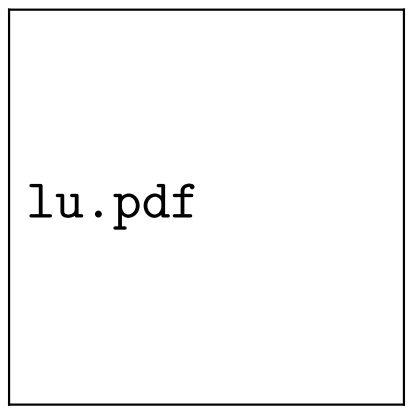
SLOPE

Presentation for the ML group at LUSEM

Johan Larsson

Department of Statistics, Lund University

May 27, 2020



Overview

Introducing SLOPE

Motivation and setting

setting

want to apply a **generalized linear model**[1] to a set of predictors $X \in \mathbb{R}^{n \times p}$ and outcome $y \in \mathbb{R}^n$

leads to finding the optimal solution ($\hat{\beta}$) to the problem

$$\text{minimize} \quad g(\beta; X, y),$$

for instance $g(\beta; X, y) \coloneqq \frac{1}{2}\|y - X\beta\|_2^2$ for OLS.

[1] least-squares, logistic, multinomial, or Poisson regression for instance

Motivation and setting

setting

want to apply a **generalized linear model**[1] to a set of predictors $X \in \mathbb{R}^{n \times p}$ and outcome $y \in \mathbb{R}^n$

leads to finding the optimal solution ($\hat{\beta}$) to the problem

$$\text{minimize} \quad g(\beta; X, y),$$

for instance $g(\beta; X, y) := \frac{1}{2}\|y - X\beta\|_2^2$ for OLS.

problem

- $p \gg n$
- believe **real** β is sparse: few signals, much noise
- want to avoid overfitting
- want efficiency

[1] least-squares, logistic, multinomial, or Poisson regression for instance

Regularization

idea
introduce regularization by constraining the problem, i.e. solve

$$\begin{aligned} \text{minimize} \quad & g(\beta; X, y) \\ \text{subject to} \quad & h(\beta) \leq t \end{aligned}$$

and choose $h(\beta)$ such that the resulting model is **sparse** by shrinking some elements in β to be *exactly* zero.

typical choices for $h(\beta)$

- ℓ_0 **norm** best subset selection
- ℓ_1 **norm** the lasso

Problems with standard methods

best subset selection

- not convex and therefore intractable for large p
- no shrinkage

lasso

- unpredictable model selection for highly correlated predictors
- can only select n coefficients

SLOPE

bogdan2015 introduced SLOPE (Sorted L-One Penalized Estimation), which solves the problem

$$\begin{aligned} &\text{minimize} && g(\beta; X, y) \\ &\text{subject to} && \sum_{i=1}^{n} \lambda_i |\beta|_{(i)} \leq t, \end{aligned}$$

where $\sum_{i=1}^{p} \lambda_i |\beta|_{(i)}$ is the **sorted** ℓ_1 **norm**, for which

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

and

$$|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}.$$

A step back: least squares

for simplicity, let's assume

$$g(\beta; X, y) = \frac{1}{2}\|X\beta - y\|_2^2,$$

i.e., we are solving (unregularized) OLS—solution available analytically through normal equations.

Constraint region for the sorted ℓ_1 norm

$\sum_{i=1}^{p} \lambda_i |\beta|_{(i)} \leq t$ defines a constraint region centered at $\mathbf{0}$

Sorted ℓ_1-regularized least squares

let's introduce regularization via the sorted ℓ_1 norm: solution $\hat{\beta}$ has to lie inside constraint region defined by

$$\sum_{i=1}^{p} \lambda_i |\beta|_{(i)} \leq t.$$

Shapes of the sorted ℓ_1 norm

choice of λ dictates shape of the constraint region

(a) $\lambda_1 = \lambda_2$ **(b)** $\lambda_1 > \lambda_2 > 0$ **(c)** $\lambda_1 > \lambda_2 = 0$

Equivalent formulations

We've so far defined SLOPE as the **constrained** optimization problem

$$\begin{aligned} &\text{minimize} && g(\beta; X) \\ &\text{subject to} && \sum_{i=1}^{p} \lambda_i |\beta|_{(i)} \leq t, \end{aligned}$$

but, with a bit of notational abuse[2], this is equivalent to the **unconstrained** problem

$$\text{minimize} \quad g(\beta; \lambda) + J(\beta; \lambda),$$

with $J(\beta; \lambda) = \sum_{i=1}^{p} \lambda_i |\beta|_{(i)}$.

[2] Redefining λ.

Clustering Property

The sorted ℓ_1 norm induces clustering: setting absolute values of coefficients to same value

Consider the two-dimensional case $y = x_1\beta_1 + x_2\beta_2 + \varepsilon$.

lasso	SLOPE
$\lambda\lvert\beta_1\rvert + \lambda\lvert\beta_2\rvert$	$\lambda_1\lvert\beta\rvert_{(1)} + \lambda_2\lvert\beta\rvert_{(2)}$

and assume x_1 and x_2 are perfectly correlated, then

- the lasso will force one of the coefficients to zero, whilst
- SLOPE (provided $\lambda_1 > \lambda_2 \geq 0$) will set them to the same value.

Selection of Regularization Sequence

Choice of λ

- λ is p-dimensional, which means there is **considerable** freedom in choosing it
- to make this problem manageable, we therefore assume that each λ_i is a function of p, n, i, and some parameters

Choice of λ: BH

Inspiration for SLOPE comes from the desire to control of false discovery rate (FDR) in a regression setting, i.e. test the hypotheses

$$H_{0,j} : \beta_j = 0 \qquad H_{1,j} : \beta_j \neq 0.$$

Benjamini–Hochberg (BH) procedure

Sort $\hat{\beta}$ in non-decreasing order according to its absolute values and reject all hypothesis $H_{(i)}$ for which $i \leq i_{\text{BH}}$, where

$$i_{\text{BH}} = \max \left\{ i \mid |\hat{\beta}|_{(i)} \geq \sigma \Phi^{-1} \left(1 - \frac{iq}{2p} \right) \right\},$$

where Φ^{-1} is the probit function.

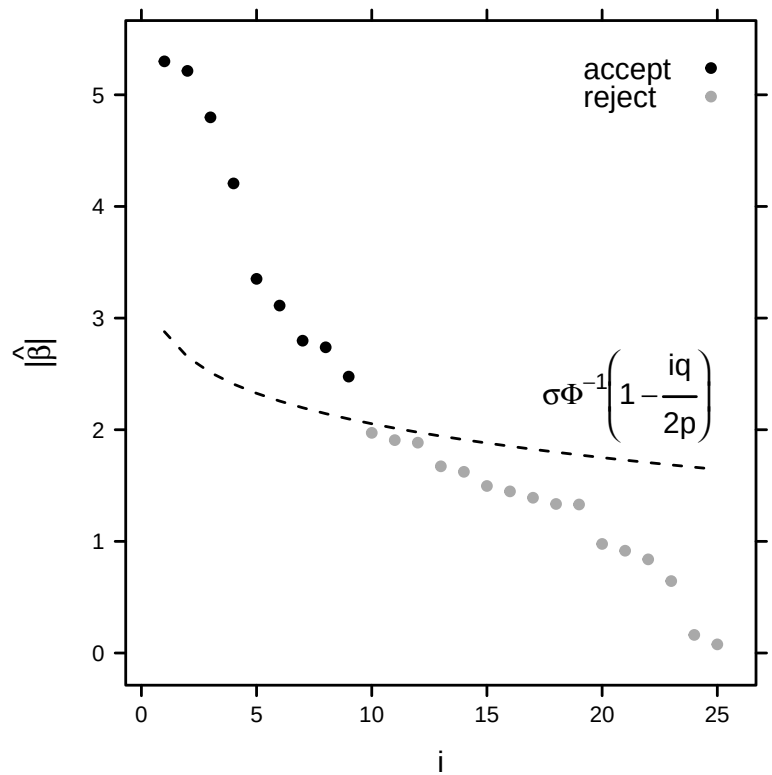


Figure 2: BH correction.

BH λ method for SLOPE

The BH method for choosing the λ sequence in SLOPE sets

$$\lambda_i = \sigma \Phi^{-1}\left(1 - \frac{qi}{2p}\right).$$

Very similar to procedure from last slide.

It turns out that this sequence promises FDR control in **orthogonal** settings (**bogdan2015**), namely

$$\text{FDR} \leq \frac{qp_0}{p},$$

where p_0 is the number of true null hypotheses. In other words, q sets an upper bound on FDR.

FDR, Power, and Prediction Error

Results from debiased[3] SLOPE and lasso and cross-validated lasso show that SLOPE controls FDR as promised and has better predictive performance.

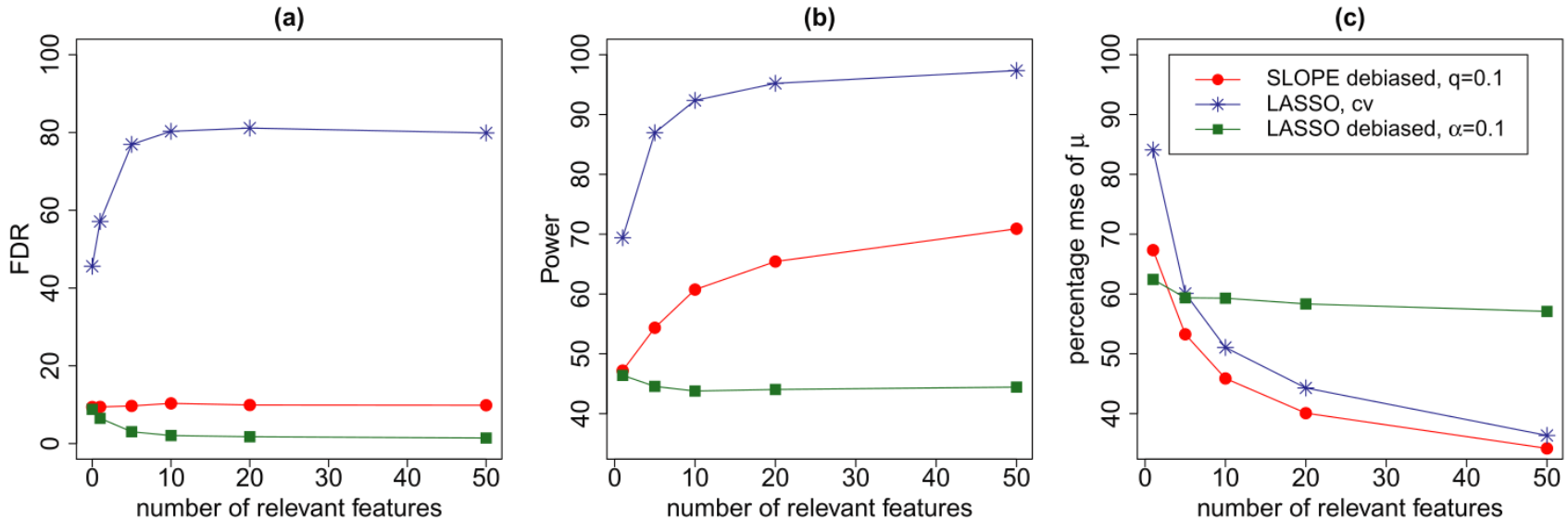


Figure 3: FDR, power, and mean-squared error (MSE) for lasso SLOPE using BH sequence. Predictors are i.i.d. generated from a normal distribution.

[3]Select support using SLOPE or lasso and estimate coefficients using standard OLS.

FDR in Gaussian design

When design is **not** orthogonal, SLOPE with the BH sequence loses control of FDR as the number of relevant predictors increase.

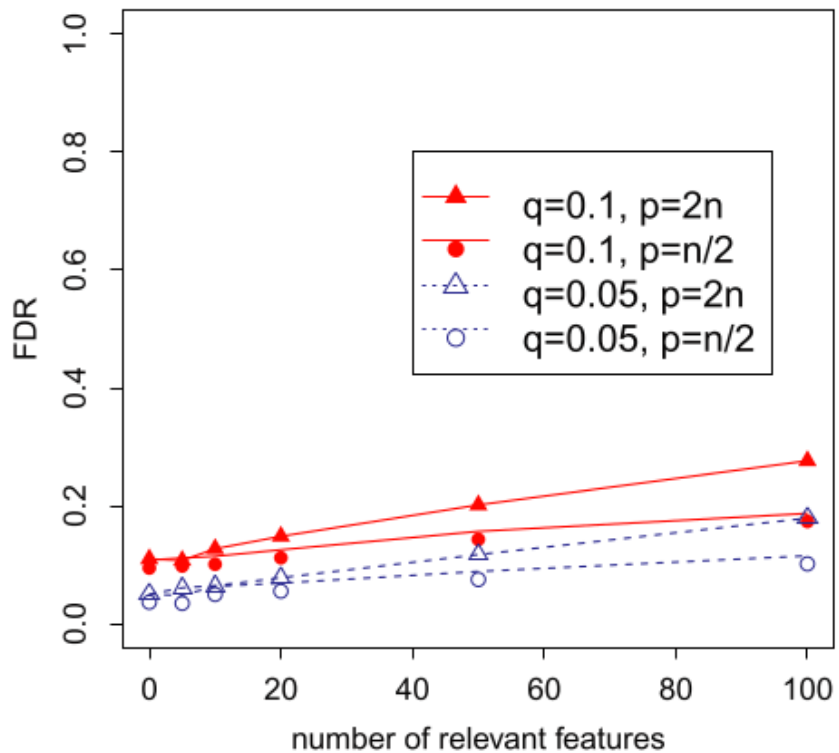


Figure 4: FDR control for SLOPE with BH sequence for Gaussian i.i.d. design. $p = 2n = 10000$.

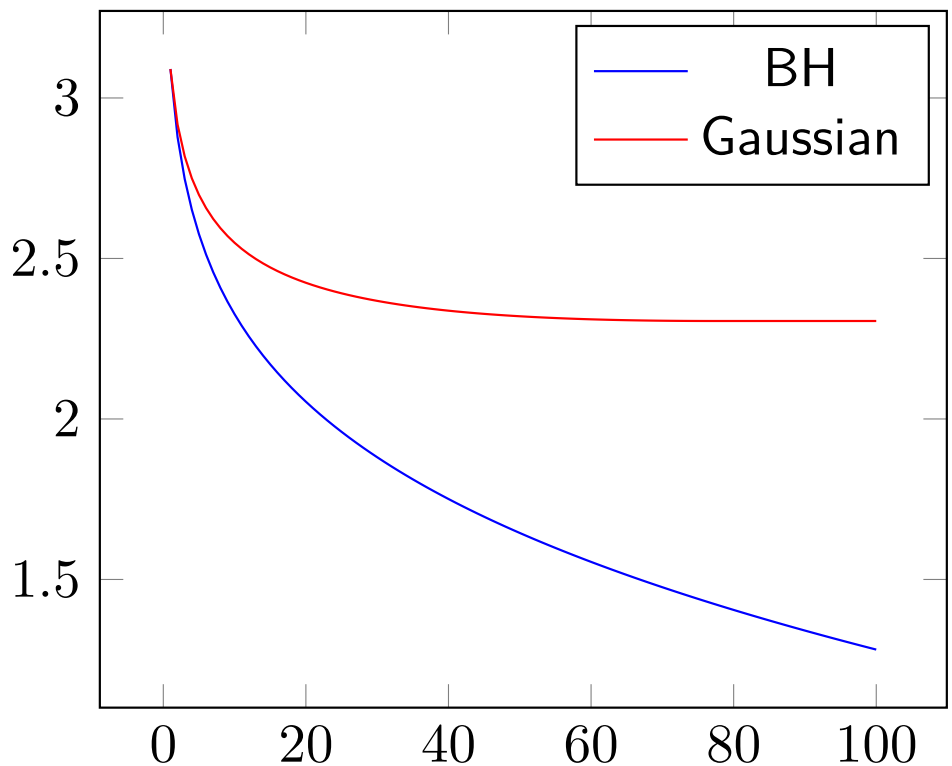
Choice of λ: Gaussian Sequence

In light of the problem with FDR control for non-orthogonal settings, **bogdan2015** consider also the **Gaussian** sequence that sets

$$\lambda_i = \min\left(\lambda_{i-1}, \lambda_i^{\mathrm{BH}} \sqrt{1 + \frac{1}{n-i}\sum_{j<i}\lambda_i^2}\right),$$

which flattens out the sequence based on n/p fraction.

Note, however, that for $p \gg n$, this sequence reduces to the lasso.



FDR Control with Gaussian Sequence

Gaussian sequence does a better job of controlling FDR in the Gaussian (but i.i.d.) case.

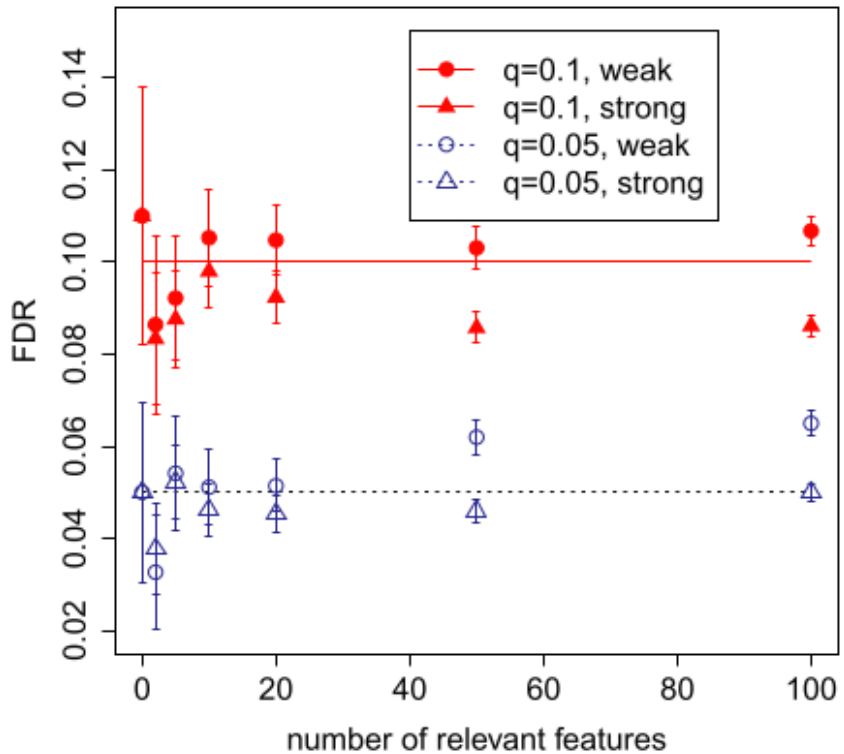


Figure 5: FDR control for SLOPE with Gaussian sequence for Gaussian i.i.d. design. $p = 2n = 10000$.

Parameter Selection for Regularization Sequence

Selecting parameters for λ sequence

We've reduced the problem of specifying the entire λ sequence to specifying parameters σ and q but **how do we choose these?**

three options

1. we know σ and can assume that predictors are not very correlated
2. use (**bogdan2015**) and estimate σ using SLOPE to select support and OLS estimates to approximate σ
3. cross-validation

Choice depends on situation. **1** is of course usually intractable, **2** works well with low correlation and $n > p$, **3** is attractive for prediction properties (but we lose FDR control).

Cross-validation and the regularization path

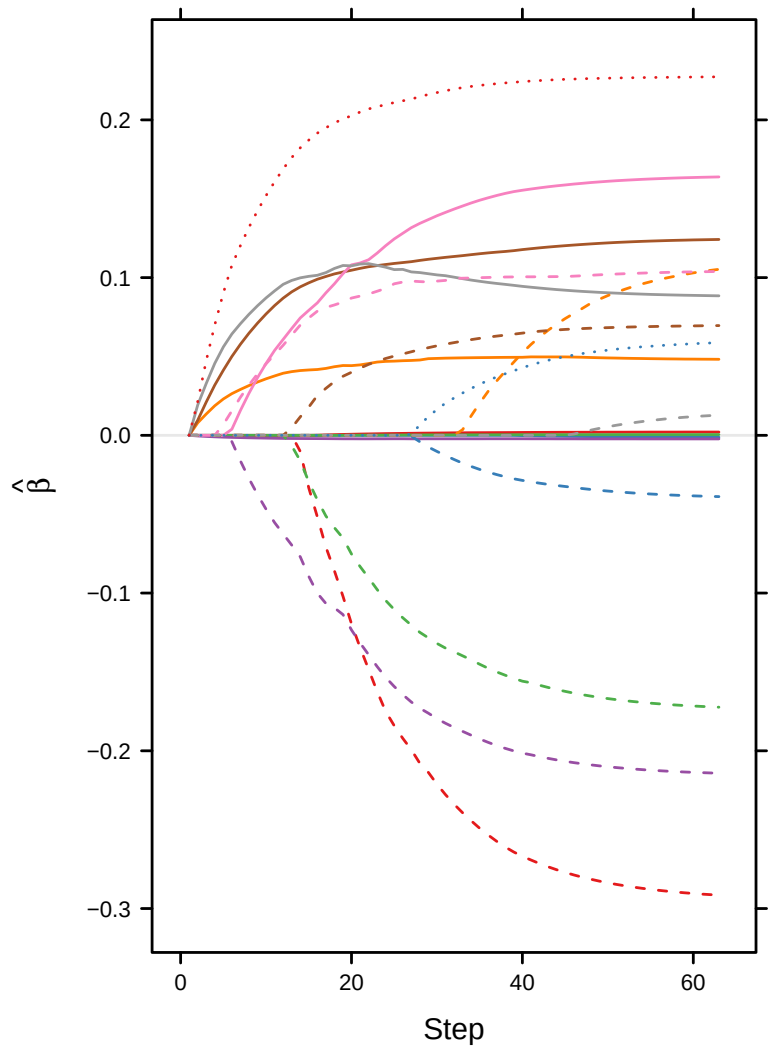
If we are interested in cross-validation to select σ and q, we will generally want to construct a **regularization path** of λ sequences,

$$\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(m)}.$$

In addition, we want $\lambda^{(1)}$ and $\lambda^{(m)}$ to yield the **intercept-only** model and **almost-saturated** model respectively.

tuning parameters

σ (scale), q (shape)



SLOPE and lasso comparisons

SLOPE and lasso regularization paths

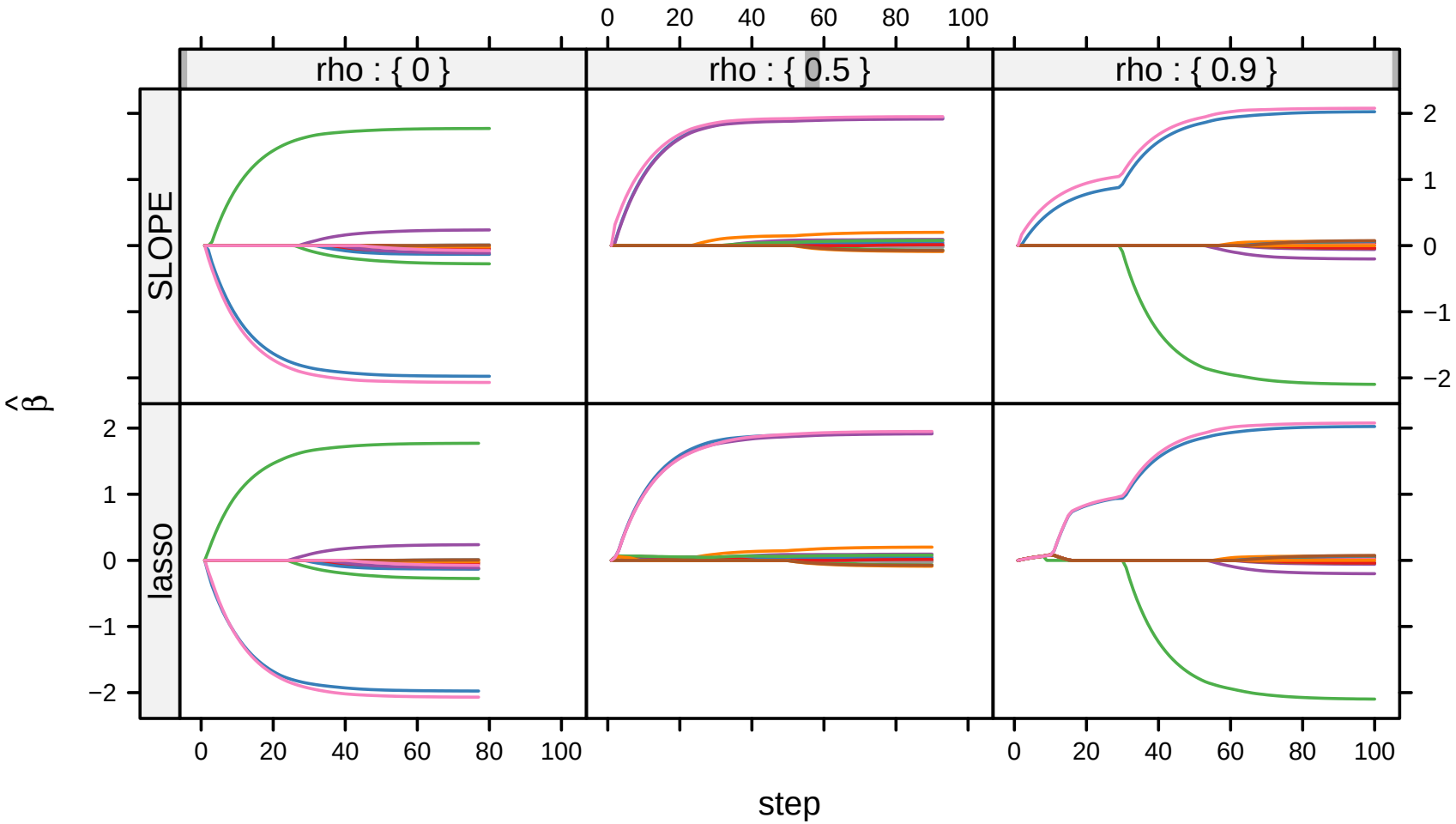


Figure 6: Comparison of SLOPE and lasso paths for correlated and uncorrelated data.

Grouping effects

Differences between lasso and SLOPE are clear in block-correlated covariance matrices.

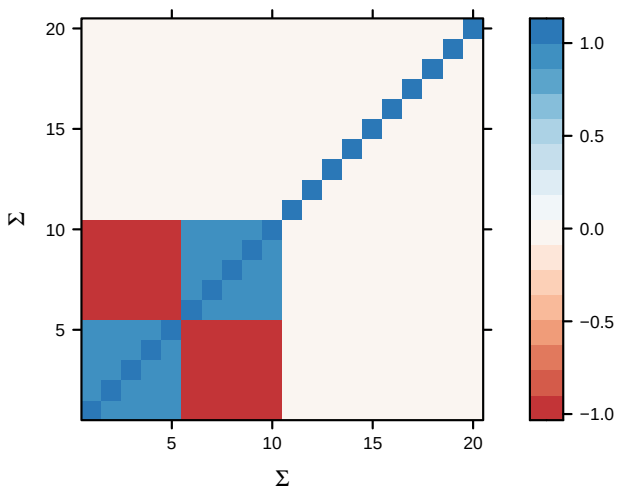


Figure 7: Correlation structure.

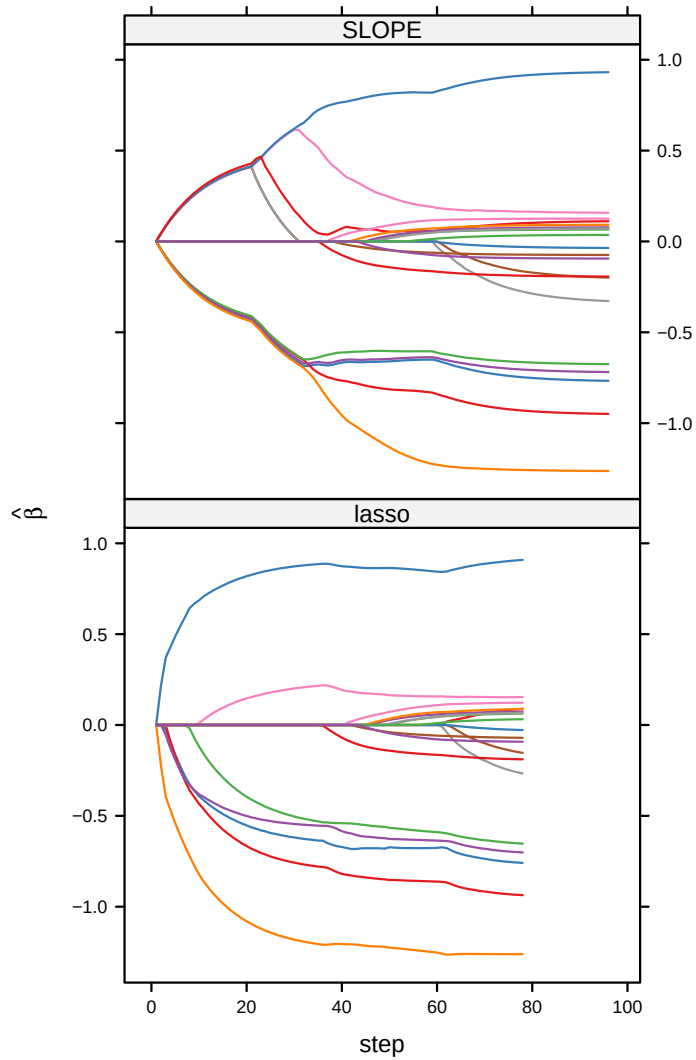


Figure 8: Regularization paths.

Performance and screening rules

Sparsity-enforcing methods, such as lasso and SLOPE, can be **very efficient**, particularly when $p \gg n$ due to **screening rules**

Screening rules are based on the idea that many solutions along the path will be sparse and that we can estimate this and **discard** predictors before fitting the model.

We have developed a screening rule for SLOPE (**larsson2020b**)

SLOPE extensions, implementation, and discussion

Extensions of SLOPE

Several extensions of SLOPE exists, all analagous to popular lasso derivatives.

Adaptive Bayesian SLOPE (ABSLOPE)

Semi-Bayesian approach that adapatively reweights penalties. See **jiang2019**.

Group SLOPE

Sorted ℓ_1 norm regularization on group level ℓ_2 norm regularization on individual level. See **brzyski2018**.

Software

Most software is still under development, not quite mature.

implementations

- SLOPE: https://CRAN.R-project.org/package=SLOPE
- Group SLOPE: https://CRAN.R-project.org/package=grpSLOPE
- ABSLOPE: https://github.com/wjiang94/ABSLOPE

Topics for discussion

- How interested are you in
  - FDR control?
  - model selection properties?
  - prediction properties?
- SLOPE penalizes stronger signals more than weaker ones—is this reasonable?
- choice of λ sequence has not been thoroughly investigated—any ideas?

References I