The Geometry of Model Recovery by Penalized and Thresholded Estimators

Patrick J C Tardivel, Tomasz Skalski, Piotr Graczyk, Ulrike Schneider

The Geometry of Model Recovery by Penalized and Thresholded Estimators

Patrick J.C. Tardivel[1], Tomasz Skalski[2,3], Piotr Graczyk[2], and Ulrike Schneider [4]

[1]Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, F-21000 Dijon, France.
[2]Université d'Angers, Angers, France
[3]Politechnika Wrocławska, Wrocław, Poland
[4]TU Wien, Wien, Austria

June 16, 2021

1 Introduction

Let us consider the linear regression model

$$Y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n\times p}$ is a design matrix, $\varepsilon \in \mathbb{R}^n$ represents a random noise having a symmetric and continuous distribution with a positive density on $\mathbb{R}^p$ (for instance, one may think that ε has iid $\mathcal{N}(0, \sigma^2)$ entries), and $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients.

Many penalized estimators of β have been broadly studied in literature *e.g.* LASSO (Chen and Donoho, 1994; Tibshirani, 1996), SLOPE (Zeng and Figueiredo, 2014; Bogdan et al., 2015), OSCAR (Bondell and Reich, 2008), Fused LASSO (Tibshirani et al., 2005), Clustered LASSO (She et al., 2010), generalized LASSO (Tibshirani et al., 2011)... When the loss function is the residual sum of squares these estimators minimize, with respect to $b \in \mathbb{R}^p$, a function of the type: $b \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - Xb\|_2^2 + \lambda \mathrm{pen}(b)$. Furthermore, the penalty term "pen" is a polyhedral gauge: a non-negative convex function which is the maximum of a finite family of linear functions and can be thus expressed as follows

$$\forall x \in \mathbb{R}^p, \mathrm{pen}(x) := \max\{0, u_1'x, \dots, u_l'x\}, \text{ for some } u_1, \dots, u_l \in \mathbb{R}^p.$$

For the above estimators, "pen" is a polyhedral norm which unit ball is a polytope except for generalized LASSO where the polyhedral gauge $\mathrm{pen} : b \in \mathbb{R}^p \mapsto \|Db\|_1$ is a semi-norm when $\{0\} \subsetneq \ker(D)$ and the unit ball $\{b \in \mathbb{R}^p : \mathrm{pen}(b) \leq 1\}$ is an unbounded polyhedron.

The most famous polyhedral penalty term is the ℓ_1 norm. This norm promoting sparsity is associated to the LASSO estimator and it is well known that this estimator has at most $\mathrm{rk}(X)$ non-null components (Osborne et al., 2000; Tibshirani, 2013). Sparsity consists in appearance of zeros in the estimator $\hat{\beta}$ and is easy to interpret: separation of relevant variables (columns of X associated to non-null components of $\hat{\beta}$) from irrelevant variables (columns of X associated to null components of $\hat{\beta}$). The literature related to penalized least squares estimators is vast and many of these estimators have interesting and relevant structures as illustrated in Vaiter et al. (2015). For instance, the Fused LASSO is a sparse and piecewise constant estimator (Tibshirani et al., 2005), the supremum norm promotes a flat estimator (some components are maximal in absolute value) (Jégou et al., 2012), and

SLOPE as well as OSCAR estimators have some clusters; namely some components of these estimators are equal in absolute value (Bondell and Reich, 2008; Figueiredo and Nowak, 2016; Schneider and Tardivel, 2020).

Even if most of theoretical results of the paper are valid when the penalized least squares estimator is not uniquely defined, we also discuss the important issue of uniqueness.

1.1 Uniform uniqueness for polyhedral gauge penalty

It is well known that a penalized least squares estimator may have a non-unique minimizer. In statistics Y is random and thus, there is an issue to provide a condition for uniqueness that is valid for all $Y \in \mathbb{R}^n$. A well known notion for uniform uniqueness (*i.e.* uniqueness valid for all $Y \in \mathbb{R}^n$), first outlined by Rosset et al. (2004); Dossal (2012), is for the design matrix X to be in general position. Actually, X being in general position is a sufficient condition for uniform uniqueness of the LASSO estimator (Tibshirani, 2013). The general position condition was also recently extended as a sufficient condition for uniform uniqueness of generalized LASSO (Ali and Tibshirani, 2019). For LASSO, Ewald and Schneider (2020) relaxed the general position condition on X on a necessary and sufficient condition for uniform uniqueness. Recently, this condition was extended as a necessary and sufficient condition for uniform uniqueness for penalized least squares estimators where the penalty term is a polyhedral norm like SLOPE, OSCAR, Fused LASSO, Clustered LASSO... (Schneider and Tardivel, 2020). Moreover, this last article illustrates that the conditions on uniform uniqueness provide insights on structures induced by LASSO and SLOPE as well as sign pattern and models for SLOPE identified by these two estimators. Specifically, we have:

i) The uniform uniqueness does not hold if and only if the row span of X intersects a face of the unit ball in dual norm (cube $[-1,1]^p$ for LASSO, sign permutahedron for SLOPE) having a dimension smaller than $\dim(\ker(X))$.

ii) Consider the LASSO estimator. A face of the cube $F = E_1 \times \cdots \times E_p$ where $E_i \in \{\{1\}, \{-1\}, [-1,1]\}$ is labeled by a sign vector $s \in \{-1,0,1\}^p$ (actually s is the barycenter of F) and there exists $y \in \mathbb{R}^n$ such that $\mathrm{sign}(\hat{\beta}^{\mathrm{lasso}}(y)) = s$ (*i.e.* the sign vector s is accessible for LASSO) if and only if the row span of X intersects F.

The dimension of F is the number of null components of s and thus based on statements i) and ii) one may notice that under the uniform uniqueness condition, the number of non-null components for LASSO is smaller than $\mathrm{rk}(X)$. This fact corroborates some well known results for LASSO estimator (Osborne et al., 2000; Tibshirani, 2013). Similar results are available for SLOPE. In particular, under the uniform uniqueness condition, the number of non-null clusters for SLOPE is smaller than $\mathrm{rk}(X)$ (Schneider and Tardivel, 2020) (a cluster is a set of components equal in absolute value). This statement corroborates a similar property given in Kremer et al. (2019).

Theorem 1 in this article provides a necessary and sufficient condition for uniform uniqueness of penalized least squares estimators when the penalty term is a polyhedral gauge. Moreover geometrical objects involved in this theorem (the row span of X and the subdifferential of pen at 0) provide insights for the development of the theory of model pattern recovery.

1.2 Model pattern recovery by penalized least squares estimators

Given $y \in \mathbb{R}^n$ and $\lambda > 0$, the set $S_{X,\lambda\mathrm{pen}}(y)$ of minimizers of a penalized least squares optimization problem is defined as follows:

$$S_{X,\lambda\mathrm{pen}}(y) := \underset{b \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y - Xb\|_2^2 + \lambda \mathrm{pen}(b). \tag{1}$$

Note that $S_{X,\lambda\mathrm{pen}}(y)$ defined in (1) is not empty; this fact is clear when pen is a norm. For generalized LASSO, the existence of a minimizer is proven in Ali and Tibshirani (2019); Dupuis and Vaiter (2019). Note that, potentially, this set is not a singleton.

To clarify the notion of "structure induced by a polyhedral gauge penalty" we are going to introduce the notion of model pattern.

Definition 1 (Model pattern)**.** *Let* pen $: \mathbb{R}^p \mapsto \mathbb{R}$ *be a polyhedral gauge. We say that* $x \in \mathbb{R}^p$ *and* $z \in \mathbb{R}^p$ *have the same model pattern with respect to* pen *when* $\partial_{\text{pen}}(x) = \partial_{\text{pen}}(z)$*, where* ∂_{pen} *represents the subdifferential of* pen*. The structure induced by a polyhedral gauge penalty* pen *is the set of all possible model patterns (or, more formally, the quotient space* $\mathbb{R}^p / \sim_{\text{pen}}$ *where* $\sim_{\text{pen}}$ *represents the equivalence relation for equality of subdifferentials).*

For the ℓ_1 norm two vectors $x, z \in \mathbb{R}^p$ have the same model pattern if and only if $\text{sign}(x) = \text{sign}(z)$. More generally, two vectors having the same model pattern with respect to a polyhedral gauge penalty share a specific structure as illustrated on many examples in section 1.6. Given X and Y, we aim at recovering the model pattern of β; for LASSO this means recovering $\text{sign}(\beta)$.

In this article, Theorem 2 gives a necessary condition (called path condition) for model pattern recovery by penalized least squares estimators. Later, in section 4, we will introduce penalized estimators relaxing this condition. Beforehand, we are going to summarize and to illustrate well known necessary conditions for sign recovery by LASSO.

1.2.1 Sign recovery by LASSO

We note $\hat{\beta}^{\text{lasso}}$ an element of $S_{X,\lambda\|.\|_1}(Y)$ (and we implicitly assume that $S_{X,\lambda\|.\|_1}(Y)$ is a singleton in this section). Of course, LASSO estimator depends on X, λ and Y and, when it is relevant, one may emphasise these dependencies. As mentioned above, the LASSO estimator is a sparse method that nullifies some of the components with positive probability, entailing that the estimator also performs so-called variable selection. Instigated by this sparsity property, an abundant literature has arisen to deal with the recovery of the location of the non-null components of β, or, more specifically, the recovery of the sign vector of β (Fuchs, 2005; Meinshausen and Bühlmann, 2006; Wainwright, 2009; Zhao and Yu, 2006; Zou, 2006).

A natural necessary condition for sign recovery by LASSO is for $\text{sign}(\beta)$ to be accessible by the LASSO, i.e. for a fixed $\lambda > 0$, there has to exist $y \in \mathbb{R}^n$ for which $\text{sign}(\hat{\beta}^{\text{lasso}}(y)) = \text{sign}(\beta)$. Otherwise, $\mathbb{P}_\varepsilon(\text{sign}(\hat{\beta}^{\text{lasso}}(Y)) = \text{sign}(\beta)) = 0$, and sign recovery is clearly impossible. A geometrical characterization of accessible sign vectors is given in Sepehri and Harris (2017); Schneider and Tardivel (2020).

When $\text{sign}(\beta)$ is accessible then the probability of sign recovery is not null (as soon as the set $\{y \in \mathbb{R}^n : \text{sign}(\hat{\beta}^{\text{lasso}}(y)) = \text{sign}(\beta)\}$ is not Lebesgue negligible). However, the accessibility of $\text{sign}(\beta)$ by LASSO does not mean that the probability of sign recovery by LASSO is close to 1 even if the non-null components of β are extremely large. Actually, the irrepresentability condition is necessary for sign recovery with a probability larger than 1/2 (Wainwright, 2009) and this condition implies accessibility. More precisely, the irrepresentability condition is satisfied when $\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} \text{sign}(\beta_I)\|_\infty \leq 1$ where $I := \{i \in [p] : \beta_i \neq 0\}$ and $\overline{I} := \{i \in [p] : \beta_i = 0\}$. Whereas irrepresentability has been widely studied as an analytical condition, we suggest on the following example a geometrical interpretation for this condition.

Example 1. *Let* $X \in \mathbb{R}^{2\times 3}$ *and* $\beta \in \mathbb{R}^3$ *as follows*

$$X := \begin{pmatrix} 5/6 & 1 & 0 \\ 1/3 & 0 & 1 \end{pmatrix} \text{ and } \beta = (\beta_1, 0, 0) \text{ where } \beta_1 > 0.$$

Actually, as illustrated on Figure 1, $\text{sign}(\beta) = (1, 0, 0)$ *is an accessible sign vector, as there exists* $y \in \mathbb{R}^2$ *such that that* $\text{sign}(\hat{\beta}^{\text{lasso}}(y)) = (1, 0, 0)$ *(e.g. by taking* $y = (1, 1)$*). On the other hand,* $\text{sign}(\beta)$ *does not satisfy the irrepresentability condition (*$\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} \text{sign}(\beta_I)\|_\infty = 30/29 > 1$*). Consequently, the probability of sign recovery by LASSO is not null (because* $\text{sign}(\beta)$ *is accessible) but smaller than 1/2 (because* $\text{sign}(\beta)$ *does not satisfy the irrepresentability condition). For this example we take* $\lambda = 1$*.*

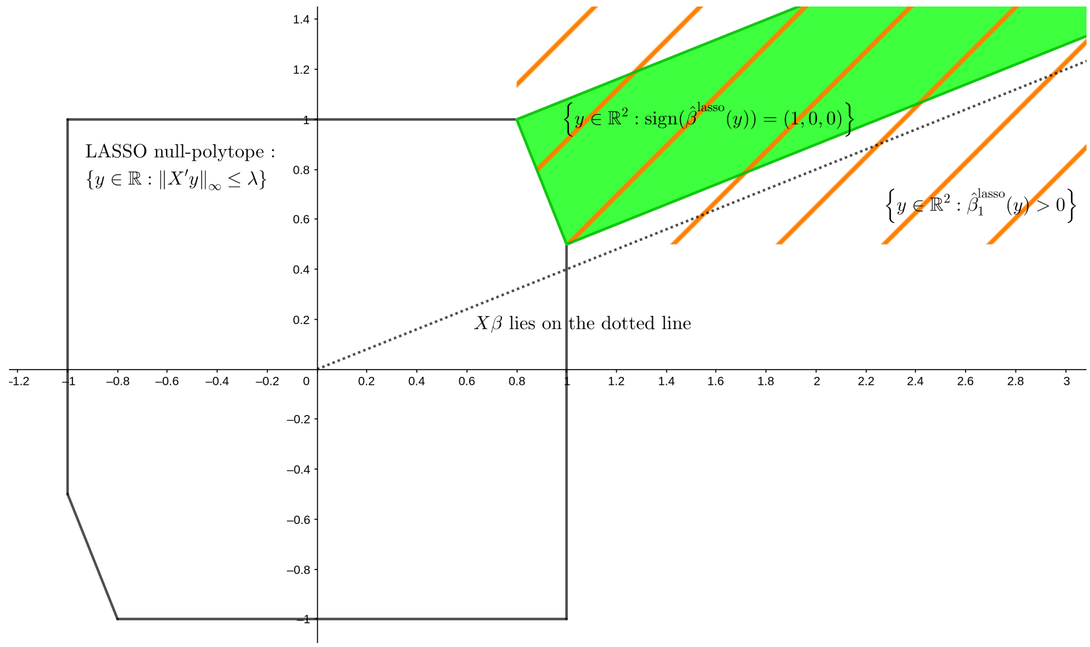

Figure 1: This figure provides sets $\{y \in \mathbb{R}^2 : \mathrm{sign}(\hat\beta^{\mathrm{lasso}}(y)) = (1,0,0)\}$ (in green) and $\{y \in \mathbb{R}^2 : \hat\beta_1^{\mathrm{lasso}}(y) > 0\}$ (in orange and hatched). Clearly the green set is not negligible with respect to the Lebesgue measure thus $\mathbb{P}(\mathrm{sign}(\hat\beta^{\mathrm{lasso}}(Y)) = (1,0,0)) > 0$ (it is the probability that Y lies on the green set). On the other hand, the probability of sign recovery cannot be large. Indeed $X\beta$ lies on the dotted half-line; in particular $X\beta \notin \{y \in \mathbb{R}^2 : \mathrm{sign}(\hat\beta^{\mathrm{lasso}}(y)) = (1,0,0)\}$ therefore $\mathbb{P}(\mathrm{sign}(\hat\beta^{\mathrm{lasso}}(Y)) = \mathrm{sign}(\beta)) \le 1/2$ since $Y = X\beta + \epsilon$ is centered in $X\beta$ and ε has a symmetric distribution. When β_1 is very large then $X\beta$, on the dotted half-line, is far from 0. Consequently, Y should lie on the hatched area with a very large probability namely the event $\hat\beta_1^{\mathrm{lasso}}(Y) > 0$ should occur with a probability close to 1. This last fact will be relevant in section 1.3.1 to illustrate that in this simple example, in contrast to LASSO, thresholded LASSO can recover $\mathrm{sign}(\beta)$ with a large probability.

1.2.2 Other results towards model pattern recovery

Generalized LASSO. By substituting the ℓ_1 norm by a polyhedral gauge $\mathrm{pen} = \|D.\|_1$, one constructs an estimator $\hat\beta \in S_{X,\lambda\|D.\|_1}(Y)$ where $D\hat\beta$ has some null components. It is a reason why the generalized LASSO is frequently used for structure recovery. Of course, the structure induced by generalized LASSO depends on the matrix D.

For instance, when D is a matrix such that $Db = (b_2 - b_1, \dots, b_p - b_{p-1})'$ (denoted D^{tv} below) then the penalty term $\|D.\|_1$ promotes neighbor components of $\hat\beta$ being equal and entailing that this estimator can recover the jump set: $\{i \in [p-1] : \beta_i \ne \beta_{i+1}\}$ (Hütter and Rigollet, 2016). Actually, articles by (Qian and Jia, 2016; Owrang et al., 2017) provide theoretical properties for jump set recovery under an irrepresentability condition.

Model subspace recovery. More generally, for a wide class of penalty terms including polyhedral gauges, Vaiter et al. (2015) showed that an irrepresentability condition is a sufficient condition for model subspace recovery by penalized least squares estimators. The notion of model subspace is related to the notion of model pattern. Specifically, the model subspace of $x \in \mathbb{R}^p$ is a vector subspace of $\mathbb{R}^p$ perpendicular to $\partial_{\mathrm{pen}}(x)$. Thus clearly two vectors having the same model pattern share the same model subspace. For the ℓ_1 norm two vectors $x, z \in \mathbb{R}^p$ have the same model subspace when $\mathrm{supp}(x) = \mathrm{supp}(z)$. In the particular case of LASSO, Theorem 6 in Vaiter et al.

(2015) shows that $\|X_{\bar{I}}'X_I(X_I'X_I)^{-1}\mathrm{sign}(\beta_I)\|_\infty < 1$ is a sufficient condition for model subspace recovery, *i.e.* the recovery of $\mathrm{supp}(\beta)$. Whereas correct, this statement is not optimal. Indeed when $\|X_{\bar{I}}'X_I(X_I'X_I)^{-1}\mathrm{sign}(\beta_I)\|_\infty < 1$ it is well known that LASSO actually can recover $\mathrm{sign}(\beta)$ (and a fortiori $\mathrm{supp}(\beta)$) (Wainwright, 2009). That is the reason why, in this article, we decided to focus on the notion of model pattern and not to retain the notion of model subspace from Vaiter et al. (2015).

The path condition as well as the irrepresentability condition can be relaxed using thresholded estimators as explained hereafter.

1.3 Model pattern recovery by a thresholded estimator

Theorems 3 and 4 generalize results known for LASSO (see the following subsection) to a wide class of penalized estimators. Specifically, we prove in this paper that "thresholded" penalized least squares estimators can recover the model pattern of β under a weaker condition than penalized least squares estimators (which are not thresholded). Now we introduce the notion of "thresholded" estimator.

Definition 2 ("Thresholded" penalized least squares estimator)**.**
Let pen *be a polyhedral gauge,* $X \in \mathbb{R}^{n\times p}$*,* $y \in \mathbb{R}^n$ *and* $\lambda > 0$*. Given* $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$*, we say that* $\hat{u}$ *is a "thresholded" estimator of* $\hat{\beta}$ *if* $\partial_{\mathrm{pen}}(\hat{\beta}) \subset \partial_{\mathrm{pen}}(\hat{u})$*.*

Definition 2 will be illustrated on many examples in section 4.

1.3.1 Sign recovery by thresholded LASSO

Hereafter, we provide a brief presentation of results known for thresholded LASSO.

Given a threshold $\tau \geq 0$, we remind that thresholded LASSO $\hat{\beta}^{\mathrm{lasso},\tau}$ is defined as follows

$$\hat{\beta}_i^{\mathrm{lasso},\tau} = \begin{cases} \hat{\beta}_i^{\mathrm{lasso}} \text{ if } |\hat{\beta}_i^{\mathrm{lasso}}| > \tau, \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Note that whatever $\tau \geq 0$ we have $\partial_{\|.\|_1}(\hat{\beta}^{\mathrm{lasso}}) \subset \partial_{\|.\|_1}(\hat{\beta}^{\mathrm{lasso},\tau})$ and thus $\hat{\beta}^{\mathrm{lasso},\tau}$ is a "thresholded" estimator of $\hat{\beta}^{\mathrm{lasso}}$ in the sense of Definition 2.

It is well known that thresholded LASSO does not have the same statistical properties as LASSO (Meinshausen and Yu, 2009; Weinstein et al., 2020). Concerning sign recovery, accessibility condition is a necessary condition for sign recovery by thresholded LASSO. Indeed, Tardivel and Bogdan (2018) recently proved that if $\mathrm{sign}(\hat{\beta}^{\mathrm{lasso},\tau}) = \mathrm{sign}(\beta)$, then $\mathrm{sign}(\beta)$ is accessible for the LASSO. Moreover, contrarily to LASSO, thresholded LASSO can recover the sign of β with a large probability under the accessibility condition (even if the irrepresentability condition is not satisfied) as soon as non-null components of β are sufficiently large. This nice property for sign recovery by thresholded LASSO remains true for thresholded basis pursuit (Saligrama and Zhao, 2011; Descloux and Sardy, 2020; Descloux et al., 2020). In Example 1 when β_1 is very large then $\hat{\beta}_1^{\mathrm{lasso}} > 0$ occurs with a large probability and thus thresholded LASSO can recover $\mathrm{sign}(\beta)$ with a probability near to 1 once $\hat{\beta}_1^{\mathrm{lasso}}$ is stochastically larger than both $|\hat{\beta}_2^{\mathrm{lasso}}|$ and $|\hat{\beta}_3^{\mathrm{lasso}}|$. In particular, Figure 2 illustrates that thresholded LASSO can recover the sign of β when $\beta = (20, 0, 0)$ and $\varepsilon \sim \mathcal{N}(0, I_2)$. This article generalizes this toy example and shows that thresholded penalized estimators can recover the model pattern of β under weaker condition than penalized estimators (which are not thresholded).

1.4 Notations

Hereafter, we give some notations that we are going to use in this article.

- Given a matrix $X \in \mathbb{R}^{n\times p}$, X' represents the transpose of the matrix X, $\ker(X)$ represents the null space of X: $\ker(X) := \{z \in \mathbb{R}^p : Xz = 0\}$ and $\mathrm{row}(X)$ represents the vector space spanned by rows of X: $\mathrm{row}(X) := \{X'z : z \in \mathbb{R}^n\}$.

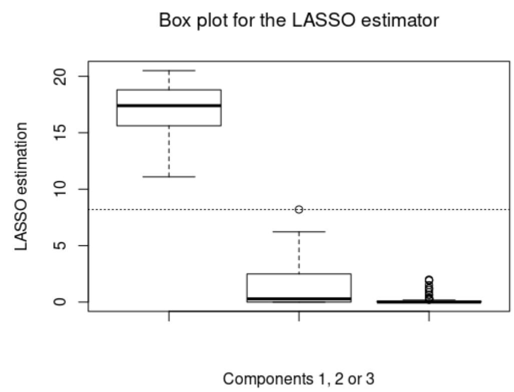

Figure 2: This figure provides box plots for components $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$ based on 1000 replicates for LASSO estimator. One may observe that when components for LASSO are filtered out with an appropriate threshold (in this example $\tau = 8$) then thresholded LASSO recovers $\mathrm{sign}(\beta)$ almost perfectly. This is in stark contrast with regular LASSO since $\mathbb{P}_\varepsilon(\mathrm{sign}(\hat{\beta}^{\mathrm{lasso}}(Y)) = (1,0,0)) = 0,17$.

- Given $p \in \mathbb{N}$, the notation $[p]$ represents the set of integers $\{1, \ldots, p\}$.
- The notation $\|.\|_0$ represents the ℓ_0 "norm"; the number of non-null components. Namely, for $x \in \mathbb{R}^p$, $\|x\|_0 := |\{i \in [p] : x_i \neq 0\}|$.
- When it is convenient, the component i^{th} of $x \in \mathbb{R}^p$ is denoted $[x]_i$.
- Given $x \in \mathbb{R}^p$ and τ the notation x^τ represents the thresholded vector

$$x^\tau := (x_1 \mathbf{1}(|x_1| > \tau), \ldots, x_p \mathbf{1}(|x_p| > \tau)).$$

- The support of $x \in \mathbb{R}^p$ is the set: $\mathrm{supp}(x) := \{i \in [p] : x_i \neq 0\}$.
- The notation $B_\infty(a, r)$ represents the ball for the ℓ_∞ norm centered in a with radius r.

1.5 Polyhedral gauges, polyhedral norms

Let $u_1, \ldots, u_l \in \mathbb{R}^p$. We remind that a polyhedral gauge pen is a function defined by

$$\forall x \in \mathbb{R}^p, \mathrm{pen}(x) = \max\{0, u_1' x, \ldots, u_l' x\}.$$

By definition a polyhedral gauge is a non-negative convex function (Rockafellar, 1997; Mousavi and Shen, 2019). A symmetric and positive polyhedral gauge (*i.e.* $\mathrm{pen}(x) = \mathrm{pen}(-x)$ and $\mathrm{pen}(x) = 0 \Rightarrow x = 0$)) is called a polyhedral norm. Hereafter we present some examples of polyhedral norms.

The ℓ_1 norm: This polyhedral norm is defined as follows

$$\forall x \in \mathbb{R}^p, \|x\|_1 = \sum_{i=1}^{p} |x_i| = \max\{u' x : u \in \{-1, 1\}^p\}.$$

The ℓ_∞ norm: Let $e_1, \ldots, e_p$ be the canonical basis in $\mathbb{R}^p$. This polyhedral norm is defined as follows

$$\forall x \in \mathbb{R}^p, \|x\|_\infty = \max\{|x_1|, \ldots, |x_p|\} = \max\{e_1' x, -e_1' x, \ldots, e_p' x, -e_p' x\}.$$

The sorted ℓ_1 norm: Let $w = (w_1, \dots, w_p) \in \mathbb{R}^p$ where $w_1 > 0$ and $w_1 \geq \dots \geq w_p \geq 0$. The sorted ℓ_1 norm $\|.\|_w$ is defined as follows:

$$\forall x \in \mathbb{R}^p, \|x\|_w = \sum_{i=1}^{p} w_i |x|_{(i)},$$

where $(.)$ is a permutation on $[p]$ such that $|x|_{(1)} \geq \dots \geq |x|_{(p)}$. Note that when weights satisfy $w = (1, \dots, 1)$ then the sorted ℓ_1 norm coincides with the ℓ_1 norm and when $w = (1, 0, \dots, 0)$ then the sorted ℓ_1 norm coincides with the ℓ_∞ norm. Let $\mathcal{S}_p$ be the set of permutations on $[p]$. The following equality, based on the rearrangement inequality, shows that the sorted ℓ_1 norm is a polyhedral norm

$$\forall x \in \mathbb{R}^p, \|x\|_w = \max \left\{ \sum_{i=1}^{p} s_i w_{\pi(i)} x_i : s_1, \dots, s_p \in \{-1, 1\} \text{ and } \pi \in \mathcal{S}_p \right\}.$$

Remark 1. *The composition of polyhedral gauge with a linear map is still a polyhedral gauge. For example, for generalized LASSO, the penalty term is the polyhedral gauge $x \in \mathbb{R}^p \mapsto \|Dx\|_1$ when $D \in \mathbb{R}^{m \times p}$. Note that, when $\{0\} \subsetneq \ker(D)$, this gauge is not a norm but just a semi-norm.*

Hereafter we present two matrices D, which are relevant for generalized LASSO (this list is not exhaustive).

- Let $p \geq 2$ and let $D^{\text{tv}} \in \mathbb{R}^{(p-1) \times p}$ be the first order difference matrix defined as follows

$$D^{\text{tv}} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

- Let $p \geq 3$ and let $D^{\text{tf}} \in \mathbb{R}^{(p-2) \times p}$ be the second order difference matrix defined as follows

$$D^{\text{tf}} = \begin{pmatrix} -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2 & -1 \end{pmatrix}.$$

  The ℓ_1 trend filtering (Kim et al., 2009) is actually a generalized LASSO with the penalty term being $\|D^{\text{tf}}.\|_1$.

1.6 Sub-gradients, subdifferentials and model patterns

We remind the reader of the definition on subgradient and subdifferential. The following can be found for instance in Hiriart-Urruty and Lemarechal (1993):
For a function $f : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a *subgradient of f at $x \in \mathbb{R}^p$* if

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

The set of all subgradients of f at x is called the *subdifferential of f at x*, denoted by $\partial_f(x)$.
In this article, we only consider continuous convex functions and thus the set of subgradients is a non-empty convex set.

Example 2. *The subdifferential of the ℓ_1 norm at $x \in \mathbb{R}^p$ is given by*

$$\partial_{\|.\|_1}(x) = \partial_{|.|}(x_1) \times \cdots \times \partial_{|.|}(x_p) \text{ where } \partial_{|.|}(t) = \begin{cases} \{1\} \text{ if } t > 0 \\ [-1,1] \text{ if } t = 0 \\ \{-1\} \text{ if } t < 0 \end{cases}$$

The subdifferential of the ℓ_∞ norm at 0 is the unit ball of the ℓ_1 norm and for $x \in \mathbb{R}^p$ where $x \neq 0$ this subdifferential is equal to

$$\partial_{\|.\|_\infty}(x) = \left\{ s \in \mathbb{R}^p : \|s\|_1 = 1 \text{ and } \begin{cases} s_i x_i \geq 0 \text{ if } |x_i| = \|x\|_\infty \\ 0 \text{ otherwise} \end{cases} \right\}.$$

Finally, note that for the polyhedral gauge $x \in \mathbb{R}^p \mapsto \|Dx\|_1$ we have $\partial_{\|D.\|_1}(x) = D'\partial_{\|.\|_1}(Dx)$.

The subdifferential of the sorted ℓ_1 norm, not reminded above, has a more complex expression than subdifferentials of both the ℓ_1 and ℓ_∞ norms and is given in Dupuis and Tardivel (2021); Schneider and Tardivel (2020).

Now, we want to illustrate that two vectors $x, z \in \mathbb{R}^p$ having the same subdifferential with respect to a polyhedral gauge share a common model pattern.

Model pattern for the ℓ_1 norm: The sign vector $\text{sign}(x) \in \{-1, 0, 1\}^p$ is defined as follows

$$\text{sign}(x) = (\text{sign}(x_1), \ldots, \text{sign}(x_p)) \text{ where } \text{sign}(t) := \begin{cases} 1 \text{ if } t > 0 \\ 0 \text{ if } t = 0 \\ -1 \text{ if } t < 0 \end{cases}$$

Subdifferentials $\partial_{\|.\|_1}(x) = \partial_{\|.\|_1}(z)$ are equal if and only if $\text{sign}(x) = \text{sign}(z)$. In words, two vectors having the same subdifferential with respect to the ℓ_1 norm share the same sets of positive components, null components and negative components.

Example: for $x = (1.45, -0.38, 1.56, 0, -2.76)$ then $\text{sign}(x) = (1, -1, 1, 0, -1)$.

Model pattern for the ℓ_∞ norm: The vector $\text{sign}^\infty(x)$ element of the finite alphabet $\{-1, *, 1\}^p$ is defined as follows

$$\text{sign}^\infty(0) = (*, \ldots, *) \text{ and for } x \neq 0,\ \forall i \in [p], [\text{sign}^\infty(x)]_i := \begin{cases} 1 \text{ if } x_i = \|x\|_\infty \\ * \text{ if } |x_i| < \|x\|_\infty \\ -1 \text{ if } x_i = -\|x\|_\infty \end{cases}$$

Note that the notation $*$ represents a components which is not maximal (except for $x = 0$). Subdifferentials $\partial_{\|.\|_\infty}(x) = \partial_{\|.\|_\infty}(z)$ are equal if and only if $\text{sign}^\infty(x) = \text{sign}^\infty(z)$. In words, two vectors having the same subdifferential with respect to the ℓ_∞ norm share the same sets of positive and maximal (in absolute value) components and negative and maximal (in absolute value) components.

Example: for $x = (1.45, 1.45, 0.56, 0, -1.45)$ then $\text{sign}^\infty(x) = (1, 1, *, *, -1)$.

Model pattern when the penalty term is the sorted ℓ_1 norm: We say that a vector $m \in \mathbb{Z}^p$ is a SLOPE model, if either $m = 0$, or, if for all $l \in [\|m\|_\infty]$, there exists $j \in [p]$ such that $|m_j| = l$. We denote the set of all SLOPE models of dimension p by $\mathcal{M}_p$ (note that, by definition, $\mathcal{M}_p \subset [-p : p]^p$). For example, when $p = 2$, $\mathcal{M}_2$ has 17 elements listed hereafter:

$$\mathcal{M}_2 = \{(0,0), \pm(1,0), \pm(0,1), \pm(1,1), \pm(1,-1), \pm(2,1), \pm(2,-1), \pm(1,2), \pm(1,-2)\}.$$

Moreover, for $x \in \mathbb{R}^p$, we define $\text{mdl}(x) \in \mathcal{M}_p$, called a model for SLOPE, through the following.

1) $\mathrm{sign}(\mathrm{mdl}(x)) = \mathrm{sign}(x)$

2) $|x_i| = |x_j| \implies |\mathrm{mdl}(x)_i| = |\mathrm{mdl}(x)_j|$

3) $|x_i| > |x_j| \implies |\mathrm{mdl}(x)_i| > |\mathrm{mdl}(x)_j|$

Let $w \in \mathbb{R}^p$ where $w_1 > \cdots > w_p > 0$. Then, subdifferentials $\partial_{\|.\|_w}(x) = \partial_{\|.\|_w}(z)$ are equal if and only if $\mathrm{mdl}(x) = \mathrm{mdl}(z)$. In words, two vectors having the same subdifferential with respect to the sorted ℓ_1 norm share 1) the same sign, 2) the same clusters (components equal in absolute value) and 3) clusters for these two vectors have the same order.

Example: for $x = (3.1, -1.2, 0.5, 0, 1.2, -3.1)$ then $\mathrm{mdl}(x) = (3, -2, 1, 0, 2, -3)$.

Model pattern for the polyhedral gauge $\|D^{\mathrm{tv}}.\|_1$: Let $p \geq 2$. The vector $\mathrm{jump}(x)$ element of the finite alphabet $\{\nearrow, \rightarrow, \searrow\}^{p-1}$ is defined as follows

$$\forall i \in [p-1], \mathrm{jump}(x)_i := \begin{cases} \nearrow & \text{if } x_{i+1} > x_i \\ \rightarrow & \text{if } x_{i+1} = x_i \\ \searrow & \text{if } x_{i+1} < x_i \end{cases}$$

Subdifferentials $\partial_{\|D^{\mathrm{tv}}.\|_1}(x) = \partial_{\|D^{\mathrm{tv}}.\|_1}(z)$ are equal if and only if $\mathrm{jump}(x) = \mathrm{jump}(z)$. Namely, two vectors x, z having the same subdifferential with respect to the $\|D^{\mathrm{tv}}.\|_1$ share the same sets of positive jumps $\{i \in [p-1] : x_i < x_{i+1}\} = \{i \in [p-1] : z_i < z_{i+1}\}$ and negative jumps $\{i \in [p-1] : x_i > x_{i+1}\} = \{i \in [p-1] : z_i > z_{i+1}\}$.

Example: for $x = (1.45, 1.45, 0.56, 0.56, -0.45, 0.35)$ then $\mathrm{jump}(x) = (\rightarrow, \searrow, \rightarrow, \searrow, \nearrow)$.

Model pattern for the polyhedral gauge $\|D^{\mathrm{tf}}.\|_1$: Let $p \geq 3$. The vector $\mathrm{knot}(x)$ element of the finite alphabet $\{l, cx, cv\}^{p-2}$ is defined as follows

$$\forall i \in [2 : p-1], \mathrm{knot}(x)_i := \begin{cases} cx \text{ if } x_i < (x_{i+1} - x_{i-1})/2 \\ l \text{ if } x_i = (x_{i+1} - x_{i-1})/2 \\ cv \text{ if } x_i > (x_{i+1} - x_{i-1})/2 \end{cases}$$

Let us consider the piecewise linear curve $C_x := \cup_{i=1}^{p-1}[(i, x_i), (i+1, x_{i+1})]$. Note that $[\mathrm{knot}(x)]_i$ is equal to l (resp. cx or cv) when in the neighborhood of i the curve C_x is linear (resp. convex or concave). Subdifferentials $\partial_{\|D^{\mathrm{tf}}.\|_1}(x) = \partial_{\|D^{\mathrm{tf}}.\|_1}(z)$ are equal if and only if $\mathrm{knot}(x) = \mathrm{knot}(z)$. Namely, two vectors x, z having the same subdifferential with respect to the $\|D^{\mathrm{tf}}.\|_1$ share the same sets of convex points: $\{i \in [2 : p-1] : \mathrm{knot}(x)_i = cx\}$ and concave points: $\{i \in [2 : p-1] : \mathrm{knot}(x)_i = cv\}$.

Example: Figure 3 provides an illustration of $\mathrm{knot}(x)$ for a particular x.

2 Necessary and sufficient condition for uniform uniqueness

In Theorem 1 we are going to provide a necessary and sufficient condition for uniform uniqueness of optimization problem (1). Actually, this theorem generalizes to polyhedral gauges Theorem 1 in Schneider and Tardivel (2020) which only covers the particular case of "pen" being a polyhedral norm. In particular, Theorem 1 relaxes the coercivity of norm and thus provides a necessary and sufficient uniqueness for uniform uniqueness of generalized LASSO.

Theorem 1 (Necessary and sufficient condition for uniform uniqueness)**.** *Let $X \in \mathbb{R}^{n\times p}$ and $\lambda > 0$. Let* pen *be a polyhedral gauge defined as follows: $\forall x \in \mathbb{R}^p, \mathrm{pen}(x) = \max\{u_1'x, \ldots, u_l'x\}$. Then there*

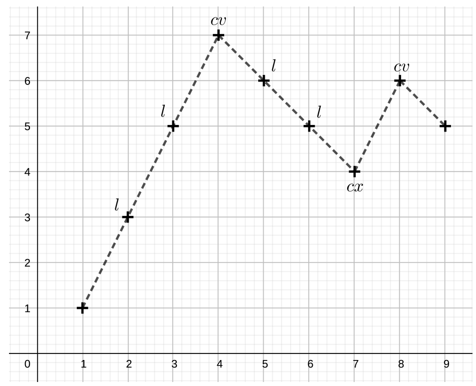

Figure 3: In this figure the dotted curve represents C_x described above for $x = (1, 3, 5, 7, 6, 5, 4, 6, 5)$. Moreover, $\text{knot}(x) = (l, l, cv, l, l, cx, cv)$.

exists $y \in \mathbb{R}^n$ for which the minimizer of the function

$$b \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - Xb\|_2^2 + \lambda \text{pen}(b), \tag{3}$$

is not unique (i.e $S_{X,\lambda\text{pen}}(y)$ is not a singleton) if and only if $\text{row}(X)$ *intersects a face of the polytope* $B^* := \text{conv}\{u_1, \ldots, u_l\}$ *whose dimension*[1] *is smaller than* $\dim(\ker(X))$.

Note that, according to Lemma 2 in the appendix, B^* is the subdifferential of pen at 0 and in the particular case where pen is a norm, B^* is the unit ball of the dual norm.

Several examples where the uniform uniqueness does not occur are given in Schneider and Tardivel (2020) in the particular case where the penalty term is a polyhedral norm. Hereafter, we provide an example where the uniform uniqueness does not occur for generalized LASSO. Clearly, for every $y \in \mathbb{R}^n$, the set of generalized LASSO minimizers $S_{X,\|D.\|_1}(y)$ is unbounded once $\ker(X) \cap \ker(D) \neq \{0\}$ and thus the uniform uniqueness does not occur. Consequently, $\ker(X) \cap \ker(D) = \{0\}$ is a necessary condition for uniform uniqueness but not a sufficient condition as illustrated on the following example.

Example 3. *Let us consider the optimization problem*

$$\arg\min_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \|Db\|_1 \text{ where } X = \begin{pmatrix} 1 & -1 & 3/4 \\ 0 & -1 & 1/4 \end{pmatrix} \quad , \quad D = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

Note that $\ker(X) \cap \ker(D) = \{0\}$ *and one may notice the following equality*

$$\|Db\|_1 = \max\{u'Db : u \in \{-1, 1\}^2\} = \max\{\pm(b_1 - b_3), \pm(b_1 - 2b_2 + b_3)\}.$$

According to the above equality, $B^* = \text{conv}\{\pm(1, 0, -1), \pm(1, -2, 1)\}$ *and clearly* $(1, -2, 1)' \in \text{row}(X)$. *Consequently, according to Theorem 1, the uniform uniqueness does not occur. Now, for* $y = (2, 1)'$, *let us illustrate that* $|S_{X,\|D.\|_1}(y)| > 1$. *Let* $\hat{\beta} = (1, 0, 0)'$ *then* $\hat{\beta} \in S_{X,\|D.\|_1}(y)$. *Indeed, the following equality holds*

$$X'(y - X\hat{\beta}) = (1, -2, 1)' = D' \begin{pmatrix} 1 \\ -1 \end{pmatrix} \in \partial_{\|D.\|_1}(\hat{\beta}) = D'\partial_{\|.\|_1}(D\hat{\beta}) = D' \begin{pmatrix} 1 \\ [-1, 1] \end{pmatrix}.$$

Moreover, let $\bar{\beta} = (1/3, 1/3, 4/3)$ *then* $X\bar{\beta} = X\hat{\beta}$ *and* $\|D\bar{\beta}\|_1 = \|D\hat{\beta}\|_1$ *implying thus* $\bar{\beta} \in S_{X,\|D.\|_1}(y)$. *Consequently* $|S_{X,\|D.\|_1}(y)| > 1$.

[1] The dimension of a face F is defined as the dimension of the affine hull of F.

In this example, extremal points of the set of minimizers are $\hat{\beta}$ and $\bar{\beta}$. More generally, according to Dupuis and Vaiter (2019), extremal points of $S_{X,\|D.\|_1}(y)$ can be explicitly computed. This description is relevant when the set of minimizers is not a singleton.

Remark 2. *When $\hat{\beta} \in S_{X,\text{pen}}(y)$ then $X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta})$. Therefore, clearly,* $\text{row}(X)$ *intersects the set $\partial_{\text{pen}}(\hat{\beta})$ which is actually a face of the polytope B^*. Indeed, $\partial_{\text{pen}}(\hat{\beta}) = \text{conv}(\{u_i\}_{i\in I})$ where $I = \{i \in [p] : u_i'\hat{\beta} = \text{pen}(\hat{\beta})\}$. Consequently, when the uniform uniqueness occurs, by Theorem 1 we have* $\dim(\partial_{\text{pen}}(\hat{\beta})) \geq \dim(\ker(X))$ *or equivalently* $\text{codim}(\partial_{\text{pen}}(\hat{\beta})) \leq \text{rk}(X)$.

Hereafter we provide some simple formulas for the dimension/codimension of the set $\partial_{\text{pen}}(x)$ for an arbitrary $x \in \mathbb{R}^p$. These facts will be relevant for Corollary 1.

The ℓ_1 norm: The dimension (resp. codimension) of the set $\partial_{\|.\|_1}(x)$ is equal to $|\{i \in [p] : x_i = 0\}|$ (resp. $\|x\|_0$).

The ℓ_∞ norm: When $x \neq 0$, the dimension (resp. codimension) of the set $\partial_{\|.\|_\infty}(x)$ is equal to $|\{i : [\text{sign}^\infty(x)]_i \in \{-1,1\}\}| - 1$ (resp. $1 + |\{i : [\text{sign}^\infty(x)]_i = *\}|$).

The sorted ℓ_1 norm: When $w = (w_1, \dots, w_p)$ where $w_1 > \dots > w_p > 0$, the dimension (resp. codimension) of the set $\partial_{\|.\|_w}(x)$ is equal to $p - \|\text{mdl}(x)\|_\infty$ ($\|\text{mdl}(x)\|_\infty$), where $\|\text{mdl}(x)\|_\infty$ is the number of non-null clusters of x (Schneider and Tardivel, 2020).

The polyhedral gauge $\|D.\|_1$: When $\ker(D') = \{0\}$ then $\dim(\partial_{\|D.\|_1}(x)) = \dim(D'\partial_{\|.\|_1}(Dx)) = \dim(\partial_{\|.\|_1}(Dx)) = |\{i \in [m] : [Dx]_i = 0\}|$. Thus the codimension of $\partial_{\|D.\|_1}(x)$ is equal to $p - |\{i \in [m] : [Dx]_i = 0\}| = p - (m - \|Dx\|_0) = p - m + \|Dx\|_0$.

- The codimension of the set $\partial_{\|D^{\text{tv}}.\|_1}(x)$ is equal to $1 + |\{i \in [p-1] : [\text{jump}(x)]_i \in \{\nearrow, \searrow\}|$ namely the codimension of $\partial_{\|D^{\text{tv}}.\|_1}(x)$ is the number of jumps plus 1.
- The codimension of the set $\partial_{\|D^{\text{tf}}.\|_1}(x)$ is equal to $2 + |\{i \in [p-2] : [\text{knot}(x)]_i \in \{cv, cx\}|$ namely the codimension of $\partial_{\|D^{\text{tf}}.\|_1}(x)$ is the number of knots of the vector x plus 2.

Under the uniform uniqueness, when $\hat{\beta} \in S_{X,\text{pen}}(y)$ then $\text{codim}(\partial_{\text{pen}}(\hat{\beta})) \leq \text{rk}(X)$. Depending on the penalty term and as illustrated above, $\text{codim}(\partial_{\text{pen}}(\hat{\beta}))$ is related respectively to the number of: non-null components, non-null clusters, non maximal components, jumps or knots. Especially when $\text{codim}(\partial_{\text{pen}}(\hat{\beta})) \leq \text{rk}(X) \leq n \ll p$ these numbers above are small compared to p which means that $\hat{\beta}$ is "sparse". Corollary 1 provides a precise meaning for this notion of "sparsity" for $\|.\|_\infty$ and for generalized Lasso.

Corollary 1. *Let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and* pen *be a polyhedral gauge on $\mathbb{R}^p$ and let us assume that the uniform uniqueness holds, i.e. for every $y \in \mathbb{R}^n$ the set $S_{X,\lambda\text{pen}}(y) = \{\hat{\beta}\}$ is a singleton.*

i) If $\text{pen} = \|.\|_\infty$, *for every $y \in \mathbb{R}^n$ we have*

$$|\{i \in [p] : |\hat{\beta}_i| < \|\hat{\beta}\|_\infty\}| \leq \max\{\text{rk}(X) - 1, 0\}.$$

ii) If $\text{pen} = \|D.\|_1$ *where $D \in \mathbb{R}^{m\times p}$ and* $\ker(D') = \{0\}$, *then for every $y \in \mathbb{R}^n$ we have*

$$\|D\hat{\beta}\|_0 \leq \text{rk}(X) + m - p.$$

Actually, $\text{rk}(X) + m - p < 0$ implies that $\ker(X) \cap \ker(D) \neq \{0\}$ and thus the uniform uniqueness assumption is not valid in Corollary 1. Let us provide some examples of Corollary 1.

- For the regular LASSO, when $D = Id_p$, according to Corollary 1 one recovers the well known fact that LASSO minimizer $\hat{\beta}$ satisfies $\|\hat{\beta}\|_0 \leq \text{rk}(X)$ under uniqueness (Osborne et al., 2000).

- Let $D = D^{\text{tv}}$ be the first difference matrix. When the uniform uniqueness occurs then the number of jumps of the unique generalized LASSO minimizer $\hat{\beta}$ is smaller than $\text{rk}(X) - 1$ *i.e.*

$$|\{i \in [p-1] : [\text{jump}(\hat{\beta})]_i \in \{\nearrow, \searrow\}\}| \leq \text{rk}(X) - 1.$$

- Let $D = D^{\text{tf}}$ be the second difference matrix. When the uniform uniqueness occur then the number of knots of the unique generalized LASSO minimizer $\hat{\beta}$ is smaller than $\text{rk}(X) - 2$, *i.e.*

$$|\{i \in [p-1] : [\text{knot}(\hat{\beta})]_i \in \{cx, cv\}\}| \leq \text{rk}(X) - 2.$$

Penalized gauges used in practice are symmetric (*i.e* $\text{pen}(x) = \text{pen}(-x)$). Under this assumption $\ker(\text{pen}) := \{x \in \mathbb{R}^p : \text{pen}(x) = 0\}$ is a vector subspace of $\mathbb{R}^p$. Proposition 1 shows that when $\dim \ker(\text{pen}) > n$ then for every $y \in \mathbb{R}^n$ the set $S_{X,\text{pen}}(y)$ is not bounded. Conversely, according to Proposition 1 when $\dim \ker(\text{pen}) \leq n$ then the set of matrices $X \in \mathbb{R}^{n \times p}$ for which the uniform uniqueness does not occur is negligible with respect to the Lebesgue measure.

Proposition 1. *Let* pen *be a symmetric polyhedral gauge on* $\mathbb{R}^p$ *and* $\lambda > 0$.

i) If $\dim(\ker(\text{pen})) > n$ *then for every* $X \in \mathbb{R}^{n \times p}$ *we have* $\{0\} \subsetneq \ker(\text{pen}) \cap \ker(X)$ *and thus for every* $y \in \mathbb{R}^n$ *the set* $S_{X,\lambda\text{pen}}(y)$ *is unbounded.*

ii) Let μ *be the Lebesgue measure on* $\mathbb{R}^{n \times p}$*. If* $\dim(\ker(\text{pen})) \leq n$ *then the following equality holds*

$$\mu\left(\left\{X \in \mathbb{R}^{n \times p} : \exists y \in \mathbb{R}^n \text{ with } |S_{X,\lambda\text{pen}}(y)| > 1\right\}\right) = 0.$$

3 Model pattern recovery

3.1 Necessary condition for model pattern recovery: accessibility

In the following definition we introduce the notion of accessible model. This definition generalizes to a broad class of penalized estimators the notions of accessible sign vector (Sepehri and Harris, 2017; Schneider and Tardivel, 2020) and accessible model for SLOPE (Schneider and Tardivel, 2020).

Definition 3 (Accessible model pattern by penalized estimators)**.** *Let* $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$ *and* pen *be a polyhedral gauge. We say that* $\beta \in \mathbb{R}^p$ *has an accessible model pattern with respect to* X *and* λpen*, if there exists* $y \in \mathbb{R}^n$ *and* $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ *such that* $\partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\beta)$.

Of course, by definition, when $\gamma \in \mathbb{R}^p$ is such that $\partial_{\text{pen}}(\gamma) = \partial_{\text{pen}}(\beta)$ then the model of γ is accessible with respect to X and pen if and only if the model β is accessible with respect to X and pen. When pen is the ℓ_1 norm scaled by a tuning parameter $\lambda > 0$, namely $\text{pen} := \lambda\|.\|_1$ then, the above definition coincides with the notion of accessibility of sign vectors with respect to X. When pen is the sorted ℓ_1 norm, namely $\text{pen} := \|.\|_w$ for some $w \in \mathbb{R}^p$ where $w_1 > \cdots > w_p > 0$ then, the above definition coincides with the notion of accessible SLOPE models with respect to X. Proposition 2 provides a geometric and an analytic characterization of accessible models.

Proposition 2 (Characterization of accessible models)**.** *Let* $X \in \mathbb{R}^{n \times p}$ *and* $\text{pen} : \mathbb{R}^p \to \mathbb{R}$ *be a polyhedral gauge.*

1) Geometric characterization: The model pattern of $\beta \in \mathbb{R}^p$ *is accessible with respect to* X *and* λpen *if and only if*

$$\text{row}(X) \cap \partial_{\text{pen}}(\beta) \neq \varnothing.$$

2) Analytic characterization: The model pattern of $\beta \in \mathbb{R}^p$ *is accessible with respect to* X *and* λpen *if and only if for every* $\gamma \in \mathbb{R}^p$ *the implication*

$$X\beta = X\gamma \implies \text{pen}(\gamma) \geq \text{pen}(\beta)$$

holds.

Based on Proposition 2, it is clear that the notion of accessible model does not depend on the tuning parameter $\lambda > 0$.

3.2 Necessary condition for model pattern recovery: path condition

The solution path for a penalized estimator is the function $0 < \lambda \mapsto \hat{\beta}(\lambda)$ where $\hat{\beta}(\lambda) \in S_{X,\lambda\text{pen}}(y)$ for fixed $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ (usually $S_{X,\lambda\text{pen}}(y)$ is assumed being a singleton). The solution path for generalized LASSO or OSCAR and Clustered LASSO is studied in, respectively, Tibshirani et al. (2011); Takahashi and Nomura (2020). Definition 4 is based on the notion of solution path. Note that Definition 4 does not require uniform uniqueness.

Definition 4 (Path condition)**.** *Let* pen *be a polyhedral gauge, $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. We say that the model of β satisfies the path condition with respect to X and* pen *when*

$$\exists \lambda > 0, \exists \hat{\beta} \in S_{X,\lambda\text{pen}}(X\beta) \text{ such that } \partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\beta).$$

For instance, $\beta = 0$ satisfies the path condition with respect to X and pen. Indeed, in that case $X\beta = 0$ and clearly $0 \in S_{X,\lambda\text{pen}}(0)$. In other words, the path condition means that in the noiseless case when $Y = X\beta$ in the solution path, one may pick a minimizer having the same model pattern as β.

The path condition is illustrated on Figure 4 in the particular case where X is the 2×3 matrix given in Example 1 and $\beta = (10, 0, 0)'$. The above definition does not provide analytic expression for checking

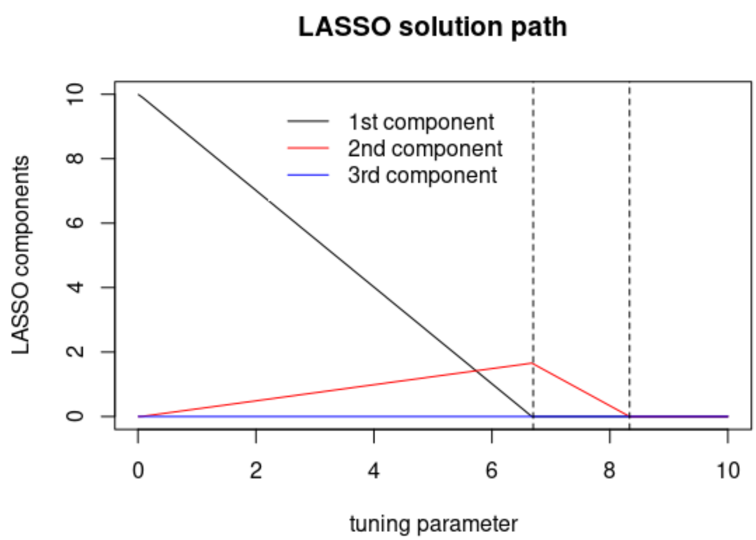

Figure 4: This figure provides curves of the functions $\lambda > 0 \mapsto \hat{\beta}_1^{\text{lasso}}(\lambda)$ (black curve), $\lambda > 0 \mapsto \hat{\beta}_2^{\text{lasso}}(\lambda)$ (red curve), $\lambda > 0 \mapsto \hat{\beta}_3^{\text{lasso}}(\lambda)$ (blue curve). Note that sign(β) does not satisfy the path condition. Indeed $\text{sign}(\hat{\beta}^{\text{lasso}}(\lambda)) = (1, 1, 0)$ when $\lambda \in (0, 6.70)$, $\text{sign}(\hat{\beta}^{\text{lasso}}(\lambda)) = (0, 1, 0)$ when $\lambda \in [6.70, 8.33)$ and $\text{sign}(\hat{\beta}^{\text{lasso}}(\lambda)) = (0, 0, 0)$ when $\lambda \geq 8.33$. Consequently, for every $\lambda > 0$, $\text{sign}(\hat{\beta}^{\text{lasso}}(\lambda)) \neq (1, 0, 0)$.

the path condition but some formulas are given in the literature. For example, when pen $= \|.\|_1$ Bühlmann and Van de Geer (2011) clearly illustrate that when $\|X_{\overline{I}}'X_I(X_I'X_I)^{-1}\text{sign}(\beta_I)\|_\infty > 1$ (where $I = \text{supp}(\beta)$, X_I and $X_{\overline{I}}$ are matrices whose columns are respectively $(X_i)_{i\in I}$ and $(X_i)_{i\notin I}$) the path condition does not hold whereas when $\|X_{\overline{I}}'X_I(X_I'X_I)^{-1}\text{sign}(\beta_I)\|_\infty < 1$ then the path condition holds. Thus, the well known "irrepresentability condition" for LASSO can be thought of as an analytical shortcut for checking the path condition. Actually, Figure 4 confirms this fact. Indeed in the above example, we have $\|X_{\overline{I}}'X_I(X_I'X_I)^{-1}\text{sign}(\beta_I)\|_\infty = 30/29 > 1$ and based on Figure 4 one may observe that the path condition does not hold for β. This article does not aim at providing analytical shortcuts for checking the path condition (we refer to Vaiter et al. (2015) for interested reader). Actually we want to show that

a) The path condition is necessary for model recovery by penalized estimator (see Theorem 2).

b) "Thresholded" penalized estimator recovers the model of β under much weaker condition than the path condition (see section 4).

Theorem 2. *Let $Y = X\beta + \varepsilon$ where $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and ε has a symmetric distribution (i.e the distribution of ε coincides with the distribution of $-\varepsilon$). Let* pen *be a polyhedral gauge. If β does not satisfy the path condition with respect to X and* pen *then the following inequality holds*

$$\mathbb{P}(\exists\lambda_0 > 0\ \exists\hat{\beta} \in S_{X,\lambda_0\mathrm{pen}}(Y) \text{ such that } \partial_{\mathrm{pen}}(\hat{\beta}) = \partial_{\mathrm{pen}}(\beta)) \le 1/2.$$

According to Theorem 2, when the path condition does not hold for the LASSO (for example, when $\|X_{\overline{I}}'X_I(X_I'X_I)^{-1}\mathrm{sign}(\beta_I)\|_\infty > 1$) then the following inequality is true

$$\mathbb{P}(\exists\lambda > 0\ \exists\hat{\beta} \in S_{X,\lambda\|.\|_1}(Y) \text{ such that } \mathrm{sign}(\hat{\beta}) = \mathrm{sign}(\beta)) \le 1/2.$$

This result is more accurate than the one given in Theorem 2 in Wainwright (2009) which, for a given $\lambda > 0$, provides the inequality $\mathbb{P}(\mathrm{sign}(\hat{\beta}^{\mathrm{lasso}}(\lambda)) = \mathrm{sign}(\beta)) \le 1/2$.

Clearly, when β satisfies the path condition with respect to X and pen then β is accessible with respect to X and pen (by taking $y = X\beta$ in the definition of accessibility). In the following section, we are going to prove that "thresholded" penalized least squares estimators recover the model of β under the accessibility condition.

4 Necessary and sufficient condition for model recovery by "thresholded" penalized estimators

Generally speaking, a practitioner performing model pattern recovery has some beliefs on the model pattern of the unknown parameter β. For instance, if the practitioner thinks that β is sparse and uses LASSO as an estimator promoting sparsity then it is quite natural for him to threshold small component of $\hat{\beta}^{\mathrm{lasso}}$ and thus to consider the thresholded LASSO estimator $\hat{\beta}^{\mathrm{lasso},\tau}$ for some threshold $\tau \ge 0$. Moreover, theoretically, when the threshold is appropriately selected, thresholded LASSO allows to recover $\mathrm{sign}(\beta)$ under weaker condition as LASSO (Tardivel and Bogdan, 2018). We aim at generalizing this nice property of thresholded LASSO to a broad class of penalized estimators. Before introducing the notion of “thresholded” estimator, let us stress that for any threshold $\tau \ge 0$, the inclusion $\partial_{\|.\|_1}(\hat{\beta}^{\mathrm{lasso}}) \subset \partial_{\|.\|_1}(\hat{\beta}^{\mathrm{lasso},\tau})$ occurs. This last inclusion is the keystone concept to introduce the notion of “thresholded” estimator as illustrated hereafter:

a) The estimator $\hat{\beta} \in S_{X,\lambda\|.\|_\infty}(Y)$ promotes a flat estimation (some components are maximal in absolute value) then, once $|\hat{\beta}_i| < \|\hat{\beta}\|_\infty$ and $|\hat{\beta}_i| \approx \|\hat{\beta}\|_\infty$, it is quite natural to neglect this approximation and to consider that “$|\hat{\beta}_i| = \|\hat{\beta}\|_\infty$”. Let us call $\hat{u}$ the estimator taking into account this approximation and obtained after modifying slightly $\hat{\beta}$ then, $\partial_{\|.\|_\infty}(\hat{\beta}) \subset \partial_{\|.\|_\infty}(\hat{u})$.

b) SLOPE estimator promotes clusters (a cluster is a set of components equal in absolute value) then, once $|\hat{\beta}_j^{\mathrm{slope}}| \approx |\hat{\beta}_i^{\mathrm{slope}}|$, it is quite natural to neglect this approximation and to consider that “$|\hat{\beta}_i^{\mathrm{slope}}| = |\hat{\beta}_j^{\mathrm{slope}}|$”. Let us call $\hat{u}$ the estimator taking into account this approximation and obtained after modifying slightly $\hat{\beta}^{\mathrm{slope}}$ then, $\partial_{\|.\|_w}(\hat{\beta}^{\mathrm{slope}}) \subset \partial_{\|.\|_w}(\hat{u})$.

c) The estimator $\hat{\beta} \in S_{X,\lambda\|D^{\mathrm{tv}}.\|_1}(Y)$ promotes neighbor components being equal then, once $\hat{\beta}_i \approx \hat{\beta}_{i+1}$, it is quite natural to neglect this approximation and to consider that “$\hat{\beta}_i = \hat{\beta}_{i+1}$”. Let us call $\hat{u}$ the estimator taking into account this approximation and obtained after modifying slightly $\hat{\beta}$ then, $\partial_{\|D^{\mathrm{tv}}.\|_1}(\hat{\beta}) \subset \partial_{\|D^{\mathrm{tv}}.\|_1}(\hat{u})$.

The list of examples described above leads to define the notion of "thresholded" penalized least squares estimator given in the Introduction in Definition 2.

In Theorem 3 we prove that the accesibility of the model pattern of β with respect to X and pen is a necessary and sufficient condition for thresholded estimators to recover the model pattern $\partial_{\text{pen}}(\beta)$. A geometric statement shows that thresholded estimators can recover the model pattern of β as soon as: 1) the model pattern of β is accessible with respect to X and pen

2) components of β are large enough.

Theorem 3. *Let* pen *be a polyhedral gauge,* $X \in \mathbb{R}^{n\times p}$ *and* $\beta \in \mathbb{R}^p$.

Necessary condition for model recovery: *If the model pattern of β is not accessible with respect to X and* pen *then for any $y \in \mathbb{R}^n$, $\lambda > 0$ and $\hat\beta \in S_{X,\lambda\text{pen}}(y)$ we have $\partial_{\text{pen}}(\hat\beta) \not\subset \partial_{\text{pen}}(\beta)$. Consequently, no "thresholded" estimator can recover the model pattern of β.*

Sufficient condition for model recovery (asymptotic): *Suppose that the uniform uniqueness holds. Let $\varepsilon \in \mathbb{R}^n$ and set $y^k = X(k\beta)+\varepsilon$. If the model pattern of β is accessible with respect to X and* pen *then*

$$\exists k_0 \in \mathbb{N} \text{ such that } \forall k \geq k_0, \exists \tau \geq 0 \text{ where } \begin{cases} \forall u \in B_\infty(\hat\beta(y^k),\tau), \partial_{\text{pen}}(u) \subset \partial_{\text{pen}}(\beta), \\ \exists \hat u \in B_\infty(\hat\beta(y^k),\tau), \partial_{\text{pen}}(\hat u) = \partial_{\text{pen}}(\beta) \supset \partial_{\text{pen}}(\hat\beta(y^k)). \end{cases}$$

Consequently, there exists a "thresholded" penalized least squares estimator $\hat u \in B_\infty(\hat\beta(y^k),\tau)$ of $\hat\beta(y^k)$ which recovers the model pattern of β. The dimension $\dim(\partial_{\text{pen}}(\hat u))$ *is maximal among elements of $B_\infty(\hat\beta(y^k),\tau)$.*

Observe that the sufficient condition given in Theorem 3 remains true when the ball with respect to the supremum norm is substituted by a ball with respect to another norm. However, we believe that for the supremum norm the explicit computation of the "thresholded" estimator $\hat u$ is simpler as illustrated hereafter. For instance, for LASSO, $\hat u$ can be taken as equal to the thresholded LASSO estimator (2). Consequently, Theorem 3 corroborates Theorem 1 in Tardivel and Bogdan (2018), which proves that thresholded LASSO recovers the sign pattern of β once the accessibility condition holds. Similarly as thresholded LASSO, when pen $= \|.\|_\infty$, an estimator $\hat u$ from Theorem 3 can be explicitly computed as illustrated on Algorithm 1.

Algorithm 1 thresholded penalized least squares estimator when the penalty term is the ℓ_∞ norm:

Require: estimation: $\hat\beta$, threshold $\tau \geq 0$.
  if $\|\hat\beta\|_\infty \leq \tau$ **then**
    $\hat u \leftarrow 0$.
  else

$$\forall i \in [p], \hat u_i \leftarrow \begin{cases} \|\hat\beta\|_\infty - \tau \text{ if } \|\hat\beta\|_\infty - 2\tau \leq \hat\beta_i \leq \|\hat\beta\|_\infty, \\ -\|\hat\beta\|_\infty + \tau \text{ if } -\|\hat\beta\|_\infty \leq \hat\beta_i \leq -\|\hat\beta\|_\infty + 2\tau, \\ \hat\beta_i \text{ otherwise.} \end{cases}$$

  end if
  return $\text{sign}^\infty(\hat u)$.

4.1 Necessary and sufficient condition for model recovery by "thresholded" Generalized LASSO

The generalized LASSO estimator $\hat\beta \in S_{X,\lambda\|D.\|_1}(Y)$ promotes null components in $D\hat\beta$.

Thus, when $[D\hat{\beta}]_i \approx 0$, it is quite natural to neglect this approximation and to consider that "$[D\hat{\beta}]_i = 0$". Moreover, one may observe that for any threshold $\tau \geq 0$ we have

$$\partial_{\|D.\|_1}(\hat{\beta}) = D'\partial_{\|.\|_1}(D\hat{\beta}) \subset D'\partial_{\|.\|_1}([D\hat{\beta}]^\tau).$$

This last inclusion suggests to recover the model pattern of β by thresholding components of $D\hat{\beta}$. One may observe that, given $\mathrm{sign}(D\beta)$, one easily recovers $\partial_{\|D.\|_1}(\beta)$. Therefore, hereafter, we focus on the sign pattern recovery of $D\beta$. Specifically, Theorem 4 shows that the equality $\mathrm{sign}(D\beta) = \mathrm{sign}([D\hat{\beta}]^\tau)$ occurs for an appropriate threshold $\tau \geq 0$ if and only if the model pattern of β is accessible with respect to X and $\|D.\|_1$.

Theorem 4 (Necessary and sufficient condition for model pattern recovery by generalized LASSO)**.** *Let $D \in \mathbb{R}^{m\times p}$, $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$.*

Necessary condition for model pattern recovery: *If the model pattern of β is not accessible with respect to X and $\|D.\|_1$ then for every $y \in \mathbb{R}^n$, $\lambda > 0$ and $\hat{\beta} \in S_{X,\lambda\|D.\|_1}(y)$ the following statements hold:*

i) $\partial_{\|D.\|_1}(\hat{\beta}) \not\subset \partial_{\|D.\|_1}(\beta)$.

ii) Whatever $\tau \geq 0$ we have $\mathrm{sign}([D\hat{\beta}]^\tau) \neq \mathrm{sign}(D\beta)$.

Sufficient condition for model pattern recovery (asymptotic): *Let us assume that for any $y \in \mathbb{R}^n$ the set $S_{X,\lambda\|D.\|_1}(y)$ contains a unique minimizer $\hat{\beta}(y)$. Let $\varepsilon \in \mathbb{R}^n$ and set $y^k = X(k\beta) + \varepsilon$. If the model pattern of β is accessible with respect to X and $\|D.\|_1$ then*

iii)

$$\exists k_0 \in \mathbb{N}, \forall k \geq k_0, \exists \tau \geq 0 \text{ such that } \mathrm{sign}([D\hat{\beta}(y^k)]^\tau) = \mathrm{sign}(D\beta).$$

iv) Let π be a permutation of $\mathcal{S}_p$ such that $|[D\hat{\beta}(y^k)]_{\pi(1)}| \geq |[D\hat{\beta}(y^k)]_{\pi(2)}| \geq \cdots \geq |[D\hat{\beta}(y^k)]_{\pi(p)}|$. Then the following family of nested subsets contains $\mathrm{supp}(D\beta)$*:*

$$\exists k_0 \in \mathbb{N} \text{ such that } \forall k \geq k_0, \text{ we have } \mathrm{supp}(D\beta) \in \{\varnothing, \{\pi(1)\}, \{\pi(1), \pi(2)\}, \ldots, \mathrm{supp}(D\hat{\beta}(y^k))\}.$$

Note that the construction of the "thresholded" $\hat{u}$ given in Theorem 3 is substituted by thresholding $D\hat{\beta}$ in Theorem 4 and these two approaches are both useful to recover the model pattern of β. Whereas Theorem 3, dealing with polyhedral gauges, is more general than Theorem 4, we believe that this last result is more intuitive for generalized LASSO.

Given $\hat{\beta} \in S_{X,\lambda\|D.\|_1}(y)$, point iii) suggests to recover the sign of $D\beta$ by thresholding components of $D\hat{\beta}$ whereas iv) suggests to recover the support $D\beta$ by constructing a nested family of subsets for $\mathrm{supp}(D\hat{\beta})$.[2] For LASSO, a suggestion for selecting the threshold is given in Tardivel and Bogdan (2018) and for basis pursuit (basis pursuit can be seen as LASSO with an infinitely small tuning parameter λ) a methodology for choosing the threshold is given in Descloux and Sardy (2020). Otherwise, in the particular case where $D = I_p$, model selection criterion can be useful to recover $\mathrm{supp}(D\beta)$ (Pokarowski et al., 2019).

In this article we provide neither a threshold τ nor a model selection criterion for model recovery. Instead, our aim is to prove that the fundamental property of the existence of a "thresholded" estimator recovering the model pattern of β is equivalent to its accessibility (see Theorems 3 and 4). Thus, "thresholded" estimators are more relevant for model pattern recovery than estimators which are not thresholded. This is illustrated on numerical experiments in section XXXX.

[2] Point iv) in Theorem 4 could be rewritten as a "nested" family of model patterns containing the model pattern of β when $k \geq k_0$. However, the fact that $\mathrm{supp}(D\beta)$ is included in a nested family of subsets seems more intuitive.

Acknowledgments

We would like to thank Samuel Vaiter for his insightful comments on the paper. The work of Patrick J.C. Tardivel has been supported by the EIPHI Graduate School (contract ANR-17-EURE-0002).

5 Appendix

Let $S \in \mathbb{R}^p$ be an arbitrary set. Hereafter, we are going to use the following notations:

- The set $\mathrm{conv}(S)$ denotes for the convex hull of S; it is the intersection of all convex sets in $\mathbb{R}^p$ containing S.
- The set $\mathrm{aff}(S)$ denotes for the affine space spanned by S; it is the intersection of all affine spaces in $\mathbb{R}^p$ containing S.
- The set $\mathrm{vect}(S)$ denotes for the vector space spanned by S; it is the intersection of all vector spaces in $\mathbb{R}^p$ containing S. Note that $\mathrm{vect}(S) = \mathrm{aff}(S \cup \{0\})$.

When E is an affine space and x is an arbitrary element in E, then $\overrightarrow{E}$ is denoted as the vector space $\overrightarrow{E} := \{z - x : z \in E\}$ where Note that $\overrightarrow{E}$ does not depend on $x \in E$. The dimension of E equals $\dim(\overrightarrow{E})$. Finally, when $E = \mathrm{aff}(S)$, we just write $\overrightarrow{\mathrm{aff}}(S)$ for this vector space.

5.1 Facts about polytopes

We report some basic definitions and facts on polytopes, which we will use throughout the article and, in particular, in the proofs in subsequent sections. The following can, for instance, be found in the excellent textbooks by Gruber (2007) and Ziegler (2012).

A set $P \subseteq \mathbb{R}^p$ is called a polytope, if it is the convex hull of a finite set of points in $\mathbb{R}^p$, namely,

$$P = \mathrm{conv}(\{v_1, \ldots, v_k\}), \text{ where } v_1, \ldots, v_k \in \mathbb{R}^p.$$

The *dimension* $\dim(P)$ of a polytope is given as the dimension of $\mathrm{aff}(P)$, the affine subspace spanned by P. A *face* F of P is any subset $F \subseteq P$ that satisfies

$$F = \{x \in P : a'x = b\}, \text{ where } P \subseteq \{x \in \mathbb{R}^p : a'x \leq b\},$$

for some $a \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Such an inequality $a'x \leq b$ is called a *valid inequality* of P. Note that $F = \varnothing$ and $F = P$ are faces of P and that any face F is again a polytope. A non-empty face where $F \neq P$ is called *proper*. A point $x_0 \in P$ lies in $\mathrm{relint}(P)$, the *relative interior* of P, if x_0 is not contained in a proper face of P. We now list a number of useful facts about polytopes involving the above definitions, which are used throughout the article. These properties can either be found explicitly or as a straightforward consequence of properties listed in the above mentioned references.

Lemma 1. *Let $P \in \mathbb{R}^p$ be a polytope given by $P = \mathrm{conv}(\{v_1, \ldots, v_k\})$, where $v_1, \ldots, v_k \in \mathbb{R}^{p \times k}$. The following properties hold.*

1) *If F and $\tilde{F}$ are faces of P, then so is $F \cap \tilde{F}$.*
2) *Let L be an affine line contained in the affine span of P. If $L \cap \mathrm{relint}(P) \neq \varnothing$ then L intersects a proper face of P.*

5.2 Facts about polyhedral gauges

Lemma 2, useful to prove Theorem 1, reminds some well known facts on subdifferential in the particular case where the convex function is the maximum of a finite family of linear functions (see *e.g.* Hiriart-Urruty and Lemarechal (1993) for some related results).

Lemma 2. *Let ϕ be a convex function defined as follows*

$$\forall x \in \mathbb{R}^p, \phi(x) := \max\{u_1'x, \dots, u_l'x\}, \text{ for some } u_1, \dots, u_l \in \mathbb{R}^p.$$

Let us remind that $B^* := \mathrm{conv}(\{u_1, \dots, u_l\})$.

i) Let $I(x) = \{i \in [l] : u_i'x = \phi(x)\}$ for $x \in \mathbb{R}^p$. Then

$$\partial_\phi(x) = \mathrm{conv}(\{u_i\}_{i \in I(x)}) = \{s \in B^* : s'x = \mathrm{pen}(x)\}.$$

ii) If F is a face of B^* then $F = \partial_\phi(x)$ for some $x \in \mathbb{R}^p$

Proof. i) The following equality is well known (see *e.g.* Hiriart-Urruty and Lemarechal (1993))

$$\partial_\phi(x) = \mathrm{conv}(\{u_i\}_{i \in I(x)}).$$

If $i \in I(x)$, then, by definition of $I(x)$, $u_i'x = \phi(x)$, therefore $u_i \in \{s \in B^* : s'x = \phi(x)\}$. Finally, because $\{s \in B^* : s'x = \phi(x)\}$ is a convex set, one may deduce that

$$\mathrm{conv}(\{u_i\}_{i \in I(x)}) \subset \{s \in B^* : s'x = \phi(x)\}.$$

Conversely, let $s \in B^*$ such that $s \notin \mathrm{conv}(\{u_i\}_{i \in I(x)})$. Therefore $s = \sum_{i=1}^l \alpha_i u_i$ where $u_1 \geq 0, \dots, u_l \geq 0$, $\sum_{i=1}^l \alpha_i = 1$ and $\alpha_{i_0} > 0$ for some $i_0 \notin I(x)$ (since $s \notin \mathrm{conv}(\{u_i\}_{i \in I(x)})$). Since $\forall i \in [l], u_i'x \leq \phi(x)$ and $u_{i_0}'x < \phi(x)$, the following equality occurs:

$$s'x = \sum_{i=1}^l \alpha_i u_i'x < \phi(x).$$

Consequently $s \notin \{s \in B^* : s'x = \phi(x)\}$ and thus the following inclusion occurs

$$\{s \in B^* : s'x = \phi(x)\} \subset \mathrm{conv}(\{u_i\}_{i \in I(x)}).$$

ii) Let $F := \{s \in B^* : a's = b\}$ be a face of B^*, where $a's \leq b$ is a valid inequality for all $s \in B^*$. Let us prove that $\partial_\phi(a) = F$. According to i) we obtain $\partial_\phi(a) = \{s \in B^* : s'a = \phi(a)\}$. Moreover, by definition, for every $i \in [l]$ we have $u_i'a \leq \phi(a)$ and thus the inequality $a's \leq \phi(a)$ is valid for all $s \in B^*$. Finally, because two supporting hyperplanes having the same normal vector coincide we may deduce that $\phi(a) = b$ and thus $\partial_\phi(a) = F$. □

Before finally showing Theorem 1, the following lemma states that the fitted values are unique over all solutions of the penalized problem for a given y. It is a generalization of Lemma 1 in Tibshirani (2013), who proves this fact for the special case of the LASSO.

Lemma 3. *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and* pen *be a polyhedral gauge. Then $X\hat{\beta} = X\tilde{\beta}$ and $\mathrm{pen}(\hat{\beta}) = \mathrm{pen}(\tilde{\beta})$ for all $\hat{\beta}, \tilde{\beta} \in S_{X,\mathrm{pen}}(y)$.*

Proof. Assume that $X\hat{\beta} \neq X\tilde{\beta}$ for some $\hat{\beta}, \tilde{\beta} \in S_{X,\mathrm{pen}}(y)$ and let $\check{\beta} = (\hat{\beta} + \tilde{\beta})/2$. Because the function $\mu \in \mathbb{R}^n \mapsto \|y - \mu\|_2^2$ is strictly convex, one may deduce that

$$\|y - X\check{\beta}\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \frac{1}{2}\|y - X\tilde{\beta}\|_2^2.$$

Consequently,

$$\frac{1}{2}\|y - X\check{\beta}\|_2^2 + \|\check{\beta}\| < \frac{1}{2}\left(\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \mathrm{pen}(\hat{\beta}) + \frac{1}{2}\|y - X\tilde{\beta}\|_2^2 + \mathrm{pen}(\tilde{\beta})\right),$$

which contradicts both $\hat{\beta}$ and $\tilde{\beta}$ being minimizers. Finally, $X\hat{\beta} = X\tilde{\beta}$ clearly implies $\mathrm{pen}(\hat{\beta}) = \mathrm{pen}(\tilde{\beta})$. □

5.3 Proof of Theorem 1

In this proof the notation $h^{\perp}$ denotes for the hyperplane $\{x \in \mathbb{R}^p : h'x = 0\}$.

Proof. ($\Longleftarrow$) Assume that there exists a face $F = \partial_{\text{pen}}(\hat{\beta})$ of $B^* = \text{conv}\{u_1, \dots, u_l\}$ that intersects $\text{row}(X)$ and satisfies $\dim(F) < \dim(\ker(X))$. By Lemma 2, every face of B^* can be written as $\partial_{\text{pen}}(\hat{\beta})$ for a particular point $\hat{\beta}$. Then we may pick $z \in \mathbb{R}^n$ with $X'z \in F$, which exists by assumption, and fix $y = X\hat{\beta} + \lambda z$. Then $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ due to the following equality

$$\frac{1}{\lambda}X'(y - X\hat{\beta}) = X'z \in \partial_{\text{pen}}(\hat{\beta}).$$

Note that, according to Lemma 2, $\partial_{\text{pen}}(\hat{\beta}) = \text{conv}\{u_i\}_{i\in I}$ where $I := \{i \in [l] : u_i'\hat{\beta} = \text{pen}(\hat{\beta})\}$ (and thus $u_i'\hat{\beta} < \text{pen}(\hat{\beta})$ when $i \notin I$). We now construct $\tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ different from $\hat{\beta}$. To construct $\tilde{\beta}$ it suffices to pick $h \in \ker(X)$ where $h \neq 0$ such that $u_i'h = 0$ for all $i \in I$.
First case:
Let $0 \in \text{aff}\{u_i\}_{i\in I}$ (thus $\text{aff}\{u_i\}_{i\in I} = \text{vect}\{u_i\}_{i\in I}$). Then $\dim(F) = \dim(\text{aff}\{u_i\}_{i\in I}) = \dim(\text{vect}\{u_i\}_{i\in I}) < \dim(\ker(X))$ implying thus $\{0\} \subsetneq \ker(X) \cap (\text{vect}\{u_i\}_{i\in I})^{\perp}$ and consequently one may pick $h \in \ker(X)$ where $h \neq 0$ for which $u_i'h = 0$ for all $i \in I$.
Second case:
Let $0 \notin \text{aff}\{u_i\}_{i\in I}$ and let $v = X'z$. Then, by construction, $v \in \text{row}(X) \cap \text{conv}\{u_i\}_{i\in I}$ and note that $v \neq 0$. Consequently, we have $\ker(X) = \text{row}(X)^{\perp} \subset v^{\perp}$ and $(\text{vect}\{u_i\}_{i\in I})^{\perp} \subset v^{\perp}$. Because the following inequality holds

$$\dim(\text{vect}\{u_i\}_{i\in I}) = \dim(\text{aff}\{u_i\}_{i\in I}) + 1 \leq \dim(\ker(X)),$$

then $\dim(\ker(X)) + \dim((\text{vect}\{u_i\}_{i\in I})^{\perp}) \geq p$. Consequently, if $\ker(X) \cap (\text{vect}\{u_i\}_{i\in I})^{\perp} = \{0\}$ then $\ker(X) + (\text{vect}\{u_i\}_{i\in I})^{\perp} = \mathbb{R}^p$ which is not possible since both $\ker(X) \subset v^{\perp}$ and $(\text{vect}\{u_i\}_{i\in I})^{\perp} \subset v^{\perp}$. Consequently $\{0\} \subsetneq \ker(X) \cap (\text{vect}\{u_i\}_{i\in I})^{\perp}$ and thus one may pick $h \in \ker(X)$ where $h \neq 0$ such that $u_i'h = 0$ for all $i \in I$.
Up to scale h, one may assume that $h \neq 0$ is small enough so that $u_i'(\hat{\beta} + h) < \text{pen}(\hat{\beta})$ once $i \notin I$. Therefore, the following equality occurs

$$\text{pen}(\hat{\beta} + h) = \max\{u_1'(\hat{\beta} + h), \dots, u_l'(\hat{\beta} + h)\} = \max\{u_i'(\hat{\beta} + h)\}_{i\in I} = \text{pen}(\hat{\beta}).$$

Consequently, by taking $\tilde{\beta} = \hat{\beta} + h$ we clearly obtain that $\tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ with $\tilde{\beta} \neq \hat{\beta}$.

($\Longrightarrow$) Let us assume that there exists $y \in \mathbb{R}^n$ and $\hat{\beta}, \tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ with $\beta \neq \tilde{\beta}$. We then have

$$\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta}) \quad \text{and} \quad \frac{1}{\lambda}X'(y - X\tilde{\beta}) \in \partial_{\text{pen}}(\tilde{\beta}).$$

According to Lemma 3, $X\hat{\beta} = X\tilde{\beta}$, thus $\frac{1}{\lambda}X'(y - X\hat{\beta}) = \frac{1}{\lambda}X'(y - X\tilde{\beta})$. Consequently, one may deduce that $\text{row}(X)$ intersects the face $\partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$. Let $F^* = \text{conv}\{u_i\}_{i\in I^*}$ be a face of $\partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$ of smallest dimension among faces intersecting $\text{row}(X)$. By the minimality of $dim(F^*)$, $\text{row}(X)$ intersects the relative interior of F^*, namely there exists $z \in \mathbb{R}^n$ such that $X'z$ lies on F^* but not on a proper face of F^*. Now, we are going to show that if $\dim(F^*) = dim(\text{aff}\{u_i\}_{i\in I^*}) \geq \dim(\ker(X))$ then $\text{row}(X)$ intersects a proper face of F^* leading to a contradiction.

Indeed, let $h = \hat{\beta} - \tilde{\beta} \neq 0$. Clearly, $h \in \ker(X)$. Moreover, because $\text{pen}(\hat{\beta}) = \text{pen}(\tilde{\beta})$ (according to Lemma 3) and because $u_i \in \partial_{\text{pen}}(\hat{\beta}) \cap \partial_{\text{pen}}(\tilde{\beta})$ for every $i \in I^*$, then, according to Lemma 2, the following equality holds

$$\forall i \in I^*, \quad u_i'h = u_i'\hat{\beta} - u_i'\tilde{\beta} = \text{pen}(\hat{\beta}) - \text{pen}(\tilde{\beta}) = 0.$$

Therefore $h \in (\text{vect}\{u_i\}_{i\in I^*})^\perp \cap \ker(X)$. Let us assume that $\dim(\text{aff}\{u_i\}_{i\in I^*}) \geq \dim(\ker(X))$. Then the following inequality holds

$$p = \dim(\text{row}(X)) + \dim(\ker(X)) \leq \dim(\text{row}(X)) + \dim(\text{aff}\{u_i\}_{i\in I^*}).$$

If $\text{row}(X) \cap \overrightarrow{\text{aff}}\{u_i\}_{i\in I^*} = \{0\}$ then $\text{row}(X) + \overrightarrow{\text{aff}}\{u_i\}_{i\in I^*} = \mathbb{R}^p$. However, the last equality cannot hold since $\text{row}(X) = \ker(X)^\perp \subset h^\perp$ and $\overrightarrow{\text{aff}}\{u_i\}_{i\in I^*} \subset \text{vect}\{u_i\}_{i\in I^*} \subset h^\perp$. Consequently, there exists $0 \neq v \in \text{row}(X) \cap \overrightarrow{\text{aff}}\{u_i\}_{i\in I^*}$.
The affine line $L := \{X'z + tv : t \in \mathbb{R}\} \subset \text{row}(X)$ intersects the relative interior of F^* (at $t = 0$) and lies in $\text{aff}(F^*)$, so L intersects the border of F^*. Finally, one may deduce that $\text{row}(X)$ intersects a proper face of F^* which gives a contradiction. □

5.4 Proof of Corollary 1

Proof. Because $\hat{\beta}$ is a minimizer, then $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta})$. Thus, clearly $\text{row}(X)$ intersects the face of $\partial_{\text{pen}}(\hat{\beta})$. Consequently, according to Theorem 1 we have $\dim(\partial_{\text{pen}}(\hat{\beta})) \geq \dim(\ker(X))$.

i) $\text{pen} = \|.\|_\infty$. If $\text{rk}(X) = 0$, then X is the null matrix. Thus $S_{X,\lambda\text{pen}}(y) = \{0\}$ and consequently $|\{i \in [p] : |\hat{\beta}_i| < \|\hat{\beta}\|_\infty\}| = 0$. Now suppose that $\text{rk}(X) \geq 1$. Then

$$\begin{aligned}
& \dim(\partial_{\|.\|_\infty}(\hat{\beta})) \geq \dim(\ker(X)), \\
\Leftrightarrow \quad & \underbrace{p - \dim(\partial_{\|.\|_\infty}(\hat{\beta}))}_{=\text{codim}(\partial_{\|.\|_\infty}(\hat{\beta}))} \leq \underbrace{p - \dim(\ker(X))}_{=\text{rk}(X)}, \\
\Leftrightarrow \quad & |\{i \in [p] : |\hat{\beta}_i| < \|\hat{\beta}\|_\infty\}| \leq \text{rk}(X) - 1.
\end{aligned}$$

ii) $\text{pen} = \|D.\|_1$. Then the following equivalence holds

$$\begin{aligned}
& \dim(\partial_{\|D.\|_1}(\hat{\beta})) \geq \dim(\ker(X)), \\
\Leftrightarrow \quad & \dim(D'\partial_{\|.\|_1}(D\hat{\beta})) \geq \dim(\ker(X)), \\
\Leftrightarrow \quad & m - \|D\hat{\beta}\|_0 \geq p - \text{rk}(X), \\
\Leftrightarrow \quad & \|D\hat{\beta}\|_0 \leq \text{rk}(X) + m - p.
\end{aligned}$$

□

Proof of Proposition 1

Lemma 4. *Let* pen *be a symmetric polyhedral gauge defined by*

$$\forall x \in \mathbb{R}^p, \text{pen}(x) = \max\{u_1'x, \dots, u_l'x\}.$$

Let F be a face of $B^* = \text{conv}(\{u_1, \dots, u_l\})$. If $0 \in F$ then $F = B^*$

Proof. According to Lemma 2, there exists $x \in \mathbb{R}^p$ such that $F = \partial_{\text{pen}}(x)$. If $0 \in F$ then pen reaches its minimum at x and thus $0 = \text{pen}(0) \geq \text{pen}(x) \geq 0$. Consequently, $x \in \ker(\text{pen})$ therefore whatever $i \in [l]$ we have $u_i'x \leq 0$ and by symmetry (since $\text{pen}(x) = \text{pen}(-x)$) we also have $-u_i'x \leq 0$ implying thus $x \in \text{vect}(\{u_i\}_{1\leq i\leq l})^\perp$. By definition of pen and since $u_i'x = 0$ the following equality occurs

$$\forall z \in \mathbb{R}^p, \text{pen}(z) \geq u_i'z = \text{pen}(x) + u_i'(z - x)$$

Consequently, whatever $i \in [l]$ we have, $u_i \in \partial_{\text{pen}}(x)$. Since $\partial_{\text{pen}}(x)$ is a convex set one may deduce that $B^* \subset F$. Finally, according to Lemma 2, $F \subset B^*$ and thus $F = B^*$. □

Proof of Proposition 1. i) Let $X \in \mathbb{R}^{n\times p}$. Then $\dim(\ker(X)) \geq p-n$, therefore when $\dim(\ker(\text{pen})) > n$ we have $\{0\} \subsetneq \ker(\text{pen}) \cap \ker(X)$. Consequently, for any $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ we have $\hat{\beta} + (\ker(\text{pen}) \cap \ker(X)) \subset S_{X,\lambda\text{pen}}(y)$. Therefore, $S_{X,\lambda\text{pen}}(y)$ is unbounded.

ii) Let $u_1, \ldots, u_l \in \mathbb{R}^p$ and let us set

$$\forall x \in \mathbb{R}^p, \text{pen}(x) = \max\{u_1'x, \ldots, u_l'x\}.$$

Since pen is symmetric then the following equivalences occur

$$h \in \ker(\text{pen}) \Leftrightarrow \forall i \in [l], u_i'h = 0 \Leftrightarrow \text{vect}(\{u_i\}_{1\leq i\leq l})^{\perp}.$$

Because for every $x \in \mathbb{R}^p, \text{pen}(x) \geq \text{pen}(0) = 0$ then $0 \in B^* = \text{conv}(\{u_i\}_{1\leq i\leq l})$. Thus $\dim(B^*) = \dim(\text{aff}\{u_i\}_{i\in[l]}) = \dim(\text{vect}\{u_i\}_{i\in[l]}) = p - \dim(\ker(\text{pen}))$. The uniform uniqueness does not hold if and only if $\text{row}(X)$ intersects a face F of B^* such that $\dim(F) < \dim(\ker(X))$. When $\dim(\ker(\text{pen})) \leq n$ then $\dim(B^*) \geq p-n \geq \dim(\ker(X))$. Consequently, a face F intersected by $\text{row}(X)$ where $\dim(F) < \dim(\ker(X))$ is a proper face of B^* and thus, according to Lemma 4, $0 \notin F$ (thus $\text{row}(X)$ which contains 0 does not trivially intersects F). Finally, the conclusion of this proof is exactly similar as the proof of Proposition 1 in Schneider and Tardivel (2020). □

Proof of Proposition 2

The following lemma generalizes Proposition 4.1 from Gilbert (2017) that is stated for the ℓ_1-norm to an arbitrary norm. This lemma is used in the proof of Proposition 2.

Lemma 5. *Let $s \in \mathbb{R}^p$ and ϕ be a convex function on $\mathbb{R}^p$. The vector space* $\text{row}(X)$ *intersects $\partial_\phi(s)$ if and only if the following holds.*

$$\forall b \in \mathbb{R}^p \quad Xb = Xs \implies \phi(b) \geq \phi(s) \tag{4}$$

Proof. Consider the function $f_s : \mathbb{R}^p \to \{0, \infty\}$ given by

$$f_s(b) = \begin{cases} 0 & Xb = Xs \\ \infty & \text{else.} \end{cases}$$

Then (4) holds for b if and only if s is a minimizer of the function $b \mapsto \phi(b) + f_s(b)$. Since we have $\partial_{f_s}(b) = \text{row}(X)$ whenever $Xb = Xs$, we can deduce that the implication (4) occurs if and only if

$$0 \in \text{row}(X) + \partial_\phi(s) \iff \text{row}(X) \cap \partial_\phi(s) \neq \varnothing.$$

□

Proof of Proposition 2. ($\implies$) When the model pattern $\partial_{\text{pen}}(\beta)$ of β is accessible with respect to X and λpen then, there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ such that $\partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\beta)$. Because $\hat{\beta}$ is a minimizer then $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta})$. Consequently, by setting $z = (y - X\hat{\beta})/\lambda$, one may deduce that $X'z \in \partial_{\text{pen}}(\beta) \cap \text{row}(X)$ (geometric characterization). Or, equivalently, by Lemma 5, whenever $X\gamma = X\beta$, we have $\text{pen}(\gamma) \geq \text{pen}(\beta)$ (analytic characterization).

($\impliedby$) If $\text{row}(X)$ intersects the face $\partial_{\text{pen}}(\beta)$ (geometric characterization) or, equivalently by the Lemma 5, if $X\gamma = X\beta$ implies $\text{pen}(\gamma) \geq \text{pen}(\beta)$ (analytic characterization), then there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\text{pen}}(\beta)$.Set $y = X\beta + \lambda z$. Then $\beta \in S_{X,\lambda\text{pen}}(y)$ implies that $\partial_{\text{pen}}(\beta)$ is accessible with respect to X and λpen. □

Proof of Theorem 2

Lemma 6. *Let $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^n$. The following set is empty or is convex*

$$V_\beta := \{y \in \mathbb{R}^n : \exists\lambda > 0 \quad \exists\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y) \text{ such that } \partial_{\mathrm{pen}}(\hat{\beta}) = \partial_{\mathrm{pen}}(\beta)\}.$$

Note that once a vector $y \in \mathbb{R}^n$ lies in the set V_β, in the solution path (see (4)) there is a tuning parameter $\lambda > 0$ and a minimizer $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$ such that $\partial_{\mathrm{pen}}(\hat{\beta}) = \partial_{\mathrm{pen}}(\beta)$.

Proof. Let us assume that $V_\beta \neq \varnothing$. Let $y, \tilde{y} \in V_\beta$. Then there exist $\lambda > 0$ and $\tilde{\lambda} > 0$ such that $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$ and $\tilde{\beta} \in S_{X,\tilde{\lambda}\mathrm{pen}}(\tilde{y})$ where $\partial_{\mathrm{pen}}(\hat{\beta}) = \partial_{\mathrm{pen}}(\tilde{\beta}) = \partial_{\mathrm{pen}}(\beta)$. Consequently, the following expression occurs

$$X'(y - X\hat{\beta}) \in \lambda\partial_{\mathrm{pen}}(\beta) \text{ and } X'(\tilde{y} - X\tilde{\beta}) \in \tilde{\lambda}\partial_{\mathrm{pen}}(\beta).$$

Let $\check{y} = \alpha y + (1-\alpha)\tilde{y}$, $\check{\lambda} = \alpha\lambda + (1-\alpha)\tilde{\lambda}$ and $\check{\beta} = \alpha\hat{\beta} + (1-\alpha)\tilde{\beta}$ and let us prove that $\check{y} \in V_\beta$. Because $\partial_{\mathrm{pen}}(\beta)$ is a convex set then $\partial_{\mathrm{pen}}(\check{\beta}) = \partial_{\mathrm{pen}}(\beta)$ and one may deduce the following equality

$$X'\left(\check{y} - X(\check{\beta})\right) = \alpha X'(y - X\hat{\beta}) + (1-\alpha)X'(\tilde{y} - X\tilde{\beta}) \in (\lambda\alpha + \tilde{\lambda}(1-\alpha))\partial_{\mathrm{pen}}(\beta) = \check{\lambda}\partial_{\mathrm{pen}}(\check{\beta}).$$

Consequently $\check{\beta} \in S_{X,\check{\lambda}\mathrm{pen}}(\check{y})$ and since $\partial_{\mathrm{pen}}(\check{\beta}) = \partial_{\mathrm{pen}}(\beta)$, one may deduce that $(\alpha y + (1-\alpha)\tilde{y}) \in V_\beta$ which implies that V_β is a convex set. ☐

Proof of Theorem 2. Clearly, by definition of the set V_β, we have

$$X\beta \notin V_\beta = \{y \in \mathbb{R}^n : \exists\lambda > 0\ \exists\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y) \text{ such that } \partial_{\mathrm{pen}}(\hat{\beta}) = \partial_{\mathrm{pen}}(\beta)\}.$$

Consequently, because V_β is convex (or empty), one may deduce that the following event always holds $\{X\beta + \varepsilon \notin V_\beta\} \cup \{X\beta - \varepsilon \notin V_\beta\}$ for any observation of $\varepsilon \in \mathbb{R}^p$. When $V_\beta = \varnothing$, this equality is clear. Otherwise, when V_β is a non empty convex set, if $\{X\beta + \varepsilon \in V_\beta\} \cap \{X\beta - \varepsilon \in V_\beta\}$ then, by convexity, one may deduce that $X\beta \in V_\beta$ leading to a contradiction. Consequently,

$$\begin{aligned} 1 &= \mathbb{P}(\{X\beta + \varepsilon \notin V_\beta\} \cup \{X\beta - \varepsilon \notin V_\beta\}) \\ &\leq \mathbb{P}(\{X\beta + \varepsilon \notin V_\beta\}) + \mathbb{P}(\{X\beta - \varepsilon \notin V_\beta\}) = 2\mathbb{P}(\{X\beta + \varepsilon \notin V_\beta\}). \end{aligned}$$

Consequently

$$1/2 \geq \mathbb{P}(\{X\beta + \varepsilon \in V_\beta\}) = \mathbb{P}(\exists\lambda > 0\ \exists\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(Y) \text{ such that } \partial_{\mathrm{pen}}(\hat{\beta}) = \partial_{\mathrm{pen}}(\beta)).$$

☐

5.5 Proof of Theorems 3 and 4

5.5.1 Necessary conditions of Theorems 3 and 4

Proof: Necessary condition of Theorem 3. Let us assume that $\partial_{\mathrm{pen}}(\hat{\beta}) \subset \partial_{\mathrm{pen}}(\beta)$. Because $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$ then $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\mathrm{pen}}(\hat{\beta}) \subset \partial_{\mathrm{pen}}(\beta)$ and consequently, $\mathrm{row}(X)$ intersects $\partial_{\mathrm{pen}}(\beta)$. Therefore, according to Proposition 2, $\partial_{\mathrm{pen}}(\beta)$ is accessible with respect to X and pen leading to a contradiction. ☐

Lemma 7 is useful to prove that in Theorem 4, i) implies ii).

Lemma 7. *Let $x \in \mathbb{R}^p$ and $\tau \geq 0$ then $\partial_{\|.\|_1}(x) \subset \partial_{\|.\|_1}(x^\tau)$.*

Proof. When $p = 1$, clearly $\partial_{|.|}(x) = \partial_{|.|}(x^\tau)$ once $xx^\tau > 0$ and $\partial_{|.|}(x) \subset \partial_{|.|}(x^\tau)$ once $xx^\tau = 0$. Now, when $p \geq 1$, since the subdifferential of the ℓ_1 norm is a Cartesian product of subdifferentials for the absolute value, the following equality holds

$$\partial_{\|.\|_1}(x) = \partial_{|.|}(x_1) \times \cdots \times \partial_{|.|}(x_p) \subset \partial_{|.|}(x_1^\tau) \times \cdots \times \partial_{|.|}(x_p^\tau) = \partial_{\|.\|_1}(x^\tau).$$

☐

Proof: Necessary condition of Theorem 4. i) is straightforward from the proof of the necessary condition given in Theorem 3 by taking pen $= \|D.\|_1$.

Let us prove that $i) \Rightarrow ii)$. Indeed, if $\mathrm{sign}([D\hat{\beta}]^\tau) = \mathrm{sign}(D\beta)$ for some $\tau \geq 0$ then according to Lemma 7 the following inclusion occurs

$$\partial_{\|D.\|_1}(\hat{\beta}) = D'\partial_{\|.\|_1}(D\hat{\beta}) \subset D'\partial_{\|.\|_1}([D\hat{\beta}]^\tau) = D'\partial_{\|.\|_1}(D\beta) = \partial_{\|D.\|_1}(\beta).$$

Consequently i) does not hold which achieves the proof. $\square$

Sufficient conditions of Theorems 3 and 4

Lemmas 8 and 9 are useful to prove that both in Theorems 3 and 4, asymptotically, $\hat{\beta}(y^k)/k$ converges to β when k tends to $+\infty$.

First, before to provide these lemmas, we remind that given a closed convex set $C \subset \mathbb{R}^p$ and $x \in C$ the asymptotic cone is the following set (Hiriart-Urruty and Lemarechal, 1993):

$$C_\infty := \{d \in \mathbb{R}^p : x + td \in C \ \forall t > 0\}.$$

Moreover, the following statements hold

- The set C_∞ does not depend on $x \in C$.
- Given C and K two closed convex sets where $C \cap K \neq \varnothing$ then $(C \cap K)_\infty = C_\infty \cap K_\infty$.
- A closed convex set C is compact if and only if $C_\infty = \{0\}$

Lemma 8. *Let* pen *be a polyhedral gauge on $\mathbb{R}^p$, $X \in \mathbb{R}^{n\times p}$, $b \in \mathrm{col}(X)$, $K_1 \geq 0$ and $K_2 \geq 0$. If $\ker(X) \cap \ker(\mathrm{pen}) = \{0\}$ then the set $C := \{u \in \mathbb{R}^p : \mathrm{pen}(u) \leq K_1 \text{ and } \|Xu - b\|_2 \leq K_2\}$ is compact.*

Proof. Clearly C is a closed and convex set. To prove that C is compact, it is enough to show that the asymptotic cone of C is $\{0\}$. Let $C^{\mathrm{pen}} := \{u \in \mathbb{R}^p : \mathrm{pen}(u) \leq K_1\}$ and $C^X := \{u \in \mathbb{R}^p : \|Xu - b\|_2 \leq K_2\}$ then, the asymptotic cones of C^{pen} and C^X are respectively $C^{\mathrm{pen}}_\infty = \ker(\mathrm{pen})$ and $C^X_\infty = \ker(X)$. Therefore $C_\infty = (C^{\mathrm{pen}} \cap C^X)_\infty = C^{\mathrm{pen}}_\infty \cap C^X_\infty = \ker(\mathrm{pen}) \cap \ker(X)$. Consequently, if $\ker(\mathrm{pen}) \cap \ker(X) = \{0\}$ then $C_\infty = \{0\}$ and thus C is compact. $\square$

Lemma 9. *Let $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$,* pen *be a polyhedral gauge on $\mathbb{R}^p$ and let us assume that for every $y \in \mathbb{R}^n$ the set $S_{X,\lambda\mathrm{pen}}(y)$ has a unique minimizer $\hat{\beta}(y)$. Let $\beta \in \mathbb{R}^p$ and let $\varepsilon \in \mathbb{R}^n$ and set $y^k = X(k\beta) + \varepsilon$. If β is accessible with respect to X and* pen *then*

$$\lim_{k\to+\infty} \hat{\beta}(y^k)/k = \beta.$$

Proof. Because $\hat{\beta}(y^k)$ is a minimizer of $S_{X,\lambda\mathrm{pen}}(y^k)$ then the following inequality occurs

$$\frac{1}{2}\|y^k - X\hat{\beta}(y^k)\|_2^2 + \lambda\mathrm{pen}(\hat{\beta}(y^k)) \leq \frac{1}{2}\|y^k - X(k\beta)\|_2^2 + \lambda\mathrm{pen}(k\beta).$$

Since $y^k - X(k\beta) = \varepsilon$ one may deduce the following inequalities

$$\begin{aligned} & \lambda\mathrm{pen}(\hat{\beta}(y^k)) \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\mathrm{pen}(k\beta), \\ \Rightarrow \quad & \mathrm{pen}(\hat{\beta}(y^k)/k) \leq \frac{\|\varepsilon\|_2^2}{2\lambda k} + \mathrm{pen}(\beta), \\ \Rightarrow \quad & \limsup_{k\to+\infty} \mathrm{pen}(\hat{\beta}(y^k)/k) \leq \mathrm{pen}(\beta). \end{aligned} \tag{5}$$

Consequently, the sequence $\left(\text{pen}(\beta(y^k)/k)\right)_{k\in\mathbb{N}^*}$ is bounded. In addition, the Cauchy-Schwarz inequality gives the following implications

$$\begin{aligned}
&\frac{1}{2}\|\varepsilon + X(k\beta) - X\hat\beta(y^k)\|_2^2 + \lambda\text{pen}(\hat\beta(y^k)) \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\text{pen}(k\beta),\\
\Rightarrow\quad & -\|\varepsilon\|_2\|X(k\beta) - X\hat\beta(y^k)\|_2 + \frac{1}{2}\|X(k\beta) - X\hat\beta(y^k)\|_2^2 + \lambda\text{pen}(\hat\beta(y^k)) \leq \lambda\text{pen}(k\beta),\\
\Rightarrow\quad & -\frac{\|\varepsilon\|_2}{k}\left\|X\left(\frac{\hat\beta(y^k)}{k} - \beta\right)\right\|_2 + \frac{1}{2}\left\|X\left(\frac{\hat\beta(y^k)}{k} - \beta\right)\right\|_2^2 + \frac{1}{k}\lambda\text{pen}\left(\frac{\hat\beta(y^k)}{k}\right) \leq \frac{\lambda\text{pen}(\beta)}{k}. \quad (6)
\end{aligned}$$

Let $l \in [0, +\infty]$ be the superior limit of the following sequence

$$\left(\left\|X\left(\hat\beta(y^k)/k - \beta\right)\right\|_2\right)_{k\in\mathbb{N}^*}. \tag{7}$$

According to (5) and (6) the following inequality occurs

$$\limsup_{k\to+\infty} \frac{\lambda\text{pen}(\beta) - \lambda\text{pen}(\beta(y^k)/k)}{k} = 0 \geq \begin{cases} l^2/2 \text{ if } l < +\infty \\ +\infty \text{ if } l = +\infty \end{cases}.$$

Consequently, $l = 0$ and thus sequence (7) is also bounded.

Due to the uniform uniqueness we have $\ker(\text{pen}) \cap \ker(X) = \{0\}$ and thus, according to Lemma 8, the sequence $((\hat\beta(y^k)/k)_{k\in\mathbb{N}^*}$ is bounded. Therefore, to prove that $\lim_{k\to+\infty}\hat\beta(y^k)/k = \beta$ it is sufficient to show that β is the unique limit point of this sequence. Let us extract a subsequence $(\hat\beta(y^{\phi(k)})/\phi(k))_{k\in\mathbb{N}^*}$ converging to $\gamma \in \mathbb{R}^p$ (where $\phi : \mathbb{N}^* \to \mathbb{N}^*$ is an increasing function). By (5) one may deduce that $\text{pen}(\gamma) \leq \text{pen}(\beta)$. Moreover by (6) one may deduce the following limit

$$\lim_{k\to+\infty}\left\|X\left(\frac{\hat\beta(y^{\phi(k)}}{\phi(k)} - \beta\right)\right\|_2^2 = \|X(\gamma - \beta)\|_2^2 = 0 \text{ (since other terms of (6) converge to 0).}$$

Finally γ satisfies the following equality and inequality

$$X\gamma = X\beta \text{ and } \text{pen}(\gamma) \leq \text{pen}(\beta).$$

Let us show that the unique element satisfying the above equality and inequality is $\gamma = \beta$. Because the model pattern of β is accessible then there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\text{pen}}(\beta)$. Let $y = X\beta + \lambda z$ then $\beta \in S_{X,\lambda\text{pen}}(y)$. Consequently, if there exists $\gamma \neq \beta$ such that $X\beta = X\gamma$ and $\text{pen}(\gamma) \leq \text{pen}(\beta)$ then, one may deduce that $\gamma \in S_{X,\lambda\text{pen}}(y)$ and thus $S_{X,\lambda\text{pen}}(y)$ is not a singleton leading to a contradiction. Consequently, $\gamma = \beta$ and thus

$$\lim_{k\to+\infty}\frac{\hat\beta(y^k)}{k} = \beta.$$

$\square$

Proof: Sufficient condition of Theorem 4. iii) Let $q := \min\{|[D\beta]_i| : i \in \text{supp}(D\beta)\} > 0$. Because $\hat\beta(y^k)/k$ converges to β when k tends to $+\infty$ then $\frac{1}{k}D\hat\beta(y^k)$ tends to $D\beta$ and consequently the following inequality holds

$$\exists k_0 \in \mathbb{N} \text{ such that } \forall k \geq k_0, \left\|\frac{1}{k}D\hat\beta(y^k) - D\beta\right\|_\infty < q/2.$$

Let $\tau = qk/2$ and let us show that $\text{sign}([D\beta(y^k)]^\tau) = \text{sign}(D\beta)$ once $k \geq k_0$.
When $i \notin \text{supp}(D\beta)$ then

$$\forall k \geq k_0, \left|\frac{1}{k}[D\hat\beta(y^k)]_i\right| < q/2 \Rightarrow \forall k \geq k_0, \left|[D\hat\beta(y^k)]_i\right| < \tau \Rightarrow \forall k \geq k_0, \text{sign}([D\hat\beta(y^k)]_i^\tau) = \text{sign}([D\beta]_i) = 0$$

When $[D\beta]_i > 0$ (thus $[D\beta]_i \geq q$) then

$$\forall k \geq k_0, \left|\frac{1}{k}[D\hat{\beta}(y^k)]_i - [D\beta]_i\right| < q/2 \quad \Rightarrow \quad [D\hat{\beta}(y^k)]_i > k([D\beta]_i - q/2) \geq \tau$$
$$\Rightarrow \quad \mathrm{sign}([D\hat{\beta}^r]_i^\tau) = \mathrm{sign}([D\beta]_i) = 1.$$

When $[D\beta]_i < 0$ (thus $[D\beta]_i \leq -q$) then

$$\forall k \geq k_0, \left|\frac{1}{k}[D\hat{\beta}(y^k)]_i - [D\beta]_i\right| < q/2 \quad \Rightarrow \quad [D\hat{\beta}(y^k)]_i < k([D\beta]_i + q/2) \leq -\tau$$
$$\Rightarrow \quad \mathrm{sign}([D\hat{\beta}^r]_i^\tau) = \mathrm{sign}([D\beta]_i) = -1.$$

iv) Let us remind that π be a permutation of $\mathcal{S}_p$ such that $|[D\hat{\beta}(y^k)]_{\pi(1)}| \geq |[D\hat{\beta}(y^k)]_{\pi(2)}| \geq \cdots \geq |[D\hat{\beta}(y^k)]_{\pi(p)}|$. Clearly, for every $\tau \geq 0$ the support of $[D\hat{\beta}(y^k)]^\tau$ lies onto the following family of nested subsets:

$$\mathrm{supp}([D\hat{\beta}(y^k)]^\tau) \in \{\varnothing, \{\pi(1)\}, \{\pi(1), \pi(2)\}, \ldots, \mathrm{supp}(D\hat{\beta}(y^k))\}.$$

Consequently, if iii) occurs, i.e. if $\mathrm{sign}([D\hat{\beta}(y^k)]^\tau) = \mathrm{sign}(D\beta)$ for some threshold τ then $\mathrm{supp}([D\hat{\beta}(y^k)]^\tau) = \mathrm{supp}(D\beta)$ and consequently, the support of $D\beta$ lies into the family of nested subsets

$$\mathrm{supp}(D\beta) \in \{\varnothing, \{\pi(1)\}, \{\pi(1), \pi(2)\}, \ldots, \mathrm{supp}(D\hat{\beta}(y^k))\}.$$

□

Proof of the sufficient condition for Theorem 3 is based on Lemma 9 and on Lemma 10 given hereafter.

Lemma 10. *Let* pen *be a polyhedral gauge defined by* $\forall x \in \mathbb{R}^p, \mathrm{pen}(x) = \max\{u_i'x, \ldots, u_l'x\}$ *and* $\beta \in \mathbb{R}^p$*. Then, there exists* $\tau > 0$ *depending on* β *such that*

$$\forall z \in B_\infty(\beta, \tau), \partial_{\mathrm{pen}}(z) \subset \partial_{\mathrm{pen}}(\beta). \tag{8}$$

Proof. Let $I := \{i \in [l] : u_i'\beta = \mathrm{pen}(x)\}$ then, according to Lemma... $\partial_{\mathrm{pen}}(\beta) = \mathrm{conv}\{u_i\}_{i\in I}$. Because the following inequalities are true

$$\forall i \notin I, u_i'\beta < \mathrm{pen}(\beta),$$

then by continuity of linear functions and by continuity of pen, one may pick $\tau > 0$ small enough such that

$$\forall z \in B_\infty(\beta, \tau), \forall i \notin I, u_i'z < \mathrm{pen}(z).$$

Consequently, whatever $z \in B_\infty(\beta, \tau)$ we have $J := \{i \in [l] : u_i'z = \mathrm{pen}(z)\} \subset I$ and thus

$$\partial_{\mathrm{pen}}(z) = \mathrm{conv}\{u_i\}_{i\in J} \subset \mathrm{conv}\{u_i\}_{i\in I} = \partial_{\mathrm{pen}}(\beta).$$

□

Proof: Sufficient condition of Theorem 3. According to Lemma 10 there exists $\tau_0 > 0$ such that whatever $z \in B_\infty(\beta, \tau_0)$ we have $\partial_{\mathrm{pen}}(z) \subset \partial_{\mathrm{pen}}(\beta)$. According to Lemma 9 for $\hat{\beta}(y^k)/k$ converges to β when k tends to $+\infty$. Consequently,

$$\exists k_0 \in \mathbb{N} \text{ such that } \forall k \geq k_0, \|\hat{\beta}(y^k)/k - \beta\|_\infty < \tau_0/2.$$

Consequently, for $k \geq k_0$ we have

$$\begin{cases} \forall u \in B_\infty(\hat{\beta}(y^k)/k, \tau_0/2), \partial_{\mathrm{pen}}(u) \subset \partial_{\mathrm{pen}}(\beta) \\ \exists \hat{u} \in B_\infty(\hat{\beta}(y^k)/k, \tau_0/2), \partial_{\mathrm{pen}}(\hat{u}) = \partial_{\mathrm{pen}}(\beta) \end{cases}$$

Since, whatever $t > 0$ and whatever $x \in \mathbb{R}^p$ we have $\partial_{\text{pen}}(x) = \partial_{\text{pen}}(tx)$ then, one may deduce that

$$\begin{cases} \forall u \in B_\infty(\hat{\beta}(y^k), k\tau_0/2), \partial_{\text{pen}}(u) \subset \partial_{\text{pen}}(\beta) \\ \exists \hat{u} \in B_\infty(\hat{\beta}(y^k), k\tau_0/2), \partial_{\text{pen}}(\hat{u}) = \partial_{\text{pen}}(\beta) \end{cases}$$

Consequently, one achieves the proof by taking $\tau = k\tau_0/2$. □

References

A. Ali and R. J. Tibshirani. The generalized lasso problem and uniqueness. *Electronic Journal of Statistics*, 13:2307–2347, 2019.

M. Bogdan, E. van den Berg, W. S. C. Sabatti, and E. J. Candès. Slope – adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9:1103–1140, 2015.

H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64:115–123, 2008.

P. Bühlmann and S. Van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, Heidelberg, 2011.

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

P. Descloux and S. Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. *Journal of Computational and Graphical Statistics*, pages 1–29, 2020.

P. Descloux, C. Boyer, J. Josse, A. Sportisse, and S. Sardy. Robust Lasso-zero for sparse corruption and model selection with missing covariates. Technical Report 2005.05628, arXiv, 2020.

C. Dossal. A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization. *Comptes Rendus Mathematique*, 350(1-2):117–120, 2012.

X. Dupuis and P. Tardivel. Proximal operator for the sorted ℓ_1 norm with application to testing procedures based on slope. 2021.

X. Dupuis and S. Vaiter. The geometry of sparse analysis regularization. Technical Report 1907.01769, arXiv, 2019.

K. Ewald and U. Schneider. Model selection properties and uniqueness of the Lasso estimator in low and high dimensions. *Electronic Journal of Statistics*, 14:944–969, 2020.

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

J.-J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

J. C. Gilbert. On the solution uniqueness characterization in the l1 norm and polyhedral gauge recovery. *Journal of Optimization Theory and Applications*, 172:70–101, 2017.

P. Gruber. *Convex and Discrete Geometry.* Springer, Heidelberg, 2007.

J.-B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms I: Fundamentals*, volume 305. Springer, Heidelberg, 1993.

J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146. PMLR, 2016.

H. Jégou, T. Furon, and J.-J. Fuchs. Anti-sparse coding for approximate nearest neighbor search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2029–2032. IEEE, 2012.

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 Trend Filtering. *SIAM Rev.*, 51(2):339–360, 2009. ISSN 0036-1445. doi: 10.1137/070690274.

P. Kremer, D. Brzyski, M. Bogdan, and S. Paterlini. Sparse index clones via the sorted ℓ_1-norm. Technical Report 3412061, Social Science Research Network, 2019.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.

S. Mousavi and J. Shen. Solution uniqueness of convex piecewise affine functions based optimization with applications to constrained l_1 minimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:1–56, 2019.

M. Osborne, B. Presnell, and B. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9:319–337, 2000.

A. Owrang, M. Malek-Mohammadi, A. Proutiere, and M. Jansson. Consistent change point detection for piecewise constant signals with normalized fused lasso. *IEEE signal processing letters*, 24(6): 799–803, 2017.

P. Pokarowski, W. Rejchel, A. Soltys, M. Frej, and J. Mielniczuk. Improving lasso for model selection and prediction. *arXiv preprint arXiv:1907.03025*, 2019.

J. Qian and J. Jia. On stepwise pattern recovery of the fused lasso. *Computational Statistics & Data Analysis*, 94:221–237, 2016.

R. Rockafellar. *Convex Analysis.* Princeton University Press, 1997.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5:941–973, 2004.

V. Saligrama and M. Zhao. Thresholded basis pursuit: Lp algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements. *IEEE Transactions on Information Theory*, 57:1567–1586, 2011.

U. Schneider and P. Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. Technical Report 2004.09106, arxiv, 2020.

A. Sepehri and N. Harris. The accessible lasso models. *Statistics*, 51:711–721, 2017.

Y. She et al. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4:1055–1096, 2010.

A. Takahashi and S. Nomura. Efficient path algorithms for clustered lasso and oscar. *arXiv preprint arXiv:2006.08965*, 2020.

P. Tardivel and M. Bogdan. On the sign recovery by lasso, thresholded lasso and thresholded basis pursuit denoising. Technical Report 1812.05723, arxiv, 2018.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

R. J. Tibshirani. The Lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

R. J. Tibshirani, M. Sanders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, 67:91–108, 2005.

R. J. Tibshirani, J. Taylor, et al. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011.

S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287, 2015.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

A. Weinstein, W. J. Su, M. Bogdan, R. F. Barber, and E. J. Candès. A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*, 2020.

X. Zeng and M. Figueiredo. Decreasing weighted sorted ℓ_1 regularization. *IEEE Signal Processing Letters*, 21:1240–1244, 2014.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

G. Ziegler. *Lectures on Polytopes*, volume 152. Springer, New York, 2012.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.