Full Length Article

Pattern recovery by SLOPE

Małgorzata Bogdan [a,b], Xavier Dupuis [c], Piotr Graczyk [d], Bartosz Kołodziejek [e], Tomasz Skalski [d,f,*], Patrick Tardivel [c], Maciej Wilczyński [f]

[a] *Institute of Mathematics, University of Wrocław, pl. Grunwaldzki 2/4, Wrocław, 50-384, Poland*
[b] *Department of Statistics, Lund University, Holger Crafoords Ekonomicentrum 1, Tycho Brahes väg 1, Lund, SE-220 07, Sweden*
[c] *Université Bourgogne Europe, CNRS, IMB UMR 5584, Dijon, 21000, France*
[d] *Laboratoire de Mathématiques LAREMA, Université d'Angers, 2 Boulevard Lavoisier, Angers, 49045, France*
[e] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, Warsaw, 00-662, Poland*
[f] *Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, Wrocław, 50-370, Poland*

ARTICLE INFO

ABSTRACT

SLOPE is a popular method for dimensionality reduction in high-dimensional regression. Its estimated coefficients can be zero, yielding sparsity, or equal in absolute value, yielding clustering. As a result, SLOPE can eliminate irrelevant predictors and identify groups of predictors that have the same influence on the response. The concept of the SLOPE pattern allows us to formalize and study its sparsity and clustering properties. In particular, the SLOPE pattern of a coefficient vector captures the signs of its components (positive, negative, or zero), the clusters (groups of coefficients with the same absolute value), and the ranking of those clusters. This is the first paper to thoroughly investigate the consistency of the SLOPE pattern. We establish necessary and sufficient conditions for SLOPE pattern recovery, which in turn enable the derivation of an irrepresentability condition for SLOPE given a fixed design matrix X. These results lay the groundwork for a comprehensive asymptotic analysis of SLOPE pattern consistency.

1. Introduction

High-dimensional data is currently ubiquitous in many areas of science and industry. Efficient extraction of information from such data sets often requires dimensionality reduction based on identifying the low-dimensional structure behind the data generation process. In this article we focus on a particular statistical model describing the data: the linear regression model

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where $Y \in \mathbb{R}^n$ is a vector of responses, $X \in \mathbb{R}^{n \times p}$ is a design matrix, $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients and $\varepsilon \in \mathbb{R}^n$ is a random noise.

It is well-known that the classical least squares estimator of β is BLUE (the best linear unbiased estimator) when the design matrix X is of full column rank. However, it is also well-known that this estimator often exhibits a large variance and a large mean squared estimation error, especially when p is large or when the columns of X are strongly correlated. Moreover, it is not uniquely determined

---

* Corresponding author.
E-mail addresses: malgorzata.bogdan@uwr.edu.pl (M. Bogdan), xavier.dupuis@u-bourgogne.fr (X. Dupuis), piotr.graczyk@univ-angers.fr (P. Graczyk), bartosz.kolodziejek@pw.edu.pl (B. Kołodziejek), tomasz.skalski@pwr.edu.pl (T. Skalski), patrick.tardivel@u-bourgogne.fr (P. Tardivel), maciej.wilczynski@pwr.edu.pl (M. Wilczyński).

when $p > n$. Therefore, scientists often resort to the penalized least squares estimators of the form,

$$\hat{\beta} = \arg \min_{b \in \mathbb{R}^p} \left\{ \|Y - Xb\|_2^2 + C \operatorname{pen}(b) \right\}, \tag{1.2}$$

where $C > 0$ and pen is the penalty on the model complexity. Typical examples of the penalties include $\operatorname{pen}(\beta) = \ell_0(\beta) = \#\{i : \beta_i \neq 0\}$, which appears in popular model selection criteria such as AIC [1], BIC [2], RIC [3], mBIC [4] or EBIC [5], or the ℓ_2 or ℓ_1 norms, resulting in famous ridge [6,7] or LASSO [8,9] estimators. In cases where the penalty function is not differentiable, penalized estimators usually possess the dimensionality reduction properties as illustrated *e.g.* in [10]. For instance, LASSO may yield some zero components [11,12] and thus its dimensionality reduction property is straightforward: elimination of irrelevant predictors.

However, in a variety of applications one is interested not only in eliminating variables which are not important but also in merging similar values of regression coefficients. The prominent statistical example is the multiple regression with categorical variables at many levels, where one may substantially reduce the model dimension and improve the estimation and prediction properties by merging regression coefficients corresponding to "similar" levels (see *e.g.* [13–17]). Another well-known example of advantages resulting from merging different model parameters are modern Convolutional Neural Networks (CNN), where the "parameter sharing" has allowed to "dramatically lower the number of unique model parameters and to significantly increase network sizes without requiring a corresponding increase in training data" [18].

In this article, we explore the dimensionality reduction properties of the well-known convex optimization method, the Sorted L-One Penalized Estimator (SLOPE) [19–21]. SLOPE has gained considerable attention due to its rich statistical properties (see, *e.g.*, [20,22–24] for false discovery rate control under various settings, and [25–27] for results on the minimax rates of estimation and prediction).

Following [19,20], we define the SLOPE estimator as the solution to the optimization problem

$$\min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - Xb\|_2^2 + \sum_{i=1}^p \lambda_i |b|_{(i)} \right\}, \tag{1.3}$$

where $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$ denote the absolute values of the components of b sorted in nonincreasing order, and $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ is a sequence of tuning parameters satisfying $\lambda_1 > 0$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

The SLOPE estimator is arguably the most significant penalized estimator developed in recent years. It can be viewed as an extension of the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) [28], where the tuning parameter Λ has components that decrease arithmetically. It is also closely related to the Pairwise Absolute Clustering and Sparsity (PACS) method [29]. In this context, the term "clustering" reflects the fact that some components of the OSCAR, PACS, and SLOPE estimators may have the same absolute value, while the terms "sparsity" and "shrinkage" indicate that some components of these estimators can be exactly zero.

SLOPE is also an extension of LASSO whose penalty term is $\lambda \| \cdot \|_1$ (*i.e.*, when $\Lambda = (\lambda, \ldots, \lambda)'$ with $\lambda > 0$). Note that contrarily to SLOPE with a decreasing sequence Λ, LASSO does not exhibit clusters. Clustering and sparsity properties for both OSCAR and SLOPE are intuitively illustrated by drawing the elliptic contour lines of the residual sum of squares $b \mapsto \|Y - Xb\|_2^2$ (when $\ker(X) = \{0\}$) together with the balls of the sorted ℓ_1 norm (see, *e.g.*, Figure 2 in [28], Figure 1 in [21] or Figure 3 in [30]). Known theoretical properties of SLOPE include its ability to cluster correlated predictors [28,31], as well as predictors with a similar influence on the L_2 loss function [32]. Specifically, when X is orthogonal, SLOPE may also cluster components of β equal in absolute value [33]. Therefore, dimensionality reduction properties of SLOPE are due to elimination of irrelevant predictors and grouping predictors having the same influence on Y. Note that, contrary to fused LASSO [34], a cluster for SLOPE does not have, in broad generality, adjacent components.

The clustering properties of SLOPE offer several advantages. One of the most important is its ability to reduce the problem's dimensionality from p to the number of clusters, thereby lowering variance and enhancing the stability of the estimator. The practical benefits of these clustering effects have been demonstrated, for example, in [32], where SLOPE proved effective for sparse portfolio selection. In this setting, SLOPE regularization not only yields sparse and well-diversified portfolios but also improves out-of-sample performance and reduces trading costs by minimizing portfolio turnover. Unlike LASSO, which encourages sparsity but may inconsistently handle similar assets, SLOPE promotes the grouping of assets with comparable risk-return profiles, resulting in portfolios that are both sparse and structured.

These theoretical and practical strengths highlight the importance of developing a rigorous mathematical foundation for the clustering properties of the SLOPE estimator — a goal this article seeks to accomplish.

The key concept for analyzing the clustering properties of SLOPE is the SLOPE pattern, which was first introduced in [35]. It allows to describe the structure (sparsity and clusters) induced by SLOPE. The SLOPE pattern extracts from a given vector:

a) The sign of each component (positive, negative, or zero),
b) The clusters (*i.e.*, indices of components with equal absolute values),
c) The hierarchy among the clusters.

The notion of a SLOPE pattern is stronger and substantially more informative than various other structures, such as the model subspace [10,36] or the sets of irrelevant or clustered components [29]. Specifically, two vectors that share the same SLOPE pattern also share the same model subspace and have identical sets of zero components as well as components equal in absolute value.

Note that for a given regression model Eq. (1.1) the SLOPE pattern depends on relative scaling of different variables. In the situations where there are no clear reasons or rules for selection of specific measurement units, we suggest defining the SLOPE

pattern with respect to the standardized design matrix. Note that standardizing explanatory variables is also a standard solution for a similar problem of scale-dependent definition of principal components in PCA.

This article focuses on recovering the pattern of β by SLOPE. From a mathematical perspective, the main result is Theorem 3.1, which specifies two conditions (named positivity and subdifferential conditions) characterizing pattern recovery by SLOPE in both noisy and noiseless settings. A byproduct of Theorem 3.1 is the SLOPE irrepresentability condition: a necessary and sufficient condition for pattern recovery in the noiseless case. The word "irrepresentability" is a tribute to works written a decade ago on sign recovery by LASSO [37–41]. However, when deriving the irrepresentability condition for SLOPE we developed a substantially different mathematical framework, which paves the way for similar analyses of other penalized estimators. Even in the case of the LASSO (see Remark 3.2), the sign recovery characterization provided by Theorem 3.1 is new and could simplify the proofs of well-known results regarding the LASSO irrepresentability condition. Furthermore, Theorem 3.1 provides a sufficient, though not necessary, condition under which a SLOPE solution shares the same model subspace for the sorted ℓ_1 norm as β, and correctly identifies the sets of irrelevant or clustered components of β. In this way we strengthen the results of [10,29,36]. Finally, the proposed positivity and subdifferential conditions are crucial in developing an algorithm for computing the solution path of SLOPE [42] or to study pattern recovery by proximal-thresholded SLOPE [43, Theorem 2.2].

In Theorem 4.1 we consider a noisy case and under the open SLOPE irrepresentability condition (a condition slightly stronger than the SLOPE irrepresentability condition) we prove that the probability of pattern recovery by SLOPE tends to 1 as soon as X is fixed and gaps between distinct absolute values of β diverge to infinity. Additionally, in Theorems 4.2 and 4.3 we apply the SLOPE irrepresentability condition to derive results on the asymptotic pattern recovery by SLOPE when the number of variables p is fixed and the sample size n diverges to infinity.

While the SLOPE ability to identify the pattern of the vector of regression coefficients β is interesting by itself, the related reduction of model dimension also brings an advantage in terms of precision of β estimation. This phenomenon is illustrated in Fig. 1, which presents the difference in precision of LASSO, Fused LASSO and SLOPE estimators, when some of the regression coefficients are equal to each other. In this example $n = 100$, $p = 200$, and the rows of the design matrix are generated as independent binary Markov chains, with $\mathbb{P}(X_{i1} = 1) = \mathbb{P}(X_{i1} = -1) = 0.5$ and $\mathbb{P}(X_{i(j+1)} \neq X_{ij}) = 1 - \mathbb{P}(X_{i(j+1)} = X_{ij}) = 0.0476$. This value corresponds to the probability of the crossover event between genetic markers spaced every 5 centimorgans and our design matrix can be viewed as an example of 100 independent haplotypes, each resulting from a single meiosis event. In this example, the correlation between columns of the design matrix decays exponentially, $\rho(X_{.i}, X_{.j}) = 0.9048^{|i-j|}$. The design matrix is then standardized, so that each column has a zero mean and a unit variance, and the response variable is generated according to the linear model Eq. (1.1) with $\beta_1 = \ldots = \beta_{30} = 40$, $\beta_{31} = \ldots = \beta_{200} = 0$ and $\sigma = 5$. In this experiment the data matrix X and the regression model are constructed such that the LASSO irrepresentability condition holds. The tuning parameter for LASSO is selected as the smallest value of λ for which LASSO can properly identify the sign of β. Similarly, the tuning parameter Λ is designed such that the SLOPE irrepresentability condition holds and Λ is multiplied by the smallest constant for which SLOPE properly returns the SLOPE pattern. The selected tuning parameters for LASSO and SLOPE are represented in the left panel of Fig. 1. For both LASSO and SLOPE, the proposed tuning parameters are close to the values minimizing the mean squared estimation error. The fused LASSO was performed using the *fusedlasso* function from the *genlasso* library in *R*. The tuning parameters λ and γ were manually selected, so as to minimize the mean squared estimation error. Since in

tuning parameters
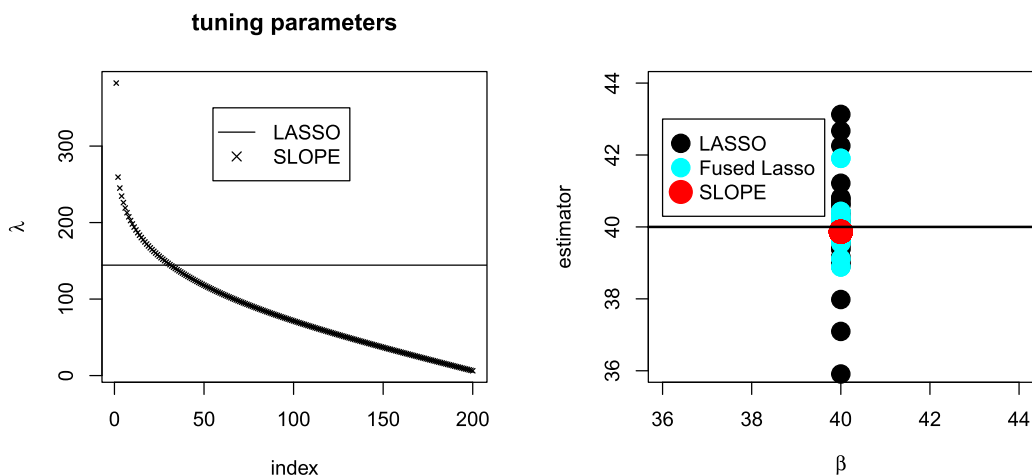
Fig. 1. Comparison of LASSO, Fused LASSO and SLOPE when the cluster structure is present in the data. Here $n = 100$, $p = 200$, the rows of X matrix are simulated as independent binary Markov chains, with the transition probability 0.0476 (corresponding to 5 centimorgans genetic distance). The correlation between i^{th} and j^{th} column of X decays exponentially as $0.9048^{|i-j|}$. First $k = 30$ columns of X are associated with Y and their nonzero regression coefficients are all equal to 40 (other details are provided in the text). Left panel represents the value of the tuning parameter for LASSO (solid line) and the sequence of tuning parameters for SLOPE (crosses). The sequences are selected such that both LASSO and SLOPE recover their corresponding patterns with a minimal bias. Right panel represents LASSO, Fused LASSO and SLOPE estimates. The tuning parameters for Fused LASSO were selected manually as to minimize the estimation error.

this example all methods properly estimate the null components of β, the right panel in Fig. 1 illustrates only the accuracy of the estimation of the nonzero coefficients. Here we can observe that the SLOPE ability to identify the cluster structure leads to superior estimation properties. SLOPE estimates the regression coefficient vector β with virtually no error, while the LASSO estimates range roughly between 36 and 44, and the Fused LASSO estimates fall between approximately 38.8 and 41.9. As a result, the squared error of the SLOPE estimate is more than twenty times smaller than that of the Fused LASSO, and over 100 times smaller than that of the LASSO (0.53 *vs* 13.25 *vs* 63.4).

1.1. Structure of the paper

Section 2 introduces the concept of the SLOPE pattern, which captures the sparsity and clustering structure in the regression coefficients. It also defines key mathematical tools such as the pattern matrix, clustered design matrix, and subdifferential of the sorted ℓ_1 norm.

Theorem 3.1, the main result of Section 3, provides necessary and sufficient conditions for exact SLOPE pattern recovery in both noisy and noiseless settings. This extends previous work by introducing a "SLOPE irrepresentability condition," generalizing the LASSO irrepresentability framework. Even for the LASSO case (see Remark 3.2), the sign recovery characterization in Theorem 3.1 is new and may simplify existing proofs.

Section 4 investigates the conditions under which SLOPE recovers the true pattern, either as the signal strength increases or as the sample size grows, given appropriate tuning. It introduces the notion of open irrepresentability, a stronger version of the standard irrepresentability condition, which guarantees asymptotic pattern recovery in high-dimensional settings.

Section 5 presents simulation studies that validate the theoretical results from Sections 3 and 4. The simulations demonstrate that appropriate tuning of the SLOPE penalty yields high probabilities of correct pattern recovery, and they show that SLOPE outperforms LASSO and Fused LASSO when the true regression vector exhibits clustering.

We conclude the paper with a discussion in Section 6.

The appendix provides detailed proofs of the main theorems. It also includes auxiliary results such as a law of iterated logarithm for strong consistency and computational verifications of irrepresentability conditions.

2. Preliminaries and basic notions on clustering properties by SLOPE

The SLOPE pattern, whose definition is recalled hereafter, is the central notion in this article.

Definition 2.1. Let $b \in \mathbb{R}^p$. The SLOPE pattern of b, $\mathrm{patt}(b)$, is defined by

$$\mathrm{patt}(b)_i = \mathrm{sign}(b_i)\,\mathrm{rank}(|b|)_i, \quad \forall i \in \{1, \ldots, p\}$$

where $\mathrm{rank}(|b|)_i \in \{0, 1, \ldots, k\}$, k is the number of nonzero distinct values in $\{|b_1|, \ldots, |b_p|\}$, $\mathrm{rank}(|b|)_i = 0$ if $b_i = 0$, $\mathrm{rank}(|b|)_i > 0$ if $|b_i| > 0$ and $\mathrm{rank}(|b|)_i < \mathrm{rank}(|b|)_j$ if $|b_i| < |b_j|$.

We denote by $\mathcal{P}_p^{\mathrm{SLOPE}} = \mathrm{patt}(\mathbb{R}^p)$ the set of SLOPE patterns.

Example 2.2. For $a = (4.7, -4.7, 0, 1.8, 4.7, -1.8)'$ we have $\mathrm{patt}(a) = (2, -2, 0, 1, 2, -1)'$. For $b = (1.2, -2.3, 3.5, 1.2, 2.3, -3.5)'$ we have $\mathrm{patt}(b) = (1, -2, 3, 1, 2, -3)'$.

Definition 2.3. Let $0 \neq M = (M_1, \ldots, M_p)' \in \mathcal{P}_p^{\mathrm{SLOPE}}$ with $k = \|M\|_\infty$ nonzero clusters. The pattern matrix $U_M \in \mathbb{R}^{p \times k}$ is defined as follows

$$(U_M)_{ij} = \mathrm{sign}(M_i)\mathbf{1}_{(|M_i|=k+1-j)}, \qquad i \in \{1, \ldots, p\}, j \in \{1, \ldots, k\}.$$

Hereafter, the notation $|M|_\downarrow = (|M|_{(1)}, \ldots, |M|_{(p)})'$ represents the components of M ordered non-increasingly by absolute value.

Example 2.4. If $M = (-2, 1, 0, -1, 2)'$, then

$$U_M = \begin{pmatrix} -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 \end{pmatrix}' \text{ and } U_{|M|_\downarrow} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}'.$$

For $k \geq 1$ we denote by $\mathbb{R}^{k+} = \{\kappa \in \mathbb{R}^k : \kappa_1 > \ldots > \kappa_k > 0\}$. Definition 2.3 implies that for $0 \neq M \in \mathcal{P}_p^{\mathrm{SLOPE}}$ and $k = \|M\|_\infty$, for $b \in \mathbb{R}^p$ we have

$$\mathrm{patt}(b) = M \iff \text{there exists } \kappa \in \mathbb{R}^{k+} \text{ such that } b = U_M \kappa.$$

2.1. Clustered matrix $\tilde{X}_M$ and clustered parameter $\tilde{\Lambda}_M$

Definition 2.5. Let $X \in \mathbb{R}^{n \times p}$, $\Lambda \in \mathbb{R}^{p+}$ and $M \in \mathcal{P}_p^{\mathrm{SLOPE}}$. The clustered matrix is defined by $\tilde{X}_M = XU_M$. The clustered parameter is defined by $\tilde{\Lambda}_M = (U_{|M|_\downarrow})'\Lambda$.

If $M = \mathrm{patt}(\beta)$ for $\beta \in \mathbb{R}^p$ satisfies $\|M\|_\infty < p$, then the pattern $M = (M_1, \ldots, M_p)'$ leads naturally to reduce the dimension of the design matrix X in the regression problem, by replacing X by $\tilde{X}_M$. Actually, if $\mathrm{patt}(\beta) = M$, then $X\beta = XU_M\kappa = \tilde{X}_M\kappa$ for $\kappa \in \mathbb{R}^{k+}$. In particular,

(i) null components $M_i = 0$ lead to discard the column X_i from the design matrix X,

(ii) a cluster $K \subset \{1, \ldots, p\}$ of M (components of M equal in absolute value) leads to replacing the columns $(X_i)_{i \in K}$ by one column equal to the signed sum: $\sum_{i \in K} \text{sign}(M_i) X_i$.

Example 2.6. Let $X = (X_1 | X_2 | X_3 | X_4 | X_5)$, $M = (1, 2, -2, 0, 1)'$ and $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)' \in \mathbb{R}^{5+}$. Then the clustered matrix and the clustered parameter are given hereafter:

$$\tilde{X}_M = (X_2 - X_3 | X_1 + X_5) \text{ and } \tilde{\Lambda}_M = \begin{pmatrix} \lambda_1 + \lambda_2 \\ \lambda_3 + \lambda_4 \end{pmatrix}.$$

2.2. Sorted ℓ_1 norm, dual sorted ℓ_1 norm and subdifferential

The sorted ℓ_1 norm is defined as follows:

$$J_\Lambda(b) = \sum_{i=1}^p \lambda_i |b|_{(i)}, \quad b \in \mathbb{R}^p,$$

where $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ are the sorted components of b with respect to the absolute value. Given a norm $\| \cdot \|$ on $\mathbb{R}^p$, we recall that the dual norm $\| \cdot \|^*$ is defined by $\|b\|^* = \max\{v'b : \|v\| \leq 1\}$ for any $b \in \mathbb{R}^p$. In particular, the dual sorted ℓ_1 norm has an explicit expression given in [44] and recalled hereafter:

$$J_\Lambda^*(b) = \max \left\{ \frac{|b|_{(1)}}{\lambda_1}, \frac{\sum_{i=1}^2 |b|_{(i)}}{\sum_{i=1}^2 \lambda_i}, \ldots, \frac{\sum_{i=1}^p |b|_{(i)}}{\sum_{i=1}^p \lambda_i} \right\}, \quad b \in \mathbb{R}^p. \tag{2.1}$$

Related to the dual norm, the subdifferential of a norm $\| \cdot \|$ at b is recalled below (see *e.g.* [45] pages 167 and 180)

$$\partial \| \cdot \|(b) = \{v \in \mathbb{R}^p : \|z\| \geq \|b\| + v'(z - b) \ \forall z \in \mathbb{R}^p\},$$
$$= \{v \in \mathbb{R}^p : \|v\|^* \leq 1 \text{ and } v'b = \|b\|\}. \tag{2.2}$$

For the sorted ℓ_1 norm, geometric descriptions of the subdifferential at $b \in \mathbb{R}^p$ have been given in the particular case where $b_1 \geq \ldots \geq b_p \geq 0$ [35,46,47]. Hereafter, for an arbitrary $b \in \mathbb{R}^p$, Proposition 2.1 provides a new and useful formula for the subdifferential of the sorted ℓ_1 norm. This representation is the crux of the mathematical content of the present paper.

Proposition 2.1. Let $b \in \mathbb{R}^p$ and $M = \text{patt}(b)$. Then we have the following formula:

$$\partial J_\Lambda(b) = \left\{ v \in \mathbb{R}^p : J_\Lambda^*(v) \leq 1 \text{ and } U_M'v = \tilde{\Lambda}_M \right\}. \tag{2.3}$$

In Proposition A.2 we derive a simple characterization of elements in $\partial J_\Lambda(b)$. The notion of SLOPE pattern is related to the subdifferential via the following result.

Proposition 2.2. Let $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ where $\lambda_1 > \ldots > \lambda_p > 0$ and $a, b \in \mathbb{R}^p$. We have $\text{patt}(a) = \text{patt}(b)$ if and only if $\partial J_\Lambda(a) = \partial J_\Lambda(b)$.

A proof of Proposition 2.2 can be found in [35]. In the Appendix, we provide an independent proof, which is based on Proposition 2.1.

From now on, to comply with Proposition 2.2, we assume that the tuning parameter $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ satisfies

$$\lambda_1 > \ldots > \lambda_p > 0.$$

2.3. Characterization of SLOPE minimizers

The SLOPE estimator is a minimizer of the following optimization problem:

$$S_{X,\Lambda}(Y) = \arg \min_{b \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - Xb\|_2^2 + J_\Lambda(b) \right\}. \tag{2.4}$$

In this article we do not assume that $S_{X,\Lambda}(Y)$ contains a unique element and potentially $S_{X,\Lambda}(Y)$ can be a non-trivial compact and convex set. Note however that cases in which $S_{X,\Lambda}(Y)$ is not a singleton are very rare. Indeed, the set of matrices $X \in \mathbb{R}^{n \times p}$ for which there exists a $Y \in \mathbb{R}^n$ where $S_{X,\Lambda}(Y)$ is not a singleton has a null Lebesgue measure on $\mathbb{R}^{n \times p}$ [35]. If $\ker(X) = \{0\}$, then $S_{X,\Lambda}(Y)$ consists of one element. Recall that a convex function f attains its minimum at a point b if and only if $0 \in \partial f(b)$. Since $\partial \frac{1}{2} \|Y - Xb\|_2^2 = \{-X'(Y - Xb)\}$, the SLOPE estimator satisfies the following characterization:

$$\hat{\beta} \in S_{X,\Lambda}(Y) \quad \Leftrightarrow \quad X'(Y - X\hat{\beta}) \in \partial J_\Lambda(\hat{\beta}).$$

3. Characterization of pattern recovery by SLOPE

The characterization of pattern recovery by SLOPE given in Theorem 3.1 is a crucial result in this article. We recall that $\tilde{P}_M = (\tilde{X}_M')^+ \tilde{X}_M' = \tilde{X}_M \tilde{X}_M^+$ is the orthogonal projection onto $\text{col}(\tilde{X}_M)$, where A^+ represents the Moore-Penrose pseudo-inverse of the matrix A (see *e.g.* [48]).

Theorem 3.1. Let $X \in \mathbb{R}^{n \times p}$, $0 \neq \beta \in \mathbb{R}^p$, $Y = X\beta + \varepsilon$ for $\varepsilon \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{p+}$. Let $M = \mathrm{patt}(\beta) \in \mathcal{P}_p^{\mathrm{SLOPE}}$ and $k = \|M\|_\infty$. Define

$$\pi = X'(\tilde{X}_M')^+ \tilde{\Lambda}_M + X'(I_n - \tilde{P}_M)Y. \tag{3.1}$$

There exists $\hat{\beta} \in S_{X,\Lambda}(Y)$ with $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$ if and only if the two conditions below hold true:

$$\begin{cases} \text{there exists } s \in \mathbb{R}^{k+} \text{ such that } \tilde{X}_M' Y - \tilde{\Lambda}_M = \tilde{X}_M' \tilde{X}_M s, & \text{(positivity condition)} \\ \pi \in \partial J_\Lambda(M). & \text{(subdifferential condition)} \end{cases}$$

If the positivity and subdifferential conditions are satisfied, then $\hat{\beta} = U_M s \in S_{X,\Lambda}(Y)$ and $\pi = X'(Y - X\hat{\beta})$.

Remark 3.1.

(i) When X is deterministic and ε has a $\mathrm{N}(0, \sigma^2 I_n)$ distribution, then the event of pattern recovery by SLOPE is the intersection of statistically independent events:

$$A = \{\omega \in \Omega : \text{there exists } s \in \mathbb{R}^{k+} \text{ such that } \tilde{X}_M' Y(\omega) - \tilde{\Lambda}_M = \tilde{X}_M' \tilde{X}_M s\},$$
$$B = \{\omega \in \Omega : \pi(\omega) \in \partial J_\Lambda(M)\}.$$

Indeed, since $\tilde{X}_M' = \tilde{X}_M' \tilde{P}_M$ then $\tilde{X}_M' Y(\omega)$ depends on $\varepsilon_A(\omega) = \tilde{P}_M \varepsilon(\omega)$. Moreover, $\pi(\omega)$ depends on $\varepsilon_B(\omega) = (I_n - \tilde{P}_M)\varepsilon(\omega)$. Since $\tilde{P}_M$ is an orthogonal projection, ε_A and ε_B have a null covariance matrix. But ε is Gaussian and hence ε_A and ε_B are independent. Therefore events A and B are independent.

(ii) Under the positivity condition, the subdifferential condition is equivalent to $J_\Lambda^*(\pi) \leq 1$. Indeed, observe that $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}_M')$ (or equivalently, $\tilde{X}_M'(\tilde{X}_M')^+ \tilde{\Lambda}_M = \tilde{\Lambda}_M$) is necessary for the positivity condition. In view of Eq. (2.3), using the definition of π, we see that $U_M' \pi = \tilde{\Lambda}_M$ is equivalent to $\tilde{X}_M'(\tilde{X}_M')^+ \tilde{\Lambda}_M = \tilde{\Lambda}_M$. This follows from the fact that $\tilde{P}_M$ is the projection matrix onto the vector subspace $\mathrm{col}(\tilde{X}_M)$, and thus $0' = [(I_n - \tilde{P}_M)\tilde{X}_M]' = U_M' X'(I_n - \tilde{P}_M)$.

(iii) The assertion of Theorem 3.1 cannot be strengthened. Indeed, if $S_{X,\Lambda}(Y)$ contains more than one element, then two different minimizers may have different SLOPE patterns.

Even though many theoretical properties on sign recovery by LASSO are known (see *e.g.* [39]), we believe that it is relevant to give a characterization of sign recovery by LASSO similar to the characterization of pattern recovery by SLOPE given in Theorem 3.1. Such a characterization could simplify proofs of well-known results on LASSO irrepresentability condition.

Remark 3.2. Let $0 \neq S \in \{-1, 0, 1\}^p$ and $k = \|S\|_1$ (k is the number of nonzero components of S). The signed matrix $U_S \in \mathbb{R}^{p \times k}$ is defined by $U_S = (\mathrm{diag}(S))_{\mathrm{supp}(S)}$ where $\mathrm{diag}(S) \in \mathbb{R}^{p \times p}$ is a diagonal matrix and $(\mathrm{diag}(S))_{\mathrm{supp}(S)}$ denotes the submatrix of $\mathrm{diag}(S)$ obtained by keeping columns corresponding to indices in $\mathrm{supp}(S)$. Observe that for any $0 \neq \beta \in \mathbb{R}^p$ there exists a unique $S \in \{-1, 0, 1\}^p$ and a unique $\kappa_0 \in (0, \infty)^k$ such that $\beta = U_S \kappa_0$. Define the reduced matrix $\tilde{X}_S$ and reduced parameter $\tilde{\lambda}_S$ by

$$\tilde{X}_S = XU_S \text{ and } \tilde{\lambda}_S = \lambda 1_k, \text{ where } 1_k = (1, \ldots, 1)' \in \mathbb{R}^k.$$

Moreover, for $b = (b_1, \ldots, b_p) \in \mathbb{R}^p$ define $\mathrm{sign}(b) = (\mathrm{sign}(b_1), \ldots, \mathrm{sign}(b_p))$. Similarly to the proof of Theorem 3.1, one may prove that the necessary and sufficient conditions for the LASSO sign recovery (*i.e.*, the existence of estimator $\hat{\beta}^{\mathrm{LASSO}}$ such that $\mathrm{sign}(\hat{\beta}^{\mathrm{LASSO}}) = \mathrm{sign}(\beta) = S$) are the following

$$\begin{cases} \text{there exists } \kappa \in \mathbb{R}_+^k \text{ such that } \tilde{X}_S' Y - \tilde{\lambda}_S = \tilde{X}_S' \tilde{X}_S \kappa, & \text{(positivity condition)} \\ X'(\tilde{X}_S')^+ 1_k + \frac{1}{\lambda} X'(I_n - \tilde{X}_S \tilde{X}_S^+)Y \in \partial \|\cdot\|_1(S). & \text{(subdifferential condition)} \end{cases}$$

In the noiseless case, when $\varepsilon = 0$ and $Y = X\beta$, the subdifferential condition reduces to $X'(\tilde{X}_S')^+ 1_k \in \partial \|\cdot\|_1(S)$ (or equivalently, $\|X'(\tilde{X}_S')^+ 1_k\|_\infty \leq 1$ and $1_k \in \mathrm{col}(\tilde{X}_S')$). Moreover, when $\ker(X_S) = \{0\}$ then $1_k \in \mathrm{col}(\tilde{X}_S')$ occurs and $\|X'(\tilde{X}_S')^+ 1_k\|_\infty \leq 1$ is equivalent to $\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S_I\|_\infty \leq 1$ where $I = \mathrm{supp}(S)$, $\overline{I} = \{1, \ldots, p\} \setminus I$ and X_I (resp. $X_{\overline{I}}$) denotes the submatrix of X obtained by keeping columns corresponding to indices in I (resp $\overline{I}$). This latter expression is known as the irrepresentability condition [37,40,41].

3.1. SLOPE irrepresentability condition

As illustrated by Fuchs [37] (Theorem 2), Bühlmann and van de Geer [49] (Theorem 7.1) and also recalled in Remark 3.2, the irrepresentability condition is related to sign recovery by LASSO in the noiseless case, *i.e.*, when the noise $\varepsilon = 0$. Analogously, analyzing pattern recovery by SLOPE in the noiseless case allows to introduce the SLOPE irrepresentability condition. This condition will be very useful in the remainder of the article when the noise term ε is no longer zero. Corollaries 3.2 and 3.3, which provide a characterization of pattern recovery by SLOPE in the noiseless case (as defined in [50]), follow as consequences of Theorem 3.1.

Corollary 3.2. Let $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$ where $\mathrm{patt}(\beta) = M \neq 0$. In the noiseless case, when $Y = X\beta$, the following statements are equivalent:

(i) There exists $\Lambda \in \mathbb{R}^{p+}$ and $\hat{\beta} \in S_{X,\Lambda}(X\beta)$ such that $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$.
(ii) For all $\lambda_1^0 > 0$, there exists $\Lambda \in \mathbb{R}^{p+}$ with $\lambda_1 < \lambda_1^0$ and $\hat{\beta} \in S_{X,\Lambda}(X\beta)$ such that $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$.
(iii) There exists $\Lambda \in \mathbb{R}^{p+}$ such that $X'(\tilde{X}_M')^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$ (or equivalently $J_\Lambda^*(X'(\tilde{X}_M')^+ \tilde{\Lambda}_M) \leq 1$ and $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}_M')$).
(iv) For all $\lambda_1^0 > 0$, there exists $\Lambda \in \mathbb{R}^{p+}$ with $\lambda_1 < \lambda_1^0$ such that $X'(\tilde{X}_M')^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$.

Typically, for penalized estimators, the penalty term is scaled by a tuning parameter $\alpha > 0$. The following corollary addresses the tuning of the SLOPE penalty J_Λ.

Corollary 3.3. When the penalty term $J_\Lambda(\cdot)$, with a fixed $\Lambda \in \mathbb{R}^{p+}$, is scaled by a parameter $\alpha > 0$, the following statements are equivalent:

 (i) There exists $\alpha > 0$ and $\hat{\beta} \in S_{X,\alpha\Lambda}(X\beta)$ such that $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$.
 (ii) There exists $\alpha_0 > 0$ such that for all $\alpha \in (0, \alpha_0)$ there exists $\hat{\beta} \in S_{X,\alpha\Lambda}(X\beta)$ for which $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$.
(iii) $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$.

From now on, given $M = \mathrm{patt}(\beta)$, we refer to the following inequality and inclusion as the SLOPE irrepresentability condition:

$$J_\Lambda^*\big(X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M\big) \leq 1 \text{ and } \tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M). \tag{3.2}$$

Remark 3.3.

 (i) When $\ker(\tilde{X}_M) = \{0\}$, we have $X'(\tilde{X}'_M)^+ = X'\tilde{X}_M(\tilde{X}'_M\tilde{X}_M)^{-1}$, and consequently, the SLOPE irrepresentability condition becomes

$$J_\Lambda^*(X'\tilde{X}_M(\tilde{X}'_M\tilde{X}_M)^{-1}\tilde{\Lambda}_M) \leq 1.$$

 (ii) A geometric interpretation of $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M$ is provided in the Appendix, see Appendix D.

Example 3.4. We give two illustrations in the particular case where $\Lambda = (4, 2)'$, $\beta = (5, 0)'$, $\bar{\beta} = (5, 3)'$ and $X = (X_1|X_2) \in \mathbb{R}^{n \times 2}$ such that

$$X'X = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

- The SLOPE irrepresentability condition does not occur when $\beta = (5, 0)'$. Indeed, $M = \mathrm{patt}(\beta) = (1, 0)'$, $\tilde{X}_M = X_1$ (thus $\tilde{X}'_M\tilde{X}_M = 1$) and $\tilde{\Lambda}_M = \lambda_1 = 4$. Therefore

$$J_\Lambda^*(X'(\tilde{X}'_M)^+\tilde{\Lambda}_M) = J_\Lambda^*(X'\tilde{X}_M(\tilde{X}'_M\tilde{X}_M)^{-1}\tilde{\Lambda}_M) = J_\Lambda^*(4X'\tilde{X}_M) = 6.4/6 > 1.$$

- The SLOPE irrepresentability condition occurs when $\bar{\beta} = (5, 3)'$. Indeed, $M = \mathrm{patt}(\bar{\beta}) = (2, 1)'$, $\tilde{X}_M = X$ and $\tilde{\Lambda}_M = \Lambda$. Therefore $\ker(\tilde{X}_M) = \{0\}$ and

$$J_\Lambda^*(X'(\tilde{X}'_M)^+\tilde{\Lambda}_M) = J_\Lambda^*(X'X(X'X)^{-1}\Lambda) = J_\Lambda^*(\Lambda) = 1 \leq 1.$$

Fig. 2 confirms graphically that SLOPE irrepresentability condition does not occur for β (resp. occurs for $\bar{\beta}$). Note that, in this setup, the SLOPE solution is unique (since $\ker(X) = \{0\}$); we denote by $\hat{\beta}(\alpha)$ the unique element of $S_{X,\alpha\Lambda}(X\beta)$ and the SLOPE solution path refers to the function $\alpha \in (0, \infty) \mapsto \hat{\beta}(\alpha)$.
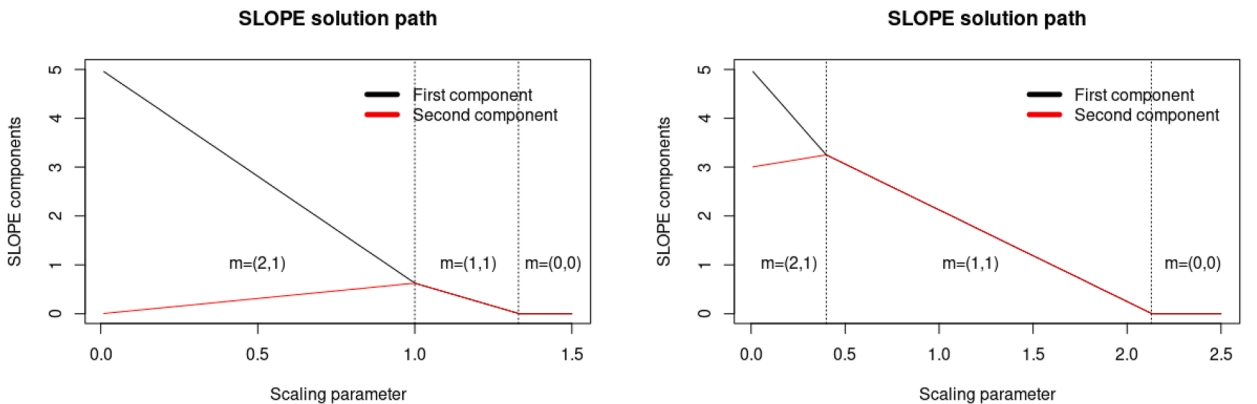
Fig. 2. On the left, the signal is $\beta = (5, 0)'$. Based on this figure one may observe that the pattern of β cannot be recovered by SLOPE in the noiseless case. Indeed, for $\alpha \in (0, 1)$ we have $\mathrm{patt}(\hat{\beta}(\alpha)) = (2, 1)'$; when $\alpha \in [1, 4/3)$ we have $\mathrm{patt}(\hat{\beta}(\alpha)) = (1, 1)'$ and when $\alpha > 4/3$ then $\hat{\beta}(\alpha) = 0$. Consequently, for every $\alpha > 0$ we have $\mathrm{patt}(\hat{\beta}(\alpha)) \neq \mathrm{patt}(\beta) = (1, 0)'$. On the right, the signal is $\bar{\beta} = (5, 3)'$. Based on this figure one may observe that $\mathrm{patt}(\bar{\beta})$ is recovered by SLOPE in the noiseless case. Indeed, for $\alpha \in (0, 0.4)$ we have $\mathrm{patt}(\hat{\beta}(\alpha)) = (2, 1)' = \mathrm{patt}(\bar{\beta})$.

4. Asymptotic probability on pattern recovery and pattern consistency

From now on, in the definition of SLOPE Eq. (2.4), we consider that the penalty term $J_\Lambda(b)$, with fixed $\Lambda \in \mathbb{R}^{p+}$, is multiplied by a scaling parameter $\alpha > 0$. We denote by $S_{X,\alpha\Lambda}(Y)$ the set of SLOPE solutions. This scaling parameter may, for instance, vary in $(0, \infty)$ for the solution path, or it can be chosen depending on the standard error of the noise. In this section we consider two asymptotic scenarios and establish conditions on tuning parameters for which the pattern of β is recovered. In Section 4.1 we consider the case where gaps between distinct absolute values of β diverge and in Section 4.2 the case where the sample size n diverges. The proofs rely on Theorem 3.1. We show that the positivity and subdifferential conditions are satisfied under our settings. It turns out that for the positivity condition the tuning parameter cannot be too large, while for the subdifferential condition it cannot be too small. In this way we consider a tuning parameter of the form $\alpha\Lambda$, where $\Lambda \in \mathbb{R}^{p+}$ is fixed and α varies. We determine the assumptions for the sequence (α) for which both positivity and subdifferential conditions hold true, *i.e.*, for which the pattern is recovered.

4.1. X is a fixed matrix

The subdifferential condition, given in Theorem 3.1, says that a vector π defined in Eq. (3.1) belongs to $\partial J_{\alpha\Lambda}(M)$, where α is a scaling parameter. This condition is equivalent to requiring that a vector $\pi_\alpha := \pi/\alpha$ is an element of $\partial J_\Lambda(M)$. We denote the vector π/α by

$$\pi_\alpha = X'(\tilde{X}'_M)^+\tilde{\Lambda}_M + \frac{1}{\alpha}X'(I_n - \tilde{P}_M)Y = X'(\tilde{X}'_M)^+\tilde{\Lambda}_M + \frac{1}{\alpha}X'(I_n - \tilde{P}_M)\varepsilon, \tag{4.1}$$

where in the latter equality we have used the fact that $(I_n - \tilde{P}_M)$ is an orthogonal projection onto $\mathrm{col}(\tilde{X}_M)^\perp$ and therefore $(I_n - \tilde{P}_M)X\beta = (I_n - \tilde{P}_M)\tilde{X}_M s = 0$, where $\beta = U_M s$ and $s \in \mathbb{R}^{\|M\|_\infty +}$.

By Theorem 3.1, the probability of pattern recovery by SLOPE is upper bounded by

$$\mathbb{P}\big(\exists\hat{\beta} \in S_{X,\alpha\Lambda}(Y) \text{ such that } \mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)\big) \leq \begin{cases} \mathbb{P}\big(J_\Lambda^*(\pi_\alpha) \leq 1\big), \\ 0 \text{ if } \tilde{\Lambda}_M \notin \mathrm{col}(\tilde{X}'_M). \end{cases} \tag{4.2}$$

Note that the condition $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)$ and $J_\Lambda^*(\pi_\alpha) \leq 1$ is necessary for pattern recovery by SLOPE, but not equivalent to it. Therefore, the inequality in Eq. (4.2) is, in general, not an equality. The first point in Theorem 4.1 shows that the probability of pattern recovery matches with the upper bound Eq. (4.2) when the gaps between the different absolute values of components of β are large enough. The last point establishes pattern consistency for SLOPE. The formulation of the theorem involves the notion of the relative interior of a set, which we recall below for completeness.

Definition 4.1. The affine hull of a set $C \subseteq \mathbb{R}^n$, denoted by $\mathrm{aff}(C)$, is the set of all affine combinations of points in C:

$$\mathrm{aff}(C) = \big\{\theta_1 x_1 + \cdots + \theta_k x_k : x_1, \ldots, x_k \in C, \theta_1 + \cdots + \theta_k = 1\big\}.$$

Definition 4.2. The relative interior of a set C, denoted $\mathrm{ri}(C)$, is the interior of C relative to its affine hull:

$$\mathrm{ri}(C) = \big\{x \in C : B(x, r) \cap \mathrm{aff}(C) \subseteq C \text{ for some } r > 0\big\},$$

where $B(x, r) = \{y \mid \|y - x\| \leq r\}$ is the ball of radius r centered at x, under any norm $\|\cdot\|$. All norms define the same relative interior.

Theorem 4.1. Let $X \in \mathbb{R}^{n\times p}$, $0 \neq M \in \mathcal{P}_p^{\mathrm{SLOPE}}$, and $\Lambda = (\lambda_1, \ldots, \lambda_p)' \in \mathbb{R}^{p+}$. Consider a sequence of signals $(\beta^{(r)})_{r\geq 1}$ with pattern M:

$$\beta^{(r)} = U_M s^{(r)} \quad \text{with} \quad s^{(r)} \in \mathbb{R}^{k+} \text{ and } k = \|M\|_\infty,$$

whose strength is increasing in the following sense:

$$\Delta_r = \min_{1\leq i < k} \big(s_i^{(r)} - s_{i+1}^{(r)}\big) \xrightarrow{r\to\infty} \infty, \text{ with the convention } s_{k+1}^{(r)} = 0$$

and let $Y^{(r)} = X\beta^{(r)} + \varepsilon$, where ε is a vector in $\mathbb{R}^n$.

(i) Sharpness of the upper bound: Let $\alpha > 0$. If ε is random, then the upper bound Eq. (4.2) is asymptotically reached:

$$\lim_{r\to\infty} \mathbb{P}\big(\exists\hat{\beta} \in S_{X,\alpha\Lambda}(Y^{(r)}) \text{ such that } \mathrm{patt}(\hat{\beta}) = M\big) = \begin{cases} \mathbb{P}\big(J_\Lambda^*(\pi_\alpha) \leq 1\big), \\ 0 \text{ if } \tilde{\Lambda}_M \notin \mathrm{col}(\tilde{X}'_M). \end{cases}$$

(ii) Pattern consistency: If $\alpha_r \to \infty$, $\alpha_r/\Delta_r \to 0$ as $r \to \infty$ and

$$X'(\tilde{X}'_M)^+\tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M)),$$

then for any $\varepsilon \in \mathbb{R}^n$ we have

$$\exists r_0 > 0 \ \forall r \geq r_0 \ \exists \hat{\beta} \in S_{X,\alpha_r\Lambda}(Y^{(r)}) \text{ such that } \mathrm{patt}(\hat{\beta}) = M.$$

Remark 4.3.

(i) The condition $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$, called open irrepresentability condition, is slightly stronger than the irrepresentability condition $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$. Note that the tight gap between these conditions is not specific to SLOPE. For instance, for LASSO, the irrepresentability condition which is sufficient for support recovery in the noisy case is stronger than the weak irrepresentability condition for the noiseless case (see [49] pages 190–192 and 244).

(ii) For the open irrepresentability condition we must check that the cardinality of the set of equalities among the p inequalities corresponding to $J_\Lambda^*(X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M) \leq 1$, see Eq. (2.1), is exactly $\|M\|_\infty$. That is, $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$ is equivalent to the following computationally verifiable conditions:

$$
\begin{cases}
J_\Lambda^*(X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M) \leq 1 \text{ and } \tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M), \\
\left| \left\{ i \in \{1, \dots, p\} : \ \sum_{j=1}^{i} |X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M|_{(j)} = \sum_{j=1}^{i} \lambda_j \right\} \right| = \|M\|_\infty.
\end{cases}
\tag{4.3}
$$

This equivalence follows from Proposition A.2.

(iii) Let us assume that the distributions of ε and $-\varepsilon$ are equal. Because the unit ball of the dual sorted ℓ_1 norm is convex, when $J_\Lambda^*(X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M) > 1$ then, independently of $\alpha > 0$, the probability of pattern recovery is smaller than $1/2$, namely

$$
\mathbb{P}\big(\exists \hat{\beta} \in S_{X,\alpha\Lambda}(Y) \text{ such that } \mathrm{patt}(\hat{\beta}) = M \big) \leq 1/2.
$$

This inequality corroborates Theorem 4.6 in [50]. For LASSO, a similar inequality on the probability of sign recovery is given in [39].

(iv) In Section 5, we illustrate that, under the open irrepresentability condition, one may select $\alpha > 0$ to fix the asymptotic probability of pattern recovery at a level arbitrarily close to 1 (a similar result for LASSO is given in [51]).

4.2. X is random, p is fixed, n tends to infinity

In this section we discuss asymptotic properties of the SLOPE estimator in the low-dimensional regression model in which p is fixed and the sample size n tends to infinity.

For each $n \geq p$ we consider a linear regression problem

$$
Y_n = X_n \beta + \varepsilon_n,
\tag{4.4}
$$

where $X_n \in \mathbb{R}^{n \times p}$ is a random design matrix. We now list our assumptions:

A. $\varepsilon_n = (\epsilon_1, \dots, \epsilon_n)'$, where $(\epsilon_i)_i$ are i.i.d. centered with finite variance.

B1. A sequence of design matrices $X_1, X_2, \dots$ satisfies the condition

$$
\frac{1}{n} X'_n X_n \xrightarrow{\mathbb{P}} C,
\tag{4.5}
$$

where C is a deterministic positive definite symmetric $p \times p$ matrix.

B2. For each $j = 1, \dots, p$,

$$
\frac{\max_{i=1,\dots,n} |X_{ij}^{(n)}|}{\sqrt{\sum_{i=1}^{n} (X_{ij}^{(n)})^2}} \xrightarrow{\mathbb{P}} 0.
$$

C. $(X_n)_n$ and $(\epsilon_n)_n$ are independent.

We will consider a sequence of tuning parameters $(\Lambda_n)_n$ defined by

$$
\Lambda_n = \alpha_n \Lambda,
$$

where $\Lambda \in \mathbb{R}^{p+}$ is fixed and $(\alpha_n)_n$ is a sequence of positive numbers.

Let $\hat{\beta}_n^{\mathrm{SLOPE}}$ be an element from the set $S_{X_n, \Lambda_n}(Y_n)$ of SLOPE minimizers. Under assumption B1, for large n with high probability, the set $S_{X_n, \Lambda_n}(Y_n)$ consists of one element. Indeed, we have

$$
\mathbb{P}\big(\ker(X_n) = \{0\}\big) = \mathbb{P}\big(X'_n X_n \text{ is positive definite}\big) \xrightarrow{n \to \infty} 1
$$

and $\ker(X_n) = \{0\}$ ensures the existence of the unique SLOPE minimizer. In a natural setting, the strong consistency of $\hat{\beta}_n^{\mathrm{SLOPE}}$ can be characterized in terms of behaviour of the tuning parameter, see Theorem C.2 or [33, Th. 4.1]. At this point we note that if Eq. (4.5) holds almost surely, then condition $\alpha_n/n \to 0$ ensures that $\hat{\beta}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \beta$. Thus, if β does not have any clusters nor zeros, *i.e.*, $\|\mathrm{patt}(\beta)\|_\infty = p$, then the $\alpha_n/n \to 0$ suffices for $\mathrm{patt}(\hat{\beta}_n^{\mathrm{SLOPE}}) \xrightarrow{a.s.} \mathrm{patt}(\beta)$. However, if $\|\mathrm{patt}(\beta)\| < p$, then the situation is more complex as we shall show below.

The first of our asymptotic results concerns the consistency of the pattern recovery by the SLOPE estimator. We note that condition B2 is not necessary for the SLOPE pattern recovery. This assumption was introduced to ensure the existence of a Gaussian vector in Theorem 4.2(i).

The formulation of the following theorem involves the notion of the pattern matrix U_M, as defined in Definition 2.3.

Theorem 4.2. Under the assumptions A, B1, C, the following statements hold true.

(i) If $B2$ is additionally satisfied and moreover $\alpha_n = \sqrt{n}$, then

$$\lim_{n\to\infty} \mathbb{P}\big(\mathrm{patt}(\hat{\beta}_n^{\mathrm{SLOPE}}) = \mathrm{patt}(\beta)\big) = \mathbb{P}\big(J_\Lambda^*(Z) \leq 1\big),$$

where $Z \sim \mathrm{N}(CU_M(U_M' C U_M)^{-1}\bar{\Lambda}_M, \sigma^2[C - CU_M(U_M' C U_M)^{-1}U_M' C])$.

(ii) Assume

$$CU_M(U_M' C U_M)^{-1}\bar{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M)). \tag{4.6}$$

The pattern of SLOPE estimator is consistent, *i.e.*

$$\mathrm{patt}(\hat{\beta}_n^{\mathrm{SLOPE}}) \xrightarrow{\mathbb{P}} \mathrm{patt}(\beta),$$

if and only if

$$\lim_{n\to\infty} \frac{\alpha_n}{n} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\alpha_n}{\sqrt{n}} = \infty.$$

(iii) The condition

$$J_\Lambda^*\big(CU_M(U_M' C U_M)^{-1}\bar{\Lambda}_M\big) \leq 1 \tag{4.7}$$

is necessary for pattern consistency of SLOPE estimator.

The random vector Z belongs to the smallest affine space containing $\partial J_\Lambda(b)$, *i.e.*, $\mathrm{aff}(\partial J_\Lambda(b)) = \{v \in \mathbb{R}^p : U_M' v = \bar{\Lambda}_M\}$, see Lemma A.3.

Condition Eq. (4.6) is the open SLOPE irrepresentability condition in the $n \to \infty$ regime. The above result should be compared with [40, Theorem 1], where the same conditions on the LASSO tuning parameter ensure consistency of sign recovery by the LASSO estimator. Below we make a step further and consider the strong consistency of SLOPE pattern recovery by $\hat{\beta}_n^{\mathrm{SLOPE}}$. Although this was not Zhao's and Yu's main focus, it can be deduced from [40, Theorem 1] that if for $c \in (0,1)$ the LASSO tuning parameter λ_n satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \to \infty$, then under the strong LASSO irrepresentability condition, one has $\mathrm{sign}(\hat{\beta}_n^{\mathrm{LASSO}}) \xrightarrow{a.s.} \mathrm{sign}(\beta)$. Even though the patterns are discrete objects, as the underlying probability space is uncountable, the convergence in probability does not imply the almost sure convergence. We show below that if $\alpha_n/n \to 0$ and $\alpha_n/\sqrt{n} \to \infty$, then $\mathrm{patt}(\hat{\beta}_n^{\mathrm{SLOPE}})$ is not strongly consistent and one actually has to impose a slightly stronger condition Eq. (4.8).

For the purpose of the a.s. convergence, we strengthen the assumption on design matrices:

B'. Assume that the rows of X_n are independent and that each row of X_n has the same law as ξ, where ξ is a random vector whose components are linearly independent a.s. and that $\mathbb{E}[\xi_i^2] < \infty$ for $i = 1, \ldots, p$.

Remark 4.4. Under B', by the strong law of large numbers, we have $n^{-1}X_n' X_n \xrightarrow{a.s.} C$, where $C = (C_{ij})_{ij}$ with $C_{ij} = \mathbb{E}[\xi_i\xi_j]$. Moreover, C is positive definite if and only if the random variables $(\xi_1, \ldots, \xi_p)$ are linearly independent a.s. Indeed, for $t \in \mathbb{R}^p$ we have $t'Ct = \mathbb{E}[(\sum_{i=1}^p t_i\xi_i)^2] > 0$ if and only if $\sum_{i=1}^p t_i\xi_i \neq 0$ a.s. for all $t \in \mathbb{R}^p \setminus \{0\}$.

Since B' ensures that Eq. (4.5) holds a.s., it also implies that for large n, almost surely there exists a unique SLOPE minimizer. We denote this element by $\hat{\beta}_n^{\mathrm{SLOPE}}$.

Theorem 4.3. Under A, B' and C assume that a sequence $(\alpha_n)_n$ satisfies

$$\lim_{n\to\infty} \frac{\alpha_n}{n} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\alpha_n}{\sqrt{n \log\log n}} = \infty. \tag{4.8}$$

If Eq. (4.6) holds, then the sequence $(\hat{\beta}_n^{\mathrm{SLOPE}})_n$ recovers almost surely the pattern of β asymptotically, *i.e.*,

$$\mathrm{patt}(\hat{\beta}_n^{\mathrm{SLOPE}}) \xrightarrow{a.s.} \mathrm{patt}(\beta). \tag{4.9}$$

Remark 4.5. Assume that Eq. (4.6) is satisfied and set $\alpha_n = c\sqrt{n \log\log n}$ for $c > 0$. Then Eq. (4.8) is not satisfied and with positive probability, the true SLOPE pattern is not recovered. See also Appendix B, where we present more refined results on the strong consistency of the SLOPE pattern. The $\log\log n$ correction in Eq. (4.8) comes from the law of iterated logarithm.

5. Simulation study

This simulation study aims at illustrating Theorems 4.1 and 4.2. Hereafter, we consider the linear regression model $Y = X\beta + \varepsilon$, where $X \in \mathbb{R}^{n\times p}$ and $\varepsilon \in \mathbb{R}^n$ has i.i.d. $\mathrm{N}(0,1)$ entries. Up to a constant, we choose components of $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ as expected values of ordered standard Gaussian statistics. Let $Z_{(1)} \geq \ldots \geq Z_{(p)}$ be ordered statistics of i.i.d. $\mathrm{N}(0,1)$ random variables. An approximation of $\mathbb{E}[Z_{(i)}]$ for some $i \in \{1, \ldots, p\}$, denoted $E(i,p)$, is given hereafter (see [52] and references therein)

$$E(i,p) = -\Phi^{-1}\left(\frac{i - 0.375}{p + 1 - 0.750}\right),$$

where Φ is the cumulative distribution function of an $\mathrm{N}(0,1)$ random variable. We set

$$\Lambda = (\lambda_1, \ldots, \lambda_p) \quad \text{with} \quad \lambda_i = E(i,p) + E(p-1,p) - 2E(p,p). \tag{5.1}$$

5.1. Sharp upper bound when X is orthogonal

This example illustrates Theorem 4.1, which concerns the limiting probability of pattern recovery as signal strength tends to infinity. We assume that $p = 100$, c is a positive real number, $X \in \mathbb{R}^{n \times p}$ is orthogonal ($X'X = I_{100}$), and $\beta \in \mathbb{R}^p$ is defined as follows:

$$\beta_1 = \ldots = \beta_{25} = c, \quad \beta_{26} = \ldots = \beta_{50} = -c/2, \quad \beta_{51} = \ldots = \beta_{100} = 0 . \tag{5.2}$$

To compute the value $\alpha_{0.95}$ of the scaling parameter for which the upper bound is 0.95 we note that π_α is a Gaussian vector having a

$$\mathrm{N}\big(X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M, \alpha^{-2} X'(I - \tilde{X}_M \tilde{X}_M^+)X\big)$$

distribution. Moreover, since $M = \mathrm{patt}(\beta)$ satisfies: $M_1 = \ldots = M_{25} = 2$, $M_{26} = \ldots = M_{50} = -1$ and $M_{51} = \ldots = M_{100} = 0$ we have

$$X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M = \mu,$$

where $\mu_1 = \ldots = \mu_{25} = \frac{1}{25} \sum_{i=1}^{25} \lambda_i$, $\mu_{26} = \ldots = \mu_{50} = -\frac{1}{25} \sum_{i=26}^{50} \lambda_i$,
$\mu_{51} = \ldots = \mu_{100} = 0$, and

$$X'(I_n - \tilde{X}_M \tilde{X}_M^+)X = \begin{pmatrix} \Sigma & 0 & 0 \\ 0 & \Sigma & 0 \\ 0 & 0 & I_{p/2} \end{pmatrix},$$

where Σ is the matrix of the dimension $p/4 \times p/4$ given by

$$\Sigma = \begin{pmatrix} 1 - 4/p & -4/p & \ldots & -4/p \\ -4/p & 1 - 4/p & \ddots & \vdots \\ \vdots & \ddots & \ddots & -4/p \\ -4/p & \ldots & -4/p & 1 - 4/p \end{pmatrix}. \tag{5.3}$$

The matrix Σ appears twice in the covariance structure, as both nonzero clusters have the same size of $p/4$.

Since the open SLOPE irrepresentability condition holds, there exists the value $\alpha_{0.95}$ such that

$$\mathbb{P}\big(J_\Lambda^*(\pi_{\alpha_{0.95}}) \le 1\big) = 0.95.$$

In practice, we simulated 50 000 instances of the random vector $Z \sim \mathrm{N}(0, X'(I - \tilde{X}_M \tilde{X}_M^+)X)$ and identified the value $\alpha_{0.95} = 9.45$, such that

$$\mathbb{P}\left(J_\Lambda^*\left(\mu + \frac{1}{\alpha_{0.95}} Z\right) \le 1\right) \approx 0.95.$$

Fig. 3 illustrates that indeed the probability of pattern recovery in the model Eq. (5.2) by SLOPE with a regularizing sequence 9.45Λ converges to 0.95 as c increases to infinity.

Probability of pattern recovery
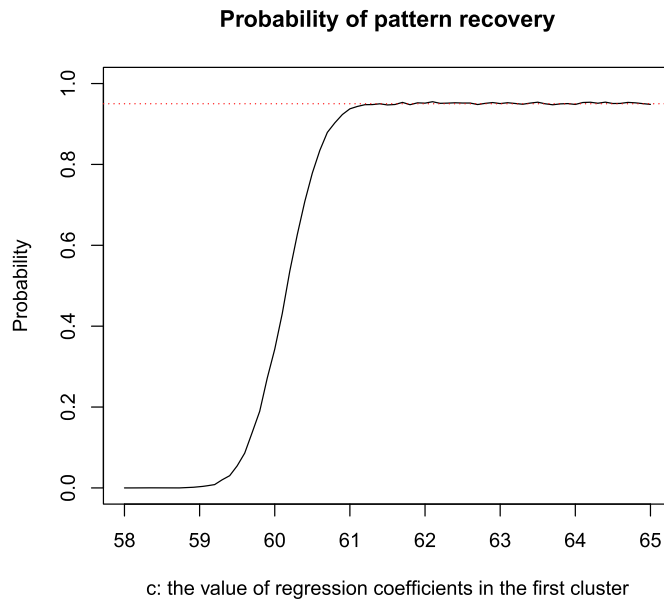
Fig. 3. Probability of the pattern recovery in the model Eq. (5.2) by SLOPE with a tuning parameter $\alpha\Lambda$, where $\alpha = 9.45$ and Λ is given in Eq. (5.1).

5.2. Limiting probability when $n \to \infty$

In this section, we illustrate Theorem 4.2, which describes the asymptotic performance of SLOPE as the sample size n tends to infinity. We consider a setting in which both the predictors and the regression coefficients exhibit a clustered structure.

Specifically, we assume that the design covariance matrix for $p = 100$ regressors is block-diagonal:

$$C = \begin{pmatrix} \Sigma & 0 & 0 & 0 \\ 0 & \Sigma & 0 & 0 \\ 0 & 0 & \Sigma & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix} \tag{5.4}$$

where Σ is a compound symmetry matrix of dimension 25×25, with $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = 0.8$ for $i \neq j$.

The true regression coefficient vector $\beta \in \mathbb{R}^p$ is defined as:

$$\beta_1 = \ldots = \beta_{25} = 30, \quad \beta_{26} = \ldots = \beta_{50} = -30, \quad \beta_{51} = \ldots = \beta_{100} = 0.$$

Thus, in this example, the pattern $M = \mathrm{patt}(\beta)$ satisfies:

$$M_1 = \ldots = M_{25} = 1, \quad M_{26} = \ldots = M_{50} = -1, \quad M_{51} = \ldots = M_{100} = 0.$$

Furthermore, we have:

$$CU_M(U'_M C U_M)^{-1}\tilde{\Lambda}_M = \mu,$$

where

$$\mu_1 = \ldots = \mu_{25} = \frac{1}{50}\sum_{i=1}^{50}\lambda_i, \quad \mu_{26} = \ldots = \mu_{50} = -\mu_1, \quad \mu_{51} = \ldots = \mu_{100} = 0.$$

Finally, the covariance matrix of the vector Z in Theorem 4.2 takes the form:

$$\Sigma_Z = C - CU_M(U'_M C U_M)^{-1}U'_M C = \begin{pmatrix} \Sigma - U & U & 0 & 0 \\ U & \Sigma - U & 0 & 0 \\ 0 & 0 & \Sigma & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix}, \tag{5.5}$$

where U is the 25×25 matrix in which all entries are equal to $(1 + 24 \cdot 0.8)/50$.

By simulating $50\,000$ instances of the multivariate normal vector $Z \sim \mathrm{N}(0, \Sigma_Z)$ we found the value $\alpha = 2.89$, such that $\mathbb{P}\left(J_\Lambda^\star\left(\mu + \frac{1}{\alpha}Z\right) \leq 1\right) \approx 0.95$

According to Theorem 4.2 SLOPE with the tuning sequence $2.89\Lambda\sqrt{n}$ – where Λ is specified in Eq. (5.1) – should recover the true pattern with the probability 0.95 as $n \to \infty$. This phenomenon is illustrated in Fig. 4, where the probability of the pattern recovery stabilizes at 0.95 for $n \geq 1500$.

Additionally, Fig. 5 illustrates the root mean square error (RMSE) of the SLOPE estimator and compares it to the RMSE of both the ordinary least squares (LS) estimator and the debiased SLOPE estimator. The latter is obtained by performing a least squares fit using the reduced model selected by SLOPE, i.e., using the design matrix $\tilde{X} = XU_{\hat{M}}$.

As shown in the figure, SLOPE consistently outperforms LS in terms of RMSE in this example. Moreover, the estimation accuracy can be further improved by debiasing SLOPE – specifically, by applying least squares estimation within the reduced model. In the setting considered here, this two-stage version of SLOPE achieves near-perfect performance, with a negligible estimation error.
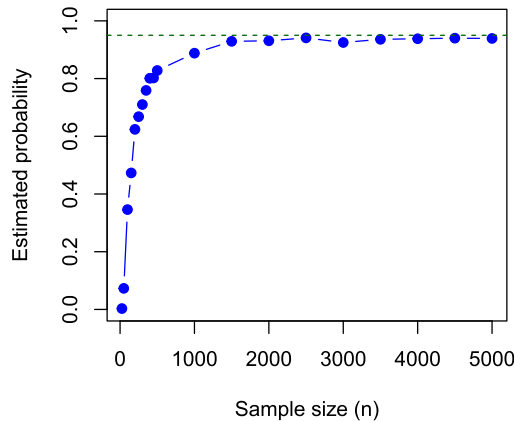
Fig. 4. Estimates of probability of pattern recovery by SLOPE as a function of n. The scaling parameter $\alpha_{0.95} = 2.89$ is chosen to fix the limiting probability of pattern recovery at 0.95.
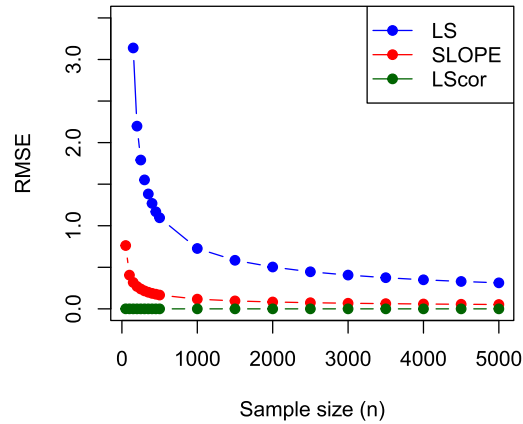
Fig. 5. Estimated Root Mean Squares of the estimators obtained by Least Squares (LS), SLOPE with the sequence of parameters as in Fig. 4 (SLOPE) and the Least Squares estimators using SLOPE dimensionality reduction, *i.e.*, $\bar{X} = XU_{\hat{M}}$ (LScor).
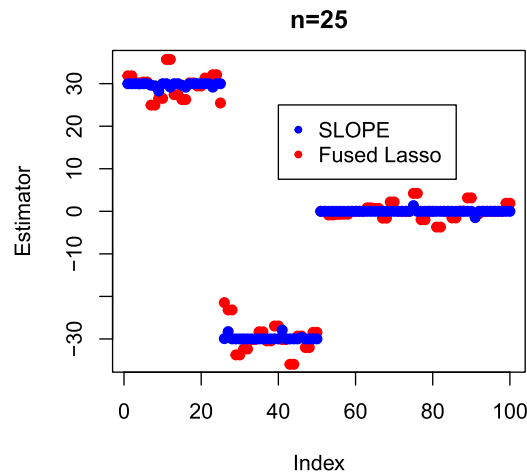

Fig. 6. Comparison of SLOPE and Fused LASSO estimators for $n = 25$ and $p = 100$. Among the 100 regression coefficients, 50 are nonzero: the first 25 are equal to 30, and the next 25 are equal to -30.

5.3. Small n performance

In Fig. 6 we present a representative example of the performance of SLOPE for the setting from Section 5.2 and a very small $n = 25$. We compare SLOPE to Fused LASSO, since the coefficients are clustered according to the proximity of variables in the design matrix. For SLOPE we use the sequence of tuning parameters as proposed in Section 5.2, while the Fused LASSO is tuned manually to minimize the estimation error. We used the *fusedlasso* function from the *genlasso* library and manually selected $\gamma = 0.25$ as the value for which we obtained the smallest RMSE over the range of λ values automatically proposed by the *fusedlasso* algorithm. We can observe that while SLOPE cannot precisely estimate the pattern, shrinkage towards the common absolute mean in both clusters allows for obtaining a very precise estimation of β, which is substantially more accurate than the fused lasso estimator. We believe that this is due to the fact that SLOPE effectively shrinks both clusters towards the same absolute value, while fused LASSO does not have this advantage.

6. Discussion

In this article we make an important step in understanding the clustering properties of SLOPE and we have shown that the irrepresentability condition provides theoretical guarantees for SLOPE pattern recovery. However, this by no means closes the topic of the SLOPE pattern recovery. Similarly to the irrepresentability condition for LASSO, the SLOPE irrepresentability condition is rather stringent and imposes a strict restriction on the number of nonzero clusters in β. On the other hand, in [51] it is shown that a much weaker condition for LASSO is required to separate the estimators of the null components of β from the estimators of nonzero regression coefficients. This condition, called accessibility (also called identifiability), requires that the vector β has the minimal ℓ_1 norm among all vectors γ such that $X\beta = X\gamma$. Thus, when the accessibility condition is satisfied one can recover the sign of β by
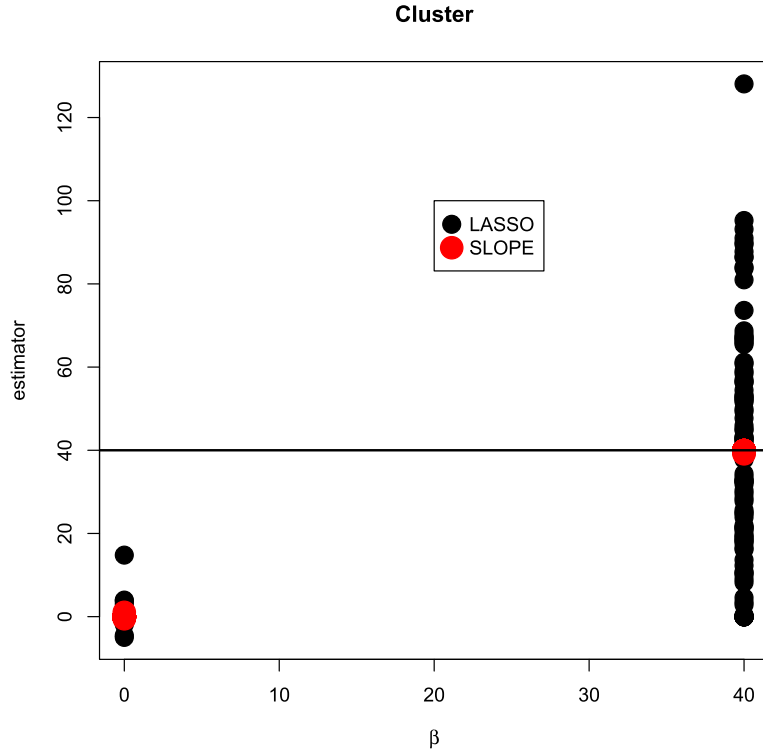
Fig. 7. Comparison of LASSO and SLOPE when the cluster structure is present in the data. Here $n = 100$, $p = 200$, and the correlation between i^{th} and j^{th} column of X is equal to $0.9048^{|i-j|}$. First $k = 100$ columns of X are associated with Y and their nonzero regression coefficients are all equal to 40. The SLOPE and LASSO irrepresentability conditions are not satisfied, but SLOPE, contrary to LASSO, satisfies the accessibility condition.

thresholding LASSO estimates. Empirical results from [51] suggest that this weaker condition is also sufficient for the sign recovery by the adaptive LASSO [41]. In this case rescaling the design matrix according to the initial estimates of regression coefficients modifies the original irrepresentability condition, so it can be satisfied for a given specific true sign vector of regression coefficients. In the recent article [50] it is shown that a similar result holds for SLOPE, whose accessibility condition holds if the vector β has the smallest sorted ℓ_1 norm among all vectors γ such that $X\beta = X\gamma$. In [50] or in [43, Theorem 2.2] it is shown that when the accessibility condition is satisfied then applying the proximal operator of the sorted ℓ_1 norm to SLOPE allows to recover the pattern of the regression coefficients. Fig. 7 illustrates this phenomenon and shows that the accessibility condition for SLOPE can be much less restrictive than the accessibility condition for LASSO. In this example the matrix X and the vector Y are generated as in example illustrated in Fig. 1 and the only difference is that now first $k = 100 = n$ regression coefficients are all equal to 40. In this situation the accessibility condition for LASSO is not satisfied and LASSO cannot properly separate the null and nonzero regression coefficients. Also, despite the selection of the tuning parameter so as to minimize the squared estimation error, the precision of LASSO estimates is very poor. As far as SLOPE is concerned, the irrepresentability condition is not satisfied but the accessibility condition holds. Thus, while SLOPE cannot properly identify the pattern, it estimates β with such a good precision that the difference between the estimated and the true pattern is hardly visible on the graph. These favorable ranking and estimation properties of the SLOPE method enable pattern recovery through appropriately selected thresholded versions of SLOPE. We also expect that the mathematical understanding of SLOPE irrepresentability condition presented in this article will lead to the development of efficient adaptive versions of SLOPE, with improved estimation and pattern recovery properties.

The results presented in this article pave the way for a full understanding of the SLOPE pattern recovery properties. We expect that our SLOPE irrepresentability condition will be a basic block for proving further results on the pattern recovery of SLOPE and adaptive SLOPE in the high-dimensional regime. We also look forward to research on other statistical models and loss functions. One specific focus of interest is the graphical SLOPE (see [53,54]), which could be used for identification of colored graphical models [55], with specific parameter sharing patterns in the precision matrix. Such repetitive patterns occur naturally in many situations, *e.g.*, in the case of the autoregressive type of dependence between variables in the database or when variables are influenced by the same structural factors. We believe that an efficient exploitation of these unknown patterns by SLOPE will lead to a great reduction of the number of parameters and improvement of the graphical models estimation properties.

Finally, we would like to recall that an interest in identifying the parameter sharing patterns goes beyond classical parametric models and is prevalent also in the modern machine learning community. As mentioned in the introduction, the prominent example is provided by the Convolutional Neural Networks (CNN), where the "parameter sharing" has made it possible to dramatically improve

computational and statistical efficiency. While the parameter sharing in CNN is driven entirely by the expert knowledge, regularization by SLOPE allows to identify and exploit patterns based on the data. In principle one can also use SLOPE in the Bayesian context and combine the information in the data with the imprecise prior knowledge on possible parameter sharing patterns (see [56] for the preliminary version of adaptive Bayesian SLOPE). It is expected that recent developments in efficient implementations of the SLOPE optimization algorithm (see, *e.g.* [57,58]) will soon allow for an integration of SLOPE regularization with the deep neural network architectures.

Data availability

Data will be made available on request.

Acknowledgment

Appendix A. Proofs

A.1. Proof of Proposition 2.1

Note that if $M = 0$, then the statement holds by Eq. (2.2). Thus we may later assume that $M \neq 0$. To ease the notation, we write $\tilde{\Lambda}$ instead of $\tilde{\Lambda}_M$. The elements of $\tilde{\Lambda}$ are denoted by $\tilde{\Lambda}_l$, $l = 1, \ldots, k$. Let $k = \|M\|_\infty$. Before proving Proposition 2.1 note that, by assumption, there exists $s \in \mathbb{R}^{k+}$ such that $b = U_M s$. Consequently, $|b|_\downarrow = U_{|M|_\downarrow} s$ and thus

$$J_\Lambda(b) = \lambda_1 |b|_{(1)} + \ldots + \lambda_p |b|_{(p)} = \Lambda' U_{|M|_\downarrow} s = \tilde{\Lambda}' s = s_1 \tilde{\Lambda}_1 + \ldots + s_k \tilde{\Lambda}_k.$$

Moreover, with $p_l = |\{i : |M_i| \geq k + 1 - l\}|$, we have $\tilde{\Lambda}_l = \lambda_{p_{l-1}+1} + \ldots + \lambda_{p_l}$, $l = 1, \ldots, k$.

Proof of Proposition 2.1. First we prove the inclusion $\partial J_\Lambda(b) \subset \{v \in \mathbb{R}^p : J_\Lambda^*(v) \leq 1 \text{ and } U_M' v = \tilde{\Lambda}\}$. Let $v \in \partial J_\Lambda(b)$. Since $J_\Lambda^*(v) \leq 1$ (see Eq. (2.2)) then, by definition of the dual sorted ℓ_1 norm, for all $j \in \{1, 2, \ldots, p\}$ we have $\sum_{i=1}^j |v|_{(i)} \leq \sum_{i=1}^j \lambda_i$. It remains to prove that $U_M' v = \tilde{\Lambda}$. For all $l \in \{1, \ldots, k\}$ we have the following inequality

$$\sum_{i=1}^l [U_M' v]_i = \sum_{i \,:\, |M_i| \geq k+1-l} \text{sign}(M_i) v_i \leq \sum_{i \,:\, |M_i| \geq k+1-l} |v_i|$$

$$\leq \sum_{i=1}^{p_l} |v|_{(i)} \leq \sum_{i=1}^{p_l} \lambda_i = \sum_{i=1}^l \tilde{\Lambda}_i. \tag{A.1}$$

Note that

$$b'v = (U_M s)'v = \sum_{i=1}^k s_i [U_M' v]_i = \sum_{l=1}^{k-1} (s_l - s_{l+1}) \sum_{i=1}^l [U_M' v]_i + s_k \sum_{i=1}^k [U_M' v]_i$$

$$\leq \sum_{l=1}^{k-1} (s_l - s_{l+1}) \sum_{i=1}^l \tilde{\Lambda}_i + s_k \sum_{i=1}^k \tilde{\Lambda}_i = \sum_{l=1}^k s_l \tilde{\Lambda}_l = J_\Lambda(b).$$

Moreover, since $v \in \partial J_\Lambda(b)$, we have $b'v = J_\Lambda(b)$ (see Eq. (2.2)). Therefore

$$\sum_{i=1}^l [U_M' v]_i = \sum_{i=1}^l \tilde{\Lambda}_i \quad \text{for} \quad l = 1, \ldots, k$$

and thus the inequalities given in Eq. (A.1) are the equalities. Thus

$$[U_M' v]_l = \tilde{\Lambda}_l \quad \text{for} \quad l = 1, \ldots, k$$

and hence that $U_M' v = \tilde{\Lambda}$.

Now we prove the other inclusion, $\partial J_\Lambda(b) \supset \{v \in \mathbb{R}^p : J_\Lambda^*(v) \leq 1 \text{ and } U_M' v = \tilde{\Lambda}_M\}$. Assume that $v \in \mathbb{R}^p$ satisfies $J_\Lambda^*(v) \leq 1$ and $U_M' v = \tilde{\Lambda}$. To prove that $v \in \partial J_\Lambda(b)$ it remains to establish that $b'v = J_\Lambda(b)$ (see Eq. (2.2)). Since $b = U_M s$, we have

$$b'v = (U_M s)'v = s' U_M' v = s' \tilde{\Lambda} = J_\Lambda(b).$$

$\square$

A.2. Proof of Proposition 2.2

Lemma A.1. Let $\Lambda \in \mathbb{R}^{p+}$ and $b \in \mathbb{R}^p$. If $\Lambda \in \partial J_\Lambda(b)$ then $b_1 \geq \ldots \geq b_p \geq 0$.

Proof. Let us assume that $b_i < 0$ for some $i \in \{1, \ldots, p\}$. For

$$\check{\pi} = (\lambda_1, \ldots, \lambda_{i-1}, -\lambda_i, \lambda_{i+1}, \ldots, \lambda_p)$$

we have $J_\Lambda^*(\check{\pi}) \leq 1$ and one may deduce that

$$\Lambda' b < \check{\pi}' b \leq \max\{\pi' b : J_\Lambda^*(\pi) \leq 1\} = J_\Lambda(b).$$

Consequently $\Lambda \notin \partial J_\Lambda(b)$ leading to a contradiction. Let us assume that $b_i < b_j$ for some $1 \leq i < j \leq p$. Let us define $\check{\pi}$, where $J_\Lambda^*(\check{\pi}) \leq 1$, as follows

$$\check{\pi}_k = \begin{cases} \lambda_k & \text{if } k \neq i, k \neq j, \\ \lambda_j & \text{if } k = i, \\ \lambda_i & \text{if } k = j, \end{cases} \quad k = 1, \ldots, p.$$

Since $\lambda_i > \lambda_j$, by the rearrangement inequality we have $\lambda_i b_i + \lambda_j b_j < \lambda_j b_i + \lambda_i b_j$. Thus, one may deduce the following inequality

$$\Lambda' b < \check{\pi}' b \leq \max\{\pi' b : \pi \in \mathbb{R}^p \ J_\Lambda^*(\pi) \leq 1\} = J_\Lambda(b).$$

Consequently $\Lambda \notin \partial J_\Lambda(b)$ leading to a contradiction. $\square$

Let ψ be an orthogonal transformation defined by

$$\psi : b \in \mathbb{R}^p \mapsto (v_1 b_{r(1)}, \ldots, v_p b_{r(p)})$$

where $v_1, \ldots, v_p \in \{-1, 1\}$ and r is a permutation on $\{1, \ldots, p\}$. Before proving Proposition 2.2 let us recall that for any $a, b \in \mathbb{R}^p$ we have $J_\Lambda(b) = J_\Lambda(\psi(b))$, $J_\Lambda^*(b) = J_\Lambda^*(\psi(b))$ and $b'a = \psi(b)'\psi(a)$ implying thus $\partial J_\Lambda(\psi(b)) = \psi(\partial J_\Lambda(b))$.

Proof of Proposition 2.2. If $\mathrm{patt}(a) = \mathrm{patt}(b)$ then, according to Proposition 2.1, $\partial J_\Lambda(a) = \partial J_\Lambda(b)$. Let us set $M = \mathrm{patt}(a)$ and $\tilde{M} = \mathrm{patt}(b)$, it remains to prove that if $\partial J_\Lambda(a) = \partial J_\Lambda(b)$ then $M = \tilde{M}$. Since the subdifferential $\partial J_\Lambda(a)$ depends on a only through its pattern, then by Proposition 2.1 we have $\partial J_\Lambda(a) = \partial J_\Lambda(M)$ and similarly $\partial J_\Lambda(b) = \partial J_\Lambda(\tilde{M})$.

First let us assume that $M = |M|_\downarrow$ namely $M_1 \geq M_2 \geq \ldots \geq M_p \geq 0$. In this case, $M'\Lambda = J_\Lambda(M)$ and hence $\Lambda = (\lambda_1, \ldots, \lambda_p)' \in \partial J_\Lambda(M)$. Since $\partial J_\Lambda(M) = \partial J_\Lambda(\tilde{M})$, it follows from Lemma A.1 that $\tilde{M}_1 \geq \ldots \geq \tilde{M}_p \geq 0$, because $\Lambda \in \partial J_\Lambda(\tilde{M})$. To prove that $M = \tilde{M}$, first let us establish that $M_p = \tilde{M}_p = 0$ or $M_p = \tilde{M}_p = 1$. If $M_p = 0$ and $\tilde{M}_p = 1$ then, let us set $\check{\pi} = (\lambda_1, \ldots, \lambda_{p-1}, 0)'$, where $J_\Lambda^*(\check{\pi}) \leq 1$. Because

$$J_\Lambda(M) = \Lambda' M = \check{\pi}' M \text{ and } J_\Lambda(\tilde{M}) = \Lambda' \tilde{M} > \check{\pi}' \tilde{M}$$

we have $\check{\pi} \in \partial J_\Lambda(M)$ and $\check{\pi} \notin \partial J_\Lambda(\tilde{M})$ which provides a contradiction. We proceed analogously for $M_p = 1$ and $\tilde{M}_p = 0$. To complete the proof that $M = \tilde{M}$, let us establish that $M_i = M_{i+1}$ and $\tilde{M}_i = \tilde{M}_{i+1}$ or $M_i > M_{i+1}$ and $\tilde{M}_i > \tilde{M}_{i+1}$. If $M_i = M_{i+1}$ and $\tilde{M}_i > \tilde{M}_{i+1}$ then, let us define $\check{\pi}$, where $J_\Lambda^*(\check{\pi}) \leq 1$, as follows

$$\check{\pi}_k = \begin{cases} \lambda_k & \text{if } k \neq i, k \neq i+1, \\ \lambda_{i+1} & \text{if } k = i, \\ \lambda_i & \text{if } k = i+1, \end{cases} \quad k = 1, \ldots, p.$$

Since $\lambda_i M_i + \lambda_{i+1} M_{i+1} = \lambda_{i+1} M_i + \lambda_i M_{i+1}$ and $\lambda_i \tilde{M}_i + \lambda_{i+1} \tilde{M}_{i+1} > \lambda_{i+1} \tilde{M}_i + \lambda_i \tilde{M}_{i+1}$ then

$$J_\Lambda(M) = \Lambda' M = \check{\pi}' M \text{ and } J_\Lambda(\tilde{M}) = \Lambda' \tilde{M} > \check{\pi}' \tilde{M}.$$

Consequently $\check{\pi} \in \partial J_\Lambda(M)$ and $\check{\pi} \notin \partial J_\Lambda(\tilde{M})$ which provides a contradiction. We proceed analogously for $M_i > M_{i+1}$ and $\tilde{M}_i = \tilde{M}_{i+1}$. Finally, if $M \neq |M|_\downarrow$ then let us pick an orthogonal transformation ψ as defined above for which $\psi(M) = |M|_\downarrow$. Since $\partial J_\Lambda(M) = \partial J_\Lambda(\tilde{M})$ implies that $\partial J_\Lambda(\psi(M)) = \partial J_\Lambda(\psi(\tilde{M}))$, the first part of the proof establishes that $\psi(\tilde{M}) = \psi(M)$ and thus $M = \tilde{M}$. $\square$

Recall that $J_\Lambda^*(x) \leq 1$ if and only if

$$|x|_{(1)} + \ldots + |x|_{(j)} \leq \lambda_1 + \ldots + \lambda_j, \qquad j = 1, \ldots, p. \tag{A.2}$$

The following result follows from the proof of Proposition 2.1.

Proposition A.2. Assume $x \in \mathbb{R}^p$ satisfies $J_\Lambda^*(x) \leq 1$ and let $b \in \mathbb{R}^p$. Then, x belongs to $\partial J_\Lambda(b)$ if and only if the following three conditions hold true:

1. If $b_i \neq 0$, then $\mathrm{sign}(x_i) = \mathrm{sign}(b_i)$,
2. If $|b_i| > |b_j|$ then $|x_i| \geq |x_j|$,
3. The equalities hold in Eq. (A.2) for $j \in \{n_1, n_2, \ldots, n_k\}$, where $n_j = |\{i : |M_i| \geq k + 1 - j\}|$ with $(M_1, \ldots, M_p)' = \mathrm{patt}(b)$.

A.3. Proof of Theorem 3.1

Proof of Theorem 3.1. *Necessity.* Let us assume that there exists $\hat{\beta} \in S_{X,\Lambda}(Y)$ with $\mathrm{patt}(\hat{\beta}) = M$. Consequently, $\hat{\beta} = U_M s$ for some $s \in \mathbb{R}^{k+}$.

By Proposition 2.2, $X'(Y - X\hat{\beta}) \in \partial J_\Lambda(\hat{\beta}) = \partial J_\Lambda(M)$. Multiplying this inclusion by U'_M, due to Eq. (2.3), we get $\tilde{X}'_M(Y - X\hat{\beta}) = \tilde{\Lambda}_M$ and so

$$\tilde{X}'_M Y - \tilde{\Lambda}_M = \tilde{X}'_M X\hat{\beta} = \tilde{X}'_M \tilde{X}_M s. \tag{A.3}$$

The positivity condition is proven.

We apply $(\tilde{X}'_M)^+$ from the left to Eq. (A.3) and use the fact that $\tilde{P}_M = (\tilde{X}'_M)^+ \tilde{X}'_M$ is the projection onto $\mathrm{col}(\tilde{X}_M)$. Since $X\hat{\beta} \in \mathrm{col}(\tilde{X}_M)$, we have $\tilde{P}_M X\hat{\beta} = X\hat{\beta}$. Thus,

$$\tilde{P}_M Y - (\tilde{X}'_M)^+ \tilde{\Lambda}_M = X\hat{\beta}.$$

The above equality gives the subdifferential condition:

$$\partial J_\Lambda(M) \ni X'(Y - X\hat{\beta}) = X'(Y - (\tilde{P}_M Y - (\tilde{X}'_M)^+ \tilde{\Lambda}_M)) \tag{A.4}$$
$$= X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M + X'(I_n - \tilde{P}_M)Y = \pi.$$

Sufficiency. Assume that the positivity condition and the subdifferential conditions hold true. Then, by the positivity condition, one may pick $s \in \mathbb{R}^{k+}$ for which

$$\tilde{\Lambda}_M = \tilde{X}'_M Y - \tilde{X}'_M \tilde{X}_M s. \tag{A.5}$$

Let us show that $U_M s \in S_{X,\Lambda}(Y)$. By definition of U_M, we have $\mathrm{patt}(U_M s) = M$ thus $\partial J_\Lambda(U_M s) = \partial J_\Lambda(M)$. Moreover, using Eqs. (A.4) and (A.5) one may deduce

$$\partial J_\Lambda(U_M s) \ni \pi = X'(Y - \tilde{P}_M Y + (\tilde{X}'_M)^+ \tilde{\Lambda}_M)$$
$$= X'(Y - \tilde{P}_M Y + (\tilde{X}'_M)^+ (\tilde{X}_M Y - \tilde{X}'_M \tilde{X}_M s))$$
$$= X'(Y - XU_M s).$$

Consequently $U_M s \in S_{X,\Lambda}(Y)$. $\quad\square$

A.4. Proof of Corollaries 3.2 and 3.3

Proof of Corollary 3.2. We will prove the implications $i) \Rightarrow iii) \Rightarrow iv) \Rightarrow ii) \Rightarrow i)$.

$i) \Rightarrow iii)$: Suppose there exist $\Lambda \in \mathbb{R}^{p+}$ and $\hat{\beta} \in S_{X,\Lambda}(X\beta)$ such that $\mathrm{patt}(\hat{\beta}) = \mathrm{patt}(\beta)$. Then, by Theorem 3.1 and since $\varepsilon = 0$, the subdifferential condition reads as: $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$.

$iii) \Rightarrow iv)$: The condition $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$ remains true when $\Lambda \in \mathbb{R}^{p+}$ is scaled by a scalar parameter $\alpha > 0$. Indeed

$$X'(\tilde{X}'_M)^+ (\alpha\tilde{\Lambda})_M = \alpha X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \alpha\partial J_\Lambda(M) = \partial J_{\alpha\Lambda}(M)$$

Therefore, up to scaling of Λ, for any $\lambda_1^0 > 0$ there exists $\Lambda \in \mathbb{R}^{p+}$ with $\lambda_1 < \lambda_1^0$, such that $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \partial J_\Lambda(M)$.

$iv) \Rightarrow ii)$: To prove that SLOPE can recover the pattern of β in the noiseless case, it remains to show that the positivity condition holds. Since $\beta = U_M s$ for some $s \in \mathbb{R}^{k+}$, where $k = \|M\|_\infty$, and $Y = X\beta$, we have

$$\tilde{X}'_M Y - \tilde{\Lambda}_M = \tilde{X}'_M \tilde{X}_M s - \tilde{\Lambda}_M.$$

Therefore, for λ_1 sufficiently small, we have $\tilde{X}'_M Y - \tilde{\Lambda}_M \in \tilde{X}'_M \tilde{X}_M \mathbb{R}^{k+}$, which proves the positivity condition.

$ii) \Rightarrow i)$: This implication follows directly by construction. $\quad\square$

Proof of Corollary 3.3. The proof of Corollary 3.3 follows by an analogous argument. $\quad\square$

A.5. Proof of Theorem 4.1

Lemma A.3. Let $0 \neq b \in \mathbb{R}^p$ and $M = \mathrm{patt}(b)$. Then the smallest affine space containing $\partial J_\Lambda(b)$ is $\mathrm{aff}(\partial J_\Lambda(b)) = \{v \in \mathbb{R}^p : U'_M v = \tilde{\Lambda}_M\}$.

Proof. According to Proposition 2.1 we have

$$\mathrm{aff}(\partial J_\Lambda(b)) \subset \{v \in \mathbb{R}^p : U'_M v = \tilde{\Lambda}_M\}.$$

Moreover, according to Theorem 4 in [35] we have

$$\dim(\mathrm{aff}(\partial J_\Lambda(b))) = \|M\|_\infty = \dim(\{v \in \mathbb{R}^p : U'_M v = \tilde{\Lambda}_M\}),$$

which achieves the proof. $\quad\square$

Proof of Theorem 4.1. *(i) Sharpness of the upper bound.* According to Theorem 3.1, pattern recovery by SLOPE is equivalent to have simultaneously the positivity condition and the subdifferential condition satisfied. The upper bound Eq. (4.2) coincides with the probability of the subdifferential condition. Thus to prove that this upper bound is sharp, it remains to show that the probability of the positivity condition tends to 1 when r tends to ∞. Clearly the upper bound is reached when $\tilde{\Lambda}_M \notin \mathrm{col}(\tilde{X}'_M)$ thus we assume hereafter that $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)$. Recall that $\beta^{(r)} = U_M s^{(r)}$ for $s^{(r)} \in \mathbb{R}^{k+}$ and thus $\tilde{X}'_M Y^{(r)} = \tilde{X}'_M \tilde{X}_M s^{(r)} + \tilde{X}'_M \varepsilon$. As $\tilde{X}'_M (\tilde{X}'_M)^+ = \tilde{X}'_M \tilde{X}_M (\tilde{X}'_M \tilde{X}_M)^+$ is the projection on $\mathrm{col}(\tilde{X}'_M)$, we obtain

$$
\begin{aligned}
\tilde{X}'_M Y^{(r)} - \alpha_r \tilde{\Lambda}_M &= \tilde{X}'_M \tilde{X}_M s^{(r)} - \alpha_r \tilde{\Lambda}_M + \tilde{X}'_M \varepsilon \\
&= \tilde{X}'_M \tilde{X}_M s^{(r)} - \alpha_r \tilde{X}'_M \tilde{X}_M (\tilde{X}'_M \tilde{X}_M)^+ \tilde{\Lambda}_M + \tilde{X}'_M \tilde{X}_M (\tilde{X}'_M \tilde{X}_M)^+ \tilde{X}'_M \varepsilon \\
&= \tilde{X}'_M \tilde{X}_M \Delta_r \left(\frac{1}{\Delta_r} s^{(r)} - \frac{\alpha_r}{\Delta_r} (\tilde{X}'_M \tilde{X}_M)^+ \tilde{\Lambda}_M + \frac{1}{\Delta_r} (\tilde{X}'_M \tilde{X}_M)^+ \tilde{X}'_M \varepsilon \right).
\end{aligned}
$$

Note that by the assumption on Δ_r:

- the vector $s^{(r)}/\Delta_r \in \mathbb{R}^{k+}$ is (component-wise) larger than or equal to $(k, \ldots, 1)$;
- $\lim_{r \to \infty} \alpha_r / \Delta_r = 0$ and $\lim_{r \to \infty} 1/\Delta_r = 0$.

Consequently, for r large enough we have

$$
\tilde{X}'_M Y^{(r)} - \alpha_r \tilde{\Lambda}_M \in \tilde{X}'_M \tilde{X}_M \mathbb{R}^{k+}.
$$

Since this fact is true for any realization of ε, one may deduce that

$$
\lim_{r \to \infty} \mathbb{P}\left(\tilde{X}'_M Y^{(r)} - \alpha_r \tilde{\Lambda}_M \in \tilde{X}'_M \tilde{X}_M \mathbb{R}^{k+} \right) = 1.
$$

(ii) Pattern consistency. In the proof of the previous part, we see that positivity condition occurs when r is sufficiently large. Thus it remains to prove that subdifferential condition occurs as $r \to \infty$ when $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$. First we observe that

$$
X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M + \frac{1}{\alpha_r} X'(I_n - \tilde{P}_M)\varepsilon \xrightarrow{r \to \infty} X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M. \tag{A.6}
$$

Note by Lemma A.3 that $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M + \alpha_r^{-1} X'(I_n - \tilde{P}_M)\varepsilon \in \mathrm{aff}(\partial J_\Lambda(M))$. Indeed, since $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)$ we have

$$
\underbrace{U'_M X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M}_{=\tilde{\Lambda}_M} + \frac{1}{\alpha_r} \underbrace{U'_M X'(I_n - \tilde{P}_M)\varepsilon(\omega)}_{=0} = \tilde{\Lambda}_M.
$$

The second term above is zero due to the fact that $(I_n - \tilde{P}_M)$ is an orthogonal projection onto $\mathrm{col}(\tilde{X}_M)^\perp$. When $X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M \in \mathrm{ri}(\partial J_\Lambda(M))$, due to Eq. (A.6), one may deduce that for sufficiently large r we have

$$
X'(\tilde{X}'_M)^+ \tilde{\Lambda}_M + \frac{1}{\alpha_r} X'(I_n - \tilde{P}_M)\varepsilon \in \partial J_\Lambda(M).
$$

Consequently, when r is sufficiently large, both the positivity and the subdifferential conditions occur which, by Theorem 3.1, concludes the proof. □

A.6. Proofs from Section 4.2

In this part we give proofs of Theorems 4.2 and 4.3. They are preceded by a series of simple lemmas. For reader's convenience we recall the setting of Section 4.2.

A. $\varepsilon_n = (\epsilon_1, \ldots, \epsilon_n)'$, where $(\epsilon_i)_i$ are i.i.d. centered with finite variance σ^2.

B1. $n^{-1} X'_n X_n \xrightarrow{\mathbb{P}} C > 0$.

B2. $\dfrac{\max_{i=1,\ldots,n} |X_{ij}^{(n)}|}{\sqrt{\sum_{i=1}^n (X_{ij}^{(n)})^2}} \xrightarrow{\mathbb{P}} 0$, where $X_n = (X_{ij}^{(n)})_{ij}$, for each $j = 1, \ldots, p$.

B'. Rows of X_n are i.i.d. distributed as ξ, where ξ is a random vector whose components are linearly independent a.s. and such that $\mathbb{E}[\xi_i^2] < \infty$ for $i = 1, \ldots, p$.

C. $(X_n)_n$ and $(\epsilon_n)_n$ are independent.

We consider a sequence of tuning parameters $(\Lambda_n)_n$ defined by $\Lambda_n = \alpha_n \Lambda$, where $\Lambda \in \mathbb{R}^{p+}$ is fixed and $(\alpha_n)_n$ is a sequence of positive numbers.

To ease the notation, we write the clustered matrices and clustered parameters without the subscript indicating the model M, *i.e.*, $\tilde{\Lambda} = U'_{|M|\downarrow} \Lambda$, $\tilde{\Lambda}_n = \alpha_n \tilde{\Lambda}$ and $\tilde{X}_n = X_n U_M$.

Lemma A.4.

(i) Under A, B1, B2 and C,

$$
\frac{1}{\sqrt{n}} X'_n \varepsilon_n \xrightarrow{d} Z \sim \mathrm{N}(0, \sigma^2 C). \tag{A.7}
$$

(ii) Under A, B1 and C,

$$\frac{1}{n} X_n' \varepsilon_n \xrightarrow{\mathbb{P}} 0. \tag{A.8}$$

(iii) Under A, B' and C,

$$0 < \limsup_{n \to \infty} \frac{\|X_n' \varepsilon_n\|_\infty}{\sqrt{n \log \log n}} < \infty \qquad \text{a.s.} \tag{A.9}$$

Proof. Proof of Eq. (A.7). It is enough to show that for any Borel subset $A \subset \mathbb{R}^p$ one has

$$\mathbb{P}\left(\frac{1}{\sqrt{n}} X_n' \varepsilon_n \in A \mid (X_n)_n \right) \xrightarrow{\mathbb{P}} \mathbb{P}(Z \in A). \tag{A.10}$$

Since both sides above are bounded, the convergence in probability implies convergence in L^1 and therefore establishes Eq. (A.7). To show Eq. (A.10) we will prove that for any subsequence $(n_k)_k$, there exists a sub-subsequence $(n_{k_l})_l$ for which, as $l \to \infty$,

$$\mathbb{P}\left(\frac{1}{\sqrt{n_{k_l}}} X_{n_{k_l}}' \varepsilon_{n_{k_l}} \in A \mid (X_n)_n \right) \xrightarrow{a.s.} \mathbb{P}(Z \in A). \tag{A.11}$$

Let $\mathbb{P}_\mathbf{X}$ denote the regular conditional probability $\mathbb{P}(\cdot \mid (X_n)_n)$ on $(\Omega, \mathcal{F})$. By assumptions B1 and B2, from sequences $(n_k)_k$ one can choose a subsequence $(n_{k_l})_l$ for which

$$\frac{1}{n_{k_l}} X_{n_{k_l}}' X_{n_{k_l}} \xrightarrow{a.s.} C > 0 \quad \text{and} \quad \frac{\max_{i=1,\dots,n_{k_l}} |X_{ij}^{(n_{k_l})}|}{\sqrt{\sum_{i=1}^{n_{k_l}} (X_{ij}^{(n_{k_l})})^2}} \xrightarrow{a.s} 0, \quad j = 1,\dots,p.$$

We have

$$\mathrm{Var}_\mathbf{X}\left(\frac{1}{\sqrt{n_{k_l}}} X_{n_{k_l}}' \varepsilon_{n_{k_l}} \right) = \frac{1}{n_{k_l}} \mathbb{E}\left[X_{n_{k_l}}' \varepsilon_{n_{k_l}} \varepsilon_{n_{k_l}}' X_{n_{k_l}} \mid (X_n)_n \right]$$

$$= \frac{1}{n_{k_l}} X_{n_{k_l}}' \mathbb{E}\left[\varepsilon_{n_{k_l}} \varepsilon_{n_{k_l}}' \right] X_{n_{k_l}} = \frac{\sigma^2}{n_{k_l}} X_{n_{k_l}}' X_{n_{k_l}} \xrightarrow{a.s.} \sigma^2 C > 0,$$

and one can apply multivariate Lindeberg-Feller CLT on the space $(\Omega, \mathcal{F}, \mathbb{P}_\mathbf{X})$ to prove Eq. (A.11). Alternatively, the same result follows from [59, Corollary 1.1][1], which concerns more general Central Limit Theorem for linearly negative quadrant dependent variables with weights forming a triangular array (in particular assumption B2 coincides with [59, (1.8)]).

For (ii) we observe that previous derivations imply that $\mathrm{Var}_\mathbf{X}(n^{-1} X_n' \varepsilon_n) \xrightarrow{\mathbb{P}} 0$. We deduce that $\mathbb{P}_\mathbf{X}(n^{-1} \|X_n' \varepsilon_n\| > \delta) \xrightarrow{\mathbb{P}} 0$ and hence (ii) follows upon averaging over $(X_n)_n$.

Eq. (A.9) is the law of iterated logarithm for an i.i.d. sequence $(\xi_i \epsilon_i)_i$. $\square$

Lemma A.5. Let $M = \mathrm{patt}(\beta)$. Assume $\alpha_n/n \to 0$.

(i) Under A, B1 and C, the positivity condition is satisfied for large n with high probability.
(ii) Under A, B' and C, the positivity condition is almost surely satisfied for large n.

Proof. If $M = 0$, then the positivity condition is trivially satisfied. Thus, we consider $M \neq 0$.

(i) Since $\tilde{X}_n' \tilde{X}_n$ is invertible for large n with high probability, the positivity condition is equivalent to

$$s_n := (\tilde{X}_n' \tilde{X}_n)^{-1} [\tilde{X}_n' Y_n - \tilde{\Lambda}_n] \in \mathbb{R}^{k+}.$$

Let $s_0 \in \mathbb{R}^{k+}$ be defined through $\beta = U_M s_0$, where $k = \|M\|_\infty$. We will show that if $\alpha_n/n \to 0$, then $s_n \xrightarrow{\mathbb{P}} s_0$. Since $\mathbb{R}^{k+}$ is an open set, this will imply that for large n with high probability, the positivity condition is satisfied.

First we rewrite s_n as

$$s_n = (\tilde{X}_n' \tilde{X}_n)^{-1} \tilde{X}_n' Y_n - \alpha_n (\tilde{X}_n' \tilde{X}_n)^{-1} \tilde{\Lambda}.$$

Since $\beta = U_M s_0$, we conclude $X_n \beta = X_n U_M s_0 = \tilde{X}_n s_0$, so the linear regression model takes the form $Y_n = \tilde{X}_n s_0 + \varepsilon_n$. Thus, $(\tilde{X}_n' \tilde{X}_n)^{-1} \tilde{X}_n' Y_n$ is the OLS estimator of s_0.

By assumption B and Lemma A.4, we deduce that

$$(\tilde{X}_n' \tilde{X}_n)^{-1} \tilde{X}_n' Y_n = s_0 + (n^{-1} \tilde{X}_n' \tilde{X}_n)^{-1} U_M \frac{1}{n} X_n' \varepsilon_n \xrightarrow{\mathbb{P}} s_0 + [(U_M' C U_M)^{-1} U_M] 0 = s_0.$$

To complete the proof, we note that

$$\alpha_n (\tilde{X}_n' \tilde{X}_n)^{-1} \tilde{\Lambda} = \frac{\alpha_n}{n} \left[n (\tilde{X}_n' \tilde{X}_n)^{-1} \tilde{\Lambda} \right] \xrightarrow{\mathbb{P}} 0 \left[(U_M' C U_M)^{-1} \tilde{\Lambda} \right] = 0.$$

---

[1] For our application, the assumption of nonnegative weights in [59, Corollary 1.1] is not essential.

(ii) If one assumes B' instead of B1, then $n^{-1} X'_n X_n \xrightarrow{a.s.} C$ and by Eq. (A.9), $n^{-1} X'_n \varepsilon_n \xrightarrow{a.s.} 0$. The result follows along the same lines as (i).

$\square$

For $M \neq 0$ we denote

$$\pi_n^{(1)} = X'_n (\tilde{X}'_n)^+ \tilde{\Lambda}_n, \qquad \pi_n^{(2)} = X'_n (I_n - \tilde{P}_n) Y_n,$$
$$\pi_n = \pi_n^{(1)} + \pi_n^{(2)},$$

which simplifies in the $M = 0$ case to $\pi_n = \pi_n^{(2)} = X'_n Y_n$.

Recall that the subdifferential condition is equivalent to $J^*_{\Lambda_n}(\pi_n) \leq 1$ and $\tilde{\Lambda}_n \in \mathrm{col}(\tilde{X}'_M)$ and the latter is satisfied in our setting. Since $\alpha J_\Lambda = J_{\alpha\Lambda}$, the subdifferential condition is satisfied if and only if

$$1 \geq J^*_\Lambda(\alpha_n^{-1} \pi_n) = J^*_\Lambda\left(\alpha_n^{-1} \pi_n^{(1)} + \frac{\sqrt{n}}{\alpha_n} n^{-1/2} \pi_n^{(2)}\right).$$

In view of results shown below, $\alpha_n^{-1} \pi_n^{(1)}$ converges almost surely, while $n^{-1/2} \pi_n^{(2)}$ converges in distribution to a Gaussian vector. Thus, the pattern recovery properties of SLOPE estimator strongly depend on the behavior of the sequence $(\alpha_n/\sqrt{n})_n$.

Lemma A.6. (a)

(i) Assume A, B1 and C. If $M \neq 0$, then

$$\frac{1}{\alpha_n} \pi_n^{(1)} \xrightarrow{\mathbb{P}} C U_M (U'_M C U_M)^{-1} \tilde{\Lambda}.$$

(ii) Assume A, B1, B2 and C. The sequence $\left(n^{-1/2} \pi_n^{(2)}\right)_n$ converges in distribution to a Gaussian vector Z with

$$Z \sim \mathrm{N}\left(0, \sigma^2 \left[C - C U_M (U'_M C U_M)^{-1} U'_M C\right]\right).$$

(iii) Assume A, B1 and C. If $\lim_{n\to\infty} \alpha_n/\sqrt{n} = \infty$, then $\alpha_n^{-1} \pi_n^{(2)} \xrightarrow{\mathbb{P}} 0$.

(b) Assume A, B' and C.

(i') If $M \neq 0$, then

$$\frac{1}{\alpha_n} \pi_n^{(1)} \xrightarrow{a.s.} C U_M (U'_M C U_M)^{-1} \tilde{\Lambda}.$$

(ii') If $\lim_{n\to\infty} \alpha_n/\sqrt{n \log\log n} = \infty$, then $\alpha_n^{-1} \pi_n^{(2)} \xrightarrow{a.s.} 0$.

Proof.

(i) Assumption B1 implies that

$$X'_n \tilde{X}_n (\tilde{X}'_n \tilde{X}_n)^{-1} = \frac{1}{n} X'_n X_n U_M (U'_M n^{-1} X'_n X_n U_M)^{-1} \xrightarrow{\mathbb{P}} C U_M (U'_M C U_M)^{-1}.$$

(ii) When $\beta = U_M s_0$, then the linear regression model takes the form $Y_n = \tilde{X}_n s_0 + \varepsilon_n$. Since $\tilde{P}_n$ is the projection matrix onto $\mathrm{col}(\tilde{X}_n)$, we have $(I_n - \tilde{P}_n)\tilde{X}_n = 0$. Thus,

$$n^{-1/2} \pi_n^{(2)} = n^{-1/2} X'_n (I_n - \tilde{P}_n) Y_n = n^{-1/2} X'_n (I_n - \tilde{P}_n) \varepsilon_n$$
$$= \left[I_p - X'_n X_n U_M (U'_M X'_n X_n U_M)^{-1} U'_M\right] \left[n^{-1/2} X'_n \varepsilon\right].$$

By assumption B1 we have,

$$n^{-1} X'_n X_n U_M (U'_M n^{-1} X'_n X_n U_M)^{-1} U'_M \xrightarrow{\mathbb{P}} C U_M (U'_M C U_M)^{-1} U'_M. \tag{A.12}$$

Thus, by Lemma A.4(i) and Slutsky's theorem, we obtain (ii). (iii) follows similarly as Lemma A.4(ii): with the aid of Eq. (A.12) we show that $\mathrm{Var}_{\mathbf{X}}(\alpha_n^{-1} \pi_n^{(2)}) \xrightarrow{\mathbb{P}} 0$, which implies that conditionally on $(X_n)_n$ we have $\alpha_n^{-1} \pi_n^{(2)} \xrightarrow{\mathbb{P}_{\mathbf{X}}} 0$.

Assumption B' implies that $n^{-1} X'_n X_n \xrightarrow{a.s.} C$ and thus (i') is proven in the same way as (i). (ii') follows from Eq. (A.9).

$\square$

Proof of Theorem 4.2. (i) is a direct consequence of Lemmas A.5 and A.6. Since positivity condition is satisfied for large n with high probability, for (ii) we have with $M = \text{patt}(\beta)$,

$$\lim_{n \to \infty} \mathbb{P}\big(\text{patt}(\hat{\beta}_n^{\text{SLOPE}}) = M\big) = \lim_{n \to \infty} \mathbb{P}\big(\pi_n \in \partial J_{\alpha_n \Lambda}(M)\big) = \lim_{n \to \infty} \mathbb{P}\big(\alpha_n^{-1} \pi_n \in \partial J_\Lambda(M)\big) \tag{A.13}$$

$$\geq \lim_{n \to \infty} \mathbb{P}\big(\alpha_n^{-1} \pi_n \in \text{ri}(\partial J_\Lambda(M))\big) = 1,$$

where in the last equality we use the Portmanteau Theorem, assumption Eq. (4.6) and the fact that sequence $(\alpha_n^{-1} \pi_n)_n$ converges in distribution to $CU_M(U'_M CU_M)^{-1}$ if and only if $\alpha_n / \sqrt{n} \to \infty$.

Condition Eq. (4.7) implies that $CU_M(U'_M CU_M)^{-1} \in \partial J_\Lambda(M)$. Since $(\alpha_n^{-1} \pi_n)_n$ converges in probability to $CU_M(U'_M CU_M)^{-1}$, the necessity of this condition is explained by Eq. (A.13). $\square$

Proof of Theorem 4.3. By Lemma A.5, the positivity condition is satisfied for large n almost surely. By Lemma A.6 (i) and (iii), we have

$$a_n := \frac{1}{\alpha_n} \pi_n \xrightarrow{a.s.} CU_M(U'_M CU_M)^{-1} \tilde{\Lambda} =: a_0.$$

It is easy to see that $U'_M a_n = \tilde{\Lambda}$. By the condition $a_0 \in \text{ri}(J_\Lambda(M))$ it follows that $a_n \in J_\Lambda(M)$ almost surely for sufficiently large n. Therefore $\pi_n \in J_{\Lambda_n}(M)$ for large n almost surely and thus the subdifferential condition is also satisfied. $\square$

Appendix B. Refined results on strong consistency of the SLOPE pattern

In this appendix we aim to give weaker assumptions on the design matrix than condition B', but which ensure the almost sure convergence of the pattern of $\hat{\beta}_n^{\text{SLOPE}}$.

A'. $\epsilon_n = (\epsilon_1, \ldots, \epsilon_n)'$, where $(\epsilon_i)_i$ are independent random variables such that

$$\mathbb{E}[\epsilon_n] = 0 \quad \text{and} \quad \text{Var}(\epsilon_n) = \sigma^2 \quad \text{for all } n, \quad \text{and} \quad \sup_n \mathbb{E}[|\epsilon_n|^r] < \infty \tag{B.1}$$

for some $r > 2$.

B''. A sequence of design matrices $X_1, X_2, \ldots$ satisfies the condition

$$\frac{1}{n} X'_n X_n \xrightarrow{a.s.} C, \tag{B.2}$$

where C is a deterministic positive definite symmetric $p \times p$ matrix.

With $X_n = \left(X_{ij}^{(n)} \right)_{ij}$,

$$\lim_{n \to \infty} \frac{(\log n)^\rho}{\sqrt{n}} \sup_{i,j} \left| X_{ij}^{(n)} \right| = 0 \quad \text{a.s. for all } \rho > 0 \tag{B.3}$$

and there exist nonnegative random variables $(c_i)_i$, constants $d > 2/r$ and $m_0 \in \mathbb{N}$ such that for $n > m \geq m_0$,

$$\sup_j \left[\sum_{i=1}^m \left(X_{ij}^{(n)} - X_{ij}^{(m)} \right)^2 + \sum_{i=m+1}^n \left(X_{ij}^{(n)} \right)^2 \right] \leq \left(\sum_{i=m+1}^n c_i \right)^d \quad \text{a.s.,} \tag{B.4}$$

$$\left(\sum_{i=m_0}^n c_i \right)^d = O(n) \quad \text{a.s.} \tag{B.5}$$

C. $(X_n)_n$ and $(\epsilon_n)_n$ are independent.

We note that conditions Eqs. (B.3) and (B.4) are trivially satisfied in the i.i.d. rows setting of Remark 4.4 or assumption B'. The main ingredient of the proof of the strong pattern consistency is the law of iterated logarithm Eq. (A.9) which holds trivially under B'. Below, we establish the same result under more general B''. The technical assumption Eq. (B.4) is a kind of weak continuity assumption on the rows of X_n as it says that the ℓ_2-distance between jth rows of X_n and X_m should not be too large.

Lemma B.1. Assume A', B'' and C. Then

$$\limsup_{n \to \infty} \frac{\|X'_n \epsilon_n\|_\infty}{\sqrt{n \log \log n}} < \infty \quad \text{a.s.} \tag{B.6}$$

Proof. In view of Eq. (4.5) we have for $j = 1, \ldots, p$,

$$n^{-1} A_n^{(j)} := n^{-1} \sum_{i=1}^n \left(X_{ij}^{(n)} \right)^2 = \left(n^{-1} X'_n X_n \right)_{jj} \xrightarrow{a.s.} C_{jj} > 0. \tag{B.7}$$

We apply the general law of iterated logarithm for weights forming a triangular array from [60]. The result follows directly from [60, Theorem 1]. Defining $a_{ni}^{(j)} := X_{ij}^{(n)}$ for $i = 1, \dots, n$, $j = 1, \dots, p$, $n \geq 1$ and 0 otherwise, we have

$$(X_n' \varepsilon_n)_j = \sum_{i=-\infty}^{\infty} a_{ni}^{(j)} \epsilon_i$$

and therefore we fall within the framework of [60, Eq. (1.3)]. Then, Eqs. (B.1)–(B.5) coincide with [60, (1.2), (1.6), (1.7), (1.8)] respectively. Let $\mathbb{P}(\cdot | (X_n)_n)$ be a regular conditional probability. Then, applying [60, Theorem 1 (i)] on the probability space $(\Omega, \mathcal{F}, \mathbb{P}_X)$ to our sequence we obtain that for $j = 1, \dots, n$,

$$\mathbb{P}\left(\limsup_{n \to \infty} \frac{\left| (X_n' \varepsilon_n)_j \right|}{\sqrt{2 A_n^{(j)} \log \log A_n^{(j)}}} \leq \sigma \Big| (X_n)_n \right) = 1 \qquad \text{a.s.}$$

Averaging over $(X_n)_n$ and using Eq. (B.7) again, we obtain the assertion. $\square$

Theorem B.2. Assume A', B" and C. Suppose that $(\alpha_n)_n$ satisfies

$$\lim_{n \to \infty} \frac{\alpha_n}{n} = 0 \qquad \text{and} \qquad \lim_{n \to \infty} \frac{\alpha_n}{\sqrt{n \log \log n}} = \infty.$$

If Eq. (4.6) is satisfied, then $\text{patt}(\hat{\beta}_n^{\text{SLOPE}}) \xrightarrow{a.s.} \text{patt}(\beta)$.

Comments:

a) Under reasonable assumptions (see *e.g.* [60, Theorem 1 (iii)]) one can show that

$$\limsup_{n \to \infty} \frac{\|X_n' \varepsilon_n\|_\infty}{\sqrt{n \log \log n}} > 0 \qquad \text{a.s.}$$

Since $\alpha_n^{-1} X_n' \varepsilon_n \xrightarrow{a.s.} 0$ is necessary for the a.s. pattern recovery, we can show that the condition $\alpha_n / \sqrt{n \log \log n} \to \infty$ cannot be weakened. Thus, the gap between the convergence in probability and the a.s. convergence is integral to the problem and in general cannot be reduced.

b) One can relax assumption B" by imposing stronger conditions on the error ε_n. *E.g.* if ε_n is Gaussian, then one can use results from [61]. We note that [61] offers a very similar result as [60], but their assumptions are not quite comparable, see [61, Section 3 i)] for detailed discussion.

c) For Gaussian errors, one can consider a more general setting where one does not assume any relation between ε_n and ε_{n+1}, *i.e.*, the error need not be incremental. For orthogonal design such approach was taken in [33]. It is proved there that one obtains the a.s. SLOPE pattern consistency with the second limit condition of Theorem B.2 replaced by $\lim_{n \to \infty} \alpha_n / \sqrt{n \log n} = \infty$. This result can be generalized to non-orthogonal designs.

Appendix C. Strong consistency of SLOPE estimator

Lemma C.1. Assume that $\varepsilon_n = (\epsilon_1, \dots, \epsilon_n)'$ with $(\epsilon)_i$ i.i.d., centered and having finite variance. Suppose

$$\frac{1}{n} X_n' X_n \xrightarrow{a.s.} C > 0. \tag{C.1}$$

and that $(\varepsilon_n)_n$ and $(X_n)_n$ are independent. Then $n^{-1} X_n' \varepsilon_n \xrightarrow{a.s.} 0$.

Proof. Let $\mathbb{P}(\cdot \mid (X_n)_n)$ denote the regular conditional probability. By [62, Th. 1.1] applied to a sequence $(n^{-1} X_n' \varepsilon_n)_j$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P}(\cdot \mid (X_n)_n))$, we obtain

$$\mathbb{P}\left(\lim_{n \to \infty} n^{-1} (X_n' \varepsilon_n)_j = 0 \mid (X_n)_n \right) = 1, \qquad j = 1, \dots, p, \qquad \text{a.s.}$$

Thus, applying the expectation to both sides above we obtain the assertion. $\square$

Theorem C.2. Assume that $Y_n = X_n \beta + \varepsilon_n$, where $\beta \in \mathbb{R}^p$, $\varepsilon_n = (\epsilon_1, \dots, \epsilon_n)'$ with $(\epsilon)_i$ i.i.d., centered and finite variance. Suppose Eq. (C.1) and that $(\varepsilon_n)_n$ and $(X_n)_n$ are independent. Let $\Lambda_n = (\lambda_1^{(n)}, \dots, \lambda_p^{(n)})'$. Then, for large n, $S_{X_n, \Lambda_n}(Y_n) = \{\hat{\beta}_n^{\text{SLOPE}}\}$ almost surely.

If $\beta \neq 0$, then $\hat{\beta}_n^{\text{SLOPE}} \xrightarrow{a.s.} \beta$ if and only if

$$\lim_{n \to \infty} \frac{\lambda_1^{(n)}}{n} = 0. \tag{C.2}$$

If $\beta = 0$ and Eq. (C.2) holds true, then $\hat{\beta}_n^{\text{SLOPE}} \xrightarrow{a.s.} 0$.

Proof of Theorem C.2. The assumption Eq. (C.1) implies that the matrix $X_n' X_n$ is positive definite for large n almost surely and hence ensuring that $\ker(X_n) = \{0\}$. It is known that under trivial kernel, the set of SLOPE minimizers contains one element only.

By Proposition 2.1, $\hat{\beta}_n^{\text{SLOPE}}$ is the SLOPE estimator of β in a linear regression model $Y_n = X_n \beta + \varepsilon_n$ if and only if for $\pi_n = X_n'(Y_n - X_n \hat{\beta}_n^{\text{SLOPE}})$ we have

$$J_\Lambda^*(\pi_n) \leq 1 \tag{C.3}$$

and

$$U_{M_n}' \pi_n = \tilde{\Lambda}_n, \tag{C.4}$$

where $M_n = \text{patt}(\hat{\beta}_n^{\text{SLOPE}})$ and $\tilde{\Lambda}_n = U_{|M_n|\downarrow}' \Lambda_n$. By the definition of π_n we have

$$\hat{\beta}_n^{\text{SLOPE}} = (X_n' X_n)^{-1} X_n' Y_n - (X_n' X_n)^{-1} \pi_n = \hat{\beta}_n^{\text{OLS}} - \left(\frac{1}{n} X_n' X_n\right)^{-1}\left(\frac{1}{n}\pi_n\right).$$

Since in our setting $\hat{\beta}_n^{\text{OLS}}$ is strongly consistent, $\hat{\beta}_n^{SLOPE} \xrightarrow{a.s.} \beta$ if and only if $(n^{-1} X_n' X_n)^{-1}(n^{-1}\pi_n) \xrightarrow{a.s.} 0$. In view of Eq. (C.1), we have $(n^{-1} X_n' X_n)^{-1}(n^{-1}\pi_n) \xrightarrow{a.s.} 0$ if and only if $n^{-1}\pi_n \xrightarrow{a.s.} 0$.

Assume $n^{-1}\lambda_1^{(n)} \to 0$. By Eq. (C.3) we have $\|\pi^n\|_\infty \leq \lambda_1^{(n)}$, which gives

$$\left\|\frac{\pi_n}{n}\right\|_\infty \leq \frac{\lambda_1^{(n)}}{n} \to 0.$$

Therefore, Eq. (C.2) implies that $\hat{\beta}_n^{SLOPE} \xrightarrow{a.s.} \beta$.

Now assume that $\beta \neq 0$ and $\hat{\beta}_n^{\text{SLOPE}}$ is strongly consistent, *i.e.*, $n^{-1}\pi_n \xrightarrow{a.s.} 0$. Then, Eq. (C.4) gives

$$p\|\pi_n\|_\infty \geq \|U_{M_n}' \pi_n\|_\infty = \|\tilde{\Lambda}_n\|_\infty \geq \lambda_1^{(n)} \tag{C.5}$$

provided $M_n \neq 0$. Applying Eq. (C.3) for $\hat{\beta}_n^{SLOPE} = 0$, we note that $M_n(\omega) = 0$ if and only if

$$J_{n^{-1}\Lambda_n}^*\left(n^{-1} X_n(\omega)' Y_n(\omega)\right) \leq 1.$$

In view of Lemma C.1, it can be easily verified that $n^{-1} X_n' Y_n \xrightarrow{a.s.} C\beta$. Since

$$\left\|\frac{1}{n}\pi_n\right\|_\infty \geq \left\|\frac{1}{n}\pi_n\right\|_\infty \mathbf{1}_{(M_n=0)} = \left\|\frac{1}{n} X_n' Y_n\right\|_\infty \mathbf{1}_{(M_n=0)},$$

we see that for $\beta \neq 0$, we have $M_n \neq 0$ for large n almost surely. Thus, for $\beta \neq 0$ we eventually obtain for large n

$$\frac{\lambda_1^{(n)}}{n} \leq p\left\|\frac{\pi_n}{n}\right\|_\infty \quad \text{a.s.}$$

$\square$

Appendix D. Geometric interpretation of $X'(\tilde{X}_M')^+ \tilde{\Lambda}_M$

Let $0 \neq \beta \in \mathbb{R}^p$ where $\text{patt}(\beta) = M$. For a SLOPE minimizer $\hat{\beta} \in S_{X,\alpha\Lambda}(X\beta)$ the following occurs:

$$\frac{1}{\alpha} X' X(\beta - \hat{\beta}) \in \partial J_\Lambda(\hat{\beta}).$$

In addition when $\text{patt}(\hat{\beta}) = M$, then the following facts hold:

- $\beta - \hat{\beta} \in \text{col}(U_M)$, so that $\frac{1}{\alpha} X' X(\beta - \hat{\beta}) \in X' X \text{col}(U_M)$.
- $\partial J_\Lambda(\hat{\beta}) = \partial J_\Lambda(M)$.

Therefore, the noiseless pattern recovery by SLOPE clearly implies that the vector space $X' X \text{col}(U_M) = \text{col}(X' \tilde{X}_M)$ intersects $\partial J_\Lambda(M)$. Actually, the vector $\bar{\Pi} = X'(\tilde{X}_M')^+ \tilde{\Lambda}_M$ appearing in Corollary 3.2 has a geometric interpretation given in Proposition D.1.

Proposition D.1. Let $X \in \mathbb{R}^{n \times p}$, $0 \neq M \in \mathcal{P}_p^{\text{SLOPE}}$ and $\Lambda \in \mathbb{R}^{p+}$. We recall that $\tilde{X}_M = X U_M$, $\tilde{\Lambda}_M = U_{|M|\downarrow}' \Lambda$ and $\bar{\Pi} = X'(\tilde{X}_M')^+ \tilde{\Lambda}_M$. We have the following statements:

i) If $\tilde{\Lambda}_M \notin \text{col}(\tilde{X}_M')$ then $\text{aff}(\partial J_\Lambda(M)) \cap \text{col}(X' \tilde{X}_M) = \emptyset$.
ii) If $\tilde{\Lambda}_M \in \text{col}(\tilde{X}_M')$ then $\text{aff}(\partial J_\Lambda(M)) \cap \text{col}(X' \tilde{X}_M) = \{\bar{\Pi}\}$.
iii) Pattern recovery by SLOPE in the noiseless case is equivalent to $\text{col}(X' \tilde{X}_M) \cap \partial J_\Lambda(M) \neq \emptyset$.
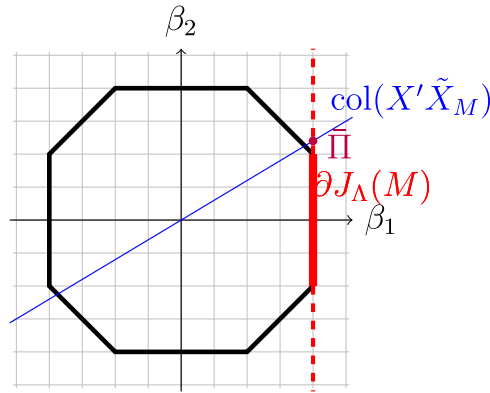
Fig. D.8. This figure illustrates $\bar{\Pi}$ in purple as the unique intersection point between $\mathrm{col}(X'\tilde{X}_M) = \mathrm{col}((1, 0.6)')$ in blue and $\mathrm{aff}(\partial J_\Lambda(M))$ in red. Since $\bar{\Pi} \notin \partial J_\Lambda(M) = \{4\} \times [-2, 2]$ then, in the noiseless case, SLOPE cannot recover $M = \mathrm{patt}(\beta) = (1,0)'$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Proof. i) We recall that, according to Lemma A.3, $\mathrm{aff}(\partial J_\Lambda(M)) = \{v \in \mathbb{R}^p : U'_M v = \tilde{\Lambda}_M\}$. If $\mathrm{aff}(\partial J_\Lambda(M)) \cap \mathrm{col}(X'\tilde{X}_M) \neq \emptyset$ then there exists $z \in \mathbb{R}^k$, where $k = \|M\|_\infty$, such that $X'\tilde{X}_M z \in \mathrm{aff}(\partial J_\Lambda(M))$. Consequently, $\tilde{\Lambda}_M = U'_M X'\tilde{X}_M z = \tilde{X}'_M \tilde{X}'_M z$ thus $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)$ which establishes i).

ii) If $\tilde{\Lambda}_M \in \mathrm{col}(\tilde{X}'_M)$ then $\bar{\Pi} \in \mathrm{aff}(\partial J_\Lambda(M))$. Indeed, since $\tilde{X}'_M (\tilde{X}'_M)^+$ is the projection on $\mathrm{col}(\tilde{X}'_M)$ we have

$$U'_M \bar{\Pi} = \tilde{X}'_M (\tilde{X}'_M)^+ \tilde{\Lambda}_M = \tilde{\Lambda}_M.$$

Moreover, since $\mathrm{col}((\tilde{X}'_M)^+) = \mathrm{col}(\tilde{X}_M)$ we deduce that $\bar{\Pi} \in \mathrm{col}(X'\tilde{X}_M)$. To prove that $\bar{\Pi}$ is the unique point in the intersection, let us prove that $\mathrm{col}(X'\tilde{X}_M) \cap \mathrm{col}(U_M)^\perp = \{0\}$. Indeed, if $v \in \mathrm{col}(X'\tilde{X}_M) \cap \mathrm{col}(U_M)^\perp$ then $v = X'\tilde{X}_M z$ for some $z \in \mathbb{R}^k$ and $U'_M v = 0$. Therefore, $\tilde{X}'_M \tilde{X}_M z = 0$, consequently $\tilde{X}_M z = 0$ and thus $v = \{0\}$. Finally, if $\Pi \in \mathrm{aff}(\partial J_\Lambda(M)) \cap \mathrm{col}(X'\tilde{X}_M)$ then $\Pi - \bar{\Pi} \in \mathrm{col}(X'\tilde{X}_M)$ and $U'_M(\Pi - \bar{\Pi}) = 0$ which implies that $\Pi = \bar{\Pi}$ and establishes ii).

According to Corollary 3.2, pattern recovery by SLOPE in the noiseless case is equivalent to $\bar{\Pi} \in \partial J_\Lambda(M)$ which is equivalent, by i) and ii), to $\mathrm{col}(X'\tilde{X}_M) \cap \partial J_\Lambda(M) \neq \emptyset$. $\square$

Example D.1.

- We observe on the right picture in Fig. 2 that the noiseless pattern recovery occurs when $\bar{\beta} = (5, 3)'$ (thus $M = \mathrm{patt}(\bar{\beta}) = (2, 1)'$). To corroborate this fact note that $\tilde{X}_M = X$ thus $\mathrm{col}(X'\tilde{X}_M) = \mathbb{R}^2$ and consequently $\mathrm{col}(X'\tilde{X}_M)$ intersects $\partial J_\Lambda(M)$.
- We observe on the left picture in Fig. 2 that the noiseless pattern recovery does not occur when $\beta = (5, 0)'$ (thus $M = \mathrm{patt}(\beta) = (1, 0)'$). To corroborate this fact, Fig. D.8 illustrates that $\mathrm{col}(X'\tilde{X}_M) = \mathrm{col}((1, 0.6)')$ does not intersect $\partial J_\Lambda(M) = \{4\} \times [-2, 2]$.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (6) (1974) 716–723.
[2] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (2) (1978) 461–464.
[3] D.P. Foster, E.I. George, The risk inflation criterion for multiple regression, Ann. Stat. 22 (4) (1994) 1947–1975.
[4] M. Bogdan, J.K. Ghosh, R.W. Doerge, Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci, Genetics 167 (2004) 989–999.
[5] J. Chen, Z. Chen, Extended Bayesian information criteria for model selection with large model spaces, Biometrika 95 (3) (2008) 759–771.
[6] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970a) 55–67.
[7] A.E. Hoerl, R.W. Kennard, Ridge regression: applications to nonorthogonal problems, Technometrics 12 (1) (1970b) 69–82.
[8] S. Chen, D. Donoho, Basis pursuit, in: Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, 1, 1994, pp. 41–44 vol.1.
[9] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. Ser. B 58 (1) (1996) 267–288.
[10] S. Vaiter, M. Golbabaee, J. Fadili, G. Peyré, Model selection with low complexity priors, Inf. Inference 4 (3) (2015) 230–287.
[11] M.R. Osborne, B. Presnell, B.A. Turlach, On the LASSO and its dual, J. Comput. Graph. Statist. 9 (2) (2000) 319–337.
[12] R.J. Tibshirani, The lasso problem and uniqueness, Electron. J. Stat. 7 (2013) 1456–1490.
[13] H.D. Bondell, B.J. Reich, Simultaneous factor selection and collapsing levels in anova, Biometrics 65 (2009) 169–177.
[14] G. Garcia-Donato, R. Paulo, Variable selection in the presence of factors: a model selection perspective, J. Amer. Statist. Assoc. (2021).
[15] A. Maj-Kańska, P. Pokarowski, A. Prochenka, Delete or merge regressors for linear model selection, Electron. J. Stat. 9 (2) (2015) 1749–1778.
[16] D. Pauger, H. Wagner, Bayesian effect fusion for categorical predictors, Bayesian Anal. 14 (2019) 341–369.
[17] B.G. Stokell, R.D. Shah, R.J. Tibshirani, Modelling high-dimensional categorical data using nonconvex fusion penalties, J. R. Stat. Soc. Ser. B. Stat. Methodol. 83 (3) (2021) 579–611.
[18] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[19] M. Bogdan, E. Van Den Berg, W. Su, E.J. Candès, Statistical estimation and testing via the sorted L1 norm, arXiv preprint arXiv:1310.1969. (2013).
[20] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E.J. Candès, SLOPE—adaptive variable selection via convex optimization, Ann. Appl. Stat. 9 (3) (2015) 1103–1140.
[21] X. Zeng, M.A.T. Figueiredo, Decreasing weighted sorted ℓ_1 regularization, IEEE Signal Process. Lett. 21 (10) (2014) 1240–1244.
[22] D. Brzyski, C.B. Peterson, P. Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, Controlling the rate of GWAS false discoveries, Genetics 205 (2017) 61–75.
[23] D. Brzyski, A. Gossmann, W. Su, M. Bogdan, Group SLOPE—adaptive selection of groups of predictors, J. Amer. Statist. Assoc. 114 (525) (2019) 419–433.

[24] M. Kos, M. Bogdan, On the asymptotic properties of SLOPE, Sankhya A 82 (2) (2020) 499–532.
[25] F. Abramovich, V. Grinshtein, High-dimensional classification by sparse logistic regression, IEEE Trans. Inform. Theory 65 (5) (2019) 3068–3079.
[26] P.C. Bellec, G. Lecué, A.B. Tsybakov, Slope meets lasso: improved oracle bounds and optimality, Ann. Statist. 46 (6B) (2018) 3603–3642.
[27] W. Su, E. Candès, SLOPE is adaptive to unknown sparsity and asymptotically minimax, Ann. Statist. 44 (3) (2016) 1038–1068.
[28] H.D. Bondell, B.J. Reich, Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR, Biometrics 64 (1) (2008) 115–123, 322–323.
[29] D.B. Sharma, H.D. Bondell, H.H. Zhang, Consistent group identification and variable selection in regression with correlated predictors, J. Comput. Graph. Statist. 22 (2) (2013) 319–340.
[30] P.J. Kremer, S. Lee, M. Bogdan, S. Paterlini, Sparse portfolio selection via the sorted ℓ_1-Norm, J. Bank. Financ. 110 (2020) 105687.
[31] M. Figueiredo, R. Nowak, Ordered weighted L1 regularized regression with strongly correlated covariates: theoretical aspects, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 775–783.
[32] P.J. Kremer, D. Brzyski, M. Bogdan, S. Paterlini, Sparse index clones via the sorted ℓ_1-norm, Quant. Financ. 22 (2) (2022) 349–366.
[33] T. Skalski, P. Graczyk, B. Kołodziejek, M. Wilczyński, Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal, Probab. Math. Statist. 42 (2) (2022) 283–302.
[34] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (1) (2005) 91–108.
[35] U. Schneider, P. Tardivel, The geometry of uniqueness, sparsity and clustering in penalized estimation, J. Mach. Learn. Res. 23 (331) (2022) 1–36.
[36] S. Vaiter, G. Peyré, J. Fadili, Model consistency of partly smooth regularizers, IEEE Trans. Inf. Theory 64 (3) (2017) 1725–1737.
[37] J.-J. Fuchs, On sparse representations in arbitrary redundant bases, IEEE Trans. Inf. Theory 50 (6) (2004) 1341–1344.
[38] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (3) (2006) 1436–1462.
[39] M.J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), IEEE Trans. Inf. Theory 55 (5) (2009) 2183–2202.
[40] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.
[41] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (476) (2006) 1418–1429.
[42] X. Dupuis, P. Tardivel, The solution path of slope, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 775–783.
[43] P. Tardivel, Étude de l'estimateur SLOPE par le prisme du schéma: Propriétés de parcimonie et d'appariement et calcul du chemin des solutions, Thèse d'habilitation, Université de Bourgogne, 2024. Translation available via the URL link, https://hal.science/hal-04528428.
[44] R. Negrinho, A. Martins, Orbit regularization, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, 27, Curran Associates, Inc., 2014, pp. 3221–3229.
[45] J.-B. Hiriart-Urruty, C. Lemaréchal, Fundamentals of Convex Analysis, Springer Science & Business Media, 2004.
[46] X. Dupuis, P.J.C. Tardivel, Proximal operator for the sorted ℓ_1 norm: application to testing procedures based on SLOPE, J. Statist. Plann. Inference 221 (2022) 1–8.
[47] P.J.C. Tardivel, R. Servien, D. Concordet, Simple expressions of the LASSO and SLOPE estimators in low-dimension, Statistics 54 (2) (2020) 340–352.
[48] A. Ben-Israel, T.N.E. Greville, Generalized Inverses, 15 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*, Springer-Verlag, New York, 2nd edition, 2003. Theory and applications.
[49] P. Bühlmann, S. Van De Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer Science & Business Media, 2011.
[50] P. Graczyk, U. Schneider, T. Skalski, P. Tardivel, A Unified Framework for Pattern Recovery in Penalized and Thresholded Estimation and its Geometry, arXiv preprint arXiv:2307.10158. (2023).
[51] P. Tardivel, M. Bogdan, On the sign recovery by least absolute shrinkage and selection operator, thresholded least absolute shrinkage and selection operator, and thresholded basis pursuit denoising, Scand. J. Stat. 49 (2022) 1636–1668.
[52] H.L. Harter, Expected values of normal order statistics, Biometrika 48 (1/2) (1961) 151–165.
[53] R. Riccobello, M. Bogdan, G. Bonaccolto, P.J. Kremer, S. Paterlini, P. Sobczyk, Graphical Modelling via the Sorted L1-Norm, arXiv preprint arXiv:2204.10403. (2022).
[54] R. Riccobello, G. Bonaccolto, P.J. Kremer, P. Sobczyk, M. Bogdan, S. Paterlini, Sparse graphical modelling for global minimum variance portfolio, Comput. Manag. Sci. 22 (2) (2025) 1–32.
[55] S. Højsgaard, S.L. Lauritzen, Graphical gaussian models with edge and vertex symmetries, J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (5) (2008) 1005–1027.
[56] W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, V. Rockova, T. Group, Adaptive bayesian SLOPE: model selection with incomplete data, J. Comput. Graph. Statist. 31(1) (2022) 113–137.
[57] J. Larsson, M. Bogdan, J. Wallin, The strong screening rule for SLOPE, in: Advances in Neural Information Processing Systems, 33, Curran Associates, Inc., 2020, pp. 14592–14603.
[58] J. Larsson, Q. Klopfenstein, M. Massias, J. Wallin, Coordinate descent for slope, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 4802–4821.
[59] M.-H. Ko, D.-H. Ryu, T.-S. Kim, Y.-K. Choi, A central limit theorem for general weighted sums of LNQD random variables and its application, Rocky Mountain J. Math. 37 (1) (2007) 259–268.
[60] T.L. Lai, C.Z. Wei, A law of the iterated logarithm for double arrays of independent random variables with applications to regression and time series models, Ann. Probab. 10 (2) (1982) 320–335.
[61] U. Stadtmüller, A note on the law of iterated logarithm for weighted sums of random variables, Ann. Probab. 12 (1) (1984) 35–44.
[62] J. Cuzick, A strong law for weighted sums of i.i.d. random variables, J. Theoret. Probab. 8 (3) (1995) 625–641.