The Geometry of Uniqueness, Sparsity and Clustering in Penalized Estimation

Ulrike Schneider, Patrick A Tardivel

The Geometry of Uniqueness, Sparsity and Clustering in Penalized Estimation

Ulrike Schneider [1] and Patrick J.C. Tardivel[2]

[1]TU Wien
[2]University of Wrocław

Abstract

We provide a necessary and sufficient condition for the uniqueness of penalized least-squares estimators whose penalty term is given by a norm with a polytope unit ball, covering a wide range of methods including SLOPE and LASSO, as well as the related method of basis pursuit. We consider a strong type of uniqueness that is relevant for statistical problems. The uniqueness condition is geometric and involves how the row span of the design matrix intersects the faces of the dual norm unit ball, which for SLOPE is given by the sign permutahedron. Further considerations based this condition also allow to derive results on sparsity and clustering features. In particular, we define the notion of a SLOPE model to describe both sparsity and clustering properties of this method and also provide a geometric characterization of accessible SLOPE models.

Keywords: SLOPE, basis pursuit, LASSO, uniqueness, sparsity, clustering, regularization, geometry, polytope.

MSC 2020: Primary 62-08; Secondary 52B12.

1 Introduction

The linear regression model $Y = X\beta + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ is a fixed matrix, $\beta \in \mathbb{R}^p$ is an unknown parameter vector, and ε is a centered random error term in $\mathbb{R}^n$, plays a central role in statistics. When $\ker(X) = \{0\}$, the ordinary least-squares estimator $\hat{\beta}^{\text{ols}} = (X'X)^{-1}X'Y$, which minimizes the residual sum of squares $\|Y - Xb\|_2^2$ with respect to $b \in \mathbb{R}^p$, is the usual estimator of β. In high dimensions, when $p > n$, and thus $\ker(X) \neq \{0\}$, the ordinary least squares estimator is no longer well-defined, as then the function $b \in \mathbb{R}^p \mapsto \|Y - Xb\|_2^2$ does not have a unique minimizer.

In this case, typically, a penalty term is added to the residual sum of squares to provide an alternative to ordinary least-squares estimation. In some cases, also the minimizer of the penalized least-squares optimization problem is not unique. Since Y is a random vector and the induced stochastic properties on the minimizer are often the object of study in a statistical framework, it is relevant to

consider a strong type of uniqueness: uniqueness for a given X that holds for all realizations[1] of Y in $\mathbb{R}^n$. In this paper, we provide a necessary and sufficient condition for uniqueness for a wide class of penalties based on a geometric criterion, as well as for the related methods of basis pursuit. Moreover, the geometry involved in this condition also yields results for model selection, i.e., sparsity and related clustering properties, which we investigate for SLOPE, LASSO, and basis pursuit.

1.1 Penalized least-squares estimators and uniqueness

The Ridge estimator, minimizing the function $b \in \mathbb{R}^p \mapsto \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_2^2$, where $\lambda > 0$ is a tuning parameter, was the first penalized estimator to appear in the statistics literature (Hoerl & Kennard, 1970; Golub et al., 1979). Due to the strict convexity of the function $b \mapsto \|b\|_2^2$, the minimizer is always unique and given by $\hat{\beta}^{\text{ridge}} = (X'X + \lambda\mathbb{I}_p)^{-1}X'Y$. This estimator is not sparse, meaning that it does not set components equal to zero almost surely. Especially when p is large, this can make the estimator more difficult to interpret compared to other methods such as LASSO or SLOPE, which do exhibit sparsity and are described in the following.

The Least Absolute Shrinkage and Selection Operator or LASSO (Chen & Donoho, 1994; Alliney & Ruzinsky, 1994; Tibshirani, 1996) is the l_1-penalized least-squares estimator defined as

$$\hat{\beta}^{\text{lasso}} = \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_1, \text{ where } \lambda > 0.$$

When $\ker(X) = \{0\}$, the function $b \in \mathbb{R}^p \mapsto \|Y - Xb\|_2^2$ is strictly convex, immediately implying the uniqueness of the LASSO minimizer. In high dimensions, $\ker(X) \neq \{0\}$ and the function $b \in \mathbb{R}^p \mapsto \|Y - Xb\|_2^2$ is not strictly convex, thus uniqueness of $\hat{\beta}^{\text{lasso}}$ is not guaranteed. A geometric description of the set of LASSO minimizers, particularly relevant when non-uniqueness occurs, is given in Dupuis & Vaiter (2019). A sufficient condition for uniqueness of the estimator for all $Y \in \mathbb{R}^n$ is for the columns of the design matrix X to be in general position. This was first outlined by Rosset et al. (2004) and later investigated by Tibshirani (2013) and Ali & Tibshirani (2019). Recently, this condition was relaxed by Ewald & Schneider (2020) to a geometric criterion that is both sufficient and necessary and which is generalized for a wide class of possible penalty terms in the present paper.

A strongly related procedure is basis pursuit, which first appeared in compressed sensing (Chen & Donoho, 1994) and is defined as

$$\hat{\beta}^{\text{bp}} = \arg\min \|b\|_1 \text{ subject to } Y = Xb,$$

provided that $Y \in \text{col}(X)$. In the noiseless case, this method allows to recover a sparse vector β (see e.g. Candès et al., 2006; Cohen et al., 2009). In the noisy case, when ε is no longer zero, the basis pursuit estimator can be viewed as the LASSO when the tuning parameter $\lambda > 0$ becomes infinitely small. Naturally, basis pursuit shares lot of properties with the LASSO estimator. For example, general position of the columns of the design matrix X is also a sufficient condition for uniqueness of

[1] Certain results in the literature (Zhang et al., 2015; Gilbert, 2017; Mousavi & Shen, 2019) provide a criterion for the uniqueness of a given minimizer. These results naturally differ strongly from the ones in the present article as they deal with a weaker notion of uniqueness.

$\hat{\beta}^{\text{bp}}$ for all $Y \in \mathbb{R}^n$ (see e.g. Dossal, 2012)[2]. However, to the best of our knowledge, a necessary and sufficient condition for this type of uniqueness has previously been unknown.

Our results also cover Sorted L-One Penalized Estimation or SLOPE (Zeng & Figueiredo, 2014; Bogdan et al., 2015), which is the penalized estimator given by

$$\hat{\beta}^{\text{slope}} = \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \frac{1}{2} \|Y - Xb\|_2^2 + \sum_{j=1}^{p} w_j |b|_{(j)},$$

where $w_1 > 0$, $w_1 \geq \cdots \geq w_p \geq 0$, and $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$. Note that the penalty term gives rise to the so-called SLOPE norm. A special case of this estimator, the Octagonal Shrinkage and Clustering Algorithm for Regression or OSCAR, has already been introduced in Bondell & Reich (2008). The SLOPE estimator is well-defined once the corresponding minimizer is unique and, similarly to the LASSO, uniqueness is obvious when $\ker(X) = \{0\}$. However, in contrast to the LASSO, no condition guaranteeing uniqueness has previously been established.

1.2 Uniqueness and polytope unit balls

In this paper, we study the problem of uniqueness of penalized estimators in a general setting, where the penalty term is not restricted the l_1- or the SLOPE norm. We describe the framework we consider in the following. Let $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, and $\|.\|$ be a norm on $\mathbb{R}^p$. Consider the solution set $S_{X,\|.\|}(y)$ to the penalized least-squares problem

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

Note that $S_{X,\|.\|}(y)$ is non-empty since the function $b \in \mathbb{R}^p \mapsto \frac{1}{2} \|y - Xb\|_2^2 + \|b\|$ is continuous and unbounded when $\|b\|$ becomes large. The penalty term may include a positive tuning parameter which can be viewed as part of the norm, for instance $\|.\| = \lambda \|.\|_1$ for the LASSO estimator. When $\|.\|$ is a norm for which $\|b + \tilde{b}\| = \|b\| + \|\tilde{b}\|$ holds if and only if $b = t\tilde{b}$ where $t \geq 0$[3], such as the l_2-norm, then $S_{X,\|.\|}(y)$ is a singleton for all $y \in \mathbb{R}^n$ and for all $X \in \mathbb{R}^{n \times p}$. This statement is a straightforward consequence of the following facts. When $\hat{\beta}, \tilde{\beta} \in S_{X,\|.\|}(y)$ we have

i) $X\hat{\beta} = X\tilde{\beta}$ (see Lemma 2 in the appendix).

ii) Since $(\hat{\beta} + \tilde{\beta})/2 \in S_{X,\|.\|}(y)$ also, $\|(\hat{\beta} + \tilde{\beta})/2\| = \|\hat{\beta}\| = \|\tilde{\beta}\| = (\|\hat{\beta}\| + \|\tilde{\beta}\|)/2$ follows.

Geometrically, such a norm $\|.\|$ possesses a unit ball $\{x \in \mathbb{R}^p : \|x\| = 1\}$ with no edges. Subsequently, the problem of uniqueness is only relevant when the unit ball of the norm under consideration contains an edge. More concretely, we restrict our attention to norms for which the unit ball $B = \{x \in \mathbb{R}^p : \|x\| \leq 1\}$ is given by a polytope. Note that this is the case for the l_1-norm, the l_∞-norm, and the SLOPE norm. Our results also cover the fused Lasso (Tibshirani et al., 2005), the clustered

[2]This reference focuses on necessary and sufficient conditions to uniquely recover a given b_0 from $y = Xb_0$ (in our notation), which is a different type of uniqueness than we consider.

[3]Typically, b and $\tilde{b}$ are not orthogonal, thus the equality in the triangular inequality does not coincide with the decomposability property described in Negahban et al. (2012).

Lasso (She, 2010), or methods with a mixed l_1,l_∞-norm penalty term (Negahban & Wainwright, 2008; Bach et al., 2012).

1.3 Sparsity and clustering: accessible models and sign estimation

As mentioned above, the LASSO estimator is a sparse method that generally sets components equal to zero with positive probability, entailing that the estimator also performs so-called model selection. In fact, when $p > n$ and the solution is unique, $\hat{\beta}^{\text{lasso}}$ contains at least $p - n$ zero components. Instigated by this sparsity property, an abundant literature has arisen to deal with the recovery of the location of the non-null components of β, or, more specifically, the recovery of the sign vector of β (Zou, 2006; Zhao & Yu, 2006; Wainwright, 2009).

A necessary condition for the recovery of $\text{sign}(\beta)$ is for this vector to be accessible by the LASSO, i.e. for a fixed $\lambda > 0$, there has to exist $Y \in \mathbb{R}^n$ for which $\text{sign}(\hat{\beta}^{\text{lasso}}) = \text{sign}(\beta)$. Otherwise, $\mathbb{P}(\text{sign}(\hat{\beta}^{\text{lasso}}) = \text{sign}(\beta)) = 0$, and recovery is clearly impossible. A geometrical characterization of accessible sign vectors is given in Sepehri & Harris (2017) under the assumption of uniqueness of LASSO solutions.

Also basis pursuit is also sometimes used for sign recovery of β (see e.g. Saligrama & Zhao, 2011; Tardivel & Bogdan, 2018; Descloux & Sardy, 2018; Descloux et al., 2020), however, the notion of accessible sign vectors has not been extended to this method before. In this article, we provide a geometric criterion for accessibility for both LASSO and basis pursuit from a different viewpoint and without the assumption of uniqueness. The geometry of our characterization is closely related to the geometrical considerations for the uniqueness of these estimators.

Finally, the SLOPE estimator is also a sparse method which additionally exhibits a clustering phenomenon, as some components may be equal in absolute value with positive probability. This property can be deduced from the explicit expressions one obtains in case the columns of X are orthogonal (Tardivel et al., 2020) and also holds in the general case. In certain applications, this clustering feature (which is not shared by the LASSO) may be of particular relevance (Kremer et al., 2019, 2020). We show how our geometric approach can be used to provide a characterization of the the clusters induced by SLOPE.

1.4 Related geometrical works

Most articles providing geometric properties in the context of penalized estimation treat the LASSO. Tibshirani & Taylor (2012) show that the LASSO residual $Y - X\hat{\beta}^{\text{lasso}}$ is the projection of Y onto the so-called LASSO null polytope $\{z \in \mathbb{R}^n : \|X'z\|_\infty \leq \lambda\}$. From this result, the authors derive an explicit formula for the Stein's unbiased risk estimate that provides an unbiased estimator for $\mathbb{E}(\|X\hat{\beta}^{\text{lasso}} - X\beta\|_2^2)$. This geometric result also lays the groundwork for selective inference (Lee et al., 2016), for deriving screening procedures (Ghaoui et al., 2012; Wang et al., 2013), and to describe the accessible LASSO models in Sepehri & Harris (2017).

For basis pursuit, geometrical considerations focus on dealing with the l_1-recovery in the noiseless case and are aimed at deriving the phase-transition curve (Donoho & Tanner, 2009).

The very recent article of Minami (2020) generalizes some results of Tibshirani & Taylor (2012) to SLOPE and shows that the number of non-null clusters (the quantity $\|\text{mdl}(\hat{\beta}^{\text{slope}})\|_\infty$ in our article) appears in the Stein's unbiased risk estimate for SLOPE estimator.

For the sake of completeness we mention that in the present paper, we provide a convex null set in Proposition 3 that generalizes the concept of the LASSO null polytope to all norm-penalized least-squares estimators, where the projection of Y onto this set yields the estimation residuals.

1.5 Notation and structure

To conclude this section, we introduce the notation used throughout this article.

We denote the set $\{1, \dots, k\}$ by $[k]$ and use $|I|$ for the cardinality of a set I. The set $\mathcal{S}_p$ contains all permutations on the set $[p]$. For a matrix A, the symbols $\text{col}(A)$ and $\text{row}(A)$ stand for the column and row space of A, respectively, whereas $\text{conv}(A)$ represents the convex hull of the columns of A. As used in previous sections already, for a number t, $\text{sign}(t)$ is given by $1, -1$, or 0 if $t > 0, t < 0$, or $t = 0$, respectively. For a vector x, $\text{sign}(x)$ is the vector containing the signs of the components of x. Finally, the symbols $\|.\|_1$, $\|.\|_2$, $\|.\|_\infty$, and $\|.\|_w$ represent the l_1-, l_2-, supremum, and the SLOPE-norm, respectively.

The remainder of this article is organized as follows. Section 2 contains the main theorem of uniqueness for penalized least-squares estimators, as well as the analogous necessary and sufficient uniqueness condition for basis pursuit. In Section 3, we investigate the model selection properties related to the geometric condition introduced in Section 2 for LASSO, BP, and SLOPE estimators, including a characterization of the SLOPE's clustering property. This section also contains a general result on the convex null set for norm-penalized least-squares estimation. All proofs are relegated to the appendix, which also contains a remainder of basic facts of subdifferentials and polytopes.

2 A necessary and sufficient condition for uniqueness of penalized problems

We start by providing the framework for the theorem on uniqueness of penalized least-squares minimization problems. For a norm $\|.\|$ on $\mathbb{R}^p$, the dual norm $\|.\|^*$ is defined by

$$\|x\|^* = \sup_{s \in \mathbb{R}^p : \|s\| \leq 1} s'x.$$

If the unit ball $B = \{x \in \mathbb{R}^p : \|x\| \leq 1\}$ is of polytope shape, the dual of B given by $B^* = \{x \in \mathbb{R}^p : \|x\|^* \leq 1\}$, the unit ball of the dual norm, is, again, a polytope. In this case, the penalty term is not differentiable and there is a strong connection between the subdifferentials $\partial_{\|.\|}(.)$ of the norm $\|.\|$ and the faces of the polytope B^*. The precise association is detailed in Appendices A.1-A.3 and this connection provides the basis for the main theorem.

Theorem 1 (Necessary and sufficient condition for uniqueness)**.** *Let $X \in \mathbb{R}^{n \times p}$ and let $\|.\|$ be a norm on $\mathbb{R}^p$ whose unit ball B is given by a polytope. Consider the penalized optimization problem*

$$S_{X,\|.\|}(y) = \underset{b \in \mathbb{R}^p}{\text{Arg min}} \frac{1}{2}\|y - Xb\|^2 + \|b\|, \tag{1}$$

where $y \in \mathbb{R}^n$. Let B^* denote the unit ball of the dual norm $\|.\|^*$. There exists $y \in \mathbb{R}^n$ with $|S_{X,\|.\|}(y)| >$

1 *if and only if* $\text{row}(X)$ *intersects a face of the dual unit ball* B^* *whose codimension is larger than* $\text{rk}(X)$.

As mentioned in the introducation, the notion of uniqueness considered in Theorem 1 is strong in the sense that it guarantees uniqueness for a given design matrix X *for all* values $y \in \mathbb{R}^n$. Such concept of uniqueness is beneficial when studying the stochastic properties of the minimizer in a statistical framework, as then y varies and a criterion independent of y is desirable. Also note that we make no assumptions on X.

If the norm $\|.\|$ involves a tuning parameter λ, the uniqueness of the corresponding penalized problem does not depend on the particular choice of λ. The parameter simply scales B and subsequently B^* and does not affect which faces are intersected by the vector space $\text{row}(X)$.

Theorem 1 generalizes Theorem 14 given in Ewald & Schneider (2020), which provides a necessary and sufficient condition for the uniqueness of the LASSO minimizer: All LASSO solutions are unique if and only if $\text{row}(X)$ only intersects faces of the unit cube $[-1,1]^p$ whose codimension is less than or equal to $\text{rk}(X)$. Note that the unit cube is, indeed, the corresponding dual to the unit ball of the l_1-norm.

Example. *We illustrate the criterion from Theorem 1 for* $\|.\| = \|.\|_\infty$*, the supremum norm, in Figure 1. Let* $X = (1 \;\; 0)$*. The unit dual ball* B^* *is given by the unit cross-polytope* $\text{conv}\{\pm(1,0)', \pm(0,1)'\}$ *and we have* $\text{rk}(X) = 1$*. Clearly, the vertex* $(1,0)'$ *with codimension* $p - 0 = 2 > 1 = \text{rk}(X)$ *intersects* $\text{row}(X)$*, so that one can pick* $y \in \mathbb{R}$ *for which the set of minimizers* $S_{X,\|.\|_\infty}(y)$ *is not a singleton. In Figure 1(a), we illustrate this fact for* $S_{X,\|.\|_\infty}(2)$.

Also consider $X = (1 \;\; 1)$*. Because* $\text{row}(X)$ *does not intersect any vertex of* $\text{conv}\{\pm(1,0)', \pm(0,1)'\}$*, the solution set* $S_{X,\|.\|_\infty}(y)$ *is always a singleton. In Figure 1(b), we illustrate this fact for* $S_{X,\|.\|_\infty}(2)$.

2.1 The related problem of basis pursuit

As mentioned before, the methods of LASSO and basis pursuit (BP) are closely related, as the BP problem can be thought of a LASSO problem with vanishing tuning parameter. More concretely, the setting for BP is the following. Let $X \in \mathbb{R}^{n \times p}$ and let $y \in \text{col}(X)$. The set $S_{X,\text{bp}}(y)$ of BP minimizers is defined as

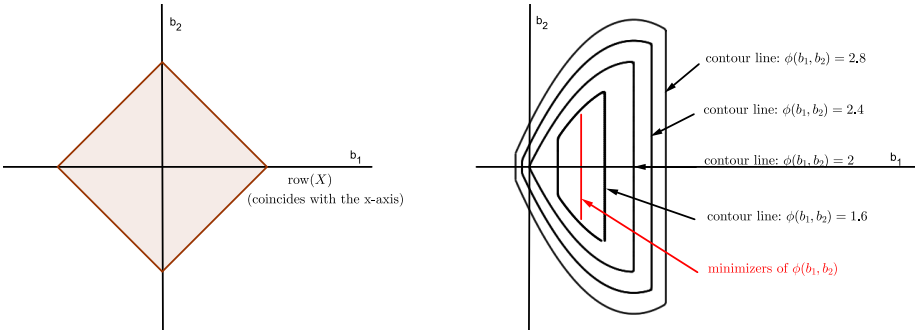
$$S_{X,\text{bp}}(y) = \text{Arg min} \|b\|_1 \text{ subject to } Xb = y.$$

The following theorem shows that, indeed, as BP is a limiting case of the LASSO, the corresponding uniqueness condition – which is independent of the choice of tuning parameter as discussed above – carries over to the BP problem.

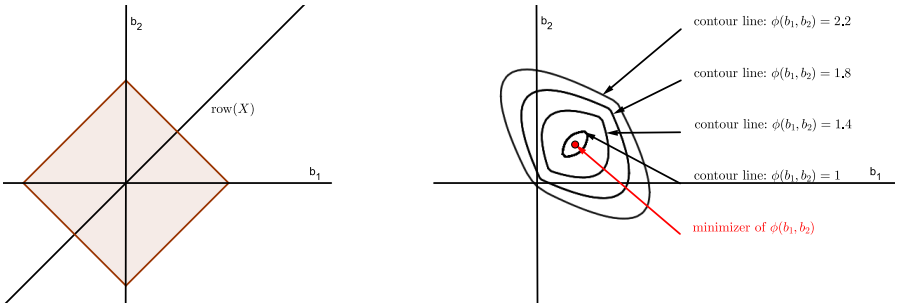
Theorem 2. *Let* $X \in \mathbb{R}^{n \times p}$*. There exists* $y \in \text{col}(X)$ *for which* $|S_{X,\text{bp}}(y)| > 1$ *if and only if* $\text{row}(X)$ *intersects a face of the unit cube* $[-1,1]^p$ *whose codimension is larger than* $\text{rk}(X)$.

We illustrate Theorem 2 in Figures 2(a) and 2(b).

In the following proposition, we show that the necessary and sufficient condition given in Theorem 1 and therefore also the one given in Theorem 2 is weak. More precisely, we establish that the set of $X \in \mathbb{R}^{n \times p}$ for which the necessary and sufficient condition given in Theorem 1 does not hold, is negligible with respect to the Lebesgue measure.

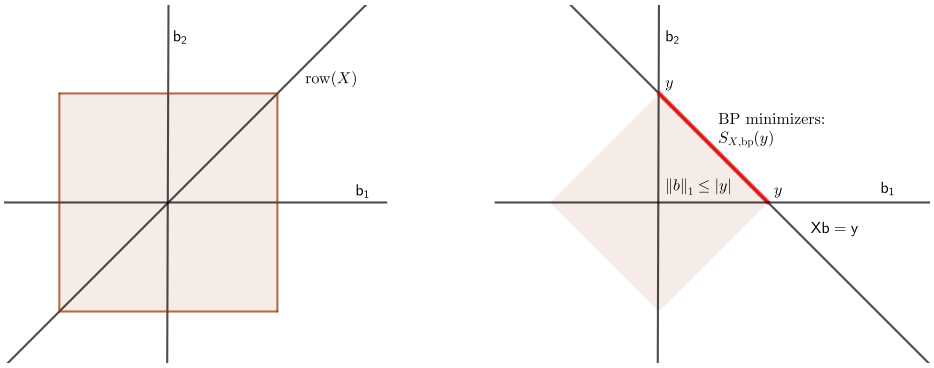


(a) Let $X = (1 \;\; 0)$. On the left-hand side, we see that $\mathrm{row}(X)$ intersects a vertex of the cross-polytope whose codimension is 2 and thus is larger than $\mathrm{rk}(X) = 1$. Therefore, there exists $y \in \mathbb{R}$ for which $S_{X,\|\cdot\|_\infty}(y)$ is not a singleton. On the right-hand side, the contour lines of the objective function $\phi(b_1, b_2) = 0.5(2 - b_1)^2 + \max\{|b_1|, |b_2|\}$ show that the set $S_{X,\mathrm{bp}}(2)$ (in red), indeed, contains infinitely many points.

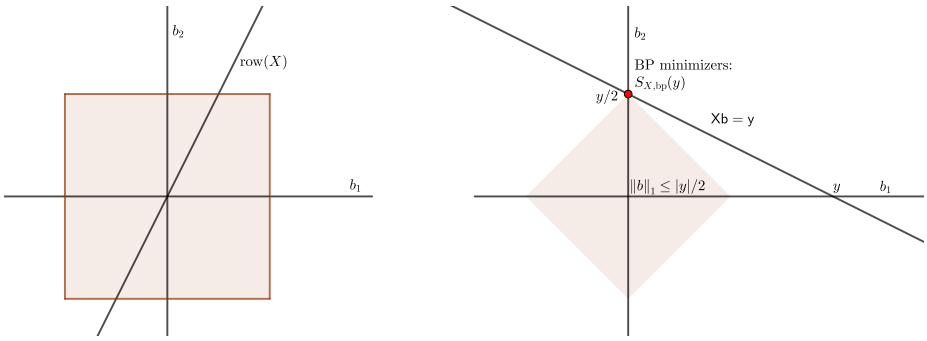


(b) Let $X = (1 \;\; 1)$. On the left-hand side, we see that $\mathrm{row}(X)$ does not intersect any face of the cross-polytope whose codimension is larger than $\mathrm{rk}(X) = 1$ (such faces are the vertices in this example). Therefore, the set $S_{X,\|\cdot\|_\infty}(y)$ is a singleton for all $y \in \mathbb{R}$. On the right-hand side, the contour lines of the objective function $\phi(b_1, b_2) = 0.5(2 - b_1)^2 + \max\{|b_1|, |b_2|\}$ show that the set $S_{X,\|\cdot\|_\infty}(2)$ (in red) does, indeed, only contain a single point.

Figure 1: Illustration of Theorem 1 for the supremum norm.



(a) Let $X = (1 \;\; 1)$. On the left-hand side, we see that $\mathrm{row}(X)$ intersects a face of the unit square whose codimension 2 is larger than $\mathrm{rk}(X) = 1$ (which are the vertices in this example). Therefore, by Theorem 2, there exists $y \in \mathbb{R}$ for which the BP minimizer is not unique. The right-hand side illustrates that, indeed, for an arbitrary $y \in \mathbb{R} \setminus \{0\}$, the set $S_{X,\mathrm{bp}}(y)$ (the red segment) is not a singleton.



(b) Illustration of Let $X = (1 \;\; 2)$. On the left-hand side, we see that $\mathrm{row}(X)$ does not intersect any face of the unit square whose codimension is larger than $\mathrm{rk}(X) = 1$ (which are the vertices in this example). Therefore, by Theorem 2, the BP minimizer is unique for all $y \in \mathbb{R}$. The right-hand side illustrates that for an arbitrary $y \in \mathbb{R}$, the set $S_{X,\mathrm{bp}}(y)$ (in red) is, indeed, a singleton.

Figure 2: Illustration of Theorem 2.

Proposition 1. *Let μ be the Lebesgue measure on $\mathbb{R}^{n\times p}$ and let $\|.\|$ be a norm on $\mathbb{R}^p$ whose unit ball is given by a polytope. The following equality holds*

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

The following corollary is then straightforward given the fact that the LASSO, which is covered by Theorem 1, and BP share the same the characterization for uniqueness.

Corollary 1. *Let μ be the Lebesgue measure on $\mathbb{R}^{n\times p}$, then the following equality holds*

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

By taking the appropriate norms in Proposition 1, and by Corollary 1, one may deduce that the necessary and sufficient conditions for uniqueness of BP, LASSO, and SLOPE are weak. However, one should be aware that Proposition 1 does not mean that this condition always occurs in practice! For example, for BP (or LASSO), when $p > n$ and $X \in \{-1, 1\}^{n\times p}$, one can always pick $y \in \mathrm{col}(X)$ for which the set of minimizers $S_{X,\mathrm{bp}}(y)$ is not a singleton (or, for any $\lambda > 0$, one can pick $y \in \mathbb{R}^n$ for which the set of minimizers $S_{X,\lambda\|.\|_1}(y)$ is not a singleton). Matrices having entries in $\{-1, 1\}$ appear in several theoretical works, such as Rauhut (2010) and Tardivel et al. (2018), and are used for applications in radar and wireless communication (see e.g. Romberg, 2009; Haupt et al., 2010).

3 Model selection properties

The geometric considerations around Theorems 1 and 2 can also provide insights on the model selection aspects of the method under consideration. The keystone is to associate a model with face of the polytope B^*, the unit ball of the dual norm. For LASSO and BP in Section 3.1, we exploit the fact that each face of the unit cube corresponds to a sign vector and show that the faces intersected by the row span of X provide the accessible sign vectors for these estimators. We take a similar, but more sophisticated approach for SLOPE in Section 3.2 where the models we consider also carry information about the clustering phenomenon of the method.

In Section 3.3, we take a different angle and characterize the SLOPE null polytope and its connection to the sparsity and clustering property of this method. For the LASSO, it is known that the estimation residuals are the projection of y onto the LASSO null polytope. We also further generalize this fact to arbitrary norm-penalized least-squares estimation.

3.1 Accessible sign vectors for LASSO and BP

We start by introducing the notion of accessible sign vectors for LASSO and BP problems.

Definition 1 (Accessible sign vectors for LASSO and BP)**.** *Let $X \in \mathbb{R}^{n\times p}$, $\sigma \in \{-1, 0, 1\}^p$, and $\lambda > 0$. We say that σ is an accessible sign vector for LASSO (or BP) with respect to X, if there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\|.\|_1}(y)$ (or there exists $y \in \mathrm{col}(X)$ and $\hat{\beta} \in S_{X,\mathrm{bp}}(y)$, respectively), such that* $\mathrm{sign}(\hat{\beta}) = \sigma$.

The following theorem provides a geometric characterization of accessible sign vectors for LASSO and BP based on faces of the unit cube $[-1,1]^p$ and the vector space $\mathrm{row}(X)$. First, note that subdifferential calculus of the l_1-norm at $\sigma \in \{-1,0,1\}^p$ gives

$$\partial_{\|.\|_1}(\sigma) = E_1 \times \cdots \times E_p \text{ with } E_j = \begin{cases} \{\sigma_j\} & |\sigma_j| = 1 \\ [-1,1] & \sigma_j = 0, \end{cases}$$

where $\partial_{\|.\|_1}(x)$ denotes the subdifferential of the l_1-norm at $x \in \mathbb{R}^p$, see Appendices A.1 and A.3 for more details. Therefore, the mapping $\sigma \mapsto \partial_{\|.\|_1}(\sigma)$ is a bijection between sign vectors in $\{-1,0,1\}^p$ and faces of the unit cube in $\mathbb{R}^p$. We let $F_1(\sigma) = \partial_{\|.\|_1}(\sigma)$ in the following. For completeness, Theorem 3 also contains an analytic characterization of accessibility.

Theorem 3 (Characterization of accessible LASSO and BP sign vectors)**.** *Let $X \in \mathbb{R}^{n\times p}$ and $\lambda > 0$.*

1) *Geometric characterization: A sign vector $\sigma \in \{-1,0,1\}^p$ is accessible for LASSO or BP with respect to X if and only if* $\mathrm{row}(X)$ *intersects the face $F_1(\sigma)$.*

2) *Analytic characterization: A sign vector $\sigma \in \{-1,0,1\}^p$ is accessible for LASSO or BP with respect to X if and only if the implication*

$$Xb = X\sigma \implies \|b\|_1 \geq \|\sigma\|_1$$

holds.

The analytic characterization for accessible sign vectors is, in fact, closely related to the identifiability condition given in Tardivel & Bogdan (2018), in which the inequality above is replaced by a strict inequality. In high-dimensional linear regression, this condition is necessary and sufficient for sign recovery of thresholded LASSO and thresholded BP (Tardivel & Bogdan, 2018), as well as for so-called thresholded justice pursuit (Descloux et al., 2020), a method closely related to BP. We point out that the analytic characterization allows to check accessibility of a particular sign vector simply by solving a BP problem, which in turn gives insight on whether the corresponding face of the unit cube is intersected by $\mathrm{row}(X)$. In practice, one does not even need an accurate numerical solver to check whether a sign vector $\sigma \in \{-1,0,1\}^p$ is accessible, when the BP problem is uniquely solvable: if we are given an approximate minimizer $\tilde{\beta}$ for the BP problem with $y = X\sigma$ that satisfies $\|\tilde{\beta} - \hat{\beta}\|_\infty < 1/2$, where $\hat{\beta}$ is the exact minimizer, it suffices to check whether $\mathrm{sign}(\mathrm{round}(\tilde{\beta})) = \sigma$, where round(.) rounds componentwise to the closest integer. In that case, σ is accessible, whereas σ is not accessible if $\mathrm{sign}(\mathrm{round}(\tilde{\beta})) \neq \sigma$, as outlined in Corollary 3 in Appendix A.6. This approach to check accessibility was used in Tardivel & Bogdan (2018) to derive the so-called identifiability curve.

Note that Theorem 3 reveals that whether a sign vector is accessible for LASSO does not depend on the value of the tuning parameter λ. We also point out that Theorems 1 and 3 allow to deduce that the number of non-null components of the LASSO is always less than or equal to $\mathrm{rk}(X)$ when the solutions are unique. Indeed, if the LASSO minimizer is unique, according to Theorem 1, $\mathrm{row}(X)$ does not intersect a face of $[-1,1]^p$ associated to a sign vector having more than $\mathrm{rk}(X)$ non-null components, i.e., a face whose codimension is larger than $\mathrm{rk}(X)$. This implies that only sign vectors with at most

$\mathrm{rk}(X)$ components different to zero are accessible. For the LASSO, this is a refined version of the well-known fact that, in case the estimator is unique, at most n components can be non-zero (see e.g. Tibshirani, 2013; Osborne et al., 2000). A similar approach for SLOPE is developed in the following, geometrically characterizing that the number of non-null clusters of SLOPE minimizers is less than or equal to $\mathrm{rk}(X)$ in case of uniqueness.

3.2 Accessible models for SLOPE

We now turn to accessible models for SLOPE, whose norm is given by $\|b\|_w = \sum_{j=1}^p w_j |b|_{(j)}$, where $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$, as introduced before. For the remainder of Section 3, we assume that the weight vector w of the satisfies

$$w_1 > \cdots > w_p > 0,$$

i.e., that all components non-zero and strictly decreasing. (This assumption is not needed for applying Theorem 1 to SLOPE, since $w_1 > 0$ and decreasing components are sufficient for $\|.\|_w$ to be a norm.) We introduce a more sophisticated notion of a "model" chosen by SLOPE compared to sign vectors that can account for the clustering property which is not shared by LASSO or BP.

Definition 2. *We say that a vector $m \in \mathbb{Z}^p$ is a SLOPE model, if either $m = 0$, or, if for all $l \in [\|m\|_\infty]$, there exists $j \in [p]$ such that $|m_j| = l$. We denote the set of all SLOPE models of dimension p by $\mathcal{M}_p$. Moreover, for $x \in \mathbb{R}^p$, we define $\mathrm{mdl}(x) \in \mathcal{M}_p$ through the following.*

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

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

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

Example. *For $x = (3.1, -1.2, 0, -3.1)'$, we have $\mathrm{mdl}(x) = (2, -1, 0, -2)'$. For $x \in \mathbb{R}^4$ with $\mathrm{mdl}(x) = (0, 2, 1, -2)'$, we have $\mathrm{sign}(x) = (0, 1, 1, -1)'$ and $|x_2| = |x_4| > |x_3| > x_1 = 0$. The set of all SLOPE models in $\mathbb{R}^2$ is given by*

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

The main geometric object of study in this section is the sign permutahedron, which constitutes the dual of the SLOPE norm unit ball (Proposition 8 in Appendix A.7) and is defined as

$$P_w^{\pm} = \mathrm{conv}\left\{(\sigma_1 w_{\pi(1)}, \ldots, \sigma_p w_{\pi(p)})' : \sigma_1, \ldots, \sigma_p \in \{-1, 1\}, \pi \in \mathcal{S}_p\right\}.$$

The shape of this polytope is illustrated in Figure 3 (in two dimensions) and in Figure 4 (in three dimensions). Also of importance will be the permutahedron, defined by

$$P_w = \mathrm{conv}\left\{(w_{\pi(1)}, \ldots, w_{\pi(p)})' : \pi \in \mathcal{S}_p\right\}.$$

The permutahedron is, in fact, a face of the sign permutahedron $P_w^{\pm}$. We denote the subdifferential of the SLOPE norm at $x \in \mathbb{R}^p$ by $\partial_{\|.\|_w}(x)$. Any $\partial_{\|.\|_w}(x)$ is a face of $P_w^{\pm}$, which we shall denote by $F_w(x)$ in the following.

SLOPE models m having only positive components can be interpreted as an ordered partition of $[p]$, where the the smallest and largest element of this partition is the set $\{j : m_j = 1\}$ and the set $\{j : m_j = \|m\|_\infty\}$, respectively. It is well known that there is a one-to-one relationship between the elements of an ordered partition and the faces of the permutahedron (see e.g. Maes & Kappen, 1992; Simion, 1997; Ziegler, 2012). Instigated by this, we show in Theorem 4 that this result can, indeed, be extended to a one-to-one relationship between all SLOPE models and the non-empty faces of the sign permutahedron, which we denote by $\mathcal{F}_0(P_w^{\pm})$.

Theorem 4. *The mapping $m \in \mathcal{M}_p \mapsto F_w(m) = \partial_{\|.\|_w}(m)$ is a bijection between the SLOPE models $\mathcal{M}_p$ and $\mathcal{F}_0(P_w^{\pm})$, the non-empty faces of the sign permutahedron $P_w^{\pm}$. In addition, the following holds.*

1) The codimension of $F_w(m)$ is given by $\|m\|_\infty$.

2) We have $F_w(x) = F_w(\mathrm{mdl}(x))$.

The assumption that components of w are strictly decreasing and non-zero is important. For example, if $w_1 = \cdots = w_p > 0$, the sign permutahedron is just a cube and clearly, there is no one-to-one relationship between the set SLOPE models and the set of faces of the cube. A similar situation arises if w contains zero components. As can be seen when $p = 2$ and $w_2 = 0$, the SLOPE norm is the supremum norm and the corresponding dual unit ball is the unit cross-polytope in $\mathbb{R}^2$, whose faces cannot bijectively be mapped to $\mathcal{M}_2$ given in the example above.

Example. *We now describe the faces $F_w(m)$, $m \in \mathcal{M}_2$, of the sign permutahedron $P_w^{\pm}$ when $w = (3.5, 1.5)'$. In the following, we use the fact that – up to an orthogonal transformation described in Lemma 5 – $F_w(m)$ is equal to $F_w(\tilde{m})$ for some $\tilde{m}$, a non-negative and non-increasing SLOPE model. The relationship between the SLOPE models $m \in \mathcal{M}_2$ and faces of the sign permutahedron $P_w^{\pm}$ are listed below and illustrated in Figure 3. Note that* $\mathrm{codim}(F_w(m)) = \|m\|_\infty$.

model $\tilde{m}$	*face $F_w(\tilde{m})$*	*codim.*	*faces $F_w(m)$ isometric to $F_w(\tilde{m})$*
$\tilde{m} = (0,0)'$	*sign permutahedron:* $P_w^{\pm}$	*0*	–
$\tilde{m} = (1,0)'$	*segment:* $\{3.5\} \times [-1.5, 1.5]$	*1*	$m \in \{(-1,0)', \pm(0,1)'\}$
$\tilde{m} = (1,1)'$	*permutahedron:* P_w	*1*	$m \in \{(-1,-1)', \pm(1,-1)'\}$
$\tilde{m} = (2,1)'$	*point:* $(3.5, 1.5)'$	*2*	$m \in \{(-2,-1)', \pm(2,-1)', \pm(1,2)', \pm(1,-2)'\}$

Analogously to the accessible sign vectors for LASSO and BP, for a given X, we introduce the notion of accessible SLOPE models.

Definition 3 (Accessible SLOPE model)**.** *Let $X \in \mathbb{R}^{n \times p}$ and $m \in \mathcal{M}_p$. We say that m is an accessible SLOPE model with respect to X if*

$$\exists y \in \mathbb{R}^n \text{ and } \exists \hat{\beta} \in S_{X,\|.\|_w}(y) \text{ such that } \mathrm{mdl}(\hat{\beta}) = m.$$

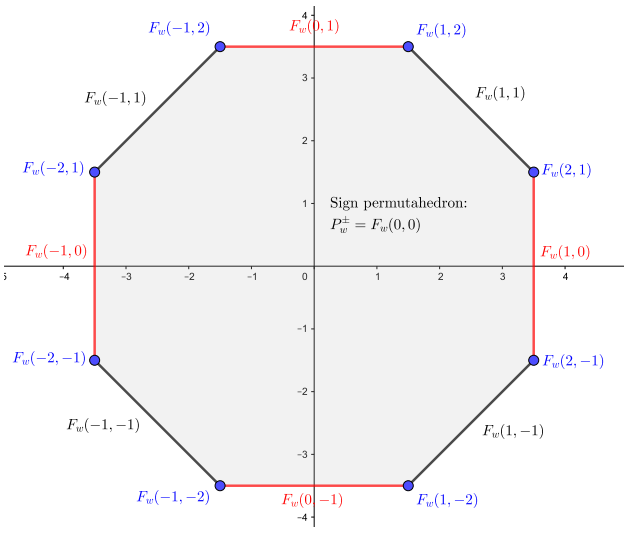


Figure 3: Illustration of the relationship between the SLOPE models and the faces of the sign permutahedron $P_w^{\pm}$ for $w = (3.5, 1.5)'$ through subdifferential calculus, see Proposition 6 in Appendix A.3 and Proposition 8 in Appendix A.7. Note that $F_w(m) = \partial_{\|.\|_w}(m)$. Faces having the same color are isometric. One may notice that $\operatorname{codim}(F_w(m)) = \|m\|_\infty$.

We now provide a geometric and analytic characterization of accessible SLOPE models.

Theorem 5 (Characterization of accessible SLOPE models)**.** *Let $X \in \mathbb{R}^{n\times p}$.*

1) *Geometric characterization: A SLOPE model $m \in \mathcal{M}_p$ is accessible with respect to X if and only if* $\operatorname{row}(X)$ *intersects the face $F_w(m)$.*

2) *Analytic characterization: A SLOPE model $m \in \mathcal{M}_p$ is accessible with respect to X if and only if the implication*

$$Xb = Xm \implies \|b\|_w \geq \|m\|_w$$

holds.

We point out that the analytic characterization allows to check accessibility of a particular SLOPE model by in fact minimizing a BP-like problem where the l_1-norm is replaced by the SLOPE norm. This in turn can give insight on whether the corresponding face of the sign permutahedron is intersected by $\operatorname{row}(X)$.

Also note that the set of accessible SLOPE models is invariant by scaling w with a constant, since $\operatorname{row}(X)$ intersects $F_w(m)$ if and only if $\operatorname{row}(X)$ intersects $F_{\lambda w}(m)$ with $\lambda > 0$. The following corollary, which is in line with Theorem 2.1 very recently given in Kremer et al. (2019), is a straightforward consequence of Theorems 1, 4 and 5.

Corollary 2. *Let $X \in \mathbb{R}^{n\times p}$. If* $\operatorname{row}(X)$ *does not intersect any face of $P_w^{\pm}$ with codimension larger than* $\operatorname{rk}(X)$*, then for all $y \in \mathbb{R}^n$, $\hat{\beta}_w(y)$, the unique element of $S_{X,\|.\|_w}(y)$, satisfies* $\|\operatorname{mdl}(\hat{\beta}_w(y))\|_\infty \leq \operatorname{rk}(X)$.

Corollary 2 generalizes the well known fact that, when uniqueness occurs, the LASSO minimizer has less than $\mathrm{rk}(X)$ non-null components. Indeed, the above corollary shows that when the SLOPE minimizer is unique, the number of non-null clusters is less than or equal to $\mathrm{rk}(X)$.

Example. *We illustrate the criterion for accessible SLOPE models from Theorem 5 for* $w = (5.5, 3.5, 1.5)'$ *and* X *given by*

$$X = \begin{pmatrix} 8 & 5 & 8 \\ 10 & 1.25 & -6 \end{pmatrix}.$$

Table 1 lists all accessible non-null SLOPE models ($m = 0$ *is always accessible through* $y = 0$*), the geometric illustration is shown in Figure 4.*

colour	type	intersection $\neq \varnothing$	face intersected isometric to	SLOPE models
orange	segments	$\mathrm{row}(X) \cap F_w(\pm(1,0,0))$	$\{5.5\} \times P^{\pm}_{(3.5,1.5)}$	$\pm(1,0,0)$
red	segments	$\mathrm{row}(X) \cap F_w(\pm(1,1,1))$	$P_{(5.5,3.5,1.5)}$	$\pm(1,1,1)$
black	segments	$\mathrm{row}(X) \cap F_w(\pm(0,0,1))$	$\{5.5\} \times P^{\pm}_{(3.5,1.5)}$	$\pm(0,0,1)$
pink	segments	$\mathrm{row}(X) \cap F_w(\pm(-1,0,1))$	$P_{(5.5,3.5)} \times [-1.5, 1.5]$	$\pm(-1,0,1)$
purple	points	$\mathrm{row}(X) \cap F_w(\pm(2,0,-1))$	$\{5.5\} \times \{3.5\} \times [-1.5, 1.5]$	$\pm(2,0,-1)$
green	points	$\mathrm{row}(X) \cap F_w(\pm(2,1,1))$	$\{5.5\} \times P_{(3.5,1.5)}$	$\pm(2,1,1)$
blue	points	$\mathrm{row}(X) \cap F_w(\pm(1,1,2))$	$\{5.5\} \times P_{(3.5,1.5)}$	$\pm(1,1,2)$
yellow	points	$\mathrm{row}(X) \cap F_w(\pm(-1,0,2))$	$\{5.5\} \times \{3.5\} \times [-1.5, 1.5]$	$\pm(-1,0,2)$

Table 1: Accessible SLOPE models with respect to $X = \left(\begin{smallmatrix} 8 & 5 & 8 \\ 10 & 1.25 & -6 \end{smallmatrix}\right)$ and $w = (5.5, 3.5, 1.5)'$.

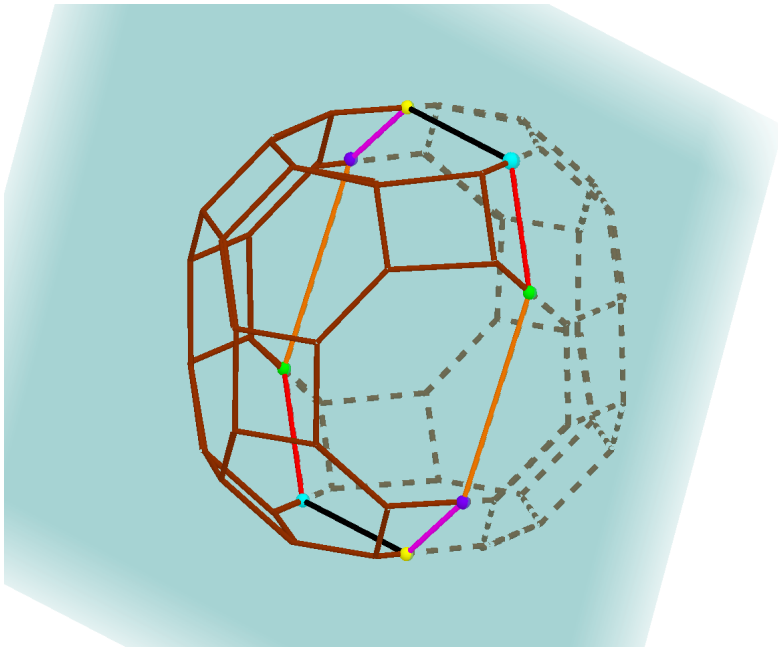


Figure 4: Illustration of the sign permutahedron $P_w^{\pm}$ (in brown) and the plane $\mathrm{row}(X)$ (in light blue). Because $\mathrm{rk}(X) = 2$ and $\mathrm{row}(X)$ does not intersect any vertex of $P_w^{\pm}$ (the faces with codimension equal to 3), the SLOPE estimator $\hat{\beta}_w(y)$ is unique for all values of $y \in \mathbb{R}^2$. Colored segments and points are the intersections between $\mathrm{row}(X)$ and the faces of $P_w^{\pm}$, determining the accessible SLOPE models shown in Table 1. For example, $m = (2, 1, 1)'$ is an accessible SLOPE model, which implies that there exists $y \in \mathbb{R}^2$ for which the SLOPE minimizer $\hat{\beta}_w(y)$ satisfies $\hat{\beta}_w(y)_1 > \hat{\beta}_w(y)_2 = \hat{\beta}_w(y)_3 > 0$. In addition, since $m = (2, 1, 0)'$ is not an accessible model, one cannot pick $y \in \mathbb{R}^2$ for which the SLOPE minimizer satisfies $\hat{\beta}_w(y)_1 > \hat{\beta}_w(y)_2 > \hat{\beta}_w(y)_3 = 0$.

3.3 The SLOPE null polytope and a general result

In the previous section, we gave a description of accessible SLOPE models based on the intersection of $\mathrm{row}(X)$ with the sign permutahedron $P_w^{\pm}$. In this section, our aim is the following: Given an accessible model $m \in \mathcal{M}_p$, we want to provide the set of $y \in \mathbb{R}^n$ for which there exists $\hat{\beta} \in S_{X,\|.\|_w}(y)$ with $\mathrm{mdl}(\hat{\beta}) = m$. In other words, we want to describe the set

$$A_w(m) = \{y \in \mathbb{R}^n : \exists \hat{\beta} \in S_{X,\|.\|_w}(y) \text{ where } \mathrm{mdl}(\hat{\beta}) = m\}.$$

Note that when the SLOPE minimizer is unique, the sets $A_w(m)$ and $A_w(\tilde{m})$ are disjoint for $m \neq \tilde{m}$, whereas $A_w(m) \cap A_w(\tilde{m}) \neq \varnothing$ might occur in case of non-uniqueness. Clearly, the empty model $m = 0$ is accessible. The corresponding set $A_w(0)$, called the *SLOPE null polytope*, given by

$$A_w(0) = \{y \in \mathbb{R}^n : \|X'y\|_w^* \leq 1\}$$

by Proposition 7. This is the set of all y such that $X'y \in P_w^{\pm}$, which is again a polytope. The proposition below shows that the faces of this polytope $N_w(m) = \{f \in \mathbb{R}^n : X'f \in F_w(m)\}$ for the *accessible* SLOPE models m are the cornerstone to describe the sets $A_w(m)$.

Proposition 2. *Let $X \in \mathbb{R}^{n\times p}$. The SLOPE model $m \in \mathcal{M}_p$ is an accessible SLOPE model if and only if $N_w(m) = \{f \in \mathbb{R}^n : X'f \in F_w(m)\} \neq \varnothing$. In that case, the set $A_w(m)$ is given by*

$$A_w(m) = \{y = f + Xb : f \in N_w(m), \mathrm{mdl}(b) = m\}.$$

Note that Proposition 2 yields another characterization of accessible SLOPE models, namely that m is accessible if and only if $N_w(m)$ is a non-empty face of the SLOPE null polytope. In case of non-uniqueness, different models may yield the same face, so one should be aware that there is no bijection between the accessible SLOPE models and the faces of the SLOPE null polytope. Also note that if $\mathrm{rk}(X) = n$ and we are given the intersection between $\mathrm{row}(X)$ and $F_w(m)$ for some accessible SLOPE model m, we can write $N_w(m) = (XX')^{-1}X(\mathrm{row}(X) \cap F_w(m))$ since

$$f \in N_w(m) \iff X'f \in \mathrm{row}(X) \cap F_w(m) \iff f \in (XX')^{-1}X(\mathrm{row}(X) \cap F_w(m)).$$

Example. *Figure 4 illustrates the accessible SLOPE models from Theorem 5 for $w = (5.5, 3.5, 1.5)'$ and*

$$X = \begin{pmatrix} 8 & 5 & 8 \\ 10 & 1.25 & -6 \end{pmatrix}.$$

Now, for every accessible SLOPE model, Figure 5 below provides the set $A_m = \{y \in \mathbb{R}^2 : \exists \hat{\beta} \in S_{X,\|.\|_w}(y) \text{ where } \mathrm{mdl}(\hat{\beta}) = m\}$ and the SLOPE null polytope.

Note that the SLOPE null polytope $A_w(0)$ can also be interpreted as the set of SLOPE residuals in the sense that $\hat{u} = y - X\hat{\beta}$ is the projection of y onto $A_w(0)$ whenever $\hat{\beta} \in S_{X,\|.\|_w}(y)$ (Minami, 2020). Or put differently again, we can decompose y as $y = X\hat{\beta} + \hat{u}$, where $X\hat{\beta}$ is the SLOPE fit and $\hat{u} \in A_w(0)$, the set of all values that lead to a zero SLOPE minimizer.

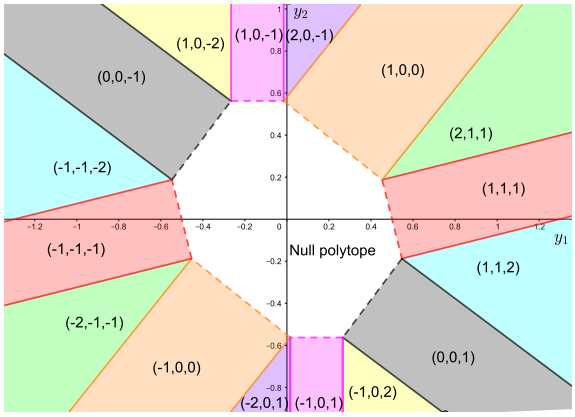


Figure 5: Illustration of the SLOPE null polytope and the accessible models for $X = \begin{pmatrix} 8 & 5 & 8 \\ 10 & 1.25 & -6 \end{pmatrix}$ and $w = (5.5, 3.5, 1.5)'$. The resulting accessible models are $\{\pm(1,0,0), \pm(1,1,1), \pm(0,0,1), \pm(-1,0,1), \pm(2,0,-1), \pm(2,1,1), \pm(1,1,2), \pm(-1,0,2)\}$, each associated with a face of the polytope. Depicted also are the sets $A_w(m) = \{y \in \mathbb{R}^2 : \exists \hat{\beta} \in S_{X,\|.\|_w}(y) \text{ with } \mathrm{mdl}(\hat{\beta}) = m\}$ for each accessible model.

This property is well known also for the LASSO, (c.f. Tibshirani & Taylor, 2012). In fact, it is straightforward to see from Proposition 7 that the same considerations hold for all problems as defined in (1). For completeness, we summarize this in the following proposition which holds for arbitrary norms.

Proposition 3. *Let $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$ and let $\|.\|$ be a norm on $\mathbb{R}^p$. Define the convex null set $A_\varnothing = \{u \in \mathbb{R}^n : \|X'u\|^* \leq 1\}$. We then have $S_{X,\|.\|}(u) = \{0\}$ for all $u \in A_\varnothing$, and any $\hat{\beta} \in S_{X,\|.\|}(y)$ satisfies $y = X\hat{\beta} + \hat{u}$ with $\hat{u} \in A_\varnothing$. Moreover, $\hat{u}$ is the projection of y onto $A_\varnothing$.*

4 Acknowledgments

We would like to thank Jan Mielniczuk and Świątosław Gal for their insightful comments on the paper.

A Appendix – Proofs

In the appendix, we additionally make use of the following notation. Let A be a matrix. We use the symbol A_j to denote the j-the column of A. For an index set I, A_I is the matrix containing columns with indices in I only. For a vector x, $\mathrm{supp}(x)$ contains the indices of the non-zero components of x. The symbol $|x|_{(j)}$ denotes the j-th order statistic of the absolute values of the components of x, i.e., $|x|_{(1)} \geq |x|_{(2)} \geq \ldots$. Let $l, k \in \mathbb{N}$ with $l \leq k$, then $[l:k]$ denotes the set $\{l, l+1, \ldots, k\}$. We let $\mathbf{1}_m$ stand for the vector $(1, \ldots, 1)' \in \mathbb{R}^m$. All inequalities involving vectors are understood componentwise.

A.1 Facts about subdifferentials

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

$$f(z) \geq f(x) + s'(z - x) \;\; \forall z \in \mathbb{R}^p. \tag{2}$$

The set of all subgradients of f at x, which is a convex set, is called the *subdifferential of f at x*, denoted by $\partial_f(x)$. It is straightforward to characterize the minimizer of a function in the following way

$$x^* \in \operatorname{Arg\,min} f \iff 0 \in \partial_f(x^*). \tag{3}$$

While convexity of f is not necessary for the above statement, the use of subdifferentials is an especially important tool when this is the case. Given that f is convex, subdifferentiability is also a local property in the sense that for any $\delta > 0$, we have

$$s \in \partial_f(x) \iff f(x + h) \geq f(x) + s'h \text{ for all } h : \|h\|_\infty \leq \delta. \tag{4}$$

A.2 Facts about polytopes

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

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

$$P_\mathcal{V} = \operatorname{conv}(V_1, \ldots, V_k) = \operatorname{conv}(V)$$

for $V = (V_1 \ldots V_k) \in \mathbb{R}^{p \times k}$. A set $P_\mathcal{H} \subseteq \mathbb{R}^p$ is called an $\mathcal{H}$-polyhedron, if it is the intersection of a finite number of half-spaces, namely,

$$P_\mathcal{H} = \bigcap_{l=1}^{m} \{x \in \mathbb{R}^p : A_l' x \leq b_l\} = \{x \in \mathbb{R}^p : A'x \leq b\},$$

for some $A = (A_1 \ldots A_m) \in \mathbb{R}^{p \times m}$ and $b \in \mathbb{R}^m$. A bounded $\mathcal{H}$-polyhedron is called $\mathcal{H}$-polytope. A set $P \subseteq \mathbb{R}^p$ is an $\mathcal{H}$-polytope if and only if it is a $\mathcal{V}$-polytope. We therefore simply use the term *polytope* in the following. The *dimension* $\dim(P)$ of a polytope is given by the dimension of $\operatorname{aff}(P)$, the affine subspace spanned by P, and its codimension by $\operatorname{codim}(P) = p - \dim(P)$. A *face* F of P is any subset $F \subseteq P$ that satisfies

$$F = \{x \in P : a'x = b_0\}, \text{ where } P \subseteq \{x : a'x \leq b_0\},$$

for some $a \in \mathbb{R}^p$ and $b_0 \in \mathbb{R}$. Such an inequality $a'x \leq b_0$ is called a *valid inequality* of P. Note that $F = \varnothing$ and $F = P$ are faces of P and that any face F is again a polytope. A face $F \neq P$ is called *proper*. A face of dimension 0 is called *vertex*, and we denote the set of all vertices of P by $\operatorname{vert}(P)$. This set satisfies $\operatorname{vert}(P) \subseteq \{V_1, \ldots, V_k\}$, where $P = \operatorname{conv}(V_1, \ldots, V_k)$. A point $x_0 \in P$ lies in $\operatorname{relint}(P)$,

the *relative interior* of P, if x_0 is not contained in a proper face of P. Finally, the *(polar) dual* of P is defined as

$$P^* = \{s \in \mathbb{R}^p : s'x \leq 1 \, \forall x \in P\},$$

which is again a polytope. We now list a number of useful facts about polytopes involving the above definitions, which are used throughout the article. These properties can either be found explicitly or as a straightforward consequence of properties listed in the above mentioned references.

Proposition 4. *Let $P \in \mathbb{R}^p$ be a polytope given by $P = \text{conv}(V)$, where $V = (V_1, \ldots, V_k) \in \mathbb{R}^{p \times k}$, and denote by P^* the dual of P. For simplicity, we assume that $\text{vert}(P) = \{V_1, \ldots, V_k\}$. Moreover, let $0 \in P$. The following properties hold.*

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

2) *For any face F of P, $F = \text{conv}(\text{vert}(P) \cap F)$.*

3) *Let D be an affine line contained in the affine span of P. If $D \cap \text{relint}(P) \neq \varnothing$ then D intersects a proper face of P.*

4) *We can write $P^* = \{s \in \mathbb{R}^p : V's \leq \mathbf{1}_k\}$.*

5) *Any face F^* of P^* can be written as $F^* = \{s \in P^* : V_I's = \mathbf{1}_{|I|}\}$ for some $I \subseteq [k]$.*

6) *Let $I \subseteq [k]$. $F = \text{conv}(V_I)$ is a face of $P \iff F^* = \{s \in P^* : V_I's = \mathbf{1}_{|I|}\}$ is a face of P^*, where I is the maximal index set in this representation.*

   In this case, F^* is the dual of F (and vice versa), and $\text{codim}(F^*) = \text{rk}(V_I)$.

A.3 Facts about subdifferentials of norms with polytope unit balls

We now consider subdifferentials of norms and list several properties in the following. In particular, we show in Proposition 5 that the subdifferential of a norm evaluated at zero is simply given by the unit ball of the corresponding dual norm, a fact that will be used throughout subsequent proofs. Proposition 6 then shows that all faces of this dual norm unit ball can be represented by a subdifferential of the original norm, provided that this norm is such that its unit ball, and therefore also the unit ball of its dual norm, are given by a polytope. Lemma 1 contains a technical result needed for the proof of Theorem 1.

A version of the following proposition – which holds independently of the shape of the unit ball of the norm under consideration – can also be found in Hiriart-Urruty & Lemarechal (1993).

Proposition 5. *Let $\|.\|$ be a norm on $\mathbb{R}^p$, and let $\|.\|^*$ denote the dual norm. Then the following holds.*

1) *The subdifferential of $\|.\|$ at 0 is given by*

$$\partial_{\|.\|}(0) = \{s \in \mathbb{R}^p : \|s\|^* \leq 1\}.$$

2) In general, the subdifferential of $\|.\|$ at x is given by

$$\partial_{\|.\|}(x) = \{s \in \mathbb{R}^p : \|s\|^* \leq 1, s'x = \|x\|\}.$$

Proof. It suffices to show *2)*. By definition, we have

$$\partial_{\|.\|}(x) = \{s \in \mathbb{R}^p : \|v\| \geq \|x\| + s'(v - x)\ \forall v \in \mathbb{R}^p\}$$

Take $s \in \partial_{\|.\|}(x)$. When $v = 0$, we get $s'x \geq \|x\|$. When $v = 2x$, we may deduce that $s'x \leq \|x\|$, implying that $s'x = \|x\|$ must hold. This also implies $\|v\| \geq s'v$ for all $v \in \mathbb{R}^p$, so that $s \in B^*$, yielding

$$\partial_{\|.\|}(x) \subseteq \{s \in B^* : s'x = \|x\|\}.$$

To see that also the converse is true, take any $s \in B^*$ satisfying $s'x = \|x\|$. Now, take any $v \in \mathbb{R}^p$. Clearly $\|v\| \geq s'v = \|x\| + s'(v - x)$, implying that

$$\{s \in B^* : s'x = \|x\|\} \subseteq \partial_{\|.\|}(x).$$

□

Proposition 6. *Let $\|.\|$ be a norm whose unit ball B is the polytope* $\mathrm{conv}(V)$ *for some $V = (V_1 \dots V_k) \in \mathbb{R}^{p \times k}$. Let $F \subseteq B^*$, where B^* is the dual norm unit ball, with $F \neq \varnothing$. Then*

$$F \text{ is a face of } B^* \iff F = \partial_{\|.\|}(x) \text{ for some } x \in \mathbb{R}^p.$$

Proof. ($\Longrightarrow$) If $F = B^*$, then $x = 0$ by Proposition 5. If F is a proper face, we can write $F = \{s \in B^* : V_I's = \mathbf{1}_{|I|}\}$ for some $I \subseteq [k]$, where I is the maximal set satisfying this. Let $x = \sum_{l \in I} V_l$. Since $x/|I| \in \mathrm{conv}(V_I)$, a proper and non-empty face of B, we have $\|x\| = |I|$. Note that for $s \in B^*$, we have $s'V_l \leq 1$, so that

$$s \in \partial_{\|.\|}(x) \iff s'x = \sum_{l \in I} V_l's = \|x\| = |I| \iff V_l's = 1\ \ \forall l \in I \iff s \in F.$$

($\Longleftarrow$) If $F = \partial_{\|.\|}(x)$, then $F = \{s \in B^* : s'x = \|x\|\}$ by Proposition 5. Since $(x/\|x\|)'s \leq 1$ clearly is a valid inequality for all $s \in B^*$, F is a face of B^*. □

Lemma 1. *Let $\|.\|$ be a norm whose unit ball B is the polytope* $\mathrm{conv}(V)$ *for some $V = (V_1 \dots V_k) \in \mathbb{R}^{p \times k}$. Let $F = \{s \in B^* : V_I's = \mathbf{1}_{|I|}\}$ be a face of B^*, the dual norm unit ball, and let I be the maximal set satisfying this. Then the following holds.*

$$F \subseteq \partial_{\|.\|}(b) \implies b \in \mathrm{col}(V_I).$$

Proof. Since $b/\|b\| \in B = \mathrm{conv}(V)$, we can write $b = \sum_{l=1}^k \alpha_l V_l$ with $\alpha_l \geq 0$ and $\sum_{l=1}^k \alpha_l = \|b\|$. Since

$\partial_{\|.\|}(b) = \{s \in B^* : s'b = \|b\|\}$ and $s'V_l \leq 1$, we have for $A = \text{supp}(\alpha)$ and any $s \in \partial_{\|.\|}(b)$

$$\|b\| = s'b = \sum_{l \in A} \alpha_l s' V_l \leq \sum_{l \in A} \alpha_l = \|b\|.$$

This implies that $s'V_l = 1$ for all $l \in \text{supp}(\alpha)$, which, since $F \subseteq \partial_{\|.\|}(b)$, yields $\text{supp}(\alpha) \subseteq I$. □

A.4 Proofs of Theorems 1 and 2

The proofs of Theorems 1 and 2 follow a similar outline, with the proof of Theorem 2 being more accessible. We therefore start with the latter one.

A.4.1 Characterization of BP minimizers and proof of Theorem 2

The following characterization of BP minimizers will prove useful in the following. It can be found in Zhang et al. (2015) and Gilbert (2017), as well as in general form in Mousavi & Shen (2019).

Let $y \in \text{col}(X)$ and let $\hat{\beta}$ satisfy $X\hat{\beta} = y$ then, $\hat{\beta} \in S_{X,\text{bp}}(y)$ if and only if

$$\exists z \in \mathbb{R}^n \text{ such that } \begin{cases} \|X'z\|_\infty \leq 1, \\ X_j' z = \text{sign}(\hat{\beta}_j) \ \ \forall j \in \text{supp}(\hat{\beta}). \end{cases} \tag{5}$$

Proof of Theorem 2.

($\Longleftarrow$) Let us assume that $\text{row}(X)$ intersects a face F of $[-1,1]^p$ whose codimension is larger than $\text{rk}(X)$. We show that one can find some $y \in \text{col}(X)$ for which $S_{X,\text{bp}}(y)$ is not a singleton.

The face F can be written as $F = E_1 \times \cdots \times E_p$, where $E_j \in \{\{-1\}, \{1\}, [-1,1]\}$ for $j \in [p]$. Now, let $J = \{j \in [p] : |E_j| = 1\}$, the set of indices of sets E_j that are singletons. We have $\text{codim}(F) = |J|$ and, by assumption, $|J| > \text{rk}(X)$. Now define $\hat{\beta} \in \mathbb{R}^p$ by setting

$$\hat{\beta}_j = \begin{cases} 1 & E_j = \{1\} \\ -1 & E_j = \{-1\} \\ 0 & j \notin J. \end{cases}$$

Clearly, $\text{supp}(\hat{\beta}) = J$. Set $y = X\hat{\beta}$. Since $\text{row}(X)$ intersects F, there exists $z \in \mathbb{R}^n$ such that $X'z \in F$. This implies that $\|X'z\|_\infty \leq 1$ and $X_j'z = \hat{\beta}_j = \text{sign}(\hat{\beta}_j)$ for any $j \in \text{supp}(\hat{\beta}) = J$. Therefore, by (5), $\hat{\beta} \in S_{X,\text{bp}}(y)$.

To show that $\hat{\beta}$ is not a unique minimizer, we provide $\tilde{\beta} \in \mathbb{R}^p$ with $\tilde{\beta} \neq \hat{\beta}$, $X\tilde{\beta} = y$ and $\|\tilde{\beta}\|_1 = \|\hat{\beta}\|_1$. Since $|J| > \text{rk}(X)$, the columns of X_J are linearly dependent, so that we can pick $h \in \ker(X)$, $h \neq 0$ such that $\text{supp}(h) \subseteq J$ and $\|h\|_\infty < 1$. Since $\|h\|_\infty < 1$, $\text{sign}(\hat{\beta} + h) = \text{sign}(\hat{\beta}) = \hat{\beta}$. Let $\tilde{\beta} = \hat{\beta} + h$. Note that $X\tilde{\beta} = X\hat{\beta} = y$ and that

$$\begin{aligned} \|\tilde{\beta}\|_1 &= \sum_{j=1}^p \text{sign}(\hat{\beta}_j + h_j)(\hat{\beta}_j + h_j) = \sum_{j=1}^p \text{sign}(\hat{\beta}_j)\hat{\beta}_j + \sum_{j \in J} \hat{\beta}_j h_j = \|\hat{\beta}\|_1 + \sum_{j \in J} (X'z)_j h_j \\ &= \|\hat{\beta}\|_1 + z'Xh = \|\hat{\beta}\|_1, \end{aligned}$$

implying that $\tilde{\beta} \in S_{X,\mathrm{bp}}(y)$ also.

($\Longrightarrow$) We assume that $\hat{\beta}, \tilde{\beta} \in S_{X,\mathrm{bp}}(y)$ with $\hat{\beta} \neq \tilde{\beta}$ for some $y \in \mathrm{col}(X)$. We need to show that there exists a face F of $[-1,1]^p$ with $F \cap \mathrm{row}(X) \neq \varnothing$ and $\mathrm{codim}(F) > \mathrm{rk}(X)$. Consider $F = E_1 \times \cdots \times E_p$ and $\tilde{F} = \tilde{E}_1 \times \cdots \times \tilde{E}_p$ with

$$E_j = \begin{cases} \{\mathrm{sign}(\hat{\beta}_j)\} & \text{if } j \in \mathrm{supp}(\hat{\beta}) \\ [-1,1] & \text{if } j \notin \mathrm{supp}(\hat{\beta}) \end{cases} \quad \text{and} \quad \tilde{E}_j = \begin{cases} \{\mathrm{sign}(\tilde{\beta}_j)\} & \text{if } j \in \mathrm{supp}(\tilde{\beta}) \\ [-1,1] & \text{if } j \notin \mathrm{supp}(\tilde{\beta}). \end{cases}$$

Note that for any two minimizers $\hat{\beta}$ and $\tilde{\beta}$, we have $\hat{\beta}_j \tilde{\beta}_j \geq 0$ for all $j \in [p]$, since otherwise $\check{\beta} = (\hat{\beta}+\tilde{\beta})/2$ satisfies $X\check{\beta} = X\hat{\beta} = X\tilde{\beta}$ as well as $\|\check{\beta}\|_1 < \|\hat{\beta}\|_1 = \|\tilde{\beta}\|_1$, which would lead to a contradiction. We therefore have $\mathrm{supp}(\check{\beta}) = \mathrm{supp}(\hat{\beta}) \cup \mathrm{supp}(\tilde{\beta})$. Note that by a convexity argument, $\check{\beta} \in S_{X,\mathrm{bp}}(y)$ also, so that by (5), there exists $\check{z} \in \mathbb{R}^n$ with $\|X'\check{z}\|_\infty \leq 1$ and $X_j'\check{z} = \mathrm{sign}(\check{\beta}_j)$ for all $j \in \mathrm{supp}(\check{\beta})$. Moreover, $X'\check{z} \in F \cap \tilde{F}$ holds. Now, let F_0 be a face of the face $F \cap \tilde{F}$ of smallest dimension that still intersects $\mathrm{row}(X)$. We write $F_0 = E_{0,1} \times \cdots \times E_{0,p}$ and let $J_0 = \{j \in [p] : |E_{0,j}| = 1\}$. Note that $\mathrm{row}(X)$ must intersect F_0 in its relative interior $\mathrm{relint}(F_0)$ where

$$\mathrm{relint}(F_0) = \mathrm{relint}(E_{0,1}) \times \cdots \times \mathrm{relint}(E_{0,p}) \quad \text{where } \mathrm{relint}(E_{0,j}) = \begin{cases} E_{0,j} & j \in J_0 \\ (-1,1) & j \notin J_0, \end{cases}$$

since otherwise $\mathrm{row}(X)$ intersects a proper face of F_0, which contradicts the assumption that F_0 is of minimal dimension. We now need to show that $\mathrm{codim}(F_0) = |J_0| > \mathrm{rk}(X)$. Assume that $|J_0| \leq \mathrm{rk}(X)$. The columns of X_{J_0} are linearly dependent since $X_{J_0}\hat{\beta}_{J_0} = X\hat{\beta} = X\tilde{\beta} = X_{J_0}\tilde{\beta}_{J_0}$ with $\hat{\beta}_{J_0} \neq \tilde{\beta}_{J_0}$, since both $\mathrm{supp}(\hat{\beta})$ and $\mathrm{supp}(\tilde{\beta})$ are subsets of $\mathrm{supp}(\check{\beta}) \subseteq J_0$. We therefore have

$$\dim(\mathrm{col}(X_{J_0})) < |J_0| \leq \mathrm{rk}(X) = \dim(\mathrm{col}(X)) \quad \text{and} \quad \mathrm{col}(X)^\perp \subsetneqq \mathrm{col}(X_{J_0})^\perp.$$

This implies that we can pick $u \in \mathrm{col}(X_{J_0})^\perp \setminus \mathrm{col}(X)^\perp$ so that $X_{J_0}'u = 0$, but $X'u \neq 0$. Pick $z_0 \in \mathbb{R}^n$ with $X'z_0 \in \mathrm{relint}(F_0)$. The affine line $\{X'(z_0 + tu) : t \in \mathbb{R}\} \subseteq \mathrm{row}(X)$ intersects the relative interior $\mathrm{relint}(F_0)$ and is included in the affine span of F_0 by construction of u. Therefore, by Proposition 4, $\mathrm{row}(X)$ intersects a proper face of F_0, yielding a contradiction. □

A.4.2 Characterization of penalized minimizers and proof of Theorem 1

In the particular and well-studied case in which the norm of the penalized problem is the l_1-norm, the solutions to the corresponding optimization problem can be characterized by the Karush-Kuhn-Tucker (KKT) conditions for the LASSO, which can be summarized as follows, see for instance, Bühlmann & Van de Geer (2011).

$$\begin{aligned} \hat{\beta} \in S_{X,\lambda\|.\|_1}(y) \quad &\Longleftrightarrow \quad \|X'(y - X\hat{\beta})\|_\infty \leq \lambda \text{ and } X_j'(y - X\hat{\beta}) = \lambda \mathrm{sign}(\hat{\beta}_j)\ \forall j \in \mathrm{supp}(\hat{\beta}) \qquad (6) \\ &\Longleftrightarrow \quad \|X'(y - X\hat{\beta})\|_\infty \leq \lambda \text{ and } \hat{\beta}'X'(y - X\hat{\beta}) = \lambda\|\hat{\beta}\|_1 \end{aligned}$$

In the above, the supremum-norm is the dual to the l_1-norm. We can generalize the above characterization for solutions to the penalized problem from (1) in the following proposition. Note that in our

notation, the tuning parameter λ is part of the norm $\|.\|$.

Proposition 7. *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$. We have $\hat{\beta} \in S_{X,\|.\|}(y)$ if and only if*

$$\|X'(y - X\hat{\beta})\|^* \leq 1 \text{ and } \hat{\beta}'X'(y - X\hat{\beta}) = \|\hat{\beta}\|.$$

Proof of Proposition 7. Using subdifferential calculus, the proof a straightforward consequence of (3) and Proposition 5.

$$\begin{aligned}\hat{\beta} \in S_{X,\|.\|}(y) &\iff 0 \in X'(X\hat{\beta} - y) + \partial_{\|.\|}(\hat{\beta}) \iff X'(y - X\hat{\beta}) \in \partial_{\|.\|}(\hat{\beta}) \\ &\iff \|X'(y - X\hat{\beta})\|^* \leq 1 \text{ and } \hat{\beta}'X'(y - X\hat{\beta}) = \|\hat{\beta}\|.\end{aligned}$$

□

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

Lemma 2. *Let $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$. Then $X\hat{\beta} = X\tilde{\beta}$ for all $\hat{\beta}, \tilde{\beta} \in S_{X,\|.\|}(y)$.*

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

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

Consequently,

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

which contradicts both β and $\tilde{\beta}$ being minimizers. □

Proof of Theorem 1.

Throughout the proof, let $B = \text{conv}(V)$ with $V = (V_1 \dots V_k) \in \mathbb{R}^{p\times k}$.

($\Longleftarrow$) Assume that there exists a face F of B^* that intersects $\text{row}(X)$ (so that F is non-empty) and satisfies $\text{codim}(F) > \text{rk}(X)$ (so that F is proper). This implies that there exists $I \subseteq [k]$ such that

$$F = \{s \in B^* : V_I's = \mathbf{1}_{|I|}\},$$

where I is the maximal index set satisfying this relationship. Moreover, this implies that $\text{conv}(V_I)$ is a proper, non-empty face of B and that we have $\|s\|^* = 1$ for all $s \in F$ and $\|v\| = 1$ for all $v \in \text{conv}(V_I)$. We show that non-unique solutions exist. Define $\hat{\beta} = \sum_{l\in I} V_l$ and observe that $\|\beta\| = \|I\|\|\sum_{l\in I} V_l/|I|\| = |I|$. Pick $z \in \mathbb{R}^n$ with $X'z$, which exists by assumption, and set $y = X\hat{\beta} + z$. Then $\hat{\beta} \in S_{X,\|.\|}(y)$ by Proposition 7, since

$$\|X'(y - X\hat{\beta})\|^* = \|X'z\|^* = 1 \quad \text{and} \quad \hat{\beta}'(X'(y - X\hat{\beta}) = \hat{\beta}'X'z = \sum_{l\in I} V_l'X'z = |I| = \|\hat{\beta}\|.$$

We now construct $\tilde{\beta} \in S_{X,\|.\|}(y)$ with $\tilde{\beta} \neq \hat{\beta}$. Since $\operatorname{codim}(F_I) = \dim(\operatorname{col}(V_I)) > \operatorname{rk}(X)$, we can pick $h \in \operatorname{col}(V_I) \cap \ker(X)$ with $h \neq 0$. Scale h such that for $h = \sum_{l \in I} c_l V_l$, we have $\max_{l \in I} |c_l| < 1$, and define $\tilde{\beta} = b + h \neq \beta$. Clearly, we have $X\tilde{\beta} = X\beta$. Note that $1 + c_l \geq 0$ and let $\gamma = \sum_{l \in I}(1 + c_l) > 0$. We also have

$$\|\tilde{\beta}\| = \gamma \, \| \sum_{l \in I} \frac{1 + c_l}{\gamma} V_l \| = \gamma = \sum_{l \in I} (1 + c_l) = |I| + \sum_{l \in I} c_l (X'z)' V_l = |I| + (X'z)'h = |I| = \|\hat{\beta}\|,$$

proving that $\tilde{\beta} \in S_{X,\|.\|}(y)$ also.

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

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

Because $X\hat{\beta} = X\tilde{\beta}$ by Lemma 2, one may deduce that $\operatorname{row}(X)$ intersects the face $\partial_{\|.\|}(\hat{\beta}) \cap \partial_{\|.\|}(\tilde{\beta})$. Now, let F^* be a face of $\partial_{\|.\|}(\hat{\beta}) \cap \partial_{\|.\|}(\tilde{\beta})$ of smallest dimension that intersects $\operatorname{row}(X)$ and write

$$F^* = \{s \in B^* : V_I' s = \mathbf{1}_{|I|}\},$$

where I is the largest index set $I \subseteq [k]$ satisfying this relationship. If $\operatorname{codim}(F) = \dim(\operatorname{col}(V_I)) \leq \operatorname{rk}(X)$, consider the following. Note that we can pick $u \in \mathbb{R}^n$ for which $X'u \neq 0$ and $X'u \in \operatorname{col}(V_I)^\perp$. For this, let $I_0 \subseteq I$ be such that the columns of V_{I_0} are linearly independent, and $\operatorname{col}(V_{I_0}) = \operatorname{col}(V_I)$. By Lemma 1, we have $\hat{\beta}, \tilde{\beta} \in \operatorname{col}(V_{I_0})$, so that we get

$$XV_{I_0}\gamma = X\beta = X\tilde{\beta} = XV_{I_0}\tilde{\gamma}$$

with $\gamma \neq \tilde{\gamma}$, implying that the columns of XV_{I_0} are linearly dependent. But this means that

$$\operatorname{rk}(XV_I) = \dim(\operatorname{col}(XV_I)) = \dim(\operatorname{col}(XV_{I_0})) < |I_0| = \dim(\operatorname{col}(V_{I_0})) = \dim(\operatorname{col}(V_I)) \leq \operatorname{rk}(X).$$

Therefore, $\operatorname{col}(XV_I) \subsetneqq \operatorname{col}(X)$ and, consequently, $\operatorname{col}(X)^\perp \subsetneqq \operatorname{col}(XV_I)^\perp$, so that we can pick $u \in \operatorname{col}(XV_I)^\perp \setminus \operatorname{col}(X)^\perp$ for which $X'u \neq 0$ and $X'u \in \operatorname{col}(V_I)^\perp$. Also note that $X'z \in F^*$ for some $z \in \mathbb{R}^n$ and that $X'z$ lies in the relative interior $\operatorname{relint}(F^*)$, as otherwise, $\operatorname{row}(X)$ would intersect a face of $\partial_{\|.\|}(\hat{\beta}) \cap \partial_{\|.\|}(\tilde{\beta})$ of smaller dimension. The affine line $\{X'(z + tu) : t \in \mathbb{R}\} \subseteq \operatorname{row}(X)$ intersects $\operatorname{relint}(F^*)$ and is included in the affine span of F^* by construction. Therefore, by Proposition 4, $\operatorname{row}(X)$ intersects a proper face of F^*, yielding a contradiction. □

A.5 Proof of Proposition 1

We turn to proving Proposition 1. Note that a set is negligible with respect to the Lebesgue measure on $\mathbb{R}^{n \times p}$ if and only if it is negligible with respect to the standard Gaussian measure on $\mathbb{R}^{n \times p}$. Therefore, to establish Proposition 1, it suffices to prove the equality

$$\mathbb{P}_Z \left(\exists y \in \mathbb{R}^n, |S_{Z,\|.\|}(y)| > 1 \right) = 0, \text{ where } Z \in \mathbb{R}^{n \times p} \text{ has iid } \mathcal{N}(0,1) \text{ entries.} \tag{7}$$

Note that $\mathrm{rk}(Z) = \min\{n, p\}$ almost surely. Therefore, when $n \geq p$, $\ker(Z) = 0$ almost surely and $S_{Z,\|.\|}(y)$ is a singleton almost surely. We use the following lemma to establish (7), where $\mathbb{N}$ stands for the (positive) natural numbers.

Lemma 3. *Let $n \in \mathbb{N}$, $q \geq n+1$, and $v \in \mathbb{R}^q$ where $v \neq 0$ is a fixed vector. If $Z = (Z_1, \ldots, Z_n) \in \mathbb{R}^{q \times n}$ has iid $\mathcal{N}(0,1)$ entries, then $\mathbb{P}_Z(v \in \mathrm{col}(Z)) = 0$.*

Proof. We first prove the result for $q = n + 1$. If $v \in \mathrm{col}(Z)$ then

$$\det(Z_1, \ldots, Z_n, v) = 0 \iff \det(Z_1/\|Z_1\|_2, \ldots, Z_n/\|Z_n\|_2, v/\|v\|_2) = 0.$$

Now, because the columns $Z_1/\|Z_1\|_2, \ldots, Z_n/\|Z_n\|_2$ follow a uniform distribution on the l_2-unit sphere, we can deduce that the distribution of the random variable $\det(Z_1/\|Z_1\|_2, \ldots, Z_n/\|Z_n\|_2, v/\|v\|_2)$ is equal to the distribution of $\det(Z_1/\|Z_1\|_2, \ldots, Z_n/\|Z_n\|_2, \zeta/\|\zeta\|_2)$. Here, ζ follows a $\mathbb{N}(0, \mathbb{I}_{n+1})$ distribution, independent from $Z_1, \ldots, Z_n$ as conditioning on $\zeta = v$ does not change the distribution. Finally, the random variable

$$\det(Z_1/\|Z_1\|_2, \ldots, Z_n/\|Z_n\|_2, \zeta/\|\zeta\|_2) = \frac{1}{\|Z_1\|_2 \times \cdots \times \|Z_n\|_2 \times \|\zeta\|_2} \det(Z_1, \ldots, Z_n, \zeta)$$

is non-zero almost surely. This implies $\mathbb{P}_Z(v \in \mathrm{col}(Z)) = 0$. When $q > n+1$, let $I \subseteq [q]$ with $|I| = n+1$ and $v_I \neq 0$. Consequently, $v_I \in \mathrm{col}(\tilde{Z})$, where $\tilde{Z} \in \mathbb{R}^{(n+1)\times n}$ is obtained by keeping the rows of Z with indices in I. Therefore, $P_Z(v \in \mathrm{col}(Z)) \leq P_{\tilde{Z}}(v_I \in \mathrm{col}(\tilde{Z})) = 0$, which concludes the proof. □

Proof of Proposition 1. If $n \leq p$, we are done. If $p > n$, let F_0 be a proper face of B^* such that $\mathrm{codim}(F_0) = q > n$. Note that $0 \notin \mathrm{aff}(F_0)$, the affine space spanned by F_0. There exists $A \in \mathbb{R}^{q \times p}$ with orthonormal rows and $v \in \mathbb{R}^q$, $v \neq 0$ such that $\mathrm{aff}(F_0) = \{x \in \mathbb{R}^p : Ax = v\}$. Since $AA' = \mathbb{I}_p$, $AZ' \in \mathbb{R}^{q \times n}$ has iid $\mathcal{N}(0,1)$ entries. Thus, by Lemma 3, we have

$$\mathbb{P}_Z\left(\mathrm{row}(Z) \cap F_0 \neq \varnothing\right) \leq \mathbb{P}_Z(\mathrm{row}(Z) \cap \mathrm{aff}(F_0) \neq \varnothing) = \mathbb{P}_Z(v \in \mathrm{col}(AZ')) = 0. \tag{8}$$

According to Theorem 1 and since $\mathrm{rk}(Z) = n$ almost surely, the following equalities hold.

$$\begin{aligned}
\mathbb{P}_Z\left(\exists y \in \mathbb{R}^n, |S_{Z,\|.\|}(y)| > 1\right) &= \mathbb{P}_Z\left(\bigcup_{\substack{F \in \mathcal{F}(P) \\ \mathrm{codim}(F) > \mathrm{rk}(Z)}} \{\mathrm{row}(Z) \cap F \neq \varnothing\}\right) \\
&= \mathbb{P}_Z\left(\bigcup_{\substack{F \in \mathcal{F}(P) \\ \mathrm{codim}(F) > n}} \{\mathrm{row}(Z) \cap F \neq \varnothing\}\right) = 0.
\end{aligned}$$

The last equality is a consequence of (8). □

A.6 Proof of Theorem 3

The following lemma generalizes Proposition 4.1 from Gilbert (2017) that is stated for the l_1-norm to an arbitrary norm. This lemma is used in the proof of both Theorem 3 and Theorem 5.

Lemma 4. *Let $s \in \mathbb{R}^p$ and $\|.\|$ be a norm on $\mathbb{R}^p$. The vector space* $\mathrm{row}(X)$ *intersects $\partial_{\|.\|}(s)$ if and only if the following holds.*

$$Xb = Xs \implies \|b\| \geq \|s\| \tag{9}$$

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

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

Then (9) holds for b if and only if s is a minimizer of the function $b \mapsto \|b\| + f_s(b)$. Since we have $\partial_{f_s}(b) = \mathrm{row}(X)$ whenever $Xb = Xs$, we can deduce that the implication (9) occurs if and only if

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

□

Proof of Theorem 3. ($\implies$) Let σ be an accessible sign vector for LASSO. Then there exists $y \in \mathbb{R}^n$ and $\hat{\beta} \in S_{X,\lambda\|.\|_1}(y)$ such that $\mathrm{sign}(\hat{\beta}) = \sigma$. According to the characterization of LASSO minimizers in (6), by setting $z = (y - X\hat{\beta})/\lambda$, one may deduce that $X'z \in F_1(\sigma)$. If σ is an accessible sign vector for BP, there exists $y \in \mathrm{col}(X)$ and $\hat{\beta} \in S_{X,\mathrm{bp}}(y)$ with $\mathrm{sign}(\hat{\beta}) = \sigma$. According to the characterization of BP minimizers in (5), there exisits $z \in \mathbb{R}^n$ such that $X'z \in F_1(\sigma)$. Therefore, $\mathrm{row}(X)$ intersects $F_1(\sigma) = \partial_{\|.\|_1}(\sigma)$ (geometric characterization), or, equivalently, by Lemma 4, whenever $Xb = X\sigma$, we have $\|b\|_1 \geq \|\sigma\|_1$ (analytic characterization).

($\impliedby$) If $\mathrm{row}(X)$ intersects the face $F_1(\sigma)$ (geometric characterization) or, equivalently, if $Xb = X\sigma$ implies $\|b\|_1 \geq \|\sigma\|_1$ (analytic characterization), then there exists $f \in F_1(\sigma)$ and $z \in \mathbb{R}^n$ such that $X'z = f$. Note that $j \in \mathrm{supp}(\sigma)$ implies that $f_j = \sigma_j = \mathrm{sign}(\sigma_j)$. Set $y = \lambda z + X\sigma$ We show that $\sigma \in S_{X,\|.\|_1}(y)$. We have

$$\begin{cases} \|X'(y - X\sigma)\|_\infty = \lambda\|X'z\|_\infty \leq \lambda, \\ X_j'(y - X\sigma) = \lambda X_j'z = \lambda f_j = \lambda\sigma_j = \lambda\mathrm{sign}(\sigma_j) \quad \forall j \in \mathrm{supp}(\sigma), \end{cases}$$

so that according to the characterization of LASSO minimizers in (6), we have $\sigma \in S_{X,\|.\|_1}(y)$, implying that σ is accessible for LASSO. For BP, set $y = X\sigma$ and note that, according to the characterization of BP minimizers in (5), $\sigma \in S_{X,\mathrm{bp}}(y)$, implying that σ is also accessible for BP. □

Corollary 3. *Let $X \in \mathbb{R}^{n\times p}$, $\sigma \in \{-1, 0, 1\}^p$ and assume that $\hat{\beta}$ is the unique solution to the BP problem $S_{X,\mathrm{bp}}(y)$ with $y = X\sigma$. Let $\tilde{\beta} \in \mathbb{R}^p$ satisfy $\|\tilde{\beta} - \hat{\beta}\|_\infty < 1/2$. We then have that*

$$\sigma \textit{ is accessible} \iff \mathrm{sign}(\mathrm{round}(\tilde{b})) = \sigma,$$

where round(.) *rounds componentwise to the nearest integer.*

Proof. ($\Longrightarrow$) If σ is accessible, by the analytic characterization in Theorem 3, $\hat{\beta} = \sigma$. Since $\|\tilde{\beta} - \sigma|_\infty < 1/2$, we get $\text{sign}(\text{round}(\tilde{\beta})) = \text{round}(\tilde{\beta}) = \sigma$.

($\Longleftarrow$) If σ is not accessible, we have $\text{sign}(\hat{\beta}) \neq \sigma$. Using $\|\tilde{\beta} - \hat{\beta}\|_\infty < 1/2$, we can show that

$$F_1(\text{sign}(\hat{\beta})) = \partial_{\|.\|_1}(\text{sign}(\hat{\beta})) \subseteq \partial_{\|.\|_1}(\text{sign}(\text{round}(\tilde{\beta}))) = F_1(\text{sign}(\text{round}(\tilde{\beta}))).$$

Since $\text{row}(X)$ intersects $F_1(\text{sign}(\hat{\beta}))$ by the geometric characterization in Theorem 3, $\text{sign}(\text{round}(\tilde{\beta}))$ is accessible. But then $\text{sign}(\text{round}(\tilde{b})) \neq \sigma$ must hold. □

A.7 Proof of Theorem 4

Theorem 4 states that there is a bijection between the SLOPE models and the faces of the sign permutahedron. The basis for proving this is the fact that the sign permutahedron is the dual of the SLOPE norm unit ball, and that any face of it is given by a subdifferential of the SLOPE norm by Proposition 6.

We start by proving the following proposition which shows that the subdifferential of the SLOPE norm at zero is, indeed, the sign permutahedron, and also characterizes the subdifferential of the SLOPE norm for certain values of x.

Proposition 8. *The subdifferential $F_w(x) = \partial_{\|.\|_w}(x)$ of the SLOPE norm exhibits the following properties.*

1) We have $F_w(0) = P_w^{\pm}$.

2) For any $x \in \mathbb{R}^p$ with $x_1 = \cdots = x_p > 0$, we have $F_w(x) = P_w$.

3) For any $x \in \mathbb{R}^p$ with $x_1 \geq \cdots \geq x_k > x_{k+1} \geq \cdots \geq x_p \geq 0$, we have

$$F_w(x) = F_{w_{[k]}}(x_{[k]}) \times F_{w_{[k+1:p]}}(x_{[k+1:p]}).$$

4) Let $0 < k_1 < \cdots < k_l < p$ be an arbitrary subdivision of $[0:p]$, then for any $x \in \mathbb{R}^p$ with $x_1 = \cdots = x_{k_1} > x_{k_1+1} = \cdots = x_{k_2} > \cdots > x_{k_l+1} = \cdots = x_p \geq 0$, we have $\text{codim}\,(F_w(\text{mdl}(x))) = \|\text{mdl}(x)\|_\infty$ *and*

$$F_w(x) = F_w(\text{mdl}(x)) = \begin{cases} P_{w_{[k_1]}} \times \cdots \times P_{w_{[k_{l-1}+1:k_l]}} \times P_{w_{[k_l+1:p]}} & \text{if } x_p > 0 \\ P_{w_{[k_1]}} \times \cdots \times P_{w_{[k_{l-1}+1:k_l]}} \times P^{\pm}_{w_{[k_l+1:p]}} & \text{if } x_p = 0. \end{cases}$$

Proof. 1) By Proposition 5, we may show that $P_w^{\pm} = B^*$.

($\subseteq$) Take any vertex $W = (\sigma_1 w_{\pi(1)}, \ldots, \sigma_p w_{\pi(p)})'$ of $P_w^{\pm}$ and any $x \in \mathbb{R}^p$ with $\|x\|_w \leq 1$. We have

$$W'x = \sum_{j=1}^{p} \sigma_j w_{\pi(j)} x_j \leq \sum_{j=1}^{p} |x_j| w_{\pi(j)} \leq \sum_{j=1}^{p} w_j |x|_{(j)} = \|x\|_w \leq 1$$

and therefore $W \in B^*$. By convexity, $P_w^{\pm} \subseteq B^*$ follows.

($\supseteq$) Let $a'x \leq b_0$ for some $a \in \mathbb{R}^p$ and $b_0 \in \mathbb{R}$ be a valid inequality of $P_w^{\pm}$. We show that this is a valid inequality of B^* also: Let W be the vertex of $P_w^{\pm}$ defined by $W_j = \operatorname{sign}(a_j) w_{\pi^{-1}(j)}$, where the permutation π satisfies $|a_{\pi(1)}| \geq \cdots \geq |a_{\pi(p)}|$. For any $s \in B^*$, we have

$$a's \leq \|a\|_w = \sum_{j=1}^{p} |a_{\pi(j)}| w_j = \sum_{j=1}^{p} \operatorname{sign}(a_j) a_j w_{\pi^{-1}(j)} = a'W \leq b_0.$$

Since $P_w^{\pm}$ can be written as the (finite) intersection of half-spaces, $P_w^{\pm} \supseteq B^*$ follows.

2) According to Proposition 5 and 1), we have

$$F_w(x) = \left\{ s \in P_w^{\pm} : \sum_{j=1}^{p} s_j = \sum_{j=1}^{p} w_j \right\}.$$

A vertex $W = (\sigma_1 w_{\pi(1)}, \ldots, \sigma_p w_{\pi(p)})'$ of $P_w^{\pm}$ with $\sigma \in \{-1, 1\}^p$ and $\pi \in \mathcal{S}_p$ then fulfills $W \in F_w$ if and only if $\sigma_1 = \cdots = \sigma_p = 1$. Convexity then yields $F_w(x) = P_w$.

3) ($\subseteq$) Let $s \in F_w(x)$. We show that $s_{[k]} \in F_{w_{[k]}}(x_{[k]})$ and $s_{[k+1:p]} \in F_{w_{[k+1:p]}}(x_{[k+1:p]})$. Let $e = \frac{x_k - x_{k+1}}{2} > 0$ and $h \in \mathbb{R}^p$ with $\|h\|_\infty < e$. Since the k largest components of $x+h$ are $\{x_j + h_j\}_{j \in [k]}$, we have

$$\|x + h\|_w = \|(x+h)_{[k]}\|_{w_{[k]}} + \|(x+h)_{[k+1:p]}\|_{w_{[k+1:p]}}.$$

Now, take $h \in \mathbb{R}^p$ such that $\|h\|_\infty < e$ and $h_{k+1} = \cdots = h_p = 0$. Using the above identity and the definition of $F_w(x)$, one may deduce that

$$\begin{aligned} \|(x+h)_{[k]}\|_{w_{[k]}} &= \|x + h\|_w - \|x_{[k+1:p]}\|_{w_{[k+1:p]}} \\ &\geq \|x\|_w + s'h - \|x_{[k+1:p]}\|_{w_{[k+1:p]}} = \|x_{[k]}\|_{w_{[k]}} + \sum_{j=1}^{k} s_j h_j. \end{aligned}$$

We therefore obtain that

$$\|x_{[k]} + h\|_{w_{[k]}} \geq \|x_{[k]}\|_{w_{[k]}} + s'_{[k]} h$$

for all $h \in \mathbb{R}^k$ satisfying $\|h\|_\infty < e$. By (4), we conclude $s_{[k]} \in F_{w_{[k]}}(x_{[k]})$. To show that $s_{[k+1:p]} \in F_{w_{[k+1:p]}}(x_{[k+1:p]})$, one can proceed in a similar manner.

($\supseteq$) For $s \in F_{w_{[k]}}(x_{[k]}) \times F_{w_{[k+1:p]}}(x_{[k+1:p]})$, we clearly have

$$s'x = \sum_{i=1}^{k} s_i x_i + \sum_{i=k+1}^{p} s_i x_i = \|x_{[k]}\|_{w_{[k]}} + \|x_{[k+1:p]}\|_{w_{[k+1:p]}} = \|x\|_w,$$

so that $s \in F_w(x)$ follows.

4) For $x \in \mathbb{R}^p$ with $x_1 = \cdots = x_{k_1} > \cdots > x_{k_l+1} = \cdots = x_p$, $\operatorname{mdl}(x)$ is clearly given by

$$\begin{cases} \operatorname{mdl}(x)_1 = \cdots = \operatorname{mdl}(x)_{k_1} = l+1 > \cdots > \operatorname{mdl}(x)_{k_l+1} = \cdots = \operatorname{mdl}(x)_p = 1 & \text{if } x_p > 0 \\ \operatorname{mdl}(x)_1 = \cdots = \operatorname{mdl}(x))_{k_1} = l > \cdots > \operatorname{mdl}(x)_{k_l+1} = \cdots = \operatorname{mdl}(x)_p = 0 & \text{if } x_p = 0. \end{cases}$$

According to 1), 2) and 3), it is clear that

$$F_w(x) = F_w(\mathrm{mdl}(x)) = \begin{cases} P_{w_{[k_1]}} \times \cdots \times P_{w_{[k_{l-1}+1:k_l]}} \times P_{w_{[k_l+1:p]}} & if x_p > 0 \\ P_{w_{[k_1]}} \times \cdots \times P_{w_{[k_{l-1}+1:k_l]}} \times P^{\pm}_{w_{[k_l+1:p]}} & if x_p = 0. \end{cases}$$

Since the codimension of a permutahedron is equal to 1 (see Maes & Kappen, 1992; Simion, 1997), the one of sign permutahedron is equal to 0, and since the (co-)dimensions of the individual (sign) permutahedra can simply be added up, we have $\mathrm{codim}\,(F_w(x)) = \|\mathrm{mdl}(x)\|_\infty$. □

Proposition 8 lays the groundwork by essentially proving Theorem 4 for all SLOPE models with non-negative and non-decreasing components. We denote this set of models by $\mathcal{M}_p^{\geq,+}$, given by

$$\mathcal{M}_p^{\geq,+} = \{m \in \mathcal{M}_p : m_1 \geq \cdots \geq m_p \geq 0\}.$$

In order to extend this proposition to all SLOPE models in $\mathcal{M}_p$, we introduce the following group of linear transformations.

Definition 4. *Let $\sigma \in \{-1, 1\}^p$, let $\pi \in \mathcal{S}_p$. We define the map*

$$\phi_{\sigma,\pi} : x \in \mathbb{R}^p \mapsto (\sigma_1 x_{\pi(1)}, \ldots, \sigma_p x_{\pi(p)})'$$

and denote by $\mathcal{G} = \{\phi_{\sigma,\pi} : \sigma \in \{-1, 1\}^p, \pi \in \mathcal{S}_p\}$.

The set $\mathcal{G}$ is a finite sub-group of the group of orthogonal transformations on $\mathbb{R}^p$. We list a number of straight-forward properties of $\mathcal{G}$ in the following lemma.

Lemma 5. *Let $x, v \in \mathbb{R}^p$, $\phi \in \mathcal{G}$, and let $\sigma \in \{-1, 1\}^p$ and $\pi \in \mathcal{S}_p$. Then the following holds.*

1) $x'v = \phi(x)'\phi(v)$

2) $\|x\|_w = \|\phi(x)\|_w$

3) $\|x\|_\infty = \|\phi(x)\|_\infty$

4) $\phi(\mathcal{M}_p) = \mathcal{M}_p$ *and* $\phi(P_w^{\pm}) = P_w^{\pm}$

5) $\mathrm{mdl}(\phi(x)) = \phi(\mathrm{mdl}(x))$

6) $\phi_{\sigma,\pi}^{-1} = \phi_{\sigma,\pi^{-1}} \in \mathcal{G}$

7) If, for $m \in \mathcal{M}_p$, $|m_{\pi(1)}| \geq \cdots \geq |m_{\pi(p)}|$ and $\sigma_j m_{\pi(j)} = |m_{\pi(j)}|$ for all $j \in [p]$, then $\phi_{\sigma,\pi}(m) \in \mathcal{M}_p^{\geq,+}$.

Lemma 6. *Let $\phi \in \mathcal{G}$ and $x \in \mathbb{R}^p$. We then have*

$$\phi^{-1}\left(F_w(\phi(x))\right) = F_w(x) \quad \text{and} \quad F_w(\phi(x)) = \phi\left(F_w(x)\right).$$

Proof. The two statements are equivalent, we show the second one. Let $s \in P_w^{\pm}$. Then

$$\begin{aligned} s \in F_w(\phi(x)) &\iff s'\phi(x) = \|\phi(x)\|_w \iff \phi^{-1}(s)'x = \|x\|_w \\ &\iff \phi^{-1}(s) \in F_w(x) \iff s \in \phi(F_w(x)) \end{aligned}$$

by Proposition 5 and Lemma 5. □

We are now equipped to prove Theorem 4.

Proof of Theorem 4.

We start by proving *1)* and *2)* before showing that the map is a bijection.

1) Let $m \in \mathcal{M}_p$ and let $\phi \in \mathcal{G}$ such that $\phi(m) \in \mathcal{M}_p^{\geq,+}$. According to Lemma 6, and because ϕ is an isomorphism on $\mathbb{R}^p$, we have

$$\operatorname{codim}(F_w(m)) = \operatorname{codim}\left(\phi^{-1}\left(F_w(\phi(m))\right)\right) = \operatorname{codim}\left(F_w(\phi(m))\right) = \|\phi(m)\|_\infty = \|m\|_\infty.$$

2) Let $x \in \mathbb{R}^p$ and let $\phi \in \mathcal{M}_p$ such that $\phi(x)_1 \geq \cdots \geq \phi(x)_p \geq 0$. According to Lemma 6 and Proposition 8, the following equalities hold

$$F_w(x) = \phi^{-1}\left(F_w(\phi(x))\right) = \phi^{-1}\left(F_w\left(\operatorname{mdl}(\phi(x))\right)\right) = \phi^{-1}\left(F_w\left(\phi\left(\operatorname{mdl}(x)\right)\right)\right) = F_w(\operatorname{mdl}(x)).$$

We now show that the mapping under consideration is indeed a bijection between $\mathcal{M}_p$ and $\mathcal{F}_0$.

(surjection) According to Proposition 6, a non-empty face of $P_w^{\pm}$ can be expressed as $F_w(x)$ for some $x \in \mathbb{R}^p$. According to *2)* above, we have $F_w(x) = F_w(\operatorname{mdl}(x))$ for $\operatorname{mdl}(x) \in \mathcal{M}_p$.

(injection) Note that Proposition 8 shows that the mapping is injective on $\mathcal{M}_p^{\geq,+}$. To prove that it remains injective on all of $\mathcal{M}_p$, we show that $|\mathcal{M}_p| \leq |\mathcal{F}_0|$. For this, we need several definitions. For $m \in \mathcal{M}_p$, let $\operatorname{stab}_{\mathcal{G}}(m) = \{\phi \in \mathcal{G} : \phi(m) = m\}$ and $\operatorname{orb}_{\mathcal{G}}(m) = \{\phi(m) : \phi \in \mathcal{G}\}$, the stabilizer and orbit of m, respectively, with respect to $\mathcal{G}$. For $m \in \mathcal{M}_p$, there exists $\phi \in \mathcal{G}$ such that $\phi(m) \in \mathcal{M}_p^{\geq,+}$. Therefore, the orbit-stabilizer formula gives

$$\mathcal{M}_p = \bigcup_{m \in \mathcal{M}_p^{\geq,+}} \operatorname{orb}_{\mathcal{G}}(m) \implies |\mathcal{M}_p| \leq \sum_{m \in \mathcal{M}_p^{\geq,+}} |\operatorname{orb}_{\mathcal{G}}(m)| = \sum_{m \in \mathcal{M}_p^{\geq,+}} \frac{|\mathcal{G}|}{|\operatorname{stab}_{\mathcal{G}}(m)|}.$$

We also look at stabilizer and orbit when $\mathcal{G}$ operates on $\mathcal{F}_0$. For a face $F \in \mathcal{F}_0$, let $\operatorname{stab}_{\mathcal{G}}(F) = \{\phi \in \mathcal{G} : \phi(F) = F\}$ and $\operatorname{orb}_{\mathcal{G}}(F) = \{\phi(F) : \phi \in \mathcal{G}\}$. We first show that if $\operatorname{orb}_{\mathcal{G}}(F_w(m)) \cap \operatorname{orb}_{\mathcal{G}}(F_w(\tilde{m})) \neq \varnothing$ for some $m, \tilde{m} \in \mathcal{M}_p^{\geq,+}$, $m = \tilde{m}$ follows. Let us assume that $F_w(\tilde{m}) = \phi(F_w(m))$ for some $\phi \in \mathcal{G}$. Note that $\phi(F_w(m)) = F_w(\phi(m))$ by Lemma 6. Since $w \in F_w(m)$ and $w \in F_w(\tilde{m}) = F_w(\phi(m))$, we have

$$w'm = \|m\|_w = \|\phi(m)\|_w = w'\phi(m),$$

where the second-last equality holds by Lemma 5 and the last equality holds since $m \in \mathcal{M}_p^{\geq,+}$. Now, if $\phi(m) \neq m$, $\phi(m)'m < \|m\|_w$ follows since the components of w are positive and strictly decreasing. But that would contradict the above, so $\phi(m) = m$ must hold. Consequently, $F_w(\tilde{m}) = F_w(m)$, which in turn implies $\tilde{m} = m$ by Proposition 8.

Now, let $m \in \mathcal{M}_p^{\geq,+}$ and let us show that $\text{stab}_\mathcal{G}(m) = \text{stab}_\mathcal{G}(F_w(m))$. The inclusion $\text{stab}_\mathcal{G}(m) \subseteq \text{stab}_\mathcal{G}(F_w(m))$ immediately follows from

$$\begin{aligned}\phi \in \text{stab}_\mathcal{G}(m) &\implies F_w(m) = \phi^{-1}(F_w(\phi(m))) = \phi^{-1}(F_w(m)) \implies \phi(F_w(m)) = F_w(m) \\ &\implies \phi \in \text{stab}_\mathcal{G}(F_w(m)).\end{aligned}$$

To show $\text{stab}_\mathcal{G}(F_w(m)) \subseteq \text{stab}_\mathcal{G}(m)$, let $\phi \in \text{stab}_\mathcal{G}(F_w(m))$ and note that $F_w(m) = \phi(F_w(m)) = F_w(\phi(m))$. Since $m \in \mathcal{M}_p^{\geq,+}$, this implies that $w \in F_w(m) = F_w(\phi(m))$, so that the same reasoning as above yields $m = \phi(m)$ and $\phi \in \text{stab}_\mathcal{G}(m)$.

To conclude, note that since the orbits $\text{orb}_\mathcal{G}(F_w(m))$ with $m \in \mathcal{M}_p^{\geq,+}$ are disjoint, and since $\text{stab}_\mathcal{G}(m) = \text{stab}_\mathcal{G}(F_w(m))$, we may deduce that

$$|\mathcal{M}_p| \leq \sum_{m \in \mathcal{M}_p^{\geq,+}} \frac{|\mathcal{G}|}{|\text{stab}_\mathcal{G}(F_w(m))|} = \sum_{m \in \mathcal{M}_p^{\geq,+}} \left|\text{orb}_\mathcal{G}(F_w(m))\right| = \Big| \bigcup_{m \in \mathcal{M}_p^{\geq,+}} \text{orb}_\mathcal{G}\left(F_w(m)\right) \Big| \leq |\mathcal{F}_0|.$$

□

A.8 Proof of Theorem 5

Proof. ($\implies$) If m is an accessible SLOPE model, then

$$\exists y \in \mathbb{R}^n, \exists \hat{\beta} \in S_{X,\|.\|_w}(y) \text{ such that } \text{mdl}(\hat{\beta}) = m.$$

By Theorem 4, we may deduce that $\partial_{\|.\|_w}(\hat{\beta}) = F_w(\hat{\beta}) = F_w(m)$. Consequently,

$$0 \in X'(X\hat{\beta} - y) + \partial_{\|.\|_w}(\hat{\beta}) = F_m \implies X'(y - X\hat{\beta}) \in F_w(m).$$

Therefore, $\text{row}(X)$ intersects $F_w(m)$ (geometric characterization), or, equivalently, by Lemma 4, whenever $Xb = Xm$ we have $\|b\|_w \geq \|m\|_w$ (analytic characterization).

($\impliedby$) If $\text{row}(X)$ intersects the face $F_w(m)$ (geometric characterization), or, equivalently, whenever $Xb = Xm$ we have $\|b\|_w \geq \|m\|_w$ (analytic characterization), there exists $z \in \mathbb{R}^n$ such that $X'z = f \in F_w(m)$. We set $y = z + Xm$ and show that $m \in S_{X,\|.\|_w}(y)$. We have

$$\|X'(y - Xm)\|_w^* = \|f\|_w^* \leq 1 \text{ and } m'X'(y - Xm) = m'f = \|m\|_w,$$

which, by Proposition 7, yields $m \in S_{X,\|.\|_w}(y)$. □

A.9 Proof of Proposition 2

Proof. By Theorem 5, we know that

$$m \in \mathcal{M}_p \text{ is accessible} \iff \text{row}(X) \cap F_w(m) \neq \varnothing \iff \exists f \in \mathbb{R}^n : X'f \in F_w(m) \iff f \in N_w(m),$$

which proves the first statement. Now, let $y = f + Xb$, where $f \in N_w(m)$ and $b \in \mathbb{R}^p$ such that $\mathrm{mdl}(b) = m$. Note that

$$\|X'(y - Xb)\|_w^* = \|X'f\|_w^* \leq 1 \text{ and } b'X'(y - Xb) = b'X'f = \|b\|_w^*,$$

where the first inequality holds since $X'f \in F_w(m)$, a face of $P_w^{\pm}$, and the latter one by applying Proposition 6 after noticing that $X'f \in F_w(m) = F_w(b) = \partial_{\|.\|_w}(b)$ by Theorem 4. Proposition 7 then yields $b \in S_{X,\|.\|_w}(y)$, so that $y \in A_w(m)$.

Conversely, let $y \in A_w(m)$ and let $\hat{\beta} \in S_{X,\|.\|_w}(y)$ so that $\mathrm{mdl}(\hat{\beta}) = m$. Then $y - X\hat{\beta} \in N_w(m)$ since by Proposition 7, we have

$$X'(y - X\hat{\beta}) \in \partial_{\|.\|_w}(\hat{\beta}) = F_w(m),$$

where the last equality holds by Theorem 4. □

A.10 Proof of Proposition 3

Proof. Note that by Proposition 7 we have that $\hat{\beta} \in S_{X,\|.\|}(y)$ if and only if we have

$$\|X'(y - X\hat{\beta})\|^* \leq 1 \text{ and } \hat{\beta}'(y - X\hat{\beta}) = \|\hat{\beta}\|.$$

Consequently, when $\|X'u\|^* \leq 1$ it is clear that $0 \in S_{X,\|.\|}(u)$ implying that $S_{X,\|.\|}(u) = \{0\}$ as all elements of $S_{X,\|.\|}(u)$ must have the same norm. Now, let $u \in A_{\varnothing}$ and remember that $\hat{u} = y - X\hat{\beta}$. The following inequality

$$(y - \hat{u})'(u - \hat{u}) = \underbrace{\hat{\beta}'X'u}_{\leq \|\hat{\beta}\|} - \underbrace{\hat{\beta}'X'(y - X\hat{\beta})}_{= \|\hat{\beta}\|} \leq 0$$

shows that, indeed, $\hat{u}$ is the projection of y onto the convex null set $A_{\varnothing}$. □

References

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

Alliney, S. & Ruzinsky, A. (1994). An algorithm for the minimization of mixed l_1 and l_2 norms with applications to bayesian estimation. *IEEE Transactions on Signal Processing* **42**, 618–627.

Bach, F., Jenatton, R., Mairal, J. & Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* **4**, 1–106.

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

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

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

Candès, E., Romberg, J. & Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics* **59**, 1207–1223.

Chen, S. & Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, vol. 1.

Cohen, A., Dahmen, W. & DeVore, R. (2009). Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society* **22**, 211–231.

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

Descloux, P. & Sardy, S. (2018). Model selection with lasso-zero: Adding straw to the haystack to better find needles. Tech. Rep. 1805.05133, arXiv.

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

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

Dupuis, X. & Vaiter, S. (2019). The geometry of sparse analysis regularization. Tech. Rep. 1907.01769, arXiv.

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

Ghaoui, L. E., Viallon, V. & Rabbani, T. (2012). Safe feature elimination in sparse supervised learning. *Pacific Journal of Optimization* **8**, 667–698.

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

Golub, G., Heath, M. & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

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

Haupt, J., Bajwa, W., Raz, G. & Nowak, R. (2010). Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory* **56**, 5862–5875.

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

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **12**, 55–67.

Kremer, P., Brzyski, D., Bogdan, M. & Paterlini, S. (2019). Sparse index clones via the sorted ℓ_1-norm. Tech. Rep. 3412061, Social Science Research Network.

Kremer, P. J., Lee, S., Bogdan, M. & Paterlini, S. (2020). Sparse portfolio selection via the sorted ℓ_1-norm. *Journal of Banking and Finance* **110**, 105687.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference with an application to the Lasso. *Annals of Statistics* **44**, 907–927.

Maes, M. & Kappen, B. (1992). On the permutahedron and the quadratic placement problem. *Philips Journal of Research* **46**, 267–292.

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

Mousavi, S. & Shen, J. (2019). 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.

Negahban, S., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.

Negahban, S. N. & Wainwright, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of $l_{1,\infty}$-regularization. In *21st International Conference on Neural Information Processing Systems*.

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

Rauhut, H. (2010). Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery* **9**, 1–92.

Romberg, J. (2009). Compressive sensing by random convolution. *SIAM Journal of Imaging Sciences* **2**, 1098–1128.

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

Saligrama, V. & Zhao, M. (2011). 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.

Sepehri, A. & Harris, N. (2017). The accessible lasso models. *Statistics* **51**, 711–721.

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

Simion, R. (1997). Convex polytopes and enumeration. *Advances in Applied Mathematics* **18**, 149–180.

Tardivel, P. & Bogdan, M. (2018). On the sign recovery by lasso, thresholded lasso and thresholded basis pursuit denoising. Tech. Rep. 1812.05723, arxiv.

Tardivel, P., Servien, R. & Concordet, D. (2018). Sparsest representations and approximations of an underdetermined linear system. *Inverse Problems* **34**.

Tardivel, P., Servien, R. & Concordet, D. (2020). Simple expressions of the LASSO and SLOPE estimators in small-dimension. *Statistics* **54**, 340–352.

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

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

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

Tibshirani, R. J. & Taylor, J. (2012). Degrees of freedom in lasso problems. *Annals of Statistics* **40**, 1198–1232.

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

Wang, J., Zhou, J., Wonka, P. & Ye, J. (2013). Lasso screening rules via dual polytope projection. In *Advances in Neural Information Processing Systems*.

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

Zhang, H., Yin, W. & Cheng, L. (2015). Necessary and sufficient conditions of solution uniqueness in 1-norm minimization. *Journal of Optimization Theory and Applications* **164**, 109–122.

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

Ziegler, G. (2012). *Lectures on Polytopes*, vol. 152. New York: Springer.

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