The Strong Screening Rule for SLOPE

Mathematical Methods of Modern Statistics 2

Johan Larsson[1] Małgorzata Bogdan[1,2] Jonas Wallin[1]

[1]Department of Statistics, Lund University,

[2]Department of Mathematics, University of Wroclaw

June 16, 2020

Sorted L-One Penalized Estimation (SLOPE)

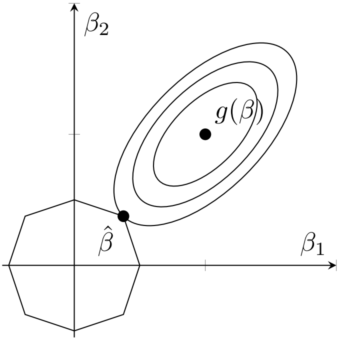
The SLOPE (**bogdan2015**) estimate is

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \{g(\beta) + J(\beta; \lambda)\}$$

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

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0, \qquad |\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}.$$

Equivalent to an inequality-constrained convex optimization problem



Motivation for screening rules

- we are interested in a **path** of penalties $\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(m)}$, many of which will lead to **sparse** solutions

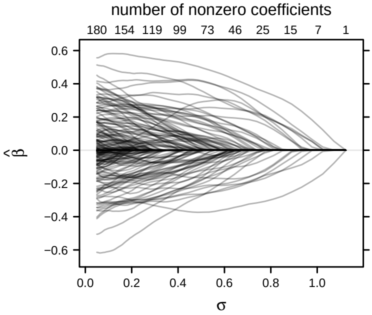


Figure 1: SLOPE path with $n = 200$, $p = 20000$. σ indicates strength of regularization.

Motivation for screening rules

- we are interested in a **path** of penalties $\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(m)}$, many of which will lead to **sparse** solutions
- **basic idea**: what if we could, based on a relatively **cheap** test, determine which predictors will be inactive before fitting the model?

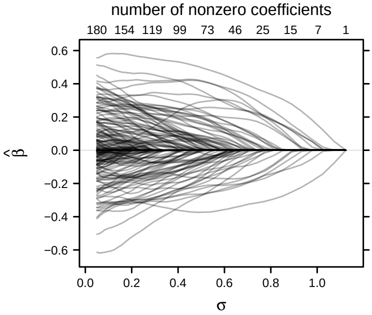


Figure 1: SLOPE path with $n = 200$, $p = 20000$. σ indicates strength of regularization.

Motivation for screening rules

- we are interested in a **path** of penalties $\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(m)}$, many of which will lead to **sparse** solutions
- **basic idea**: what if we could, based on a relatively **cheap** test, determine which predictors will be inactive before fitting the model?
- it turns out that we can, using screening rules!

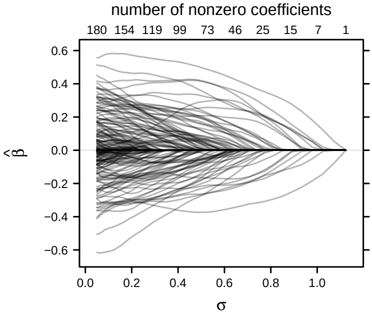


Figure 1: SLOPE path with $n = 200$, $p = 20000$. σ indicates strength of regularization.

Strong screening rule for SLOPE

Assume that we have a solution for $\lambda^{(k-1)}$ and want the solution at $\lambda^{(k)}$.

Using the optimality condition for the SLOPE problem,

$$\mathbf{0} \in \nabla g\big(\beta(\lambda^{(k)})\big) + \partial J\big(\beta(\lambda^{(k)}); \lambda^{(k)}\big),$$

where ∂J is the subgradient, we can determine which predictors will be active at $\lambda^{(k)}$.

Strong screening rule for SLOPE

Assume that we have a solution for $\lambda^{(k-1)}$ and want the solution at $\lambda^{(k)}$.

Using the optimality condition for the SLOPE problem,

$$\mathbf{0} \in \nabla g\big(\beta(\lambda^{(k)})\big) + \partial J\big(\beta(\lambda^{(k)}); \lambda^{(k)}\big),$$

where ∂J is the subgradient, we can determine which predictors will be active at $\lambda^{(k)}$.

our contribution
approximate $\nabla g\big(\beta(\lambda^{(k)})\big)$ and use optimality criterion (as if $\nabla g\big(\beta(\lambda^{(k)})\big)$ was known) to discard predictors

Strong screening rule for SLOPE

Assume that we have a solution for $\lambda^{(k-1)}$ and want the solution at $\lambda^{(k)}$.

Using the optimality condition for the SLOPE problem,

$$\mathbf{0} \in \nabla g\big(\beta(\lambda^{(k)})\big) + \partial J\big(\beta(\lambda^{(k)}); \lambda^{(k)}\big),$$

where ∂J is the subgradient, we can determine which predictors will be active at $\lambda^{(k)}$.

our contribution
approximate $\nabla g\big(\beta(\lambda^{(k)})\big)$ and use optimality criterion (as if $\nabla g\big(\beta(\lambda^{(k)})\big)$ was known) to discard predictors

violations

- violations—incorrectly discarding predictors—may occur
- can always be caught by checking optimality conditions (and refitting if present)
- are **so rare** that, in practice, the benefits from using the rule far outweigh the costs it incurs

Efficiency for real data

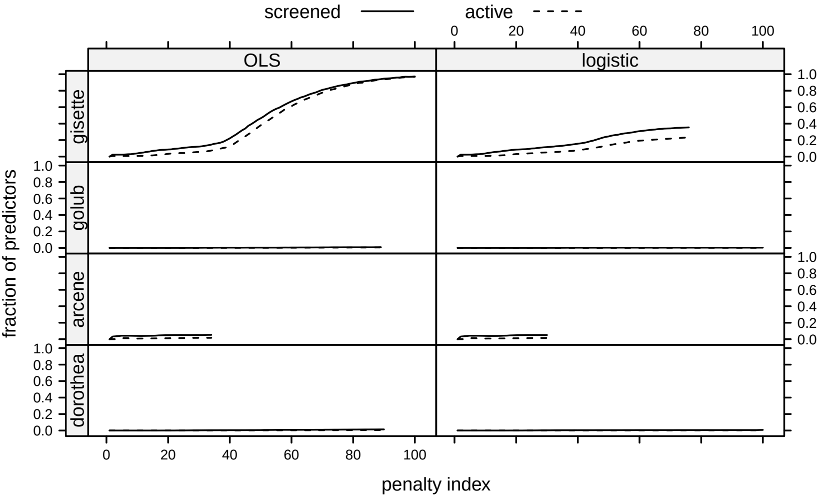


Figure 2: Efficiency for real data sets. The dimensions of the predictor matrices are 100×9920 (arcene), 800×88119 (dorothea), 6000×4955 (gisette), and 38×7129 (golub).

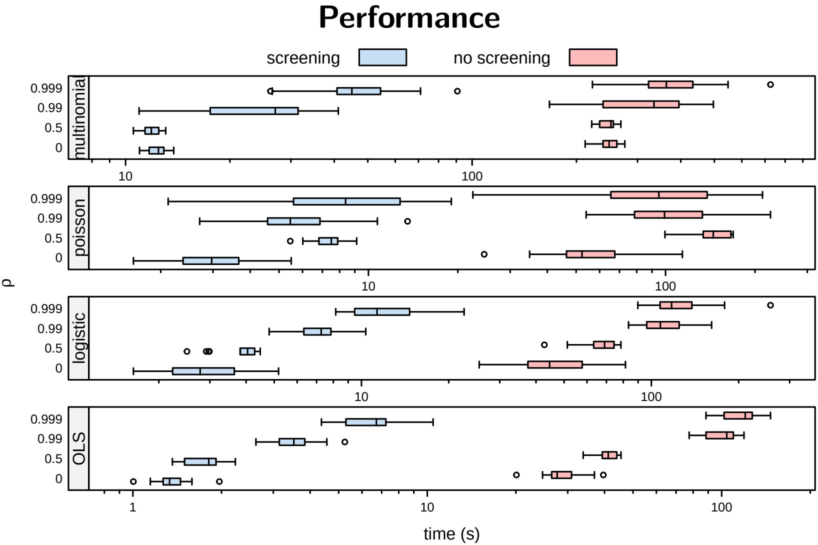


Figure 3: Performance benchmarks for various generalized linear models with $X \in \mathbb{R}^{200\times 20000}$. Predictors are autocorrelated through an $\mathrm{AR}(1)$ process with correlation ρ.

References I