Pattern Recovery in Penalized and Thresholded Estimation and its Geometry

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

Pattern Recovery in Penalized and Thresholded Estimation and its Geometry*

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

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

March 3, 2023

Abstract

For many penalized estimators the penalty term is a real-valued polyhedral gauge such as LASSO, SLOPE, OSCAR, PACS, fused LASSO, clustered LASSO and generalized LASSO. This article focuses on the subdifferential recovery at β with respect to a penalty term (also called pattern recovery), where β is an unknown parameter of regression coefficients. For LASSO, when the penalty term is the ℓ_1 norm, the pattern of β only depends on the sign of β and sign recovery by LASSO is a well known topic in the litterature. We generalize the notion of pattern recovery and illustrate it for many examples of real-valued polyhedral gauge penalty. We provide theoretical guarantees for pattern recovery; in particular the "noiseless recovery condition" is necessary for a probability of recovery larger than 1/2. This condition may be relaxed using thresholded penalized least squares estimators; a new class of estimators generalizing thresholded LASSO. Indeed, we show that the "accessibility condition", a weaker condition than the "noiseless recovery condition", is necessary and asymptotically sufficient for pattern recovery by a thresholded penalized least squares estimator.

1 Introduction

We 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 random noise having a symmetric and continuous distribution with a positive density on $\mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ is the vector of unknown regression coefficients.

Penalized estimation of β has been studied extensively in literature. This includes methods such as the LASSO (Chen and Donoho, 1994; Tibshirani, 1996), SLOPE (Zeng and Figueiredo, 2014; Bogdan et al., 2015; Negrinho and Martins, 2014), OSCAR (Bondell and Reich, 2008), fused LASSO (Tibshirani et al., 2005), clustered LASSO (She, 2010), PACS (Sharma et al., 2013) and generalized LASSO (Tibshirani and Taylor, 2012). When the loss function is the residual sum of squares, these estimators minimize the function given by $b \in \mathbb{R}^p \mapsto \frac{1}{2}\|y - Xb\|_2^2 + \lambda \mathrm{pen}(b)$. The penalty term "pen" is a real-valued

*Previous versions of this article were also circulated under the title "The Geometry of Model Recovery by Penalized and Thresholded Estimators".

†The order of authors is alphabetical.

polyhedral gauge, i.e., a non-negative and positively homogeneous convex function that vanishes at 0 and whose unit ball is given by a (possibly unbounded) polyhedron.

Each of these estimators typically exhibits a particular structure, as illustrated, e.g., in Vaiter et al. (2015). The LASSO is sparse (some components of this estimator may be equal zero), the fused LASSO is sparse and some adjacent components are equal (Tibshirani et al., 2005), the supremum norm promotes clustering of components that are maximal in absolute value (Jégou et al., 2012), and SLOPE as well as OSCAR display further clustering phenomena where certain components may be equal in absolute value (Bondell and Reich, 2008; Figueiredo and Nowak, 2016; Schneider and Tardivel, 2022; Bogdan et al., 2022; Skalski et al., 2022). In this article, we use a general geometric approach to characterize the different structures inherent to these methods.

1.1 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

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

The solution set $S_{X,\lambda\mathrm{pen}}$ is always non-empty when pen is a real-valued polyhedral gauge. This can be learned from Proposition 3 in Appendix A[1]. Note that $S_{X,\lambda\mathrm{pen}}$ does not have to be a singleton and we treat uniqueness by giving a necessary and sufficient condition for it in Section 5. We now introduce the notion of a pattern equivalence class, a central concept to this article.

Definition 1 (Pattern equivalence class)**.** *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$*. We say that* β *and* $\tilde{\beta}$ *with* $\beta, \tilde{\beta} \in \mathbb{R}^p$ *share the same pattern with respect to* pen *if*

$$\partial_{\mathrm{pen}}(\beta) = \partial_{\mathrm{pen}}(\tilde{\beta}),$$

where ∂_{pen} *denotes the subdifferential of* pen*. We then write* $\beta \overset{\mathrm{pen}}{\sim} \tilde{\beta}$*. The set of all elements of* $\mathbb{R}^p$ *sharing the same pattern as* β *is called the pattern equivalence class* C_β*.*

We prove that the pattern equivalence classes coincide with the normal cones of the polar dual to the unit ball of pen in Theorem 4 in Appendix C.4. This interesting property is illustrated in Figures 1-4 in Section 2.2.

For the ℓ_1 norm two vectors $\beta, \tilde{\beta} \in \mathbb{R}^p$ have the same pattern if and only if $\mathrm{sign}(\beta) = \mathrm{sign}(\tilde{\beta})$. More generally, two vectors having the same pattern with respect to a real-valued polyhedral gauge penalty share a specific structure as illustrated on many examples in section 2.2. Given X and Y, we aim at recovering the pattern of β; for LASSO this means recovering $\mathrm{sign}(\beta)$.

In this article, Theorem 1 gives a necessary condition (called noiseless recovery condition) for 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 well known necessary conditions for sign recovery by LASSO.

Sign recovery by LASSO

We note $\hat{\beta}^{\mathrm{LASSO}}$ as a unique element of $S_{X,\lambda\|.\|_1}(Y)$ (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

[1]The existence of a minimizer is clear when pen is a norm. For the special case of the generalized LASSO (in which pen is not a norm), existence is shown in Ali and Tibshirani (2019) or Dupuis and Vaiter (2019). However, these proofs do not carry over to the general case.

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, the sign recovery is impossible. A geometrical characterization of accessible sign vectors is given in Sepehri and Harris (2017); Schneider and Tardivel (2022). 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 $\overline{I} := \{i \in [p] : \beta_i = 0\}$.

Other results towards pattern recovery

SLOPE. The notions of accessibility condition and irrepresentability condition for SLOPE have been recently introduced respectively in Schneider and Tardivel (2022) and in Bogdan et al. (2022). In particular, in this last article, similarly as for LASSO, when the SLOPE irrepresentability condition does not occur, the probability of pattern recovery is smaller than 1/2.

Generalized LASSO. By substituting the ℓ_1 norm by a real-valued polyhedral gauge $\text{pen} = \|D.\|_1$, 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 pattern recovery. Of course, the pattern induced by generalized LASSO depends on the matrix D.

For instance, when D is a matrix such that $Db = (b_2 - b_1, \ldots, b_p - b_{p-1})'$ (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: $\{i \in [p-1] : \beta_i \neq \beta_{i+1}\}$ (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 real-valued 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 pattern. Specifically, the model subspace of $x \in \mathbb{R}^p$ is a vector subspace of $\mathbb{R}^p$ perpendicular to $\partial_{\text{pen}}(x)$. For the ℓ_1 norm two vectors $x, z \in \mathbb{R}^p$ have the same model subspace when $\{i \in [p] : x_i \neq 0\} = \{i \in [p] : z_i \neq 0\}$. In the particular case of LASSO, Theorem 6 in Vaiter et al. (2015) shows that $\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} \text{sign}(\beta_I)\|_\infty < 1$ is a sufficient condition for model subspace recovery, *i.e.* the recovery of $\{i \in [p] : \beta_i \neq 0\}$. Whereas correct, this statement is not optimal. Indeed when $\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} \text{sign}(\beta_I)\|_\infty < 1$ it is well known that LASSO actually can recover $\text{sign}(\beta)$ (and a fortiori $\{i \in [p] : \beta_i \neq 0\}$) (Wainwright, 2009). Whereas we do not retain the notion of model subspace in this article; in supplementary material, we prove that the linear span of a pattern equivalence class coincides with the model subspace.

The noiseless recovery condition as well as the irrepresentability condition can be relaxed using structured estimators as explained hereafter.

1.2 Pattern recovery by a structured estimator

Theorem 2 generalizes 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, with a large probability, the pattern of β under a weaker condition than penalized least squares estimators (which are not thresholded). Now we introduce the notion of structured estimator.

Definition 2 (Thresholded penalized least squares estimator)**.**
Let pen *be a real-valued 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 $\hat{\beta}^{\mathrm{thr}}$ is a structured estimator of $\hat{\beta}$ if $\partial_{\mathrm{pen}}(\hat{\beta}) \subseteq \partial_{\mathrm{pen}}(\hat{\beta}^{\mathrm{thr}})$.*

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

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 for every $\tau \geq 0$ we have $\partial_{\|.\|_1}(\hat{\beta}^{\mathrm{LASSO}}) \subseteq \partial_{\|.\|_1}(\hat{\beta}^{\mathrm{LASSO},\tau})$ and thus $\hat{\beta}^{\mathrm{LASSO},\tau}$ is a structured 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, the accessibility condition is necessary for sign recovery by thresholded LASSO. Indeed, Tardivel and Bogdan (2022) recently proved that if $\mathrm{sign}(\hat{\beta}^{\mathrm{LASSO},\tau}) = \mathrm{sign}(\beta)$, then $\mathrm{sign}(\beta)$ is accessible for the LASSO. Moreover, they also proved that, 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, 2021; Descloux et al., 2022).

1.3 Notation

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: $\mathrm{row}(X) = \{X'z : z \in \mathbb{R}^n\}$. The defect of X is $\mathrm{def}(X) = \dim(\ker(X))$.
- Given $p \in \mathbb{N}$, the notation $[p]$ represents the set of integers $\{1, \dots, p\}$.
- Given $x \in \mathbb{R}^p$ and τ the notation x^τ represents the thresholded vector

$$x^\tau = \left(x_1 \mathbf{1}\left(|x_1| > \tau\right), \dots, x_p \mathbf{1}\left(|x_p| > \tau\right)\right).$$

- The notation $\overline{B}_\infty(a, r)$ represents the closed ball for the ℓ_∞ norm centered in a with radius r.

2 Examples of polyhedral gauges and examples of pattern equivalence class

2.1 Real-valued polyhedral gauges, polyhedral norms

It is known that a real-valued polyhedral gauge pen can be written as the maximum of linear functions as follows (Rockafellar, 1997; Mousavi and Shen, 2019)

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

Note that a polyhedral gauge whose unit ball $\{x \in \mathbb{R}^p : \text{pen}(x) \leq 1\}$ is a bounded and symmetric with respect to the origin polyhedron is a polyhedral norm. Examples of polyhedral norms are the ℓ_1 norm: $\|x\|_1 = \sum_{i=1}^p |x_i|$, the supremum norm: $\|x\|_\infty = \max\{|x_1|, \ldots, |x_p|\}$ and the sorted ℓ_1 norm: $\|x\|_w = \sum_{i=1}^p w_i |x|_{(i)}$, where $w_1 > 0$, $w_1 \geq \cdots \geq w_p \geq 0$ and $(.)$ is a permutation on $[p]$ such that $|x|_{(1)} \geq \cdots \geq |x|_{(p)}$.

Furthermore, the composition of real-valued polyhedral gauge with a linear map is still a real-valued polyhedral gauge. For example, for generalized LASSO, the penalty term is the real-valued polyhedral gauge $x \in \mathbb{R}^p \mapsto \|Dx\|_1$ where $D \in \mathbb{R}^{m \times p}$. Note that, when $\{0\} \subsetneq \ker(D)$, the function $x \in \mathbb{R}^p \mapsto \|Dx\|_1$ is not a norm but only 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

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

- 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

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

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

2.2 Subgradients, subdifferentials and patterns

We remind the reader of the definition on subgradient and subdifferential. The following can be found for instance in Hiriart-Urruty and Lemarechal (2001):
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 1. *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 real-valued polyhedral gauge $x \in \mathbb{R}^p \mapsto \|Dx\|_1$ we have $\partial_{\|D.\|_1}(x) = D' \partial_{\|.\|_1}(Dx)$ (see Hiriart-Urruty and Lemarechal (2001) page 184).

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 (2022); Schneider and Tardivel (2022). Now, we want to illustrate that two vectors $x, z \in \mathbb{R}^p$ having the same subdifferential with respect to a real-valued polyhedral gauge share a common pattern.

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

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

Subdifferentials $\partial_{\|.\|_1}(x) = \partial_{\|.\|_1}(z)$ are equal if and only if $\operatorname{sign}(x) = \operatorname{sign}(z)$. For instance when $x = (1.45, -0.38, 1.56, 0, -2.76)$ then $\operatorname{sign}(x) = (1, -1, 1, 0, -1)$.

Graphical illustration when $p = 2$ (Figure 1):

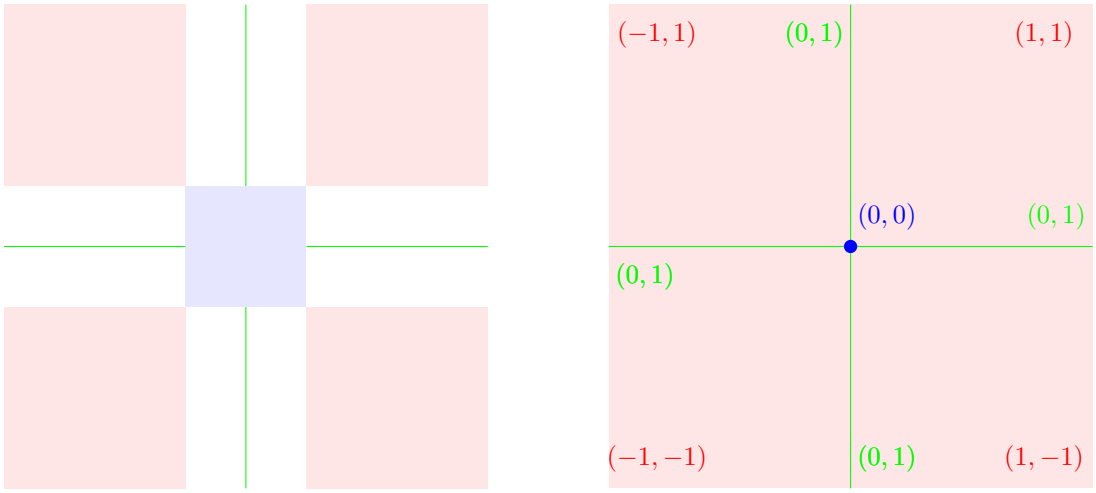

Figure 1: In this figure $\operatorname{pen}(b) = |b_1| + |b_2|$. On the left the blue polytope is $B^* = \partial_{\operatorname{pen}}(0) = \operatorname{conv}\{\pm(1,1), \pm(1,-1)\}$ (B^* is the unit ball of the ℓ_∞ norm). Red and green (unbounded) sets are preimages, with respect to the projection onto B^*, of vertices and edge centers. The picture on the right provides $\operatorname{sign}(x) \in \{(0,0), \pm(1,0), \pm(0,1), \pm(1,1), \pm(1,-1)\}$ depending on the localization of $x \in \mathbb{R}^2$.

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

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

Note that the notation $*$ represents a components which is null or not maximal in absolute value. Subdifferentials $\partial_{\|.\|_\infty}(x) = \partial_{\|.\|_\infty}(z)$ are equal if and only if $\operatorname{sign}^\infty(x) = \operatorname{sign}^\infty(z)$. For instance when $x = (1.45, 1.45, 0.56, 0, -1.45)$ then $\operatorname{sign}^\infty(x) = (1, 1, *, *, -1)$.

Graphical illustration when $p = 2$ (Figure 2):

Pattern for the sorted ℓ_1 norm: Let $x \in \mathbb{R}^p$. The SLOPE pattern of x, $\operatorname{patt}(x)$, is defined by

$$\operatorname{patt}(x)_i = \operatorname{sign}(x_i) \operatorname{rank}(|x|)_i, \quad \forall i \in [p]$$

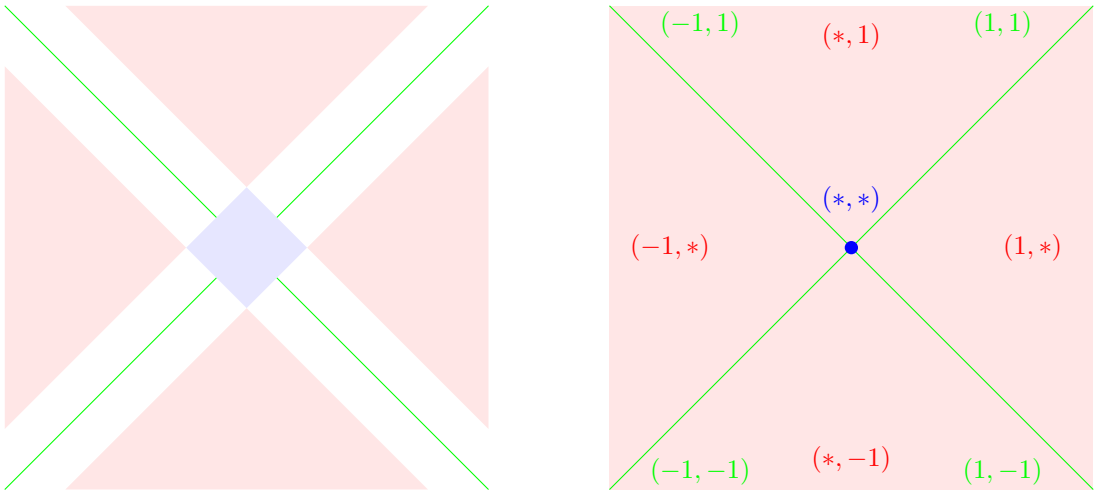

Figure 2: In this figure $\mathrm{pen}(b) = \max\{|b_1|, |b_2|\}$. On the left the blue polytope is $B^* = \partial_{\mathrm{pen}}(0) = \mathrm{conv}\{\pm(1,0), \pm(0,1)\}$ (B^* is the unit ball of the ℓ_1 norm). Red and green (unbounded) sets are preimages, with respect to the projection onto B^*, of vertices and edge centers. The picture on the right provides $\mathrm{sign}^\infty(x) \in \{(*,*), (*,\pm 1), (\pm 1, *), (\pm 1, \pm 1)\}$ depending on the localization of $x \in \mathbb{R}^2$.

where $\mathrm{rank}(|x|)_i \in \{0, 1, \ldots, k\}$, k is the number of nonzero distinct values in $\{|x_1|, \ldots, |x_p|\}$, $\mathrm{rank}(|x|)_i = 0$ if $x_i = 0$, $\mathrm{rank}(|x|)_i > 0$ if $|x_i| > 0$ and $\mathrm{rank}(|x|)_i < \mathrm{rank}(|x|)_j$ if $|x_i| < |x_j|$.
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 $\mathrm{patt}(x) = \mathrm{patt}(z)$. For instance when $x = (3.1, -1.2, 0.5, 0, 1.2, -3.1)$ then $\mathrm{patt}(x) = (3, -2, 1, 0, 2, -3)$.

Graphical illustration when $p = 2$ (Figure 3):

Pattern for the real-valued 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 $\mathrm{jump}(x) = \mathrm{jump}(z)$. For instance when $x = (1.45, 1.45, 0.56, 0.56, -0.45, 0.35)$ then $\mathrm{jump}(x) = (\rightarrow, \searrow, \rightarrow, \searrow, \nearrow)$.

Graphical illustration when $p = 2$ (Figure 4):

Pattern for the real-valued 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 $C := \cup_{i=1}^{p-1}[(i, x_i), (i+1, x_{i+1})]$. 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 $\mathrm{knot}(x) = \mathrm{knot}(z)$. For instance, Figure 5 provides an illustration of $\mathrm{knot}(x)$ for a particular $x \in \mathbb{R}^9$.

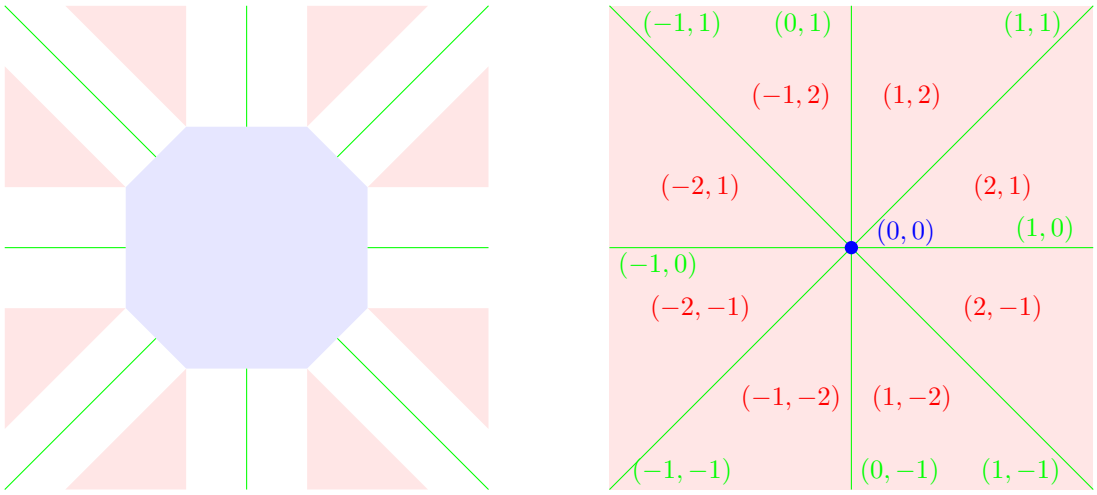

Figure 3: In this figure $\mathrm{pen}(b) = w_1|b|_{(1)} + w_2|b|_{(2)}$ for some $w_1 > w_2 > 0$. On the left the blue polytope is $B^* = \partial_{\mathrm{pen}}(0) = \mathrm{conv}\{\pm(w_1, w_2), \pm(w_1, -w_2), \pm(w_2, w_1), \pm(w_2, -w_1)\}$ (B^*, also called the signed permutahedron, is the unit ball of the dual sorted ℓ_1 norm). Red and green (unbounded) sets are preimages, with respect to the projection onto B^*, of vertices and edge centers. The picture on the right provides $\mathrm{patt}(x) \in \{(0,0), \pm(1,0), \pm(0,1), \pm(1,1), \pm(1,-1), \pm(1,2), \pm(1,-2), \pm(2,1), \pm(2,-1)\}$ depending on the localization of $x \in \mathbb{R}^2$.

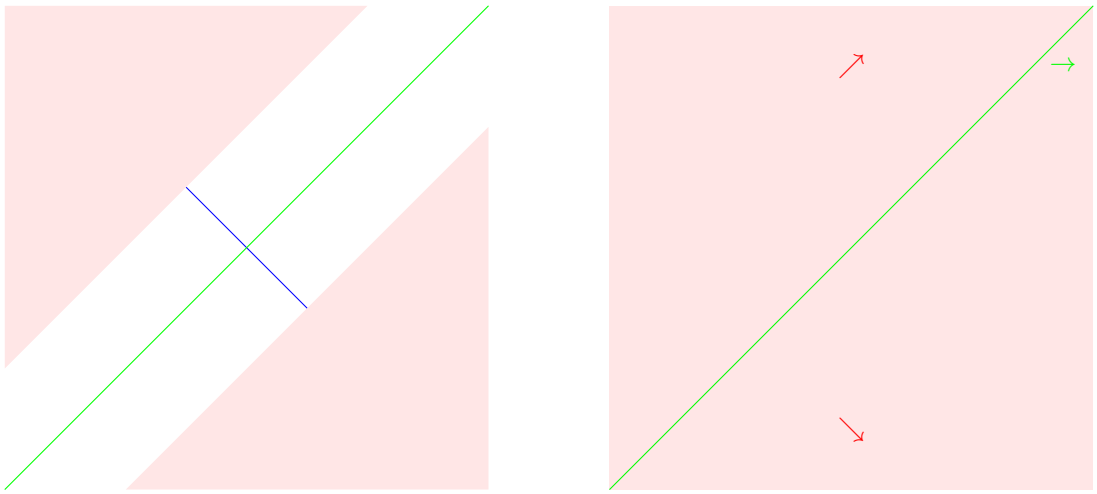

Figure 4: In this figure $\mathrm{pen}(b) = |b_1 - b_2|$. On the left the blue polytope is $B^* = \partial_{\mathrm{pen}}(0) = \mathrm{conv}\{\pm(1 - 1)\}$. Red and green (unbounded) sets are preimages, with respect to the projection onto B^*, of vertices and edge center. The picture on the right provides $\mathrm{jump}(x) \in \{\nearrow, \rightarrow, \searrow\}$ depending on the localization of $x \in \mathbb{R}^2$.

Illustration when $p = 3$: When $p = 3$ then $\mathrm{pen}(b) = |b_1 - 2b_2 + b_3|$ and $B^* = \partial_{\mathrm{pen}}(0) = \mathrm{conv}\{\pm(1, -2, 1)\}$. The three normal cones of the segment B^* are the half-space $\{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 \leq (x_1 + x_3)/2\}$ associated to the vertex $(1, -2, 1)$, $\{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 = (x_1 + x_3)/2\}$ a vector plane perpendicular to B^* and the half-space $\{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 \geq (x_1 + x_3)/2\}$ associated to the vertex $(-1, 2, -1)$. Therefore, the sets $\{x \in \mathbb{R}^3 : \mathrm{knot}(x) = cx\}$, $\{x \in \mathbb{R}^3 : \mathrm{knot}(x) = l\}$ and $\{x \in \mathbb{R}^3 : \mathrm{knot}(x) = cx\}$ are relative interior of normal cones of B^*.

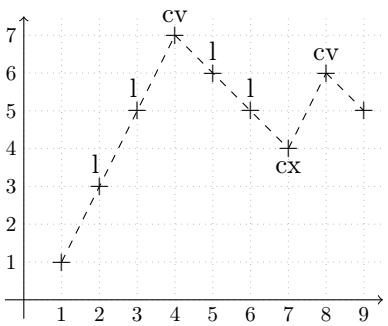

Figure 5: In this figure the dotted curve represents C described above for $x = (1, 3, 5, 7, 6, 5, 4, 6, 5)$. Here $\text{knot}(x) = (l, l, cv, l, l, cx, cv)$.

As illustrated above, $B^* = \partial_{\text{pen}}(0)$ is a a polytope in $\mathbb{R}^p$. It is known that relative interiors of normal cones of a polytope provide a partition in $\mathbb{R}^p$ (see *e.g.* Theorem 4.13 page 17 in Ewald (1996)). In Appendix C.4 we prove that the partition given by relative interior normal cones of B^* coincides with equivalent classes for the relation $\overset{\text{pen}}{\sim}$.

3 Pattern recovery in penalized estimation

3.1 Accessibility: a necessary condition for pattern recovery with a positive probability

We introduce the notion of accessible patterns in the following definition. This definition generalizes the notion of accessible sign vectors (Sepehri and Harris, 2017; Schneider and Tardivel, 2022) and accessible patterns for SLOPE (Schneider and Tardivel, 2022) to a broad class of penalized estimators.

Definition 3 (Accessible pattern)**.** *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 pattern with respect to X and λ*pen*, if there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ such that $\hat{\beta} \overset{\text{pen}}{\sim} \beta$.*

When pen is the ℓ_1-norm scaled by a tuning parameter $\lambda > 0$, i.e., $\text{pen} = \lambda\|.\|_1$ the above definition coincides with the notion of accessibility of sign vectors with respect to X. When pen is the sorted ℓ_1-norm, i.e., $\text{pen} = \|.\|_w$ for some $w \in \mathbb{R}^p$ with $w_1 > \cdots > w_p > 0$, the above definition coincides with the notion of accessible SLOPE patterns with respect to X. Proposition 1 provides both a geometric and an analytic characterization of accessible patterns.

Proposition 1 (Characterization of accessible patterns)**.** *Let $X \in \mathbb{R}^{n\times p}$ and* pen $: \mathbb{R}^p \to \mathbb{R}$ *be a real-valued polyhedral gauge.*

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

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

*2) Analytic characterization: The pattern of $\beta \in \mathbb{R}^p$ is accessible with respect to X and λ*pen *if and only if for any $b \in \mathbb{R}^p$ the implication*

$$X\beta = Xb \implies \text{pen}(\beta) \leq \text{pen}(b)$$

holds.

Based on Proposition 1, it is clear that the notion of accessibility does not depend on the tuning parameter λ.

3.2 The noiseless recovery condition: a necessary condition for pattern recovery with a probability larger than 1/2

The solution path for a penalized estimator is the curve $0 < \lambda \mapsto \hat{\beta}_\lambda$ where $\hat{\beta}_\lambda$ is the unique element of $S_{X,\lambda\text{pen}}(y)$ for fixed $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$. The solution path for the generalized LASSO or OSCAR and Clustered LASSO is studied in Tibshirani and Taylor (2012) or Takahashi and Nomura (2020), respectively. Definition 4 is based on this notion of a solution path. Note that Definition 4 does not require uniqueness of estimator.

Definition 4 (Noiseless recovery condition)**.** *Let* pen *be a real-valued polyhedral gauge,* $X \in \mathbb{R}^{n\times p}$ *and* $\beta \in \mathbb{R}^p$*. We say that the pattern of* β *satisfies the noiseless recovery condition with respect to* X *and* pen *if*

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

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

The noiseless recovery condition is illustrated for the LASSO in Figure 6 in the particular case where X and β are given hereafter

$$X = \begin{pmatrix} 5/6 & 1 & 0 \\ 1/3 & 0 & 1 \end{pmatrix} \text{ and } \beta = (10, 0, 0)'.$$

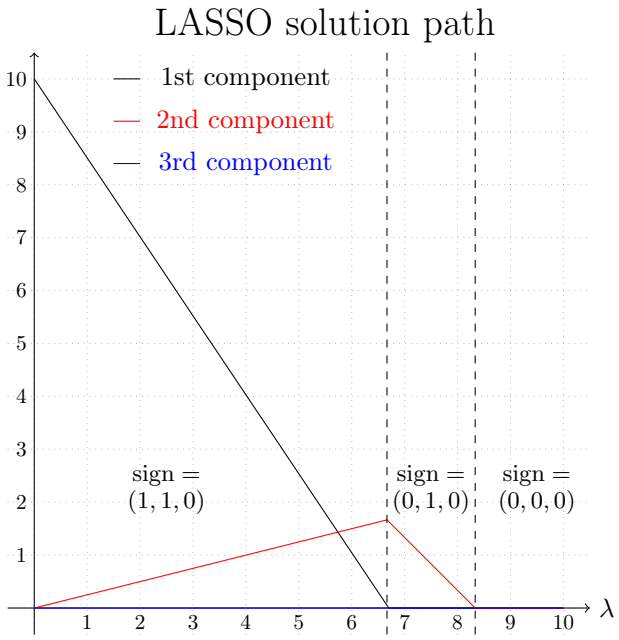

Figure 6: This figure provides curves of the functions $\lambda > 0 \mapsto (\hat{\beta}_\lambda^{\text{LASSO}})_1$ (black curve), $\lambda > 0 \mapsto (\hat{\beta}_\lambda^{\text{LASSO}})_2$ (red curve), $\lambda > 0 \mapsto (\hat{\beta}_\lambda^{\text{LASSO}})_3$ (blue curve). Note that $\text{sign}(\beta)$ does not satisfy the noiseless recovery condition. Indeed, $\text{sign}(\hat{\beta}_\lambda^{\text{LASSO}}) = (1, 1, 0)'$ for $\lambda \in (0, \lambda_1)$ (where $\lambda_1 = 20/3$), $\text{sign}(\hat{\beta}_\lambda^{\text{LASSO}}) = (0, 1, 0)'$ for $\lambda \in [\lambda_1, \lambda_2)$ (where $\lambda_2 = 25/3$) and $\text{sign}(\hat{\beta}_\lambda^{\text{LASSO}})) = (0, 0, 0)'$ for $\lambda \geq \lambda_2$. Consequently, for every $\lambda > 0$, $\text{sign}(\hat{\beta}_\lambda^{\text{LASSO}}) \neq (1, 0, 0)'$.

In supplementary material, we provides a geometrical characterisation of the noiseless recovery condition. Neither the above definition nor the geometrical characterisation provide an analytic ex-

pression for checking the noiseless recovery condition, but some formulas are given in the literature. For example, when pen $= \|.\|_1$, the noiseless recovery condition is equivalent to

$$\|X'(X_I')^+\text{sign}(\beta_I)\|_\infty \leq 1 \text{ and } \text{sign}(\beta_I) \in \text{row}(X_I), \tag{3}$$

where $I = \{i \in [p] : \beta_i \neq 0\}$ and X_I is the matrix whose columns are $(X_j)_{j\in I}$. Note that under the minor assumption that $\ker(X_I) = \{0\}$ then $\text{sign}(\beta_I) \in \text{row}(X_I)$ occurs and the expression (3) coincides with the irrepresentability condition: $\|X_{\overline{I}}'X_I(X_I'X_I)^{-1}\text{sign}(\beta_I)\|_\infty \leq 1$ where $X_{\overline{I}}$ is a matrix whose columns are $(X_j)_{j\notin I}$ (Bühlmann and Van de Geer, 2011; Wainwright, 2009; Zou, 2006; Zhao and Yu, 2006). Thus, the well known *irrepresentability condition* for LASSO can be thought of as an analytical shortcut for checking the noiseless recovery condition. Figure 6 confirms this. 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 6, one may observe that the noiseless recovery condition does not hold for β. For the sorted ℓ_1 norm when $M = \text{patt}(\beta)$, the noiseless recovery condition is equivalent to

$$\|X'(\tilde{X}_M')^+\tilde{W}_M\|_w^* \leq 1 \text{ and } \tilde{W}_M \in \text{row}(\tilde{X}_M),$$

where $\|.\|_w^*$ is the dual sorted ℓ_1 norm, $\tilde{X}_M$ is the clustered matrix and $\tilde{W}_M$ is the clustered parameter (see Bogdan et al. (2022) for more details). In appendix we also provide an analytic characterisation of noiseless pattern recovery when the penalty term is the supremum norm. However, this article does not aim at providing a list of analytical shortcuts for checking the noiseless recovery condition. In fact, we want to show that

a) The noiseless recovery condition is a necessary condition for pattern recovery with a probability larger than 1/2, see Theorem 1.

b) Thresholded penalized estimators recover the pattern of β under much weaker condition than the noiseless recovery condition, see Section 4.

Theorem 1. *Let $Y = X\beta + \varepsilon$ where $X \in \mathbb{R}^{n\times p}$ is a fixed matrix, $\beta \in \mathbb{R}^p$ and ε follows a symmetric distribution. Let* pen *be a real-valued polyhedral gauge. If β does not satisfy the noiseless recovery condition with respect to X and* pen*, then*

$$\mathbb{P}\left(\exists\lambda > 0\ \exists\hat{\beta} \in S_{X,\lambda\text{pen}}(Y) \text{ such that } \hat{\beta} \overset{\text{pen}}{\sim} \beta\right) \leq 1/2.$$

By Theorem 1, if the noiseless recovery condition does not hold for the LASSO (for example, when $\|X_{\overline{I}}'X_I(X_I'X_I)^{-1}\text{sign}(\beta_I)\|_\infty > 1$), the following holds

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

This above result is stronger than the one given in Theorem 2 in Wainwright (2009) which shows that $\mathbb{P}(\text{sign}(\hat{\beta}^{\text{LASSO}}(\lambda)) = \text{sign}(\beta)) \leq 1/2$ for fixed $\lambda > 0$.

Clearly, if β satisfies the noiseless recovery condition with respect to X and pen, β is accessible with respect to X and pen by taking $y = X\beta$ in the definition of accessibility. In the following section, we show that thresholded penalized least-squares estimators also recover the pattern of β under the accessibility condition.

4 Pattern recovery by thresholded penalized estimators

In practice, some additional information about β may be priorly known, e.g. its sparsity. Therefore it is quite natural to threshold small components of $\hat{\beta}^{\text{LASSO}}$ and so consider the thresholded LASSO estimator $\hat{\beta}^{\text{LASSO},\tau}$ for some threshold $\tau \geq 0$. Moreover, if the threshold is appropriately selected, the estimator allows to recover $\text{sign}(\beta)$ under weaker conditions than LASSO itself (Tardivel and

Bogdan, 2022). We aim at generalizing this property to a broader class of penalized estimators. Before introducing the notion of a structured estimator, recall that for any threshold $\tau \geq 0$, the inclusion $\partial_{\|.\|_1}(\hat{\beta}^{\text{LASSO}}) \subseteq \partial_{\|.\|_1}(\hat{\beta}^{\text{LASSO},\tau})$ occurs. This last inclusion is the keystone concept to introduce the notion of a structured estimator as defined in Definition 2 in the introduction. Some heuristic examples are listed hereafter:

(a) The penalty term $\|\cdot\|_\infty$ promotes clustering of components that are maximal in absolute value: Once $|\hat{\beta}_j| < \|\hat{\beta}\|_\infty$ but $|\hat{\beta}_j| \approx \|\hat{\beta}\|_\infty$, it is quite natural to set $|\hat{\beta}_j| = \|\hat{\beta}\|_\infty$. Let $\hat{\beta}^{\text{thr}}$ be the estimator taking into account this approximation, obtained after slightly modifying $\hat{\beta}$. Then $\partial_{\|.\|_\infty}(\hat{\beta}) \subseteq \partial_{\|.\|_\infty}(\hat{\beta}^{\text{thr}})$.

(b) The sorted ℓ_1 norm penalty promotes clustering of components equal in absolute value: Once $|\hat{\beta}^{\text{SLOPE}}_j| \approx |\hat{\beta}^{\text{SLOPE}}_i|$, it is quite natural to set $|\hat{\beta}^{\text{SLOPE}}_i| = |\hat{\beta}^{\text{SLOPE}}_j|$. Let $\hat{\beta}^{\text{thr}}$ be the estimator taking into account this approximation and obtained after modifying slightly $\hat{\beta}^{\text{SLOPE}}$. Then, $\partial_{\|.\|_w}(\hat{\beta}^{\text{SLOPE}}) \subseteq \partial_{\|.\|_w}(\hat{\beta}^{\text{thr}})$.

(c) The penalty term $\|D^{tv}\cdot\|$ promotes neighboring components to be equal: Once $\hat{\beta}_j \approx \hat{\beta}_{j+1}$, it is quite natural to set $\hat{\beta}_j = \hat{\beta}_{j+1}$. Let $\hat{\beta}^{\text{thr}}$ be the estimator taking into account this approximation and obtained after modifying slightly $\hat{\beta}$. Then, $\partial_{\|D^{\text{tv}}.\|_1}(\hat{\beta}) \subseteq \partial_{\|D^{\text{tv}}.\|_1}(\hat{\beta}^{\text{thr}})$.

The notion of accessibility introduced for penalized estimators in Section 3 also covers the structured estimators as can be learned from the proposition below.

Proposition 2. *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. We have*

$$\begin{aligned} &\exists y \in \mathbb{R}^n,\ \exists \hat{\beta} \in S_{X,\lambda\text{pen}}(y) \text{ such that } \hat{\beta} \overset{\text{pen}}{\sim} \beta \\ &\quad \iff \exists y \in \mathbb{R}^n, \exists \hat{\beta} \in S_{X,\lambda\text{pen}}(y) \text{ such that } \partial_{\text{pen}}(\hat{\beta}) \subseteq \partial_{\text{pen}}(\beta). \end{aligned}$$

According to Propositions 1 and 2, if there exists $b \in \mathbb{R}^p$ such that $Xb = X\beta$ and $\text{pen}(b) < \text{pen}(\beta)$, then for any $y \in \mathbb{R}^n$, $\lambda > 0$, and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ we have $\partial_{\text{pen}}(\hat{\beta}) \not\subseteq \partial_{\text{pen}}(\beta)$. Consequently, no penalized nor thresholded penalized estimator can recover the pattern of β.

On the other hand, if $\text{pen}(b) \geq \text{pen}(\beta)$ for all $b \in \mathbb{R}^p$ with $Xb = X\beta$, then both penalized and structured penalized estimator can recover the pattern of β with different "choices" of y. However, in practice, a statistician does not aim at picking the appropriate y to recover the pattern of β, but instead uses the response of a linear regression model as a particular y to infer this pattern.

In this direction, by Theorem 1, if $Y = X\beta + \varepsilon$, the noiseless recovery condition (a stronger condition than the accessibility condition) is necessary for recovering the pattern of β via a penalized estimator with probability larger than $1/2$. In Theorem 2, we relax the stringent noiseless recovery condition by considering a structured estimator. Before stating this theorem, we introduce a following class of structured estimators.

Definition 5 (τ-thresholded penalized estimator)**.** *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ and $\lambda \geq 0$. Given $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$, we say that $\hat{\beta}^{\text{str},\tau}$ is a τ-structured estimator of $\hat{\beta}$ if*

1) $\partial_{\text{pen}}(\hat{\beta}) \subseteq \partial_{\text{pen}}(\hat{\beta}^{\text{str},\tau})$,

2) $\|\hat{\beta} - \hat{\beta}^{\text{str},\tau}\|_\infty \leq \tau$,

3) $\dim(\partial_{\text{pen}}(b)) \leq \dim(\partial_{\text{pen}}(\hat{\beta}^{\text{str},\tau}))$ *for all b with $\|\hat{\beta} - b\|_\infty \leq \tau$.*

The thresholded LASSO is, in fact, an example of τ-structured estimator with threshold τ. Another example of a τ-structured estimator when the penalty term is the supremum norm, is given in Algorithm 1. Theorem 2 shows that a structured estimator recovers the pattern of β under the assumption that $Xb = X\beta$ implies $\text{pen}(b) \geq \text{pen}(\beta)$ and that the signal is "large enough", as is formalized in the following theorem.

Theorem 2. *Let* pen *be a real-valued polyhedral gauge, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, and $\lambda > 0$. Assume that uniform uniqueness holds, i.e. for any $y \in \mathbb{R}^n$, the set $S_{X,\lambda\mathrm{pen}}(y)$ contains the unique minimizer $\hat{\beta}(y)$. For $\varepsilon \in \mathbb{R}^n$ and for $r \in \mathbb{N}$ set $y^{(r)} = X(r\beta) + \varepsilon$. If $\mathrm{pen}(b) \geq \mathrm{pen}(\beta)$ for any $b \in \mathbb{R}^p$ with $Xb = X\beta$, then there exists $r_0 \in \mathbb{N}$ and $\tau \geq 0$ such that for all $r \geq r_0$*

$$\begin{cases} \partial_{\mathrm{pen}}(b) \subseteq \partial_{\mathrm{pen}}(\beta) \text{ for any } b \in \overline{B}_\infty(\hat{\beta}(y^{(r)}), \tau) \\ \exists b_0 \in \overline{B}_\infty(\hat{\beta}(y^{(r)}), \tau) \text{ such that } b_0 \overset{\mathrm{pen}}{\sim} \beta \end{cases}$$

Consequently, a τ-thresholded penalized estimator $\hat{\beta}^{\mathrm{str},\tau}(y^{(r)})$ recovers the pattern of β.

Similar results in which non-null components are large enough (i.e., $r \geq r_0$ in Theorem 2) are given in Tardivel and Bogdan (2022) and Descloux et al. (2022). In particular, Theorem 2 corroborates Theorem 1 in Tardivel and Bogdan (2022), which proves that the thresholded LASSO estimator recovers the sign of β once the accessibility condition holds and non-null components of β are large enough. Similarly as thresholded LASSO, when $\mathrm{pen} = \|.\|_\infty$, a τ-estimator can be explicitly computed as illustrated in Algorithm 1 below.

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**
    $\hat{\beta}^{\mathrm{str},\tau} \leftarrow 0$.
  else

$$\forall j \in [p]\ \hat{\beta}_j^{\mathrm{str},\tau} \leftarrow \begin{cases} \|\hat{\beta}\|_\infty - \tau & \text{if } \hat{\beta}_j \geq \|\hat{\beta}\|_\infty - 2\tau \text{ and } \hat{\beta}_j \geq 0, \\ -\|\hat{\beta}\|_\infty + \tau & \text{if } \hat{\beta}_j \leq -\|\hat{\beta}\|_\infty + 2\tau \text{ and } \hat{\beta}_j < 0, \\ \hat{\beta}_j & \text{otherwise.} \end{cases}$$

  end if
  return $\hat{\beta}^{\mathrm{str},\tau}$

5 A necessary and sufficient condition for uniform uniqueness

Since uniqueness is an assumption in Theorem 2, we provide a a necessary and sufficient condition for uniform uniqueness of the penalized optimization problem (1) in Theorem 3. This theorem relaxes the coercivity condition for the penalty term needed in Theorem 1 in Schneider and Tardivel (2022) and extends the result to encompass methods such as the generalized LASSO.

Theorem 3 (Necessary and sufficient condition for uniform uniqueness)**.** *Let $X \in \mathbb{R}^{n\times p}$ and $\lambda > 0$. Let* pen *be a real-valued polyhedral gauge, i.e.,* $\mathrm{pen}(x) = \max\{u_1'x, \ldots, u_k'x\}$ *for some $u_1, \ldots, u_k \in \mathbb{R}^p$ with $u_1 = 0$. Let*

$$S_{X,\lambda\mathrm{pen}}(y) = \operatorname*{Arg\,min}_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda\mathrm{pen}(b). \tag{4}$$

Then the solution to the above minimization problem is unique, i.e, $S_{X,\lambda\mathrm{pen}}(y)$ is a singleton for all $y \in \mathbb{R}^n$ if and only if $\mathrm{row}(X)$ *does not intersect a face of the polytope $B^* = \mathrm{conv}\{u_1, \ldots, u_k\}$ whose dimension*[2] *is strictly less than* $\mathrm{def}(X)$.

Note that a face F of B^* satisfies

$$\dim(F) < \mathrm{def}(X) \iff \mathrm{codim}(F) > \mathrm{rk}(X),$$

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

where $\mathrm{codim}(F) = p - \dim(F)$. We now illustrate some cases of non-uniqueness occurring for the generalized LASSO with $\mathrm{pen}(b) = \|Db\|_1$ for some $D \in \mathbb{R}^{m\times p}$. Clearly, the set of generalized LASSO minimizers $S_{X,\lambda\|D.\|_1}(y)$ is unbounded for every $y \in \mathbb{R}^n$ once $\ker(X) \cap \ker(D) \neq \{0\}$. Consequently, $\ker(X) \cap \ker(D) = \{0\}$ is a necessary condition for uniform uniqueness, yet, it is not sufficient, as illustrated in the example below.

Example 2. *An example of generalized LASSO optimization problem for which the set of minimizers is not restricted to a singleton is given in Barbara et al. (2019). We recall this example hereafter:*

$$\operatorname*{Arg\,min}_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \frac{1}{2}\|Db\|_1 \text{ where } X = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & 1 \\ \sqrt{2} & 0 & 0 \end{pmatrix},\ D = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \text{ and } y = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Note that $S_{X,\frac{1}{2}\|D.\|_1}(y) = \mathrm{conv}\{(0,1/2,0)', (0,0,1/2)'\}$ *(Barbara et al., 2019). Since*

$$\|Db\|_1 = \max\{\pm(4b_1 + 2b_2 + 2b_3), \pm(2b_1 + 2b_2), \pm(2b_1 + 2b_3)\}$$

then $B^* = \mathrm{conv}\{\pm(4,2,2)', \pm(2,2,0)', \pm(2,0,2)'\}$. *Because the vertex* $F = (4,2,2)'$ *is an element of* $\mathrm{row}(X)$ *and satisfies* $\dim(F) = 0 < 1 = \mathrm{def}(X)$ *then, according Theorem 3, the uniform uniqueness cannot hold. This complies with the fact that* $S_{X,\frac{1}{2}\|D.\|_1}(y)$ *is not a singleton.*

When $\ker(X) \cap \ker(D) = \{0\}$, in broad generality, the set of generalized LASSO minimizers is a polytope (a bounded polyhedron) (Barbara et al., 2019) and extremal points can be explicitly computed (Dupuis and Vaiter, 2019). This description is relevant when the set of minimizers is not a singleton.

6 Numerical experiments

Below, in our simulations, we consider the linear regression model $Y = X\beta + \varepsilon$ where:

- The matrix $X = (X_1|\dots|X_{150}) \in \mathbb{R}^{100\times 150}$ has iid $\mathcal{N}(0, 1/100)$ entries.
- The random noise $\varepsilon \in \mathbb{R}^n$ has iid $\mathcal{N}(0,1)$ entries.

Hereafter, given a set $S \subseteq [p]$, the notation X_S represents a matrix whose columns are $(X_i)_{i\in S}$.

Numerical experiments for LASSO

For LASSO, the noiseless recovery condition and the accessibility condition depend on β through $\mathrm{sign}(\beta) \in \{-1,0,1\}^p$. Moreover, since the distribution of X is invariant by a) changing the sign of a column and b) by columns' permutation then, the probability that a $k-$sparse vector satisfies the noiseless recovery condition is given by

$$\mathbb{P}_X(\|X'(X_I')^+ 1_k\|_\infty \leq 1 \text{ and } 1_k \in \mathrm{row}(X_I)), \text{ where } I = [k] \text{ and } 1_k = (1,\dots,1)' \in \mathbb{R}^k.$$

Moreover, the accessibility condition is satisfied with probability

$$\mathbb{P}_X(\min\{\|\gamma\|_1 : X\gamma = X_I 1_k\} = k).$$

Figure 7 provides these probabilities as functions of k: the number of nonzero components.

Figure 8 illustrates sign recovery properties by LASSO and thresholded LASSO for a particular observation of $X \in \mathbb{R}^{100\times 150}$, a particular observation of $Y \in \mathbb{R}^{100}$ and when $\beta \in \mathbb{R}^{150}$ is a $k-$sparse parameter with $k \in \{4, 30\}$, $\beta_1 = \dots = \beta_{k/2} = 20$ and $\beta_{k/2+1} = \dots = \beta_k = -20$. For the LASSO estimator, we consider the following setting:

- LASSO with a large tuning parameter $\lambda = 2\sqrt{2\log(150)}$ (as suggested by Candès and Plan (2009)).
- LASSO with a small tuning parameter; the one provided by SURE formula, which for a given X and Y minimizes the function $\lambda > 0 \mapsto \frac{1}{2}\|Y - X\hat{\beta}^{\mathrm{LASSO}}(\lambda)\|_2^2 + |\{i \in [p] : \hat{\beta}_i^{\mathrm{LASSO}}(\lambda) \neq 0\}|$ (see *e.g.* Tibshirani and Taylor (2012) or Vaiter et al. (2017)).

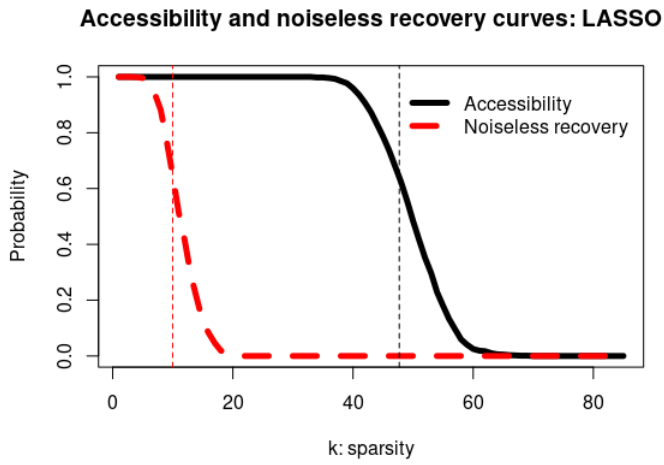

Figure 7: These curves provide the probability of the noiseless recovery condition and the probability of the accessibility condition as functions of the support size k. The value $k = 50/\log(150) = 9.9$ (Wainwright, 2009) is, approximately, the preimage of 0.5 for the noiseless recovery curve and $k = 100\rho_{DT}(2/3) = 47.8$, where ρ_{DT} is the phase transition curve (Donoho and Tanner, 2009b,a), is approximately the preimage of 0.5 for the accessibility curve.

Numerical experiments when the penalty term is the supremum norm

The noiseless recovery condition and the accessibility condition depend on β through $\text{sign}^{\infty}(\beta) \in \{-1, *, 1\}^p$. Moreover, since the distribution of X is invariant by a) changing the sign of a column and b) by columns' permutation then, the probability that a non-null vector having k non-maximal components in absolute value satisfies the noiseless recovery condition is given by

$$\mathbb{P}_X(\tilde{X}'(\tilde{X}')^+ e_1 = e_1) \text{ where } \tilde{X} = (\tilde{X}_1|X_I) \text{ with } \tilde{X}_1 = \sum_{i=1}^{p-k} X_i \text{ and } I = \{p-k+1, \ldots, p\}.$$

Note that an explicit formula for checking the noiseless recovery condition is given in supplementary material. Moreover, the accessibility condition is satisfied with probability

$$\mathbb{P}_X(\min\{\|\gamma\|_\infty : X\gamma = \tilde{X}_1\} = 1).$$

Figure 9 provides both the probability of the accessibility condition and the probability of the noiseless recovery condition as functions of k: the number of non-maximal components in absolute value.

In my opinion the values of β, λ, X and ε, that have been used in Figure 9, should be added.

In Figure 10 we illustrate the pattern recovery properties by a penalized least squares estimator and a thresholded penalized least squares estimator where the penalty term is the supremum norm. Specifically, $\beta \in \mathbb{R}^{150}$ satisfies $\beta_1 = \cdots = \beta_{60} = 20$, $\beta_{61} = \cdots = \beta_{120} = -20$ and $\beta_{121} = \cdots = \beta_{150} = 0$. The tuning parameter is selected as follows:

- The tuning parameter is given by the SURE formula which, for a given X and Y, minimizes the function $\lambda > 0 \mapsto \frac{1}{2}\|Y - X\hat{\beta}_\lambda\|_2^2 + \text{card}(\{i \in [p] : |\hat{\beta}_i| < \|\hat{\beta}_\lambda\|_\infty\})$ where $\hat{\beta}_\lambda$ is the penalized least squares estimator (see *e.g.* Minami (2020) or Vaiter et al. (2017)).

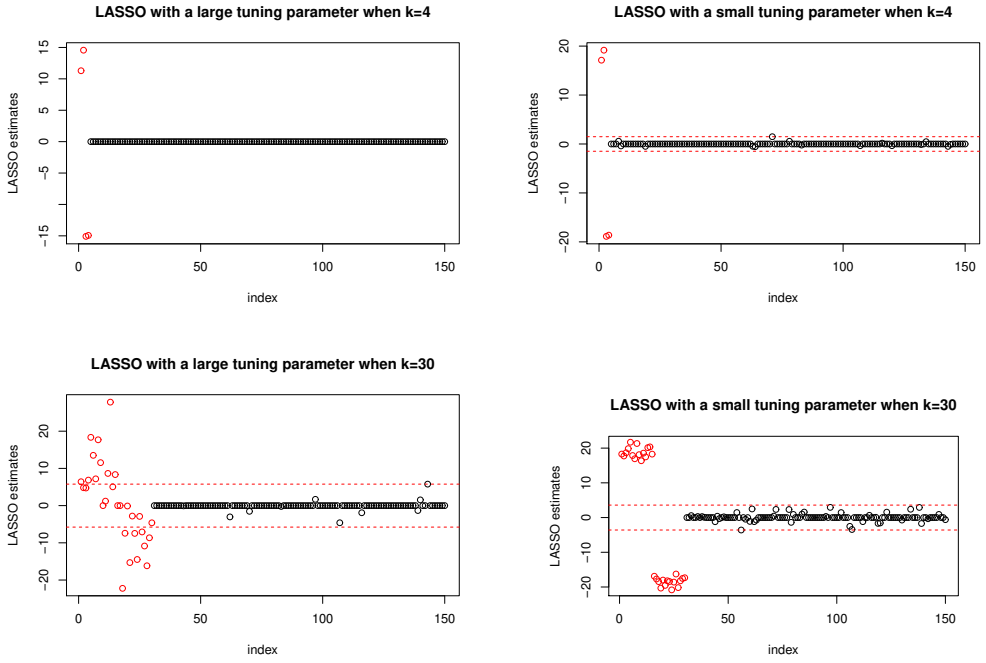

Figure 8: Illustrations of sign recovery by LASSO and thresholded LASSO. On the top, when $k = 4$, both the noiseless recovery condition and the accessibility condition hold. Thus, both LASSO and thresholded LASSO can recover the sign of β. With the large tuning parameter $\lambda = 2\sqrt{2\log(150)}$ the sign of β is recovered both by LASSO and thresholded LASSO (top left). When the tuning parameter is small (computed by SURE), some null components of β are not correctly estimated at 0 (black points which do not lie on the x-axis), but there exists a threshold, for which the thresholded LASSO recovers the sign of β (top right). On the bottom, when $k = 30$, the accessibility condition holds but the noiseless recovery condition does not hold, thus thresholded LASSO can recover the sign of β but LASSO cannot. When the tuning parameter is large: $\lambda = 2\sqrt{2\log(150)}$, both LASSO and thresholded LASSO fail to recover the sign of β (bottom left). When the tuning parameter is small, some null components of β are not correctly estimated at 0 but there exists a threshold, for which the thresholded LASSO recovers the sign of β (bottom right).

Acknowledgments

We would like to thank Samuel Vaiter, Mathurin Massias and Abderrahim Jourani for their insightful comments on the paper. Patrick Tardivel's institute has been supported by the EIPHI Graduate School (contract ANR-17-EURE-0002). The work of Tomasz Skalski has been supported by a French Government Scholarship.

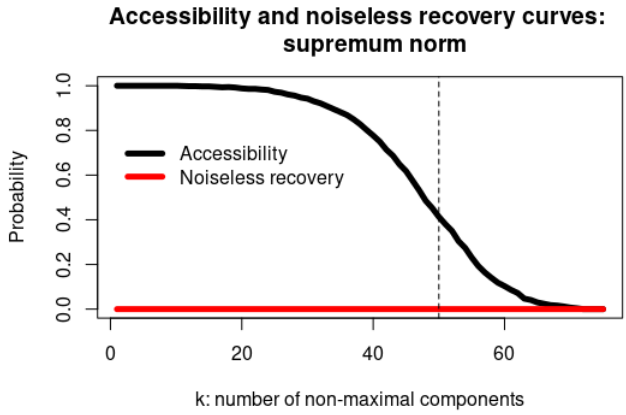

Figure 9: These curves provide the probability of the noiseless recovery and the probability of the accessibility condition as functions of the number of non-maximal components in absolute value k. The value $k = 50$ (Amelunxen et al., 2014) provides, approximately, the preimage of 0.5 for the accessibility curve.

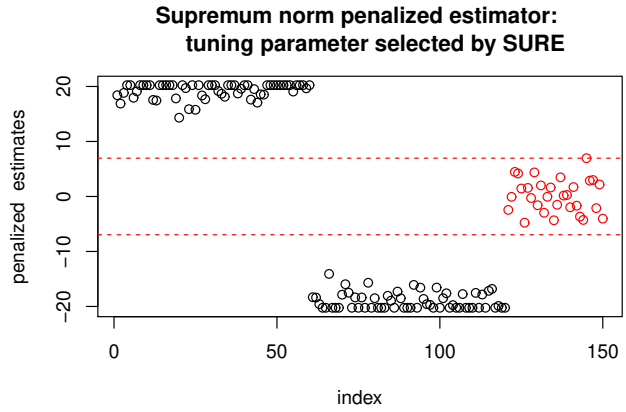

Figure 10: Illustrations of pattern recovery by a penalized estimator and a thresholded penalized estimator where the penalty term is the supremum norm. When $k = 30$, the accessibility condition holds, but the noiseless recovery condition does not hold. Thus, as illustrated on this picture, a thresholded penalized estimator can recover the pattern of β but a penalized estimator cannot.

A Appendix – Existence of the minimizer

We show that the optimization problem of interest in this article always has a minimizer.

Proposition 3. *Let* $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\mathrm{pen}(x) = \max\{u_1'x, \ldots, u_l'x\}$ *where* $u_1, \ldots, u_l \in \mathbb{R}^p$ *with* $u_1 = 0$. *For*

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

the optimization problem $\min_{b\in\mathbb{R}^p} f(b)$ *has at least one minimizer.*

For the remainder of this section, without loss of generality, we set $\lambda = 1$ since otherwise, this parameter can be absorbed into the penalty function. The proof of Proposition 3 relies on the following

two lemmas.

Lemma 1. *Let the assumptions of Proposition 3 hold and let $(\beta_m)_{m\in\mathbb{N}}$ be a minimizing sequence of f:*

$$\lim_{m\to\infty} f(\beta_m) = \inf_{b\in\mathbb{R}^p} f(b).$$

Then also $(X\beta_m)_{m\in\mathbb{N}}$ and $(\mathrm{pen}(\beta_m))_{m\in\mathbb{N}}$ converge. Moreover, these limits do not depend on the minimizing sequence.

Proof. The sequence $(X\beta_m)_{n\in\mathbb{N}}$ is bounded. Otherwise, $\|y-X\beta_m\|_2^2$ would be unbounded also, contradicting $\inf\{f(b) : b \in \mathbb{R}^p\} \leq f(0) < \infty$. Let $\tilde{\beta}_m$ be another minimizing sequence. Note that also $X\tilde{\beta}_m$ is bounded. Now extract arbitrary converging subsequences $(X\beta_{n_m})_{m\in\mathbb{N}}$ and $(X\tilde{\beta}_{\tilde{n}_m})_{m\in\mathbb{N}}$ with limits l and $\tilde{l}$, respectively. Note that $(\beta_{n_m})_{m\in\mathbb{N}}$ and $(\tilde{\beta}_{\tilde{n}_m})_{m\in\mathbb{N}}$ are still minimizing sequences so that also $\mathrm{pen}(\beta_{n_m})$ and $\mathrm{pen}(\tilde{\beta}_{\tilde{n}_m})$ must converge. We now show that $l = \tilde{l}$. If $l \neq \tilde{l}$, set $\bar{\beta}_m = (\beta_{n_m} + \tilde{\beta}_{\tilde{n}_m})/2$. By the above considerations, $(f(\bar{\beta}_m))_{m\in\mathbb{N}}$ is convergent. Since the function $z \in \mathbb{R}^n \mapsto \|y-z\|_2^2$ is strictly convex and pen is convex, we may deduce that

$$\begin{aligned}
\limsup_{m\to\infty} f(\bar{\beta}_m) &\leq \frac{1}{2}\|y-(l+\tilde{l})/2\|_2^2 + \limsup_{m\to\infty} \mathrm{pen}(\bar{\beta}_m) \\
&< \frac{1}{2}\left(\|y-l\|_2^2/2 + \|y-\tilde{l}\|_2^2/2\right) + \lim_{m\to\infty} \mathrm{pen}(\beta_{n_m})/2 + \lim_{m\to\infty} \mathrm{pen}(\tilde{\beta}_{\tilde{n}_m})/2 \\
&= \frac{1}{2}\lim_{m\to\infty} f(\beta_{n_m}) + \frac{1}{2}\lim_{m\to\infty} f(\tilde{\beta}_{\tilde{n}_m}) = \inf_{b\in\mathbb{R}^p} f(b),
\end{aligned}$$

yielding a contradiction. Since the selection of convergent subsequences was arbitrary, this implies that $(X\beta_m)_{m\in\mathbb{N}}$ and $(X\tilde{\beta}_m)_{m\in\mathbb{N}}$ share a unique limit point and that the sequences $(\mathrm{pen}(\beta_m))_{m\in\mathbb{N}}$ and $(\mathrm{pen}(\tilde{\beta}_m))_{m\in\mathbb{N}}$ converges as well. □

We remark that Lemma 1 also holds for any non-negative, convex function in place of the polyhedral gauge pen.

Lemma 2. *Let the assumptions of Proposition 3 hold and let $\gamma \geq 0$. The optimization problem*

$$\min_{b\in\mathbb{R}^p} \|y-Xb\|_2^2 \quad \text{subject to} \quad \mathrm{pen}(b) \leq \gamma \tag{5}$$

has at least one minimizer.

Proof. Let $P_\gamma = \{b \in \mathbb{R}^p : \mathrm{pen}(b) \leq \gamma\}$ be the closed and convex feasible region of (5). We set $z = Xb$ and note that the linearly transformed set XP_γ is still closed and convex. Therefore, the minimization problem

$$\min \ \|y-z\|_2^2 \quad \text{subject to} \quad z \in XP_\gamma$$

has a unique solution $\hat{z} \in XP_\gamma$, namely, the projection of y onto XP_γ. Consequently, $\hat{z} = X\hat{b}$ for some $\hat{b} \in P_\gamma$, where $\hat{b}$ is not necessarily unique. Finally, $\hat{b}$ clearly is a solution of the optimization problem (5). □

Before we turn to the proof of Proposition 3, we make the following observations. Note that we can decompose the polyhedron $P_\gamma = \{b \in \mathbb{R}^p : \mathrm{pen}(b) \leq \gamma\} = \{b \in \mathbb{R}^p : u_1'b, \ldots, u_l'b \leq \gamma\}$, where $\gamma \geq 0$, into the sum of a polyhedral cone (the so-called recession cone of P_γ) and a polytope, (see, e.g., Ziegler, 2012, Theorem 1.2 and Proposition 1.12). For $\gamma = 1$, we can therefore write

$$P_1 = \{b \in \mathbb{R}^p : u_1'b \leq 0, \ldots, u_l'b \leq 0\} + E,$$

where E is a polytope and therefore bounded. For arbitrary $\gamma \geq 0$, we then write

$$P_\gamma = P_0 + \gamma E. \tag{6}$$

Proof of Proposition 3. Let $(\beta_m)_{m\in N}$ be a minimizing sequence of f. By Lemma 1, both sequences $(X\beta_m)_{m\in\mathbb{N}}$ and $(\text{pen}(\beta_m))_{m\in\mathbb{N}}$ converge to, say, l and γ, respectively. This implies that

$$\frac{1}{2}\|y-l\|_2^2+\gamma=\inf_{b\in\mathbb{R}^p} f(b).$$

Let $\hat{\beta}$ be an arbitrary solution of (5). We prove that $f(\hat{\beta})=\|y-l\|_2^2+\gamma$. For this, we distinguish the following two cases.

1) Assume that $\gamma>0$. For n large enough so that $\text{pen}(\beta_m)>0$, we set u_n as

$$u_m=\frac{\gamma}{\text{pen}(\beta_m)}\beta_m.$$

Clearly, $\text{pen}(u_m)=\gamma$ so that $u_m\in P_\gamma$. Consequently, by definition of $\hat{\beta}$, we have $\|y-X\hat{\beta}\|_2^2\leq\|y-Xu_m\|_2^2$ and $\text{pen}(\hat{\beta})\leq\gamma$, so that

$$f(\hat{\beta})=\frac{1}{2}\left\|y-X\hat{\beta}\right\|_2^2+\text{pen}(\hat{\beta})\leq\frac{1}{2}\|y-Xu_m\|_2^2+\gamma\longrightarrow\frac{1}{2}\|y-l\|_2^2+\gamma$$

as $m\to\infty$, implying $f(\hat{\beta})=\inf\{f(b):b\in\mathbb{R}^p\}$.

2) Assume that $\gamma=0$. Using (6), we can write $\beta_m=u_m+\text{pen}(\beta_m)v_m$ with $u_m\in P_0$ and $v_m\in E$, where E is bounded. Since $X\beta_m\to l$ and $\text{pen}(\beta_m)v_n\to 0$ one may deduce that also $Xu_m\to l$, yielding

$$f(\hat{\beta})=\frac{1}{2}\left\|y-X\hat{\beta}\right\|_2^2\leq\frac{1}{2}\|y-Xu_m\|_2^2\longrightarrow\frac{1}{2}\|y-l\|_2^2$$

as $m\to\infty$ implying again that $f(\hat{\beta})=\inf\{f(b):b\in\mathbb{R}^p\}$ which completes the proof. □

B Appendix – Facts about polytopes and polyhedral gauges

We recall some basic definitions and facts about polytopes which we will use throughout the proofs. The following can be found in textbooks such as 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 $\{v_1,\ldots,v_k\}\subseteq\mathbb{R}^p$, namely,

$$P=\text{conv}\{v_1,\ldots,v_k\}.$$

The *dimension* $\dim(P)$ of a polytope is defined as the dimension of $\text{aff}(P)$, the affine subspace spanned by P. An inequality $a'x\leq c$ is called a *valid inequality* of P if $P\subseteq\{x\in\mathbb{R}^p:a'x\leq c\}$. A *face* F of P is any subset $F\subseteq P$ that satisfies

$$F=\{x\in P:a'x=c\}\text{ for some }a\in\mathbb{R}^p\text{ and }c\in\mathbb{R}\text{ with }P\subseteq\{x\in\mathbb{R}^p:a'x\leq c\}.$$

Note that $F=\varnothing$ and $F=P$ are faces of P and that any face F is again a polytope. A non-empty face F with $F\neq P$ is called *proper*. A point $x_0\in P$ lies in $\text{ri}(P)$, the *relative interior* of P, if x_0 is not contained in a proper face of P. We state two useful properties of faces in the following lemma.

Lemma 3. *Let $P\subseteq\mathbb{R}^p$ be a polytope given by $P=\text{conv}\{v_1,\ldots,v_k\}$, where $v_1,\ldots,v_k\in\mathbb{R}^p$. The following properties hold.*

i) If F and $\tilde{F}$ are faces of P, then so is $F\cap\tilde{F}$.

ii) Let L be an affine line contained in the affine span of P. If $L\cap\text{ri}(P)\neq\varnothing$, then L intersects a proper face of P.

Lemma 4 characterizes the connection between a certain class of convex functions (which encompasses polyhedral gauges) and the faces of a related polytope. The lemma is needed to prove Theorem 3.

Lemma 4. *Let $v_1, \ldots, v_k \in \mathbb{R}^p$, P be the polytope $P = \text{conv}\{v_1, \ldots, v_k\}$ and ϕ be the convex function defined by*

$$\phi(x) = \max\{v_1'x, \ldots, v_k'x\} \text{ for } x \in \mathbb{R}^p.$$

Then the subdifferential of ϕ at x is a face of P and is given by

$$\partial_\phi(x) = \text{conv}\{v_l : l \in I_\phi(x)\} = \{s \in P : s'x = \phi(x)\}, \text{ where } I_\phi(x) = \{l \in [k] : v_l'x = \phi(x)\}.$$

Conversely, let F be a non-empty face of P. Then $F = \partial_\phi(x)$ for some $x \in \mathbb{R}^p$.

Proof. The fact that $\partial_\phi(x) = \text{conv}\{u_l : l \in I_\phi(x)\}$ can be found in Hiriart-Urruty and Lemarechal (2001, p. 183). To prove the second equality, we consider the following. If $l \in I_\phi(x)$, by definition of $I_\phi(x)$, $v_l'x = \phi(x)$ and thus $v_l \in \{s \in P : s'x = \phi(x)\}$. Since $\{s \in P : s'x = \phi(x)\}$ is a convex set, one may deduce that

$$\text{conv}\{v_l : l \in I_\phi(x)\} \subseteq \{s \in P : s'x = \phi(x)\}.$$

Conversely, assume $s \in P$ is such that $s \notin \text{conv}\{v_l : l \in I_\phi(x)\}$. We then have $s = \sum_{l=1}^k \alpha_l v_l$ where $\alpha_1, \ldots, \alpha_k \geq 0$, $\sum_{l=1}^k \alpha_l = 1$ and $\alpha_{l_0} > 0$ for some $l_0 \notin I_\phi(x)$. Since $v_l'x \leq \phi(x)$ for all $l \in [k]$ and $u_{l_0}'x < \phi(x)$, we also get

$$s'x = \sum_{l=1}^k \alpha_l v_l'x < \phi(x).$$

Consequently, $s \notin \{s \in P : s'x = \phi(x)\}$ and thus

$$\{s \in P : s'x = \phi(x)\} \subseteq \text{conv}\{v_l : l \in I_\phi(x)\}.$$

Therefore, $\partial_\phi(x) = \text{conv}\{v_l : l \in I_\phi(x)\} = \{s \in P : s'x = \phi(x)\}$.

Now we show that the subdifferentials of ϕ are the (non-empty) faces of P. Let $x \in \mathbb{R}^p$. By definition of ϕ, $v_l'x \leq \phi(x)$ for every $l \in [k]$ so that the inequality $x's \leq \phi(x)$ is valid for all $s \in P$. This implies that $\partial_\phi(x)$ is a non-empty face of P. Conversely, let $F = \{s \in P : a's = c\}$ be a non-empty face of P where $a \in \mathbb{R}^p$, $c \in \mathbb{R}$ and $a's \leq c$ is a valid inequality for all $s \in P$. We prove that $F = \partial_\phi(a)$. For this, take any $s \in F$. We get $a's = c$ as well as $a's \leq \phi(a)$ as shown above, implying that $c \leq \phi(a)$. Analogously, for any $s \in \partial_\phi(a)$, $a's = \phi(a)$ as well as $a's \leq c$ since $\partial_\phi(a) \subseteq P$, yielding $\phi(a) \leq c$. Therefore we may deduce that $\phi(a) = c$ and thus $F = \partial_\phi(a)$. □

C Appendix – Proofs

We use the following notation in the appendix.

- Let S be a subset of $\mathbb{R}^p$ then
  - By $\text{conv}(S)$ we denote the convex hull of a subset S. In particular when $S = \{x_1, \ldots, x_l\}$ then
$$\text{conv}(S) = \left\{ \sum_{i=1}^l \alpha_i x_i : \alpha_1 \geq 0, \ldots, \alpha_l \geq 0, \sum_{i=1}^l \alpha_i = 1 \right\}.$$
  - By $\text{aff}(S)$ we denote the affine hull of a subset S. In particular when $S = \{x_1, \ldots, x_l\}$ then
$$\text{aff}(S) = \left\{ \sum_{i=1}^l \alpha_i x_i : \sum_{i=1}^l \alpha_i = 1 \right\}.$$
  - By $\overrightarrow{\text{aff}}(S)$ we denote the vector space parallel to $\text{aff}(S)$. In particular when $S = \{x_1, \ldots, x_l\}$ then
$$\overrightarrow{\text{aff}}(S) = \left\{ \sum_{i=1}^l \alpha_i x_i : \sum_{i=1}^l \alpha_i = 0 \right\}.$$

- By $\text{lin}(S)$ we denote the vector span of S.
- By $\text{ri}(S)$ we denote the relative interior of S.

- Given a matrix $A \in \mathbb{R}^{n\times p}$, $\text{col}(A)$ represents the vector space spanned by columns of A: $\text{col}(A) = \{Az : z \in \mathbb{R}^p\}$.
- The orthogonal complement of a vector space V is denoted $V^\perp$. In the particular case where $V = \text{lin}(v)$, for some $v \in \mathbb{R}^p$, then $v^\perp$ represents $\text{lin}(v)^\perp$, the hyperplane orthogonal to v.

C.1 Proof of Theorem 3 from Section 5

The following lemma, also needed to show Theorem 3, states that the fitted values are unique over all non-unique solutions of the penalized problem for a given y. It is a generalization of Lemma 1 in Tibshirani (2013), which shows this fact for the special case of the LASSO.

Lemma 5. *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, $\lambda > 0$ and* pen *be a polyhedral gauge. Then $X\hat{\beta} = X\tilde{\beta}$ and $\text{pen}(\hat{\beta}) = \text{pen}(\tilde{\beta})$ for all $\hat{\beta}, \tilde{\beta} \in S_{X,\text{pen}}(y)$.*

Proof. Assume that $X\hat{\beta} \neq X\tilde{\beta}$ for some $\hat{\beta}, \tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ 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,

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

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

Proof. ($\Longrightarrow$) Assume that there exists a face F of $B^* = \text{conv}\{u_1, \dots, u_k\}$ that intersects $\text{row}(X)$ and satisfies $\dim(F) < \text{def}(X)$. By Lemma 4, $F = \partial_{\text{pen}}(\hat{\beta})$ for some $\hat{\beta} \in \mathbb{R}^p$. Let $z \in \mathbb{R}^n$ with $X'z \in F$, which exists by assumption. Now let $y = X\hat{\beta} + \lambda z$. Note that $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ since

$$0 \in X'X\hat{\beta} - X'y + \lambda\partial_{\text{pen}}(\hat{\beta}) \iff \frac{1}{\lambda}X'(y - X\hat{\beta}) = X'z \in \partial_{\text{pen}}(\hat{\beta}).$$

We now construct $\tilde{\beta} \in S_{X,\lambda\text{pen}}(y)$ with $\tilde{\beta} \neq \hat{\beta}$. According to Lemma 4, $\partial_{\text{pen}}(\hat{\beta}) = \text{conv}\{u_l : l \in I\}$ where $I = I_{\text{pen}}(\hat{\beta}) = \{l \in [k] : u_l'\hat{\beta} = \text{pen}(\hat{\beta})\}$ and thus $u_l'\hat{\beta} < \text{pen}(\hat{\beta})$ whenever $l \notin I$. We now show that it is possible to pick $h \in \ker(X)$ with $h \neq 0$ but $u_l'h = 0$ for all $l \in I$. We then make h small enough such that $u_l'(\hat{\beta} + h) \leq \text{pen}(\hat{\beta})$ still holds for all $l \notin I$, which in turn implies that $\text{pen}(\hat{\beta} + h) = \max\{u_l'\hat{\beta} : l \in I\} = \text{pen}(\hat{\beta})$. This, together with $X\hat{\beta} = X(\hat{\beta} + h)$, yields $\hat{\beta} \neq \tilde{\beta} = \hat{\beta} + h \in S_{X,\lambda\text{pen}}(y)$ also. We now show that $\ker(X) \cap \text{col}(U)^\perp \neq \{0\}$, where $U = (u_l)_{l\in I} \in \mathbb{R}^{p\times|I|}$. For this, we distinguish two cases:

1) Assume that $0 \in \text{aff}\{u_l : l \in I\}$. Then $\text{aff}\{u_l : l \in I\} = \text{col}(U)$ and $\dim(F) = \text{rk}(U) < \text{def}(X)$. This implies that

$$\dim(\ker(X)) + \dim(\text{col}(U)^\perp) > p,$$

which proves what was claimed.

2) Assume that $0 \notin \text{aff}\{u_l : l \in I\}$. Note that this implies that $v = X'z \in \text{row}(X) \cap \text{conv}\{u_l : l \in I\}$ satisfies $X'z \neq 0$. We also have $\text{rk}(U) = \dim(\text{aff}\{u_l : l \in I\}) + 1 = \dim(F) + 1 \leq \text{def}(X)$ which implies that

$$\dim(\ker(X)) + \dim(\text{col}(U)^\perp) \geq p.$$

If $\ker(X) \cap \mathrm{col}(U)^\perp = \{0\}$, then $\mathbb{R}^p = \ker(X) \oplus \mathrm{col}(U)^\perp$. But we also have $\ker(X) \subseteq v^\perp$ as well as $\mathrm{col}(U)^\perp \subseteq v^\perp$, yielding a contradiction and proving the claim.

($\Longleftarrow$) Assume that there exists $y \in \mathbb{R}^n$ such that $\hat{\beta}, \tilde{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$ with $\hat{\beta} \neq \tilde{\beta}$. We then have

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

According to Lemma 5, $X\hat{\beta} = X\tilde{\beta}$, thus $\frac{1}{\lambda}X'(y - X\hat{\beta}) = \frac{1}{\lambda}X'(y - X\tilde{\beta})$. Consequently, one may deduce that $\mathrm{row}(X)$ intersects the face $\partial_{\mathrm{pen}}(\hat{\beta}) \cap \partial_{\mathrm{pen}}(\tilde{\beta})$. Let $F^* = \mathrm{conv}\{u_l : l \in I^*\}$ be a face of $\partial_{\mathrm{pen}}(\hat{\beta}) \cap \partial_{\mathrm{pen}}(\tilde{\beta})$ of smallest dimension among all faces of $\partial_{\mathrm{pen}}(\hat{\beta}) \cap \partial_{\mathrm{pen}}(\tilde{\beta})$ intersecting $\mathrm{row}(X)$. By minimality of $\dim(F^*)$, $\mathrm{row}(X)$ intersects the relative interior of F^*, namely, there exists $z \in \mathbb{R}^n$ such that $v = X'z$ lies in F^*, but not on a proper face of F^*. We will now show that if $\dim(F^*) \geq \mathrm{def}(X)$, then $\mathrm{row}(X)$ intersects a proper face of F^*, yielding a contradiction.

For this, first observe that $\dim(F^*) = \dim(\mathrm{aff}\{u_l : l \in I^*\})$ and that we can write the affine space $\mathrm{aff}\{u_l : l \in I^*\} = u_{l_0} + \mathrm{col}(\tilde{U}^*)$ where $l_0 \in I^*$ and $\tilde{U}^* = (u_l - u_{l_0})_{l \in I^* \setminus \{l_0\}} \in \mathbb{R}^{p \times |I^*|-1}$, implying that $\dim(F^*) = \mathrm{rk}(\tilde{U}^*)$.

Now let $h = \hat{\beta} - \tilde{\beta} \neq 0$. Clearly, $h \in \ker(X)$. Moreover, since $\mathrm{pen}(\hat{\beta}) = \mathrm{pen}(\tilde{\beta})$ by Lemma 5, and since $u_l \in \partial_{\mathrm{pen}}(\hat{\beta}) \cap \partial_{\mathrm{pen}}(\tilde{\beta})$ for all $l \in I^*$, by Lemma 4, we get

$$u_l'h = u_l'\hat{\beta} - u_l'\tilde{\beta} = \mathrm{pen}(\hat{\beta}) - \mathrm{pen}(\tilde{\beta}) = 0 \; \forall l \in I^*.$$

Therefore, $h \in \ker(X) \cap \mathrm{col}(U^*)^\perp$, where $U^* = (u_l)_{l \in I^*} \in \mathbb{R}^{p \times |I^*|}$. Assume that $\dim(F^*) \geq \mathrm{def}(X)$. Then

$$\dim(\mathrm{row}(X)) + \dim(\mathrm{col}(\tilde{U}^*)) \geq \mathrm{rk}(X) + \mathrm{def}(X) = p.$$

If $\mathrm{row}(X) \cap \mathrm{col}(\tilde{U}^*) = \{0\}$, then $\mathbb{R}^p = \mathrm{row}(X) \oplus \mathrm{col}(\tilde{U}^*)$. However, the last relationship cannot hold since $\mathrm{row}(X) = \ker(X)^\perp \subseteq h^\perp$ as well as $\mathrm{col}(\tilde{U}^*) \subseteq \mathrm{col}(U^*) \subseteq h^\perp$, where $h \neq 0$. Consequently, there exists $0 \neq \tilde{v} \in \mathrm{row}(X) \cap \mathrm{col}(\tilde{U}^*)$. The affine line $L = \{X'z + t\tilde{v} : t \in \mathbb{R}\} \subseteq \mathrm{row}(X)$ intersects the relative interior of F^* at $t = 0$ and clearly lies in $\mathrm{aff}(F^*) = u_{l_0} + \mathrm{col}(\tilde{U}^*)$, since $X'z \in F^*$ and $\tilde{v} \in \mathrm{col}(\tilde{U}^*)$. Therefore, L must intersect a proper face of F^* by Lemma 3. But then also $\mathrm{row}(X)$ intersects a proper face of F^*, which yields the required contradiction. □

C.2 Proofs for Section 3

Proof of Proposition 1

The following lemma can be seen as generalizing Proposition 4.1 from Gilbert (2017) from the ℓ_1-norm to all convex functions.

Lemma 6. *Let $\beta \in \mathbb{R}^p$ and ϕ be a convex function on $\mathbb{R}^p$. Then* $\mathrm{row}(X)$ *intersects $\partial_\phi(\beta)$ if and only if, for any $b \in \mathbb{R}^p$, the following implication holds*

$$X\beta = Xb \implies \phi(\beta) \leq \phi(b). \tag{7}$$

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

$$\iota_\beta(b) = \begin{cases} 0 & \text{when } Xb = X\beta \\ \infty & \text{else.} \end{cases}$$

Then (7) holds for any $b \in \mathbb{R}^p$ if and only if β is a minimizer of the function $b \mapsto \phi(b) + \iota_\beta(b)$. It is straightforward to show that $\partial_{\iota_\beta}(\beta) = \mathrm{row}(X)$. We can therefore deduce that the implication (7) holds for any $b \in \mathbb{R}^p$ if and only if

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

□

Proof of Proposition 1. By Lemma 6, the geometric characterization of accessible patterns is equivalent to the analytic one. We show the geometric characterization.

($\Longrightarrow$) When the pattern of β is accessible with respect to X and λpen, there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ such that $\hat{\beta} \overset{\text{pen}}{\sim} \beta$. Because $\hat{\beta}$ is a minimizer, $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\beta)$, so that, clearly, row$(X)$ intersects $\partial_{\text{pen}}(\beta)$.

($\Longleftarrow$) If row(X) intersects the face $\partial_{\text{pen}}(\beta)$, then there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\text{pen}}(\beta)$. For $y = X\beta + \lambda z$, we have $\frac{1}{\lambda}X'(y - X\beta) = X'z$, so that $\beta \in S_{X,\lambda\text{pen}}(y)$, and the pattern of β is accessible with respect to X and λpen. □

Proof of Theorem 1

Lemma 7. *Let $\phi : \mathbb{R}^p \to \mathbb{R}$ be the polyhedral gauge defined as*

$$\phi(x) = \max\{u_1'x, \ldots, u_k'x\} \text{ for some } u_1, \ldots, u_k \in \mathbb{R}^p$$

If $\partial_\phi(x) = \partial_\phi(v)$, we have $\partial_\phi(x) = \partial_\phi(\alpha x + (1-\alpha)v) = \partial_\phi(v)$ for all $\alpha \in [0,1]$.

Proof. Let $s \in \partial_\phi(x) = \partial_\phi(v)$. Since s is a subgradient at x and at v, the following two inequalities hold

$$\begin{aligned}\phi(\alpha x + (1-\alpha)v) &\geq \phi(x) - (1-\alpha)s'(x - v)\\ \phi(\alpha x + (1-\alpha)v) &\geq \phi(v) + \alpha s'(x - v).\end{aligned}$$

Multiplying the first inequality by α, the second by $(1-\alpha)$ and adding them, we get

$$\phi(\alpha x + (1-\alpha)v) \geq \alpha\phi(x) + (1-\alpha)\phi(v).$$

Using the convexity of ϕ, we arrive at

$$\phi(\alpha x + (1-\alpha)v) = \alpha\phi(x) + (1-\alpha)\phi(v).$$

By Lemma 4 we have $\partial_\phi(x) = \text{conv}\{u_l : l \in I\}$, where $I_\phi(x) = \{l \in [k] : u_l'x = \phi(x)\}$. Therefore, if $u_l \in \partial_\phi(x) = \partial_\phi(v)$, then $u_l'x = \phi(x)$ and $u_l'v = \phi(v)$, thus

$$u_l'(\alpha x + (1-\alpha)v) = \alpha\phi(x) + (1-\alpha)\phi(v) = \phi(\alpha x + (1-\alpha)v).$$

Consequently, $u_l \in \partial_\phi(\alpha x + (1-\alpha)v)$. On the other hand, if $u_l \notin \partial_\phi(x)$, then $u_l'x < \phi(x)$ and $u_l'v < \phi(v)$, thus

$$u_l'(\alpha x + (1-\alpha)v) < \alpha\phi(x) + (1-\alpha)\phi(v) = \phi(\alpha x + (1-\alpha)v).$$

Consequently, $u_l \notin \partial_\phi(\alpha x + (1-\alpha)v)$ and the claim follows. □

Lemma 8. *Let $X \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$. The following set is convex*

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

Note that V_β may be empty.

Proof. Assume that $V_\beta \neq \varnothing$. Let $y, \tilde{y} \in V_\beta$. Then there exist $\lambda > 0$ and $\tilde{\lambda} > 0$ such that $\hat{\beta} \in S_{X,\lambda\text{pen}}(y)$ and $\tilde{\beta} \in S_{X,\tilde{\lambda}\text{pen}}(\tilde{y})$ with $\partial_{\text{pen}}(\hat{\beta}) = \partial_{\text{pen}}(\tilde{\beta}) = \partial_{\text{pen}}(\beta)$. Consequently,

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

Let $\alpha \in (0,1)$ and $\check{y} = \alpha y + (1-\alpha)\tilde{y}$. Define $\check{\lambda} = \alpha\lambda + (1-\alpha)\tilde{\lambda}$ and $\check{\beta} = \alpha\hat{\beta} + (1-\alpha)\tilde{\beta}$. We show that $\check{y} \in V_\beta$. Indeed, observe that

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

By Lemma 7, $\partial_{\text{pen}}(\check{\beta}) = \partial_{\text{pen}}(\alpha\hat{\beta} + (1-\alpha)\tilde{\beta}) = \partial_{\text{pen}}(\beta)$, so that $\check{\beta} \in S_{X,\check{\lambda}\text{pen}}(\check{y})$ also, which proves the claim. □

Proof of Theorem 1. Assume that the noiseless recovery condition does not hold for β. Then $X\beta \notin V_\beta$, where V_β is defined as in Lemma 8. Consequently, by convexity of V_β, we have $X\beta + \varepsilon \notin V_\beta$ or $X\beta - \varepsilon \notin V_\beta$ for any realization of $\varepsilon \in \mathbb{R}^n$. Therefore

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

Consequently,

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

□

C.3 Proof for Section 4

Proof of Proposition 2

Proof of Proposition 2. We only need to prove the implication ($\Longleftarrow$), as the other direction is obvious. Assume that $\partial_{\mathrm{pen}}(\hat{\beta}) \subseteq \partial_{\mathrm{pen}}(\beta)$. Since $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$, we have $\frac{1}{\lambda}X'(y - X\hat{\beta}) \in \partial_{\mathrm{pen}}(\hat{\beta}) \subseteq \partial_{\mathrm{pen}}(\beta)$. Consequently, $\mathrm{row}(X)$ intersects $\partial_{\mathrm{pen}}(\beta)$ which implies that the pattern of β is accessible with respect to X and pen by Proposition 1. Consequently, there exists $y \in \mathbb{R}^n$ and there exists $\hat{\beta} \in S_{X,\lambda\mathrm{pen}}(y)$ for which $\hat{\beta} \overset{\mathrm{pen}}{\sim} \beta$. □

Proof of Theorem 2

Lemmas 9 and 10 are used to prove Theorem 2 which claims that, asymptotically, $\hat{\beta}(y^{(r)})/r$ converges to β when r tends to ∞.

Before stating these lemmas, note that for a non-empty closed and convex set $C \subseteq \mathbb{R}^p$ and $x \in C$, the asymptotic cone is defined as (cf. Hiriart-Urruty and Lemarechal, 2001)

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

Moreover, the following statements hold.

- The set C_∞ does not depend on the choice of $x \in C$.
- A non-empty closed and convex set C is compact if and only if $C_\infty = \{0\}$.

Lemma 9. *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$*,* $X \in \mathbb{R}^{n\times p}$*,* $v \in \mathrm{col}(X)$*. Let* $K_1 \geq 0$*,* $K_2 \geq 0$ *be large enough such that* $C = \{b \in \mathbb{R}^p : \mathrm{pen}(b) \leq K_1, \|Xb - v\|_2 \leq K_2\}$ *is non-empty. If* $\ker(X) \cap \ker(\mathrm{pen}) = \{0\}$ *then, the set* C *is compact.*

Proof. Clearly, C is closed and convex. If $\mathrm{pen}(d) > 0$ or if $Xd \neq 0$ then $d \notin C_\infty$. Consequently, $C_\infty \subset \ker(X) \cap \ker(\mathrm{pen}) = \{0\}$ and thus C is compact. □

Lemma 10. *Let* $X \in \mathbb{R}^{n\times p}$*,* $\lambda > 0$*,* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$ *and assume that uniform uniqueness holds for* (1)*. Let* $\beta \in \mathbb{R}^p$*,* $\varepsilon \in \mathbb{R}^n$ *and set* $y^{(r)} = X(r\beta) + \varepsilon$*. If* β *is accessible with respect to* X *and* pen*, then*

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

Proof. Since $\hat{\beta}(y^{(r)})$ is a minimizer of $S_{X,\lambda\mathrm{pen}}(y^{(r)})$, the following inequality holds

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

Since $y^{(r)} - X(r\beta) = \varepsilon$, one may deduce that

$$\begin{aligned} \lambda \text{pen}(\hat{\beta}(y^{(r)})) &\leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda \text{pen}(r\beta) \\ \implies \text{pen}(\hat{\beta}(y^{(r)})/r) &\leq \frac{\|\varepsilon\|_2^2}{2\lambda r} + \text{pen}(\beta) \\ \implies \limsup_{r\to\infty} \text{pen}(\hat{\beta}(y^{(r)})/r) &\leq \text{pen}(\beta). \end{aligned} \tag{8}$$

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

$$\begin{aligned} &\frac{1}{2}\|\varepsilon + X(r\beta) - X\hat{\beta}(y^{(r)})\|_2^2 + \lambda \text{pen}(\hat{\beta}(y^{(r)})) \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda \text{pen}(r\beta) \\ &\quad \implies -\|\varepsilon\|_2 \|X(r\beta) - X\hat{\beta}(y^{(r)})\|_2 + \frac{1}{2}\|X(r\beta) - X\hat{\beta}(y^{(r)})\|_2^2 \leq \lambda \text{pen}(r\beta) - \lambda \text{pen}(\hat{\beta}(y^{(r)})) \\ &\quad \implies -\|\varepsilon\|_2/r \left\|X\left(\hat{\beta}(y^{(r)})/r - \beta\right)\right\|_2 + \frac{1}{2}\left\|X\left(\hat{\beta}(y^{(r)})/r - \beta\right)\right\|_2^2 \leq \lambda \text{pen}(\beta)/r - \lambda/r \, \text{pen}\left(\hat{\beta}(y^{(r)})/r\right). \end{aligned} \tag{9}$$

Let $\alpha \in [0, \infty]$ be the limes superior of the sequence

$$\left(\left\|X\left(\hat{\beta}(y^{(r)})/r - \beta\right)\right\|_2\right)_{r\in\mathbb{N}}. \tag{10}$$

By (9) we get

$$\limsup_{r\to\infty} \frac{\lambda \text{pen}(\beta) - \lambda \text{pen}(\hat{\beta}(y^{(r)})/r)}{r} \;\geq\; \begin{cases} \alpha^2/2 & \text{if } \alpha < \infty \\ \infty & \text{if } \alpha = \infty. \end{cases}$$

Moreover, by (8) we get

$$\limsup_{r\to\infty} \frac{\lambda \text{pen}(\beta) - \lambda \text{pen}(\hat{\beta}(y^{(r)})/r)}{r} = 0$$

We can conclude that $\alpha = 0$ and that the sequence (10) converges to 0.

Due to uniform uniqueness, we have $\ker(\text{pen}) \cap \ker(X) = \{0\}$ and thus, by Lemma 9, the sequence $(\hat{\beta}(y^{(r)})/r)_{r\in\mathbb{N}}$ is bounded. Therefore, to prove that $\lim_{r\to\infty} \hat{\beta}(y^{(r)})/r = \beta$, it suffices to show that β is the unique accumulation point of this sequence. We extract a subsequence $(\hat{\beta}(y^{\phi(r)})/\phi(r))_{r\in\mathbb{N}}$ converging to $\gamma \in \mathbb{R}^p$ (where $\phi : \mathbb{N} \to \mathbb{N}$ is an increasing function). By (8), one may deduce that $\text{pen}(\gamma) \leq \text{pen}(\beta)$. Moreover, we get that

$$0 = \lim_{r\to\infty} \left\|X\left(\hat{\beta}(y^{(\phi(r))}/\phi(r) - \beta\right)\right\|_2^2 = \|X(\gamma - \beta)\|_2^2.$$

Finally, γ satisfies

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

and we show that the only $\gamma \in \mathbb{R}^p$ satisfying the above is $\gamma = \beta$. Because the pattern of β is accessible, 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 $\text{pen}(\gamma) \leq \text{pen}(\beta)$, one may deduce that $\gamma \in S_{X,\lambda\text{pen}}(y)$ also, contradicting the uniform uniqueness assumption. Consequently, $\gamma = \beta$ and

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

□

Finally, the proof of the sufficient condition in Theorem 2 is based on Lemma 10 and on Lemma 11 given below.

Lemma 11. *Let* pen *be a real-valued polyhedral gauge on* $\mathbb{R}^p$*. Then, there exists* $\tau > 0$ *depending on* β *such that*

$$\partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta) \text{ for all } b \in \overline{B}_\infty(\beta, \tau).$$

Proof. Let $I = \{l \in [k] : u_l'\beta = \text{pen}(\beta)\}$. By Lemma 4, $\partial_{\text{pen}}(\beta) = \text{conv}\{u_l\}_{l\in I}$. Since

$$u_l'\beta < \text{pen}(\beta)\ \forall l \notin I,$$

and since linear functions and the gauge pen are continuous, one may pick $\tau > 0$ small enough such that

$$u_l'b < \text{pen}(b)\ \forall l \notin I,\ \forall b \in \overline{B}_\infty(\beta, \tau).$$

Consequently, for any $b \in \overline{B}_\infty(\beta, \tau)$, we have $J = \{l \in [k] : u_l'b = \text{pen}(b)\} \subseteq I$ and thus

$$\partial_{\text{pen}}(b) = \text{conv}\{u_l\}_{l\in J} \subseteq \text{conv}\{u_l\}_{l\in I} = \partial_{\text{pen}}(\beta).$$

□

Proof of Theorem 2. By Lemma 11, there exists $\tau_0 > 0$ such that for any $b \in \overline{B}_\infty(\beta, \tau_0)$ we have $\partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta)$. By Lemma 10, $\hat{\beta}(y^{(r)})/r$ converges to β when r tends to ∞. Consequently, we have

$$\exists r_0 \in \mathbb{N} \text{ such that } \forall r \geq r_0, \|\hat{\beta}(y^{(r)})/r - \beta\|_\infty \leq \tau_0/2.$$

Consequently, for $r \geq r_0$ we have

$$\begin{aligned} &\forall b \in \overline{B}_\infty(\hat{\beta}(y^{(r)})/r, \tau_0/2),\ \partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta) \text{ and} \\ &\exists \tilde{b} \in \overline{B}_\infty(\hat{\beta}(y^{(r)})/r, \tau_0/2), \partial_{\text{pen}}(\tilde{b}) = \partial_{\text{pen}}(\beta). \end{aligned}$$

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

$$\begin{aligned} &\forall b \in \overline{B}_\infty(\hat{\beta}(y^{(r)}), r\tau_0/2), \partial_{\text{pen}}(b) \subseteq \partial_{\text{pen}}(\beta) \\ &\exists \tilde{b} \in \overline{B}_\infty(\hat{\beta}(y^{(r)}), r\tau_0/2), \partial_{\text{pen}}(\tilde{b}) = \partial_{\text{pen}}(\beta) \end{aligned}$$

Consequently, the claim follows by taking $\tau = r\tau_0/2$. □

C.4 Pattern equivalence classes are relative interior normal cones of B^*

Hereafter, we remind the definition of a normal cone (see *e.g.* Hiriart-Urruty and Lemarechal (2001) page 65). Let K be a closed convex set in $\mathbb{R}^p$ and $x \in K$. The normal cone of K at x is

$$N_K(x) := \{s \in \mathbb{R}^p : s'(z - x) \leq 0\ \forall z \in K\}.$$

The following fact is particularly relevant when K is a polytope. Let F be a face of K. If $x \in \text{ri}(F)$ and $z \in \text{ri}(F)$ then $N_K(x) = N_K(z)$ (see *e.g.* Ewald (1996) page 16).

We already know that $\partial_{\text{pen}}(x)$ is a face of the polytope B^* (see Lemma 4). According to the above property, the normal cone $N_{B^*}(s)$ does not depend on $s \in \text{ri}(\partial_{\text{pen}}(x))$ and we denote it $N_{B^*}(\partial_{\text{pen}}(x))$. The following notation represents the pattern equivalence class of an arbitrary $x \in \mathbb{R}^p$ for $\overset{\text{pen}}{\sim}$

$$C_x := \left\{w \in \mathbb{R}^p : w \overset{\text{pen}}{\sim} x\right\}.$$

Our objective is to prove Theorem 4.

Theorem 4. *Let $x \in \mathbb{R}^p$ then $C_x = \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x)))$.*

To prove the above theorem we are going to use some lemmas listed hereafter.

Lemma 12. *Let $x \in \mathbb{R}^p$ then $C_x \subset N_{B^*}(\partial_{\mathrm{pen}}(x))$*

Proof. Let $w \in C_x$ and $s \in \mathrm{ri}(\partial_{\mathrm{pen}}(x))$. Since $s \in \partial_{\mathrm{pen}}(x) = \partial_{\mathrm{pen}}(w)$ then, for $z \in B^*$, we have

$$w'(z-s) = \underbrace{w'z}_{\leq \mathrm{pen}(w)} - \underbrace{w's}_{=\mathrm{pen}(w)} \leq \mathrm{pen}(w) - \mathrm{pen}(w) = 0$$

Consequently, $w \in N_{B^*}(s) = N_{B^*}(\partial_{\mathrm{pen}}(x))$. □

Lemma 13. *Let $x \in \mathbb{R}^p$ then $lin(N_{B^*}(\partial_{\mathrm{pen}}(x))) \subseteq \overrightarrow{\mathrm{aff}}(\partial \mathrm{pen}(x))^{\perp}$.*

Proof. Let $v \in \mathrm{ri}(\partial_{\mathrm{pen}}(x))$, $s \in \mathrm{ri}(\partial_{\mathrm{pen}}(x))$ (then $N_{B^*}(s) = N_{B^*}(v) = N_{B^*}(\partial_{\mathrm{pen}}(x))$) and $z \in N_{B^*}(\partial_{\mathrm{pen}}(x))$. Since $z \in N_{B^*}(s)$ we have

$$z'(w-s) \leq 0 \quad \forall w \in B^*.$$

In particular $z'(v-s) \leq 0$. Moreover, since $z \in N_{B^*}(v)$, we have

$$z'(w-v) \leq 0 \quad \forall w \in B^*.$$

In particular $z'(s-v) \leq 0$. Therefore, $z'(s-v) = 0$. Finally, since $N_{B^*}(\partial_{\mathrm{pen}}(x))$ is perpendicular to the set $\{s-v : s, v \in \mathrm{ri}(\partial_{\mathrm{pen}}(x))\}$, one may deduce that

$$\mathrm{lin}(N_{B^*}(s)) \subseteq \mathrm{lin}(\{s-v : s, v \in \mathrm{ri}(\partial_{\mathrm{pen}}(x))\})^{\perp} \subseteq \overrightarrow{\mathrm{aff}}(\partial \mathrm{pen}(x))^{\perp}.$$

□

Lemma 14. *Let $x \in \mathbb{R}^p$ then $C_x \subset \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x)))$*

Proof. Let $w \in C_x$ and let $z \in B(w,\epsilon) \cap \mathrm{aff}(N_{B^*}(\partial_{\mathrm{pen}}(x)))$ where $\epsilon > 0$. Let us show that for $\epsilon > 0$ small enough $z \in C_x$. According to Lemma 11, for $\epsilon > 0$ small enough we have $z \in B(w,\epsilon)$ implies $\partial_{\mathrm{pen}}(z) \subseteq \partial_{\mathrm{pen}}(w) = \partial_{\mathrm{pen}}(x)$. Moreover, if $\partial_{\mathrm{pen}}(z) \subsetneq \partial_{\mathrm{pen}}(x)$ then pick $u \in \partial_{\mathrm{pen}}(z)$ and $v \in \partial_{\mathrm{pen}}(x) \setminus \partial_{\mathrm{pen}}(z)$. Since $u - v \in \overrightarrow{\mathrm{aff}}(\partial_{\mathrm{pen}}(x))$, since $x \in N_{B^*}(\partial_{\mathrm{pen}}(x))$ (Lemma 1) and $z \in \mathrm{aff}(N_{B^*}(\partial_{\mathrm{pen}}(x))) = \mathrm{lin}(N_{B^*}(\partial_{\mathrm{pen}}(x)))$ then, according to Lemma 2, we have $x'(u-v) = 0 = z'(u-v)$. Consequently, $u'z = \mathrm{pen}(z) = v'z$ and thus $v \in \partial_{\mathrm{pen}}(z)$ which leads to a contradiction. Therefore $z \in C_x$. □

Proof of Theorem 4. As proved in Lemma 4 any non-empty face of B^* can be written as $\partial_{\mathrm{pen}}(x)$ for some $x \in \mathbb{R}^p$. Consequently, one may chose a subset $E \subset \mathbb{R}^p$ for which $\phi : z \in E \mapsto \partial_{\mathrm{pen}}(z)$ is a bijection between E and non-empty faces of B^*. Note that E is finite (since a polytope has a finite number of faces) and let us set $E = \{x_1, \ldots, x_l\}$ then we have the following properties

- Because $\partial_{\mathrm{pen}}(x_i) \neq \partial_{\mathrm{pen}}(x_j)$ once $i \neq j$ then $C_{x_i} \cap C_{x_j} = \varnothing$. Moreover, let $x \in \mathbb{R}^p$ then $\partial_{\mathrm{pen}}(x)$ is a face of B^*. Therefore $\partial_{\mathrm{pen}}(x) = \partial_{\mathrm{pen}}(x_i)$ for some $i \in [l]$ thus $x \in C_{x_i}$. Consequently, $C_{x_1}, \ldots, C_{x_l}$ is a partition of $\mathbb{R}^p$.
- Relative interior normal cones of a polytope provides a partition (see *e.g.* Ewald page 17). Consequently, $\mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x_1))), \ldots, \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x_l)))$ is also a partition of $\mathbb{R}^p$.

Because $C_{x_i} \subset \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x_i)))$ partitions given above coincide. Consequently, for all $i \in [l]$, we have $C_{x_i} = \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x_i)))$. Finally, for all $x \in \mathbb{R}^p$ we have $x \in C_{x_i}$ for some $i \in [l]$ and thus

$$C_x = C_{x_i} = \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x_i))) = \mathrm{ri}(N_{B^*}(\partial_{\mathrm{pen}}(x))).$$

□

References

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

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3:224–294, 2014.

A. Barbara, J. Abderrahim, and S. Vaiter. Maximal solutions of sparse analysis regularization. *Journal of Optimization Theory and Applications*, 180:374–396, 2019.

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

M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, and M. Wilczyński. Pattern recovery by SLOPE. Technical Report 2203.12086, arXiv, 2022.

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.

E. Candès and Y. Plan. Near-ideal model selection by l_1 minimization. *Annals of Statistics*, 37: 2145–2177, 2009.

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*, 30:530–543, 2021.

P. Descloux, C. Boyer, J. Josse, A. Sportisse, and S. Sardy. Robust Lasso-zero for sparse corruption and model selection with missing covariates. *Scandinavian Journal of Statistics*, 49:1605–1635, 2022.

D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A*, 367:4273–4293, 2009a.

D. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22:1–53, 2009b.

X. Dupuis and P. J. C. Tardivel. Proximal operator for the sorted ℓ_1 norm: Application to testing procedures based on SLOPE. *Journal of Statistical Planning and Inference*, 221:1–8, 2022.

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

Günter Ewald. *Combinatorial Convexity and Algebraic Geometry*. Springer, 1996.

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

J.-J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51:3601–3608, 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. *Fundamentals of Convex Analysis*. Springer, Heidelberg, 2001.

J.-C. Hütter and P. Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146, 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, 2012.

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51:339–360, 2009.

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.

K. Minami. Degrees of freedom in submodular regularization: A computational perspective of Stein's unbiased risk estimate. *Journal of Multivariate Analysis*, 175:104546, 2020.

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.

R. Negrinho and A. Martins. Orbit regularization. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

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: 799–803, 2017.

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

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

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. *Journal of Machine Learning Research*, 23:1–36, 2022.

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

D.B. Sharma, H.D. Bondell, and H.H. Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Computational Statistics and Data Analysis*, 22:319–340, 2013.

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

T. Skalski, P. Graczyk, B. Kołodziejek, and M. Wilczyński. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal. Technical Report 2202.08573, arxiv, 2022.

A. Takahashi and S. Nomura. Efficient path algorithms for clustered Lasso and OSCAR. Technical Report 2006.08965, arxiv, 2020.

P. Tardivel and M. Bogdan. On the sign recovery by least absolute shrinkage and selection operator, thresholded least absolute shrinkage and selection operator and thresholded basis pursuit denoising. *Scandinavian Journal of Statistics*, page to appear, 2022.

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 and J. Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40:1198–1232, 2012.

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.

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

S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 69:791–832, 2017.

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 model-x knockoffs with ℓ_p-regularized statistics. Technical Report 2007.15346, arxiv, 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.M. 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.