EVIC 2023

Column Generation-Based Decomposition for Large-Scale Feature Selection Problems

Nicolás Acevedo, Fernando Ordoñez

Departamento de Ingeniería Industrial, Universidad de Chile.

1. Introduction

Feature selection is central to constructing fast, accurate, and interpretable models from large-scale data. Many solutions for this problem exist, but not all of them can handle large-scale instances at reasonable times without compromising solution quality. We devised an efficient alternative method for large-scale feature selection problems, utilizing a state-of-the-art decomposition method for a second-order cone formulation.

2. Motivation

The following plot shows a fast increase in the execution times of some large-scale instances of unconstrained LASSO.

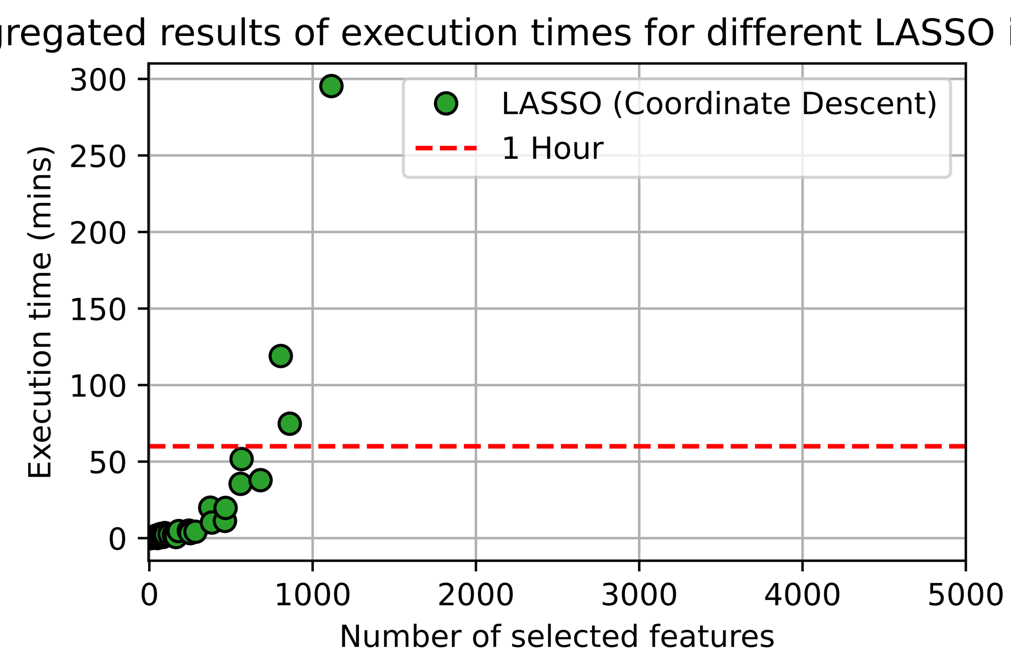


Figura 1: *Instances with* 10000 *obsservations and total # of features in* $\{1000, 2000, \ldots, 5000\}$.

3. Methodology

The decomposition method of [1] is a Column Generation technique to address large-scale conic optimization problems. We propose the following **Second-Order Cone Problem** (SOCP) formulation for feature selection on a linear model, inspired by [2] and [3].

$$\text{(SOCP)} \quad \omega = \min_{\beta,z,u,\xi} \quad \xi^2 + \tau \sum_{i=1}^{m} z_i + \kappa \sum_{i=1}^{m} u_i \tag{1a}$$

$$\text{s.t.} \quad \|y - X\beta\|_2 \le \xi, \tag{1b}$$

$$\beta_i^2 \le z_i u_i, \qquad i = 1, \ldots, m, \tag{1c}$$

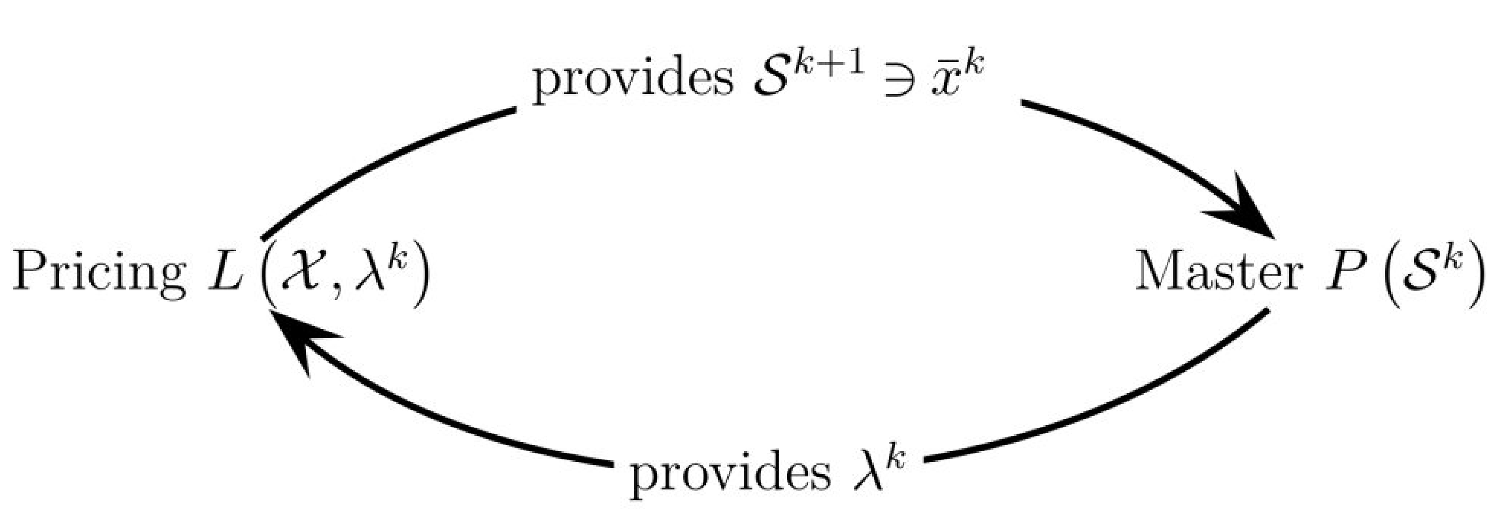
$$u \in \mathbb{R}_+^m, \tag{1d}$$

$$z \in \mathbb{R}_+^m, \tag{1e}$$

were we use the classic notation for linear models $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times m}$.

The **decomposition method** splits the problem in 2 subproblems, for iterations $k \ge 0$:

1. **Master problem:** $P(\mathcal{S}^k)$. Problem (1), but on a subset $\mathcal{S}^k$ of the feasible set of the original problem.
   - It returns the solution $x^k = (\beta^k, z^k, u^k)$ and the dual solution λ^k associated with the constraint (1b).
2. **Pricing problem:** $L(\mathcal{X}, \lambda^k)$. The Lagrangian relaxation of constraint (1b).
   - It returns the solution $\overline{x}^k = (\overline{\beta}^k, \overline{z}^k \overline{u}^k)$, to add to $\mathcal{S}^{k+1}$.

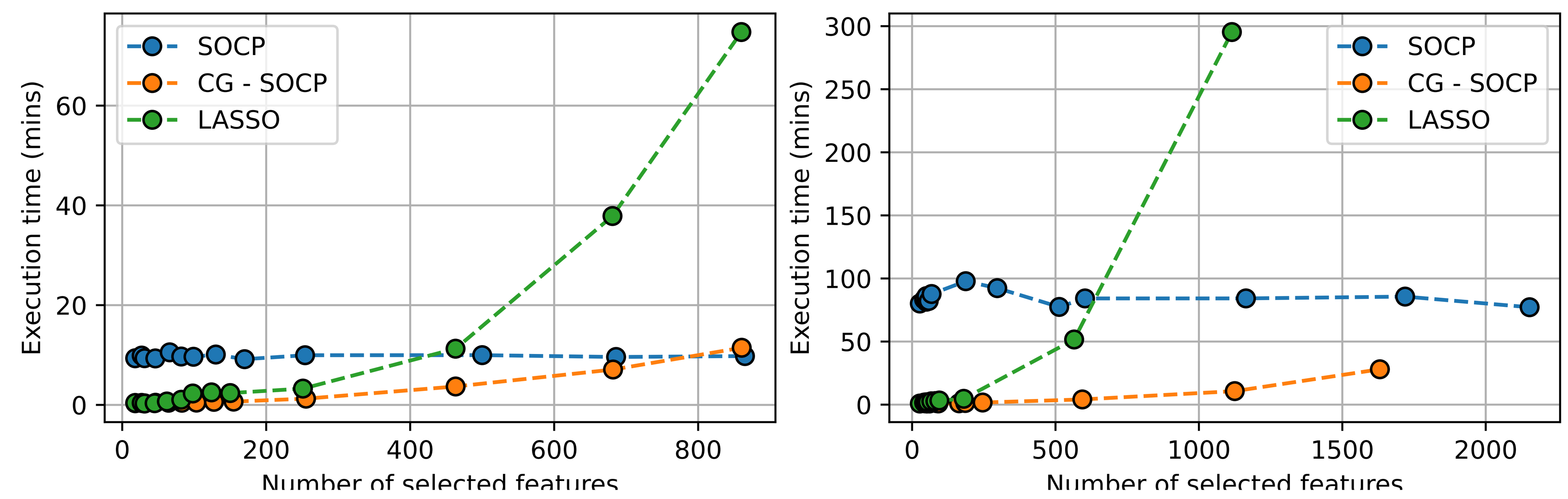


Proposition 1 $\forall \tau, \kappa > 0$, penalization parameters of SOCP (1), $\exists \lambda = 2\sqrt{\tau\kappa}$, penalization parameter of the unconstrained LASSO $\min_\beta\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\}$, such that SOCP (1) and LASSO are equivalent.

4. Results & Discussion

For the results, CG refers to the column generation decomposition, and LASSO refers to the unconstrained LASSO solved via Coordinate Descent.

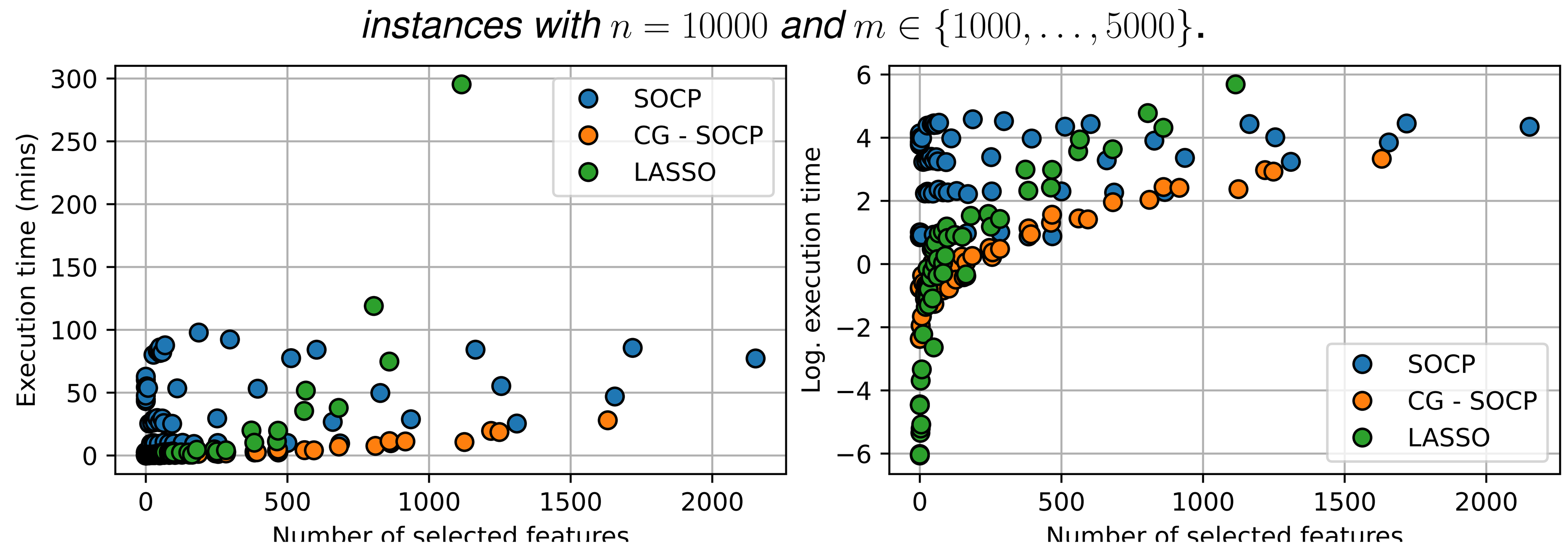
Mean execution times results in terms of the number of selected features, for two different instances with $n = 10000$.



(a) $m = 2000$ (b) $m = 5000$

The SOCP (2) solves the entire problem independent of the penalty, so its execution time is almost constant. The coordinate descent method is the fastest when approximately 75 features at selected at most. The decomposition method solves smaller instances of the SOCP (2) and it is fastest when approximately 75 to 40% of the features are selected.

Aggregated results of execution times in terms of number of selected features, for instances with $n = 10000$ *and* $m \in \{1000, \ldots, 5000\}$.



(a) *Normal times* (b) *Log. times*

Neither coordinate descent nor the decomposition method execution times appear to depend on the size of the instances, as they have a clear dependence on the number of selected features only among the different instances. Therefore, the method should solve the problem in competitive times independent of its size, for certain penalties.

References

[1] Chicoisne, R. (2023). Computational aspects of column generation for nonlinear and conic optimization: classical and linearized schemes. *Computational Optimization and Applications*, 84(3):789–831.

[2] Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813 – 852.

[3] Küçükyavuz, S., Shojaie, A., Manzour, H., Wei, L., and Wu, H.-H. (2020). Consistent second- order conic integer programming for learning bayesian networks. arXiv:2005.14346.

Acknowledgements

This work was supported by ANID–Subdirección de Capital Humano/Magíster Nacional 2022–Folio 22220861.