PROBABILITY
AND
MATHEMATICAL STATISTICS

Published online *.*.2022
doi:10.37190/0208-4147.00066
Online First version

PATTERN RECOVERY AND SIGNAL DENOISING BY SLOPE WHEN THE DESIGN MATRIX IS ORTHOGONAL*

BY

TOMASZ SKALSKI (WROCŁAW AND ANGERS), PIOTR GRACZYK (ANGERS),
BARTOSZ KOŁODZIEJEK (WARSZAWA), AND MACIEJ WILCZYŃSKI (WROCŁAW)

Abstract. Sorted ℓ_1 Penalized Estimator (SLOPE) is a relatively new convex regularization method for fitting high-dimensional regression models. SLOPE allows the reduction of the model dimension by shrinking some estimates of the regression coefficients completely to zero or by equating the absolute values of some nonzero estimates of these coefficients. This allows one to identify situations where some of true regression coefficients are equal. In this article we will introduce the SLOPE pattern, i.e., the set of relations between the true regression coefficients, which can be identified by SLOPE. We will also present new results on the strong consistency of SLOPE estimators and on the strong consistency of pattern recovery by SLOPE when the design matrix is orthogonal and illustrate advantages of the SLOPE clustering in the context of high frequency signal denoising.

2020 Mathematics Subject Classification: Primary 62J05; Secondary 62J07.

Key words and phrases: linear regression, SLOPE, signal denoising.

1. INTRODUCTION

1.1. Introduction and motivations. Linear Multiple Regression concerns the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{Y} \in \mathbb{R}^n$ is an output vector, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a fixed design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of predictors and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a noise vector. The primary goal is to estimate $\boldsymbol{\beta}$. In the low-dimensional setting, i.e., when the number p of predictors is not larger than the numbers n of explanatory variable and $\boldsymbol{X}$ is of full rank, the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ has an exact formula $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. For practical reasons there is an urge to avoid the high-dimensionality curse, therefore we want the estimate to be sparse, i.e., to be describable by a smaller number of parameters. Several solutions were proposed

* The first author was supported by a French Government Scholarship. Research of the first and the second author was supported by Centre Henri Lebesgue, program ANR-11-LABX-0020-0.

to deal with that problem. One of them, the Least Absolute Shrinkage and Selection Operator (LASSO [7, 24]) involves penalizing the residual sum of squares $\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$ with the ℓ_1 norm of $\widehat{\boldsymbol{\beta}}$ multiplied by a tuning parameter λ:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}} := \mathop{\arg\min}_{\mathbf{b}\in\mathbb{R}^p}\left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_1\right].$$

The LASSO estimator is not unbiased, but is a shrinkage estimator which shrinks some $\widehat{\beta}_j^{\mathrm{LASSO}}$ completely to zero, resulting in a sparser estimate. In the case of $\boldsymbol{X}$ being an orthogonal matrix, i.e. $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, the exact formula for $\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ introduced by Tibshirani [24] is based on $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$:

$$\widehat{\beta}_i^{\mathrm{LASSO}} = \mathrm{sign}(\widehat{\beta}_i^{\mathrm{OLS}})\max\{|\widehat{\beta}_i^{\mathrm{OLS}}| - \lambda, 0\}.$$

Another approach to reduce the dimensionality is the Sorted ℓ_1 Penalized Estimator (SLOPE [3, 2, 25]), which not only generalizes the LASSO method, but also allows one to clusterize the similar coefficients of $\boldsymbol{\beta}$. In SLOPE, the ℓ_1 norm is replaced by its sorted version $J_{\boldsymbol{\Lambda}}$, which depends on the tuning vector $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ with $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$:

$$J_{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) := \sum_{i=1}^{p} \lambda_i |\boldsymbol{\beta}|_{(i)},$$

where $\{|\boldsymbol{\beta}|_{(i)}\}_{i=1}^p$ is a decreasing permutation of the absolute values of $\beta_1, \ldots, \beta_p$:

$$\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}} := \mathop{\arg\min}_{\mathbf{b}\in\mathbb{R}^p}\left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b})\right].$$

The case of $\boldsymbol{\Lambda}$ being an arithmetic sequence was studied by Bondell and Reich [5] and called the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR). The special case of SLOPE with $\lambda_1 = \cdots = \lambda_p > 0$ is LASSO. For $\boldsymbol{\Lambda} = (0, \ldots, 0)$ we obtain the OLS estimator.

Clustering the predictors allows for additional dimension reduction by identifying variables with the same absolute values of regression coefficients. One can recently observe the rise of interest in methods which cluster highly correlated predictors [6, 12, 14, 17, 18, 23]. SLOPE is ideal for this task, since it is capable of identifying the low-dimensional structure, which is called the SLOPE pattern, defined by Schneider and Tardivel in terms of the subdifferential of the SLOPE norm $J_{\boldsymbol{\Lambda}}$; see Remark 1.1. By convention, in this article we let $\mathrm{sign}(0) = 0$. We will denote by k the number of clusters of $\boldsymbol{patt}(\boldsymbol{b}) = (m_1, \ldots, m_p)'$, i.e., the number of non-zero components of $|\boldsymbol{b}|$.

DEFINITION 1.1 (SLOPE pattern [4]). The *SLOPE pattern* is a function

$$\boldsymbol{patt} : \mathbb{R}^p \to \mathbb{Z}^p \quad \text{such that} \quad patt(\boldsymbol{b})_i = \mathrm{sign}(b_i)\,rank(|b_i|),$$

where the rank of $|b_i| \neq 0$ is defined to be the number of $|c_j|$'s satisfying $|b_i| \geqslant |c_j|$, where $|c_1|, \ldots, |c_k|$, $k \leqslant p$, are distinct non-zero values among $|b_1|, \ldots, |b_p|$. We adopt the convention that $rank(0) = 0$.

We denote by $\mathcal{M}_p$ the set of all possible SLOPE patterns of $\boldsymbol{b} \in \mathbb{R}^p$.

FACT 1.1 (Basic properties of SLOPE pattern [20]).

(a) *For every* $1 \leqslant l \leqslant \|\boldsymbol{patt}(\boldsymbol{b})\|_\infty$ *there exists* j *such that* $|patt(\boldsymbol{b})_j| = l$,

(b) $\operatorname{sign}(\boldsymbol{patt}(\boldsymbol{b})) = \operatorname{sign}(\boldsymbol{b})$ (*sign preservation*),

(c) $|b_i| = |b_j| \Rightarrow |patt(\boldsymbol{b})_i| = |patt(\boldsymbol{b})_j|$ (*cluster preservation*),

(d) $|b_i| > |b_j| \Rightarrow |patt(\boldsymbol{b})_i| > |patt(\boldsymbol{b})_j|$ (*hierarchy preservation*).

EXAMPLE 1.1. $\boldsymbol{patt}(4, 0, -1.5, 1.5, -4) = (2, 0, -1, 1, -2)$.

REMARK 1.1 (Subdifferential description of the SLOPE pattern [20]). Let $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)$ satisfy $\lambda_1 > \cdots > \lambda_p > 0$. Then

$$\boldsymbol{patt}(\boldsymbol{b}_1) = \boldsymbol{patt}(\boldsymbol{b}_2) \iff \partial_{J_{\boldsymbol{\Lambda}}}(\boldsymbol{b}_1) = \partial_{J_{\boldsymbol{\Lambda}}}(\boldsymbol{b}_2),$$

where $\partial_f(\boldsymbol{b})$ is the subdifferential of the function $f : \mathbb{R}^p \to \mathbb{R}$ at $\boldsymbol{b}$, i.e.

$$\partial_f(\boldsymbol{b}) = \{v \in \mathbb{R}^p : f(\boldsymbol{z}) \geqslant f(\boldsymbol{b}) + v'(\boldsymbol{z} - \boldsymbol{b})\ \forall \boldsymbol{z} \in \mathbb{R}^p\}.$$

The subdifferential approach may be applied to a wider class of penalizers that are polyhedral gauges (see [22]).

DEFINITION 1.2 (Pattern recovery by SLOPE). We say that the SLOPE estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ *recovers the pattern* of $\boldsymbol{\beta}$ when

$$\boldsymbol{patt}(\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}) = \boldsymbol{patt}(\boldsymbol{\beta}).$$

The clustering properties of SLOPE have been studied before in [5, 11], but the researchers considered strongly correlated predictors, used in financial mathematics to group the assets with respect to their partial correlation with the hedge fund return times series [13]. In our article we assume the orthogonal design

$$\boldsymbol{X}'\boldsymbol{X} = n\mathbb{I}_p. \tag{1.1}$$

This is a classical and natural assumption in case of experimential data [24]. Moreover, in the asymptotic case, where $n \to \infty$ and p is fixed, it is usually supposed that $\boldsymbol{X}'\boldsymbol{X}/n \to \boldsymbol{C} > 0$ [26, 27]. In (1.1) the design matrix $\boldsymbol{X}$ is orthogonal. Then the Euclidean norm of each n-dimensional column of $\boldsymbol{X}$ equals n. If it was 1, the terms of $\boldsymbol{X}$ would approach zero for large n, which is not natural. Such matrices are widely used in signal analysis [19, 8]. For general $\boldsymbol{X}$ the problem is considered in the companion article [4].

To study the properties of SLOPE we often use the closed unit ball $C_{\mathbf{\Lambda}}$ in the dual norm of $J_{\mathbf{\Lambda}}$, which was studied e.g. by Zeng and Figueiredo [25]. This dual ball is described explicitly as a signed permutahedron (see e.g. [16, 20])

$$C_{\mathbf{\Lambda}} = \Big\{\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p) \in \mathbb{R}^p : \sum_{j \leqslant i} |\pi|_{(j)} \leqslant \sum_{j \leqslant i} \lambda_j,\ i = 1, \ldots, p\Big\}. \tag{1.2}$$

In this article we prove novel results on the strong consistency of SLOPE both in estimation and in pattern recovery. We also introduce a new method, based on the minimax approach, to find the relationship between $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$.

1.2. Outline of the paper. In Section 2 we derive connections between $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ in the orthogonal design. We use the minimax theorem of Sion [1]. In Section 3 we focus on the properties of $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$. We use the geometric interpretation of SLOPE to explain its ability to identify the SLOPE pattern, and we provide new theoretical results on the support recovery and clustering properties using a representation of SLOPE as a function of the ordinary least squares (OLS) estimator. A similar approach to LASSO was used by Ewald and Schneider [10].

To analyze the asymptotic properties of the SLOPE estimator, e.g. its consistency, we have to assume that the sample size n tends to infinity. Therefore, in Section 4 we define a sequence of linear regression models

$$\boldsymbol{Y}^{(n)} = \boldsymbol{X}^{(n)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n^{(n)}.$$

In this sequence, the response vector $\boldsymbol{Y}^{(n)} \in \mathbb{R}^n$, the design matrix $\boldsymbol{X}^{(n)} \in \mathbb{R}^{n\times p}$ and the error term $\boldsymbol{\varepsilon}^{(n)} = (\varepsilon_1^{(n)}, \ldots, \varepsilon_n^{(n)})' \in \mathbb{R}^n$ vary with n. The error term has a normal $N(0, \sigma^2 \boldsymbol{I}_n)$ distribution. We make no assumptions about relations between $\boldsymbol{\varepsilon}^{(n)}$ and $\boldsymbol{\varepsilon}^{(m)}$ for $n \neq m$. In this paper we consider the specific, but statistically important model in which $n \geqslant p$ and the columns of $\boldsymbol{X}$ are pairwise orthogonal. The orthogonality assumption allows us to derive, by simple techniques, relatively precise results on the SLOPE estimator (e.g. Theorem 3.1), which seem unavailable when the columns of $\boldsymbol{X}$ are not orthogonal.

Substantially more difficult techniques based on subdifferential calculus are developed in [4]. These techniques are used there to establish properties of the SLOPE estimator in the general case, when the columns of $\boldsymbol{X}$ are not orthogonal, the sequence of error terms $\boldsymbol{\varepsilon}^{(n)}$ is incremental, and p may be much larger than n. In the asymptotic theorem proved in [4] under different assumptions on $\boldsymbol{X}_n'\boldsymbol{X}_n$ stronger restrictions on the tuning λ_n are considered. We provide conditions under which the SLOPE estimator is strongly consistent. Additionally, when for each n the design matrix is orthogonal, we provide conditions on the sequence of tuning parameters such that SLOPE is strongly consistent in pattern recovery. In Section 5 we show applications of SLOPE clustering in terms of high frequency signal denoising and illustrate them with simulations. The Appendix contains the proofs of the technical results.

2. APPROACH BY MINIMAX THEOREM

2.1. Technical results. Let r_{SLOPE} denote the minimum value of the SLOPE criterion, attained by $\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}$, i.e.

$$r_{\text{SLOPE}} := \min_{\mathbf{b}\in\mathbb{R}^p} \left[\tfrac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b})\right] = \tfrac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}\|_2^2 + J_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}).$$

Since

$$\|\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}\|_2 \leqslant \sqrt{p}\,\|\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}\|_\infty, \qquad \lambda_1\|\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}\|_\infty \leqslant J_{\boldsymbol{\Lambda}}(\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}) \leqslant r_{\text{SLOPE}},$$

it follows that

$$\lambda_1\|\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}\|_2 \leqslant \sqrt{p}\,r_{\text{SLOPE}} \leqslant \sqrt{p}\left[\tfrac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\mathbf{0}\|_2^2 + J_{\boldsymbol{\Lambda}}(\mathbf{0})\right] = \frac{\sqrt{p}}{2}\|\boldsymbol{Y}\|_2^2.$$

We immediately get the following result.

COROLLARY 2.1. $\|\widehat{\boldsymbol{\beta}}^{\text{SLOPE}}\|_2^2 \leqslant M_0$, *where* $M_0 = \left(\frac{p\,\|\boldsymbol{Y}\|_2^4}{4\lambda_1^2}\right)$.

By this corollary, we can clearly limit our search to vectors $\boldsymbol{\beta}$ from the compact set $\mathcal{M} := \left\{\boldsymbol{b}\in\mathbb{R}^p : \|\boldsymbol{b}\|_2^2 \leqslant M_0\right\}$. Therefore, we can equivalently define a SLOPE solution by

$$\widehat{\boldsymbol{\beta}}^{\text{SLOPE}} = \arg\min_{\mathbf{b}\in\mathcal{M}}\left[\tfrac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b})\right]. \tag{2.1}$$

PROPOSITION 2.1. *Let $C_{\boldsymbol{\Lambda}}$ be the unit ball in the dual SLOPE norm. Then, for each $\boldsymbol{b}\in\mathbb{R}^p$,*

$$J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} \boldsymbol{\pi}'\boldsymbol{b}. \tag{2.2}$$

The proof is a simple application of the definition of the dual norm and the reflexivity of $(\mathbb{R}^p, J_{\boldsymbol{\Lambda}}) = (\mathbb{R}^p, J_{\boldsymbol{\Lambda}}^*)^*$. Thus

$$J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \|\boldsymbol{b}\|_{(\mathbb{R}^p, J_{\boldsymbol{\Lambda}})} = \sup_{\mathbf{x}\colon J_{\boldsymbol{\Lambda}}^*(\mathbf{x})\leqslant 1} \mathbf{x}'\boldsymbol{b}.$$

REMARK 2.1. (a) A different, longer proof is given in [3, Proposition 1.1].

(b) Formula (2.2) holds in much greater generality for Lovász extensions in place of the $J_{\boldsymbol{\Lambda}}$ norm [15]. We thank the anonymous referee for pointing this out.

2.2. Saddle point. Let $r : \mathcal{M} \times C_{\mathbf{\Lambda}} \to \mathbb{R}$ be defined by

$$r(\boldsymbol{b}, \boldsymbol{\pi}) := \tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \boldsymbol{\pi}'\boldsymbol{b}.$$

As an immediate consequence of (2.1) and Proposition 2.1 we obtain

$$\begin{aligned} r_{\mathrm{SLOPE}} &= \min_{\mathbf{b}\in\mathbb{R}^p} \left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\mathbf{\Lambda}}(\boldsymbol{b})\right] = \min_{\mathbf{b}\in\mathcal{M}} \left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\mathbf{\Lambda}}(\boldsymbol{b})\right] \\ &= \min_{\mathbf{b}\in\mathcal{M}} \max_{\boldsymbol{\pi}\in C_{\mathbf{\Lambda}}} \left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \boldsymbol{\pi}'\boldsymbol{b}\right] = \min_{\mathbf{b}\in\mathcal{M}} \max_{\boldsymbol{\pi}\in C_{\mathbf{\Lambda}}} r(\boldsymbol{b}, \boldsymbol{\pi}). \end{aligned}$$

It turns out that maximization over $\boldsymbol{\pi} \in C_{\mathbf{\Lambda}}$ and minimization over $\boldsymbol{b} \in \mathcal{M}$ can be interchanged without affecting the result. To see this, note that both $C_{\mathbf{\Lambda}}$ and $\mathcal{M}$ are convex and compact. Moreover, for each fixed $\boldsymbol{\pi} \in C_{\mathbf{\Lambda}}$, $r(\boldsymbol{b}, \boldsymbol{\pi})$ is a convex continuous function of $\boldsymbol{b} \in \mathcal{M}$, and for each fixed $\boldsymbol{b} \in \mathcal{M}$, $r(\boldsymbol{b}, \boldsymbol{\pi})$ is concave with respect to $\boldsymbol{\pi} \in C_{\mathbf{\Lambda}}$ (in fact, linear). Therefore, all assumptions of Sion's minimax theorem are fulfilled (see [1, p. 218]) and thus there exists a saddle point $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*) \in \mathcal{M} \times C_{\mathbf{\Lambda}}$ such that

$$\begin{aligned} \max_{\boldsymbol{\pi}\in C_{\mathbf{\Lambda}}} \min_{\mathbf{b}\in\mathcal{M}} r(\boldsymbol{b}, \boldsymbol{\pi}) &= \min_{\mathbf{b}\in\mathcal{M}} r(\boldsymbol{b}, \boldsymbol{\pi}^*) = r(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*) \\ &= \max_{\boldsymbol{\pi}\in C_{\mathbf{\Lambda}}} r(\boldsymbol{\beta}^*, \boldsymbol{\pi}) = \min_{\mathbf{b}\in\mathcal{M}} \max_{\boldsymbol{\pi}\in C_{\mathbf{\Lambda}}} r(\boldsymbol{b}, \boldsymbol{\pi}) = r_{\mathrm{SLOPE}}. \end{aligned}$$

In the next section we shall see that the first coordinate of any saddle point $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)$ is SLOPE estimator.

2.3. SLOPE solution when $\boldsymbol{X}$ has full column rank. Since for each fixed $\boldsymbol{\pi} \in C_{\mathbf{\Lambda}}$, the function $r(\boldsymbol{b}, \boldsymbol{\pi})$ is convex with respect to $\boldsymbol{b} \in \mathcal{M}$, any point $\boldsymbol{b}_{\boldsymbol{\pi}} \in \mathcal{M}$ at which the gradient $\frac{\partial r(\boldsymbol{b},\boldsymbol{\pi})}{\partial \boldsymbol{b}}$ is zero, is a global minimum point. If we rewrite $r(\boldsymbol{b}, \boldsymbol{\pi})$ as

$$r(\boldsymbol{b}, \boldsymbol{\pi}) = \tfrac{1}{2}\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{b} + \tfrac{1}{2}\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{\pi}'\boldsymbol{b}$$

and differentiate with respect to $\boldsymbol{b}$, we obtain

$$\frac{\partial r(\boldsymbol{b}, \boldsymbol{\pi})}{\partial \boldsymbol{b}} = -\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) + \boldsymbol{\pi}.$$

Equating this gradient to $\mathbf{0}$ gives the following equation for the optimum point $\boldsymbol{b}_{\boldsymbol{\pi}}$:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi}. \tag{2.3}$$

Substituting this into the equation for $r(\boldsymbol{b}_{\boldsymbol{\pi}}, \boldsymbol{\pi})$, we find that

$$\begin{aligned} r(\boldsymbol{b}_{\boldsymbol{\pi}}, \boldsymbol{\pi}) &= \tfrac{1}{2}\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{Y} + \tfrac{1}{2}\boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} + \boldsymbol{\pi}'\boldsymbol{b}_{\boldsymbol{\pi}} \\ &= \tfrac{1}{2}\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} + \boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{\pi} - \tfrac{1}{2}\boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} \\ &= \tfrac{1}{2}\boldsymbol{Y}'\boldsymbol{Y} - \tfrac{1}{2}\boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} = \tfrac{1}{2}\boldsymbol{Y}'\boldsymbol{Y} - \tfrac{1}{2}\boldsymbol{b}_{\boldsymbol{\pi}}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} \\ &= \tfrac{1}{2}\boldsymbol{Y}'\boldsymbol{Y} - \tfrac{1}{2}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi})'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi}). \end{aligned}$$

Let $p_j = |\{i : |m_i| = k + 1 - j\}|$ be the number of elements of the jth cluster of $\boldsymbol{\beta}$, $P_j = \sum_{i\leqslant j} p_i$ and $P_{k+1} = p$.

LEMMA 2.1. *Assume that $\boldsymbol{X}$ has full column rank, i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. Let $\boldsymbol{\pi}^* = (\pi_1^*, \dots, \pi_p^*)' \in C_{\boldsymbol{\Lambda}}$ be any solution of*

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}}{\arg\min}[(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi})'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi})\,]$$

and let $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)'$ be the corresponding point from $\mathcal{M}$ given by

$$\boldsymbol{\beta}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi}^*).$$

Then $(\boldsymbol{\pi} - \boldsymbol{\pi}^*)'\boldsymbol{\beta}^* \leqslant 0$ for all $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$ and hence

(a) $\mathrm{sign}(\beta_i^*) \cdot \mathrm{sign}(\pi_i^*) \geqslant 0$, *$i = 1, \dots, p$,*

(b) $(|\pi_1^*|, \dots, |\pi_p^*|)$ *and* $(|\beta_1^*|, \dots, |\beta_p^*|)$ *are similarly sorted, i.e. if* $|(patt(\boldsymbol{\beta}))_i| = k + 1 - j$, *then* $|\boldsymbol{\pi}^*|_i \in \{|\boldsymbol{\pi}^*|_{(P_{j-1}+1)}, \dots, |\boldsymbol{\pi}^*|_{(P_j)}\}$,

(c) *for any permutation τ satisfying $|\beta_{\tau(1)}^*| \geqslant \cdots \geqslant |\beta_{\tau(p)}^*|$, if there is a $k \in \{2, \dots, p\}$ such that $\sum_{i=1}^{k-1} |\pi_{\tau(i)}^*| < \sum_{i=1}^{k-1} \lambda_i$ and $|\pi_{\tau(k)}^*| > 0$, then $|\beta_{\tau(k-1)}^*| = |\beta_{\tau(k)}^*|$.*

The proof is given in the Appendix. An immediate consequence of the lemma is the following result.

LEMMA 2.2. *Assume that $\boldsymbol{X}$ has full column rank, i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. The point $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)$ defined as in Lemma* 2.1 *is a saddle point of the function $r(\boldsymbol{b}, \boldsymbol{\pi})$.*

The proof is given in the Appendix. We use the last lemma to prove the main result of this section.

THEOREM 2.1. *Assume that $\boldsymbol{X}$ has full column rank, i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. Let the point $\boldsymbol{\beta}^*$ be defined as in Lemma* 2.1. *Then $\boldsymbol{\beta}^*$ is the SLOPE estimator of $\boldsymbol{\beta}$.*

Proof. Using the fact that

$$\max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta}^*, \boldsymbol{\pi}) = \min_{\mathbf{b}\in\mathcal{M}} \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{b}, \boldsymbol{\pi})$$

(see the previous lemma) we have

$$\begin{aligned}\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{\beta}^*) &= \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} \left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + \boldsymbol{\pi}'\boldsymbol{\beta}^*\right] \\ &= \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta}^*, \boldsymbol{\pi}) = \min_{\mathbf{b}\in\mathcal{M}} \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{b}, \boldsymbol{\pi}) = \min_{\mathbf{b}\in\mathbb{R}^p} \left[\tfrac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b})\right]. \quad \blacksquare\end{aligned}$$

COROLLARY 2.2. *In the linear model satisfying $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ we have*

$$\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}} = \frac{1}{n}\boldsymbol{\pi}^* = \frac{1}{n}\underset{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}}{\arg\min}\,\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\pi}\|_2^2 = \underset{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}/n}}{\arg\min}\,\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\pi}\|_2^2,$$

is the proximal projection of $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ onto $C_{\boldsymbol{\Lambda}/n}$.

Projections onto $C_{\boldsymbol{\Lambda}}$ are widely used in [15] in the study of the notion of degrees of freedom. However, Corollary 2.2 is not stated there explicitly.

REMARK 2.2. Assume that $\boldsymbol{X}$ has full column rank, i.e. $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. For each $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$, the point $\boldsymbol{b}_{\boldsymbol{\pi}}$ defined in (2.3) is in $\{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}\|_2^2 \leqslant M\}$ for $M > \max\{M_0, M_1\}$ with

$$M_1 := \max_{\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}} \|(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi})\|_2^2.$$

3. PROPERTIES OF SLOPE IN THE ORTHOGONAL DESIGN

3.1. SLOPE vs. OLS. By Theorem 2.1 and Corollary 2.2, when $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, the orthogonal projection of the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{Y}$ onto the unit ball $C_{\boldsymbol{\Lambda}/n}$ is equal to $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$. For $\boldsymbol{\Lambda} = (200, 100)'$ and $n = 50$ this property is illustrated in Figure 1. The figure represents $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ (black arrows) depending on the localization of $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ in the orthogonal design. For $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ being the blue point (for colors, see the pdf file) located in the area labelled by $(1, 0)$ the first component of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ is positive and the second is null. For $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ being the yellow point in the area labelled by $(-1, 1)$ both components of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ have equal absolute value (clusterization), but their signs are opposite. For $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ being the red point in the area labelled by $(1, 2)$ both components of $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ are positive and the first component is smaller than the second one. The blue polytope is the

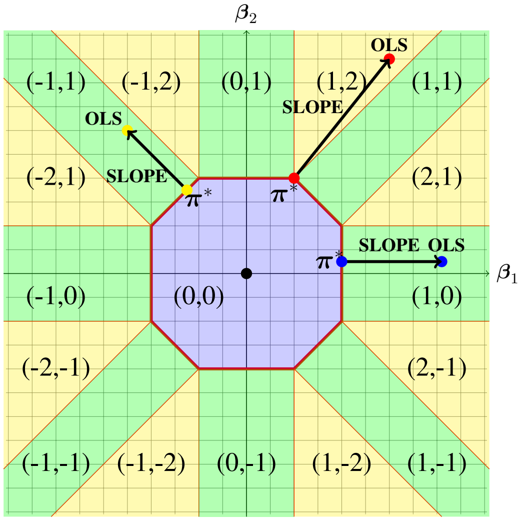


Figure 1. The dual unit ball $C_{\boldsymbol{\Lambda}/n}$ for $\boldsymbol{\Lambda} = (200, 100)'$ and examples of $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ in the orthogonal design for $n = 50$ and $p = 2$. The labels of each colored set refer to the pattern of $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ for $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ lying in this set. The arrows point from $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ to $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$.

dual SLOPE unit ball $C_{\mathbf{\Lambda}}$, and the labels

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

associated to the areas of this figure correspond to all SLOPE patterns for $n = 50$ and $p = 2$. In the orthogonal design, one may also explicitly compute the SLOPE estimator. Indeed, by Corollary 2.2, $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ is the image of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ by the proximal operator of the SLOPE norm. Therefore, this operator has a closed form formula [2, 21, 9]. This explicit expression gives an analytical way to learn that SLOPE solution is sparse and built of clusters.

LEMMA 3.1. *In the linear model satisfying $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ we have*

$$\underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\mathbf{\Lambda}}(\boldsymbol{b})\right] = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2}\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{b}\|_2^2 + J_{\mathbf{\Lambda}}(\boldsymbol{b})\right]. \tag{3.1}$$

As (3.1) is not proven in [3, (1.14)], we give the proof in the Appendix.

The next theorem gives a sufficient condition for the clustering effect of the SLOPE estimator in the orthogonal design.

THEOREM 3.1. *Consider a linear model with orthogonal design $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$. Let π be a permutation of $(1, \ldots, p)$ such that*

$$|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(1)}| \geqslant \cdots \geqslant |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(p)}|.$$

For $i \in \{1, \ldots, p-1\}$,

$$\textit{if } |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i)}| - |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i+1)}| \leqslant \frac{\lambda_i - \lambda_{i+1}}{n}, \textit{ then } |\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}_{\pi(i)}| = |\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}_{\pi(i+1)}|.$$

Proof. By Lemma 3.1, in the orthogonal design, $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ is the proximal map of $J_{\mathbf{\Lambda}/n}(\cdot)$ at $\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. The result may be inferred from [2, Lemma 2.3]. ■

We now derive necessary and sufficient conditions under which SLOPE in the orthogonal design recovers the support of the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, i.e. $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}_i = 0 \Leftrightarrow \boldsymbol{\beta}_i = 0$.

THEOREM 3.2. *Under the orthogonal design $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, let π be a permutation of $(1, \ldots, p)$ satisfying $|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(1)}| \geqslant \cdots \geqslant |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(p)}|$. Without loss of generality suppose that* $\mathrm{supp}(\boldsymbol{\beta}) = \{1, \ldots, p_0\}$ *with $p_0 < p$. A necessary and sufficient condition for SLOPE to identify the set of relevant covariables is:*

(a) $\displaystyle\min_{1\leqslant i\leqslant p_0} |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_i| > \max_{p_0+1\leqslant i\leqslant p} |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_i|,$

(b) $\displaystyle\sum_{i=k}^{p_0} |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i)}| > \frac{1}{n}\sum_{i=k}^{p_0} \lambda_i$ *for $k = 1, \ldots, p_0$,*

(c) $\displaystyle\sum_{i=p_0+1}^{k} |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i)}| \leqslant \frac{1}{n}\sum_{i=p_0+1}^{k} \lambda_i$ *for $k = p_0 + 1, \ldots, p$.*

Proof. The result may be inferred from the properties of the proximal SLOPE [3, Lemmas 2.3 and 2.4] and from Lemma 3.1. ■

4. ASYMPTOTIC PROPERTIES OF SLOPE

In this section we discuss several asymptotic properties of SLOPE estimators in the low-dimensional regression model in which p is fixed and the sample size n tends to infinity. For each $n \geqslant 1$ we consider a linear model

$$\boldsymbol{Y}^{(n)} = \boldsymbol{X}^{(n)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{(n)}, \tag{4.1}$$

where $\boldsymbol{Y}^{(n)} = (y_1^{(n)}, \ldots, y_n^{(n)})' \in \mathbb{R}^n$ is a vector of observations, $\boldsymbol{X}^{(n)} \in \mathbb{R}^{n\times p}$ is a deterministic design matrix with rank$(\boldsymbol{X}^{(n)}) = p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is a vector of unknown regression coefficients and $\boldsymbol{\varepsilon}^{(n)} = (\varepsilon_1^{(n)}, \ldots, \varepsilon_n^{(n)})' \in \mathbb{R}^n$ is a noise term, which has a normal $N(0, \sigma^2\boldsymbol{I}_n)$ distribution. We make no assumptions about the dependence between $\boldsymbol{\varepsilon}^{(n)}$ and $\boldsymbol{\varepsilon}^{(m)}$ for $n \neq m$. In particular, $\boldsymbol{\varepsilon}^{(n)}$ need not be a subsequence of $\boldsymbol{\varepsilon}^{(m)}$.

When defining the sequence $(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})$ of SLOPE estimators, we assume that the tuning vector varies with n. More precisely, for each $n \geqslant 1$ its coefficients $\lambda_1^{(n)} \geqslant \cdots \geqslant \lambda_p^{(n)} \geqslant 0$ are fixed and $\lambda_1^{(n)} > 0$. We denote by $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ the SLOPE estimator corresponding to the tuning vector $\boldsymbol{\Lambda}^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})'$:

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min}\big[\tfrac{1}{2}\|\boldsymbol{Y}^{(n)} - \boldsymbol{X}^{(n)}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}^{(n)}}(\boldsymbol{b})\big]. \tag{4.2}$$

4.1. Strong consistency of the SLOPE estimator. Let us recall the definition of a strongly consistent estimator $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ of $\boldsymbol{\beta}$: for all $\boldsymbol{\beta} \in \mathbb{R}^p$ we have $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \to \boldsymbol{\beta}$ almost surely.

Below we discuss consistency of the sequence $(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})$ of SLOPE estimators, defined by (4.2).

THEOREM 4.1. *Consider the linear regression model* (4.1) *and assume that*

$$\lim_n n^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)} = \boldsymbol{C},$$

where $\boldsymbol{C}$ is a positive definite matrix. Let $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$, $n \geqslant 1$, be the SLOPE estimator corresponding to the tuning vector $\boldsymbol{\Lambda}^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})'$.

(a) *If $\lim_{n\to\infty} \lambda_1^{(n)}/n = 0$, then $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}$.*

(b) *If the true parameter $\boldsymbol{\beta}$ satisfies $\lambda_0\|\boldsymbol{\beta}\|_\infty > \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta}/2$ and $\lambda_1^{(n)}/n \to 0$, then $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ does not converge to $\boldsymbol{\beta}$. Hence, $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ is not strongly consistent for $\boldsymbol{\beta}$.*

Before proving the above theorems we state a simple technical lemma. It follows quickly from the Borel–Cantelli Lemma and the tail inequality: If $Z \sim N(0,1)$, then $\mathbb{P}(Z > t) \leqslant t^{-1}e^{-t^2/2}/\sqrt{2\pi}$, $t > 0$.

LEMMA 4.1. *Assume that* $(Q_n)_{n\in\mathbb{N}}$ *is a sequence of Gaussian random variables, defined on the same probability space, which converges in distribution to* $\mathrm{N}(0,\sigma^2)$ *for some* $\sigma \in (0,\infty)$. *Then, for any* $\delta > 0$,

$$\lim_{n\to\infty} \frac{Q_n}{(\log n)^{1/2+\delta}} = 0 \quad a.s.$$

Our proof of the strong consistency of SLOPE is based on the strong consistency of the OLS estimator. The latter result is folklore and we prove it in our setting.

PROPOSITION 4.1. *Consider the linear regression model* (4.1). *If we have* $\lim_n n^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)} = \boldsymbol{C}$, *where* $\boldsymbol{C}$ *is positive definite, then* $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \xrightarrow{a.s.} \boldsymbol{\beta}$.

Proof. We have

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} - \boldsymbol{\beta} &= ((\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)} - \boldsymbol{\beta} \\ &= ((\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{\varepsilon}^{(n)}.\end{aligned}$$

Then $\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} - \boldsymbol{\beta})$ has a normal $N(0, \sigma^2(n^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1})$ distribution and its components satisfy the assumptions of Lemma 4.1. Since $(\log n)^{1/2+\delta} = o(\sqrt{n})$, we get the assertion by Lemma 4.1. ■

Proof of Theorem 4.1. (a) It follows from Theorem 2.1 that there exists a vector $\boldsymbol{\pi}_n^* \in C_{(\boldsymbol{\Lambda}^{(n)})}$ such that

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = ((\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1}((\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)} - \boldsymbol{\pi}_n^*).$$

Since $\boldsymbol{\pi}_n^*$ takes values in $\boldsymbol{C}_{\boldsymbol{\Lambda}^{(n)}}$, it follows that $\|\boldsymbol{\pi}_n^*\|_\infty \leqslant \lambda_1^{(n)}$. Hence,

$$\boldsymbol{\pi}_n^*/n \xrightarrow{\mathrm{a.s.}} \boldsymbol{0}, \tag{4.3}$$

because $\|\boldsymbol{\pi}_n^*/n\|_\infty \leqslant \lambda_1^{(n)}/n \to 0$. The assumption that $\mathrm{rank}(\boldsymbol{X}^{(n)}) = p$ implies that the matrix $(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}$ is invertible and hence the least squares estimator of $\boldsymbol{\beta}$ is unique and has the form $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} = ((\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)}$. Combining (4.3) with the fact that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \xrightarrow{\mathrm{a.s.}} \boldsymbol{\beta}$, we conclude that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} &= ((\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1}((\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)} - \boldsymbol{\pi}_n^*) \\ &= \widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} - ((\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1}\boldsymbol{\pi}_n^* \\ &= \widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} - \left(\frac{(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}}{n}\right)^{-1}\frac{\boldsymbol{\pi}_n^*}{n} \xrightarrow{\mathrm{a.s.}} \boldsymbol{\beta} - \boldsymbol{C}^{-1}\boldsymbol{0} = \boldsymbol{\beta}.\end{aligned}$$

(b) Since $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ minimizes over $\boldsymbol{b} \in \mathbb{R}^p$ the function

$$l(\boldsymbol{b}) := \tfrac{1}{2}\|\boldsymbol{Y}^{(n)} - \boldsymbol{X}^{(n)}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}^{(n)}}(\boldsymbol{b})$$

and since $\lambda_1^{(n)}\|\boldsymbol{b}\|_\infty \leqslant J_{\boldsymbol{\Lambda}^{(n)}}(\boldsymbol{b})$, it follows that

$$\begin{aligned}
0 &\leqslant l(0) - l(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) \\
&= (\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)} - \tfrac{1}{2}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} - J_{\boldsymbol{\Lambda}^{(n)}}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) \\
&\leqslant (\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)} - \tfrac{1}{2}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} - \lambda_1^{(n)}\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}\|_\infty \\
&= (\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \\
&\quad - \tfrac{1}{2}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} - \lambda_1^{(n)}\|\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}\|_\infty.
\end{aligned}$$

The last equality follows from the fact that $(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})'(\boldsymbol{X}^{(n)})'(\boldsymbol{Y}^{(n)} - \boldsymbol{X}^{(n)})\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} = 0$. Suppose to the contrary that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}$. Then, using the facts that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}$ and $\lim_n n^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)} = \boldsymbol{C}$, we have

$$0 \leqslant \frac{l(0) - l(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})}{n} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta} - \tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta} - \lambda_0\|\boldsymbol{\beta}\|_\infty = \tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta} - \lambda_0\|\boldsymbol{\beta}\|_\infty.$$

For $\lambda_0 > 0$ this provides a contradiction since the inequality $\lambda_0\|\boldsymbol{\beta}\|_\infty \leqslant \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta}$ does not hold when the value of $\boldsymbol{\beta}$ is sufficiently close to 0. ∎

REMARK 4.1. The proof of Theorem 4.1(b) does not exclude that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \to \boldsymbol{\beta}$ for $\|\boldsymbol{\beta}\|$ large enough.

However, the definition of strong consistency requires the convergence for any $\boldsymbol{\beta}$. We have proved that if the true parameter $\boldsymbol{\beta}$ satisfies $\lambda_0\|\boldsymbol{\beta}\|_\infty > \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta}/2$ and $\lim_n \lambda_1/n = \lambda_0 > 0$, then $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ does not converge to $\boldsymbol{\beta}$.

4.2. Asymptotic pattern recovery in the orthogonal design. We again consider a sequence of linear models (4.1) but this time we assume that for each n the deterministic design matrix $\boldsymbol{X}^{(n)}$ of size $n \times p$ satisfies

$$(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)} = n\boldsymbol{I}_p. \tag{4.4}$$

As usual, we assume Gaussian errors, $\boldsymbol{\varepsilon}^{(n)} \sim N(0, \sigma^2\boldsymbol{I}_n)$.

Let $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{SLOPE}}(n), \ldots, \widehat{\boldsymbol{\beta}}_p^{\mathrm{SLOPE}}(n))'$ be the SLOPE estimator defined by (4.2). With the above notation we present the main result of this section.

THEOREM 4.2. *Assume that*

$$\lim_{n\to\infty} \frac{\lambda_1^{(n)}}{n} = 0$$

and there exists $\delta > 0$ *such that*

$$\liminf_{n\to\infty} \frac{\lambda_i^{(n)} - \lambda_{i+1}^{(n)}}{\sqrt{n}\,(\log n)^{1/2+\delta}} = m > 0 \quad \textit{for } i = 1, \ldots, p-1. \tag{4.5}$$

Then

$$\boldsymbol{patt}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) \overset{a.s.}{\to} \boldsymbol{patt}(\boldsymbol{\beta}).$$

Note that the above conditions are satisfied e.g. by $\lambda_i^{(n)} = c(p+1-i)n^{2/3}$ for any constant $c > 0$.

Proof of Theorem 4.2. Without loss of generality we may assume that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\beta_1 \geqslant \cdots \geqslant \beta_p \geqslant 0$. Indeed, we can always achieve this by permuting the columns of $\boldsymbol{X}^{(n)}$ and changing their signs.

Since the space of patterns is discrete, we have to show that for large n,

$$\boldsymbol{patt}(\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) = \boldsymbol{patt}(\boldsymbol{\beta}) \quad \text{a.s.}$$

We divide the proof into the following four parts:

$$\begin{aligned}
&\text{(i)} && \beta_i = \beta_j > 0 \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) = \widehat{\beta}_j^{\mathrm{SLOPE}}(n) \text{ a.s. for large } n,\\
&\text{(ii)} && \beta_i > \beta_{i+1} \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) > \widehat{\beta}_{i+1}^{\mathrm{SLOPE}}(n) \text{ a.s. for large } n,\\
&\text{(iii)} && \beta_i = 0 \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) = 0 \text{ a.s. for large } n,\\
&\text{(iv)} && \beta_i > 0 \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) > 0 \text{ a.s. for large } n.
\end{aligned}$$

Points (ii) and (iv) follow quickly by the strong consistency of $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$. To prove (i) and (iii) we observe that for each n we are in the orthogonal design case.

Let π_n be a permutation of $(1, \ldots, p)$ satisfying

$$|\widehat{\beta}_{\pi_n(1)}^{\mathrm{OLS}}(n)| \geqslant \ldots \geqslant |\widehat{\beta}_{\pi_n(p)}^{\mathrm{OLS}}(n)|.$$

By the strong consistency of the OLS estimator, taking n sufficiently large, we may ensure that the clusters of $\boldsymbol{\beta}$ do not interlace in $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}$ in the sense that if $\beta_i > \beta_j$, then $\widehat{\beta}_i^{\mathrm{OLS}}(n) > \widehat{\beta}_j^{\mathrm{OLS}}(n)$ a.s. for n sufficiently large.

Let us now consider point (i). Let S_i denote the cluster containing $\beta_i > 0$, that is, the set $S_i = \{j \in \{1, \ldots, p\} \colon \beta_j = \beta_i\}$. In view of the ordering of $\boldsymbol{\beta}$, there exists $k_i \in \{1, \ldots, p\}$ such that

$$S_i = \big\{\pi_n(j) \colon j \in \{k_i, k_i+1, \ldots, k_i + \#S_i - 1\}\big\}.$$

We will show that if $\pi_n(k), \pi_n(k+1) \in S_i$, then for large n,

$$\widehat{\beta}_{\pi_n(k)}^{\mathrm{SLOPE}}(n) = \widehat{\beta}_{\pi_n(k+1)}^{\mathrm{SLOPE}}(n) \quad \text{a.s.,} \tag{4.6}$$

thus $\widehat{\beta}_j^{\mathrm{SLOPE}}(n) = \widehat{\beta}_k^{\mathrm{SLOPE}}(n)$ for $j, k \in S_i$, which finishes the proof of (i).

Now assume that $\pi_n(k), \pi_n(k+1) \in S_i$. Then, by Theorem 3.1, condition (4.6) is satisfied if

$$|\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(k)}(n)| - |\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(k+1)}(n)| \leqslant \frac{1}{n}(\lambda_k^{(n)} - \lambda_{k+1}^{(n)}) \tag{4.7}$$

for large n and both $\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(k)}(n)$ and $\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(k)}(n)$ have the same sign. The latter is ensured by the strong consistency of the OLS estimator and the fact that $\beta_i > 0$.

If $\pi_n(k), \pi_n(k+1) \in S_i$, then we have the bound

$$|\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(k)}(n) - \widehat{\beta}^{\mathrm{OLS}}_{\pi_n(k+1)}(n)| \leqslant \sum_{j \in S_i} |\widehat{\beta}^{\mathrm{OLS}}_j(n) - \widehat{\beta}^{\mathrm{OLS}}_i(n)|. \tag{4.8}$$

Take any $j \in S_i$. Since both $\widehat{\beta}^{\mathrm{OLS}}_j(n)$ and $\widehat{\beta}^{\mathrm{OLS}}_i(n)$ have a normal distribution with the same mean, by Lemma 4.1 we have

$$\lim_{n\to\infty} \frac{\sqrt{n}\,(\widehat{\beta}^{\mathrm{OLS}}_j(n) - \widehat{\beta}^{\mathrm{OLS}}_i(n))}{(\log n)^{1/2+\delta}} = 0 \quad \text{a.s.}$$

In view of (4.8) and (4.5), this implies that (4.7) holds true for large n. Hence, (a) follows.

It remains to establish (iii). Assume that $\beta_{p_0} > 0 = \beta_{p_0+1} = \cdots = \beta_p$. Clearly, condition (a) from Theorem 3.2 is satisfied thanks to the strong consistency of the OLS estimator. For (b), we have for $k = 1, \ldots, p_0$,

$$\frac{1}{n}\sum_{i=k}^{p_0} \lambda_i^{(n)} \leqslant p_0 \frac{\lambda_1^{(n)}}{n},$$

which converges to 0. On the other hand, the left-hand side of (b) converges a.s. to $\sum_{i=k}^{p_0} \beta_i$, which is positive. Thus, condition (b) from Theorem 3.2 holds for large n. Condition (c) there follows from Lemma 4.1. Indeed, for $\delta > 0$ and $k = p_0 + 1, \ldots, p$ we have

$$\lim_{n\to\infty} \frac{\sqrt{n}}{(\log n)^{1/2+\delta}} \sum_{i=p_0+1}^{k} |\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(i)}(n)| = \sum_{i=p_0+1}^{k} \lim_{n\to\infty} \frac{|\sqrt{n}\,\widehat{\beta}^{\mathrm{OLS}}_{\pi_n(i)}(n)|}{(\log n)^{1/2+\delta}} = 0 \text{ a.s.},$$

while

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}\,(\log n)^{1/2+\delta}} \sum_{i=p_0+1}^{k} \lambda_i^{(n)} \geqslant \sum_{i=p_0+1}^{k} \lim_{n\to\infty} \frac{\lambda_i^{(n)} - \lambda_{i+1}^{(n)}}{\sqrt{n}\,(\log n)^{1/2+\delta}} = m > 0.$$

Thus, all assumptions of Theorem 3.2 are satisfied and the proof is complete. ■

5. NUMERICAL EXPERIMENT

Below we present an application of SLOPE in signal denoising. In our example $\boldsymbol{X} \in \mathbb{R}^{300\times 100}$ is an orthogonal system of trigonometric functions, i.e.

$$X_{i,(2*j-1)} = \sin(2\pi ij/n) \quad \text{and} \quad X_{i,(2*j)} = \cos(2\pi ij/n)$$

for $i = 1,\dots,100$ and $j = 1,\dots,150$. Here $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector consisting of two clusters: 20 coordinates with absolute value 100 and 20 coordinates with absolute value 80. The absolute values of the coordinates of $\boldsymbol{\beta}$ are sorted in a decreasing way. The signs of the non-zero coordinates are chosen independently with random uniform distribution. To avoid large bias caused by the shrinkage nature of LASSO and SLOPE, we debias them by combining with the OLS method. For that reason we use the following definition of the pattern matrix $\boldsymbol{U}_M$ and the clustered design matrix $\widetilde{\boldsymbol{X}}_M$, which is based on the SLOPE pattern:

DEFINITION 5.1. Let $M \neq 0$ be a pattern in $\mathcal{M}_p$ with $k = \|M\|_\infty$ non-zero clusters. The *pattern matrix* $\boldsymbol{U}_M \in \mathbb{R}^{p\times k}$ is defined as follows:

$$(\boldsymbol{U}_M)_{ij} = \operatorname{sign}(m_i)\mathbf{1}_{(|m_i|=k+1-j)}, \qquad i \in \{1,\dots,p\},\, j \in \{1,\dots,k\}.$$

DEFINITION 5.2. Let $M \neq 0$ be a pattern in $\mathbb{R}^p$ and $k = \max\{\|M\|_\infty, 1\}$. For $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ we define the *clustered design matrix* by $\widetilde{\boldsymbol{X}}_M = \boldsymbol{X}\boldsymbol{U}_M \in \mathbb{R}^{n\times k}$.

To perform the debiased SLOPE, we begin by recovering the support and clusters of a true vector $\boldsymbol{\beta}$ with SLOPE. Then, using the SLOPE pattern M obtained, we replace the design matrix with its clustered version $\widetilde{\boldsymbol{X}}_M = \boldsymbol{X}\boldsymbol{U}_M$. Then we perform the Ordinary Least Squares regression for the model $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}_M\mathbf{b} + \boldsymbol{\varepsilon}$, where $\mathbf{b}$ consists only of distinct absolute values of $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$.

Analogously we proceed with the debiased LASSO. However, in this method we use the LASSO pattern matrix defined in the following way:

For LASSO we have the LASSO pattern that is a vector of signs [22]. For $\boldsymbol{S} \in \{-1,0,1\}^p$, $\|\boldsymbol{S}\|_1$ denotes the number of non-zero coordinates. If $\|\boldsymbol{S}\|_1 = k \geqslant 1$, then we define the corresponding pattern matrix $\boldsymbol{U}_S \in \mathbb{R}^{p\times k}$ by

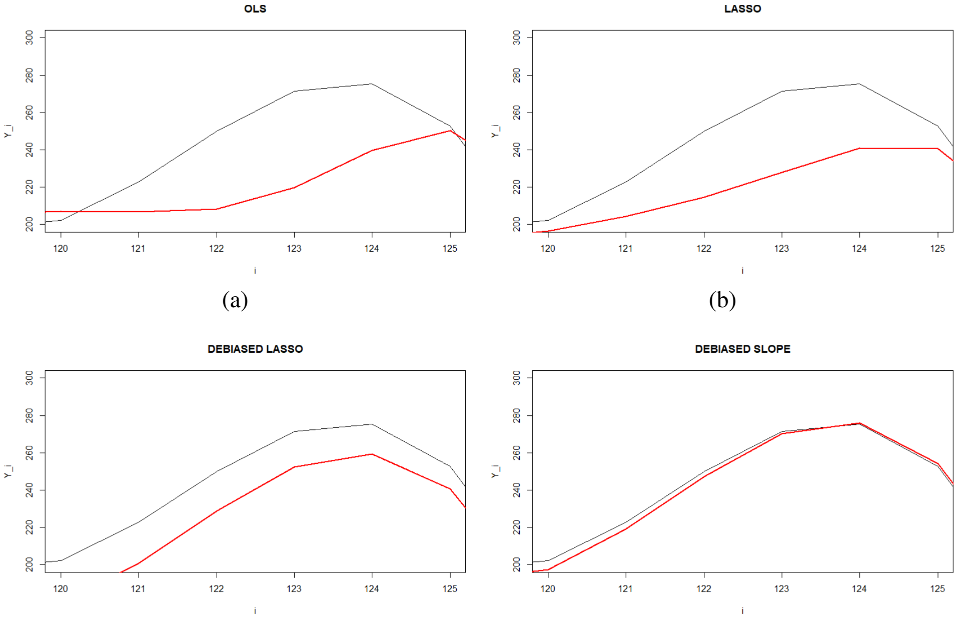
$$\boldsymbol{U}_S = \operatorname{diag}(\boldsymbol{S})_{\operatorname{supp}(\boldsymbol{S})},$$

i.e. the submatrix of $\operatorname{diag}(\boldsymbol{S})$ obtained by keeping the columns corresponding to the indices in $\operatorname{supp}(\boldsymbol{S})$. Then we define the reduced matrix $\tilde{\boldsymbol{X}}_S$ by

$$\tilde{\boldsymbol{X}}_S = \boldsymbol{X}\boldsymbol{U}_S.$$

Equivalently, we have $\tilde{\boldsymbol{X}}_S = (S_iX_i)_{i\in\operatorname{supp}(S)}$. The notion of pattern matrix also appears in [4]. In our example $\boldsymbol{\varepsilon} \in N(0,\sigma^2\boldsymbol{I}_n)$ and $\sigma = 30$.

We compare the Mean Square Error and the signal denoising of the classical OLS estimation, LASSO with the tuning parameter λ_{cv} minimizing the cross-validated error, a debiased version of LASSO with $\lambda = 5\lambda_{cv}$ and a debiased version of SLOPE with the tuning vector $\mathbf{\Lambda}$ chosen with respect to the sequence proposed below Theorem 4.2 ($\lambda_i = 0.1(p + 1 - i)n^{2/3}$).



(a) (b)

(c) (d)

Figure 2. Comparison of signal denoising by OLS (a), LASSO (b), debiased LASSO (c) and debiased SLOPE (d) on the coordinates $[120, 125]$ of the regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The black lines correspond to the true values of $\boldsymbol{X}\boldsymbol{\beta}$. The red lines correspond to the estimators $\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

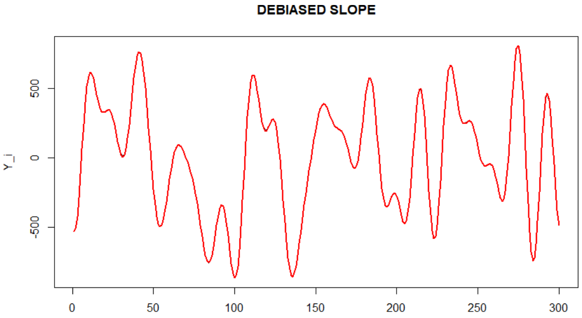
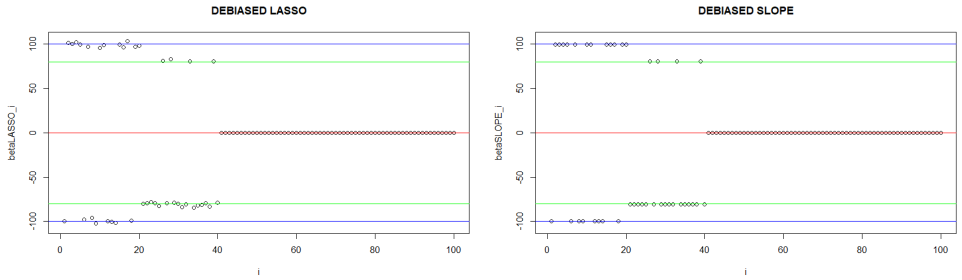


Figure 3. Signal denoising by debiased SLOPE on all coordinates of the regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The (almost overlapping) black line and the red line correspond respectively to the true values of $\boldsymbol{X}\boldsymbol{\beta}$ and to $\boldsymbol{Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$.

We also compare debiased SLOPE with debiased LASSO based on a single trial, as shown in Figure 4 and Table 1. The horizontal lines correspond to the



(a) (b)

Figure 4. Pattern recovery by debiased LASSO (left) and by debiased SLOPE (right) in the same setting as above.

true values of $\boldsymbol{\beta}$. As one may observe, in the presented setting LASSO does not recover the true support, while debiased SLOPE perfectly recovers support, sign and clusters.

Table 1. Comparison of MSE between different regression methods.

	OLS	LASSO-CV	LASSO-LS	SLOPE-LS
$MSE(\boldsymbol{\beta}, \cdot)$	613.6797	426.3705	171.7957	20.74967

6. APPENDIX

Proof of Lemma 2.1. Since the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is non-negative definite, it follows that the function $g : C_{\boldsymbol{\Lambda}} \to [0, \infty)$ defined by

$$g(\boldsymbol{\pi}) := (\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi})'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi})$$

is convex in $\boldsymbol{\pi}$. Therefore, at the point $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_p^*)'$ where g attains its global minimum over $C_{\boldsymbol{\Lambda}}$, the gradient ∇g satisfies

$$[\nabla g(\boldsymbol{\pi}^*)]'(\boldsymbol{\pi} - \boldsymbol{\pi}^*) \geqslant 0 \quad \text{for all } \boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}.$$

This implies $(\boldsymbol{\pi} - \boldsymbol{\pi}^*)'\boldsymbol{\beta}^* \leqslant 0$ for all $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$, because

$$\nabla g(\boldsymbol{\pi}^*) = -2(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi}^*) = -2\boldsymbol{\beta}^*.$$

In the proof of parts (a)–(c) we use the fact that $\boldsymbol{\pi}^*$ maximizes $\boldsymbol{\pi}'\boldsymbol{\beta}^*$ over $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$.

To prove (a) suppose that $\operatorname{sign}(\beta_i^*) \cdot \operatorname{sign}(\pi_i^*) < 0$ for some i and define

$$\boldsymbol{\pi} = (\pi_1^*, \ldots, \pi_{i-1}^*, -\pi_i^*, \pi_{i+1}^*, \ldots, \pi_p^*)'.$$

Then $(\boldsymbol{\pi}^*)'\boldsymbol{\beta}^* < \boldsymbol{\pi}'\boldsymbol{\beta}^*$, which is impossible since $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$.

To prove (b), consider a permutation τ of $(1,\ldots,n)$ such that the sequences $(|\pi^*_{\tau(1)}|,\ldots,|\pi^*_{\tau(p)}|)$ and $(|\beta^*_1|,\ldots,|\beta^*_p|)$ are similarly sorted. Define

$$\boldsymbol{\pi} = (s_1\cdot\pi^*_{\tau(1)}, s_2\cdot\pi^*_{\tau(2)},\ldots,s_p\cdot\pi^*_{\tau(p)}),$$

where $s_i = \operatorname{sign}(\beta^*_i)$ for $i=1,\ldots,p$. If $(|\pi^*_{\tau(1)}|,\ldots,|\pi^*_{\tau(p)}|) \neq (|\pi^*_1|,\ldots,|\pi^*_p|)$, then by the Hardy–Littlewood–Pólya rearrangement inequality,

$$\boldsymbol{\pi}'\boldsymbol{\beta}^* = \sum_{i=1}^{p} |\pi^*_{\tau(i)}|\,|\boldsymbol{\beta}^*_i| > \sum_{i=1}^{p} |\pi^*_i|\,|\boldsymbol{\beta}^*_i| \geqslant (\boldsymbol{\pi}^*)'\boldsymbol{\beta}^*,$$

which is impossible since $\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}$.

Finally, to prove (c), suppose that $\sum_{i=1}^{k-1}|\pi^*_{\tau(i)}| < \sum_{i=1}^{k-1}\lambda_i$ and $|\pi^*_{\tau(k)}|>0$. In this case there is a sufficiently small $\delta>0$ such that

$$\boldsymbol{\pi} = (\pi^*_1,\ldots,\pi^*_{i-2},\pi^*_{i-1}+\delta s_{i-1},\pi^*_i-\delta s_i,\pi^*_{i+1},\ldots,\pi^*_p)'\in C_{\boldsymbol{\Lambda}}.$$

If $|\beta^*_{\tau(k-1)}| > |\beta^*_{\tau(k)}|$ then

$$\boldsymbol{\pi}'\boldsymbol{\beta}^* = (\boldsymbol{\pi}^*)'\boldsymbol{\beta}^* + \delta(|\beta^*_{\tau(k-1)}| - |\beta^*_{\tau(k)}|) > (\boldsymbol{\pi}^*)'\boldsymbol{\beta}^*,$$

which is impossible. ∎

Proof of Lemma 2.2. First we note that for all $\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}$,

$$\begin{aligned} r(\boldsymbol{\beta}^*,\boldsymbol{\pi}) &= \tfrac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + \boldsymbol{\pi}'\boldsymbol{\beta}^* \\ &= \tfrac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + (\boldsymbol{\pi}^*)'\boldsymbol{\beta}^* + (\boldsymbol{\pi}-\boldsymbol{\pi}^*)'\boldsymbol{\beta}^* \\ &= r(\boldsymbol{\beta}^*,\boldsymbol{\pi}^*) + (\boldsymbol{\pi}-\boldsymbol{\pi}^*)'\boldsymbol{\beta}^* \leqslant r(\boldsymbol{\beta}^*,\boldsymbol{\pi}^*), \end{aligned}$$

where the inequality follows from the fact that $(\boldsymbol{\pi}-\boldsymbol{\pi}^*)'\boldsymbol{\beta}^*\leqslant 0$ for all $\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}$; see the proof of Lemma 2.1. Therefore, $\max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta}^*,\boldsymbol{\pi}) = r(\boldsymbol{\beta}^*,\boldsymbol{\pi}^*)$. Moreover, from the definition of $\boldsymbol{\beta}^*$ it can be seen that $r(\boldsymbol{\beta}^*,\boldsymbol{\pi}^*) = \min_{\beta\in\mathcal{M}} r(\boldsymbol{\beta},\boldsymbol{\pi}^*)$. These two facts imply that

$$\begin{aligned} \min_{\beta\in\mathcal{M}}\max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta},\boldsymbol{\pi}) &\leqslant \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta}^*,\boldsymbol{\pi}) = r(\boldsymbol{\beta}^*,\boldsymbol{\pi}^*) \\ &= \min_{\beta\in\mathcal{M}} r(\boldsymbol{\beta},\boldsymbol{\pi}^*) \leqslant \max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}}\min_{\beta\in\mathcal{M}} r(\boldsymbol{\beta},\boldsymbol{\pi}). \end{aligned}$$

Since

$$\max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}}\min_{\beta\in\mathcal{M}} r(\boldsymbol{\beta},\boldsymbol{\pi}) \leqslant \min_{\beta\in\mathcal{M}}\max_{\boldsymbol{\pi}\in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta},\boldsymbol{\pi})$$

(by the max-min inequality), we have equality throughout. This completes the proof. ∎

Poof of Lemma 3.1. Observe that

$$\frac{1}{n}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{Y} - \frac{2}{n}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{b},$$
$$\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{b}\|_2^2 = \frac{1}{n^2}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{Y} - \frac{2}{n}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{b}.$$

Therefore the two optimization problems differ by $\frac{1}{2n}(\boldsymbol{Y}'\boldsymbol{Y} - \frac{1}{n}\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{Y})$, which does not depend on $\boldsymbol{b}$, which implies their equivalence. ∎

Acknowledgments. The authors thank Małgorzata Bogdan, Patrick Tardivel, Wojciech Rejchel and Tomasz Żak for references, comments and discussion. The authors are grateful to the anonymous referees for improving the paper and for simplifying the proof of Lemma 3.1.

REFERENCES

[1] J.-P. Aubin, *Mathematical Methods of Game and Economic Theory*, North-Holland, 1980.

[2] M. Bogdan, E. van den Berg, C. Sabatti, W. Su and E. J. Candès, *SLOPE – Adaptive variable selection via convex optimization*, Ann. Appl. Statist. 9 (2015), 1103–1140.

[3] M. Bogdan, E. van den Berg, W. Su and E. J. Candès, *Statistical estimation and testing via the sorted ℓ_1 norm*, arXiv:1310.1969 (2013).

[4] M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel and M. Wilczyński, *Pattern recovery by SLOPE*, arXiv:2203.12086 (2022).

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

[6] H. D. Bondell and B. J. Reich, *Simultaneous factor selection and collapsing levels in ANOVA*, Biometrics 65 (2009), 169–177.

[7] S. Sh. Chen and D. L. Donoho, *Basis pursuit*, in: Proc. 1994 28th Asilomar Conference on Signals, Systems and Computers, IEEE, 1994, 41–44.

[8] S. Sh. Chen, D. L. Donoho and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. 20 (1998), 33–61.

[9] X. Dupuis and P. Tardivel, *Proximal operator for the sorted l_1 norm: Application to testing procedures based on SLOPE*, hal-03177108v2 (2021).

[10] K. Ewald and U. Schneider, *Uniformly valid confidence sets based on the Lasso*, Electron. J. Statist. 12 (2018), 1358–1387.

[11] M. A. T. Figueiredo and R. Nowak, *Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects*, in: Proc. 19th Int. Conf. on Artificial Intelligence and Statistics, Proc. Mach. Learning Res. 51, 2016, 930–938.

[12] J. Gertheiss and G. Tutz, *Sparse modeling of categorial explanatory variables*, Ann. Appl. Statist. 4 (2010), 2150–2180.

[13] P. Kremer, D. Brzyski, M. Bogdan and S. Paterlini, *Sparse index clones via the sorted ℓ_1-norm*, Quant. Finance 22 (2022), 349–366.

[14] A. Maj-Kańska, P. Pokarowski and A. Prochenka, *Delete or merge regressors for linear model selection*, Electron. J. Statist. 9 (2015), 1749–1778.

[15] K. Minami, *Degrees of freedom in submodular regularization: a computational perspective of Stein's unbiased risk estimate*, J. Multivariate Anal. 175 (2020), art. 104546, 22 pp.

[16] R. Negrinho and A. F. T. Martins, *Orbit regularization*, in: Advances in Neural Information Processing Systems 27, 2014, 9 pp.

[17] Sz. Nowakowski, P. Pokarowski and W. Rejchel, *Group Lasso merger for sparse prediction with high-dimensional categorical data*, arXiv:2112.11114 (2021).

[18] M.-R. Oelker, J. Gertheiss and G. Tutz, *Regularization and model selection with categorical predictors and effect modifiers in generalized linear models*, Statist. Model. 14 (2014), 157–177.

[19] K. R. Rao, N. Ahmed and M. A. Narasimhan, *Orthogonal transforms for digital signal processing*, in: Proc. 18th Midwest Symposium on Circuits and Systems, 1975, 1–6.

[20] U. Schneider and P. Tardivel, *The geometry of uniqueness, sparsity and clustering in penalized estimation*, arXiv:2004.09106 (2020).

[21] P. Tardivel, R. Servien and D. Concordet, *Simple expression of the LASSO and SLOPE estimators in low-dimension*, Statistics 54 (2020), 340–352.

[22] P. Tardivel, T. Skalski, P. Graczyk and U. Schneider, *The geometry of model recovery by penalized and thresholded estimators*, hal-03262087 (2021).

[23] B. G. Stokell, R. D. Shah and R. J. Tibshirani, *Modelling high-dimensional categorical data using nonconvex fusion penalties*, arXiv:2002.12606 (2021).

[24] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Statist. Soc. Ser. B. Statist. Methodology 101 (1996), 167–188.

[25] X. Zeng and M. A. T. Figueiredo, *Decreasing weighted sorted ℓ_1 regularization*, IEEE Signal Process. Lett. 21 (2014), 1240–1244.

[26] P. Zhao and B. Yu, *On model selection consistency of Lasso*, J. Mach. Learn. Res. 7 (2006), 2541–2563.

[27] H. Zou, *The adaptive lasso and its oracle properties*, J. Amer. Statist. Assoc. 101 (2006), 1418–1429.

Tomasz Skalski
Faculty of Pure and Applied Mathematics
Wrocław University of Science and Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
and
Laboratoire de Mathématiques LAREMA
Université d'Angers
2 Boulevard Lavoisier
49045 Angers Cedex 01, France
E-mail: tomasz.skalski@pwr.edu.pl

Piotr Graczyk
Laboratoire de Mathématiques LAREMA
Université d'Angers
2 Boulevard Lavoisier
49045 Angers Cedex 01, France
E-mail: graczyk@univ-angers.fr

Bartosz Kołodziejek
Faculty of Mathematics and Information Science
Warsaw University of Technology
Koszykowa 75
00-662 Warszawa, Poland
E-mail: b.kolodziejek@mini.pw.edu.pl

Maciej Wilczyński
Faculty of Pure and Applied Mathematics
Wrocław University of Science and Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland
E-mail: maciej.wilczynski@pwr.edu.pl

Received 21.9.2021;
accepted 13.10.2022