Simple expressions of the LASSO and SLOPE estimators in small-dimension

Patrick J C Tardivel, Rémi Servien, Didier Concordet

Simple expressions of the LASSO and SLOPE estimators in small-dimension

Patrick J.C. Tardivel*, Rémi Servien and Didier Concordet
TOXALIM, Université de Toulouse, INRA, ENVT, Toulouse, France.

Abstract

We study the LASSO and SLOPE estimators when the design X satisfies $\ker(X) = \mathbf{0}$. We state that, even if the design is not orthogonal, even if residuals are correlated, up to a transformation, the LASSO and SLOPE estimators have a simple expression based on the best linear unbiased estimator.

Keywords: Best linear unbiased estimator, LASSO, SLOPE.

1 Introduction

Let us consider the following small-dimensional linear model

$$Y = X\beta^* + \varepsilon, \tag{1}$$

where X is a $n \times p$ fixed design matrix with $\ker(X) = \mathbf{0}$ (thus $n \geq p$ hence the adjective "small-dimensional"), $\beta^* \in \mathbb{R}^p$ is an unknown parameter and ε is a centered random vector with an invertible and known covariance matrix Γ.

The Least Absolute Shrinkage and Selection Operator (LASSO) estimator [Tibshirani, 1996] and the Sorted L-One Penalized Estimation (SLOPE) estimator [Bogdan et al., 2015] are respectively defined by

$$\hat{\beta}^{\text{lasso}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda\|\beta\|_1 \right\} \text{ and } \hat{\beta}^{\text{slope}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda_1|\beta_{[1]}| + \cdots + \lambda|\beta_{[p]}| \right\} \tag{2}$$

where, in the second expression, the tuning parameters $(\lambda_i)_{1\leq i\leq p}$ satisfies $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $[\,.\,]$ is a permutation of $\{1, \ldots, p\}$ such that $|\beta_{[1]}| \geq \cdots \geq |\beta_{[p]}|$.

It is well known that when the design X is orthogonal (so $\ker(X) = \mathbf{0}$) the LASSO is just the following soft thresholded Ordinary Least Squares (OLS) estimator [Tibshirani, 1996]

$$\hat{\beta}^{\text{lasso}} = \left(\text{sign}(\hat{\beta}_1^{\text{ols}})(|\hat{\beta}_1^{\text{ols}}| - \lambda)_+, \ldots, \text{sign}(\hat{\beta}_p^{\text{ols}})(|\hat{\beta}_p^{\text{ols}}| - \lambda)_+\right). \tag{3}$$

Popularized by the pioneer work of Tibshirani, the orthogonal design became a case study: Chzhen et al. [2017], Duan et al. [2016], G'Sell et al. [2015], Lockhart et al. [2014], Tian et al. [2015], Wen et al. [2016]. Furthermore some nice properties such as the irrepresentable condition [Bühlmann and van de Geer, 2011, Meinshausen and Bühlmann, 2006, Zhao and Yu, 2006, Zou, 2006] holds when X is orthogonal.

Similarly to the LASSO, the orthogonal design is also a case study for the SLOPE estimator [Bogdan et al., 2015, Gossmann et al., 2015, Su and Candes, 2016].

As explain previously, the orthogonal design setting appears as the ideal case. By seeking to generalize its properties to the non-orthogonal setting, we discovered a relevant orthogonalizing transformation U. Actually, in the new model $\tilde{Y} + \tilde{X}\beta^* + \tilde{\varepsilon}$, where $\tilde{Y} = UY$, $\tilde{X} = UX$ and $\tilde{\varepsilon} = U\varepsilon$, the LASSO $\tilde{\beta}^{\text{lasso}}$ has the following simple expression based on the Best Linear Unbiased Estimator (BLUE)

$$\tilde{\beta}^{\text{lasso}} = \left(\text{sign}(\hat{\beta}_1^{\text{blue}})(|\hat{\beta}_1^{\text{blue}}| - \lambda)_+, \ldots, \text{sign}(\hat{\beta}_p^{\text{blue}})(|\hat{\beta}_p^{\text{blue}}| - \lambda)_+\right). \tag{4}$$

*corresponding author: patrick.tardivel@gmail.com

Similarly to the LASSO we also obtained a simple expression for the SLOPE based on the BLUE. The transformation U is available in small dimension even if X is not orthogonal (but $\ker(X) = \mathbf{0}$) and even if components of ε are correlated

Let us point out the differences and advantages of the LASSO $\tilde{\beta}^{\text{lasso}}$ obtained after applying the orthogonalizing transformation U.

- Contrarily to the LASSO $\hat{\beta}^{\text{lasso}}$ obtained when X is orthogonal and $\varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n)$ (very often met in practice), in general the components of $\tilde{\beta}^{\text{lasso}}$, given in (4), are not independent.
- Expression of the LASSO can be rewritten, up to a transformation, as an expression based on the BLUE. As an example, one can derive from the multiple testing procedure based on the knockoff-LASSO estimator [Janson and Su, 2016] a new procedure based on the BLUE. The knockoff-LASSO is the following estimator

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X_{\text{ko}}\beta\|^2 + \lambda\|\beta\|_1.$$

  Since X_{ko} satisfies $\ker(X_{\text{ko}}) = \mathbf{0}$ [Barber and Candès, 2015], up to a transformation, the knockoff-LASSO is just a soft thresholded BLUE (where the BLUE is $(X'_{\text{ko}}\Gamma^{-1}X_{\text{ko}})^{-1}X_{\text{ko}}\Gamma^{-1}Y$, with $\Gamma = \text{var}(Y)$).

1.1 Notations

In this article, we denote J the SLOPE norm $J : \beta \in \mathbb{R}^p \mapsto \lambda_1|\beta|_{[1]} + \cdots + \lambda_p|\beta|_{[p]}$ where $|\beta_{[1]}| \geq \cdots \geq |\beta_{[p]}|$ and $\lambda_1 \geq \cdots \geq \lambda_p$ (see *e.g* Bogdan et al. [2015] for the proof that J is a norm). The OLS and BLUE estimators of the model (1), denoted $\hat{\beta}^{\text{ols}}$ and $\hat{\beta}^{\text{blue}}$, are respectively equal to

$$\hat{\beta}^{\text{ols}} := (X'X)^{-1}X'Y \text{ and } \hat{\beta}^{\text{blue}} := (X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}Y. \tag{5}$$

Whatever $t \in \mathbb{R}$, we set $(t)_+ = \max\{t, 0\}$ and $\text{sign}(t) = \mathbf{1}_{t>0} - \mathbf{1}_{t<0}$. Finally, given a subset $A \subset \mathbb{R}^p$, conv(A) is the smallest convex set containing A.

2 Orthogonalization of the design: simple form of the LASSO and SLOPE

When the design is orthogonal, some algorithms provides the SLOPE estimation [Bogdan et al., 2015] but the estimator writing is not explicit. To our knowledge, there is still no explicit formula for the SLOPE. In the following theorem, we provide the explicit expression of the SLOPE when X is orthogonal.

Theorem 1 *Let τ be a permutation of $\{1, \ldots, p\}$ ordering the components of the OLS estimator (5) namely $|\hat{\beta}^{\text{ols}}_{\tau(1)}| \geq \cdots \geq |\hat{\beta}^{\text{ols}}_{\tau(p)}|$. Let $(\hat{S}_k)_{1\leq k\leq p}$ be a sequence defined by $\forall k \in \{1, \ldots, p\}, \hat{S}_k := \sum_{i=1}^{k} \left(|\hat{\beta}^{\text{ols}}_{\tau(i)}| - \lambda_i\right)$ and let $1 \leq k_1 \leq \cdots \leq k_s = p$ be a partition of $\{1, \ldots, p\}$ such that*

$$k_1 := \max\left\{\underset{k\in\{1,\ldots,p\}}{\operatorname{argmax}}\left\{\frac{\hat{S}_k}{k}\right\}\right\} \text{ and } \forall i \in \{2, \ldots, s\}, \hat{k}_i := \max\left\{\underset{k>\hat{k}_{i-1}}{\operatorname{argmax}}\left\{\frac{\hat{S}_k - \hat{S}_{k_{i-1}}}{k - \hat{k}_{i-1}}\right\}\right\}.$$

When the design matrix X is orthogonal (i.e $X'X = Id_p$), whatever $i \in \{1, \ldots, p\}$, the components of $\hat{\beta}^{\text{slope}}$ (2) satisfy the inequality $\hat{\beta}^{\text{ols}}_i\hat{\beta}^{\text{slope}}_i \geq 0$ and

$$\left(|\hat{\beta}^{\text{slope}}_{\tau(1)}|, \ldots, |\hat{\beta}^{\text{slope}}_{\tau(p)}|\right) := \Bigg(\underbrace{\left(\frac{\hat{S}_{k_1}}{k_1}\right)_+, \ldots, \left(\frac{\hat{S}_{k_1}}{k_1}\right)_+}_{k_1 \text{ components}}, \ldots, \underbrace{\left(\frac{\hat{S}_{k_s} - \hat{S}_{k_{s-1}}}{k_s - k_{s-1}}\right)_+, \ldots, \left(\frac{\hat{S}_{k_s} - \hat{S}_{k_{s-1}}}{k_s - k_{s-1}}\right)_+}_{k_s - k_{s-1} \text{ components}}\Bigg).$$

Let us notice that when $\hat{\beta}^{\text{ols}}$ has a continuous distribution over $\mathbb{R}^p$ then almost surely the Cesàro sequence $(\hat{S}_k/k)$ reaches its maximum at a unique point, thus $k_1 := \text{argmax}\,\{\hat{S}_k/k\}$ is unique and the same argument applies for $k_2, \ldots, k_s$.

The figure 1 gives the relation between the OLS estimator and the SLOPE estimator when X is an orthogonal design matrix with $p = 2$ columns, $\lambda_1 = 2$ and $\lambda_2 = 1$.

When X is not orthogonal, the theorem 2 shows that applying the transformation $U := (X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}$ to the model (1) gives a new model in which the LASSO and the SLOPE have simple expressions depending only on the BLUE.

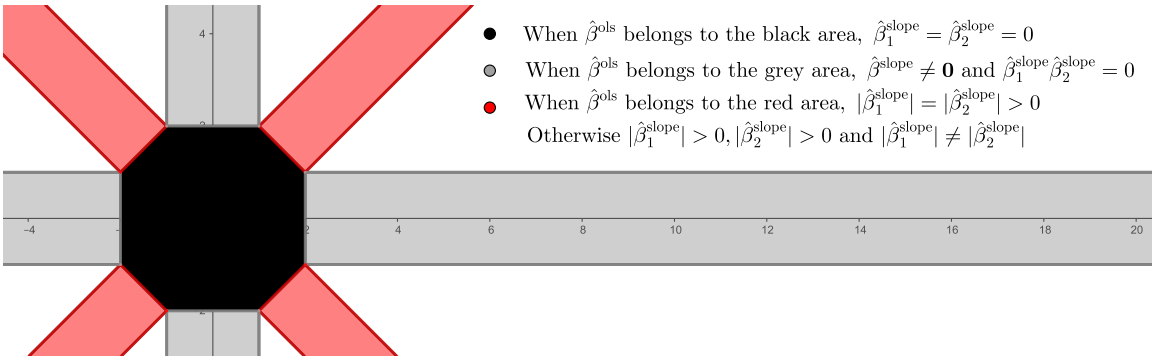

Figure 1: This figure illustrates the relation between the OLS estimator and the SLOPE. Let $\hat{S}_1, \hat{S}_2$ be defined as in the theorem 1. When $\hat{S}_1 \leq 0$ and $\hat{S}_2 \leq 0$ then $\hat{\beta}^{\text{ols}}$ is on the black area and $\hat{\beta}^{\text{slope}} = \mathbf{0}$. When $\hat{S}_1 \leq \hat{S}_2/2$ and $\hat{S}_2/2 > 0$ then $\hat{\beta}^{\text{ols}}$ is on the red area and $|\hat{\beta}_1^{\text{slope}}| = |\hat{\beta}_2^{\text{slope}}| > 0$. When $\hat{S}_1 > \hat{S}_2/2$ with $\hat{S}_1 > 0$ and $\hat{S}_2 - \hat{S}_1 < 0$ then $\hat{\beta}^{\text{ols}}$ is on the grey area, $\hat{\beta}^{\text{slope}} \neq \mathbf{0}$ and $|\hat{\beta}_1^{\text{slope}}||\hat{\beta}_2^{\text{slope}}| = 0$. Otherwise $\hat{\beta}^{\text{ols}}$ is on the white area then $|\hat{\beta}_1^{\text{slope}}| > 0, |\hat{\beta}_2^{\text{slope}}| > 0$ and $|\hat{\beta}_1^{\text{slope}}| \neq |\hat{\beta}_2^{\text{slope}}|$.

Theorem 2 *Let us apply the transformation* $U := (X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}$ *to the model (1). One obtains the new model* $\tilde{Y} := \beta^* + \tilde{\varepsilon}$ *where* $\tilde{Y} = UY$ *and* $\tilde{\varepsilon} = U\varepsilon$. *Let* $\tilde{\beta}^{\text{lasso}}$ *and* $\tilde{\beta}^{\text{slope}}$ *be the LASSO and SLOPE of this new model, namely*

$$\tilde{\beta}^{\text{lasso}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}}\ \|\tilde{Y} - \beta\|^2 + \lambda\|\beta\|_1 \text{ and } \tilde{\beta}^{\text{slope}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}}\ \|\tilde{Y} - \beta\|^2 + J(\beta).$$

i) *Whatever* $i \in \{1, \dots, p\}$, *the* i^{th} *component of the lasso is* $\tilde{\beta}_i^{\text{lasso}} = \text{sign}(\hat{\beta}_i^{\text{blue}})(|\hat{\beta}_1^{\text{blue}}| - \lambda)_+$.

ii) *Let us denote* $\hat{S}_k := \sum_{i=1}^k \left(|\hat{\beta}_{\tau(i)}^{\text{blue}}| - \lambda_i\right)$, *then whatever* $i \in \{1, \dots, p\}, \tilde{\beta}_i^{\text{slope}}\beta_i^{\text{blue}} \geq 0$ *and*

$$\left(|\tilde{\beta}_{\tau(1)}^{\text{slope}}|, \dots, |\tilde{\beta}_{\tau(p)}^{\text{slope}}|\right) := \Bigg(\underbrace{\left(\frac{\hat{S}_{k_1}}{k_1}\right)_+, \dots, \left(\frac{\hat{S}_{k_1}}{k_1}\right)_+}_{k_1 \text{ components}}, \dots, \underbrace{\left(\frac{\hat{S}_{k_s} - \hat{S}_{k_{s-1}}}{k_s - k_{s-1}}\right)_+, \dots, \left(\frac{\hat{S}_{k_s} - \hat{S}_{k_{s-1}}}{k_s - k_{s-1}}\right)_+}_{k_s - k_{s-1} \text{ components}}\Bigg).$$

Proof: After applying the U transformation, the design matrix of the new model is Id_p (thus orthogonal). Consequently the LASSO has the following expression (3)

$$(\tilde{\beta}_1^{\text{lasso}}, \dots, \tilde{\beta}_p^{\text{lasso}}) := \left(\text{sign}(\hat{\beta}_1^{\text{ols}})(|\hat{\beta}_1^{\text{ols}}| - \lambda)_+, \dots, \text{sign}(\hat{\beta}_{\text{p}}^{\text{ols}})(|\hat{\beta}_{\text{p}}^{\text{ols}}| - \lambda)_+\right).$$

The OLS estimator of the new model is $\hat{\beta}^{\text{ols}} := (Id_p' Id_p)^{-1} Id_p \tilde{Y} = \tilde{Y} = UY = \hat{\beta}^{\text{blue}}$ which provides the explicit expression of the LASSO. The same argument remains true for the SLOPE. □

When the components of ε are correlated, a traditional transformation to recover the BLUE in model (1) is to apply the linear transformation $\Gamma^{-1/2}$. Because $(\Gamma^{-1/2}X)^T(\Gamma^{-1/2}X) = X^T\Gamma^{-1}X = \text{var}(\hat{\beta}^{\text{blue}})^{-1}$, contrarily to the transformation U, the obtained design matrix $\tilde{X} = \Gamma^{-1/2}X$ is not orthogonal. Consequently, after applying the transformation $\Gamma^{-1/2}$, neither the LASSO nor the SLOPE estimators have an explicit expression. In order to relax the irrepresentable condition, Jia et al. [2015] looked at the transformation $F = PD^{-1}P^T$ where the orthogonal matrix P and the diagonal matrix D are given by the singular value decomposition of X. Applying the transformation F, one obtains a new model in which the design is orthogonal. Because after applying F, the OLS estimator $\tilde{\beta}^{\text{ols}}$ of the new model is $\tilde{\beta}^{\text{ols}} = X'Y$, the transformation F does not provide the BLUE contrarily to the transformation U.

3 Conclusion

In this article we proposed a transformation U that allows to get simple and explicit writing for the LASSO and the SLOPE. In addition, our result points out that methods using the LASSO or the SLOPE in small dimension can be derived as methods which only use the BLUE.

4 Appendix: Proof of the theorem 1

First, let us notice that when X is orthogonal the following equivalence holds

$$\hat{\beta}^{\text{slope}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}}\ \|Y - X\beta\|^2 + J(\beta) \Leftrightarrow \hat{\beta}^{\text{slope}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}}\ \|\hat{\beta}^{\text{ols}} - \beta\|^2 + J(\beta).$$

Consequently, to prove the theorem 1, one only needs to provide an explicit expression of the minimizer of the function ϕ defined hereafter

$$\forall x \in \mathbb{R}^p, \phi(x) = \|y - x\|^2 + J(x), \text{ where } y \in \mathbb{R}^p \text{ is a fixed vector.}$$

Let us notice that ϕ is a coercive and strictly convex function thus whatever $y \in \mathbb{R}^p$, ϕ has a unique minimizer. As suggested by the assumption 2.1 in the article of Bogdan et al. [2015], one can restrict the study of the function ϕ to $y_1 \geq y_2 \geq \cdots \geq y_p \geq 0$. Actually finding the minimizer in this particular case allows to recover easily the minimizer of ϕ when y is an arbitrary vector of $\mathbb{R}^p$ as explained in Bogdan et al. [2015]. Let us remind some basic notions of sub-differentiability. Let $\epsilon > 0$, let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function the sub-differential of f at the point $x \in \mathbb{R}^p$ denoted $\partial f(x)$ satisfies the following equivalence

$$s \in \partial f(x) \text{ if } \forall h \in B(0,\epsilon), f(x+h) - f(x) \geq \langle s, h\rangle \Leftrightarrow s \in \partial f(x) \text{ if } \forall h \in \mathbb{R}^p, f(x+h) - f(x) \geq \langle s, h\rangle.$$

The sub-differentiability allows to characterise the minimizer of ϕ (see *e.g* Hiriart-Urruty and Lemaréchal [2013] page 177). The point x^* is a minimizer of ϕ if and only if $\mathbf{0} \in \partial\phi(x^*)$.

The purpose of this part is to prove the proposition 1. The theorem 1 is a straightforward consequence of this proposition.

Proposition 1 *Let $\phi : x \in \mathbb{R}^p \mapsto \|y - x\|^2 + J(x)$ with $y_1 \geq \cdots \geq y_p \geq 0$, let $(S_k)_{1\leq i\leq p}$ be a sequence such that $S_k = \sum_{i=1}^k y_i - \lambda_i$ and let $1 \leq k_1 \leq \cdots \leq k_s = p$ be a partition of $\{1,\ldots,p\}$ such that*

$$k_1 := \max\left\{\underset{k\in\{1,\ldots,p\}}{\text{argmax}}\left\{\frac{S_k}{k}\right\}\right\} \text{ and } \forall i \in \{2,\ldots,s\}, k_i := \max\left\{\underset{k>k_{i-1}}{\text{argmax}}\left\{\frac{S_k - S_{k_{i-1}}}{k - k_{i-1}}\right\}\right\}.$$

Let $c_1 = S_{k_1}/k_1$ and for all $i \in \{2,\ldots,s\}, c_i = (S_{k_i} - S_{k_{i-1}})/(k_i - k_{i-1})$ and let $x^* \in \mathbb{R}^p$ be a vector such that

$$x^* = (\underbrace{(c_1)_+,\ldots,(c_1)_+}_{k_1 \text{ components}}, \underbrace{(c_2)_+,\ldots,(c_2)_+}_{k_2-k_1 \text{ components}}, \ldots, \underbrace{(c_s)_+,\ldots,(c_s)_+}_{k_s-k_{s-1} \text{ components}}), \text{ where } \forall t \in \mathbb{R}, (t)_+ = \max\{t, 0\}.$$

Then the unique minimizer of ϕ is x^*.

To prove the proposition 1, we are going to provide three lemmas. The lemma 1 gives some results about the sub-differential of J and ϕ. The lemma 2 show that $x^* = (0,\ldots,0)$ is the unique solution of ϕ once the sequence $(S_k)_{1\leq k\leq p}$ is negative.

The lemma 3 shows that that $x^* = (S_p/p,\ldots,S_p/p)$ is the unique solution of ϕ once the Cesàro sequence $(S_k/k)_{1\leq k\leq p}$ reaches its maximum at $k = p$ and $S_p > 0$.

Hereafter the SLOPE norm J is also denoted $J_{\lambda_1,\ldots,\lambda_p}$, the set of permutations of $\{1,\ldots,p\}$ is denoted $\mathfrak{S}_p$ and given $u \in \mathbb{R}^p$, the permutation $[\,.\,] \in \mathfrak{S}_p$ is such that $|u_{[1]}| \geq \cdots \geq |u_{[p]}|$.

Lemma 1 *The properties i) and ii) deal with the sub-differential of J and the property iii) deals with the sub-differential of ϕ.*

i) *If $x_1 = \cdots = x_p > 0$ then* $\text{conv}\left((\lambda_{r(1)},\ldots,\lambda_{r(p)})_{r\in\mathfrak{S}_p}\right) \subset \partial J(x)$.

ii) *If $x_1 = \cdots = x_p = 0$ then* $\text{conv}\left(\bigcup_{r\in\mathfrak{S}_p}[-\lambda_{r(1)},\lambda_{r(1)}] \times \cdots \times [-\lambda_{r(p)},\lambda_{r(p)}]\right) \subset \partial J(x)$.

iii) *Let $0 = k_0 \leq k_1 \leq \cdots \leq k_s \leq k_{s+1} = p$ be a partition of $\{1, \ldots, p\}$ such that*

$$x_{k_0+1} = \cdots = x_{k_1} > x_{k_1+1} = \cdots = x_{k_2} > \cdots > x_{k_{s-1}+1} = \cdots = x_{k_s} > x_{k_s+1} = \cdots = x_{k_{s+1}} = 0.$$

Let us define the functions $\phi_1, \ldots, \phi_{s+1}$ as follows

$$\forall j \in \{0, \ldots, s\}, \phi_{j+1}(x_{k_j+1}, \ldots, x_{k_{j+1}}) = \sum_{i=k_j+1}^{k_{j+1}} (y_i - x_i)^2 + J_{\lambda_{k_j+1}, \ldots, \lambda_{k_{j+1}}}(x_{k_j+1}, \ldots, x_{k_{j+1}})$$

Then the sub-differential of ϕ satisfies $\partial\phi_1(x_1, \ldots, x_{k_1}) \times \cdots \times \partial\phi_s(x_{k_{s-1}+1}, \ldots, x_{k_s}) \times \partial\phi_{s+1}(\mathbf{0}) \subset \partial\phi(x)$.

Proof: First, let us prove i). Because whatever $r \in \mathfrak{S}_p$ the two following expressions hold

$$\begin{array}{rcl} J(x+h) & = & \lambda_1|(x+h)_{[1]}| + \cdots + \lambda_p|(x+h)_{[p]}| \geq \lambda_{r(1)}|x_1 + h_1| + \cdots + \lambda_{r(p)}|x_p + h_p| \text{ and} \\ J(x) & = & \lambda_{r(1)}|x_1| + \cdots + \lambda_{r(p)}|x_p|, \end{array}$$

one deduces that

$$J(x+h) - J(x) \geq \lambda_{r(1)}(|x_1 + h_1| - |x_1|) + \cdots + \lambda_{r(p)}(|x_p + h_p| - |x_p|) \geq \lambda_{r(1)}h_1 + \cdots + \lambda_{r(p)}h_p.$$

Consequently, whatever $r \in \mathfrak{S}_p$ we have $(\lambda_{r(1)}, \ldots, \lambda_{r(p)}) \in \partial J(x)$. Furthermore because $\partial J(x)$ is a convex set, one deduces the result i).

Now, let us prove ii), whatever $s_1 \in [-1, 1], \ldots, s_p \in [-1, 1]$ whatever $r \in \mathfrak{S}_p$, the following inequality hold

$$J(h) - J(\mathbf{0}) = \lambda_1|h_{[1]}| + \cdots + \lambda_p|h_{[p]}| \geq \lambda_{r(1)}|h_1| + \cdots + \lambda_{r(p)}|h_p| \geq \lambda_{r(1)}s_1h_1 + \cdots + \lambda_{r(p)}s_ph_p.$$

Thus $[-\lambda_{r(1)}, \lambda_{r(1)}] \times \cdots \times [-\lambda_{r(p)}, \lambda_{r(p)}] \in \partial J(\mathbf{0})$. Because $\partial J(\mathbf{0})$ is a convex set, one deduces the result ii).

Finally, let us show iii). Let $h \in \mathbb{R}^p$ small enough so that whatever $i \in \{1, \ldots, s\}$ the inequality $x_{k_i} - \|h\|_\infty > x_{k_{i+1}} + \|h\|_\infty$ occurs (such a small h insures that the k_1^{th} largest components of $x+h$ are $x_1+h_1, \ldots, x_{k_1}+h_{k_1}$ and so on). As a consequence, the SLOPE norm satisfies the following equality

$$J_{\lambda_1, \ldots, \lambda_p}(x+h) = \sum_{i=0}^{s} J_{\lambda_{k_i+1}, \ldots, \lambda_{k_{i+1}}}(x_{k_i+1} + h_{k_i+1}, \ldots, x_{k_{i+1}} + h_{k_{i+1}}).$$

When h is small enough one deduces that whatever $u \in \partial\phi_1(x_1, \ldots, x_{k_1}) \times \cdots \times \partial\phi_s(x_{k_{s-1}+1}, \ldots, x_{k_s})$,

$$\begin{array}{rcl} \phi(x+h) - \phi(x) & = & \displaystyle\sum_{i=0}^{s} \left(\phi_{i+1}(x_{k_i+1} + h_{k_i+1}, \ldots, x_{k_{i+1}} + h_{k_{i+1}}) - \phi_{i+1}(x_{k_i+1}, \ldots, x_{k_{i+1}})\right) \\ & \geq & \displaystyle\sum_{i=0}^{s} u_{k_i+1}h_{k_i+1} \times \ldots, u_{k_{i+1}}h_{k_{i+1}} = \langle u, h \rangle. \end{array}$$

Consequently, $u \in \partial\phi(x)$ which insures that iii) holds □

Lemma 2 *Let us assume that $\forall i \in \{1, \ldots, p\}, S_i \leq 0$ then the unique minimizer of $\phi : x \in \mathbb{R}^p \mapsto \|y-x\|^2 + J(x)$ is $x^* = (0, \ldots, 0)$.*

Proof: To prove that $x^* = (0, \ldots, 0)$ is a minimizer of ϕ, it suffices to show that $\mathbf{0} \in \partial\phi(x^*)$. Let us gives the following equivalences

$$\mathbf{0} \in \partial\phi(x^*) \Leftrightarrow \mathbf{0} \in -y + x^* + \partial J(x^*) \Leftrightarrow y \in \partial J(\mathbf{0}).$$

By lemma 2, the sub-differential of ϕ at $\mathbf{0}$ contains the set C given hereafter

$$C := \text{conv}\left(\bigcup_{r \in \mathfrak{S}_p} [-\lambda_{r(1)}, \lambda_{r(1)}] \times \cdots \times [-\lambda_{r(p)}, \lambda_{r(p)}]\right) \subset \partial J(\mathbf{0}).$$

Let us remind that a closed convex set is the intersection of all closed half spaces containing it. Let $a_1x_1 + \cdots + a_px_p \geq b$ be an arbitrary closed half space containing C, to prove that $y \in C$, we are going to show that $a_1y_1 + \cdots + a_py_p \geq b$. Let us set $|a_{(1)}| \leq \cdots \leq |a_{(p)}|$ and let us denote $u_i = |a_{(i+1)}| - |a_{(i)}|$ with $i \in \{1, \ldots, p-1\}$.

Because $v := (-\lambda_p \text{sign}(a_{(1)}), \ldots, -\lambda_1 \text{sign}(a_{(p)})) \in C$ and because whatever $r \in \mathfrak{S}_p$, $(v_{r(1)}, \ldots, v_{r(p)}) \in C$, one deduces that $a_{(1)}v_1 + \cdots + a_{(p)}v_p = -\lambda_p|a_{(1)}| - \cdots - \lambda_1|a_{(p)}| \geq b$. The following implications shows that $a_1y_1 + \cdots + a_py_p \geq -\lambda_p|a_{(1)}| - \cdots - \lambda_1|a_{(p)}| \geq b$. We deduce from this last inequality that

$$\begin{aligned}
& a_1y_1 + \cdots + a_py_p \geq -\lambda_p|a_{(1)}| - \cdots - \lambda_1|a_{(p)}|, \\
\Leftrightarrow \quad & a_{(1)}y_{(1)} + \lambda_p|a_{(1)}| + \cdots + a_{(p)}y_{(p)} + \lambda_1|a_{(p)}| \geq 0, \\
\Leftrightarrow \quad & |a_{(1)}| \left(\text{sign}(a_{(1)})y_{(1)} + \lambda_p\right) + \cdots + |a_{(p)}| \left(\text{sign}(a_{(p)})y_{(p)} + \lambda_1\right) \geq 0, \\
\Leftrightarrow \quad & |a_{(1)}| \left(\sum_{i=1}^{p} \lambda_i + \sum_{i=1}^{p} \text{sign}(a_{(i)})y_{(i)} \right) + \sum_{i=1}^{p-1} u_i \left(\sum_{j=1}^{p-i} \lambda_j + \sum_{j=i}^{p} \text{sign}(a_{(j)})y_{(j)} \right) \geq 0.
\end{aligned}$$

The last expression comes from the identity $|a_{(1)}|b_1 + \cdots + |a_{(p)}|b_p = |a_{(1)}|(b_1 + \cdots + b_p) + u_1(b_2 + \cdots + b_p) + \cdots + u_{p-1}b_p$. Finally, the inequality given hereafter insures that $a_1y_1 + \cdots + a_py_p \geq b$. In other terms,

$$|a_{(1)}| \left(\sum_{i=1}^{p} \lambda_i + \sum_{i=1}^{p} \text{sign}(a_{(i)})y_{(i)} \right) + \sum_{i=1}^{p-1} u_i \left(\sum_{j=1}^{p-i} \lambda_j + \sum_{j=i+1}^{p} \text{sign}(a_{(j)})y_{(j)} \right) \geq -|a_{(1)}|S_p - \sum_{i=1}^{p-1} u_iS_i \geq 0.$$

Consequently, $y \in C$ and so $x^* = (0, \ldots, 0)$ is the unique minimizer of ϕ. □

Lemma 3 *Let us assume that $\forall i \in \{1, \ldots, p\}, S_i/i \leq S_p/p$ and $S_p > 0$ then the unique minimizer of $\phi : x \in \mathbb{R}^p \mapsto \|y - x\|^2 + J(x)$ is $x^* = (S_p/p, \ldots, S_p/p)$.*

Proof: To prove that x^* is a minimizer of ϕ, it suffices to show that $\mathbf{0} \in \partial\phi(x^*)$. Let us gives the following equivalences

$$\mathbf{0} \in \partial\phi(x^*) \Leftrightarrow \mathbf{0} \in -y + x^* + \partial J(x^*) \Leftrightarrow y - x^* \in \partial J(x^*).$$

By the lemma 1, $\text{conv}\left((\lambda_{r(1)}, \ldots, \lambda_{r(p)})_{r \in \mathfrak{S}_p}\right) \subset \partial J(x^*)$. Hereafter we are going to show $-y + x^* \in \text{conv}\left((\lambda_{r(1)}, \ldots, \lambda_{r(p)})_{r \in \mathfrak{S}_p}\right)$. Let us remind that a closed convex set is the intersection of all closed half spaces containing it. Let $a_1x_1 + \cdots + a_px_p \geq b$ be an arbitrary closed half space containing $\text{conv}\left((\lambda_{r(1)}, \ldots, \lambda_{r(p)})_{r \in \mathfrak{S}_p}\right)$ to prove that $y - x^* \in \text{conv}\left((\lambda_{r(1)}, \ldots, \lambda_{r(p)})_{r \in \mathfrak{S}_p}\right)$ it suffices to prove that $a_1(y_1 - x_1^*) + \cdots + a_p(y_p - x_p^*) \geq b$. Let us set $a_{(1)} \leq \cdots \leq a_{(p)}$ and let us denote $u_i = a_{(i+1)} - a_{(i)}$ with $i \in \{1, \ldots, p-1\}$. By definition of the half space $a_1x_1 + \cdots + a_px_p \geq b$, an appropriate permutation $r \in \mathfrak{S}_p$ insures that $a_{(1)}\lambda_1 + \cdots + a_{(p)}\lambda_p \geq b$. The following implications shows that $a_1(y_1 - x_1^*) + \cdots + a_p(y_p - x_p^*) \geq a_{(1)}\lambda_1 + \cdots + a_{(p)}\lambda_p \geq b$.

$$\begin{aligned}
& a_1(y_1 - x_1^*) + \cdots + a_p(y_p - x_p^*) \geq a_{(1)}\lambda_1 + \cdots + a_{(p)}\lambda_p, \\
\Leftrightarrow \quad & a_{(1)} \left(y_{(1)} - \frac{S_p}{p} - \lambda_1 \right) + \cdots + a_{(p)} \left(y_{(p)} - \frac{S_p}{p} - \lambda_p \right) \geq 0, \\
\Leftrightarrow \quad & a_{(1)} \underbrace{\left(\sum_{i=1}^{p} y_{(i)} - S_p - \sum_{i=1}^{p} \lambda_i \right)}_{=0} + \sum_{i=1}^{p-1} u_i \left(\sum_{j=i+1}^{p} y_{(j)} - (p-i)\frac{S_p}{p} - \sum_{j=i+1}^{p} \lambda_j \right) \geq 0.
\end{aligned}$$

The last expression comes from the identity $a_{(1)}b_1 + \cdots + a_{(p)}b_p = a_{(1)}(b_1 + \cdots + b_p) + u_1(b_2 + \cdots + b_p) + \cdots + u_{p-1}b_p$. Finally, the inequality given hereafter insures that $a_1(-y_1 + x_1^*) + \cdots + a_p(-y_p + x_p^*) \geq b$.

$$\sum_{i=1}^{p-1} u_i \left(\sum_{j=i+1}^{p} y_{(j)} - (p-i)\frac{S_p}{p} - \sum_{j=i+1}^{p} \lambda_j \right) \geq \sum_{i=1}^{p-1} u_i \left(S_p - S_i - (p-i)\frac{S_p}{p} - \right) = \sum_{i=1}^{p-1} \frac{u_i}{i} \left(\frac{S_p}{p} - \frac{S_i}{i} \right) \geq 0.$$

Consequently, $-y + x^* \in \text{conv}\left((\lambda_{r(1)}, \ldots, \lambda_{r(p)})_{r \in \mathfrak{S}_p}\right)$ thus $x^* = (S_p/p, \ldots, S_p/p)$ is the unique minimizer of ϕ. □

Proof of the proposition 1: First, let us show that $c_1 > c_2 > \cdots > c_s$. By construction $c_1 \geq c_2 \geq \cdots \geq c_s$ thus let us shows that whatever $i \in \{1, \ldots, s-1\}$ the inequality $c_{i+1} = c_i$ cannot occur. Indeed, the following equality always holds

$$\frac{S_{k_{i+1}} - S_{k_{i-1}}}{k_{i+1} - k_{i-1}} = c_{i+1}\frac{k_{i+1} - k_i}{k_{i+1} - k_{i-1}} + c_i\frac{k_i - k_{i-1}}{k_{i+1} - k_{i-1}} \text{ (by setting } k_0 = 0 \text{ and } S_{k_0} = 0).$$

Consequently, if $c_{i+1} = c_i$ thus one deduces that $k_{i+1} \in \underset{k>k_{i-1}}{\operatorname{argmax}} \left\{ \frac{S_k - S_{k_{i-1}}}{k - k_{i-1}} \right\}$. Because $k_{i+1} > k_i$ this contradicts that k_i is the largest element of $\underset{k>k_{i-1}}{\operatorname{argmax}} \left\{ \frac{S_k - S_{k_{i-1}}}{k - k_{i-1}} \right\}$.

First, let us assume that $c_1 > \cdots > c_s > 0$ then the lemma 1 insures that

$$\partial\phi(x^*) = \partial\phi_1(\underbrace{c_1, \ldots, c_1}_{k_1 \text{ components}}) \times \cdots \times \partial\phi_s(\underbrace{c_s, \ldots, c_s}_{k_s - k_{s-1} \text{ components}}).$$

The lemma 3 insures that whatever $i \in \{1, \ldots, s\}$, we have $\mathbf{0} \in \partial\phi_i(c_i, \ldots, c_i)$. Thus $\mathbf{0} \in \partial\phi(x^*)$ which insures that x^* is a minimizer of ϕ.

Now, if $0 \geq c_1 > \cdots > c_s$ then the sequence $(S_i)_{1 \leq i \leq p}$ is negative thus the lemma 2 insures that $x^* = (0, \ldots, 0)$ is a minimizer of ϕ.

Finally, if $c_1 > \cdots > c_{i_0} > 0 \geq c_{i_0+1} > \cdots > c_s$ with $i_0 \in \{1, \ldots, s-1\}$ then the lemma 1 insures that

$$\partial\phi(x^*) = \partial\phi_1(\underbrace{c_1, \ldots, c_1}_{k_1 \text{ components}}) \times \cdots \times \partial\phi_{i_0}(\underbrace{c_{i_0}, \ldots, c_{i_0}}_{k_{i_0} - k_{i_0-1} \text{ components}})\partial \times \phi_{i_0+1}(\mathbf{0}), \text{ with } \phi_{i_0+1} \text{ as in lemma 1.}$$

The lemma 3 insures that whatever $i \in \{1, \ldots, i_0\}$, we have $\mathbf{0} \in \partial\phi_i(c_i, \ldots, c_i)$. Furthermore, because $\forall i > k_{i_0}, (S_i - S_{k_{i_0}}) \leq 0$ the lemma 2 insures that $\mathbf{0} \in \partial\phi_{i_0+1}(\mathbf{0})$. Thus $\mathbf{0} \in \partial\phi(x^*)$ which insures that x^* is a minimizer of ϕ. □

Acknowledgements

This work is part of the project GMO90+ supported by the grant CHORUS 2101240982 from the Ministry of Ecology, Sustainable Development and Energy in the national research program RiskOGM. Patrick Tardivel is partly supported by a phD fellowship from GMO90+. We also received a grant for the project from the IDEX of Toulouse "Transversalité 2014".

References

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5):2055–2085, 2015.

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):1103–1140, 2015.

Peter Bühlmann and Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, 2011. ISBN 3642201911, 9783642201912.

Evgenii Chzhen, Mohamed Hebiri, and Joseph Salmon. On lasso refitting strategies. arXiv preprint arXiv:1707.05232, 2017.

Junbo Duan, Charles Soussen, David Brie, Jérôme Idier, Mingxi Wan, and Yu-Ping Wang. Generalized lasso with under-determined regularization matrices. Signal processing, 127:239–246, 2016.

Alexej Gossmann, Shaolong Cao, and Yu-Ping Wang. Identification of significant genetic variants via slope, and its extension to group slope. In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, pages 232–240. ACM, 2015.

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. Journal of the Royal Statistical Sociey: Series B (Statistical Methodology), 78(2):423–444, 2015.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. Convex Analysis and Minimization Algorithms I: Fundamentals, volume 305. Springer Science & Business Media, 2013.

Lucas Janson and Weijie Su. Familywise error rate control via knockoffs. Electronic Journal of Statistics, 10 (1):960–975, 2016.

Jinzhu Jia, Karl Rohe, et al. Preconditioning the lasso for sign consistency. Electronic Journal of Statistics, 9 (1):1150–1172, 2015.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. The Annals of Statistics, 42(2):413–468, 2014.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3):1436–1462, 2006.

Weijie Su and Emmanuel Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. The Annals of Statistics, 44(3):1038–1068, 2016.

Xiaoying Tian, Joshua R Loftus, and Jonathan E Taylor. Selective inference with unknown variance via the square-root lasso. arXiv preprint arXiv:1504.08031, 2015.

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

Chao-Kai Wen, Jun Zhang, Kai-Kit Wong, Jung-Chieh Chen, and Chau Yuen. On sparse vector recovery performance in structurally orthogonal matrices via lasso. IEEE Transactions on Signal Processing, 64(17): 4519–4533, 2016.

Peng Zhao and Bin Yu. On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563, 2006.

Hui Zou. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101 (476):1418–1429, 2006.