The Strong Screening Rule for SLOPE

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

[1]*Department of Statistics, Lund University*
[2]*Department of Mathematics, University of Wroclaw*
[*]johan.larsson@stat.lu.se

Introduction

Sorted L-One Penalized Estimation (SLOPE) [1] is a sparsity-enforcing regularization method for regression, which can be viewed as an extension of the lasso. Compared the the lasso, however, SLOPE offers better control of false discoveries, enhanced predictive performance, and handles correlated predictors in a natural fashion.

The lasso, however, is more efficient to estimate. A major reason for this is *screening rules*, which employ cheap tests to discard predictors before fitting the model, thereby reducing the dimension of the problem. When the number of predictors (p) outnumber the number of observations (n), such reductions may be dramatic—Tibshirani et al. [2] developed a screening rule for the lasso called the *strong rule* and showed that it reduces the time required to solve the lasso by orders of magnitude in the $p \gg n$ regime.

In this poster we present the strong rule for SLOPE [3]: a generalization of the strong rule for the lasso and show that it too yields considerable performance improvements, consistently reducing the time required to solve SLOPE to fractions of that required otherwise.

SLOPE

The SLOPE [1] estimate is

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

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

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

The solution to SLOPE can be seen as the projection from $g(\beta)$ onto the sorted ℓ_1 norm (Figure 1).

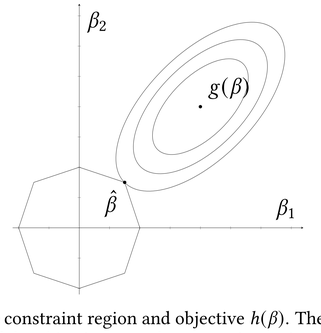

Figure 1: The constraint region and objective $h(\beta)$. The solution $\hat{\beta}$ is a projection of the objective onto the constraint region.

Motivation

We are interested in fitting a path of regularization penalties, $\lambda^{(1)} \geq \lambda^{(2)} \geq \cdots \geq \lambda^{(m)}$ with $\lambda^{(1)}$ chosen so that it yields the intercept-only model.

We know that many of the solutions are going to be sparse and that this sparsity pattern is related to the optimality condition

$$\mathbf{0} \in \nabla g(\hat{\beta}(\lambda)) + \partial J(\hat{\beta}(\lambda); \lambda)$$

where $\partial J(\hat{\beta})$ is the subdifferential for the sorted ℓ_1 norm. Before providing the definition for the $\partial J(\hat{\beta})$, let $\mathcal{A}_i$ be a set of predictors (cluster) for which the corresponding coefficients are equal in absolute value, $R(x)$ an operator that returns the ranks of elements in $|x|$, and $|x|_\downarrow$ be an operator that returns the absolute values of x sorted in non-increasing order. The subdifferential for the sorted ℓ_1 norm is then the set of all $g \in \mathbb{R}^p$ such that

$$g_{\mathcal{A}_i} = \left\{ s \in \mathbb{R}^{|\mathcal{A}_i|} \;\middle|\; \begin{cases} \operatorname{cumsum}(|s|_\downarrow - \lambda_{R(s)_{\mathcal{A}_i}}) \preceq \mathbf{0} & \text{if } \beta_{\mathcal{A}_i} = \mathbf{0}, \\ \operatorname{cumsum}(|s|_\downarrow - \lambda_{R(s)_{\mathcal{A}_i}}) \preceq \mathbf{0} & \\ \quad \wedge \sum_{j \in \mathcal{A}_i} \left(|s_j| - \lambda_{R(s)_j} \right) = 0 & \text{otherwise.} \end{cases} \right\}$$

Strong Rule for SLOPE

Suppose we are at step k on the path, i.e. have the solution at $k-1$. In this case, we could determine the active set of predictors exactly by checking the optimality condition if $\nabla g(\hat{\beta}(\lambda^{(k)}))$ was available. Because it is not, however, we substitute the gradient with an approximation based on the assumption that the gradient is a piece-wise linear function of λ with a slope no greater than one, which leads to the approximation

$$\underbrace{\underbrace{\left| \nabla g\left(\hat{\beta}(\lambda^{(k-1)}) \right) \right|_\downarrow}_{\text{previous gradient}} + \underbrace{\lambda^{(k-1)} - \lambda^{(k)}}_{\text{unit slope bound}}}_{\text{gradient prediction for step } k.}$$

This approximation is the basis of the strong rule for SLOPE, which consists of running an algorithm that

1. creates a **screened set** of predictors using the subdifferential and optimality conditions,
2. fits SLOPE and checks optimality conditions, and
3. if there are any violations in 2, includes these predictors in the screened set and refits the model.

Efficiency

We tested the screening rule on four real data sets to examine how many of the inactive predictors the rule correctly discards. The results show that the rule is able to reduce the dimension of the problem immensely (Figure 2).

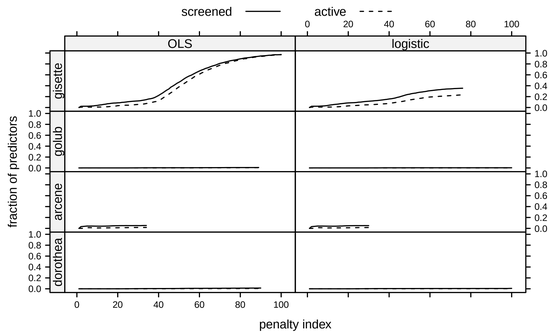

Figure 2: Efficiency for real data sets. The dimensions of the predictor matrices are 100×9920 (arcene), 800×88119 (dorothea), 6000×4955 (gisette), and 38×7129 (golub).

Performance

To test out the performance of the rule, we fit least-squares, logistic, Poisson, and multinomial models regularized with the sorted ℓ_1 norm for data sets with $n = 200$ and $p = 20000$.

The reduction in computation time for solving SLOPE is dramatic, consistently yielding more than tenfold speed increase (Figure 3).

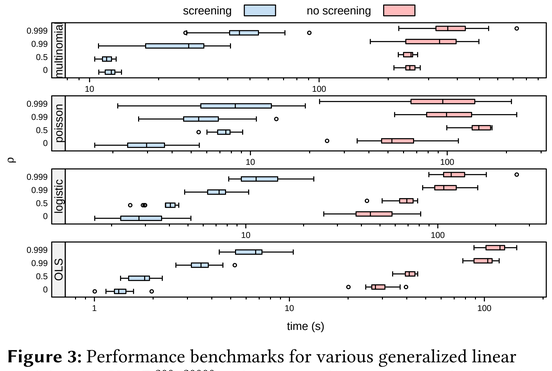

Figure 3: Performance benchmarks for various generalized linear models with $X \in \mathbb{R}^{200 \times 20000}$. Adjacent predictors are correlated with correlation ρ.

Violations of the Rule

Our screening rule is **heuristic**, which means that it may lead to violations: discarding predictors erroneously. If violations occur the model must be refit. To study the prevalence of violations of the screening rule, we conducted simulations on data from the the model $y = X\beta$ with $X \in \mathbb{R}^{p \times 2000}$ with varying p and level of corelation between predictors. Figure 4 shows that such violations are rare.

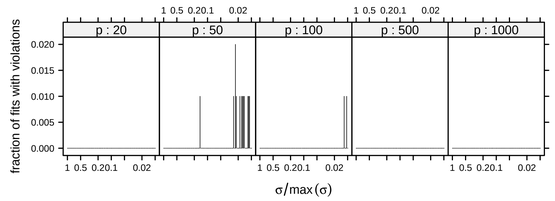

Figure 4: Violations for sorted ℓ_1 regularized least squares regression with predictors pairwise correlated with $\rho = 0.5$. $X \in \mathbb{R}^{100 \times p}$.

We also note that there were no violations of the rule in the examples showcased in Figures 2 and 3.

Additional Resources

The findings presented in this poster are available in detailed form in a preprint on ArXiv [3] along with additional results and theoretical background.

A software implementation of the strong screening rule for SLOPE is also available via the SLOPE package for R (https://CRAN.R-project.org/package=SLOPE).

References

[1] Małgorzata Bogdan et al. "SLOPE - Adaptive Variable Selection via Convex Optimization". In: *The annals of applied statistics* 9.3 (2015), pp. 1103–1140. ISSN: 1932-6157. DOI: 10.1214/15-AOAS842. pmid: 26709357.

[2] Robert Tibshirani et al. "Strong Rules for Discarding Predictors in Lasso-Type Problems". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 74.2 (Mar. 2012), pp. 245–266. ISSN: 1369-7412. DOI: 10/c4bb85.

[3] Johan Larsson, Małgorzata Bogdan, and Jonas Wallin. "The Strong Screening Rule for SLOPE". In: *arXiv:2005.03730 [cs, stat]* (May 2020). arXiv: 2005.03730 [cs, stat].

Contact

- johan.larsson@stat.lu.se
- https://larssonjohan.com
- twitter: johanlarsson86