The Solution Path of SLOPE

Patrick J C Tardivel, Xavier Dupuis

The Solution Path of SLOPE

Xavier Dupuis
Institut de Mathématiques de Bourgogne
UMR 5584 CNRS
Université de Bourgogne
F-21000 Dijon, France
xavier.dupuis@u-bourgogne.fr

Patrick Tardivel
Institut de Mathématiques de Bourgogne
UMR 5584 CNRS
Université de Bourgogne
F-21000 Dijon, France
patrick.tardivel@u-bourgogne.fr

Abstract

The SLOPE estimator has the particularity of having null components (sparsity) and components that are equal in absolute value (clustering). The number of clusters depends on the regularization parameter of the estimator. This parameter can be chosen as a trade-off between interpretability (with a small number of clusters) and accuracy (with a small mean squared error or a small prediction error). Finding such a compromise requires to compute the solution path, that is the function mapping the regularization parameter to the estimator. We provide in this article an algorithm to compute the solution path of SLOPE.

1 Introduction

The SLOPE estimator (Sorted L One Penalized Estimator [1, 21]) is defined as a solution to the following convex program:

$$\min_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \gamma \sum_{i=1}^{p} \lambda_i |b|_{\downarrow i}. \tag{1}$$

In (1), $\lambda_1 > 0$, $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ is a given sequence of penalty parameters, $\gamma > 0$ is the regularization parameter and $|b|_{\downarrow 1} \geq \cdots \geq |b|_{\downarrow p} \geq 0$ are sorted components of b in absolute value. The SLOPE estimator generalizes both the LASSO estimator (Least Absolute Shrinkage and Selection Operator [20]) for which $\lambda_1 = \cdots = \lambda_p = 1$, and the OSCAR estimator (Octagonal Shrinkage and Clustering Algorithm for Regression [3]) for which the sequence $\lambda_1, \ldots, \lambda_p$ is arithmetic.

The SLOPE estimator is gaining popularity among statisticians due to its relevant properties such as false discovery rate control [1] and dimension reduction of the regression model. The latter property comes from the structure of the solutions to the optimization problem (1), which have null components (sparsity) as well as components equal in absolute value (clustering) [14, 6, 3][1]. When y represents the random response of a linear regression model, sparsity has a well-known statistical interpretation: identification of relevant explanatory variables. Clustering also has a statistical interpretation when the design matrix X is standardized: the explanatory variables having the same regression coefficient have the same impact on the response [15]. On the other hand, without restriction on the design matrix, for a categorical variable having different levels, the equal regression coefficients represent levels that can be grouped together [17, 10]. Therefore, SLOPE estimator can identify relevant explanatory variables, group explanatory variables having the same impact on the response and, more generally, reduce the dimension of the regression model.

The solution path gives the solution of a penalized optimization problem with respect to the regularization parameter $\gamma > 0$. For the LASSO, this path shows that the number of explanatory

[1] When X is an orthogonal matrix, the solution to the problem (1) is explicit and its sparsity and clustering properties are straightforward [1, 5, 16, 19]).

Preprint. Under review.

variables selected by this estimator tends to decrease when the regularization parameter becomes large (see *e.g.* [9, 13]). Adjusting the regularization parameter γ allows a compromise between selecting a small number of explanatory variables and constructing an accurate estimator. Similarly, the construction of the solution path of SLOPE is useful to adjust the regularization parameter to have a good trade-off between interpretability (by selecting a small number of clusters of explanatory variables) and accuracy (with a small mean squared error or a small prediction error).

In this article, given a sequence of penalty parameters $\lambda_1 > \cdots > \lambda_p > 0$, we prove that the solution path of SLOPE is piecewise linear on $(0, +\infty)$, we characterize its affine components, and we provide an algorithm to compute the path.[2]

2 Basic notions on SLOPE

2.1 Sorted ℓ_1 norm and its dual norm

Definition 1 *The* sorted ℓ_1 norm *associated to* $\lambda \in \mathbb{R}^p$ *with* $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ *and* $\lambda_1 > 0$ *is defined as follows:*

$$J_\lambda(b) = \sum_{i=1}^{p} \lambda_i |b|_{\downarrow i}, \quad b \in \mathbb{R}^p,$$

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

Given a norm $\|\cdot\|$ on $\mathbb{R}^p$, we recall that its dual norm $\|\cdot\|^*$ is defined by $\|v\|^* = \max\{b'v : \|b\| \leq 1\}$, for $v \in \mathbb{R}^p$.

Remark 1 *The* dual sorted ℓ_1 norm *has an explicit expression given in [11] and reminded hereafter:*

$$J_\lambda^*(v) = \max\left\{ \frac{\|v\|_{(1)}}{\lambda_1}, \frac{\|v\|_{(2)}}{\sum_{i=1}^2 \lambda_i}, \ldots, \frac{\|v\|_{(p)}}{\sum_{i=1}^p \lambda_i} \right\}, \quad v \in \mathbb{R}^p,$$

where $\|\cdot\|_{(k)}$ *is the* $k-$*norm (the sum of the k largest components in absolute value).*

2.2 SLOPE pattern

The SLOPE pattern introduced in [14], whose definition is reminded below, is a central notion in this article.

Definition 2 *The* SLOPE pattern $\mathrm{patt}(b) \in \mathbb{Z}^p$ *of* $b \in \mathbb{R}^p$ *is defined by*

$$\mathrm{patt}(b)_i = \mathrm{sign}(b_i)\,\mathrm{rank}(|b|)_i, \quad i \in \{1, \ldots, p\},$$

where $\mathrm{rank}(|b|)_i \in \{0, 1, \ldots, k\}$*, k is the number of nonzero distinct values in* $\{|b_1|, \ldots, |b_p|\}$*,* $\mathrm{rank}(|b|)_i = 0$ *if and only if* $b_i = 0$*, and* $\mathrm{rank}(|b|)_i < \mathrm{rank}(|b|)_j$ *if* $|b_i| < |b_j|$*.*

We denote by $\mathcal{P}_p^{\mathrm{slope}} = \mathrm{patt}(\mathbb{R}^p)$ the set of SLOPE patterns. Note in the definition above that $k = \|\mathrm{patt}(b)\|_\infty$; it is the number of nonzero clusters of b.

Example 1 *Let* $b = (4.2, -1.3, 0, 1.3, 4.2)'$*. Then* $\mathrm{patt}(b) = (2, -1, 0, 1, 2)'$*.*

Definition 3 *Let* $m \in \mathbb{Z}^p$ *be a SLOPE pattern with* $k := \|m\|_\infty > 0$*. The associated* pattern matrix $U_m \in \mathbb{R}^{p \times k}$ *is defined by*

$$(U_m)_{ij} = \mathrm{sign}(m_i)\mathbf{1}_{(|m_i| = k+1-j)}, \quad i \in \{1, \ldots, p\},\ j \in \{1, \ldots, k\}.$$

For $k \geq 1$ we denote $\mathbb{R}^{k+} = \{s \in \mathbb{R}^k \colon s_1 > \ldots > s_k > 0\}$. Definition 3 is such that, for $b \in \mathbb{R}^p$ and $m \in \mathbb{Z}^p$ a SLOPE pattern with $k := \|m\|_\infty > 0$, we have

$$\mathrm{patt}(b) = m \iff \exists s \in \mathbb{R}^{k+} \text{ such that } b = U_m s.$$

Hereafter, the notation $|m|_\downarrow = (|m|_{\downarrow 1}, \ldots, |m|_{\downarrow p})'$ represents the components of m sorted non-increasingly with respect to the absolute value.

[2] This question has been addressed recently in two preprints [12, 18] where, contrary to us, it is required that $\ker(X) = \{0\}$ and the affine components are not characterized.

Example 2 *Let* $m = (2, -1, 0, 1, 2)'$. *Then*

$$U_m = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 & 0 \end{pmatrix}' \text{ and } U_{|m|_\downarrow} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}'.$$

Definition 4 *Let* $m \in \mathbb{Z}^p$ *be a SLOPE pattern with* $k := \|m\|_\infty > 0$. *The* clustered matrix $\tilde{X}_m \in \mathbb{R}^{n \times k}$ *of* $X \in \mathbb{R}^{n \times p}$ *is defined by* $\tilde{X}_m = XU_m$; *the* clustered parameter $\tilde{\lambda}_m \in \mathbb{R}^k$ *of* $\lambda \in \mathbb{R}^p$ *is defined by* $\tilde{\lambda}_m = (U_{|m|_\downarrow})'\lambda$.

Note that the dimension of the design matrix X is reduced when it is clustered as $\tilde{X}_m$ by a pattern m: a null component $m_i = 0$ leads to discard the column X_i from the design matrix X, and a cluster $K \subset \{1, \dots, p\}$ of m (set of components of m equal in absolute value) leads to replace the columns $(X_i)_{i \in K}$ by one column equal to the signed sum: $\sum_{i \in K} \operatorname{sign}(m_i) X_i$.

Example 3 *Let* $X = (X_1|X_2|X_3|X_4|X_5)$, $m = (2, -1, 0, 1, 2)'$, $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)' \in \mathbb{R}^5$. *Then the clustered matrix and the clustered parameter are given by:*

$$\tilde{X}_m = (X_1 + X_5| - X_2 + X_4) \text{ and } \tilde{\lambda}_m = \begin{pmatrix} \lambda_1 + \lambda_2 \\ \lambda_3 + \lambda_4 \end{pmatrix}.$$

2.3 Subdifferential of the sorted ℓ_1 norm

The subdifferential of a norm is related to the dual norm $\|\cdot\|^*$ via the following formula [7, p. 180]:

$$\partial\|\cdot\|(b) = \{v \in \mathbb{R}^p : \|v\|^* \leq 1 \text{ and } b'v = \|b\|\}, \quad b \in \mathbb{R}^p.$$

In particular, $\partial\|\cdot\|(b)$ is a face of the dual unit ball. For the sorted ℓ_1 norm, the above formula can be specified further with the pattern matrix and the clustered parameter associated to $m = \operatorname{patt}(b)$ for $b \neq 0$ [2, 14]:

$$\partial J_\lambda(b) = \left\{ v \in \mathbb{R}^p : J_\lambda^*(v) \leq 1 \text{ and } U_m' v = \tilde{\lambda}_m \right\}. \tag{2}$$

Remark 2 *Let* $\lambda \in \mathbb{R}^{p+}$. *The mapping* $m \mapsto \partial J_\lambda(m)$ *is a bijection between the set of SLOPE patterns and the set of faces of the unit ball of* J_λ^* *(the signed permutahedron) [14, Theorem 6]. It is no longer true when* $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ *is not a decreasing sequence. Therefore we restrict our study to the case where* $\lambda \in \mathbb{R}^{p+}$, *i.e.* $\lambda_1 > \cdots > \lambda_p > 0$. *The bijection is illustrated in Figure 1:*

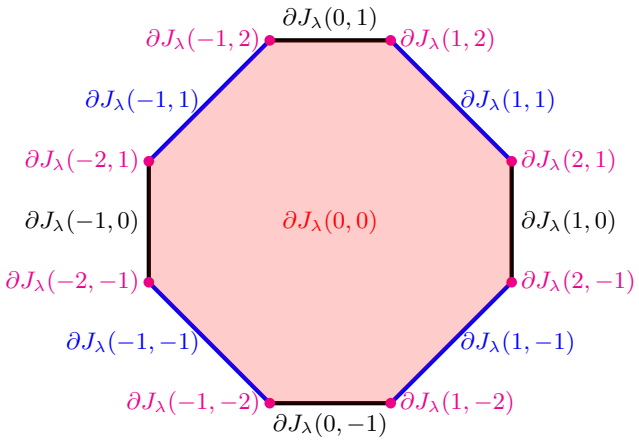

Figure 1: Bijection between

$$\mathcal{P}_2^{\text{slope}} = \{(0,0), \pm(1,0), \pm(0,1), \pm(1,1), \pm(1,-1), \pm(1,2), \pm(1,-2), \pm(2,1), \pm(2,-1)\}$$

and the set of faces of the signed permutahedron for $\lambda_1 > \lambda_2 > 0$.

3 Solution, fitted value and gradient paths

3.1 Solution set and fitted value

Given $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{p+}$, and $\gamma > 0$, we denote by $\mathcal{S}_{X,y,\lambda}(\gamma)$ (or simply $\mathcal{S}(\gamma)$ when there is no ambiguity) the set of solutions to the SLOPE optimization problem (1), namely:

$$\min_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \gamma J_\lambda(b).$$

For any $\gamma > 0$, the objective function of the above problem is continuous and coercive thus the *solution set* $\mathcal{S}(\gamma)$ is nonempty. Moreover, the *fitted value* $\widehat{\mathrm{fit}}(\gamma) = X\widehat{\beta}$ does not depend on $\widehat{\beta} \in \mathcal{S}(\gamma)$. When $\mathcal{S}(\gamma)$ is a singleton, we denote by $\widehat{\beta}(\gamma)$ its unique element. Note that uniqueness is rather a weak assumption, indeed the set $\{X \in \mathbb{R}^{n\times p} : \exists y \in \mathbb{R}^n, \exists\gamma > 0$ such that $\mathcal{S}_{X,y,\lambda}(\gamma)$ is not a singleton$\}$ has zero Lebesgue measure [14, Proposition 3].

Theorem 1 below shows that $\widehat{\mathrm{fit}}(\cdot)$ and $\widehat{\beta}(\cdot)$ are piecewise linear functions. Moreover expressions of $\widehat{\mathrm{fit}}(\cdot)$ and $\widehat{\beta}(\cdot)$ restricted to the interval I_m (depending on a SLOPE pattern m) are affine and explicit. We denote hereafter by A^+ the Moore-Penrose pseudo-inverse of a matrix A.

Theorem 1 *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^{p+}$, and $m \in \mathbb{Z}^p$ be a non-null SLOPE pattern with $k := \|m\|_\infty > 0$.*

1. *The set $I_m := \{\gamma > 0 : \exists\widehat{\beta} \in \mathcal{S}(\gamma)$ such that $\mathrm{patt}(\widehat{\beta}) = m\}$ is an interval, with the following characterization:*

$$\gamma \in I_m$$
$$\Updownarrow$$
$$\begin{cases} \exists s \in \mathbb{R}^{k+} \text{ such that } \tilde{X}'_m y - \gamma\tilde{\lambda}_m = \tilde{X}'_m\tilde{X}_m s & \text{(positivity condition),} \\ X'(\tilde{X}'_m)^+\tilde{\lambda}_m + \frac{1}{\gamma}X'(I_n - (\tilde{X}'_m)^+\tilde{X}'_m)y \in \partial J_\lambda(m) & \text{(subdifferential condition).} \end{cases}$$

   Moreover, $\widehat{\beta} := U_m s \in \mathcal{S}(\gamma)$ and $\mathrm{patt}(\widehat{\beta}) = m$ for any $s \in \mathbb{R}^{k+}$ satisfying the positivity condition at $\gamma \in I_m$.

2. *The* fitted value path *$\gamma \mapsto \widehat{\mathrm{fit}}(\gamma)$ is continuous and piecewise linear on $(0, +\infty)$, with the following affine expression on I_m:*

$$\widehat{\mathrm{fit}}(\gamma) = (\tilde{X}'_m)^+\tilde{X}'_m y - \gamma(\tilde{X}'_m)^+\tilde{\lambda}_m, \quad \gamma \in I_m.$$

3. *If $\mathcal{S}(\gamma) = \{\widehat{\beta}(\gamma)\}$ for all $\gamma > 0$, then the* solution path *$\gamma \mapsto \widehat{\beta}(\gamma)$ is continuous and piecewise linear on $(0, +\infty)$, with the following affine expression on I_m:*

$$\widehat{\beta}(\gamma) = U_m(\tilde{X}'_m\tilde{X}_m)^{-1}(\tilde{X}'_m y - \gamma\tilde{\lambda}_m), \quad \gamma \in I_m.$$

The characterization of the interval I_m above is closely related to Theorem 3.1 in [2].

3.2 Gradient path and clusters

A solution of the SLOPE optimization problem is characterized by the following two conditions

$$\widehat{\beta} \in \mathcal{S}(\gamma) \Leftrightarrow \begin{cases} J^*_\lambda(X'(y - X\widehat{\beta})) \le \gamma \\ \widehat{\beta}'X'(y - X\widehat{\beta}) = \gamma J_\lambda(\widehat{\beta}) \end{cases}$$

Note that $X'(y - X\widehat{\beta}) = X'(y - \widehat{\mathrm{fit}}(\gamma))$ is the gradient at $\widehat{\beta}$ of the sum of residual squares $b \mapsto \frac{1}{2}\|y - Xb\|_2^2$. Subsequently, we call gradient path the expression $\gamma > 0 \mapsto X'(y - \widehat{\mathrm{fit}}(\gamma))$. The set of inequalities describing the ball of radius γ for the dual sorted ℓ_1 norm which are saturated by the gradient is :

$$\mathcal{A}(\gamma) := \left\{ i \in [p] : \frac{\|X'(y - \widehat{\mathrm{fit}}(\gamma))\|_{(i)}}{\sum_{j=1}^i \lambda_j} = \gamma \right\}.$$

According to Proposition 1 below, the set $\mathcal{A}(\gamma)$ provides both the number of non-zero clusters, the size of these clusters as well as the number of non-zero components.

Proposition 1 *Let $\lambda \in \mathbb{R}^{p+}$, $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\gamma > 0$ and $\widehat{\beta} \in \mathcal{S}(\gamma)$.*

1. *Let $1 \leq k_1 \leq \cdots \leq k_l \leq p$ be a subdivision such that:*

$$|\mathrm{supp}(\widehat{\beta})| = k_l \text{ and } |\widehat{\beta}|_{\downarrow 1} = \cdots = |\widehat{\beta}|_{\downarrow k_1} > \cdots > |\widehat{\beta}|_{\downarrow k_{l-1}+1} = \cdots = |\widehat{\beta}|_{\downarrow k_l} > 0$$

(i.e. $\widehat{\beta}$ has l non-null clusters, the cluster of the largest value has k_1 elements and so on and $\widehat{\beta}$ has k_l non-null components). Then, $\{k_1, \ldots, k_l\} \subset \mathcal{A}(\gamma)$.

2. *Conversely, if $\{k_1, \ldots, k_l\} = \mathcal{A}(\gamma)$ then*

$$|\widehat{\beta}|_{\downarrow 1} = \cdots = |\widehat{\beta}|_{\downarrow k_1} \geq \cdots \geq |\widehat{\beta}|_{\downarrow k_{l-1}+1} = \cdots = |\widehat{\beta}|_{\downarrow k_l} \geq |\widehat{\beta}|_{\downarrow k_l+1} = \cdots = |\widehat{\beta}|_{\downarrow p} = 0$$

(i.e. the number of non-null clusters of $\widehat{\beta}$ is smaller or equal to l and the number of non-null components is smaller or equal to k_l).

4 Algorithms to compute the solution path

To keep this section simple we assume that $\mathcal{S}(\gamma) = \{\widehat{\beta}(\gamma)\}$ for all $\gamma > 0$. Let $J^*_\lambda(X'y) = \gamma_0 > \gamma_1 > \ldots \gamma_r > \gamma_{r+1} = 0$ be a subdivision such that $\gamma \mapsto \widehat{\beta}(\gamma)$ is affine with pattern $m^{(i)}$ on the interval (γ_{i+1}, γ_i) for $i = 0, \ldots, r$ (*i.e* the interior of $I_{m^{(i)}}$ is (γ_{i+1}, γ_i)).

First, let us explain how to compute the SLOPE solution path on $[\gamma_1, \gamma_0]$. By construction of $m^{(0)}$ the following implication holds

$$\forall \gamma \in (\gamma_1, \gamma_0), \mathrm{patt}(\widehat{\beta}(\gamma)) = m^{(0)} \Rightarrow \frac{1}{\gamma} X'(y - \widehat{\mathrm{fit}}(\gamma)) \in \partial J_\lambda(m^{(0)}).$$

Moreover, since $\gamma > 0 \mapsto \widehat{\mathrm{fit}}(\gamma)$ is continuous, $\mathrm{fit}(\gamma_0) = 0$ and $\partial J_\lambda(m^{(0)})$ is a closed set, we get

$$\frac{1}{\gamma_0} X'(y - \widehat{\mathrm{fit}}(\gamma_0)) = \frac{1}{\gamma_0} X'y \in \partial J_\lambda(m^{(0)}). \tag{3}$$

Algorithm 1 provides the pattern $M(\frac{1}{\gamma_0}X'y)$ of the smallest face of the signed permutahedron containing $\frac{1}{\gamma_0}X'y$. Therefore, by construction

$$\partial J_\lambda\left(M\left(\frac{1}{\gamma_0}X'y\right)\right) \subset \partial J_\lambda(m^{(0)}) \Rightarrow \left\|M\left(\frac{1}{\gamma_0}X'y\right)\right\|_\infty \leq \|m^{(0)}\|_\infty \tag{4}$$

According to (4), if $\frac{X'y}{\gamma_0}$ lies onto a facet of the signed permutahedron, we have $m^{(0)} = M\left(\frac{1}{\gamma_0}X'y\right)$.

Algorithm 1 Pattern of the smallest face containing a vector

Require: $\lambda \in \mathbb{R}^{p+}$ and $z \in \mathbb{R}^p$ such that $J^*_\lambda(z) \leq 1$

Define the set of saturated inequalities as follows

$$\mathcal{A}(z) := \left\{ i \in [p] : \frac{\|z\|_{(i)}}{\sum_{j=1}^i \lambda_j} = 1 \right\}.$$

Define $M(z) \in \mathcal{P}_p^{\mathrm{slope}}$ as follows

$$M(z) := \begin{cases} 0 \text{ if } \mathcal{A}(z) = \emptyset \\ \forall j \in [p], M_j(z) = \mathrm{sign}(z_j) \sum_{i \in \mathcal{A}(z)} \mathbf{1}(|z_j| \geq \lambda_i) \text{ if } \mathcal{A}(z) \neq \emptyset \end{cases}$$

return $M(z)$

Example 4 *We provide illustrations of the solution path of SLOPE when $y = (6, 2)' \in \mathbb{R}^2$, $\lambda = (4, 2) \in \mathbb{R}^{2+}$ and $X \in \mathbb{R}^{2\times 2}$ is the matrix given below*

$$X = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Largest kink γ_0: *We have $X'y = (7,5)'$, therefore $\gamma_0 = J^*_\lambda(X'y) = 2$.*

Pattern $m^{(0)}$ in the left neighborhod of γ_0: *Since $\frac{1}{\gamma_0}X'y = (3.5, 2.5)'$ lies in the relative interior of the permutahedron $\partial J_\lambda(1,1) = \mathrm{conv}\{(4,2)', (2,4)'\}$ then $m^{(0)} = M(\frac{1}{\gamma_0}X'y) = (1,1)$.*

Affine expression of $\widehat{\beta}(\gamma)$ in the left neighborhod of γ_0: *According to statement 3 in Theorem 1 when $\gamma < \gamma_0 = 2$ is sufficiently close to γ_0 we have $\widehat{\beta}(\gamma) = (\frac{8-4\gamma}{3}, \frac{8-4\gamma}{3})$.*

Algorithm 2 uses the characterisation of $I_{m^{(0)}}$, based on the positivity and subdifferential conditions, to provide both the kink γ_1 as well as the pattern $m^{(1)}$.

Algorithm 2 Computing the next kink and the next pattern

Require: $X \in \mathbb{R}^{n\times p}, \lambda \in \mathbb{R}^{p+}, \gamma_i, m^{(i)}$ and $s(\gamma) = (\tilde{X}'_{m^{(i)}}\tilde{X}_{m^{(i)}})^{-1}(\tilde{X}'_{m^{(i)}}y - \gamma\tilde{\lambda}_{m^{(i)}})$

Let $k = \|m^{(i)}\|_\infty$ and compute

$$\gamma_{\text{fuse}} = \begin{cases} \sup\{\gamma \in [0,\gamma_i) : s(\gamma) \notin \mathbb{R}^{k+}\} & \text{if the set is not empty} \\ 0 & \text{otherwise} \end{cases}.$$

if $\gamma_{\text{fuse}} = 0$ **then**

Compute γ_{split} as follows

$$\gamma_{\text{split}} = \begin{cases} \sup\{\gamma \in [0,\gamma_i) : X'(y - \tilde{X}_{m^{(i)}}s(\gamma)) \notin \gamma\partial J_\lambda(m^{(i)})\} & \text{if the set is not empty} \\ 0 & \text{otherwise} \end{cases}$$

if $\gamma_{\text{slpit}} = 0$ **then**

return The SLOPE solution path is entirely computed

else

$\gamma_{i+1} = \gamma_{\text{split}}$

Using Algorithm 1 compute $m^{(i+1)} = M(\frac{1}{\gamma_{i+1}}X'(y - \tilde{X}_{m^{(i)}}s(\gamma_{i+1})))$

return $\gamma_{i+1}, m^{(i+1)}$

end if

else if $\gamma_{\text{fuse}} > 0$ and $X'(y - \tilde{X}_{m^{(i)}}s(\gamma_{\text{fuse}})) \in \gamma_{\text{fuse}}\partial J_\lambda(M)$ **then**

$\gamma_{i+1} = \gamma_{\text{fuse}}$

$m^{(i+1)} = \mathrm{patt}(U_{m^{(i)}}s(\gamma_{\text{fuse}}))$

return $\gamma_{i+1}, m^{(i+1)}$

else

Compute γ_{i+1} as follows

$$\gamma_{i+1} = \sup\{\gamma \in [\gamma_{\text{fuse}},\gamma_i) : X'(y - \tilde{X}_{m^{(i)}}s(\gamma)) \notin \gamma\partial J_\lambda(m^{(i)})\}$$

Using Algorithm 1 compute $m^{(i+1)} = M(\frac{1}{\gamma_{i+1}}X'(y - \tilde{X}_{m^{(i)}}s(\gamma_{i+1}))$

return $\gamma_{i+1}, m^{(i+1)}$

end if

Using iteratively Algorithm 2 allows to compute entirely the SLOPE solution path[3].

5 SLOPE solution path applied on data

Below we will use a data set describing the quality of red "Vinho Verde" wines [4][4]. In this data set explanatory variables $X \in \mathbb{R}^{1599\times 11}$ are physicochemical measurements such as density, acidity, the amount of sugar and alcohol, etc and the response $y \in \mathbb{R}^{1599}$ is the wine quality score between 0

[3] An implementation in Python of these algorithms for computing the SLOPE solution path is available online: `https://anonymous.4open.science/r/slope-path-744C`

[4] This data set is available online: `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`

and 10. For this numerical experiment we take $\lambda = (11, 10, \dots, 1)$. The matrix X is mean-centered ($\forall j \in \{1, \dots, 11\}$ $\sum_{i=1}^{1599} X_{ij} = 0$) and standardized ($\forall j \in \{1, \dots, 11\}$ $\sum_{i=1}^{1599} X_{ij}^2 = 1598$).

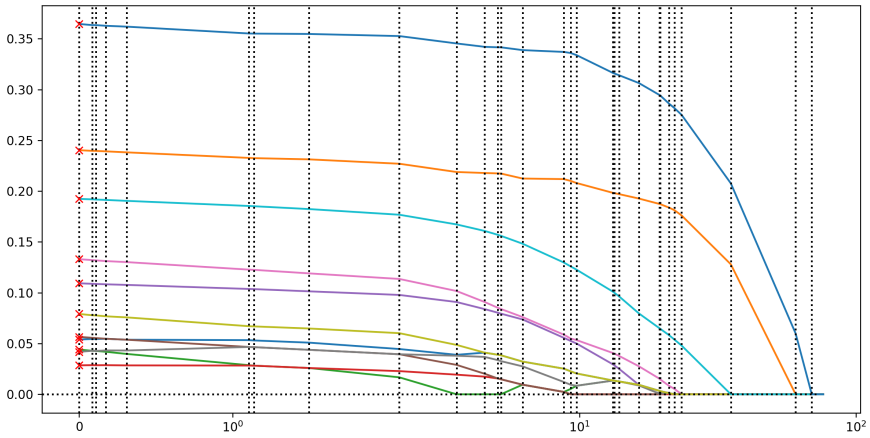

Figure 2: This figure provides the solution of SLOPE, in absolute value, as a function of $\gamma > 0$ (the x-axis is reported on the logarithm scale). One may observe that when $0.07 = \gamma^{(28)} < \gamma < \gamma^{(4)} = 17.79$, at least two components of SLOPE are equal in absolute value. Moreover, in absolute value, the SLOPE solution converges to the ordinary least squares estimator when γ tends to 0 (components of the ordinary least squares estimator in absolute value are crosses on the y-axis).

6 Conclusion and future works

One of the main result in this article is Theorem 1 proving that the SLOPE solution path is piecewise linear and providing the characterization of the intervals where the path is affine. Moreover algorithms 1 and 2 allow to compute numerically this path. The computational time of our numerical scheme depends mainly on the number of kinks. In the illustration on real data set, the number of intervals is rather small (29) and kinks are concentrated around zero. However the number of intervals where the path is affine is bounded by the number of SLOPE patterns in $\mathbb{R}^p$ and potentially, similarly as for LASSO [9], this huge upper bound might be reached for some pathological examples. In the future we are going to test our method on various data sets and if kinks are too concentrated around zero we would switch for another numerical scheme (like, for instance, the one developped in [8]) to compute the SLOPE solution path in the neighborhood of zero.

Aknowledgements

The Institut de Mathématiques de Bourgogne (IMB) receives support from the EIPHI Graduate School (contract ANR-17-EURE-0002)

7 Appendix

Proof of Theorem 1

1: I_m is an interval) Hereafter we suppose that $I_m \neq \emptyset$. Let $\gamma_0, \gamma_1 \in I_m$ and pick $\widehat{\beta}(\gamma_0) \in \mathcal{S}(\gamma_0)$, $\widehat{\beta}(\gamma_1) \in \mathcal{S}(\gamma_1)$ such that $\mathrm{patt}(\widehat{\beta}(\gamma_0)) = \mathrm{patt}(\widehat{\beta}(\gamma_1)) = m$. Let $\alpha \in [0, 1]$, $\bar{\gamma} = \alpha\gamma_0 + (1 - \alpha)\gamma_1$ and $\bar{\beta} = \alpha\widehat{\beta}(\gamma_0) + (1 - \alpha)\widehat{\beta}(\gamma_1)$ then $\mathrm{patt}(\bar{\beta}) = m$. Indeed, if $m = 0$ then clearly $\mathrm{patt}(\bar{\beta}) = 0$. Otherwise, let $k = \|m\|_\infty \geq 1$ then $\widehat{\beta}(\gamma_0) = U_m s_0$ for some $s_0 \in \mathbb{R}^{k+}$, $\widehat{\beta}(\gamma_1) = U_m s_1$ for some

$s_1 \in \mathbb{R}^{k+}$ therefore $\bar{\beta} = U_m \bar{s}$ where $\bar{s} = \alpha s_0 + (1-\alpha) s_1 \in \mathbb{R}^{k+}$. To prove that I_m is an interval it remains to show that $\bar{\beta} \in \mathcal{S}(\bar{\gamma})$. Because both $\widehat{\beta}(\gamma_0)$ and $\widehat{\beta}(\gamma_1)$ are SLOPE minimizers, we have

$$X'(y - X\widehat{\beta}(\gamma_0)) \in \gamma_0 \partial J_\lambda(m) \text{ and } X'(y - X\widehat{\beta}(\gamma_1)) \in \gamma_1 \partial J_\lambda(m).$$

By construction of $\bar{\beta}$ the following equality occurs:

$$\alpha X'(y - X\widehat{\beta}(\gamma_0)) + (1-\alpha) X'(y - X\widehat{\beta}(\gamma_1)) = X'(y - X\bar{\beta}).$$

Moreover, since $\partial J_\lambda(m)$ is a convex set, we have $\alpha\gamma_0 \partial J_\lambda(m) + (1-\alpha)\gamma_1 \partial J_\lambda(m) \subset \bar{\gamma} \partial J_\lambda(m)$. Consequently, $X'(y - X\bar{\beta}) \in \bar{\gamma} \partial J_\lambda(m) = \bar{\gamma} \partial J_\lambda(\bar{\beta})$ thus $\bar{\beta} \in \mathcal{S}(\bar{\gamma})$.

1: characterization of I_m) The proof of this characterization is closely related to the proof of Theorem 3.1 in [2].

Necessity. If $\gamma \in I_m$, then there exists $\widehat{\beta} \in \mathcal{S}(\gamma)$ such that $\mathrm{patt}(\widehat{\beta}) = m$. Consequently, $\widehat{\beta} = U_m s$ for some $s \in \mathbb{R}^{k+}$. Because $\widehat{\beta}$ is a element of $\mathcal{S}(\gamma)$ whose pattern is m then $X'(y - \widehat{\mathrm{fit}}(\gamma)) \in \gamma \partial J_\lambda(\widehat{\beta}) = \gamma \partial J_\lambda(m)$. Multiplying this inclusion by U_m', due to (2), we get $\tilde{X}_m'(y - \widehat{\mathrm{fit}}(\gamma)) = \gamma \tilde{\lambda}_m$ and so

$$\tilde{X}_m' y - \gamma \tilde{\lambda}_m = \tilde{X}_m' \widehat{\mathrm{fit}}(\gamma) = \tilde{X}_m' \tilde{X}_m s. \tag{5}$$

The positivity condition is proven.

We apply $(\tilde{X}_m')^+$ from the left to (5) and use the fact that $(\tilde{X}_m')^+ \tilde{X}_m'$ is the projection onto $\mathrm{col}(\tilde{X}_m)$. Since $\widehat{\mathrm{fit}}(\gamma) \in \mathrm{col}(\tilde{X}_m)$, we have $(\tilde{X}_m')^+ \tilde{X}_m' \widehat{\mathrm{fit}}(\gamma) = \widehat{\mathrm{fit}}(\gamma)$. Thus,

$$(\tilde{X}_m')^+ \tilde{X}_m' y - \gamma (\tilde{X}_m')^+ \tilde{\lambda}_m = \widehat{\mathrm{fit}}(\gamma).$$

The above equality gives the subdifferential condition:

$$\begin{aligned} \partial J_\lambda(m) \ni \frac{1}{\gamma} X'(y - \widehat{\mathrm{fit}}(\gamma)) &= \frac{1}{\gamma} X'(y - ((\tilde{X}_m')^+ \tilde{X}_m' y - \gamma (\tilde{X}_m')^+ \tilde{\lambda}_m)) \\ &= X'(\tilde{X}_m')^+ \tilde{\lambda}_m + \frac{1}{\gamma} X'(I_n - (\tilde{X}_m')^+ \tilde{X}_m') y. \end{aligned} \tag{6}$$

Sufficiency. Assume that the positivity condition and the subdifferential conditions hold true. Then, by the positivity condition, one may pick $s \in \mathbb{R}^{k+}$ for which

$$\gamma \tilde{\lambda}_m = \tilde{X}_m' y - \tilde{X}_m' \tilde{X}_m s. \tag{7}$$

Let us show that $U_m s \in \mathcal{S}(\gamma)$. By definition of U_m, we have $\mathrm{patt}(U_m s) = m$ thus $\partial J_\lambda(U_m s(\gamma)) = \partial J_\lambda(m)$. Moreover, using (6) and (7) one may deduce

$$\begin{aligned} \partial J_\lambda(U_m s) &\ni \frac{1}{\gamma} X'(y - (\tilde{X}_m')^+ \tilde{X}_m' y + \gamma (\tilde{X}_m')^+ \tilde{\lambda}_m) \\ &= \frac{1}{\gamma} X'(y - (\tilde{X}_m')^+ \tilde{X}_m' y + (\tilde{X}_m')^+ (\tilde{X}_m y - \tilde{X}_m' \tilde{X}_m s)) \\ &= \frac{1}{\gamma} X'(y - X U_m s). \end{aligned}$$

Consequently $U_m s \in \mathcal{S}(\gamma)$.

2 and 3: continuity) Let $\gamma \in (0, +\infty)$, $(\gamma_n)_{n \in \mathbb{N}}$ be a sequence converging to γ and $\widehat{\beta}(\gamma_n) \in S_{X, \gamma_n J_\lambda}(y)$. Both sequences $(\widehat{\beta}(\gamma_n))_{n \in \mathbb{N}}$ and $(\widehat{\mathrm{fit}}(\gamma_n))_{n \in \mathbb{N}}$ are bounded therefore, up to extract a subsequence, one may assume that both $(\widehat{\beta}(\gamma_n))_{n \in \mathbb{N}}$ and $(\widehat{\mathrm{fit}}(\gamma_n))_{n \in \mathbb{N}}$ converge respectively to a limit point $l \in \mathbb{R}^p$ and $Xl \in \mathbb{R}^n$. Let $\widehat{\beta}(\gamma) \in \mathcal{S}(\gamma)$. Because $\widehat{\beta}(\gamma_n)$ is a minimizer, the following inequality occurs.

$$\frac{1}{2} \|y - \widehat{\mathrm{fit}}(\gamma_n)\|_2^2 + \gamma_n J_\lambda(\widehat{\beta}(\gamma_n)) \leq \frac{1}{2} \|y - \widehat{\mathrm{fit}}(\gamma)\|_2^2 + \gamma_n J_\lambda(\widehat{\beta}(\gamma)).$$

Taking the limit in the above expression gives

$$\frac{1}{2} \|y - Xl\|_2^2 + \gamma J_\lambda(l) \leq \frac{1}{2} \|y - \widehat{\mathrm{fit}}(\gamma)\|_2^2 + \gamma J_\lambda(\widehat{\beta}(\gamma)).$$

Because $\widehat{\beta}(\gamma) \in \mathcal{S}(\gamma)$, one may deduce that $l \in \mathcal{S}(\gamma)$ and thus $Xl = \widehat{\mathrm{fit}}(\gamma)$. Therefore, the unique limit point of the bounded sequence $(\widehat{\mathrm{fit}}(\gamma_n))_{n\in\mathbb{N}}$ is $\widehat{\mathrm{fit}}(\gamma)$. Consequently, $\lim_{n\to+\infty} \widehat{\mathrm{fit}}(\gamma_n) = \widehat{\mathrm{fit}}(\gamma)$ and thus the function $\gamma \in (0, +\infty) \mapsto \widehat{\mathrm{fit}}(\gamma)$ is continuous. Similarly, if $\mathcal{S}(\gamma)$ is a singleton then $l = \widehat{\beta}(\gamma)$, the unique limit point of the bounded sequence $(\widehat{\beta}(\gamma_n))_{n\in\mathbb{N}}$ is $\widehat{\beta}(\gamma)$ and thus $\lim_{n\to+\infty} \widehat{\beta}(\gamma_n) = \widehat{\beta}(\gamma)$. Therefore the function $\gamma \in (0, +\infty) \mapsto \widehat{\beta}(\gamma)$ is continuous.

2) When $\gamma \in I_m$ then multiplying both side of the positivity condition by $(\tilde{X}'_m)^+$ and using the fact that $(\tilde{X}'_m)^+\tilde{X}'_m$ is the projection onto $\mathrm{col}(\tilde{X}_m)$ gives

$$(\tilde{X}'_m)^+\tilde{X}'_m y - \gamma(\tilde{X}'_m)^+\tilde{\lambda}_m = (\tilde{X}'_m)^+\tilde{X}'_m\tilde{X}_m s = \tilde{X}_m s = \widehat{\mathrm{fit}}(\gamma).$$

3) The proof of statement 3) relies on Lemma 1 proved in supplementary material.

Lemma 1 *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^{p+}$. There exists $\widehat{\beta} \in \mathcal{S}(\gamma)$ for which the pattern $m = \mathrm{patt}(\widehat{\beta})$ satisfies $\ker(\tilde{X}_m) = \{0\}$.*

Consequently, when $\gamma \in I_m$ and $\mathcal{S}(\gamma)$ is a singleton then $\ker(\tilde{X}_m) = \{0\}$, where $m = \mathrm{patt}(\widehat{\beta}(\gamma))$. Since $\tilde{X}'_m\tilde{X}_m$ is invertible, the positivity condition gives

$$\widehat{\beta}(\gamma) = U_m s = U_m(\tilde{X}'_m\tilde{X}_m)^{-1}(\tilde{X}'_m y - \gamma\tilde{\lambda}_m).$$

References

[1] 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.

[2] Małgorzata Bogdan, Xavier Dupuis, Piotr Graczyk, Bartosz Kołodziejek, Tomasz Skalski, Patrick Tardivel, and Maciej Wilczyński. Pattern recovery by slope. *arXiv preprint arXiv:2203.12086*, 2022.

[3] 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.

[4] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

[5] Xavier Dupuis and Patrick JC Tardivel. Proximal operator for the sorted ℓ_1 norm: Application to testing procedures based on slope. *Journal of Statistical Planning and Inference*, 221:1–8, 2022.

[6] Mario Figueiredo and Robert Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Artificial Intelligence and Statistics*, pages 930–938. PMLR, 2016.

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

[8] Johan Larsson, Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin. Coordinate descent for slope. *arXiv preprint arXiv:2210.14780*, 2022.

[9] Julien Mairal and Bin Yu. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th International Conference on Machine Learning*, pages 353–360, 2012.

[10] Aleksandra Maj-Kańska, Piotr Pokarowski, and Agnieszka Prochenka. Delete or merge regressors for linear model selection. 2015.

[11] Renato Negrinho and Andre Martins. Orbit regularization. *Advances in neural information processing systems*, 27, 2014.

[12] Shunichi Nomura. An exact solution path algorithm for slope and quasi-spherical oscar. *arXiv preprint arXiv:2010.15511*, 2020.

[13] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.

[14] Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. *Journal of Machine Learning Research*, 23(331):1–36, 2022.

[15] Dhruv B Sharma, Howard D Bondell, and Hao Helen Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22(2):319–340, 2013.

[16] Tomasz Skalski, Piotr Graczyk, Bartosz Kołodziejek, and Maciej Wilczyński. Pattern recovery and signal denoising by slope when the design matrix is orthogonal. *arXiv preprint arXiv:2202.08573*, 2022.

[17] Benjamin G Stokell, Rajen D Shah, and Ryan J Tibshirani. Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(3):579–611, 2021.

[18] Atsumori Takahashi and Shunichi Nomura. Efficient path algorithms for clustered lasso and oscar. *arXiv preprint arXiv:2006.08965*, 2020.

[19] 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.

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

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