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

Rémi Servien, Patrick J C Tardivel, Didier Concordet

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

Patrick J.C. Tardivel , Rémi Servien and Didier Concordet

INTHERES, Université de Toulouse, INRA, ENVT, Toulouse, France.

ARTICLE HISTORY
Compiled January 29, 2019

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 low-dimensional linear model

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

where X is a $n \times p$ fixed design matrix with $\ker(X) = \mathbf{0}$ (*i.e.* $n \geq p$), $\beta^* \in \mathbb{R}^p$ 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 and the Sorted L-One Penalized Estimation (SLOPE) estimator respectively introduced by Tibshirani [1] and Bogdan et al. [2] are defined by

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

$$\hat{\beta}^{\text{slope}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda_1 |\beta_{[1]}| + \cdots + \lambda_p |\beta_{[p]}| \right\}. \tag{3}$$

In the second expression, the tuning parameters $(\lambda_i)_{1\leq i\leq p}$ satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. The brackets $[\,.\,]$ denote 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 (*i.e* $X'X = Id_p$), the LASSO estimator leads to the following soft-thresholded Ordinary Least Squares (OLS) esti-

CONTACT Patrick Tardivel. Email: tardivel@math.uni.wroc.pl

mator [1]

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

Popularized by the pioneer work of Tibshirani, the orthogonal design became a case study [3–8]. Furthermore some properties such as the irrepresentable condition [9–12] hold when X is orthogonal. The orthogonal design is also a case study for the SLOPE estimator [2,13,14].

Whatever the considered estimator, the orthogonal design setting appears to be an ideal case. By seeking to generalize its properties to the non-orthogonal setting, we discovered a relevant orthogonalizing transformation U. Actually, if we consider the new model $\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon}$, where $\tilde{Y} = UY$, $\tilde{X} = UX$ and $\tilde{\varepsilon} = U\varepsilon$, the LASSO estimator $\tilde{\beta}$ can simply be written as a function of the Best Linear Unbiased Estimator (BLUE) in the following way:

$$\tilde{\beta} = \left(\text{sign}(\hat{\beta}_i^{\text{blue}})(|\hat{\beta}_i^{\text{blue}}| - \lambda s_i)_+\right)_{1\leq i\leq p} \text{ where } s_1 > 0, \ldots, s_p > 0. \tag{5}$$

Similarly to the LASSO estimator, we also obtained a simple expression for the SLOPE estimator based on the BLUE. In low-dimension, the U transformation giving (5) is also available when 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 estimator $\tilde{\beta}$ as described in (5).

- Contrarily to the LASSO estimator $\hat{\beta}^{\text{lasso}}$ obtained when X is orthogonal and $\varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n)$, the components of $\tilde{\beta}$ given by (5) are in general not independent.
- From methods based on the LASSO estimator one derives methods based on the BLUE. As an example, one can derive from the multiple testing procedure based on the knockoff-LASSO estimator [15] a new procedure based on the BLUE. The knockoff-LASSO is the following estimator

$$\hat{\beta}^{\text{kn-lasso}} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \frac{1}{2}\|Y - X_{\text{ko}}\beta\|^2 + \lambda\|\beta\|_1.$$

  Since X_{ko} satisfies $\ker(X_{\text{ko}}) = \mathbf{0}$ [16], 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 for example [2] 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

$$\hat{\beta}^{\text{ols}} := (X'X)^{-1}X'Y \text{ and } \hat{\beta}^{\text{blue}} := (X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}Y. \tag{6}$$

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 provide the SLOPE estimation [2] but the estimator writing is not explicit. To our knowledge, there does not exist currently any explicit formula for the SLOPE. In the following theorem, we provide the explicit expression of the SLOPE when X is orthogonal.

Theorem 2.1. *Let τ be a permutation of $\{1,\dots,p\}$ ordering the components of the OLS estimator (6) 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 $\forall k \in \{1,\dots,p\}, \hat{S}_k := \sum_{i=1}^{k}\left(|\hat{\beta}^{\text{ols}}_{\tau(i)}| - \lambda_i\right)$ and let $1 \leq k_1 \leq \cdots \leq k_s = p$ be a partition of $\{1,\dots,p\}$ such that*

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

When the design matrix X is orthogonal, whatever $i \in \{1,\dots,p\}$, the components of $\hat{\beta}^{\text{slope}}$ (2) satisfy the inequality $\hat{\beta}^{\text{ols}}_i\hat{\beta}^{\text{slope}}_i \geq 0$ and $\left(|\hat{\beta}^{\text{slope}}_{\tau(1)}|,\dots,|\hat{\beta}^{\text{slope}}_{\tau(p)}|\right)$ is given by

$$\Bigg(\underbrace{\left(\frac{\hat{S}_{k_1}}{k_1}\right)_+,\dots,\left(\frac{\hat{S}_{k_1}}{k_1}\right)_+}_{k_1 \text{ components}},\dots,\underbrace{\left(\frac{\hat{S}_{k_s}-\hat{S}_{k_{s-1}}}{k_s-k_{s-1}}\right)_+,\dots,\left(\frac{\hat{S}_{k_s}-\hat{S}_{k_{s-1}}}{k_s-k_{s-1}}\right)_+}_{k_s-k_{s-1} \text{ components}}\Bigg).$$

Let us notice that when $\hat{\beta}^{\text{ols}}$ has a continuous distribution over $\mathbb{R}^p$, the Cesàro sequence $(\hat{S}_k/k)$ almost surely reaches its maximum at a unique point. In other terms, $k_1 := \text{argmax}\,\{\hat{S}_k/k\}$ is unique and the same argument applies for $k_2,\dots,k_s$. When $\lambda_1 = \cdots = \lambda_p = \lambda$, the Cesàro sequence is non-increasing and consequently the following equality holds

$$\forall i \in \{1,\dots,p\}, \hat{\beta}^{\text{slope}}_i = \text{sign}(\hat{\beta}^{\text{ols}}_i)(|\hat{\beta}^{\text{ols}}_i| - \lambda)_+ = \hat{\beta}^{\text{lasso}}_i.$$

As a consequence, when $\lambda_1 = \cdots = \lambda_p = \lambda$ in the orthogonal setting, the formula of the SLOPE given in the theorem 2.1 coincides with the one of the LASSO given in (4).

When X is orthogonal, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 Id_n)$ and $\lambda_i = \sigma z(1 - iq/(2p)), i \in \{1,\dots,p\}$ (where $q \in (0,1)$ and $z(\eta)$ is the $\eta-$quantile of the $\mathcal{N}(0,1)$ distribution), the procedure rejecting the null hypothesis $\beta^*_i = 0$ when $\hat{\beta}^{\text{slope}}_i \neq 0$ controls the FDR at level q [2]. The explicit expression of the SLOPE shows that this procedure is close to the original Benjamini-Hochberg procedure [17] (actually, these procedures are equal when the sequence $(\hat{S}_k/k)$ is decreasing) and this provides an intuitive explanation for the FDR control.

An illustration of the relationship between the OLS estimator and the SLOPE estimator when X is orthogonal is given by the figure 1 in the special case where $p = 2$, $\lambda_1 = 2$ and $\lambda_2 = 1$.
This figure also illustrates the properties of the SLOPE: this estimator is sparse (*i.e.* some components are exactly equal to 0), and some components are equal in absolute value.

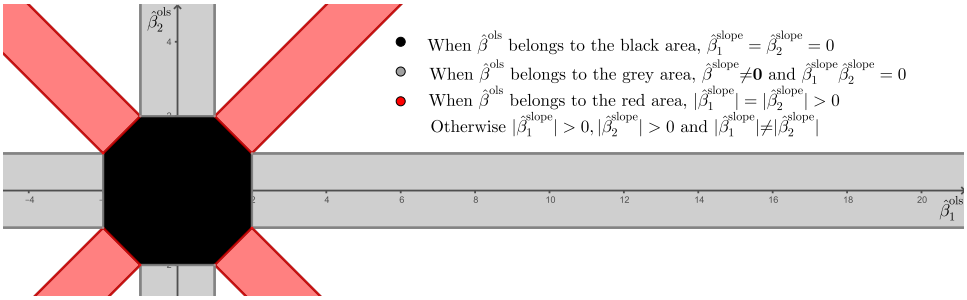

Figure 1. This figure illustrates the relationship between the OLS estimator and the SLOPE, the x-axis and y-axis represent respectively the first and second component of the OLS estimator. Let $\hat{S}_1, \hat{S}_2$ be defined as in the theorem 2.1. When $\hat{S}_1 \leq 0$ and $\hat{S}_2 \leq 0$ then $\hat{\beta}^{\text{ols}}$ 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 $|\hat{\beta}_1^{\text{slope}}| \neq |\hat{\beta}_2^{\text{slope}}|$.

Remarks: Let us point out some relevant transformations allowing us to obtain simple expressions for the LASSO estimator and the SLOPE:

- **Transformation which brings back to the orthogonal setting:** When X is not orthogonal, applying the transformation $U := (X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}$ on each side of the model (1) brings back the orthogonal setting in which $\tilde{Y} := \beta^* + \tilde{\varepsilon}$, where $\tilde{Y} = UY$ and $\tilde{\varepsilon} = U\varepsilon$. In this new model, the LASSO estimator has the expression described in (5) with $s_1 = \cdots = s_5 = 1$ and the SLOPE has the expression described in the theorem 2.1 except that the OLS estimator is substituted by the more accurate BLUE estimator.

- **Transformation which brings back the orthogonal columns setting:** When columns of X are not orthogonal (*i.e.* when $X'X$ is not diagonal), applying the transformation $U_{s_1,\ldots,s_p} := D(X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}$, where $D := \text{diag}(1/\sqrt{s_1}, \ldots, 1/\sqrt{s_p})$ and $s_1 > 0, \ldots, s_p > 0$, on each side of the model (1) returns the orthogonal columns' setting. After applying this transformation, LASSO's expression is still explicit and its expression is described by formula (5).

To avoid confusion, we call soft-thresholded BLUE the LASSO-type estimator whose formula is given in (5) and LASSO the estimator solution of (2). The soft-thresholded BLUE is easier to compute than the LASSO estimator (because the LASSO estimator computation needs to solve numerically the optimization problem described in (2)) and the soft-thresholded BLUE estimator is also easier to interpret than the LASSO estimator (contrarily to the LASSO estimator, the relationship between BLUE and the soft-thresholded BLUE is explicit). Thus it is recommended that in low-dimension the soft-thresholded BLUE should be used instead of the LASSO estimator, except if there are particular reasons for using the latter. If the goal is to estimate the non-null components of β^*, the soft-thresholded BLUE outperforms the LASSO, solution of (2), as evidenced by the following example.

Example: Let ε be a Gaussian vector of $\mathbb{R}^2$ having a $\mathcal{N}(0, Id_2)$ distribution, let $\beta^* = (t, 0) \in \mathbb{R}^2$ with $t \neq 0$, let X be the design matrix described hereafter and let $Y = X\beta^* + \varepsilon$.

$$X = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}' \text{ and thus } \hat{\beta}^{\text{blue}} \sim \mathcal{N}\left(\begin{pmatrix} t \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}\right).$$

When $u \in \mathbb{R}^p$, let us set $\text{supp}(u) := \{i \in \{1, \dots, p\} \mid u_i \neq 0\}$. Now let us denote respectively $\hat{\beta}^{\text{lasso}}(\lambda)$ and $\tilde{\beta}(\lambda)$ the LASSO estimator and the soft-thresholded BLUE to stress that these estimators depend on λ. Whatever $\lambda > 0$, and even if $|t|$ is infinitely large, the following inequality holds

$$\mathbb{P}\left(\text{supp}(\hat{\beta}^{\text{lasso}}(\lambda)) = \text{supp}(\beta^*)\right) \leq 1/2.$$

This inequality is illustrated in figure 2 when $\lambda \in \{1/2, 1\}$. More explanations about this inequality can be found in Wainwright [18] or in Tardivel and Bogdan [19].

On the other hand, if we take $s_1 = \sqrt{5}, s_2 = 1$ and λ as the $1 - \alpha$ quantile of a $\mathcal{N}(0, 1)$ distribution in (5), we have $\mathbb{P}(\text{supp}(\tilde{\beta}(\lambda)) = \text{supp}(\beta^*)) \approx 1 - \alpha$ when $|t|$ is large. Consequently, by using the soft-thresholded BLUE, the probability to recover $\text{supp}(\beta^*)$ can be arbitrarily close to 1 when holds: $|t|$ is large and α is small.

3. Conclusion

It can be seen in this article that up to a transformation, the LASSO and the SLOPE have simple and explicit writing. In addition, our results point out that methods using the LASSO or the SLOPE in low-dimension can be derived as methods which only use the BLUE.

Acknowledgements

Patrick Tardivel thanks Professor Bogdan to allow him to continue his research at the Mathematics Institute of Wrocław University. Authors thank Artur Bogdan for his careful reading.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

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'.

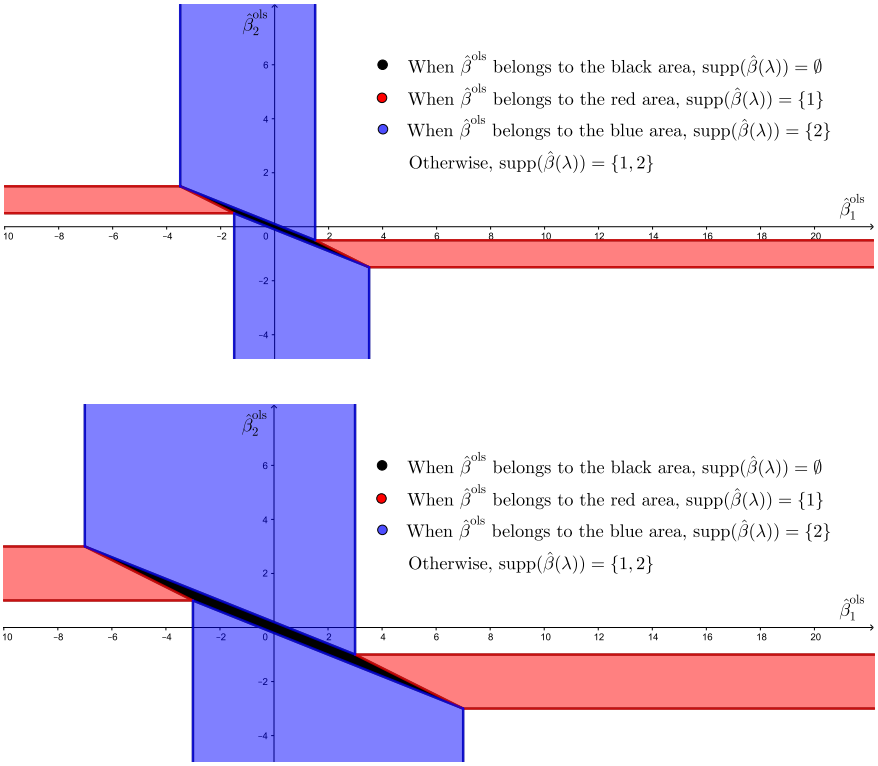

Figure 2. These figures provide the relationship between $\hat{\beta}^{\text{ols}}$ and the set of non-null components of the LASSO estimator $\text{supp}(\hat{\beta}^{\text{lasso}}(\lambda))$ (as explained in [20]) when $\lambda = 0.5$ (above) and $\lambda = 1$ (below). The x-axis and y-axis represent respectively the first and second component of the OLS estimator. One may notice that the estimator $\text{supp}(\hat{\beta}^{\text{lasso}}(\lambda))$ recovers the true set $\{1\}$ when $\hat{\beta}^{\text{ols}}$ is in the red area. Because $\hat{\beta}^{\text{ols}}$ is in the red area, this implies that $\beta_1^{\text{ols}}\beta_2^{\text{ols}} \leq 0$, and because $\mathbb{P}(\hat{\beta}_1^{\text{ols}}\hat{\beta}_2^{\text{ols}} \leq 0) \leq 1/2$, one may deduce that $\mathbb{P}(\text{supp}(\hat{\beta}^{\text{lasso}}(\lambda)) = \{1\}) \leq 1/2$.

References

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

[2] Bogdan M, van den Berg E, Sabatti C, et al. Slope - adaptive variable selection via convex optimization. The Annals of Applied Statistics. 2015;9(3):1103–1140.

[3] Chzhen E, Hebiri M, Salmon J. On lasso refitting strategies. arXiv preprint arXiv:170705232. 2017;.

[4] Duan J, Soussen C, Brie D, et al. Generalized lasso with under-determined regularization matrices. Signal processing. 2016;127:239–246.

[5] G'Sell MG, Wager S, Chouldechova A, et al. Sequential selection procedures and false discovery rate control. Journal of the Royal Statistical Sociey: Series B (Statistical Methodology). 2015;78(2):423–444.

[6] Lockhart R, Taylor J, Tibshirani RJ, et al. A significance test for the lasso. The Annals of Statistics. 2014;42(2):413–468.

[7] Tian X, Loftus JR, Taylor JE. Selective inference with unknown variance via the square-root lasso. arXiv preprint arXiv:150408031. 2015;.

[8] Wen CK, Zhang J, Wong KK, et al. On sparse vector recovery performance in structurally orthogonal matrices via lasso. IEEE Transactions on Signal Processing. 2016;64(17):4519–4533.

[9] Bühlmann P, van de Geer S. Statistics for high-dimensional data: Methods, theory and

applications. Springer; 2011.
[10] Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006;34(3):1436–1462.
[11] Zhao P, Yu B. On model selection consistency of lasso. The Journal of Machine Learning Research. 2006;7:2541–2563.
[12] Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006;101(476):1418–1429.
[13] Gossmann A, Cao S, Wang YP. 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; ACM; 2015. p. 232–240.
[14] Su W, Candes E. Slope is adaptive to unknown sparsity and asymptotically minimax. The Annals of Statistics. 2016;44(3):1038–1068.
[15] Janson L, Su W. Familywise error rate control via knockoffs. Electronic Journal of Statistics. 2016;10(1):960–975.
[16] Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. The Annals of Statistics. 2015;43(5):2055–2085.
[17] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;:289–300.
[18] Wainwright MJ. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1 constrained quadratic programming (lasso). IEEE transactions on information theory. 2009;55(5):2183–2202.
[19] Tardivel PJ, Bogdan M. On the sign recovery given by the thresholded lasso and thresholded basis pursuit. arXiv preprint arXiv:181205723. 2018;.
[20] Schneider U, Ewald K. On the distribution and model selection properties of the lasso estimator in low and high dimensions. arXiv preprint arXiv:170809608. 2017;.
[21] Boyd S, Vandenberghe L. Convex optimization. Cambridge university press; 2004.
[22] Hiriart-Urruty JB, Lemaréchal C. Convex analysis and minimization algorithms i: Fundamentals. Vol. 305. Springer Science & Business Media; 2013.

Appendix A. Proof of the theorem 2.1

Sketch of proof

Theorem 2.1 is a straightforward consequence of proposition A.1. Our goal is thus to prove this proposition, which provides an explicit expression for the minimizer of $x \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - x\|^2 + J(x)$ (where $y \in \mathbb{R}^p$ is a fixed vector such that $y_1 \geq y_2 \geq \cdots \geq y_p \geq 0$).

Looking at the output of the algorithm described in [2] allowed us to conjecture the following two minor results:

1) The null vector is the unique minimizer of $x \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - x\|^2 + J(x)$ when $(S_k)_{1\leq k\leq p}$ is negative.

2) The vector $(S_p/p, \ldots, S_p/p)$ is the unique minimizer of $x \in \mathbb{R}^p \mapsto \frac{1}{2}\|y-x\|^2+J(x)$ when the Cesàro sequence $(S_k/k)_{1\leq k\leq p}$ reaches its maximum at $k = p$ and when $S_p > 0$.

Statements 1) and 2) are proved respectively in lemma A.3 and in lemma A.4. These two lemmas are the keystones of the proof of proposition A.1.

In these lemmas we need to prove that an element u belongs to a closed convex set C. To prove this belonging, we use the fact that C is the intersection of all half-spaces that contain it (see [21] page 49). Consequently, if $a_1x_1 + \cdots + a_px_p \geq b$ is an

arbitrary half-space containing C then, to show that $u \in C$, it is enough to prove that $a_1u_1 + \cdots + a_pu_p \geq b$.

Let $v \in \mathbb{R}^p$, let $[.]$ be a permutation so that $|v_{[1]}| \geq \cdots \geq |v_{[p]}|$, and let r be an arbitrary permutation of $\{1, \ldots, p\}$. One of the key inequalities in lemma A.2 is the following one

$$\lambda_1|v_{[1]}| + \cdots + \lambda_p|v_{[p]}| \geq \lambda_{r(1)}|v_1| + \cdots + \lambda_{r(p)}|v_p|, \text{ where } \lambda_1 \geq \cdots \geq \lambda_p > 0.$$

Proof of the proposition A.1

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

$$\hat{\beta}^{\text{slope}} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2}\|Y - X\beta\|^2 + J(\beta) \Leftrightarrow \hat{\beta}^{\text{slope}} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2}\|\hat{\beta}^{\text{ols}} - \beta\|^2 + J(\beta).$$

Consequently, to prove theorem 2.1, one only needs to provide an explicit expression of the minimizer of the function ϕ defined hereafter

$$\forall x \in \mathbb{R}^p, \phi(x) = \frac{1}{2}\|y - x\|^2 + J(x), \text{ where } y \in \mathbb{R}^p \text{ is a fixed vector.}$$

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 assumption 2.1 in the article of [2], 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 makes it possible to recover easily the minimizer of ϕ when y is an arbitrary vector of $\mathbb{R}^p$ as explained in [2]. 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 $\partial f(x)$ is the convex set described hereafter.

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

The sub-differentiability makes it possible to characterise the minimizer of ϕ (see *e.g* [22] page 177). The point x^* is a minimizer of ϕ if and only if $\mathbf{0} \in \partial\phi(x^*)$.

Proposition A.1. *Let $\phi : x \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - x\|^2 + J(x)$ 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

$$x^* = (\underbrace{(c_1)_+, \ldots, (c_1)_+}_{k_1 \text{ components}}, \underbrace{(c_2)_+, \ldots, (c_2)_+}_{k_2 - k_1 \text{ components}}, \ldots, \underbrace{(c_s)_+, \ldots, (c_s)_+}_{k_s - k_{s-1} \text{ components}}).$$

Then the unique minimizer of ϕ is x^*.

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 A.2. *Properties i) and ii) deal with the sub-differential of J and property iii) deals with the sub-differential of ϕ.*

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

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

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 $j \in \{0,\ldots,s\}$ and let us define functions $\phi_1,\ldots,\phi_{s+1}$ as follows

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

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

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

one may deduce 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 may deduce 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 holds

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

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 may deduce result ii).

Finally, let us show iii). Let $h \in \mathbb{R}^p$ be 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 ensures 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 may deduce that whatever $u\in\partial\phi_1(x_1,\ldots,x_{k_1})\times\cdots\times\partial\phi_s(x_{k_{s-1}+1},\ldots,x_{k_s})$ then $u\in\partial\phi(x)$. Indeed $\phi(x+h)-\phi(x)$ is equal to

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

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

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

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

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 (.) be a permutation of $\{1,\ldots,p\}$ such that $|a_{(1)}|\leq\cdots\leq|a_{(p)}|$ and let us denote $u_i=|a_{(i+1)}|-|a_{(i)}|$ with $i\in\{1,\ldots,p-1\}$. Because $v:=(-\lambda_p\operatorname{sign}(a_{(1)}),\ldots,-\lambda_1\operatorname{sign}(a_{(p)}))\in C$ and because whatever $r\in\mathfrak{S}_p$, $(v_{r(1)},\ldots,v_{r(p)})\in C$, one may deduce that $a_{(1)}v_1+\cdots+a_{(p)}v_p=-\lambda_p|a_{(1)}|-\cdots-\lambda_1|a_{(p)}|\geq b$. The following implications show 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

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

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, because $y_1 \geq \cdots \geq y_p \geq 0$ one may deduce the inequality given hereafter which ensures that $a_1y_1+\cdots+a_py_p \geq b$. In other terms,

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

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

Lemma A.4. *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* $\phi : x \in \mathbb{R}^p \mapsto \frac{1}{2}\|y-x\|^2+J(x)$ *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 A.2, $\text{conv}\left((\lambda_{r(1)},\ldots,\lambda_{r(p)})_{r\in\mathfrak{S}_p}\right) \subset \partial J(x^*)$. 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 $(.)$ be a permutation of $\{1,\ldots,p\}$ such that $a_{(1)} \leq \cdots \leq a_{(p)}$ 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 $r \in \mathfrak{S}_p$ ensures that $a_{(1)}\lambda_1+\cdots+a_{(p)}\lambda_p \geq b$. The following implications show that $a_1(y_1-x_1^*)+\cdots+a_p(y_p-x_p^*) \geq a_{(1)}\lambda_1+\cdots+a_{(p)}\lambda_p \geq b$.

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

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, because $y_1 \geq \cdots \geq y_p \geq 0$ and because whatever $i \in \{1,\ldots,p\}, S_i/i \leq S_p/p$ one may deduce the inequalities given hereafter

which ensures that $a_1(-y_1+x_1^*)+\cdots+a_p(-y_p+x_p^*)\geq b$.

$$\begin{aligned}\sum_{i=1}^{p-1}u_i\left(\sum_{j=i+1}^{p}y_{(j)}-(p-i)\frac{S_p}{p}-\sum_{j=i+1}^{p}\lambda_j\right) &\geq \sum_{i=1}^{p-1}u_i\left(\sum_{j=i+1}^{p}(y_j-\lambda_j)-(p-i)\frac{S_p}{p}\right),\\ &\geq \sum_{i=1}^{p-1}u_i\left(S_p-S_i-(p-i)\frac{S_p}{p}\right),\\ &\geq \sum_{i=1}^{p-1}iu_i\left(\frac{S_p}{p}-\frac{S_i}{i}\right)\geq 0.\end{aligned}$$

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

Proof of the proposition A.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 show 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 $c_{i+1}=c_i$ one may deduce 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 k_i being 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 A.2 ensures that

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

Lemma A.4 ensures 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 ensures 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 lemma A.3 ensures 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 lemma A.2 ensures that

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

with ϕ_{i_0+1} as in lemma A.2. Lemma A.4 ensures that whatever $i\in\{1,\ldots,i_0\}$, we have $\mathbf{0}\in\partial\phi_i(c_i,\ldots,c_i)$. Furthermore, because $\forall i>k_{i_0},(S_i-S_{k_{i_0}})\leq 0$, lemma A.3 ensures that $\mathbf{0}\in\partial\phi_{i_0+1}(\mathbf{0})$. Thus $\mathbf{0}\in\partial\phi(x^*)$, which ensures that x^* is a minimizer of ϕ. □