



Proximal operator for the sorted l_1 norm with application to testing procedures based on SLOPE

Xavier Dupuis, Patrick J C Tardivel

► To cite this version:

Xavier Dupuis, Patrick J C Tardivel. Proximal operator for the sorted l_1 norm with application to testing procedures based on SLOPE. 2021. hal-03177108v1

HAL Id: hal-03177108

<https://hal.science/hal-03177108v1>

Preprint submitted on 22 Mar 2021 (v1), last revised 19 Feb 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Proximal operator for the sorted l_1 norm with application to testing procedures based on SLOPE

Xavier Dupuis and Patrick J.C. Tardivel

Institut de Mathématiques de Bourgogne, Université Bourgogne, Dijon, France.

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 corroborate that SLOPE as well as OSCAR perform clustering selection. Moreover, our geometric approach to prove the formula for the proximal operator provides insight to show that testing procedures based on SLOPE are more powerful than step-down testing procedures but less powerful than step-up testing procedures.

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

$$\hat{\beta} \in \operatorname{argmin}_{b \in \mathbb{R}^p} \frac{1}{2} \|Y - Xb\|_2^2 + \sum_{i=1}^p \lambda_i |b_i|_{\downarrow i} \text{ where } |b_i|_{\downarrow 1} \geq \dots \geq |b_i|_{\downarrow p}. \quad (1)$$

For OSCAR the hyper-parameter $\lambda = (\lambda_1, \dots, \lambda_p)$ has arithmetically decreasing components. SLOPE is both an extension of OSCAR where the hyper-parameter satisfies $\lambda_1 > 0$ and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ and an extension of the Least Absolute Shrinkage and Selection Operator (LASSO) (Chen and Donoho, 1994; Tibshirani, 1996). Indeed, when $\lambda_1 = \dots = \lambda_p > 0$ then the penalty $\sum_{i=1}^p \lambda_i |b_i|_{\downarrow i}$ coincides with the l_1 norm ; the well known penalty term for the LASSO. With respect to other penalized estimators, when $\lambda_1 > \dots > \lambda_p > 0$, 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 $\lambda_1 = \sigma \Phi^{-1}(1 - \alpha/2p), \dots, \lambda_p = \sigma \Phi^{-1}(1 - \alpha/2)$ (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-Hochberg's procedure and controlling the FDR (Benjamini and Hochberg, 1995).

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 backward forward proximal gradient

algorithm (see e.g. 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).

There is a particular case under which the proximal operator of the sorted l_1 norm is explicit; when $y_1 \geq \dots \geq y_p \geq 0$ and $y_1 - \lambda_1 \geq \dots \geq y_p - \lambda_p$ then the proximal operator is simply given by $((y_1 - \lambda_1)_+, \dots, (y_p - \lambda_p)_+)$. Except for this particular case, the proximal operator of the sorted l_1 norm was believed as tedious to compute explicitly. 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 $y_i - \lambda_i, \dots, y_j - \lambda_j$ non-decreasing and non-constant (where $i < j$) and then to substitute 1) (y_i, \dots, y_j) by $((\sum_{l=i}^j y_l)/(j+1-i), \dots, (\sum_{l=i}^j y_l)/(j+1-i))$ and 2) $(\lambda_i, \dots, \lambda_j)$ by $((\sum_{l=i}^j \lambda_l)/(j+1-i), \dots, (\sum_{l=i}^j \lambda_l)/(j+1-i))$. 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 sign permutahedron polytope (the sign 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 $\lambda = (\lambda_1, \dots, \lambda_p)$ where $\lambda_1 > 0$ and $\lambda_1 \geq \dots, \lambda_p \geq 0$ the sorted l_1 norm J_λ is defined as follows:

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

where $|x|_{\downarrow 1} \geq \dots \geq |x|_{\downarrow p}$ are sorted components of x with respect to the absolute value.

Given an sorted l_1 norm, the proximal operator is defined as follows:

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

Hereafter, we also use the following notation

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

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 quite easily prove the following identity

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

One may observe that the orthogonal transformation $\psi(x) = (\epsilon_1 x_{\pi(1)}, \dots, \epsilon_p x_{\pi(p)})$ where π is a permutation of $\{1, \dots, p\}$ and $\epsilon_1, \dots, \epsilon_p \in \{-1, 1\}$ 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 $y_1 \geq \dots \geq y_p \geq 0$ as already pointed out by Bogdan et al. (2015).

Proposition 1. Let $y \in \mathbb{R}^p$ such that $y_1 \geq \dots \geq y_p \geq 0$, $\lambda \in \mathbb{R}^p$ such that $\lambda_1 > 0$ and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$.

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 $\text{prox}_{J_\lambda}(y)$. Specifically, let $k \in \{1, \dots, p\}$ the largest integer for which the Cesàro sequence reaches its maximum. Then, the proximal operator satisfies the following formula:

$$\text{prox}_{J_\lambda}(y) = \begin{cases} (0, \dots, 0) & \text{if } C_k \leq 0 \\ (C_k, \dots, C_k, \text{prox}_{J_{\lambda_{k+1}, \dots, \lambda_p}}(y_{k+1}, \dots, y_p)) & \text{otherwise} \end{cases} \quad (2)$$

2.1 Proof of Proposition 1

Let $\lambda \in \mathbb{R}^p$ such that $\lambda_1 > 0$ and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$. 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: sign permutahedron The following inclusion 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 sign-permutahedron can be described as a H-polytope as follows:

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

For the H-description of the sign permutahedron one may see Godland and Kabluchko (2020) and references therein. Finally, whatever $x \in \mathbb{R}^p$, $\partial_{J_\lambda}(x) \subset P^\pm(\lambda_1, \dots, \lambda_p)$ (this fact is also reminded and proved in Lemma 2).

Sub-differential at a constant vector: permutahedron Let $c > 0$ then the following inclusion 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 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 $\bar{I} = \{k+1, \dots, p\}$ then

$$\partial_{J_\lambda}(x) = \partial_{J_{\lambda_I}}(x_I) \times \partial_{J_{\lambda_{\bar{I}}}}(x_{\bar{I}}). \quad (3)$$

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 sign-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.
- ii) 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 $x_1^* = \dots = x_k^* = c > x_{k+1}^* \geq \dots \geq x_p^* \geq 0$ 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 $x_1^* = \dots = x_k^* = C_k > x_{k+1}^* \geq \dots \geq x_p^* \geq 0$.
- v) If the unique minimizer x^* of f satisfies $x_1^* = \dots = x_k^* > x_{k+1}^* \geq \dots \geq x_p^* \geq 0$ then $x_{\bar{T}}^*$, where $\bar{T} = \{k+1, \dots, p\}$, is the minimizer of the function

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

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 $\|x\|_2^2 = \||x|_{\downarrow}\|_2^2$ we have

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

Now, clearly $x'y \leq |x|'y$ where $|x| = (|x_1|, \dots, |x_p|)$. Finally, due to the rearrangement inequality (Hardy, 1952) 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 $x_1^* \geq \dots \geq x_p^* \geq 0$.

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

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

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

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

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 $x^* = (C_p, \dots, C_p)$. In other words, we have to prove that $y - x^* \in P(\lambda_1, \dots, \lambda_p)$ where $x^* = (C_p, \dots, C_p)$. One checks hereafter that $y - x^*$ satisfies inequalities of the permutahedron:

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

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

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

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

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

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 sign permutahedron, one may deduce the following inequalities:

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

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 $x_1^* = \dots = x_k^* = C_k > x_{k+1}^* \geq \dots \geq x_p^*$. 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 $l \in \{1, \dots, p\}$ such that $x_1^* = \dots = x_l^* > x_{l+1}^* \geq \dots \geq x_p^* \geq 0$. The first part of the proof shows that the largest integer for which the Cesàro sequence reaches its maximum is l and $x_1^* = \dots = x_l^* = C_l$. Consequently $k = l$ and $x_1^* = \dots = x_k^* = C_k > x_{k+1}^* \geq \dots \geq x_p^* \geq 0$.

v) Let $I = \{1, \dots, k\}$ and let us remind that $\bar{I} = \{k+1, \dots, p\}$. Because the minimizer of f satisfies $x_1^* = \dots = x_k^* > x_{k+1}^* \geq \dots \geq x_p^* \geq 0$, according to (3), one may deduce that

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

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

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 (2) 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 $i \in \{1, \dots, p\}$) 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 $\lambda = (\lambda_1, \dots, \lambda_p)$ where $\lambda_1 > 0$ and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 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*

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

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

The following implication occurs

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

According to the above implication, a step-up procedure based on the ordinary least squares estimator and thresholds $\lambda_1, \dots, \lambda_p$ is more 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):

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

Then, according to Proposition 2, the procedure based on SLOPE is more 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 $\lambda_1 = \sigma\Phi^{-1}(1 - \alpha/2p), \dots, \lambda_p = \sigma\Phi^{-1}(1 - \alpha/2)$) then, the procedure based on SLOPE is less powerful than the (step-up) Benjamini-Hochberg's multiple testing procedure.

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 $\lambda_1 \geq \dots \geq \lambda_p$.*

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

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 sign permutahedron. Let $s \in \partial_{J_\lambda}(x)$ and let π a permutation such that $|s_{\pi(1)}| \geq \dots \geq |s_{\pi(p)}|$. Let $h = \sum_{i=1}^k \text{sign}(s_{\pi(i)})e_{\pi(i)}$ where e_1, \dots, 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:

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

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

ii) Let $s \in \partial_{J_\lambda}(x)$ and let $\mathbf{1} = (1, \dots, 1)$. Let us illustrate that s satisfies the H-description of the permutahedron. For $\eta \in \mathbb{R}$ small enough the vector $x + \eta \mathbf{1}$ is component-wise positive and clearly $(x + \eta \mathbf{1})_\downarrow = (x_{\downarrow 1} + \eta, \dots, x_{\downarrow p} + \eta)$. Consequently, the following inequality occurs:

$$J_\lambda(x - \eta \mathbf{1}) = \sum_{i=1}^p \lambda_i (x_{\downarrow i} + \eta) = J_\lambda(x) + \eta \sum_{i=1}^p \lambda_i \geq J_\lambda(x) + \eta \sum_{i=1}^p s_i \quad (4)$$

Taking $\eta > 0$ (resp. $\eta < 0$) small enough in equation (4), 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 $\{1, \dots, p\}$ such that $s_{\pi(1)} \geq \dots \geq s_{\pi(p)}$ 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 $k \in \{1, \dots, p-1\}$, $\sum_{i=1}^k s_{\downarrow i} \leq \sum_{i=1}^k \lambda_i$, which achieves to prove that $s \in P(\lambda_1, \dots, \lambda_p)$. □

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 $k_0 := \min\{i \in \{1, \dots, p\} : |\hat{\beta}^{\text{slope}}|_{\downarrow i} = 0\}$ clearly satisfies $k_0 \leq k$. Because $|\hat{\beta}^{\text{slope}}|_{\downarrow} = \text{prox}_{J_\lambda}(|\hat{\beta}^{\text{ols}}|_{\downarrow})$ then, using sub-differential rule described in (3) and Lemma 2, one may deduce the following inclusion:

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

Since a vector in $P^\pm(\lambda_{k_0}, \dots, \lambda_p)$ 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 $|\hat{\beta}^{\text{ols}}|_{\downarrow 1} > \lambda_1, \dots, |\hat{\beta}^{\text{ols}}|_{\downarrow k} > \lambda_k$.

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

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

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

Aknowledgements

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.
- 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.
- 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.
- Thomas Godland and Zakhar Kabluchko. Projections and angle sums of permutohedra and other polytopes. *arXiv preprint arXiv:2009.04186*, 2020.
- GH Hardy. j. e. littlewood and g. polya, inequalities. *Cambridge University Press*, 10:169, 1952.
- 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.
- Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.
- Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness and model selection of penalized estimators including slope, lasso, and basis pursuit. *arXiv preprint arXiv:2004.09106*, 2020.
- Patrick JC Tardivel, Rémi Servien, and Didier Concorde. 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.