Look-Ahead Screening Rules for the Lasso

EYSM 2021

Johan Larsson

Department of Statistics, Lund University

June 29, 2021

LUND
UNIVERSITY

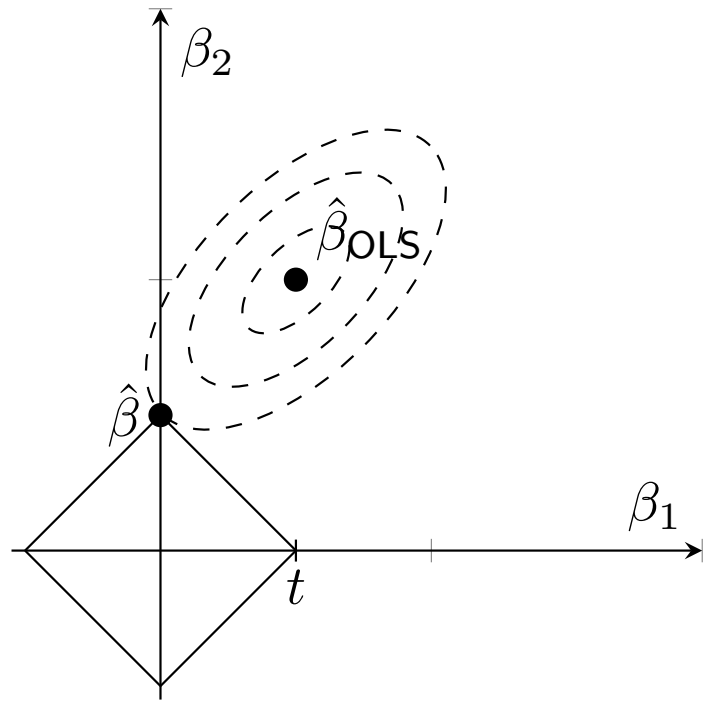
The Lasso

The lasso (Tibshirani 1996) is a type of penalized regression, represented by the following convex optimization problem:

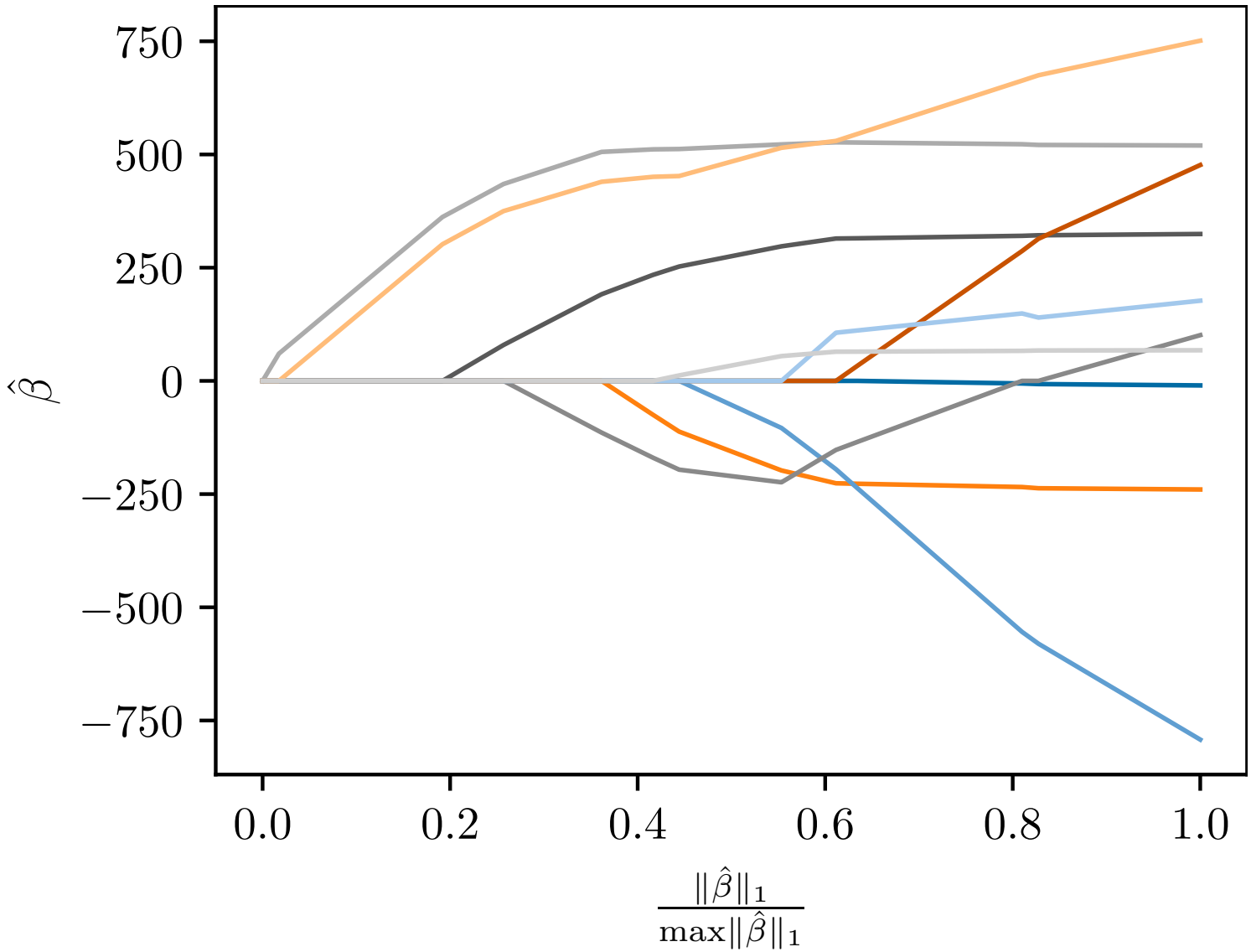
$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ P(\beta; \lambda) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1. \right\}$$

λ is a hyper-parameter that controls the level of **penalization**.

$\hat{\beta}_\lambda$ is the solution to this problem for a given λ.



The Lasso Path



Predictor Screening Rules

motivation

Many of the solution vectors, $\hat{\beta}$, along the regularization path will be **sparse**, which means some predictors (columns) in X will be **inactive**, especially if $p \gg n$.

Predictor Screening Rules

motivation

Many of the solution vectors, $\hat{\beta}$, along the regularization path will be **sparse**, which means some predictors (columns) in X will be **inactive**, especially if $p \gg n$.

basic idea

If we could, based on a relatively **cheap** test, determine which predictors will be inactive before fitting the model, we could solve the problem **much faster**.

Predictor Screening Rules

motivation
Many of the solution vectors, $\hat{\beta}$, along the regularization path will be **sparse**, which means some predictors (columns) in X will be **inactive**, especially if $p \gg n$.

basic idea
If we could, based on a relatively **cheap** test, determine which predictors will be inactive before fitting the model, we could solve the problem **much faster**.

it turns out we can!
using screening rules

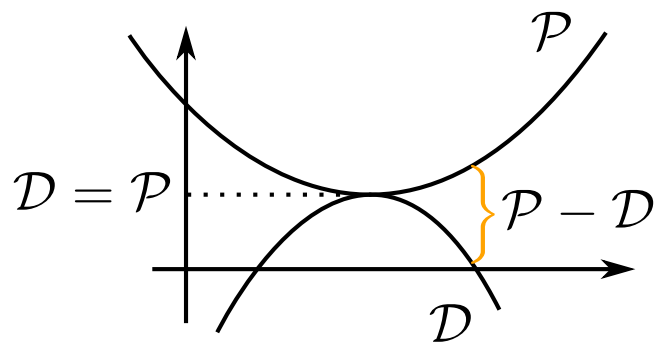
The Dual

The **dual** is a complementary problem to the primal problem. For the lasso, it is

$$\underset{\theta \in \mathbb{R}^n}{\text{maximize}} \left\{ D(\theta; \lambda) = \frac{1}{2} y^T y - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda} \right\|_2^2 . \right\}$$

The **duality gap** is the difference between the primal and dual objectives and is tight at the optimum, that is

$$P(\hat{\beta}; \lambda) - D(\hat{\theta}; \lambda) = 0.$$



This means that the dual and primal—in the case of the lasso—are related via

$$y = X\hat{\beta}(\lambda) + \lambda\hat{\theta}(\lambda).$$

KKT Conditions

The Karush–Kuhn–Tucker (KKT) stationarity condition for the lasso specify that

$$\mathbf{0} \in X^T(X\beta - y) + \lambda\partial,$$

where ∂ is the subdifferential of the ℓ_1 norm, with elements given by

$$\partial_j = \begin{cases} \{\text{sign}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_j = 0. \end{cases}$$

This means that

$$|x_j^T \hat{\theta}| < 1 \implies \hat{\beta}_j = 0.$$

Safe Screening Rules

Safe screening rules work by finding a region within which the dual-optimal solution, $\hat{\theta}$, must lie.

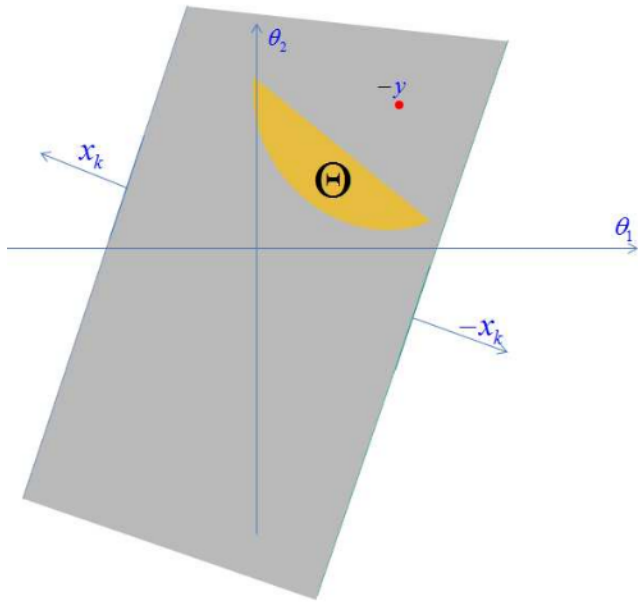


Figure 1: Safe region from original SAFE screening rule.

I.e., we replace the inequality $|x_j^T \hat{\theta}| < 1$ with an inequality that leads to the same conclusion, but doesn't involve $\hat{\theta}$.

Gap Safe Screening Rule

The Gap Safe screening rule uses the *duality gap* to define such a region, and disards the jth predictor if

$$|X^T\theta_\lambda|_j + \|x_j\|_2\sqrt{\frac{1}{\lambda_*^2}\big(\mathcal{P}(\beta_\lambda;\lambda^*) - \mathcal{D}(\theta_\lambda;\lambda^*)\big)} < 1$$

where

$$\theta_\lambda = \frac{y - X\beta_\lambda}{\max\big(|X^T(y - X\beta_\lambda)|, \lambda\big)}.$$

Dynamic Screening

Duality gap *decreases* during iterative optimization—screening becomes better and better.

Our Contribution: Look-Ahead Screening

Inequality on last slide is *quadratic*, which means we can find the next *critical* point easily:

$$\lambda_* = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where

$$\begin{aligned} a &= \left(1 - |x_j^T \theta_\lambda|\right)^2 - \frac{1}{2}\theta_\lambda^T \theta_\lambda \|x_j\|_2^2, \\ b &= \left(\theta_\lambda^T y - \|\beta_\lambda\|_1\right)\|x_j\|_2^2, \\ c &= -\frac{1}{2}\|y - X\beta_\lambda\|_2^2 \|x_j\|_2^2. \end{aligned}$$

This allows us to screen predictors for all upcoming steps.

Our Contribution: Look-Ahead Screening

Inequality on last slide is *quadratic*, which means we can find the next *critical* point easily:

$$\lambda_* = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

where

$$\begin{aligned} a &= \left(1 - |x_j^T \theta_\lambda|\right)^2 - \frac{1}{2}\theta_\lambda^T \theta_\lambda \|x_j\|_2^2, \\ b &= \left(\theta_\lambda^T y - \|\beta_\lambda\|_1\right)\|x_j\|_2^2, \\ c &= -\frac{1}{2}\|y - X\beta_\lambda\|_2^2 \|x_j\|_2^2. \end{aligned}$$

This allows us to screen predictors for all upcoming steps.

Say we are at step k along the path; then simply check the inequality for all $\lambda_{k+1}, \lambda_{k+2}, \ldots$

Example

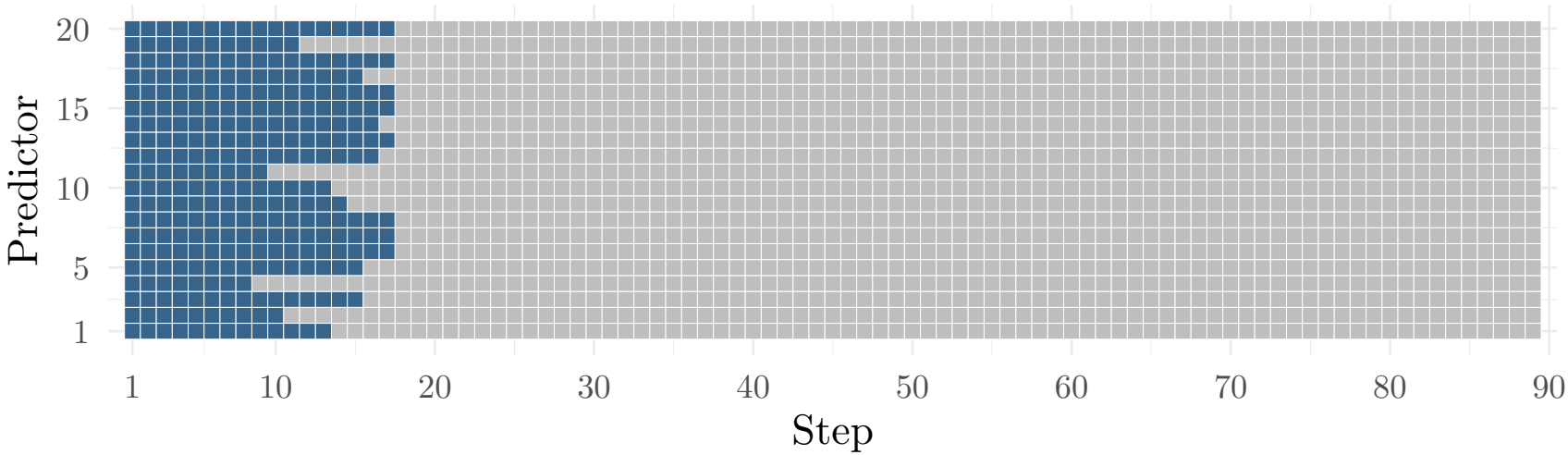


Figure 2: The predictors screened at the first step of the lasso path via look-ahead screening for a random sample of 20 predictors from the *leukemia* data set. A blue square indicates that the corresponding predictor can be discarded at the respective step.

Results on Simulated Data

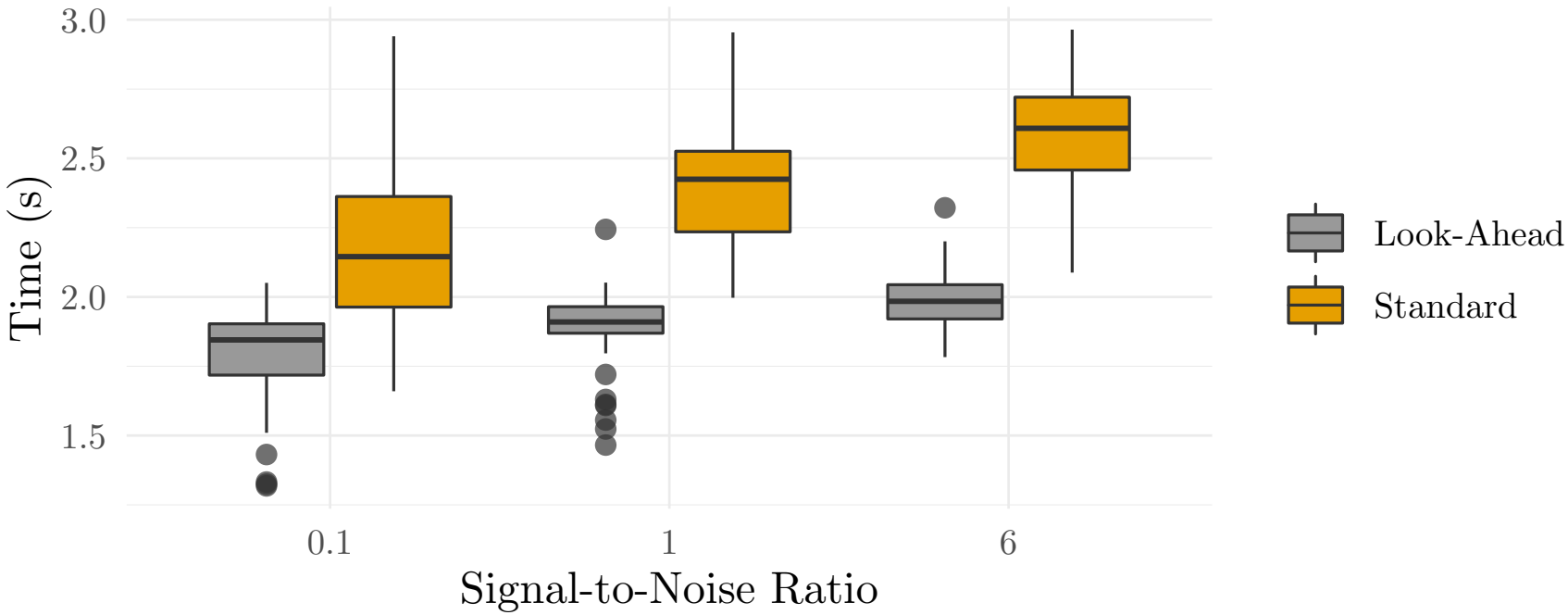


Figure 3: Standard box plots of timings to fit a full lasso path to a simulated data set with $n = 100$, $p = 50\,000$, and five true signals.

code and results available at github.com/jolars/LookAheadScreening

Conclusions

- simple idea; easily extendable to other rules with similar properties
- application comes essentially *for free*
- agnostic to solver used
- likely works for other loss function too (but we have not studied this yet)

References I


Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. ISSN: 0035-9246. JSTOR: 2346178.