Proximal operator for the sorted l 1 norm: Application to testing procedures based on SLOPE

Xavier Dupuis, Patrick J C Tardivel

Proximal operator for the sorted ℓ_1 norm: Application to testing procedures based on SLOPE

Xavier Dupuis* Patrick J.C. Tardivel*

Abstract

A decade ago OSCAR was introduced as a penalized estimator where the penalty term, the sorted ℓ_1 norm, allows to perform clustering selection. More recently, SLOPE was introduced as a penalized estimator controlling the False Discovery Rate (FDR) as soon as the hyper-parameter of the sorted ℓ_1 norm is properly selected. For both, OSCAR and SLOPE, numerical schemes to compute these estimators are based on the proximal operator of the sorted ℓ_1 norm. The main goal of this note is to provide a short and simple formula for this operator. Based on this formula one may observe that the output of the proximal operator has some components equal and thus this formula corroborates that SLOPE, as well as OSCAR, perform clustering selection. Moreover, our geometric approach to prove the formula for the proximal operator provides insights to show that testing procedures based on SLOPE are more powerful than step-down testing procedures but less powerful than step-up testing procedures.

Keywords: SLOPE, Proximal operator, Sorted ℓ_1 norm, False discovery rate.

1 Introduction

Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Bondell and Reich, 2008) and Sorted L-One Penalized Estimation (SLOPE) (Bogdan et al., 2015; Zeng and Figueiredo, 2014) are both penalized estimators based on the sorted ℓ_1 norm. First introduced in the particular case where the loss function is the residual sum of squares, these estimators are defined as follows:

$$\hat{\beta} \in \underset{b\in\mathbb{R}^p}{\operatorname{Argmin}} \frac{1}{2}\|Y - Xb\|_2^2 + \sum_{i=1}^{p} \lambda_i |b|_{\downarrow i} \text{ where } |b|_{\downarrow 1} \geq \cdots \geq |b|_{\downarrow p}.$$

For OSCAR the hyper-parameter $\lambda = (\lambda_1, \ldots, \lambda_p)$ has arithmetically decreasing components. SLOPE is both an extension of OSCAR where the hyper-parameter satisfies $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ and an extension of the Least Absolute Shrinkage and Selection Operator (LASSO) (Chen and Donoho, 1994; Tibshirani, 1996). Indeed, when $\lambda_1 = \cdots = \lambda_p > 0$ then the penalty $\sum_{i=1}^{p} \lambda_i |b|_{\downarrow i}$ coincides with the ℓ_1 norm; the well known penalty term for the LASSO. With respect to other penalized estimators, when $\lambda_1 > \cdots > \lambda_p > 0$, SLOPE (and a fortiori OSCAR) has the particularity to perform clustering selection, namely some components of the SLOPE estimator are equal in absolute value (Bondell and Reich, 2008; Schneider and Tardivel, 2020). Moreover, in the linear Gaussian model, by taking $\lambda_1 = \sigma\Phi^{-1}(1 - \alpha/2p), \ldots, \lambda_p = \sigma\Phi^{-1}(1 - \alpha/2)$ (where Φ is the standard normal cumulative distribution function) SLOPE estimator allows to controls the False Discovery Rate at level α in the particular case where X is an orthogonal matrix (Bogdan et al., 2015). Note that the above hyper-parameter also called BH sequence coincides with thresholds given in the seminal article introducing the Benjamini-Hocheberg's procedure and controlling the FDR (Benjamini and Hochberg, 1995).

*Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, F-21000 Dijon, France. Xavier.Dupuis@u-bourgogne.fr Patrick.Tardivel@u-bourgogne.fr

Numerically, like many penalized estimators where the loss function is smooth and the penalty term is non-smooth, one may solve SLOPE or OSCAR with a Forward-Backbard proximal gradient algorithm (see e.g. Combettes and Wajs (2005); Parikh and Boyd (2014)). This method relies on the computation of the proximal operator for the sorted ℓ_1 norm. This operator is also a very important tool for the development of the approximate message passing theory (Bu et al., 2020; Zhang and Bu, 2021). A statistical motivation for this operator relies on the fact that, when X is an orthogonal matrix (*i.e.* $X'X = I$), the SLOPE estimator is the image of the ordinary least squares estimator $\hat{\beta}^{\text{ols}} = X'Y$ by the proximal operator of the sorted ℓ_1 norm. Indeed, since $Y - X\hat{\beta}^{\text{ols}}$ is orthogonal to the vector space $col(X) := \{Xu : u \in \mathbb{R}^p\}$, one may deduce the following identity:

$$\hat{\beta}^{\text{slope}} = \underset{b\in\mathbb{R}^p}{\text{argmin}} \ \frac{1}{2}\|X(\hat{\beta}^{\text{ols}} - b)\|_2^2 + \sum_{i=1}^{p} \lambda_i |b|_{\downarrow i},$$

which gives the claimed formula when X is an isometry (*i.e.* $X'X = I$).

There is a particular case under which the proximal operator of the sorted ℓ_1 norm is explicit; when $y_1 \geq \cdots \geq y_p \geq 0$ and $y_1 - \lambda_1 \geq \ldots, \geq y_p - \lambda_p$ then the proximal operator is simply given by $((y_1 - \lambda_1)_+, \ldots, (y_p - \lambda_p)_+)$. Otherwise, when components of y are non-increasing and non-negative, an algorithm computing the proximal operator for the sorted ℓ_1 norm is given in Bogdan et al. (2015) (see algorithm 3). This algorithm suggests to first identify a sub-sequence $y_i - \lambda_i, \ldots, y_j - \lambda_j$ non-decreasing and non-constant (where $i < j$) and then to substitute 1) $(y_i, \ldots, y_j)$ by $((\sum_{l=i}^{j} y_l)/(j+1-i), \ldots, (\sum_{l=i}^{j} y_l)/(j+1-i))$ and 2) $(\lambda_i, \ldots, \lambda_j)$ by $((\sum_{l=i}^{j} \lambda_l)/(j+1-i), \ldots, (\sum_{l=i}^{j} \lambda_l)/(j+1-i))$. Whereas correct, we believe that this algorithm is difficult to implement because it is not easy to identify iteratively non-decreasing and non-constant sub-sequence.

The main motivation for this note is to provide a short and simple formula for the proximal operator of the sorted ℓ_1 norm. The proof for this formula is based on recent advances on sub-differential calculus for the sorted ℓ_1 norm and on the description of the signed permutahedron polytope (the signed permutahedron is the sub-differential of the sorted ℓ_1 norm at 0) (Schneider and Tardivel, 2020). We also illustrate that some sub-differential calculus rules give geometrical insights for testing procedures based on SLOPE. In particular there is a geometrical way to understand why the testing procedure based on SLOPE with the BH sequence is more conservative than the seminal Benjamini Hochberg's procedure.

1.1 Notation

Given a hyper-parameter $\lambda = (\lambda_1, \ldots, \lambda_p)$ where $\lambda_1 > 0$ and $\lambda_1 \geq \ldots, \lambda_p \geq 0$ the sorted ℓ_1 norm J_λ is defined as follows:

$$\forall x \in \mathbb{R}^p, J_\lambda(x) = \lambda_1 |x|_{\downarrow 1} + \cdots + \lambda_p |x|_{\downarrow p}$$

where $|x|_{\downarrow 1} \geq \cdots \geq |x|_{\downarrow p}$ are the sorted components of x with respect to the absolute value. A proof that J_λ is a norm is given in Bogdan et al. (2015)[1]. The proximal operator of the sorted ℓ_1 norm is defined as follows:

$$\forall y \in \mathbb{R}^p, \text{prox}_{J_\lambda}(y) = \underset{x\in\mathbb{R}^p}{\text{argmin}} \ \frac{1}{2}\|y - x\|_2^2 + J_\lambda(x).$$

We remind the reader of the definition on sub-gradient and sub-differential. The following can, for instance, be found in Hiriart-Urruty and Lemaréchal (2004).

[1] An alternative proof relies on the following identity:

$$J_\lambda(x) = \lambda_p \sum_{i=1}^{p} |x|_{\downarrow i} + (\lambda_{p-1} - \lambda_p) \sum_{i=1}^{p-1} |x|_{\downarrow i} + \cdots + (\lambda_1 - \lambda_2)|x|_{\downarrow 1};$$

since the sum of the k largest components of x in absolute value: $x \in \mathbb{R}^p \mapsto \sum_{i=1}^{k} |x|_{\downarrow i}$ is a norm (named the $k-$norm) and since $\lambda_1 \geq \cdots \geq \lambda_p$ and $\lambda_1 > 0$, J_λ is a non-negative combination of norms (with at least one positive coefficient).

Definition 1. *For a convex function $f : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a* sub-gradient of f at $x \in \mathbb{R}^p$ *if*

$$f(z) \geq f(x) + s'(z - x) \ \ \forall z \in \mathbb{R}^p.$$

The set of all sub-gradients of f at x is called the sub-differential of f at x, *denoted by $\partial_f(x)$.*

Hereafter, we also use the following notation:

- Let $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $I \subset \{1, \ldots, p\}$, the writing x_I represents the vectors $(x_i)_{i \in I}$. Moreover, the notation $x_{\downarrow 1} \geq \cdots \geq x_{\downarrow p}$ represents sorted components of x.
- The notation $\mathcal{S}_p$ represents the group of permutations in $\{1, \ldots, p\}$.
- Given a set A, the notation $\text{conv}(A)$ represents the convex hull of A.

2 Proximal operator for the sorted ℓ_1 norm

Given an orthogonal transformation (for all $u, v \in \mathbb{R}^p$, $u'v = \psi(u)'\psi(v)$ or equivalently $\psi'\psi = I$) such that whatever $u \in \mathbb{R}^p$, $J_\lambda(\psi(u)) = J_\lambda(u)$ then one may prove the following identity:

$$\text{prox}_{J_\lambda}(y) = \psi'(\text{prox}_{J_\lambda}(\psi(y))).$$

One may observe that the orthogonal transformation $\psi(x) = (\epsilon_1 x_{\pi(1)}, \ldots, \epsilon_p x_{\pi(p)})$ where π is a permutation of $\{1, \ldots, p\}$ and $\epsilon_1, \ldots, \epsilon_p \in \{-1, 1\}$ is also an isometry for the sorted ℓ_1 norm (independently of λ). Specifically when $|y|_\downarrow = \psi(y)$ then one may obtain the following equality:

$$\text{prox}_{J_\lambda}(y) = \psi'(\text{prox}_{J_\lambda}(|y|_\downarrow)).$$

Consequently, in Proposition 1, one may restrict our statement to the particular case where $y_1 \geq \cdots \geq y_p \geq 0$ as already pointed out by Bogdan et al. (2015).

Proposition 1. *Let $y \in \mathbb{R}^p$ such that $y_1 \geq \cdots \geq y_p \geq 0$, $\lambda \in \mathbb{R}^p$ such that $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Using the Cesàro sequence $(C_j)_{1 \leq j \leq p}$ where $C_j = \frac{1}{j} \sum_{i=1}^{j} (y_i - \lambda_i)$, one may compute explicitly the first components of $prox_{J_\lambda}(y)$. Specifically, let $k \in \{1, \ldots, p\}$ be the largest integer for which the Cesàro sequence reaches its maximum. Then, the proximal operator satisfies the following formula:*

$$prox_{J_\lambda}(y) = \begin{cases} (0, \ldots, 0) \text{ if } C_k \leq 0 \\ (C_k, \ldots, C_k, prox_{J_{\lambda_{k+1}, \ldots, \lambda_p}}(y_{k+1}, \ldots, y_p)) \text{ otherwise} \end{cases} . \tag{1}$$

It follows that it can be implemented recursively in a very easy (and naive) way. For saving computational time, one may use a screening rule for the proximal operator before computing formula (1) (or before using any other algorithm computing the proximal operator of the sorted ℓ_1 norm). For instance, one may use the "step-up rule" described in Proposition 2 to discard some null components of $x^* = \text{prox}_{J_\lambda}(y)$. Specifically, when $|y|_{\downarrow p} \leq \lambda_p, \ldots, |y|_{\downarrow i} \leq \lambda_i$ then the last $p + 1 - i$ components of $\text{prox}_{J_\lambda}(|y|_\downarrow)$ are null.

2.1 Proof of Proposition 1

Let $\lambda \in \mathbb{R}^p$ such that $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Sub-differential calculus of the sorted ℓ_1 norm satisfies the following properties given in Propositions 5 and 8 in Schneider and Tardivel (2020) and in Lemma A.2 in Tardivel et al. (2020).

Sub-differential at 0: signed permutahedron The following equality holds:

$$\partial_{J_\lambda}(0) = \text{conv}((\sigma_1\lambda_{\pi(1)}, \dots, \sigma_p\lambda_{\pi(p)}), \sigma_1, \dots, \sigma_p \in \{-1, 1\}, \pi \in S_p).$$

The V-polytope $P^{\pm}(\lambda_1, \dots, \lambda_p) := \text{conv}((\sigma_1\lambda_{\pi(1)}, \dots, \sigma_p\lambda_{\pi(p)}), \sigma_1, \dots, \sigma_p \in \{-1, 1\}, \pi \in S_p)$, so called signed permutahedron can be described as a H-polytope as follows:

$$P^{\pm}(\lambda_1, \dots, \lambda_p) = \left\{ x \in \mathbb{R}^p : \forall j \in \{1, \dots, p\}, \sum_{i=1}^{j} |x|_{\downarrow i} \leq \sum_{i=1}^{j} \lambda_i \right\}.$$

This polytope is actually the unit ball of the dual sorted ℓ_1 norm (Brzyski, 2015; Negrinho and Martins, 2014). The above H-description of the signed permutahedron is also given in Godland and Kabluchko (2020). Finally, for any $x \in \mathbb{R}^p$, $\partial_{J_\lambda}(x) \subset P^{\pm}(\lambda_1, \dots, \lambda_p)$ (this fact is also reminded and proved in Lemma 2).

Sub-differential at a constant vector: permutahedron Let $c > 0$ then the following equality holds:

$$\partial_{J_\lambda}(c, \dots, c) = \text{conv}((\lambda_{\pi(1)}, \dots, \lambda_{\pi(p)}), \pi \in S_p)$$

The V-polytope $P(\lambda_1, \dots, \lambda_p) := \text{conv}((\lambda_{\pi(1)}, \dots, \lambda_{\pi(p)}), \pi \in S_p)$, so called permutahedron can be described as an H-polytope as follows:

$$P(\lambda_1, \dots, \lambda_p) = \left\{ x \in \mathbb{R}^p : \sum_{i=1}^{p} x_i = \sum_{i=1}^{p} \lambda_i \text{ and } \forall j \in \{1, \dots, p-1\}, \sum_{i=1}^{j} x_{\downarrow i} \leq \sum_{i=1}^{j} \lambda_i \right\}.$$

For the H-description of the permutahedron, one may see Negrinho and Martins (2014) or Godland and Kabluchko (2020) and references therein.

Computation rule for sub-differential calculus Let $x \in \mathbb{R}^p$ such that $x_1 \geq \dots \geq x_k > x_{k+1} \geq \dots \geq x_p \geq 0$, $I = \{1, \dots, k\}$ and $\overline{I} = \{k+1, \dots, p\}$ then

$$\partial_{J_\lambda}(x) = \partial_{J_{\lambda_I}}(x_I) \times \partial_{J_{\lambda_{\overline{I}}}}(x_{\overline{I}}). \tag{2}$$

Proposition 1 is a consequence of Lemma 1 below, which mainly reminds some results hidden in Tardivel et al. (2020). Proofs are shortened and, contrarily to the seminal article of Tardivel et al. (2020), the closed-form formula for the proximal operator of SLOPE is derived from H-descriptions of permutahedron and signed permutahedron.

Lemma 1. *Let $y \in \mathbb{R}^p$ such that $y_1 \geq \dots \geq y_p \geq 0$ and let f be the following function*

$$f : x \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - x\|_2^2 + J_\lambda(x).$$

Let $(C_j)_{1 \leq j \leq p}$ be the Cesàro sequence defined by $C_j := \frac{1}{j}\sum_{i=1}^{j}(y_i - \lambda_i)$ then the following properties hold:

i) The unique minimizer of f, denoted x^*, satisfies $x_1^* \geq \dots \geq x_p^* \geq 0$.

ii) The unique minimizer of f is $x^* = (0, \dots, 0)$ if and only if the Cesàro sequence is non-positive.

iii) If the unique minimizer x^* of f satisfies $x_1^* = \dots = x_p^* = c > 0$ then $C_p = c$ and the Cesàro sequence reaches its maximum at p. Conversely if the Cesàro sequence reaches its maximum at p and $C_p > 0$ then $x^* = (C_p, \dots, C_p)$.

iv) If the unique minimizer x^* of f satisfies $x_1^* = \cdots = x_k^* = c > x_{k+1}^* \geq \cdots \geq x_p^* \geq 0$ then $c = C_k$ and the largest integer for which the Cesàro reaches its maximum is k. Conversely, if the largest integer for which the Cesàro reaches its maximum is $k < p$ and $C_k > 0$ then $x_1^* = \cdots = x_k^* = C_k > x_{k+1}^* \geq \cdots \geq x_p^* \geq 0$.

v) If the unique minimizer x^* of f satisfies $x_1^* = \cdots = x_k^* > x_{k+1}^* \geq \cdots \geq x_p^* \geq 0$ then $x_{\overline{I}}^*$, where $\overline{I} = \{k+1, \ldots, p\}$, is the minimizer of the function

$$f_{\overline{I}} : x \in \mathbb{R}^{p-k} \mapsto \frac{1}{2}\|y_{\overline{I}} - x\|_2^2 + J_{\lambda_{\overline{I}}}(x).$$

Proof. i) We remind the proof already given Bogdan et al. (2015). Let $x \in \mathbb{R}^p$. Let us prove the following inequality

$$\frac{1}{2}\|y - x\|_2^2 - J_\lambda(x) \geq \frac{1}{2}\|y - |x|_\downarrow\|_2^2 - J_\lambda(|x|_\downarrow).$$

Since $J_\lambda(x) = J_\lambda(|x|_\downarrow)$ and $\|x\|_2^2 = \||x|_\downarrow\|_2$ we have

$$\begin{aligned} & \frac{1}{2}\|y - x\|_2^2 + J_\lambda(x) \geq \frac{1}{2}\|y - |x|_\downarrow\|_2^2 + J_\lambda(|x|_\downarrow), \\ \Leftrightarrow \quad & \|y - x\|_2^2 \geq \|y - |x|_\downarrow\|_2^2, \\ \Leftrightarrow \quad & x'y \leq |x|_\downarrow' y. \end{aligned}$$

Now, clearly $x'y \leq |x|'y$ where $|x| = (|x_1|, \ldots, |x_p|)$. Finally, due to the Hardy-Littlewood-Pólya rearrangement inequality one may deduce that $|x|'y \leq |x|_\downarrow' y$. Therefore the minimizer x^* of f satisfies $f(x^*) \geq f(|x^*|_\downarrow)$. Since this minimizer is unique, one may deduce that $x^* = |x^*|_\downarrow$ and thus $x_1^* \geq \cdots \geq x_p^* \geq 0$.

ii) The minimizer of f is 0 if and only if $y - 0 \in \partial_{J_\lambda}(0) = P^{\pm}(\lambda_1, \ldots, \lambda_p)$. Due to H-description of the signed permutahedron, one may deduce the following equivalences:

$$y \in P^{\pm}(\lambda_1, \ldots, \lambda_p) \Leftrightarrow \forall j \in \{1, \ldots, p\}, \sum_{i=1}^{j} y_i \leq \sum_{i=1}^{j} \lambda_i \Leftrightarrow \forall j \in \{1, \ldots, p\}, C_j \leq 0.$$

iii) If x^* satisfies $x_1^* = \cdots = x_p^* = c > 0$ then $y - x^* \in \partial_{J_\lambda}(x^*) = P(\lambda_1, \ldots, \lambda_p)$. Consequently, due to the H-description of the permutahedron, one may deduce the following inequalities:

$$\begin{aligned} y - x^* \in P(\lambda_1, \ldots, \lambda_p) \quad \Rightarrow \quad & \begin{cases} \sum_{i=1}^{p}(y_i - x_i^*) = \sum_{i=1}^{p}(y_i - c) = \sum_{i=1}^{p} \lambda_i \text{ and} \\ \forall j < p, \sum_{i=1}^{j}(y - x^*)_{\downarrow i} = \sum_{i=1}^{j}(y_i - c) \leq \sum_{i=1}^{j} \lambda_i \end{cases}, \\ \Rightarrow \quad & c = \frac{\sum_{i=1}^{p}(y_i - \lambda_i)}{p} = C_p \text{ and } \forall j < p, \frac{\sum_{i=1}^{j}(y_i - \lambda_i)}{j} = C_j \leq c. \end{aligned}$$

Consequently, $C_p = c$ and the Cesàro sequence reaches its maximum at p.

Conversely, when the Cesàro sequence reaches its maximum at p and $C_p > 0$, let us show that $x^* = (C_p, \ldots, C_p)$. In other words, we have to prove that $y - x^* \in P(\lambda_1, \ldots, \lambda_p)$ where $x^* = (C_p, \ldots, C_p)$. One checks hereafter that $y - x^*$ satisfies inequalities of the permutahedron:

$$\begin{aligned} \sum_{i=1}^{p}(y_i - C_p) \quad &= \quad \sum_{i=1}^{p} y_i - \sum_{i=1}^{p}(y_i - \lambda_i) = \sum_{i=1}^{p} \lambda_i \text{ and,} \\ \forall j \in \{1, \ldots, p-1\}, \sum_{i=1}^{j}(y - x^*)_{\downarrow i} = \sum_{i=1}^{j}(y_i - C_p) \quad &\leq \quad \sum_{i=1}^{j}(y_i - C_j) = \sum_{i=1}^{j} y_i - \sum_{i=1}^{j}(y_i - \lambda_i) = \sum_{i=1}^{j} \lambda_i. \end{aligned}$$

iv) If the minimizer x^* of f satisfies $x_1^* = \cdots = x_k^* = c > x_{k+1}^* \geq \cdots \geq x_p^* \geq 0$ then

$$y - x^* \in \partial_{J_\lambda}(x^*) = P(\lambda_1, \ldots, \lambda_k) \times \partial_{J_{\lambda_{\overline{I}}}}(x_{\overline{I}}^*).$$

Let $I = \{1, \ldots, k\}$. Since $y_I - x_I^* \in P(\lambda_1, \ldots, \lambda_k)$, due to the H-description of the permutahedron, one may deduce the following inequalities:

$$\begin{aligned} y_I - x_I^* \in P(\lambda_1, \ldots, \lambda_k) \quad &\Rightarrow \quad \begin{cases} \sum_{i=1}^k (y_i - x_i^*) = \sum_{i=1}^k (y_i - c) = \sum_{i=1}^k \lambda_i \text{ and} \\ \forall j < k, \sum_{i=1}^j (y - x^*)_{\downarrow i} = \sum_{i=1}^j (y_i - c) \leq \sum_{i=1}^j \lambda_i \end{cases}, \\ &\Rightarrow \quad c = \frac{\sum_{i=1}^k (y_i - \lambda_i)}{k} = C_k \text{ and } \forall j < k, \frac{\sum_{i=1}^j (y_i - \lambda_i)}{j} = C_j \leq c. \end{aligned}$$

Consequently, $C_k = c$. Now, let us achieve to prove that the Cesàro sequence reaches its maximum at k. Because $y - x^*$ is an element of the signed permutahedron, one may deduce the following inequalities:

$$\forall j > k, \sum_{i=1}^j y_i - x_i^* \leq \sum_{i=1}^j |y - x^*|_{\downarrow i} \leq \sum_{i=1}^j \lambda_i \Rightarrow \forall j > k, \frac{\sum_{i=1}^j y_i - \lambda_i}{j} = C_j \leq \frac{\sum_{i=1}^j x_i^*}{j} < c.$$

Conversely, when the largest integer for which the Cesàro sequence reaches its maximum is k where $k < p$ and $C_k > 0$ then let us prove that $x_1^* = \cdots = x_k^* = C_k > x_{k+1}^* \geq \cdots \geq x_p^*$. Because $C_k > 0$ and $k < p$, according to ii) and iii), one may deduce that $x^* \neq 0$ and x^* is not constant. In addition, because components of x^* are decreasing and non-negative, one may deduce that, there exists an integer $l \in \{1, \ldots, p\}$ such that $x_1^* = \cdots = x_l^* > x_{l+1} \geq \cdots \geq x_p^* \geq 0$. The first part of the proof shows that the largest integer for which the Cesàro sequence reaches its maximum is l and $x_1^* = \cdots = x_l^* = C_l$. Consequently $k = l$ and $x_1^* = \cdots = x_k^* = C_k > x_{k+1}^* \geq \cdots \geq x_p^* \geq 0$.

v) Let $I = \{1, \ldots, k\}$ and let us remind that $\overline{I} = \{k+1, \ldots, p\}$. Because the minimizer of f satisfies $x_1^* = \cdots = x_k^* > x_{k+1}^* \geq \cdots \geq x_p^* \geq 0$, according to (2), one may deduce that

$$y - x^* \in \partial_{J_\lambda}(x^*) = \partial_{J_{\lambda_I}}(x_I^*) \times \partial_{J_{\lambda_{\overline{I}}}}(x_{\overline{I}}^*) \Rightarrow y_{\overline{I}} - x_{\overline{I}}^* \in \partial_{J_{\lambda_{\overline{I}}}}(x_{\overline{I}}^*).$$

The last belonging implies that $x_{\overline{I}}^*$ is a minimizer of $f_{\overline{I}}$.

□

3 Comparison between procedures based on the ordinary least squares estimators and procedures based on SLOPE

In this section we consider a linear regression model $Y = X\beta + \varepsilon$ where $X \in \mathbb{R}^{n \times p}$ is an orthogonal matrix, $\beta \in \mathbb{R}^p$ is an unknown parameter and $\varepsilon \in \mathbb{R}^p$ is a random noise (for instance, one may assume that ε has iid $\mathcal{N}(0, \sigma^2)$ components).

Since X is an orthogonal matrix, SLOPE estimator solution of (1) satisfies $\hat{\beta}^{\text{slope}} = \text{prox}_{J_\lambda}(\hat{\beta}^{\text{ols}})$ and moreover $|\hat{\beta}^{\text{slope}}|_\downarrow = \text{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_\downarrow)$ where $\hat{\beta}^{\text{ols}} = X'Y$. Some components of SLOPE are exactly equal to 0 and thus one may provide a testing procedure based on SLOPE by rejecting the null hypothesis $\mathcal{H}_i^0 : \beta_i = 0$ (for some $i \in \{1, \ldots, p\}$) when $\hat{\beta}_i^{\text{slope}} \neq 0$ (Bogdan et al., 2015; Kos and Bogdan, 2020). Actually, the multiple testing procedure based on SLOPE rejects at least k null hypotheses (associated to the $k-$largest components of $\hat{\beta}^{\text{ols}}$ in absolute value) when $|\hat{\beta}^{\text{slope}}|_{\downarrow k} > 0$ and this procedure rejects exactly k null hypotheses when $|\hat{\beta}^{\text{slope}}|_{\downarrow k} > 0$ and $|\hat{\beta}^{\text{slope}}|_{\downarrow k+1} = 0$.

Proposition 2 is useful to compare multiple testing procedures based on $\hat{\beta}^{\text{ols}}$ with procedures based on SLOPE.

Proposition 2. *Let $\lambda = (\lambda_1, \ldots, \lambda_p)$ where $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p$ and let us remind that, in the orthogonal setting, $|\hat{\beta}^{\text{slope}}|_{\downarrow} = \text{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_{\downarrow})$. The following implication occurs:*

$$|\hat{\beta}^{\text{ols}}|_{\downarrow 1} > \lambda_1, \ldots, |\hat{\beta}^{\text{ols}}|_{\downarrow k} > \lambda_k \quad \Rightarrow \quad |\hat{\beta}^{\text{slope}}|_{\downarrow k} > 0.$$

According to the above implication, a step-down procedure based on the ordinary least squares estimator and thresholds $\lambda_1, \ldots, \lambda_p$ is less powerful than a procedure based on SLOPE.

The following implication occurs:

$$|\hat{\beta}^{\text{slope}}|_{\downarrow k} > 0 \quad \Rightarrow \quad \exists i \geq k, |\hat{\beta}^{\text{ols}}|_{\downarrow i} > \lambda_i.$$

According to the above implication, a step-up procedure based on the ordinary least squares estimator and thresholds $\lambda_1, \ldots, \lambda_p$ is at least as powerful than a procedure based on SLOPE.

For instance, let us chose the hyper-parameter λ as in the Holm's procedure (Holm, 1979) and Hochberg's procedure (Hochberg, 1988):

$$\lambda_1 = \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right), \lambda_2 = \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2(p-1)}\right), \ldots, \lambda_p = \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Then, according to Proposition 2, the procedure based on SLOPE is at least as powerful than the Holm step-down procedure but less powerful than the Hochberg step-up procedure. Otherwise, according to the second implication, when the hyper-parameter for SLOPE is given by the BH sequence (namely $\lambda_1 = \sigma\Phi^{-1}(1-\alpha/2p), \ldots, \lambda_p = \sigma\Phi^{-1}(1-\alpha/2)$) then, the procedure based on SLOPE is less powerful than the (step-up) Benjamini-Hochberg's multiple testing procedure. In the seminal article on SLOPE (Bogdan et al., 2015) one may find the following comment: "The procedure based on SLOPE is sandwiched between the step-down and step-up procedures in the sense that it rejects at most as many hypotheses as the step-up (Benjamini-Hochberg) procedure and at least as many as the step-down cousin, also known to control the FDR (Sarkar, 2002)". Thus Proposition 2 gives a proof for this (unproven) comment.

Whereas $|\hat{\beta}^{\text{slope}}|_{\downarrow} = \text{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_{\downarrow})$, one does not need to compute explicitly $\text{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_{\downarrow})$ to determine null hypotheses rejected by the SLOPE procedure. Actually, the number of null hypotheses rejected by this procedure coincides with the number of non-null components of $\text{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_{\downarrow})$. Proposition 3 gives a simple analytical shortcut to compute exactly the number of null components for the proximal operator of the sorted ℓ_1 norm.

Proposition 3. *Let $y \in \mathbb{R}^p$ such that $y_1 \geq \cdots \geq y_p \geq 0$, $\lambda \in \mathbb{R}^p$ such that $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Using the remainder sequence $(R_j)_{1 \leq j \leq p}$ where $R_j = \sum_{i=j}^{p}(y_i - \lambda_i)$, one may compute explicitly the number of null components of $x^* = prox_{J_\lambda}(y)$.*

i) The vector x^* is positive (component-wise) if and only if $(R_j)_{1 \leq j \leq p}$ is a positive sequence.

ii) If x^* is not positive then $k_0 := \min\{i \in \{1, \ldots, p\} : x_i^* = 0\}$ is the smallest integer for which $(R_j)_{1 \leq j \leq p}$ reaches its minimum and $R_{k_0} \leq 0$. Conversely, if the smallest integer for which $(R_j)_{1 \leq j \leq p}$ reaches its minimum is k_0 and $R_{k_0} \leq 0$ then $k_0 := \min\{i \in \{1, \ldots, p\} : x_i^* = 0\}$.

Note that Lemma 3.1 in Bogdan et al. (2013) also gives a technical result on the value k_0 described in Proposition 3. However, contrarily to Proposition 3, Lemma 3.1 does not provide a shortcut to compute explicitly the number of non-null components of the proximal operator.

3.1 Proof of Proposition 2

The proof of Proposition 2 is based on two inclusions given in Lemma 2.

Lemma 2. *Let $\lambda \in \mathbb{R}^p$ such that $\lambda_1 > 0$ and $\lambda_1 \geq \cdots \geq \lambda_p$.*

i) Let $x \in \mathbb{R}^p$ then $\partial_{J_\lambda}(x) \subset \partial_{J_\lambda}(0) = P^{\pm}(\lambda_1, \ldots, \lambda_p)$.

ii) Let $x \in \mathbb{R}^p$ such that $x_1 > 0, \ldots, x_p > 0$ then $\partial_{J_\lambda}(x) \subset P(\lambda_1, \ldots, \lambda_p)$.

Proof. The proof of i) is already given in Proposition 5 of Schneider and Tardivel (2020). However, we provide hereafter a proof based on the H-description of the signed permutahedron. Let $s \in \partial_{J_\lambda}(x)$ and let π a permutation such that $|s_{\pi(1)}| \geq \cdots \geq |s_{\pi(p)}|$. Let $h = \sum_{i=1}^{k} \operatorname{sign}(s_{\pi(i)}) e_{\pi(i)}$ where $e_1, \ldots, e_p$ is the canonical basis of $\mathbb{R}^p$ and $k \leq p$. Because J_λ is a norm and s is a sub-gradient, the following inequality occurs:

$$J_\lambda(x) + \sum_{i=1}^{k} \lambda_i = J_\lambda(x) + J_\lambda(h) \geq J_\lambda(x+h) \geq J_\lambda(x) + \sum_{i=1}^{k} s_{\pi(i)} \operatorname{sign}(s_{\pi(i)}) = J_\lambda(x) + \sum_{i=1}^{k} |s|_{\downarrow i}.$$

Consequently, whatever $k \in \{1, \ldots, p\}, \sum_{i=1}^{k} |s|_{\downarrow i} \leq \sum_{i=1}^{k} \lambda_i$. Thus, $s \in P^{\pm}(\lambda_1, \ldots, \lambda_p)$.

ii) Let $s \in \partial_{J_\lambda}(x)$, $\mathbf{1} = (1, \ldots, 1)$ and $\eta \in \mathbb{R}$. Let us illustrate that s satisfies the H-description of the permutahedron. Since s is a sub-gradient of J_λ at x then by definition we have $J_\lambda(x + \eta\mathbf{1}) \geq J_\lambda(x) + \eta s'\mathbf{1} = J_\lambda(x) + \eta \sum_{i=1}^{p} s_i$. Now, when η is small enough the vector $x + \eta\mathbf{1}$ is component-wise positive and clearly $(x + \eta\mathbf{1})_\downarrow = ((x + \eta\mathbf{1})_{\downarrow 1}, \ldots, (x + \eta\mathbf{1})_{\downarrow p}) = (x_{\downarrow 1} + \eta, \ldots, x_{\downarrow p} + \eta)$. Consequently, the following inequality occurs:

$$J_\lambda(x + \eta\mathbf{1}) = \sum_{i=1}^{p} \lambda_i (x_{\downarrow i} + \eta) = J_\lambda(x) + \eta \sum_{i=1}^{p} \lambda_i \geq J_\lambda(x) + \eta \sum_{i=1}^{p} s_i.$$

Taking $\eta > 0$ (resp. $\eta < 0$) small enough in equation (3.1), one may derive that $\sum_{i=1}^{p} \lambda_i \geq \sum_{i=1}^{p} s_i$ (resp. $\sum_{i=1}^{p} \lambda_i \leq \sum_{i=1}^{p} s_i$) implying thus $\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} s_i$. Let π be a permutation in $\{1, \ldots, p\}$ such that $s_{\pi(1)} \geq \cdots \geq s_{\pi(p)}$ and let $h = \sum_{i=1}^{k} e_{\pi(i)}$ where $k < p$. Because J_λ is a norm and s is a sub-gradient, the following inequality occurs:

$$J_\lambda(x) + \sum_{i=1}^{k} \lambda_i = J_\lambda(x) + J_\lambda(h) \geq J_\lambda(x+h) \geq J_\lambda(x) + \sum_{i=1}^{k} s_{\pi(i)}$$

Consequently, whatever $k \in \{1, \ldots, p-1\}$, $\sum_{i=1}^{k} s_{\downarrow i} \leq \sum_{i=1}^{k} \lambda_i$, which achieves to prove that $s \in P(\lambda_1, \ldots, \lambda_p)$.

□

Hereafter, we provide the proof of Proposition 2.

Proof. Let us prove the first implication by contradiction. If $|\hat{\beta}^{\text{slope}}|_{\downarrow k} = 0$ then $k_0 := \min\{i \in \{1, \ldots, p\} : |\hat{\beta}^{\text{slope}}|_{\downarrow i} = 0\}$ clearly satisfies $k_0 \leq k$. Because $|\hat{\beta}^{\text{slope}}|_\downarrow = \operatorname{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_\downarrow)$ then, using sub-differential rule described in (2) and Lemma 2, one may deduce the following inclusion:

$$|\hat{\beta}^{\text{ols}}|_\downarrow - |\hat{\beta}^{\text{slope}}|_\downarrow \in \partial_{J_\lambda}(|\hat{\beta}^{\text{slope}}|_\downarrow) \subset \begin{cases} P^{\pm}(\lambda_1, \ldots, \lambda_p) \text{ if } k_0 = 1 \\ P(\lambda_1, \ldots, \lambda_{k_0 - 1}) \times P^{\pm}(\lambda_{k_0}, \ldots, \lambda_p) \text{ if } k_0 > 1 \end{cases}$$

Since a vector in $P^{\pm}(\lambda_{k_0}, \ldots, \lambda_p)$ has its first component smaller in absolute value than λ_{k_0} one may deduce that $|\hat{\beta}^{\text{ols}}|_{\downarrow k_0} \leq \lambda_{k_0}$ which contradicts that $|\hat{\beta}^{\text{ols}}|_{\downarrow 1} > \lambda_1, \ldots, |\hat{\beta}^{\text{ols}}|_{\downarrow k} > \lambda_k$.

Let us prove the second implication. If $|\hat{\beta}^{\text{slope}}|_{\downarrow k} > 0$ then $k_0 := \max\{i \in \{1, \ldots, p\} : |\hat{\beta}^{\text{slope}}|_{\downarrow i} > 0\}$ clearly satisfies $k_0 \geq k$. Using the same argument as above one may deduce the following inclusion:

$$|\hat{\beta}^{\text{ols}}|_\downarrow - |\hat{\beta}^{\text{slope}}|_\downarrow \in \partial_{J_\lambda}(|\hat{\beta}^{\text{slope}}|_\downarrow) \subset \begin{cases} P(\lambda_1, \ldots, \lambda_p) \text{ if } k_0 = p \\ P(\lambda_1, \ldots, \lambda_{k_0}) \times P^{\pm}(\lambda_{k_0 + 1}, \ldots, \lambda_p) \text{ if } k_0 < p \end{cases}$$

Since a vector in $P(\lambda_1,\dots,\lambda_{k_0})$ has its last component greater than or equal to λ_{k_0}, one may deduce that $|\hat{\beta}^{\text{ols}}|_{\downarrow k_0} - |\hat{\beta}^{\text{slope}}|_{\downarrow k_0} \geq \lambda_{k_0}$ and thus $|\hat{\beta}^{\text{ols}}|_{\downarrow k_0} > \lambda_{k_0}$.

□

3.2 Proof of Proposition 3

Proof. Before to give the proof of i), we provide another H-description of the permutahedron (Ziegler, 2012; Simion, 1997):

$$P(\lambda_1,\dots,\lambda_p) = \left\{ x \in \mathbb{R}^p : \sum_{i=1}^p x_i = \sum_{i=1}^p \lambda_i \text{ and } \forall j \in \{2,\dots,p\}, \sum_{i=j}^p x_{\downarrow i} \geq \sum_{i=j}^p \lambda_i \right\}.$$

Actually, one may easily check that the above H-description of the permutahedron is equivalent to the one given in section 2.1.

i) If $x_1^* \geq \cdots \geq x_p^* > 0$ then, according to Lemma 2, $\partial_{J_\lambda}(x^*) \subset P(\lambda_1,\dots,\lambda_p)$ and thus $y - x^* \in \partial_{J_\lambda}(x^*)$ satisfies the following inequalities

$$\begin{aligned} & \forall j \in \{1,\dots,p\}, \sum_{i=j}^p (y_i - x_i^*) \geq \sum_{i=j}^p (y - x^*)_{\downarrow i} \geq \sum_{i=j}^p \lambda_i \\ \Rightarrow \quad & \forall j \in \{1,\dots,p\}, R_j = \sum_{i=j}^p (y_i - \lambda_i) \geq \sum_{i=j}^p x_i^* > 0. \end{aligned}$$

Conversely, if x^* is not positive component-wise then $k_0 := \min\{k \in \{1,\dots,p\} : x_k^* = 0\}$ is well defined. Since $x_{k_0-1}^* > x_{k_0}^* = \cdots = x_p^* = 0$ one may deduce that $(y_{k_0},\dots,y_p) \in P^{\pm}(\lambda_{k_0},\dots,\lambda_p)$ and thus the following implication holds

$$\sum_{i=k_0}^p |y|_{\downarrow i} \leq \sum_{i=k_0}^p \lambda_i \Rightarrow R_{k_0} = \sum_{i=k_0}^p (y_i - \lambda_i) \leq 0.$$

ii) If x^* is not positive then let us show that k_0, defined above, is the smallest integer for which $(R_j)_{1\leq j\leq p}$ reaches its minimum. As already claimed above, $(y_{k_0},\dots,y_p) \in P^{\pm}(\lambda_{k_0},\dots,\lambda_p)$ and thus the following inequalities occur:

$$\forall j \geq k_0, R_{k_0} - R_j = \begin{cases} \sum_{i=k_0}^{j-1}(y_i - \lambda_i) = \sum_{i=k_0}^{j-1} |y|_{\downarrow i} - \sum_{i=k_0}^{j-1} \lambda_i \text{ if } j > k_0 \\ 0 \text{ if } j = k_0 \end{cases} \Rightarrow \forall j \geq k_0, R_{k_0} \leq R_j.$$

Therefore, once $k_0 = 1$ then k_0 is clearly the smallest integer for which the remaining sequence reaches its minimum. Otherwise, when $k_0 > 1$ since $x_1^* \geq \cdots \geq x_{k_0-1}^* > x_{k_0}^* = 0$ then, according to the sub-differential rule (2) and Lemma 2, we have $(y_1 - x_1^*,\dots,y_{k_0-1} - x_{k_0-1}^*) \in P(\lambda_1,\dots,\lambda_{k_0-1})$. Consequently, the following implication occurs

$$\begin{aligned} & \forall j \in \{1,\dots,k_0-1\}, \sum_{i=j}^{k_0-1} (y_i - x_i^*) \geq \sum_{i=j}^{k_0-1} (y - x^*)_{\downarrow i} \geq \sum_{i=j}^{k_0-1} \lambda_i \\ \Rightarrow \quad & \forall j \in \{1,\dots,k_0-1\}, \sum_{i=j}^{k_0-1} (y_i - \lambda_i) \geq \sum_{i=j}^{k_0-1} x_i^* > 0. \end{aligned}$$

Thus, according to the right-hand side, $R_j > R_{k_0}$ once $j < k_0$. Therefore k_0 is the smallest integer for which the remaining sequence reaches its minimum.

Conversely, if the remaining sequence reaches its minimum at k_0 and $R_{k_0} \leq 0$ thus according to i) x^* is not positive component-wise thus $k_1 := \min\{i \in \{1, \dots, p\} : x_i^* = 0\}$ is well defined. Finally, the first part of the proof of ii) shows that k_1 is the smallest integer for which the remaining sequence reaches its minimum and thus $k_0 = k_1$.

□

Aknowledgements

We would like to thank Małgorzata Bogdan for her insightful comments on the paper. This work has been supported by the EIPHI Graduate School (contract ANR-17-EURE-0002).

References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57 (1):289–300, 1995.

Malgorzata Bogdan, Ewout van den Berg, Weijie Su, and Emmanuel Candes. Statistical estimation and testing via the sorted l1 norm. *arXiv preprint arXiv:1310.1969*, 2013.

Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3): 1103, 2015.

Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Damian Brzyski. *Selecting relevant groups of explanatory variables viaconvex optimization methods with the false discovery rate control*. PhD thesis, Wrocław university of technology, 2015.

Zhiqi Bu, Jason M Klusowski, Cynthia Rush, and Weijie J Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537, 2020.

Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Thomas Godland and Zakhar Kabluchko. Projections and angle sums of permutohedra and other polytopes. *arXiv preprint arXiv:2009.04186*, 2020.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.

Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4): 800–802, 1988.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Michał Kos and Małgorzata Bogdan. On the asymptotic properties of slope. *Sankhya A*, 82(2):499–532, 2020.

Renato Negrinho and Andre Martins. Orbit regularization. *Advances in neural information processing systems*, 27:3221–3229, 2014.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Sanat K Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of statistics*, 30(1):239–257, 2002.

Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. *arXiv e-prints*, pages arXiv–2004, 2020.

Rodica Simion. Convex polytopes and enumeration. *Advances in Applied Mathematics*, 18(2):149–180, 1997.

Patrick JC Tardivel, Rémi Servien, and Didier Concordet. Simple expressions of the lasso and slope estimators in low-dimension. *Statistics*, 54(2):340–352, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Xiangrong Zeng and Mário AT Figueiredo. Decreasing weighted sorted l_1 regularization. *IEEE Signal Processing Letters*, 21(10):1240–1244, 2014.

Yiliang Zhang and Zhiqi Bu. Efficient designs of slope penalty sequences in finite dimension. *arXiv preprint arXiv:2102.07211*, 2021.

Günter M Ziegler. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.