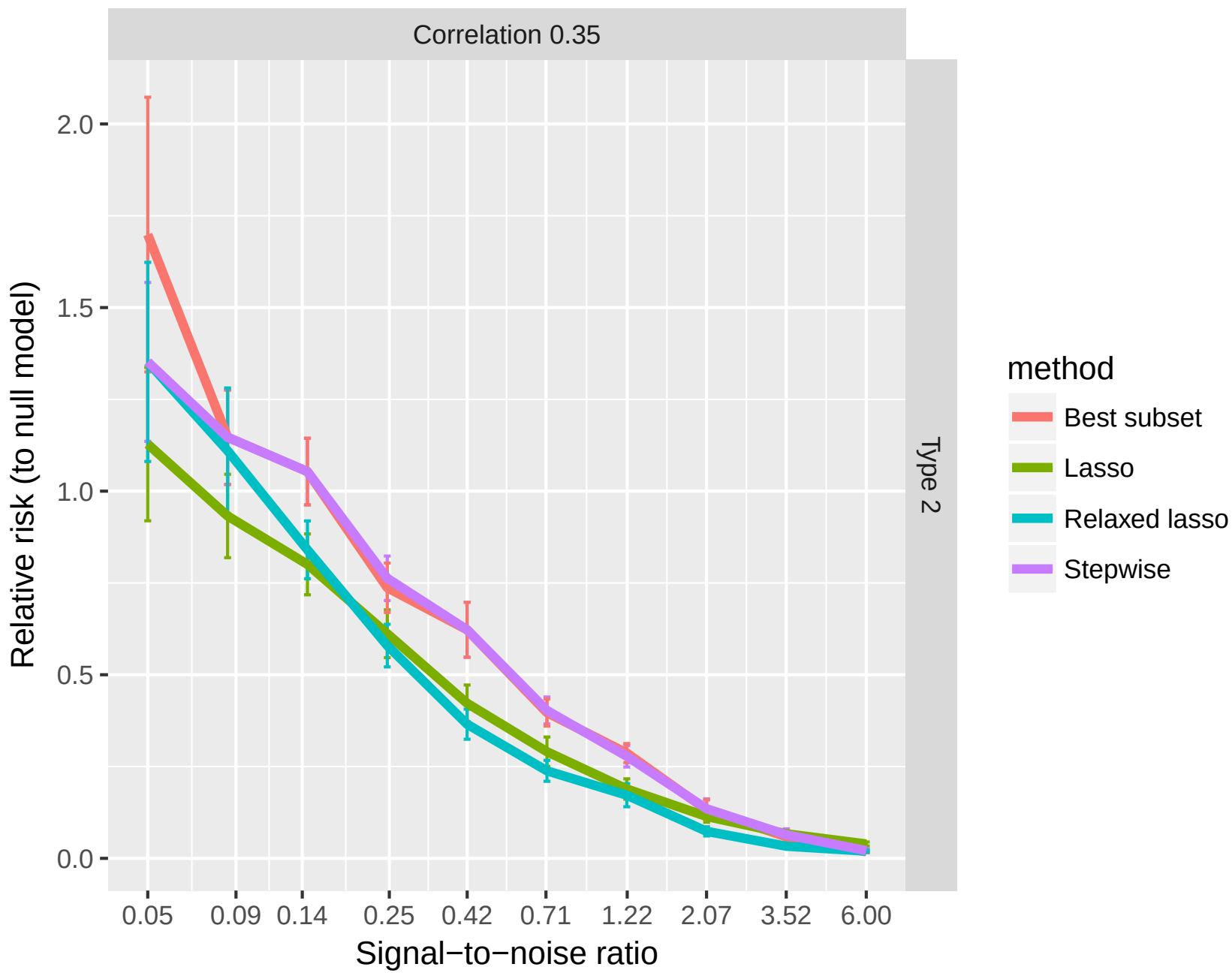
Notable features of previous plot

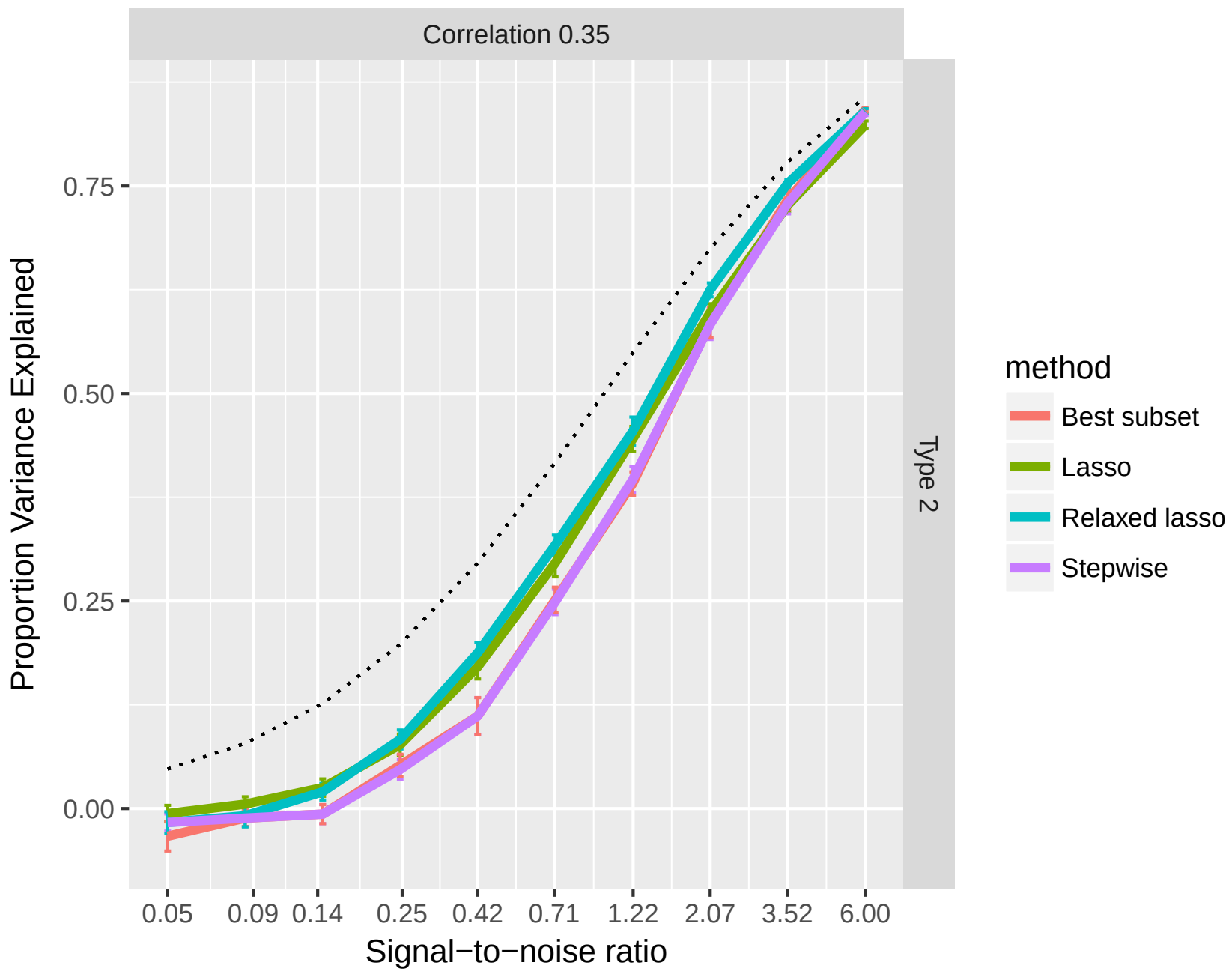
- Df for lasso is size of active set (Efron et al 2004, Zou et al 2007) — shrinkage offsets selection.
- Best-subset most aggressive, with forward stepwise just behind (in this example).
  Df can *exceed* p for BS and FS due to non-convexity (Janson et al 2005, Kaufman& Rosset 2014)
- Relaxed Lasso notably less aggressive, in particular $\hat{\beta}_\lambda^{LS}$ ($\gamma = 0$).

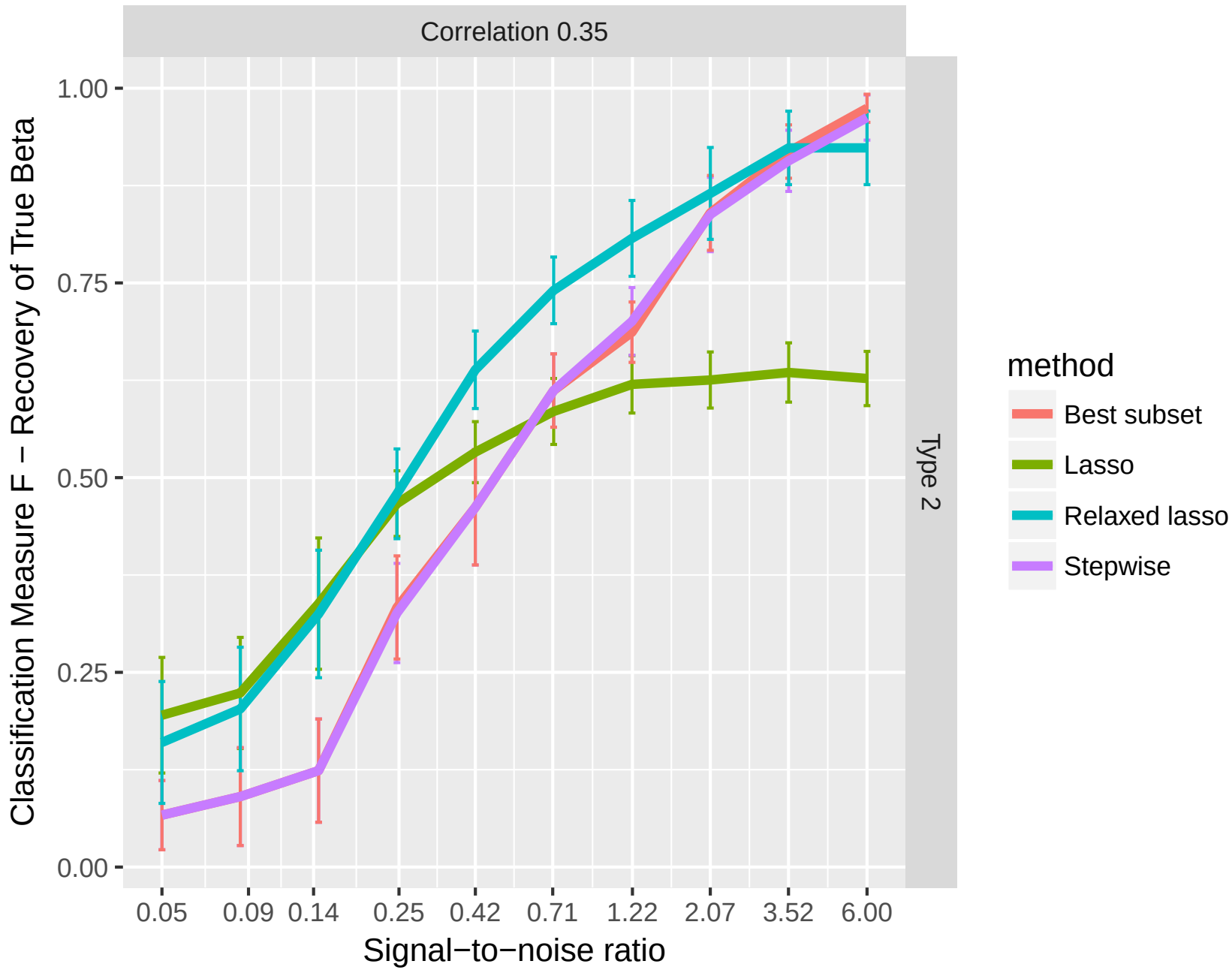
Next plots $\cdots$

- Show results over a range of SNRs
- Averaged over 10 simulations
- For each method, a validation set of same size as training set used to select the best model
- Reported errors are over infinite test set


n=70, p=30, s=5
Correlation 0.35
Type 2
(Test Error) / σ2
1.4
1.3
1.2
1.1
1.0
0.05
0.09
0.14
0.25
0.42
0.71
1.22
2.07
3.52
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=70, p=30, s=5
Correlation 0.35
Type 2
Relative risk (to null model)
2.0
1.5
1.0
0.5
0.0
0.05
0.09
0.14
0.25
0.42
0.71
1.22
2.07
3.52
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=70, p=30, s=5
Correlation 0.35
Type 2
Proportion Variance Explained
0.00
0.25
0.50
0.75
0.05
0.09
0.14
0.25
0.42
0.71
1.22
2.07
3.52
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=70, p=30, s=5
Correlation 0.35
Type 2
Classification Measure F – Recovery of True Beta
1.00
0.75
0.50
0.25
0.00
0.05
0.09
0.14
0.25
0.42
0.71
1.22
2.07
3.52
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise

Next plots ···

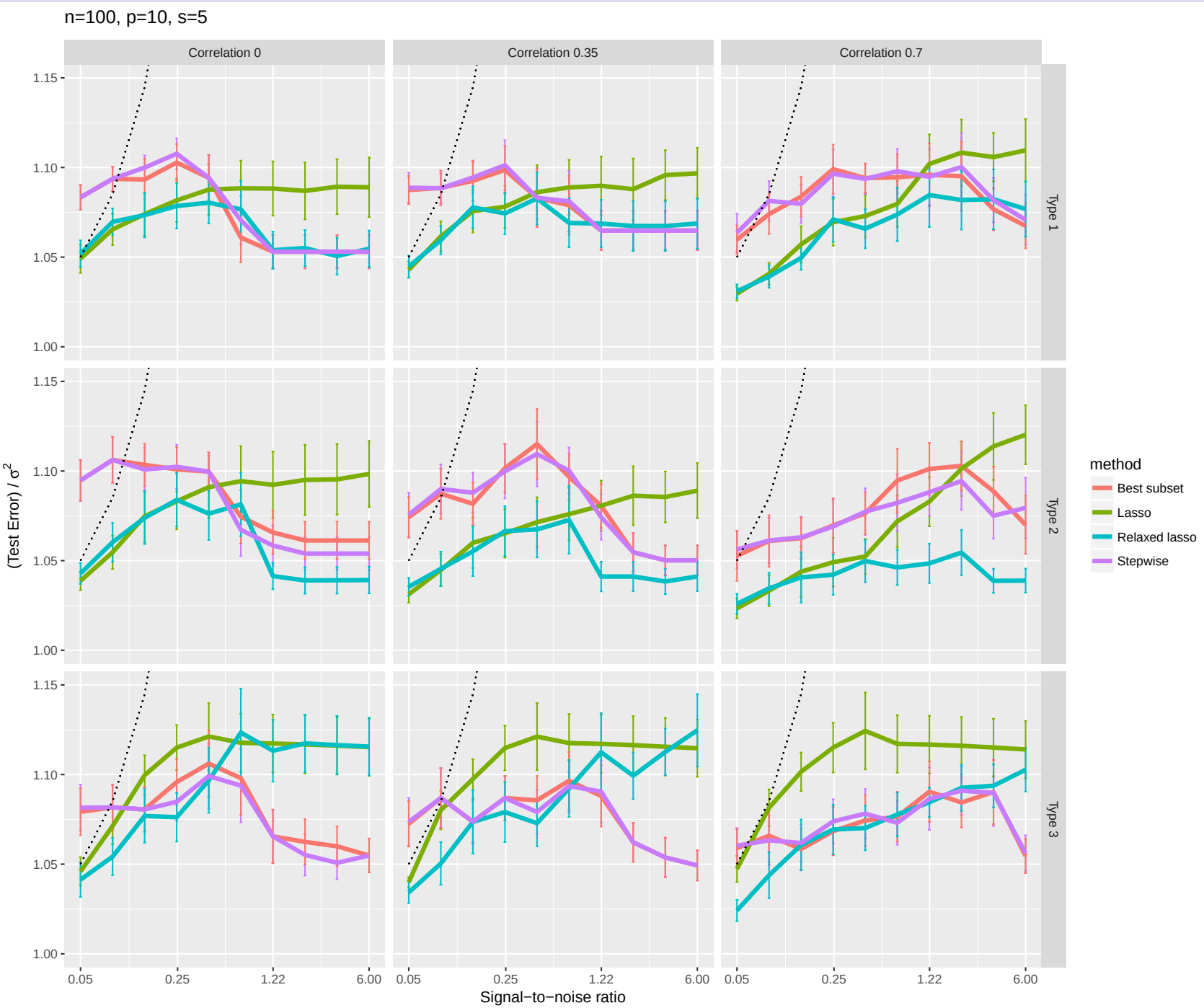
As before, but also

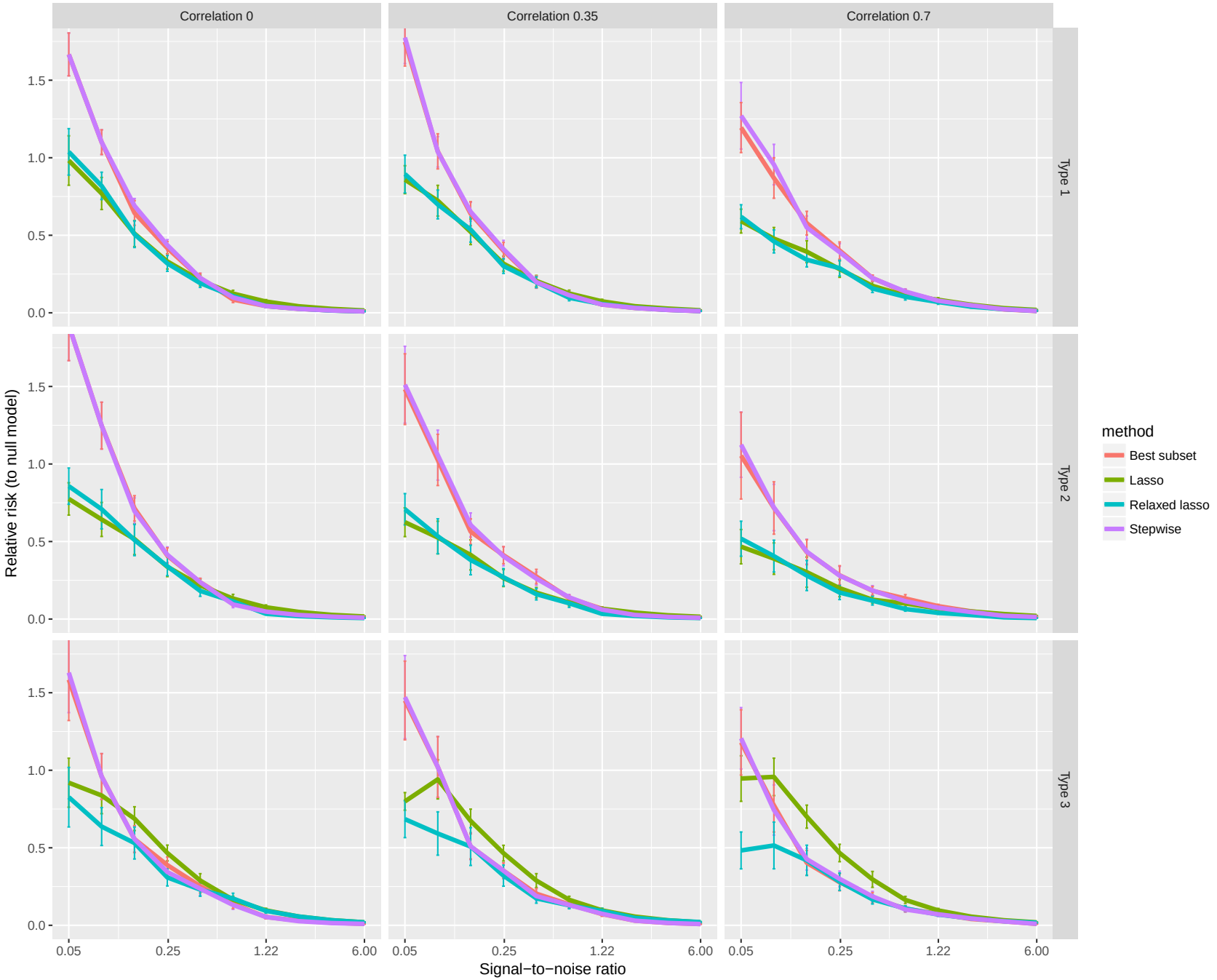
- different pairwise correlations between variables
- Different patterns of true coefficients
- Different problem sizes N, p.

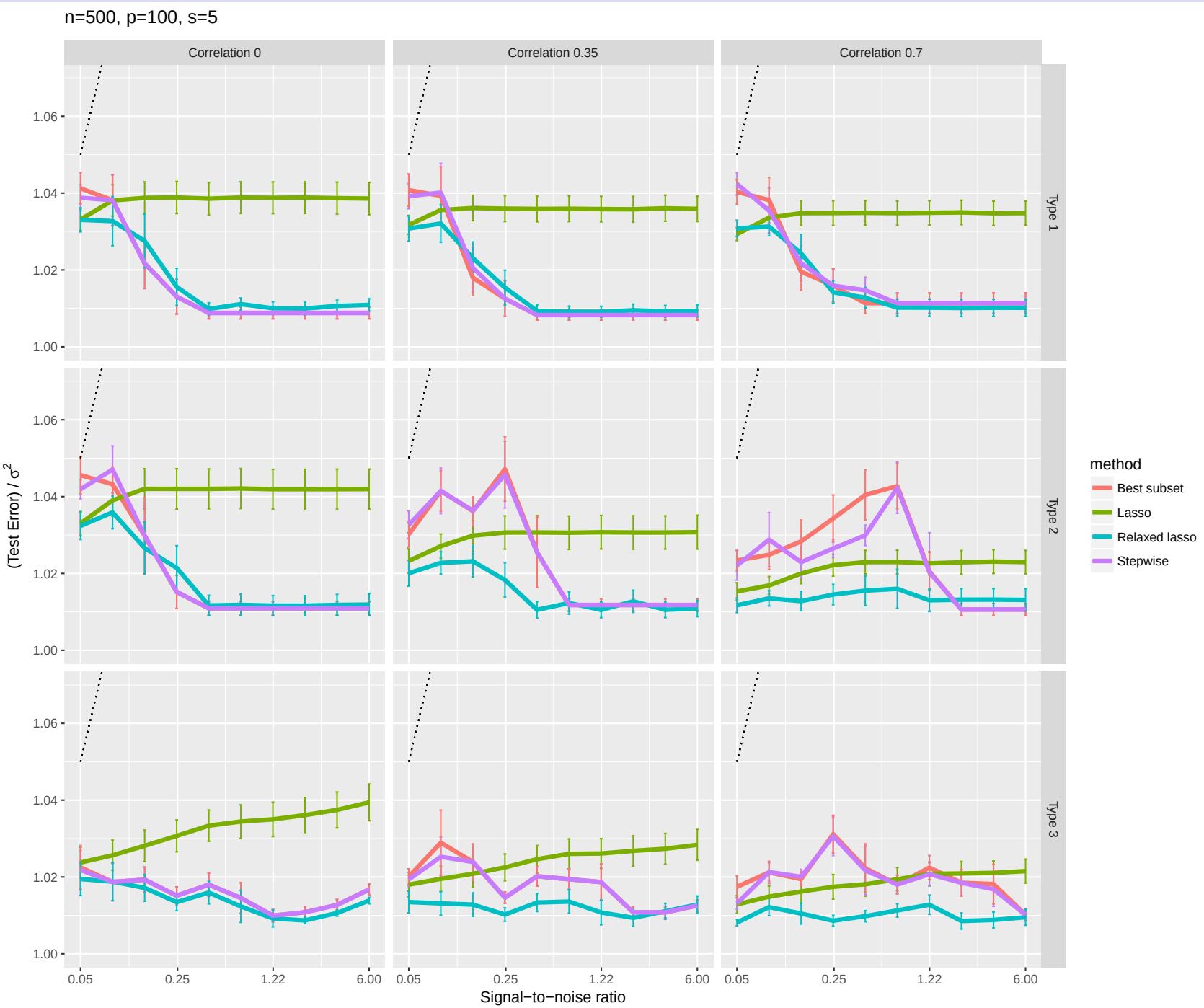
Timings

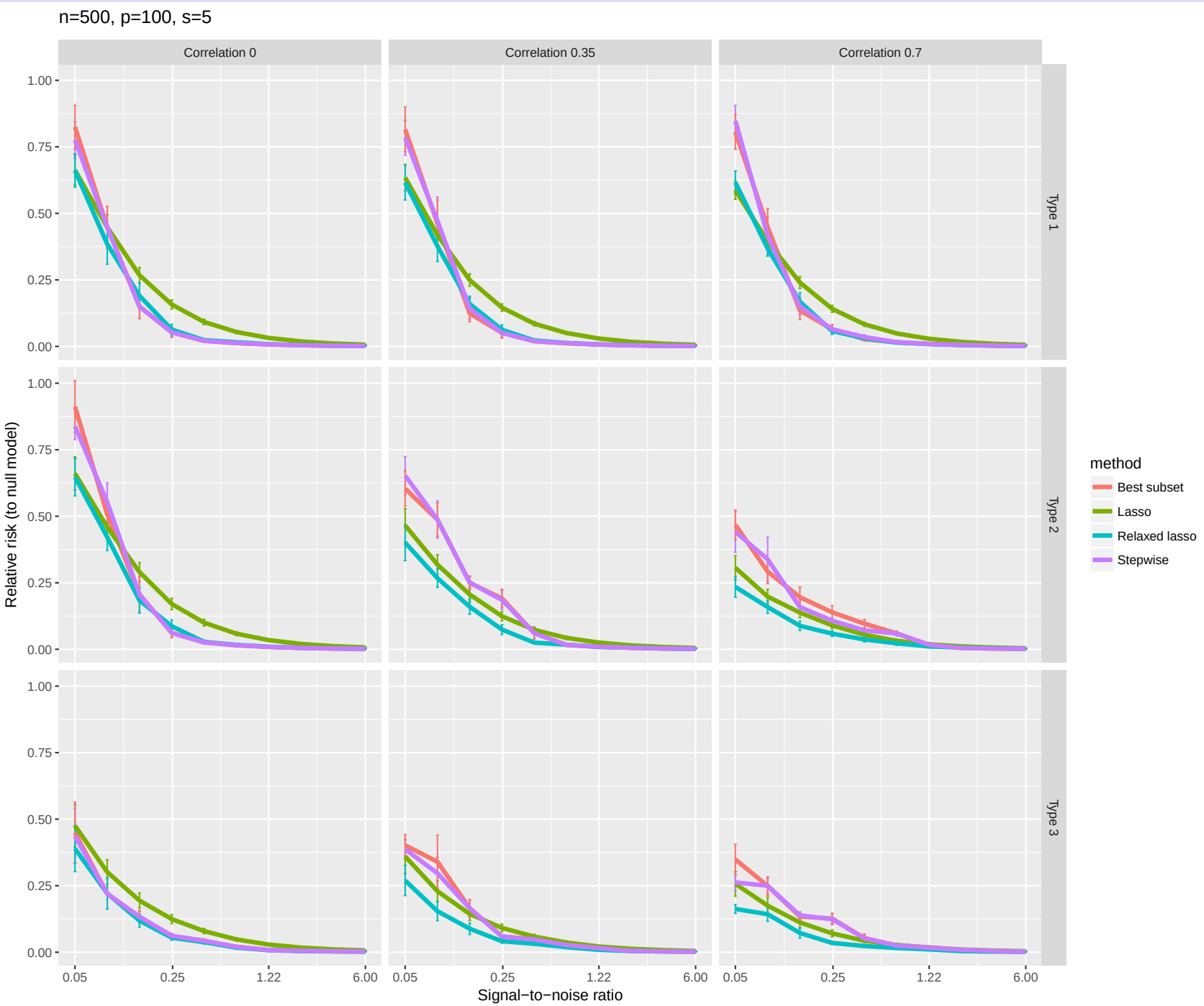
Setting		BS	FS	Lasso	R Lasso
low	n=100, p=10	3.43	0.006	0.002	0.002
medium	n=500, p=100	>**120 min**	0.818	0.009	0.009
high-5	n=50, p=1000	>**126 min**	0.137	0.011	0.011
high-10	n=100, p=1000	>**144 min**	0.277	0.019	0.021

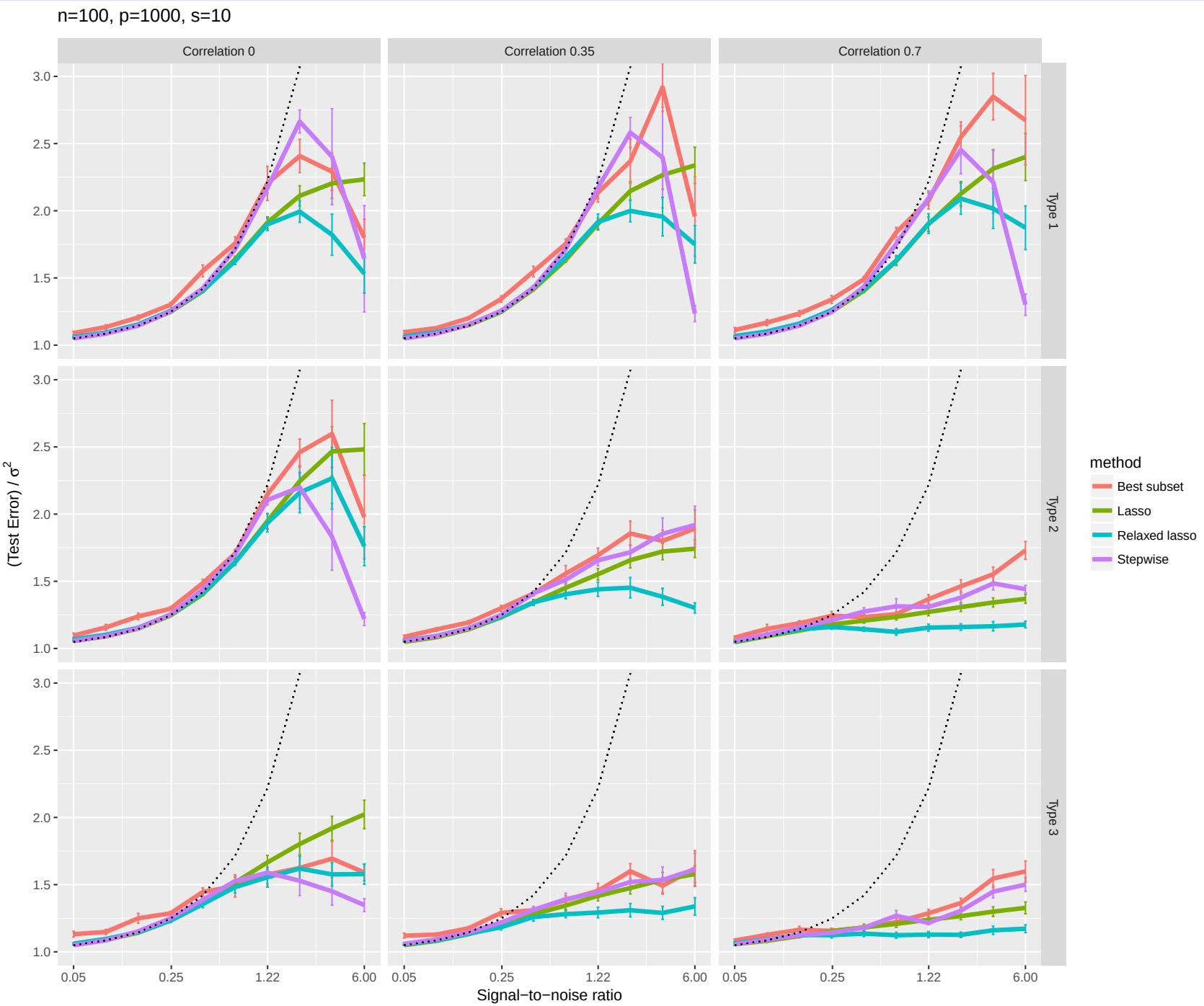
Average time in seconds to compute one path of solutions for each method, on a Linux cluster

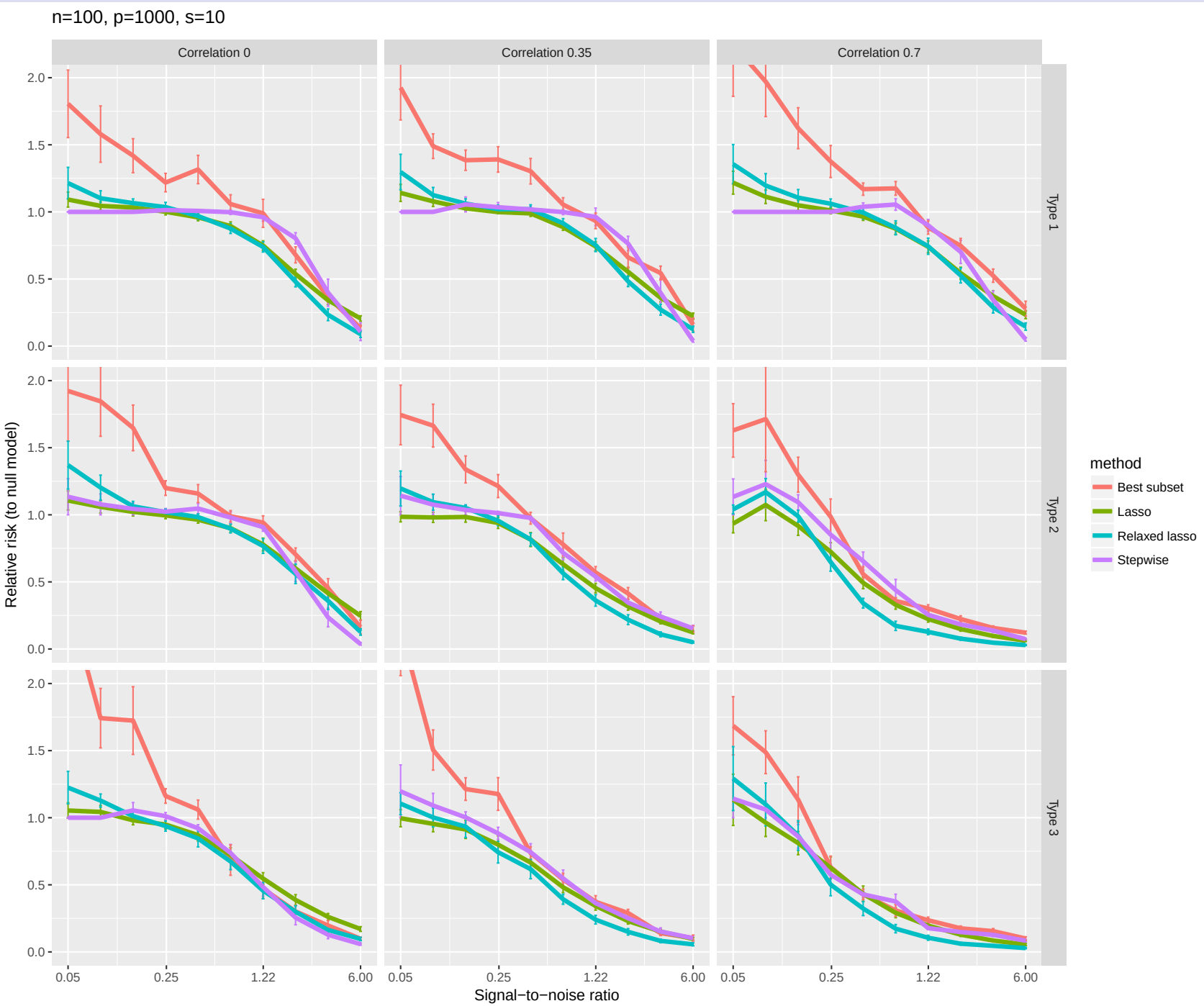

n=100, p=10, s=5
Correlation 0
Correlation 0.35
Correlation 0.7
Type 1
Type 2
Type 3
(Test Error) / σ²
1.15
1.10
1.05
1.00
0.05
0.25
1.22
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=100, p=10, s=5
Correlation 0
Correlation 0.35
Correlation 0.7
Type 1
Type 2
Type 3
Relative risk (to null model)
0.0
0.5
1.0
1.5
0.05
0.25
1.22
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=500, p=100, s=5
Correlation 0
Correlation 0.35
Correlation 0.7
Type 1
Type 2
Type 3
(Test Error) / σ²
1.00
1.02
1.04
1.06
0.05
0.25
1.22
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=500, p=100, s=5
Correlation 0
Correlation 0.35
Correlation 0.7
Type 1
Type 2
Type 3
Relative risk (to null model)
1.00
0.75
0.50
0.25
0.00
0.05
0.25
1.22
6.00
Signal−to−noise ratio
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=100, p=1000, s=10
Correlation 0
Correlation 0.35
Correlation 0.7
Type 1
Type 2
Type 3
(Test Error) / σ²
Signal−to−noise ratio
3.0
2.5
2.0
1.5
1.0
0.05
0.25
1.22
6.00
method
Best subset
Lasso
Relaxed lasso
Stepwise


n=100, p=1000, s=10
Correlation 0
Correlation 0.35
Correlation 0.7
Type 1
Type 2
Type 3
Relative risk (to null model)
Signal−to−noise ratio
0.05
0.25
1.22
6.00
0.0
0.5
1.0
1.5
2.0
method
Best subset
Lasso
Relaxed lasso
Stepwise

Outline and Summary

We consider linear regression models $\eta(X) = X^T\beta$ with potentially very large numbers of variables, and methods for selecting an informative subset.

- Revisit two baby boomers (best-subset selection and forward-stepwise selection), one millennial (lasso) and a newborn (relaxed lasso).
- Simulation study to evaluate them all over a wide range of settings.

Conclusions:

- forward stepwise very close to best subset, but much faster.
- relaxed lasso overall winner, and fastest by far.
- In wide-data settings, and low SNR, lasso can beat best subset and forward stepwise.

Paper: https://arxiv.org/abs/1707.08692

R package: https://github.com/ryantibs/best-subset/

Outline and Summary

We consider linear regression models $\eta(X) = X^T\beta$ with potentially very large numbers of variables, and methods for selecting an informative subset.

- Revisit two baby boomers (best-subset selection and forward-stepwise selection), one millennial (lasso) and a newborn (relaxed lasso).
- Simulation study to evaluate them all over a wide range of settings.

Conclusions:

- forward stepwise very close to best subset, but much faster.
- relaxed lasso overall winner, and fastest by far.
- In wide-data settings, and low SNR, lasso can beat best subset and forward stepwise.

Paper: https://arxiv.org/abs/1707.08692

R package: https://github.com/ryantibs/best-subset/

Thank you for attending!