

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

Nicolás Acevedo, Fernando Ordóñez

Departamento de Ingeniería Industrial, Universidad de Chile

Motivation

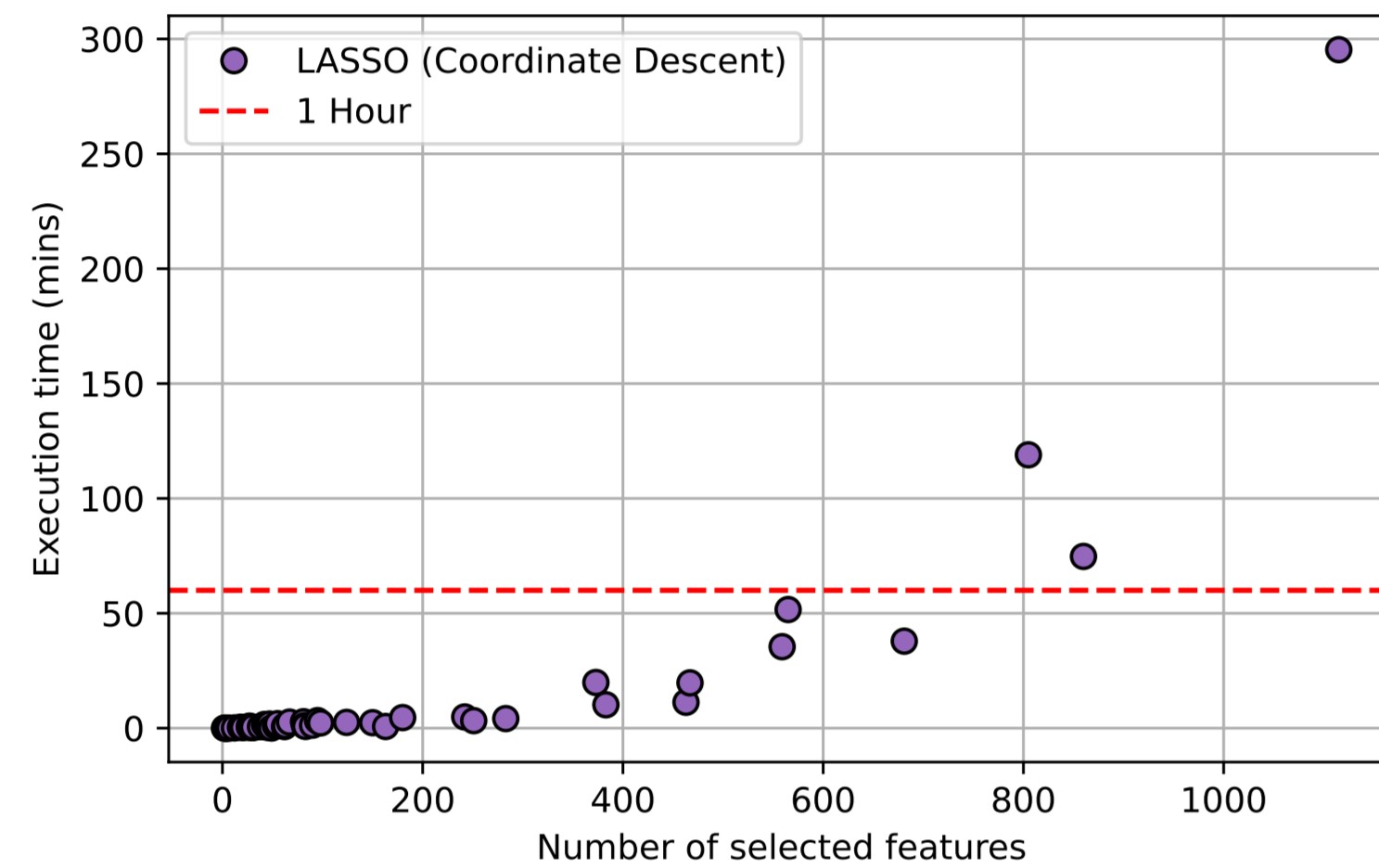
Feature selection is central to constructing fast, accurate, and interpretable models from large-scale data. Many techniques exist, but **not all can handle large-scale instances at reasonable times** without compromising solution quality.

We show the example of **unconstrained LASSO**, which solves a minimization problem of the form

$$(\text{LASSO}) \quad \min_{\beta} \quad \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \quad (1)$$

within the context of a linear model $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times m}$, $\beta \in \mathbb{R}^m$, and λ is a penalty parameter.

Figure 1. Aggregated execution time results for LASSO instances with $n = 10000$ observations and a total number of features $m \in \{1000, 2000, 3000, 4000, 5000\}$.



Objectives

1. Develop an exact **second-order cone formulation** for feature selection.
2. Decompose the conic problem with a **column generation method** for large-scale instances.
3. **Test the performance** of the decomposed problem against its original problem and state-of-the-art techniques.

Problem Formulation

Similar to LASSO, the proposed formulation minimizes the sum of squared errors of a linear model, while penalizing the number of non-zero coefficients $\beta_i \neq 0$.

It is a **Second-Order Cone Program (SOCP)**, inspired by [1][3], of the form

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

$$\text{s.t.} \quad \|y - X\beta\|_2 \leq \xi, \quad (2b)$$

$$\left\| \begin{pmatrix} u_i - z_i \\ 2\beta_i \end{pmatrix} \right\|_2 \leq u_i + z_i, \quad i = 1, \dots, m, \quad (2c)$$

$$u \in \mathbb{R}_+^m, \quad (2d)$$

$$z \in \mathbb{R}_+^m, \quad (2e)$$

where the constraints (2c) are equivalent to $\beta_i^2 \leq z_i u_i$, $i = 1, \dots, m$.

▪ **Definition:** A set $\mathcal{L}_{n+1} := \{(x, y) \in \mathbb{R}^{n+1} : \|x\|_2 \leq y\}$ is called a **second-order cone**.

LASSO equivalence: $\forall \tau, \kappa > 0$, penalty parameters of the SOCP (2), $\exists \lambda = 2\sqrt{\tau\kappa}$, penalty parameter of the unconstrained LASSO (1), such that **the two problems are equivalent**.

Decomposition Method

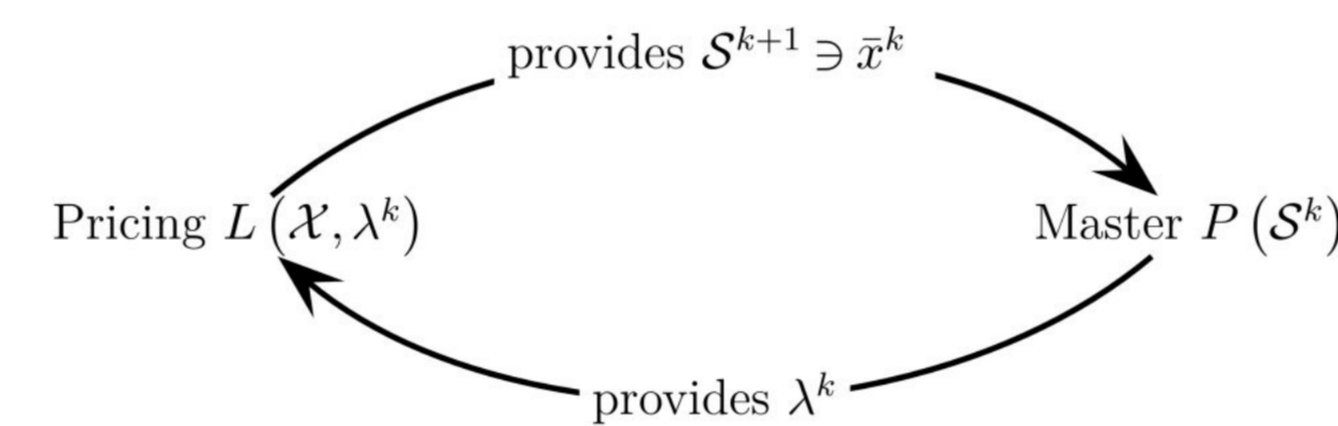
It is a **Column Generation** proposed in [2] for solving large conic problems of the form

$$(P(\mathcal{X})) \quad \omega(\mathcal{X}) = \min_{x, y} \quad f(x, y) \\ \text{s. t.} \quad x \in \mathcal{X}, \\ -g(x, y) \in \mathcal{C},$$

where \mathcal{X} is a high-dimensional set and \mathcal{C} is a cone. The **issue** with $P(\mathcal{X})$ is the **joint presence** of the difficult constraints $-g(x, y) \in \mathcal{C}$ and the magnitude of the dimension $n \gg 1$ of \mathcal{X} .

The method decomposes the problem in **2 subproblems** for each iteration $k \geq 0$:

1. **Master problem $P(\mathcal{S}^k)$:** Problem (2), but on a subset $\mathcal{S}^k \subset \mathcal{X}$ of the feasible region 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 conic constraint $-g(x, y) \in \mathcal{C}$.
2. **Pricing problem $L(\mathcal{X}, \lambda^k)$:** The Lagrangian relaxation of constraint $-g(x, y) \in \mathcal{C}$.
 - It returns the solution $\bar{x}^k = (\bar{\beta}^k, \bar{z}^k, \bar{u}^k)$ to construct \mathcal{S}^{k+1} .



SOCP Decomposition

The feasible set \mathcal{S}^k of the **master problem** can be constructed with the combination $A^k \pi$ of the initial and previous pricing solutions $\{A^s\}_{s=0}^{k-1} = \{(\bar{\beta}^s, \bar{z}^s, \bar{u}^s)\}_{s=0}^{k-1}$.

$$(P(\mathcal{S}^k)) \quad \min_{\pi, \xi} \quad \xi^2 + \tau \sum_{i=1}^m z_i^k \pi + \kappa \sum_{i=1}^m u_i^k \pi \quad (4a)$$

$$\text{s.t.} \quad \begin{pmatrix} y - X\mathbf{B}^k \pi \\ \xi \end{pmatrix} \in \mathcal{L}_{n+1}, \quad (4b)$$

$$\begin{pmatrix} \mathbf{u}_i^k - \mathbf{z}_i^k \pi \\ 2\mathbf{b}_i^k \pi \end{pmatrix} \in \mathcal{L}_3, \quad i = 1, \dots, m, \quad (4c)$$

$$\mathbf{U}^k \pi \in \mathbb{R}_+^m, \quad (4d)$$

$$\mathbf{Z}^k \pi \in \mathbb{R}_+^m, \quad (4e)$$

$$\pi \in \Pi. \quad (4f)$$

It returns the optimal solution π^* to construct $(\beta^k, z^k, u^k) = (\mathbf{B}^k \pi^*, \mathbf{Z}^k \pi^*, \mathbf{U}^k \pi^*)$.

The best performance is obtained relaxing the first conic constraint (2b) in the **pricing problem**.

$$(L(\mathcal{X}, \lambda^k)) \quad \min_{\beta, z, u, \xi} \quad \xi^2 + \tau \sum_{i=1}^m z_i + \kappa \sum_{i=1}^m u_i - \left\langle \begin{pmatrix} \psi^k \\ \mu^k \end{pmatrix}, \begin{pmatrix} y - X\beta \\ \xi \end{pmatrix} \right\rangle \quad (5a)$$

$$\text{s.t.} \quad \begin{pmatrix} u_i - z_i \\ 2\beta_i \end{pmatrix} \in \mathcal{L}_3, \quad i = 1, \dots, m, \quad (5b)$$

$$u \in \mathbb{R}_+^m, \quad (5c)$$

$$z \in \mathbb{R}_+^m, \quad (5d)$$

where $\lambda^k = (\psi^k, \mu^k) \in \mathcal{L}^{n+1}$ is the dual solution of the first conic constraint (4b) and $\pi \in \Pi$ defines the type of combination. The problem returns a new solution $(\bar{\beta}^k, \bar{z}^k, \bar{u}^k)$ to construct \mathcal{S}^{k+1} .

Results

For the following, **SOCP** refers to directly solving problem (2) with MOSEK, **CG**– L_{C_1} refers to the decomposition method relaxing constraint (2b), and **LASSO** refers to problem (1) solved via Coordinate Descent.

Figure 2. **Execution times** of instances with $m = 2000$ and $n = 10000$.

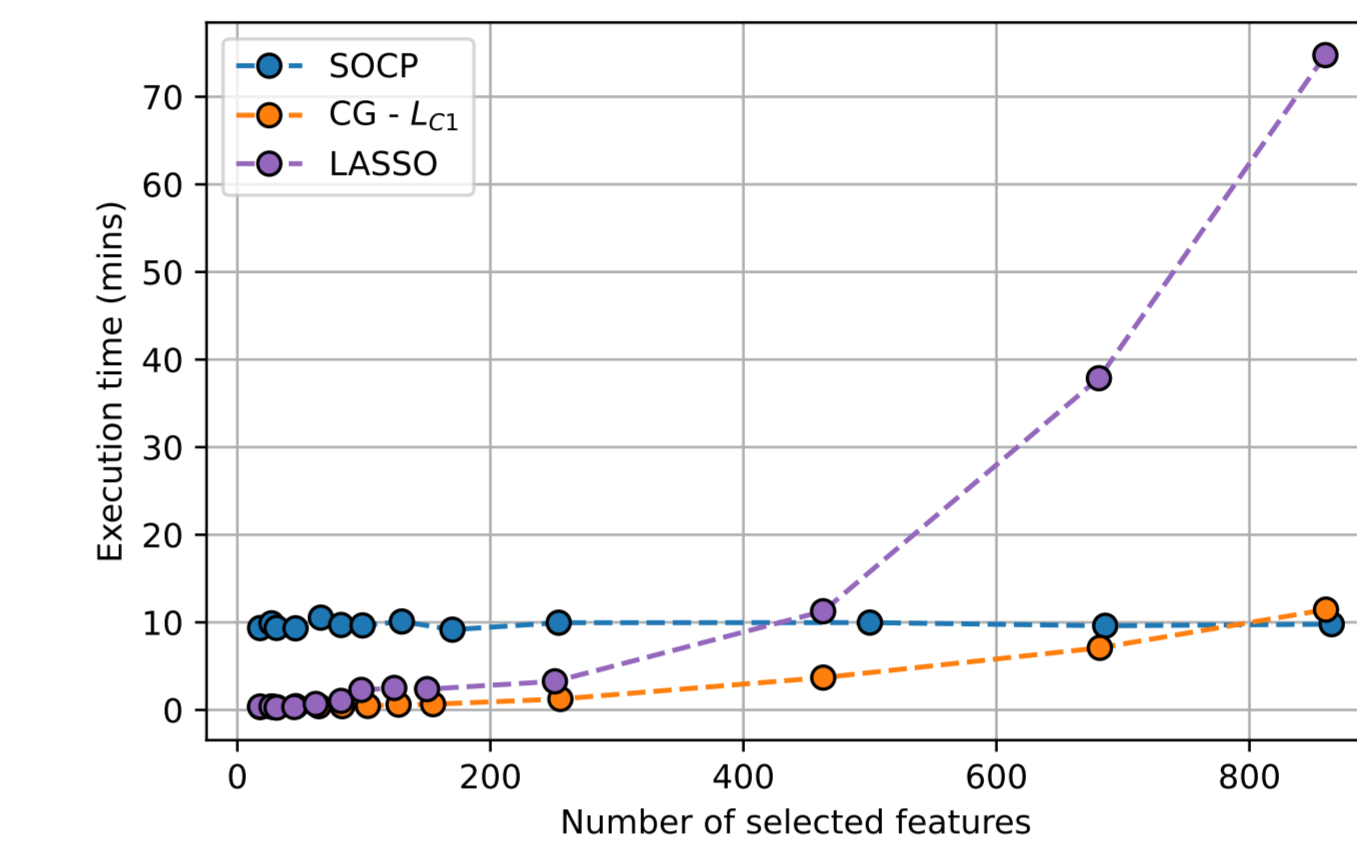
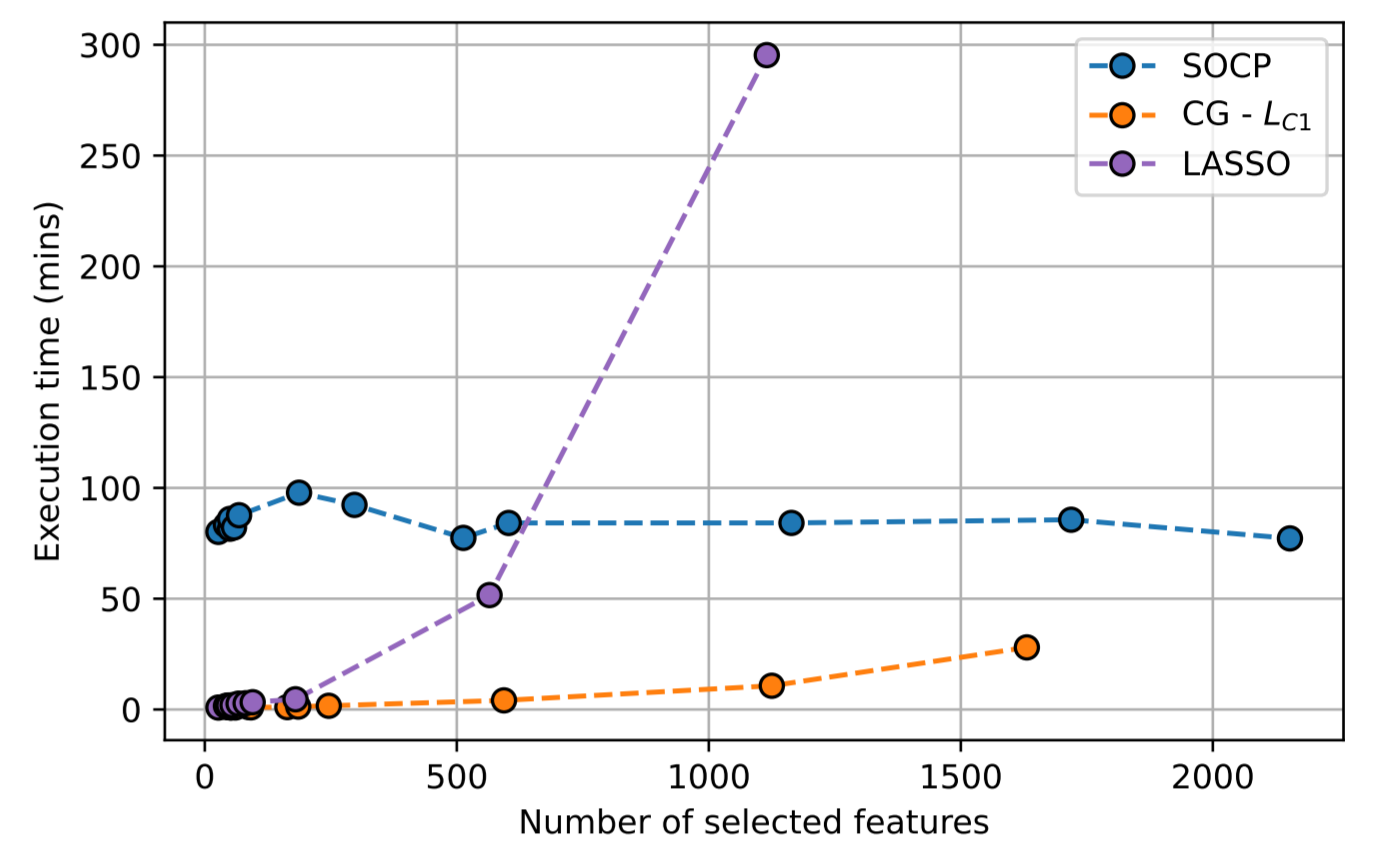
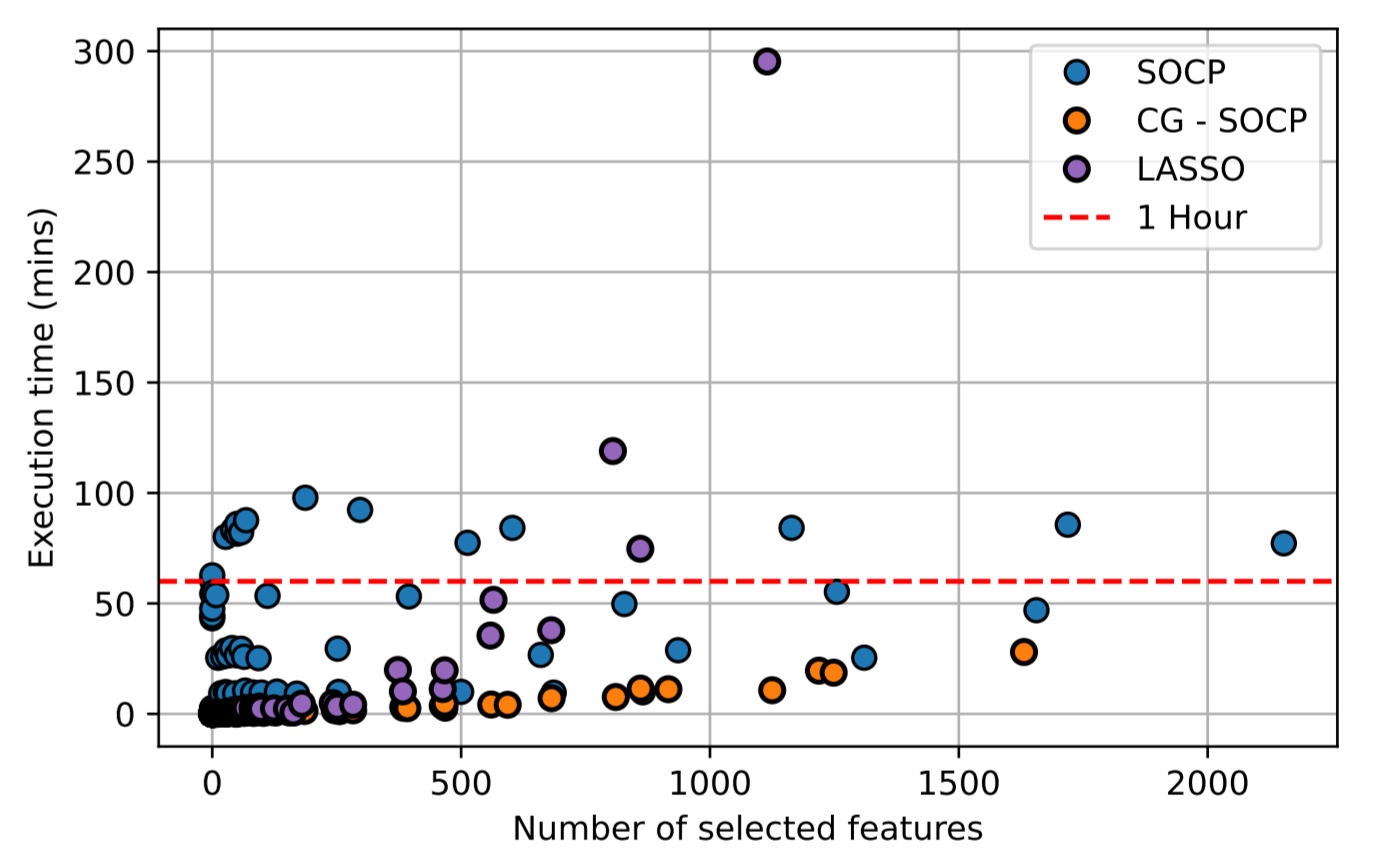
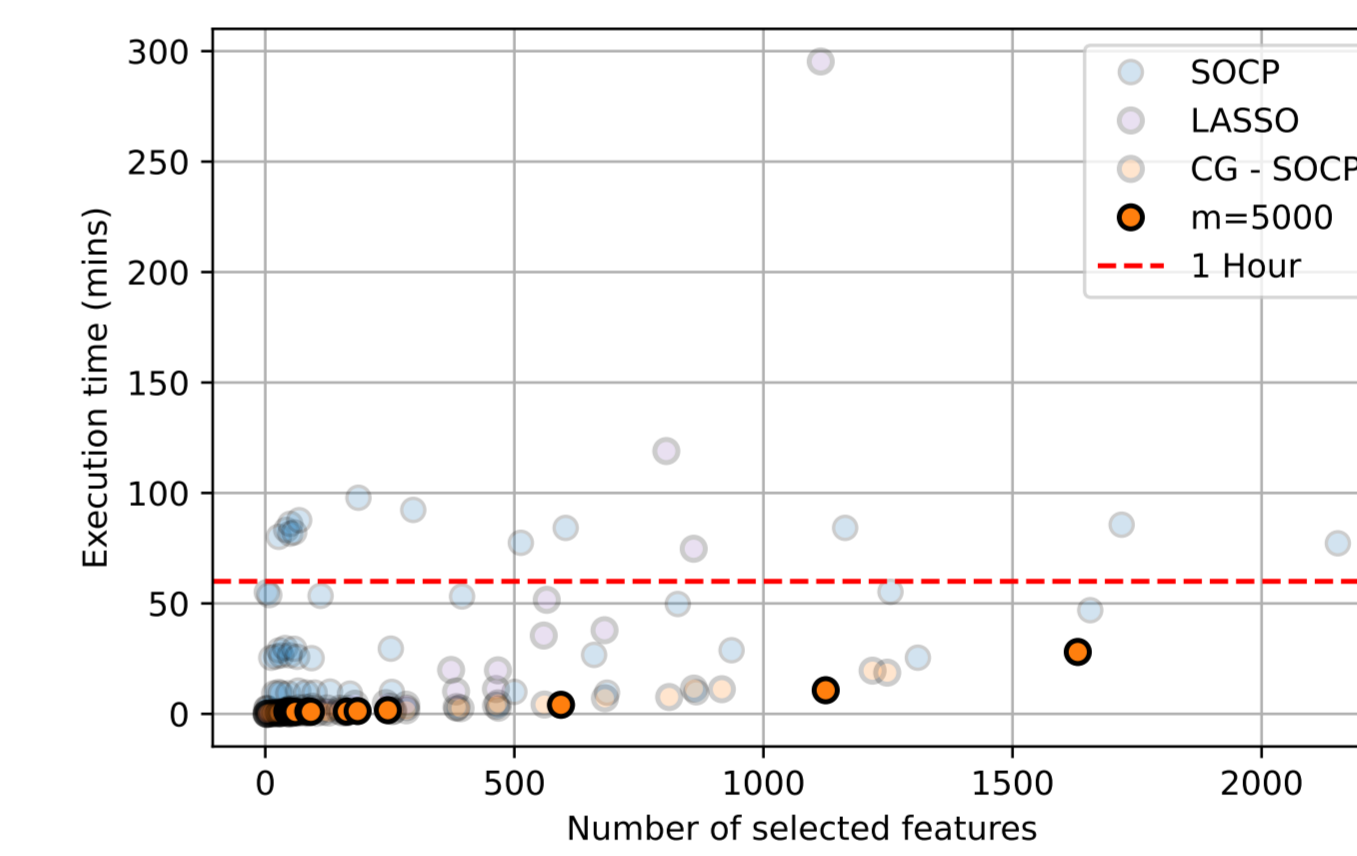


Figure 3. **Execution times** of instances with $m = 5000$ and $n = 10000$.



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 at most 75 features are selected. The decomposition method is the fastest when approximately 75 to 40% of the total features are non-zero.

Figure 4. **Aggregated execution times** of instances with $m \in \{1000, 2000, \dots, 5000\}$ and $n = 10000$.



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. Therefore, the method should solve the problem in competitive times independently of the size of the instance given specific penalty parameters.

References

- [1] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813 – 852, 2016.
- [2] Renaud Chicoisne. Computational aspects of column generation for nonlinear and conic optimization: classical and linearized schemes. *Computational Optimization and Applications*, 84(3):789–831, January 2023.
- [3] Simge Küçükyavuz, Ali Shojai, Hasan Manzour, Linchuan Wei, and Hao-Hsiang Wu. Consistent second-order conic integer programming for learning bayesian networks. *arXiv:2005.14346*, 2020.

Acknowledgments

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