**WROCŁAW UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**Faculty of Pure and Applied Mathematics**

**UNIVERSITÉ D'ANGERS**

**Laboratoire Angevin de REcherche en MAthématiques, UMR CNRS 6093**

# DOCTORAL DISSERTATION

Tomasz Skalski

# Geometric and Combinatorial Aspects of Statistical Models

Supervisor: dr hab. Maciej Wilczyński, prof. WUST

Supervisor: prof. dr hab. Piotr Graczyk

Co-advisor in France: dr Patrick Tardivel

Wrocław, March 2023

**POLITECHNIKA WROCŁAWSKA**

**Wydział Matematyki**

**UNIVERSITÉ D'ANGERS**

**Laboratoire Angevin de REcherche en MAthématiques, UMR CNRS 6093**

# ROZPRAWA DOKTORSKA

Tomasz Skalski

# Geometryczne i kombinatoryczne zagadnienia modeli statystycznych

Promotor rozprawy: dr hab. Maciej Wilczyński, prof. uczelni

Promotor rozprawy: prof. dr hab. Piotr Graczyk

Promotor pomocniczy we Francji: dr Patrick Tardivel

Wrocław, marzec 2023

**UNIVERSITÉ D'ANGERS**

**Laboratoire Angevin de REcherche en MAthématiques, UMR CNRS 6093**

**POLITECHNIKA WROCŁAWSKA**

**Wydział Matematyki**

# THÈSE DE DOCTORAT

Tomasz Skalski

# Aspects Géométriques et Combinatoires de Modeles Statistiques

Directeur de thèse: prof. Piotr Graczyk

Co-directeur de thèse: prof. Maciej Wilczyński

Co-encadrant en France: Patrick Tardivel, MCF

Wrocław, Mars 2023

# Acknowledgements

First of all, I would like to express the gratitude to my supervisors Professors Maciej Wilczyński and Piotr Graczyk for their guidance and support during the time of my PhD studies. Thank you for the opportunity of sharing the studies between Wrocław and Angers and for your encouragement to submit the results of our long and fruitful discussions for publication. I would like to thank Dr. Patrick Tardivel for agreeing to be the co-encadrant in France and to extend our research to discussions in Dijon. Many thanks to Professor Krzysztof Bogdan for his guidance from the beginning of the PhD studies, for many inspiring discussions and for initiating the co-tutelle agreement with Université d'Angers. I would like to express my sincere gratitude to Prof. Małgorzata Bogdan for introducing me to the SLOPE estimator. I owe special appreciation to the late Prof. Hélène Massam, who agreed to take care of my journey through the theory of graphical models.
I owe many thanks to other co-authors of my research articles: Dr. Hab. Bartosz Kołodziejek, Dr. Michał Bosy, Dr. Hab. Ulrike Schneider and Dr. Xavier Dupuis for broading the scope of my research and for letting me gather more knowledge. Our co-operation was a pleasure.
I would like to thank the organizers and participants of seminars related to the topic of the dissertation, which helped us to exchange the current knowledge, as well in Wrocław (Random Graphs and Discrete Structures, Mathematical Statistics, Theory of Markov Semigroups and Schrödinger Operators), as in Angers (Probability and Statistics seminar, Doctorants' seminar) and on-line (Statistical Learning Seminars).
Special thanks are due to Dr. Clément Elvira, Dr. Hab. Cédric Herzet, Dr. Hab. Wojciech Rejchel, Dr. Tomasz Stroiński, Dr. Samuel Vaiter and Prof. Piotr Zwiernik for mind-opening discussions, which helped me seeing the other branches of the currently presented research. I would also want to thank the anonymous referees of our articles for their insightful comments and for improving them.
This dissertation would be not possible to complete without a good atmosphere in the academic community. I would like to thank my colleagues from 6.06 and KNM for our long discussions on mathematics, life et al. Special thanks are due to prof. Janusz Górniak for his enormous help during the beginning of my academic adventure and to Monika Kaczmarz for her invaluable support and care of PhD students and their initiatives.
I would also like to thank everyone, who helped my to adapt in Angers, especially to Małgorzata Graczyk and Alexandra Le Petitcorps.
I would like to thank from all my heart to my family, especially to my Parents, who taught me how to count and think and supported my through the PhD, even during the harder times. Thanks are also due to all the people whom I've met while pursuing my hobbies, especially to the Lingustics Olympiad community and the Vytautas' Light Cavalry.

The major part of the research was supported by a French Government Scholarship. The research in Wrocław University of Science and Technology was supported in part by grant 049M/0010/19 from WUST. The research in Université d'Angers was supported by Centre Henri Lebesgue, program ANR-11-LABX-0020-0. Thank you for helping me to focus on research.

# Contents

# Summary

This dissertation concerns new applications of discrete geometry and combinatorics in modern statistics. First of them focuses on one of widely used remedies to the inevitable growth of data, that is the use of penalized linear regresion methods. With an aim to recover the needed properties possessed by the vector of regression coefficients, we start our discussion with the Sorted $\ell_1$ Penalized Estimator (SLOPE), which was proposed almost a decade ago. Especially, we examine the notion of the SLOPE pattern, which maintains the information about the support, sign and ranking between the regression coefficients. In particular, it preserves the clusters of coefficients with the same absolute value. In Chapter 3 we provide the conditions, under which SLOPE recovers the set of relevant covariables and the clusters when the design matrix is orthogonal. We also derive new results on the strong consistency of the SLOPE estimator and its pattern. Chapter 4 extends the discussion on SLOPE to a general class of fixed design matrices. We provide the SLOPE irrepresentability condition, which is necessary and sufficient for the pattern recovery in the noiseless case and illustrate it geometrically. Later on, we consider the case of asymptotic growth of the number of explanatory variables and of the incremental error. In Chapter 5 we study the wider class of penalized estimators, called the polyhedral gauges. It allows one to use the notions from the geometry of polyhedra to generalize the notion of the pattern and the results on its recovery. Chapter 6 is articulated around the existence of the Maximum Likelihood Estimator (MLE) for discrete exponential families. We give its new characterization based on the notion of the set of uniqueness. Later on, we inspect the size of independent identically distributed samples which is needed to ensure its existence with high probability. For that reason we use the notions from the analysis of discrete hypercubes and apply our results in the environment of random graphs. Last of the chapters connects the theory of graphical models in statistics with the notion of graph Laplacian matrices and discretized Wiener processes. The thesis is based on three already published articles and two preprints, which are available on-line.

# Streszczenie

Niniejsza rozprawa poświęcona jest nowym zastosowaniom geometrii dyskretnej i kombinatoryki w nowoczesnej statystyce. Pierwsze z nich skupione jest na jednym z popularniejszych rozwiązań na radzenie z ciągłym przyrostem danych, jest nim penalizowana regresja liniowa. Mając na celu odtworzenie potrzebnych nam własności wektora współczynników regresji, rozpoczynamy dyskusję od estymatora SLOPE (Sorted $\ell_1$ Penalized Estimator), który został wprowadzony w poprzedniej dekadzie. Szczególną uwagę poświęcamy pojęciu wzorca SLOPE, który zachowuje informację o nośniku, znaku i rankingu między współczynnikami regresji. Informuje on również o klastrach współczynników o tej samej wartości bezwzględnej. W rozdziale trzecim podajemy warunki, dla których SLOPE poprawnie odtwarza nośnik oraz klastry wektora współczynników regresji przy ortogonalnej macierzy eksperymentu. Przy tym założeniu wyprowadzamy też nowe wyniki dotyczące mocnej zgodności estymatora SLOPE i jego wzorca. Rozdział czwarty rozszerza dyskusję na temat SLOPE, pomijając założenie o ortogonalności macierzy eksperymentu. Wprowadzamy warunek niereprezentowalności dla SLOPE, który jest konieczny i dostateczny do odtworzenia wzorca w przypadku braku szumu, po czym ilustrujemy ten warunek geometrycznie. Następnie rozważamy przypadek asymptotycznego przyrostu liczby zmiennych objaśniających i szumu rosnącego inkrementalnie. W rozdziale piątym omawiamy szerszą klasę penalizowanych estymatorów zwaną polyhedral gauges. Pozwala ona na wykorzystanie twierdzeń z geometrii wielościanów do uogólnienia pojęcia wzorca i wyników dotyczących jego odtwarzania. Rozdział szósty dotyczy istnienia estymatora największej wiarogodności (MLE) w dyskretnych rodzinach wykładniczych. Podajemy jego pełną charakteryzację za pomocą pojęcia zbioru jednoznaczności. Następnie badamy rozmiar próby niezależnych zmiennych losowych o tym samym rozkładzie, która zapewnia istnienie MLE z wysokim prawdopodobieństwem. W tym celu wykorzystujemy narzędzia z analizy hipersześcianów dyskretnych i stosujemy otrzymane wyniki w modelach wykładniczych grafów losowych. Ostatni z rozdziałów skupiony jest na połączeniu między teorią modeli graficznych w statystyce, a pojęciami laplasjanu grafu oraz dyskretyzacji procesów Wienera. Rozprawa jest oparta na trzech opublikowanych artykułach oraz dwóch preprintach dostępnych on-line.

# Résumé

Cette thèse traite des applications de la combinatoire et de la géométrie discrète aux statistiques modernes. La première application porte sur l'une des approches fréquemment utilisées pour faire face au volume croissant des données, à savoir l'utilisation de méthodes de régression linéaire pénalisée. Dans le but de retrouver la structure que possède le vecteur des coefficients de régression, nous commençons notre discussion par l'estimateur SLOPE (Sorted $\ell_1$ Penalized Estimator), qui a été proposé il y a presque dix ans. En particulier, nous examinons la notion de schéma SLOPE, qui conserve l'information sur le support, le signe et le classement des coefficients de régression. En particulier, il préserve les groupes de coefficients ayant la même valeur absolue. Dans le chapitre 3, nous fournissons les conditions, sous lesquelles SLOPE récupère l'ensemble des variables pertinentes et des groupes lorsque la matrice de planification est orthogonale. Nous déduisons également de nouveaux résultats sur la forte consistance de l'estimateur SLOPE et de son schéma. Le chapitre 4 étend la discussion sur l'estimateur SLOPE à une classe générale de matrices de planification fixes. Nous fournissons la condition d'irreprésentabilité du SLOPE, qui est nécessaire et suffisante pour la recouvrement du schéma dans le cadre non-bruité et nous illustrons cette propriété géométriquement. Dans un deuxième temps, nous considérons le cas asymptotique lorsque le nombre de variables explicatives tend vers l'infini et que l'erreur est incrémentale. Dans le chapitre 5, nous étudions la classe plus large des estimateurs pénalisés, appelés les jauges polyédriques. Elle permet d'utiliser les notions issues de la géométrie des polyèdres pour généraliser la notion du schéma et les résultats sur sa récupération. Le chapitre 6 s'articule autour de l'existence de l'estimateur du maximum de vraisemblance (MLE) pour les familles exponentielles discrètes. Nous donnons sa nouvelle caractérisation basée sur la notion d'ensemble d'unicité. Par la suite, nous inspectons la taille des échantillons indépendants identiquement distribués qui est nécessaire pour assurer son existence avec une grande probabilité. Pour cela, nous utilisons les notions issues de l'analyse des hypercubes discrets et appliquons nos résultats dans l'environnement des graphes aléatoires. Le dernier des chapitres relie la théorie des schémas graphiques en statistique avec la notion de matrices laplaciennes de graphes et de processus de Wiener discrétisés. La thèse est basée sur trois articles déjà publiés et deux prépublications, qui sont disponibles en ligne.

# Wstęp

Obecnie można zauważyć szybki i nieunikniony przyrost danych, zarówno pod względem liczby obserwacji, jak i pod względem liczby sposobów, za pomocą których da się je zmierzyć. Duże zbiory danych i ich analiza rosną na znaczeniu w życiu codziennym, dzięki czemu statystyka matematyczna oraz nauki związane z analizą danych zyskują coraz większe zainteresowanie w innych działach matematyki i jej zastosowaniach. Jednak statystyka i analiza danych nie powinny być traktowane w oderwaniu od innych działów matematyki. Kluczowym punktem badań niniejszej rozprawy będzie znajdowanie nowych połączeń między współczesną statystyką, geometrią i kombinatoryką.
Dyskusja podjęta w tej rozprawie rozpoczyna się od rozważania przestrzeni euklidesowych o skończonym wymiarze. Zwracamy uwagę na przypadki, w których zbiór wartości danego estymatora można podzielić na skończoną liczbę podzbiorów w taki sposób, żeby istniała bijekcja między nami, a interesującymi nas własnościami tego estymatora. Taki sposób predykcji obserwacji jest znany jako problem klasyfikacji [32], nad którym badania zostały rozpoczęte przez Fishera [77] i który jest stosowany w prawie każdym dziale nauk związanych z przetwarzaniem danych.

## PL.1 Penalizowana regresja liniowa

Za jedno z istotnych zastosowań problemu klasyfikacji możemy uznać wersję regresji liniowej, w której zamiast dokładnego oszacowania wartości nieznanego wektora interesują nas jego wybrane właściwości. W modelu regresji liniowej wielorakiej mającym $n$ obserwacji i $p$ zmiennych objaśniających zakładamy, że zmienna objaśniana $\boldsymbol{Y} = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$ jest postaci $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, gdzie $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ jest macierzą eksperymentu, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ jest nieznanym wektorem współczynników regresji, a $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathbb{R}^n$ jest losowym wektorem błędu (szumu). Głównym zagadnieniem regresji liniowej są oszacowanie $\boldsymbol{\beta}$ oraz wydobycie jego istotnych własności. Szeroka klasa zagadnień opartych na wyborze podzbioru współrzędnych $\boldsymbol{\beta}$, które będą najlepiej spełniać oczekiwania dotyczące estymatora i jego złożoności, znany jest jako problem wyboru modelu. Więcej informacji o tym zagadnieniu można znaleźć w przeglądowym artykule [49]. Przez większość dyskusji zakładamy, że wektor błędu $\boldsymbol{\varepsilon}$ jest losowy, a jego rozkład jest ciągły i symetrycznyn. Niektóre z przedstawionych wyników dotyczą również przypadku niezaszumionego, w którym $\boldsymbol{\varepsilon} \equiv \boldsymbol{0}$, co pomoże w wyprowadzeniu wyników asymptotycznych w przypadku zaszumionym. Dzięki istniejącym wynikom dotyczącym rozkładu normalnego możemy pokazać ulepszone wyniki dla szumu gaussowskiego $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. Najbardziej klasyczna metoda regresji liniowej, metoda najmniejszych kwadratów (Ordinary Least Squares, OLS), została wprowadzona przez Legendre'a na początku XIX wieku [10, 97, 124]. Estymator OLS jest zdefiniowany jako wektor $\boldsymbol{b}$ minimalizujący odległość euklidesową między $\boldsymbol{Y}$, a $\boldsymbol{X}\boldsymbol{b}$:

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} := \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2.$$

Popularność estymatora najmniejszych kwadratów ma miejsce ze względu na swoją prostotę, dokładność przy względnie małej liczbie zmiennych [33, str. 4] i wielu innych przydatnych statystycznych własności. Dla przykładu, jeżeli macierz $\boldsymbol{X'X}$ jest odwracalna, a $\mathbf{Cov}(\boldsymbol{Y}) = \sigma^{\mathbf{2}}\boldsymbol{I}_{\mathbf{n}}$ dla $\sigma > 0$, to OLS jest najlepszym estymatorem w klasie nieobciążonych liniowych estymatorów (BLUE) [1, Theorem 2.7.1.] wektora $\boldsymbol{\beta}$. W tym przypadku istnieje wzór jawny na $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, który można w dosyć łatwy sposób wyprowadzić [1, str. 28]:

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$$

Po dodaniu założenia o gaussowskości wektora $\boldsymbol{Y}$ estymator OLS staje się estymatorem największej wiarogodności (MLE) [1, str. 28]. W przypadku, gdy macierz eksperymentu $\boldsymbol{X}$ jest ortogonalna, tzn. $\boldsymbol{X'X} = c\boldsymbol{I}_p$, $c > 0$, powyższy wzór sprowadza się do $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \frac{1}{c}\boldsymbol{X'Y}$. Przy powyższych założeniach OLS jest również mocno zgodnym estymatorem dla $\boldsymbol{\beta}$ [6]. Z drugiej strony, estymator ten nie jest jednoznacznie zdefiniowany, kiedy macierz $\boldsymbol{X'X}$ jest nieodwracalna, co ma miejsce w wysokowymiarowym przypadku, gdy $p > n$. Ponadto, przy ogólnie przyjmowanych założeniach o błędzie $\boldsymbol{\varepsilon}$, OLS nie jest estymatorem rzadkim, z prawdopodobieństwem 1 składa się on z $p$ współrzędnych o parami różnych wartościach. Mniej klasyczne założenia o wektorze $\boldsymbol{Y}$ i macierzy $\boldsymbol{X}$, przy których estymator nie jest jednoznaczny, można znaleźć między innymi w niedawno opublikowanym artykule Dupuisa i Vaitera [62]. W praktyce często wektor $\boldsymbol{\beta}$ składa się ze względnie małej liczby niezerowych współrzędnych, co w naturalny sposób sugeruje metody, które promują rzadkość wektora $\boldsymbol{\beta}$ rozumianą jako małą liczbę niezerowych współrzędnych lub opisywalność $\boldsymbol{\beta}$ za pomocą małej liczby parametrów. Istnieją różne propozycje rozwiązania powyższego problemu, m.in. porównanie pasujących modeli przy pomocy kryterium informacyjnego, np. BIC [158] lub AIC [2]. Innym podejściem jest penalizowana regresja liniowa postaci:

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + pen_\lambda(\boldsymbol{b})\right],$$

gdzie $\mathrm{pen}_\lambda(\boldsymbol{b})$ jest ustaloną nieujemną karą, którą można modyfikować za pomocą parametru $\lambda > 0$. To podejście jest wykorzystywane m.in. w neuroobrazowaniu [39], prognozowaniu cen energii elektrycznej [106, 180] oraz w matematyce finansowej w celu grupowaniu aktywów na podstawie ich korelacji częściowej z szeregiem czasowym stóp zwrotu z funduszy inwestycyjnych [116]. Pierwszym z zaproponowanych sposobów penalizowanej regresji była metoda wyboru najlepszego podzbioru [12, 102], gdzie kara $\mathrm{pen}_\lambda(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_0$ jest równa liczbie współrzędnych niezerowych wektora $\boldsymbol{b}$ pomnożonej przez $\lambda$. Kłopot z jej zastosowaniem polega na tym, że dla dużych wartości $p$ znalezienie najlepszego podzbioru w ten sposób jest problemem NP-trudnym [78]. Inną metodą na znalezienie rzadkiego rozwiązania jest metoda LASSO (skrót od Least Absolute Shrinkage and Selection Operator [47, 176]), w której karą nałożoną na estymator $\boldsymbol{b}$ jest jego norma $\ell_1$ pomnożona przez parametr $\lambda > 0$:

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

Estymator LASSO jest obciążony. Jako estymator ściągający (shrinkage estimator) sprowadza on współrzędne $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ w stronę zera. Niektóre z nich zostają ściągnięte całkowicie do zera, co przynosi rezultat w postaci rzadszego estymatora. W przypadku, gdy macierz eksperymentu $\boldsymbol{X}$ jest ortonormalna, tzn. $\boldsymbol{X'X} = \boldsymbol{I}_p$, jawny wzór na $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$, wyprowadzony w pracy Tibshiraniego [176], jest oparty na $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$:

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

Aby zapewnić istnienie rozwiązania zagadnienia penalizowanej regresji liniowej, często używana jest kara wypukła, co umożliwia wykorzystanie narzędzi z analizy wypukłej. Zainteresowanym różnymi przykładami pomysłów na modyfikacje estymatora LASSO można polecić artykuł [78]. Warto zauważyć, że kary niewypukłe również są wykorzystywane m.in. przy wyborze najlepszego pozdbioru [12, 102] lub estymatora SCAD (smoothly clipped absolute deviation [72]).

## PL.2 SLOPE

Innym sposobem na zmniejszenie wymiaru jest estymator SLOPE (Sorted $\ell_1$ Penalized Estimator [27, 26, 189]), który poza uogólnieniem metody LASSO, skleja takie same współczynniki regresji $\boldsymbol{\beta}$ oraz skorelowane kolumny macierzy $\boldsymbol{X}$. Zgodnie z nazwą, w estymatorze SLOPE norma $\ell_1$ w funkcji kary została zastąpiona przez zdefiniowaną poniżej posortowaną normę $\ell_1$:

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

gdzie $\lambda_1 > 0, \lambda_1 \geq \ldots, \lambda_p \geq 0$, natomiast $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ to wartości bezwzględne współrzędnych wektora $\boldsymbol{b}$ posortowane malejąco. Szczególny przypadek SLOPE z $\boldsymbol{\Lambda}$ będącym ciągiem arytmetycznym jest znany pod nazwą OSCAR [29]. Zauważmy, że w parametr $\lambda$ został zastąpiony w SLOPE przez nierosnący wektor $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$ parametrów (tuning vector). Możemy więc nie tylko modyfikować wielkość wektora $\boldsymbol{\Lambda}$, ale również jego kształt. Główną motywacją autorów metody SLOPE było testowanie $p$ hipotez zerowych $H_i^0 : \beta_i = 0$ i kontrola współczynnika fałszywych odkryć (FDR control) zdefiniowanego jako oczekiwana proporcja między liczbą fałszywych odrzuceń hipotezy zerowej, a łączną liczbą jej odrzuceń (w przypadku braku odrzuceń przyjmuje się FDR = 0). Ponadto, SLOPE jest uogólnieniem poprzednich metod:

- $\lambda_1 = \ldots = \lambda_p = 0 \Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{OLS}}$,
- $\lambda_1 = \ldots = \lambda_p > 0 \Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{LASSO}}$,
- $\boldsymbol{\Lambda}$ is an arithmetic sequence $\Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{OSCAR}}$.

Dokładniejszy opis historii badań nad estymatorem SLOPE znajduje się w podrozdziale 4.1.1. W naszych badaniach skupiamy się na odtworzeniu istotnych właściwości wektora $\boldsymbol{\beta}$ za pomocą estymatora SLOPE, które są zakodowane w wektorze zwanym wzorcem SLOPE. Tak jak znak wektora $\boldsymbol{\beta}$ w przypadku LASSO, wzorzec SLOPE można całkowicie opisać za pomocą subróżniczki funkcji kary. Dokładniej mówiąc, niech $k$ będzie liczbą klastrów wektora $\boldsymbol{\beta}$, tzn. liczbą niezerowych różnych wartości współrzędnych $|\boldsymbol{\beta}|$.

**Definicja PL.2.1** (Wzorzec SLOPE)**.** *Wzorzec SLOPE* $\boldsymbol{patt} : \mathbb{R}^p \to \mathbb{Z}^p$ *jest zdefiniowany następująco:*

$$\boldsymbol{patt}(\boldsymbol{b})_i = \text{sign}(b_i)rank(|b_i|),$$

*gdzie* $rank(|b_i|) \in \{1, 2, \ldots, k\}$ *jest zdefiniowany jako liczba takich* $|c_j|$*, dla których* $|b_i| \geq |c_j|$*, gdzie* $|c_1|, |c_2|, \ldots, |c_k|$*,* $k \leq p$ *są niezerowymi różnymi wartościami spośród* $|b_1|, \ldots, |b_p|$*. Przyjmujemy* $rank(0) = 0$*.*

Wzorzec SLOPE zachowuje informację nie tylko o znaku wektora, ale również o jego klastrach, tzn. zbiorach współrzędnych mających tę samą wartość bezwzględną oraz o hierarchii między tymi wartościami.

**Fakt PL.2.1** (Własności wzorca SLOPE [156])**.**
(a) Dla każdego $1 \leq l \leq \|\boldsymbol{patt}(\boldsymbol{b})\|_\infty$ istnieje takie $j$, że $|\boldsymbol{patt}(\boldsymbol{b})_j| = l$,

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

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

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

**Example PL.2.2.** $\boldsymbol{patt}((4, 0, -1.5, 1.5, -4)') = (2, 0, -1, 1, -2)'$.

Mówimy, że $\widehat{\boldsymbol{\beta}}^{SLOPE}$ odtwarza wzorzec $\boldsymbol{\beta}$, kiedy

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

W rozdziale trzecim i czwartym omawiamy nowe warunki konieczne i dostateczne na odtwarzanie wzorca SLOPE, jak również nowe wyniki o zgodności i mocnej zgodności SLOPE i jego wzorca w przypadku, gdy $n \geq p$.

### PL.2.1 Zgodność

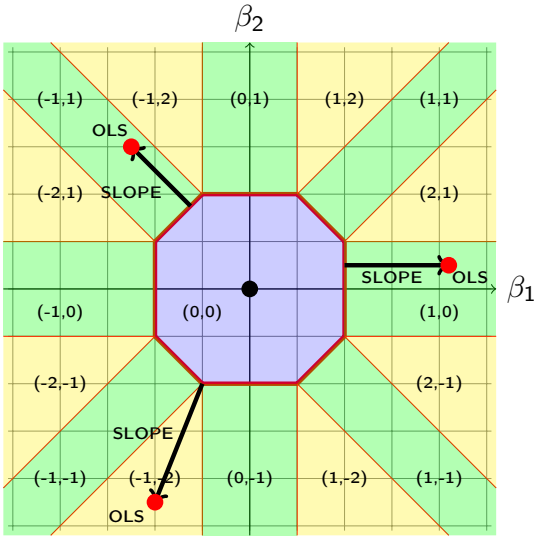
Głównym narzędziem do udowodnienia mocnej zgodności SLOPE w przypadku $n \geq p$ jest mocna zgodność estymatora największych kwadratów udowodniona m.in. w pracy Andersona i Taylora [6]. Główne wyniki dotyczące zgodności LASSO można znaleźć w następujących publikacjach:

- zgodność LASSO: Knight, Fu, 2000 [112],
- mocna zgodność LASSO: Chatterjee, Lahiri, 2011 [44],
  Główne założenia: $\lambda_n$ jest rzędu mniejszego niż $n$ oraz $\mathbb{E}|\boldsymbol{\varepsilon}| < \infty$.
- zgodność znaku LASSO: Zhao, Yu, 2006 [192].

Podrozdziały 3.4 oraz 4.6 niniejszej rozprawy dotyczą mocnej zgodności SLOPE i jego wzorca. Warto zauważyć, że w przypadku, gdy wektor $\boldsymbol{\Lambda}$ jest stały (LASSO), wzorzec SLOPE nie jest zgodny, nawet jeśli założenia o zgodności wektora znaku są spełnione. Estymator LASSO nie odtwarza klastrów, które są elementem wzorca SLOPE. Zauważmy też, że mocna zgodność estymatora LASSO (SLOPE) nie implikuje mocnej zgodności jego znaku (ani jego wzorca). Jako kontrprzykład można rozważyć ciąg $((1/n, 1/n, -1/n, -1/n)')_{n\geq 1}$, który zbiega do $(0, 0, 0, 0)'$, podczas gdy jego znak i jego wzorzec SLOPE są równe $(1, 1, -1, -1)'$ dla wszystkich $n$.

### PL.2.2 Warunek niereprezentowalności

Dla danych $n, p > 0$ zgodność zbioru niezerowych współrzędnych wektora $\hat{\boldsymbol{\beta}}^{LASSO}$ jest prawie równoważna do warunku, w którym współrzędne spoza nośnika wektora $\boldsymbol{\beta}$ nie są reprezentowane przez współrzędne doń należące [192]. Z tego powodu warunek został nazwany warunkiem niereprezentowalności (irrepresentability condition). W rozdziale 4. rozprawy wprowadzony jest analogon powyższego warunku dla estymatora SLOPE.




$\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ w przypadku ortogonalnym: $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ for $\boldsymbol{\Lambda} = (2,1)'$.

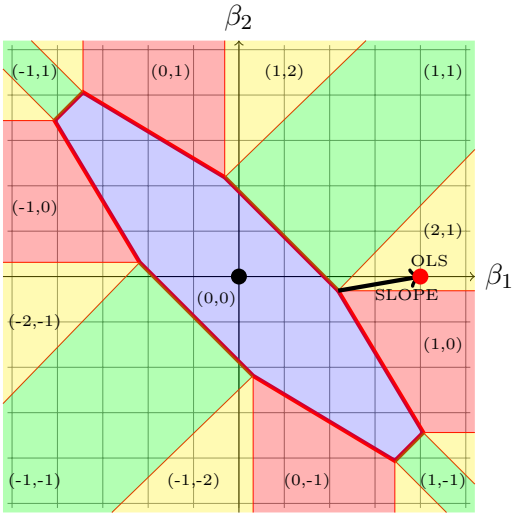
### PL.2.3 Penalizowana regresja liniowa i jej geometria

Dla lepszego zrozumienia powiązań między $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$, a $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, ilustrujemy je w przypadku niskowymiarowym $p \leq n$ dla $p = 2$. Tutaj wartości estymatora SLOPE jest równa różnicy między estymatorem najmniejszych kwadratów, a jednym z jego rzutów na $(\boldsymbol{X}'\boldsymbol{X})^{-1}C_{\boldsymbol{\Lambda}}$, gdzie $C_{\boldsymbol{\Lambda}}$ jest kulą jednostkową w normie dualnej do $J_{\boldsymbol{\Lambda}}$. Powyższy wynik ma łatwą interpretację w przypadku ortogonalnym, w którym zarówno $\boldsymbol{X}$, jak i rzut na $C_{\boldsymbol{\Lambda}}$, są ortogonalne. Ta zależność ma rezultat w postaci jawnego wzoru na $\hat{\boldsymbol{\beta}}^{SLOPE}$ w przypadku ortogonalnym, który został niedawno wprowadzony przez Tardivela, Serviena i Concordeta [175]. Więcej ilustracji związków między estymatorem SLOPE, a jego geometrią, można znaleźć w podrozdziale 4.4.

Zachodzi połączenie między estymatorami LASSO i SLOPE, a systemami pierwiastkowymi (root systems) wykorzystywanymi w analizie harmonicznej. Dokładniej mówiąc, można zauważyć, że kula jednostkowa w normie $\ell_\infty$ dualnej do kary stosowanej w LASSO oraz kula jednostkowa w normie $J^*_{\boldsymbol{\Lambda}}$ dualnej do normy SLOPE są proporcjonalne odpowiednio do powłok wypukłych orbit grup Weyla odpowiadających $p$-tej potędze systemu pierwiastkowego $A_1$ oraz systemowi $B_p$ [90, 144]. Więcej na temat systemów pierwiastkowych można znaleźć m.in. w książce Helgasona [100] oraz w publikacjach [92, 63].

Geometrię penalizowanej regresji liniowej i odtwarzania wzorców jej estymatorów można rozważać w ogólniejszym przypadku, co robimy w rozdziale piątym. Mianowicie, estymator SLOPE można zaklasyfikować do jednej ze skończonej liczby klas, gdy kara jest postaci polyhedral gauge, tzn. nieujemną funkcją wypukłą równą maksimum skończonej liczby funkcji liniowych. Istotne właściwości estymatora, również nazwane jego wzorcem (pattern), można całkowicie opisać za pomocą subróżniczki kary. Ponieważ nierówności liniowe definiujące polyhedral gauge są spełnione przez przekrój skończonej liczby związanych z nimi półprzestrzeni, zachodzi tutaj silne powiązanie z teorią wielościanów, z którą można zapoznać się. m.in. w książkach Grubera [94], Grünbauma [95], Hiriart-Urrutiego i Lemarechala [101] oraz Zieglera [193].

W rozdziale piątym uogólniamy niektóre z nowo otrzymanych wyników SLOPE na powyższą klasę estymatorów. W tym celu rozważamy pojęcie klasy równoważności wzorca (pattern equivalence class), która pozwala nam opisać odtworzenie wzorca jako odtworzenie wartości sub-




$$\hat{\boldsymbol{\beta}}^{\text{SLOPE}} \text{ oraz } \hat{\boldsymbol{\beta}}^{\text{OLS}} \text{ dla } \boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \text{ and } \boldsymbol{\Lambda} = (2,1)'.$$

różniczki kary pen.
Wprowadzamy i charakteryzujemy pojęcie osiągalności (accessibility) wzorca i wprowadzamy warunek niereprezentowalności dla polyhedral gauges. Z pomocą narzędzi z geometrii wielościanów, w szczególności pojęcia stożków normalnych ścian rozważanych wielościanów, rozważamy własności równoważności wzorca.
Następnie skupiamy się na odtwarzaniu modelu przez estymatory progowane. Są one uogólnieniem progowanego LASSO, jednak w odróżnieniu od swojego poprzednika ich celem jest zmniejszenie złożoności estymatora nie tylko poprzez sprowadzenie części współrzędnych do zera, ale również do innych jego uproszczeń związanych z danym wzorcem. Rozszerzamy też pojęcie jednorodnej jednoznaczności estymatora znanej dla norm wielościennych [156] do polyhedral gauges.

## PL.3 Dyskretne rodziny wykładnicze

W rozdziale szóstym przechodzimy z przestrzeni euklidesowych do przestrzeni ze skończoną liczbą elementów (dyskretnych). Punktem naszych zainteresowań są rodziny wykładnicze zdefiniowane na skończonej przestrzeni stanów $\mathcal{X}$, $|\mathcal{X}| = K < \infty$. Z powiązaną dyskusją na temat przeliczalnych zbiorów można zapoznać się w pracy Jacobsena [105]. My rozważamy podprzestrzeń liniową $\mathcal{B}$ przestrzeni funkcji $\mathbb{R}^{\mathcal{X}}$, do której należy funkcja stała dodatnia. Na zbiorze $\mathcal{X}$ wprowadzamy też ściśle dodatnią wagę $\mu : \mathcal{X} \to (0, \infty)$. Rodzinę wykładniczą konstruujemy następująco:
Dla rzeczywistej funkcji $\phi$ definiujemy odpowiednio jej funkcję partycji oraz log-partycji:

$$Z(\phi) = \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x), \quad \psi(\phi) = \log Z(\phi),$$

a także gęstość wykładniczą

$$p = e(\phi) = e^{\phi - \psi(\phi)} = e^{\phi} / Z(\phi).$$

Rodzina wykładnicza rozpięta przez $\mathcal{B}$ jest ukazana poniżej.

$$e(\mathcal{B}) := \{p = e(\phi) : \ \phi \in \mathcal{B}\}.$$

Dzięki temu możemy wprowadzić funkcję wiarogodności i log-wiarogodności. Ta ostatnia jest ściśle wklęsła, co zapewnia nam jednoznaczność MLE w przypadku jego istnienia. Istnienie estymatora nie jest zapewnione pomimo faktu, że funkcja wiarogodności jest ograniczona. W naszych badaniach wyprowadzamy nowe kryterium na istnienie MLE i wykorzystujemy je w rodzinach wykładniczych rozpiętych przez funkcje Rademachera i Walsha oraz w wykładniczych modelach grafów losowych.
Naszym głównym narzędziem jest pojęcie zbioru jednoznaczności, czyli takiego zbioru $U \subset \mathcal{X}$, że $\phi = 0$ jest jedyną funkcją z zadanej klasy, która jest równa zero na całym zbiorze $U$. Dokładniej, pokazujemy, że MLE istnieje dla $e(\mathcal{B})$ oraz próby i.i.d. $x_1, x_2, \ldots, x_n \in \mathcal{X}$ wtedy i tylko wtedy, gdy zbiór $\{x_1, \ldots, x_n\}$ jest jednoznaczności dla nieujemnego stożka funkcji $\mathcal{B}_+ := \{\phi \in \mathcal{B} : \ \phi \geq 0\}$. Powyższe kryterium podajemy również jako zagadnienie programowania liniowego.
Dodatkowo, w rozważanych zastosowaniach wyznaczamy asymptotyczne wielkości zbiorów jednoznaczności. W tym celu używamy klasycznych wyników z teorii grafów losowych oraz z problemu zbieracza kuponów (Coupon Collector's Problem) [118, str. 194-195], [68], and [143].
W rodzinie rozpiętej przez funkcje Walsha wykorzystujemy własności hipersześcianu $\{-1, 1\}^k$ ($k = \log_2 K$), jego podkości oraz związanego z nim grafu hipersześcianu.

## PL.4 Modele graficzne

Jednym z działów nowoczesnej statystyki wykorzystujących zarówno MLE, jak i estymatorów penalizowanych jest teoria modeli graficznych (graphical models). Model graficzny jest rodziną rozkładów prawdopodobieństwa skończonego zbioru zmiennych losowych $X_1, X_2, \ldots, X_N$, które są przedstawione za pomocą $N$ wierzchołków grafu (skierowanego lub nieskierowanego, zależnie od zastosowania). Obecność krawędzi pomiędzy dwoma wierzchołkami świadczy w modelu o zależności warunkowej między powiązanymi z nimi zmiennymi. Z gaussowskim modelem graficznym mamy do czynienia, kiedy zmienne $X_1, X_2, \ldots, X_N$ są z rozkładu normalnego. W tym przypadku pełna informacja o strukturze niezależności warunkowej między zmiennymi jest zawarta w macierzy odwrotnej do macierzy kowariancji (w macierzy precyzji) wektora losowego $\boldsymbol{X} = (X_1, \ldots, X_N)'$. Czytelnika zainteresowanego głębszym zapoznaniem się z modelami graficznymi zachęcamy do książek wprowadzających do tej teorii [122, 187, 20]. Dyskusję nad istnieniem MLE w dyskretnych modelach hierarchicznych, w tym w modelach graficznych, można znaleźć w pracy Wanga, Rauha i Massam [185]. Odpowiednik estymatora LASSO w modelach graficznych został wprowadzony około piętnastu lat temu [130, 188, 79] i jest w dalszym ciągu używany [37]. Graficzny odpowiednik SLOPE został wprowadzony niezależnie w pracy doktorskiej Sobczyka [166] oraz w pracy Mazzy-Anthony'ego, Mazourego i Coatesa [129].
Wróćmy do struktury niezależności warunkowej w modelu graficznych. Można zauważyć, że położenie zer w macierzy precyzji wektora $\boldsymbol{X}$ jest identyczne z położeniem zer w laplasjanie grafu opisującego tę strukturę. Nasza dyskusja skupiona jest na tych z procesów Markowa, których dyskretyzacje mają macierz kowariancji równą odwrotności macierzy zbliżonej do laplasjanu zadanego grafu. Jako przykład można podać sytuację, w której zadany graf jest $n$-wierzchołkową ścieżką. Wtedy dodanie 1 do lewego górnego wyrazu laplasjanu skutkuje macierzą odwrotną do $(\boldsymbol{\Sigma})_{i,j} = \min\{i, j\}$, która to jest dyskretyzacją macierzy kowariancji

klasycznego procesu Wienera.

## PL.5 Plan rozprawy

Rozprawa jest oparta na pięciu artykułach. Trzy z nich są opublikowane, jeden został wysłany do czasopisma, a kolejny jest w przygotowaniu i niebawem zostanie wysłany. Preprinty nieopublikowanych artykułów są dostępne w repozytoriach arXiv oraz HAL.

1. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal [165] (z P. Graczykiem, B. Kołodziejkiem i M. Wilczyńskim), Probability and Mathematical Statistics, 42(2):283–302, 2022

2. Pattern recovery by SLOPE [23] (z M. Bogdan, X. Dupuisem, P. Graczykiem, B. Kołodziejkiem, P. Tardivelem i M. Wilczyńskim), wysłany do recenzji `https://arxiv.org/pdf/2203.12086.pdf`

3. Pattern Recovery in Penalized and Thresholded Estimation and its Geometry [93] (z P. Graczykiem, U. Schneider i P. Tardivelem), w przygotowaniu, `https://hal.science/hal-03262087v2/document`

4. Maximum likelihood estimation for discrete exponential families and random graphs [22] (z K. Bogdanem i M. Bosym), ALEA, 19, 1045–1070 (2022)

5. Remarks on Laplacian of Graphical Models in Various Graphs [164], Proceedings, GSI 2021, Paris, France, July 21–23, 2021

Plan niniejszej rozprawy jest następujący:
Rozdział drugi zawiera podstawowe pojęcia dotyczące wyników używanych w dalszych rozdziałach.
Rozdział trzeci skupia się na metodzie SLOPE w przypadku ortogonalnej macierzy eksperymentu, tzn. $\boldsymbol{X}'\boldsymbol{X} = n\boldsymbol{I}_p$. W tym przypadku wprowadzamy również nowe wyniki dotyczące mocnej zgodności estymatora SLOPE i jego wzorca. Następnie ilustrujemy numerycznie zastosowanie SLOPE w klasteryzacji do odtwarzania sygnałów o wysokiej frekwencji.
W rozdziale czwartym omawiamy odtwarzanie wzorca SLOPE bez ograniczeń z poprzedniego rozdziału. Pokazujemy, że odtwarzanie wzorca można scharakteryzować za pomocą dwóch warunków, które nazywamy 'positivity condition' oraz 'subdifferential condition'. Następnie wprowadzamy 'irrepresentability condition' dla SLOPE, który jest uogólnieniem 'irrepresentability condition' dla LASSO [82, 41] wraz z jego geometryczną interpretacją. W dalszej kolejności podajemy bardziej wyrafinowane warunki asymptotyczne na zgodność oraz mocną zgodność estymatora SLOPE i jego wzorca.
Rozdział piąty dotyczy odtwarzania wzorca w ogólniejszym przypadku kar postaci polyhedral gauges. Wyprowadzamy warunki konieczne i dostateczne na jednorodną jednoznaczność estymatora, tzn. jego jednoznaczność dla dowolnej wartości $\boldsymbol{Y} \in \mathbb{R}^n$. Również pokazujemy kryteria na odtwarzanie wzorca dla podanych estymatorów i ich progowanych (thresholded) odmian. Ponadto, uogólniamy irrepresentability condition na penalizację poprzez polyhedral gauges oraz opisujemy relacje pomiędzy wzorcem, a modelem rozważanym w publikacji Vaitera i in. [182].
W rozdziale szóstym podajemy konieczne i dostateczne warunki na istnienie estymatora największej wiarogodności (MLE) w dyskretnych rodzinach wykładniczych. Głównym narzędziem

używanym w tym celu jest pojęcie zbiorów jednoznaczności, czyli takich pozdbiorów przestrzeni stanów, na którym funkcja z danej klasy jest równa zero wtedy i tylko wtedy, gdy jest ona stale równa zero na całej przestrzeni. Następnie wykorzystujemy nasze kryterium do klasy funkcji Rademachera i Walsha oraz do modelu wykładniczych grafów losowych (ERGM).
W rozdziale siódmym rozważamy powiązania między gaussowskimi modelami graficznymi, a macierzami kowariancji zdyskretyzowanych procesów Wienera. W tym celu używamy laplasjanu grafu, który opisuje strukturę niezależności warunkowej zadanego modelu.

# Résumé

Récemment nous pouvons observer l'émergence rapide de données massives, tant au niveau du nombre d'observations que du nombre de variables explicatives à mesurer. En raison de l'augmentation de l'influence des big data sur la vie quotidienne, la statistique mathématique et la science des données gagnent un intérêt croissant dans le domaine des mathématiques et des sciences appliquées. Cependant, elles ne doivent pas être considérées séparément des autres branches des mathématiques. Dans cette thèse, nous nous attachons à proposer de nouveaux liens entre la statistique moderne, la géométrie et la combinatoire. Nous commençons notre discussion par les espaces euclidiens de dimension finie. Nous prêtons attention aux situations où l'espace des valeurs possibles de l'estimateur peut être partitionné en un nombre fini de sous-ensembles tels qu'il existe une correspondance bijective entre les sous-ensembles et les propriétés spécifiques des estimations. Cette prédiction des propriétés d'une observation est mieux connue sous le nom de problème de classification [32], dont la recherche a été lancée par Fisher [77] et qui est appliquée dans presque tous les domaines liés aux données.

## FR.1 Régression Linéaire Pénalisée

Parmi les applications importantes du problème de la classification, on peut citer les modifications de la régression linéaire, qui ne se concentrent pas sur la valeur exacte de l'estimation mais sur ses propriétés importantes.

Dans le modèle de régression linéaire multiple ayant $n$ observations et $p$ variables explicatives, nous supposons que le vecteur de réponse $\boldsymbol{Y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ est de la forme $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ est une matrice de planification, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ est un vecteur inconnu de coefficients de régression et $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathbb{R}^n$ est un bruit aléatoire. L'objectif principal de la régression est d'estimer $\boldsymbol{\beta}$ et de retrouver ses propriétés essentielles. La abondante classe de problèmes de choix du sous-ensemble de coordonnées de $\boldsymbol{\beta}$, qui est le plus approprié pour nos demandes sur l'estimation et sa parcimonie, est mieux connue sous le nom de problème de sélection de modèle. Pour une étude sur le problème du sélection de modèle, cf. [49]. Dans la plupart de nos discussions, nous supposons que l'erreur $\boldsymbol{\varepsilon}$ a une distribution symétrique et continue. Certains de nos résultats couvrent le cas sans erreur de $\boldsymbol{\varepsilon} \equiv \boldsymbol{0}$, qui est outil pour établir des résultats asymptotiques. En raison de l'abondance des connaissances existantes sur la distribution gaussienne, nous sommes également en mesure de présenter des résultats plus efficaces sur le comportement des estimateurs avec gausienne $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. La méthode de régression linéaire la plus classique, la méthode des moindres carrés ordinaires (OLS), a été proposée par Legendre au début du XIXe siècle [10, 97, 124]. L'estimateur OLS est défini comme le minimiseur de la somme des carrés des résidus, à savoir

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} := \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2.$$

OLS a gagné sa popularité dans le cas $n \geq p$ en raison de sa simplicité, de son efficacité pour un un nombre relativement bas de variables [33, str. 4] et de nombreuses propriétés statistiques utiles. Par exemple, si la matrice $\boldsymbol{X}'\boldsymbol{X}$ est inversible et $\mathbf{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$ pour $\sigma > 0$, alors les OLS constituent le meilleur estimateur linéaire sans biais (BLUE) [1, Theorem 2.7.1.] de $\boldsymbol{\beta}$. Dans ce cas, la formule exacte de $\hat{\boldsymbol{\beta}}^{OLS}$ peut être calculée à partir de l'équation suivante pour $\hat{\boldsymbol{\beta}}^{OLS}$ peut être facilement déduite [1, str. 28]:

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Si nous supposons en plus la gaussianité de $\boldsymbol{Y}$, alors les OLS sont aussi l'estimateur du maximum de vraisemblance (MLE) de $\boldsymbol{\beta}$ [1, str. 28]. Lorsque la matrice de planification $\boldsymbol{X}$ est orthogonale, c'est-à-dire que $\boldsymbol{X}'\boldsymbol{X} = c\boldsymbol{I}_p$, $c > 0$, la formule ci-dessus se résume à $\hat{\boldsymbol{\beta}}^{OLS} = c\boldsymbol{X}'\boldsymbol{Y}$. Sous les hypothèses ci-dessus, les OLS est également un estimateur fortement consistant de $\boldsymbol{\beta}$ [6]. Cependant, cet estimateur n'est pas défini de manière unique lorsque la matrice $\boldsymbol{X}'\boldsymbol{X}$ est unité. défini lorsque la matrice $\boldsymbol{X}'\boldsymbol{X}$ n'est pas inversible, ce qui est le cas dans un cadre de haute dimension $p > n$. De plus, sous des hypothèses communes sur le terme d'erreur $\boldsymbol{\varepsilon}$, l'estimation OLS n'est pas parcimonieux, avec la probabilité 1 elle contient $p$ coordonnées de valeurs mutuellement différentes. Pour une exemple d'hypothèses inhabituelles sur $\boldsymbol{Y}$ et $\boldsymbol{X}$ et d'absence d'estimation unique, nous pouvons référer à l'article récent par Dupuis et Vaiter [62]. Comme le vrai vecteur $\boldsymbol{\beta} \in \mathbb{R}^p$ des coefficients de régression pourrait contenir beaucoup moins de coordonnées non nulles, il existe une voie naturelle pour proposer des méthodes qui favoriseraient la parcimonie de $\boldsymbol{\beta}$, c'est-à-dire un petit nombre de coordonnées non nulles, ou du moins favoriseraient les estimations de $\boldsymbol{\beta}$ qui sont descriptibles par un petit nombre de paramètres. Plusieurs solutions ont été proposées pour traiter ce problème. L'une d'entre elles consiste à comparer les modèles appropriés par un critère d'information, par exemple BIC [158] ou AIC [2]. Une autre approche consiste à utiliser la régression pénalisée de la forme

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left[ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + pen_\lambda(\boldsymbol{b}) \right],$$

où $\text{pen}_\lambda(\boldsymbol{b})$ est un pénalisateur arbitraire non négatif, modifiable en fonction de la valeur de $\lambda > 0$. Elle a été appliquée, par exemple, à l'imagerie cérébrale [39], à la prévision des prix de l'électricité [106, 180] ou encore dans les mathématiques financières pour regrouper les actifs en fonction de leur corrélation partielle avec la série chronologique des rendements des fonds spéculatifs [116]. La première présence de cette idée est la sélection du meilleur sous-ensemble [12, 102] avec $\text{pen}_\lambda(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_0$ étant un nombre de coordonnées non nulles de $\boldsymbol{b}$ multiplié par $\lambda$. Cependant, pour de grandes valeurs de $p$, trouver le meilleur sous-ensemble par cette méthode est un problème NP-difficile [78]. Pour une solution plus parcimonieuse, on peut utiliser l'estimateus LASSO (Least Absolute Shrinkage and Selection Operator [47, 176]), dans lequel la pénalité ajoutée à la somme des carrés des résidus $\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ est une norme $\ell_1$ de $\hat{\boldsymbol{\beta}}$ et le paramètre de régularisation est $\lambda > 0$:

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} := \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \left[ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_1 \right].$$

L'estimateur LASSO n'est pas sans biais, en tant qu'estimateur de rétrécissement il apporte certains coefficients $\hat{\beta}_j^{LASSO}$ vers zéro. Certaines des coordonnées sont complètement réduites à zéro, ce qui donne une estimation plus parcimonie. Lorsque la matrice de planification $\boldsymbol{X}$ est orthonormée, c'est-à-dire $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, la formule exacte de $\hat{\boldsymbol{\beta}}^{LASSO}$ trouvée par Tibshirani [176] est basée sur $\hat{\boldsymbol{\beta}}^{OLS}$:

$$\hat{\beta}_i^{\text{LASSO}} = \text{sign}(\hat{\beta}_i^{\text{OLS}}) \max\left\{|\hat{\beta}_i^{\text{OLS}}| - \lambda, 0\right\}.$$

Pour garantir l'existence d'une solution du problème de la régression linéaire pénalisée, il y a nombreux propositions que la pénalité soit convexe. Cela permet également d'appliquer les outils de l'analyse convexe. Pour une comparaison plus large d'autres modifications d'un estimateur LASSO, nous renvoyons à l'article [78]. Cependant, il existe également des applications de pénalisateurs non convexes, par exemple la sélection du meilleur sous-ensemble [12, 102] ou l'estimateur SCAD (Smoothly Clipped Absolute Deviation [72]).

## FR.2 SLOPE

Une autre approche pour réduire la dimensionnalité est l'estimateur pénalisé trié $\ell_1$ (SLOPE [27, 26, 189]), qui, en plus de généraliser la méthode LASSO, regroupe les coefficients égaux de $\boldsymbol{\beta}$ et les colonnes corrélées de la matrice de planification $\boldsymbol{X}$. Comme son nom l'indique, dans SLOPE, la norme $\ell_1$ comme pénalisateur est remplacée par la norme $\ell_1$ triée :

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

où $\lambda_1 > 0, \lambda_1 \geq \ldots \geq \lambda_p \geq 0$ et $|\boldsymbol{b}|_{(1)} \geq \ldots \geq |\boldsymbol{b}|_{(p)}$ sont les valeurs absolues des coordonnées de $\boldsymbol{b}$ triées par ordre décroissant. Le sous-cas particulier de SLOPE avec $\boldsymbol{\Lambda}$ étant une séquence arithmétique est également connu sous le nom OSCAR [29]. Il est important de noter que dans SLOPE, un seul paramètre de régularisation $\lambda$ est remplacé par un vecteur non croissant $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$ de $p$ paramètres de régularisation (le vecteur de régularisation). Ainsi, il nous permet de modifier non seulement l'échelle de $\boldsymbol{\Lambda}$, mais aussi sa forme. La principale motivation des auteurs de SLOPE était de tester les $p$ hypothèses nulles $H_0^i : \beta_i = 0$ et le contrôle du taux de fausses découvertes (le contrôle FDR), qui est défini par la proportion attendue entre le nombre de faux rejets d'hypothèses nulles et le nombre total de rejets d'hypothèses nulles (en cas d'absence de rejet, le FDR est défini comme égal à zéro). De plus, SLOPE généralise certaines des approches précédentes en régression linéaire :

- $\lambda_1 = \ldots = \lambda_p = 0 \Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{OLS}}$,
- $\lambda_1 = \ldots = \lambda_p > 0 \Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{LASSO}}$,
- $\boldsymbol{\Lambda}$ est une suite arithmétique $\Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{OSCAR}}$.

Pour une description plus détaillée d'histoire des recherches sur l'estimateur SLOPE, voir la Section 4.1.1.
Dans notre recherche, nous nous concentrons sur une autre propriété importante de SLOPE, à savoir la récouvrement des caractéristiques importantes d'un vecteur de coefficients de régression $\boldsymbol{\beta}$, appelé son SLOPE schéma. Il s'agit d'un analogue du vecteur de signe dans LASSO et, comme son prédécesseur, il peut être entièrement décrit par le sous-différentiel de la norme pénalisante. Pour être plus spécifique, laissez $k$ être le nombre de clusters de $\boldsymbol{patt}(\boldsymbol{\beta}) = (m_1, \ldots, m_p)'$ c'est-à-dire le nombre de composantes distinctes non nulles de $|\boldsymbol{\beta}|$.

**Définition FR.2.1** (SLOPE schéma)**.** *Le SLOPE schéma est une fonction* $\boldsymbol{patt} : \mathbb{R}^p \to \mathbb{Z}^p$ *telle que*

$$\boldsymbol{patt}(\boldsymbol{b})_i = \text{sign}(b_i) rank(|b_i|),$$

*où* $rank(|b_i|) \in \{1, 2, \dots, k\}$ *est défini comme le nombre de* $|c_j|$*'s satisfaisant* $|b_i| \geq |c_j|$*, où* $|c_1|, |c_2|, \dots, |c_k|$*,* $k \leq p$*, sont des valeurs distinctes non nulles parmi* $|b_1|, \dots, |b_p|$*. Nous adoptons la convention que* $rank(0) = 0$*.*

Le SLOPE schéma ne préserve pas seulement le signe d'un vecteur, mais détecte également ses clusters, c'est-à-dire les ensembles de coordonnées partageant la même valeur absolue et l'ordre entre ces valeurs absolues (ordre entre les clusters).

**Fait FR.2.1** (Propriétés de base du SLOPE schéma [156])**.**

(a) pour chaque $1 \leq l \leq \|\boldsymbol{patt}(\boldsymbol{b})\|_\infty$ il existe $j$ tel que $|\boldsymbol{patt}(\boldsymbol{b})_j| = l$,

(b) $\text{sign}(\boldsymbol{patt}(\boldsymbol{b})) = \text{sign}(\boldsymbol{b})$ (préservation du signe),

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

(d) $|b_i| > |b_j| \Rightarrow |\boldsymbol{patt}(\boldsymbol{b})_i| > |\boldsymbol{patt}(\boldsymbol{b})_j|$ (préservation de l'ordre).

**Exemple FR.2.2.** $\boldsymbol{patt}((4, 0, -1.5, 1.5, -4)') = (2, 0, -1, 1, -2)'$.

On dit que l'estimateur SLOPE $\widehat{\boldsymbol{\beta}}^{SLOPE}$ récupère le schéma de $\boldsymbol{\beta}$ lorsque

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

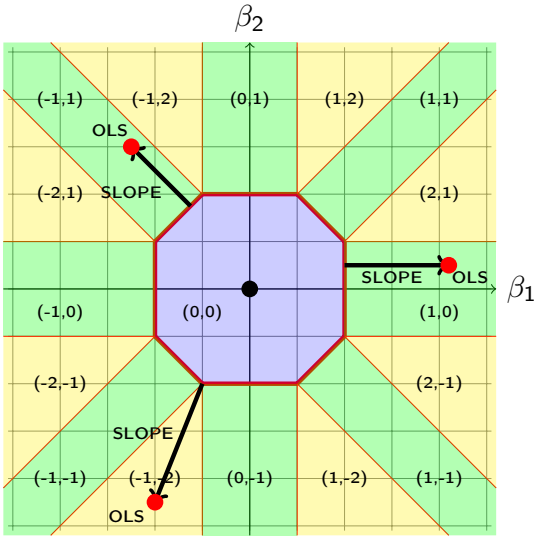
Dans les chapitres 3 et 4, nous discutons de nouvelles conditions nécessaires et suffisantes pour le recouvrement du SLOPE schéma, ainsi que de nouveaux résultats pour $n \geq p$ sur la consistance et la forte consistance de l'estimation SLOPE et de son schéma.

### FR.2.1 Consistance

L'outil principal pour montrer la consistance forte de SLOPE est la forte consistance de l'estimateur des moindres carrés, dont la preuve a été présentée par exemple dans l'article de Anderson et Taylor [6]. Les principaux résultats sur la consistance de LASSO peuvent être trouvés dans les articles suivants:

- consistance de LASSO: Knight, Fu, 2000 [112],
- forte consistance de LASSO : Chatterjee, Lahiri, 2011 [44],
  Les principales hypothèses sont que le paramètre de régularisation $\lambda_n$ est d'un ordre inférieur à $n$ et qu'il existe une espérance infinie de la valeur absolue d'un terme d'erreur.
- consistance du signe de LASSO: Zhao, Yu, 2006 [192].

Les sections 3.4 et 4.6 se concentrent sur la forte consistance de SLOPE et de son schéma. Il est important de rappeler que si le vecteur d'accord $\boldsymbol{\Lambda}$ est constant (LASSO comme cas particulier de SLOPE), alors le SLOPE schéma n'est pas consistant, même si l'hypothèse sur la consistance du vecteur de signe est satisfaite. A titre d'exemple, les clusters d'un vrai vecteur de paramètres, qui sont contenus dans un SLOPE schéma, ne sont pas préservés par LASSO. Comme autre remarque intéressante, on peut mentionner que la forte consistance de LASSO (ou SLOPE) n'implique pas nécessairement la forte consistance de son signe (de son schéma SLOPE). Comme contre-exemple, on peut remarquer que la séquence $((1/n, 1/n, -1/n, -1/n)')_{n\geq 1}$ converge vers $(0, 0, 0, 0)'$, alors que son signe et son schéma SLOPE sont égaux $(1, 1, -1, -1)'$ pour tout $n$ positif.



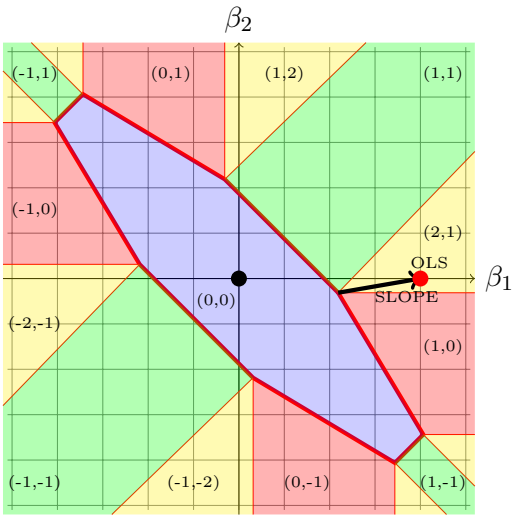

$\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$ et $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ dans un plan orthogonal: $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ pour $\boldsymbol{\Lambda} = (2, 1)'$.

### FR.2.2 Condition d'irreprésentabilité

Étant donné tout $n, p > 0$, dans LASSO, la consistance de la sélection du vrai sous-ensemble de $\boldsymbol{\beta}$ est presque équivalente à ce que les coordonnées en dehors du vrai support de $\boldsymbol{\beta}$ ne soient pas représentables par les coordonnées à l'intérieur du support [192]. Nous introduisons un analogue de cette condition d'irreprésentabilité de LASSO à SLOPE. Puis nous dérivons la caractérisation géométrique d'une estimation SLOPE.

### FR.2.3 Géométrie de la régression linéaire pénalisée

Pour une meilleure compréhension des connexions entre $\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$ et $\hat{\boldsymbol{\beta}}^{\text{OLS}}$, nous donnons une illustration d'un cas de bas dimension $p \leq n$ pour $p = 2$. Ici, l'estimation SLOPE est une différence entre l'estimation des moindres carrés et sa projection sur $(\boldsymbol{X}'\boldsymbol{X})^{-1}C_{\boldsymbol{\Lambda}}$, où $C_{\boldsymbol{\Lambda}}$ est une boule unité dans une norme double de $J_{\boldsymbol{\Lambda}}$. Ce résultat est particulièrement facile à interpréter dans le cas où $\boldsymbol{X}$ est une matrice orthogonale. Il en résulte une formule plus simple, qui a été récemment proposée par Tardivel, Servien et Concordet [175]. Pour une présentation plus large de l'application de l'approche géométrique à l'estimateur SLOPE et à sa récouvrement de schémas, nous invitons à la Section 4.4.
LASSO et SLOPE présentent une connexion avec la théorie des systèmes racines. En particulier, on peut observer qu'une boule unité en norme $\ell_\infty$, qui est duale de la norme $\ell_1$ utilisée dans LASSO et une boule unité en norme $J^*_{\boldsymbol{\Lambda}}$ sont proportionnelles, respectivement, aux coques convexes des orbites des groupes de Weyl correspondant au produit cartésien $p^{ieme}$ du système racine $A_1$ et au système racine $B_p$ [90, 144]. Pour plus d'informations sur la théorie des systèmes racine et ses recherches actuelles, nous renvoyons à un cahier de cours de Helgason [100] et à [92, 63]. La géométrie des estimateurs de régression linéaire pénalisés et leur recouvrement peuvent être inspectés dans un cas plus général, ce qui constitue l'essence du chapitre 5. En effet, l'estimateur SLOPE peut être classé dans l'une des classes d'un nombre fini de classes lorsque la pénalité est une gauge polyédrique, c'est-à-dire une fonction convexe non négative, qui est le maximum d'un nombre fini de fonctions linéaires. Les caractéristiques importantes de cette estimation, aussi appelée son schéma, peuvent être entièrement décrites avec le sous-




$$\hat{\boldsymbol{\beta}}^{\text{SLOPE}} \text{ et } \hat{\boldsymbol{\beta}}^{\text{OLS}} \text{ for } \boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix} \text{ et } \boldsymbol{\Lambda} = (2,1)'.$$

différentiel du pénaliseur. Puisque les inégalités pour les fonctions linéaires déterminant le gauge polyédrique sont satisfaites par une intersection du nombre fini de demi-espaces correspondants, ces estimateurs sont fortement liés à la théorie des polytopes, cf. les livres de Gruber [94], Grünbaum [95], Hiriart-Urruty et Lemarechal [101] ou Ziegler [193]. Dans le chapitre 5, nous généralisons certains des nouveaux résultats pour SLOPE à la classe d'estimateurs ci-dessus. Pour généraliser la notion du schéma à la classe des gauges polyédriques, nous considérons la classe d'équivalence des patrons, ce qui nous permet de désigner son recouvrement, comme le recouvrement de la sous-différentielle de pen.
Nous introduisons et caractérisons l'accessibilité du schéma et nous donnons une condition d'irreprésentabilité pour les gabarits polyédriques. Avec des outils issus de la géométrie polyédrique, en particulier avec les cônes normaux des faces des polyèdres considérés, nous discutons les propriétés de l'égalité des schémas entre vecteurs. Nous étudions également les relations entre les ensembles de vecteurs ayant le même schéma avec la notion de sous-espace modèle, cf. [182]. Plus tard, nous discutons de la récouvrement du schéma par les estimateurs pénalisés seuillés, qui sont une généralisation du LASSO seuillé, qui ne considère pas seulement le signe de l'estimateur, mais son schéma entier. Notre discussion étend également la caractérisation de l'unicité uniforme d'un estimateur de la classe des normes polyédriques [156] aux gauges polyédriques.

## FR.3 Familles exponentielles discrètes

Dans le chapitre 6, nous passons des espaces euclidiens aux espaces discrets, c'est-à-dire les espaces avec un nombre fini d'éléments. Ici, le point majeur de notre intérêt sont les familles exponentielles discrètes, que nous comprenons comme des familles exponentielles définies sur un espace fini d états $\mathcal{X}$, $|\mathcal{X}| = K < \infty$. Pour les familles sur les ensembles infinis dénombrables, voir Jacobsen [105]. Nous considérons ensuite un sous-espace linéaire $\mathcal{B}$ de l'espace des fonctions linéaires $\mathbb{R}^{\mathcal{X}}$, qui contient une fonction constante strictement positive. Nous introduisons également une fonction de poids strictement positive $\mu : \mathcal{X} \to (0, \infty)$. Nous construisons la famille exponentielle de la façon suivante:

Pour une fonction à valeur réelle $\phi$, nous définissons les fonctions de partition et de log-partition,

$$Z(\phi) = \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x), \quad \psi(\phi) = \log Z(\phi),$$

respectivement, et densité exponentielle

$$p = e(\phi) = e^{\phi - \psi(\phi)} = e^{\phi} / Z(\phi).$$

La famille exponentielle couverte par l'ensemble $\mathcal{B}$ est

$$e(\mathcal{B}) := \{p = e(\phi) : \ \phi \in \mathcal{B}\}.$$

Cela nous permet de dériver la fonction de vraisemblance et de log-vraisemblance. La fonction de log-vraisemblance est strictement concave. Par conséquent, si l'estimateur du maximum de vraisemblance (MLE) existe, il est unique. Malgré le caractère borné de la fonction de vraisemblance, l'MLE peut ne pas exister. Notre objectif principal est d'établir une nouvelle caractérisation de l'existence de l'MLE et de l'appliquer à des familles spécifiques, les familles exponentielles couvertes par les fonctions de Rademacher et de Walsh, et les familles exponentielles de graphes aléatoires.
L'outil clé dans notre discussion est la notion nouvellement introduite d'ensemble d'unicité, c'est-à-dire un tel ensemble $U \subset X$, que $\phi = 0$ est la seule fonction d'une classe donnée de fonctions qui est égale à zéro sur $U$. Pour être plus précis, nous montrons que le MLE pour $e(\mathcal{B})$ et l'échantillon i.i.d. $x_1, x_2, \ldots, x_n \in \mathcal{X}$ existe si et seulement si $\{x_1, \ldots, x_n\}$ est l'ensemble d'unicité pour le cône non négatif $\mathcal{B}_+ := \{\phi \in \mathcal{B} : \phi \geq 0\}$. Nous reformulons également ce critère sous la forme d'un problème de programmation linéaire.
De plus, pour les applications proposées, nous utilisons les limites de probabilité établies par les outils classiques de la théorie des graphes aléatoires et du problème du collecteur de coupons, voir [118, p. 194-195], [68], et [143]. Dans la famille des fonctions de Walsh, nous utilisons les propriétés de l'hypercube $\{-1, 1\}^k$ ($k = \log_2 K$), de ses sous-cubes et du graphe hypercube correspondant.

## FR.4 Modèles graphiques

L'une des branches de la statistique qui est connue à la fois pour l'utilisation de l'MLE et pour son efficacité et pour appliquer l'estimation pénalisée est la théorie des modèles graphiques. Un modèle graphique est une famille de distributions de probabilités d'une collection finie de variables aléatoires $X_1, \ldots, X_N$, qui sont codées par $N$ sommets d'un graphe (dirigé ou non dirigé). La présence (ou l'absence) d'une arête entre deux sommets renseigne sur la dépendance (ou l'indépendance) conditionnelle entre les sommets correspondants. Lorsque toutes les variables aléatoires sont gaussiennes, on parle de modèles graphiques gaussiens. Dans ce cas, la structure d'indépendance conditionnelle est entièrement codée par l'emplacement des zéros dans l'inverse de la matrice de covariance (la matrice de précision) d'un vecteur $\boldsymbol{X} = (X_1, \ldots, X_N)'$. Par savoir plus sur les modèles graphiques nous invitons à consulter les introductions existantes sur ce sujet [122, 187, 20]. Pour la discussion sur l'existence de MLE dans les modèles hiérarchiques discrets, qui incluent les modèles graphiques, nous nous référons à l'article de Wang, Rauh et Massam [185]. La version modèle graphique d'un estimateur LASSO a été proposée il y a environ quinze ans [130, 188, 79] et gagne encore en popularité, cf. par exemple [37]. La SLOPE graphique a été proposée indépendamment par Sobczyk dans sa thèse de doctorat [166] et par

Mazza-Anthony, Mazoure et Coates [129]. Revenons-en à la structure d'indépendance conditionnelle. On peut observer que l'emplacement des zéros dans la matrice de précision de $\boldsymbol{X}$ est le même que dans la matrice laplacienne d'un graphe sous-jacent. Dans notre discussion, nous examinons de plus près de tels graphes et recherchons ces processus de Markov, dont les discrétisations ont une matrice de covariance qui est l'inverse d'une matrice laplacienne un peu modifiée. À titre d'exemple, lorsqu'un graphique sous-jacent est un chemin de $n$ sommets, l'ajout de 1 à l'entrée supérieure gauche de sa matrice laplacienne donne une inverse égale à $(\boldsymbol{\Sigma})_{i,j} = \min\{i, j\}$, qui est une version discrétisée de la matrice de covariance d'un processus de Wiener.

## FR.5 Plan de la thèse

La thèse est basée sur cinq articles, dont trois sont publiés, un est en cours de révision, et un est une version étendue d'un preprint, qui peut être trouvé sur le site de HAL.

1. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal [165] (avec P. Graczyk, B. Kołodziejek et M. Wilczyński), Probability and Mathematical Statistics, 42(2):283–302, 2022

2. Pattern recovery by SLOPE [23] (avec M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, P. Tardivel et M. Wilczyński), soumis, `https://arxiv.org/pdf/2203.12086.pdf`

3. Pattern Recovery in Penalized and Thresholded Estimation and its Geometry [93] (avec P. Graczyk, U. Schneider et P. Tardivel), en préparation, `https://hal.science/hal-03262087v2/document`

4. Maximum likelihood estimation for discrete exponential families and random graphs [22] (avec K. Bogdan et M. Bosy), ALEA, 19, 1045–1070 (2022)

5. Remarks on Laplacian of Graphical Models in Various Graphs [164], Proceedings, GSI 2021, Paris, France, July 21–23, 2021

Le plan de la thèse est le suivant:
Le chapitre 2 contient les préliminaires de la recherche présentée et décrit les notions de base et nouvelles utilisées dans les chapitres suivants. Dans le chapitre 3, nous nous concentrons sur l'estimation du SLOPE dans le cas où la matrice de planification est orthogonale, c'est-à-dire, $\boldsymbol{X}'\boldsymbol{X} = n\boldsymbol{I}_p$. En ce cas, nous présentons également de nouveaux résultats sur la forte consistance des estimateurs SLOPE et sur la forte consistance de la récouvrement de schémas par SLOPE et nous illustrons les avantages du regroupement SLOPE dans le contexte du dénoissement de signaux à haute fréquence. Le chapitre 4 se concentre sur la récouvrement du SLOPE schéma sans restriction sur la matrice $\boldsymbol{X}$. Nous montrons que la récouvrement du schéma peut être déterminée par deux critères, appelés condition de positivité et condition subdifférentielle. Plus tard, nous introduisons une condition d'irreprésentabilité SLOPE, qui généralise la condition d'irreprésentabilité LASSO [82, 41] et nous l'illustrons ensuite géométriquement. Plus tard, nous dérivons les conditions asymptotiques raffinées sur la consistance et la consistance forte de l'estimation SLOPE et de son schéma. Le chapitre 5 concerne le problème de la récouvrement du schéma dans un cas général où la pénalité est une gauge polyédrique. Nous établissons la condition nécessaire et suffisante pour l'unicité uniforme de l'estimateur. Ensuite, nous dérivons des conditions pour la récouvrement du schéma des estimateurs pénalisés par une gauge polyédrique et pour leurs versions seuillées. Enfin, nous généralisons la condition d'irreprésentabilité pour les pénalisateurs étant des gauges

polyédriques et nous discutons des connexions entre un schéma et un modèle de l'article de Vaiter et al. [182].
Le chapitre 6 porte sur les conditions nécessaires et suffisantes pour l'existence de l'estimateur du maximum de vraisemblance (MLE) dans les familles exponentielles discrètes. L'outil principal de cet article est la notion d'ensembles d'unicité, c'est-à-dire les sous-ensembles d'un espace d'état sur lesquels une fonction d'une classe donnée est égale à zéro si et seulement si elle est égale à zéro sur un espace entier. Nous appliquons nos résultats à la classe des fonctions de Rademacher et de Walsh et aux modèles de graphes aléatoires exponentiels (ERGM). Dans le chapitre 7, nous abordons le lien entre les modèles graphiques gaussiens et les matrices de covariance des processus de Wiener discrétisés. Pour cela, nous utilisons la matrice laplacienne discrète d'une structure graphique sous-jacente.

# Chapter 1

# Introduction

In recent times we may observe the rapid and inevitable emergence of data collection, both in the number of observations and in the number of features to measure them. Due to the increase of the influence of big data on everyday life, mathematical statistics and data science gain growth of interest in mathematics and applied sciences. However, they should not be considered separately from other branches of mathematics. In this dissertation we focus on proposing new connections of modern statistics with geometry and combinatorics.
We start our discussion with finite-dimensional Euclidean spaces. We pay attention to those situations where the space of possible values of the estimator can be partitioned into a finite number of subsets such that there is a one-to-one correspondence between the subsets and specific properties of estimates. This prediction of the properties of an observation is better known as a classification problem [32], research on which was started by Fisher [77] and is being applied in almost all fields related to data processing.

## 1.1 Penalized Linear Regression

As one of important applications of the classification problem, we can point out the modifications of linear regression, which do not focus on the exact estimate value but on its important properties. In the multiple linear regression model having $n$ observations and $p$ explanatory variables, we assume that the response vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$ is of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown vector of regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathbb{R}^n$ is a random noise. The main objective of linear regression is to estimate $\boldsymbol{\beta}$ and recover its essential properties. The wide class of problems of choosing the subset of coordinates of $\boldsymbol{\beta}$, which is most suitable for our demands on the estimate and its sparsity, is better known as the model selection problem. For the survey on the model selection problem, cf. [49]. In most of our discussion, we assume that the error $\boldsymbol{\varepsilon}$ has a symmetric and continuous distribution. Some of our results cover the noiseless case of $\boldsymbol{\varepsilon} \equiv \mathbf{0}$, which is a strong tool for establishing asymptotic results. Due to the abundance of the existing knowledge on the Gaussian distribution, we are also able to present more efficient results on the behavior of considered estimators with Gaussian $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$. The most classical linear regression method, the Ordinary Least Squares method (OLS), was proposed by Legendre at the beginning of the nineteenth century [10, 97, 124]. The OLS estimator is defined as the minimizer of the residual sum of squares, namely

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} := \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2.$$

The OLS estimator gained its popularity in the case $n \geq p$ due to its simplicity, efficiency for a relatively small number of variables [33, p. 4] and many useful statistical properties. For example, if the matrix $\boldsymbol{X}'\boldsymbol{X}$ is invertible and $Cov(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}_n$ for $\sigma > 0$, then the OLS is the best linear unbiased estimator (BLUE) [1, Theorem 2.7.1.] of $\boldsymbol{\beta}$. In this case, the exact formula for $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ can be easily deduced [1, p. 28]:

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

If we additionally assume the gaussianity of $\boldsymbol{Y}$, then OLS is also the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ [1, p. 28]. When the design matrix $\boldsymbol{X}$ is orthogonal, i.e. $\boldsymbol{X}'\boldsymbol{X} = c\boldsymbol{I}_p$, $c > 0$, the above formula boils down to $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \frac{1}{c}\boldsymbol{X}'\boldsymbol{Y}$. Under the above assumptions, the OLS is also a strongly consistent estimator of $\boldsymbol{\beta}$ [6]. However, this estimator is not uniquely defined when the matrix $\boldsymbol{X}'\boldsymbol{X}$ is not invertible, which is the case in a high-dimensional setting $p > n$. Additionally, under common assumptions on the error term $\boldsymbol{\varepsilon}$, OLS estimate is not sparse, with probability 1 it contains $p$ coordinates of mutually different values. For an example of unusual assumptions on $\boldsymbol{Y}$ and $\boldsymbol{X}$ and no unique estimate, one may get acquainted with to a recent article by Dupuis and Vaiter [62]. As the true vector $\boldsymbol{\beta} \in \mathbb{R}^p$ of regression coefficients could contain much fewer nonzero coordinates, there is a natural pathway to propose methods that would promote the sparsity of $\boldsymbol{\beta}$, that is, a small number of nonzero coordinates, or at least to promote the estimates of $\boldsymbol{\beta}$ that are describable by a small number of parameters. Several solutions were proposed to deal with such problem. One of them is to compare the suitable models by an information criterion, for example BIC [158] or AIC [2]. Another approach is to use the penalized regression of the form

$$\hat{\boldsymbol{\beta}} := \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^p} \left[ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + pen_\lambda(\boldsymbol{b}) \right],$$

where $\mathrm{pen}_\lambda(\boldsymbol{b})$ is an arbitrary nonnegative penalizer, modifiable according to the value of $\lambda > 0$. It has been applied, for example, in brain imaging [39], forecasting electricity prices [106, 180] or in financial mathematics to group assets with respect to their partial correlation with the hedge fund return times series [116]. The first presence of this idea is the best subset selection [12, 102] with $\mathrm{pen}_\lambda(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_0$ being a number of nonzero coordinates of $\boldsymbol{b}$ multiplied by $\lambda$. However, for large values of $p$, finding the best subset by this method is an NP-hard problem [78]. For a sparser solution, one may use the Least Absolute Shrinkage and Selection Operator (LASSO [47, 176]), in which the penalty added to the residual sum of squares $\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$ is an $\ell_1$ norm of $\hat{\boldsymbol{\beta}}$ and the tuning parameter is $\lambda > 0$:

$$\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}} := \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^p} \left[ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda\|\boldsymbol{b}\|_1 \right].$$

The LASSO estimator is not unbiased, as a shrinkage estimator it brings some coefficients $\hat{\beta}_j^{\mathrm{LASSO}}$ toward zero. Some of the coordinates are being shrunk completely to zero, which results in a sparser estimate. When the design matrix $\boldsymbol{X}$ is orthonormal, i.e. $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, the exact formula for $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ found by Tibshirani [176] is based on $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$:

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

To guarantee the existence of a solution of the penalized linear regression problem, many proposals offer the penalty to be convex. This also allows one to apply the tools from convex analysis. For a broader comparison of other modifications of a LASSO estimator, we refer to the article [78]. However, there are also applications of nonconvex penalizers, e.g., the best subset selection [12, 102] or the smoothly clipped absolute deviation (SCAD) [72].

## 1.2 SLOPE

Another approach to reduce dimensionality is the Sorted $\ell_1$ Penalized Estimator (SLOPE [27, 26, 189]), which apart of generalizing the LASSO method, clusterizes the equal coefficients of $\boldsymbol{\beta}$ and correlated columns of the design matrix $\boldsymbol{X}$. As the name suggests, in SLOPE the $\ell_1$ norm as penalizer is replaced by the sorted $\ell_1$ norm:

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

where $\lambda_1 > 0, \lambda_1 \geq \ldots, \lambda_p \geq 0$ and $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ are the absolute values of coordinates of $\boldsymbol{b}$ sorted in descending order. The special subcase of SLOPE with $\boldsymbol{\Lambda}$ being an arithmetic sequence is also known as the OSCAR estimator [29]. It is important to note that in SLOPE a single tuning parameter $\lambda$ is replaced by a non-increasing vector $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$ of $p$ tuning parameters (the tuning vector). Thus, it allows us to modify not only the scaling of $\boldsymbol{\Lambda}$, but also its shape. The main motivation of the authors of SLOPE was the testing of $p$ null hypotheses $H_i^0 : \beta_i = 0$ and the control of the rate of false discoveries (the FDR control), which is defined as the expected proportion between the amount of false rejections of null hypotheses and the total amount of rejections of null hypotheses (in case of no rejections, the FDR is defined to equal zero). Moreover, SLOPE generalizes some of the previous approaches in linear regression:

- $\lambda_1 = \ldots = \lambda_p = 0 \Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{OLS}}$,
- $\lambda_1 = \ldots = \lambda_p > 0 \Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{LASSO}}$,
- $\boldsymbol{\Lambda}$ is an arithmetic sequence $\Rightarrow \hat{\boldsymbol{\beta}}^{\text{SLOPE}} = \hat{\boldsymbol{\beta}}^{\text{OSCAR}}$.

A more detailed description of the history of research on SLOPE can be found in Section 4.1.1. In our research, we focus on another important property of SLOPE, that is the recovery of important characteristics of a vector of regression coefficients $\boldsymbol{\beta}$, called its SLOPE pattern. It is an analogue of the sign vector in LASSO and, like its precedessor, it is fully describable by the subdifferential of the penalizing norm. To be more specific, let $k$ be the number of clusters of $\boldsymbol{patt}(\boldsymbol{\beta}) = (m_1, \ldots, m_p)'$ i.e., the number of distinct nonzero components of $|\boldsymbol{\beta}|$.

**Definition 1.2.1** (SLOPE pattern)**.** *The SLOPE pattern is a function* $\boldsymbol{patt} : \mathbb{R}^p \to \mathbb{Z}^p$ *such that*

$$\boldsymbol{patt}(\boldsymbol{b})_i = \text{sign}(b_i)rank(|b_i|),$$

*where* $rank(|b_i|) \in \{1, 2, \ldots, k\}$ *is defined to be the number of* $|c_j|$*'s satisfying* $|b_i| \geq |c_j|$*, where* $|c_1|, |c_2|, \ldots, |c_k|$*,* $k \leq p$*, are distinct nonzero values among* $|b_1|, \ldots, |b_p|$*. We adopt the convention that* $rank(0) = 0$*.*

The SLOPE pattern does not only preserve the sign of a vector, but also detects its clusters, i.e. the sets of coordinates sharing the same absolute value and the order between those absolute values (hierarchy between clusters).

**Fact 1.2.1** (Basic properties of SLOPE pattern [156])**.**

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

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

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

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

**Example 1.2.2.** $\boldsymbol{patt}((4, 0, -1.5, 1.5, -4)') = (2, 0, -1, 1, -2)'$.

We say that the SLOPE estimator $\widehat{\boldsymbol{\beta}}^{SLOPE}$ recovers the pattern of $\boldsymbol{\beta}$ when

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

In Chapter 3 and Chapter 4 we discuss novel necessary and sufficient conditions for the recovery of the SLOPE pattern, as well as the novel results for $n \geq p$ on the consistency and strong consistency of both the SLOPE estimate and its SLOPE pattern.

### 1.2.1 Consistency

The main tool to show the strong consistency of SLOPE is the strong consistency of the Ordinary Least Squares estimator, the proof of which was presented e.g. in the article of Anderson and Taylor [6]. Main results on the consistency of LASSO might be found in the following articles:

- consistency of LASSO: Knight, Fu, 2000 [112],
- strong consistency of LASSO: Chatterjee, Lahiri, 2011 [44],
  Main assumptions are that the tuning parameter $\lambda_n$ is of a smaller order than $n$ and that there is a finite expectation of the absolute value of an error term.
- consistency of the sign of LASSO: Zhao, Yu, 2006 [192].

Sections 3.4 and 4.6 focus on the strong consistency of SLOPE and its pattern. It is important to recall that if the tuning vector $\boldsymbol{\Lambda}$ is constant (LASSO as the special case of SLOPE), then the SLOPE pattern is not consistent, even if the assumption on the consistency of the sign vector is satisfied. As an example, the clusters of a true vector of parameters, which are contained in a SLOPE pattern, are not preserved by LASSO. As another interesting remark, it may be mentioned that the strong consistency of LASSO (or SLOPE) does not necessarily imply the strong consistency of its sign (of its SLOPE pattern). As a counterexample, one may notice that the sequence $((1/n, 1/n, -1/n, -1/n)')_{n\geq 1}$ converges to $(0, 0, 0, 0)'$, while its sign and its SLOPE pattern are equal $(1, 1, -1, -1)'$ for every positive $n$.

### 1.2.2 Irrepresentability condition

Given any $n, p > 0$, in LASSO, the consistency of selection of the true subset of $\boldsymbol{\beta}$ is almost equivalent to the coordinates outside the true support of $\boldsymbol{\beta}$ not being representable by the coordinates inside the support [192]. We introduce an analogue of this Irrepresentability Condition from LASSO to SLOPE. Then we derive the geometric characterization of a SLOPE estimate.

### 1.2.3 Geometry of penalized linear regression

For a better understanding of the connections between $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, below we give an illustration of a low-dimensional case $p \leq n$ for $p = 2$. Here, the SLOPE estimate is a difference between the least squares estimate and its projection onto $(\boldsymbol{X}'\boldsymbol{X})^{-1}C_{\boldsymbol{\Lambda}}$, where $C_{\boldsymbol{\Lambda}}$ is a unit ball in a norm dual to $J_{\boldsymbol{\Lambda}}$. This result is especially easy to interpret in the case of $\boldsymbol{X}$ being an orthogonal matrix. That resulted in an easier formula, which was recently proposed by Tardivel,

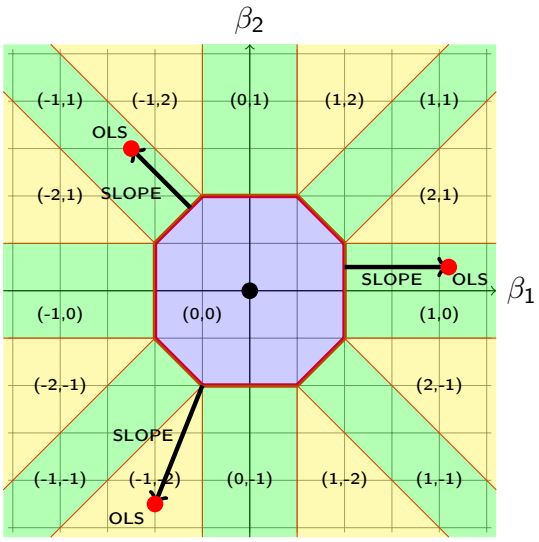



Figure 1.1: $\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ in orthogonal design: $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ for $\boldsymbol{\Lambda} = (2, 1)'$.

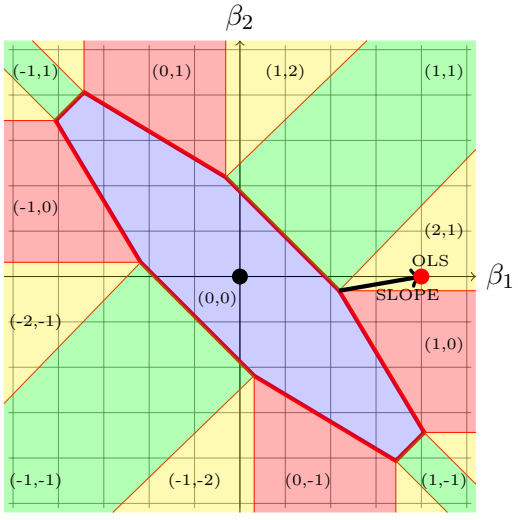



Figure 1.2: $\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ for $\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$ and $\boldsymbol{\Lambda} = (2, 1)'$.

Servien and Concordet [175]. For a wider presentation of applying the geometrical approach to the SLOPE estimator and its pattern recovery we invite the reader to Section 4.4.
LASSO and SLOPE estimators exhibit a connection to the theory of root systems. Namely, it may be observed that a unit ball in $\ell_\infty$ norm, which is dual to the $\ell_1$ norm used in LASSO and a unit ball in $J^*_{\mathbf{\Lambda}}$ norm are proportional, respectively, to the convex hulls of orbits of Weyl groups corresponding to the $p^{th}$ Cartesian product of the root system $A_1$ and of the root system $B_p$ [90, 144]. For more on the theory of root systems and its current research, we refer to a coursebook of Helgason [100] and to [92, 63].
The geometry of penalized linear regression estimators and their pattern recovery may be inspected in more general case, which is the essence of Chapter 5. Indeed, the SLOPE estimator can be classified into one of finite number of classes when the penalty is a polyhedral gauge, i.e. a nonnegative convex function, which is a maximum of a finite number of linear functions. The important features of this estimate, also named its pattern, may be fully described with the subdifferential of the penalizer. Since the inequalities for linear functions determining the polyhedral gauge are satisfied by an intersection of the finite number of corresponding half-spaces, those estimators are strongly connected with the theory of polytopes, cf. the books of Gruber [94], Grünbaum [95], Hiriart-Urruty and Lemarechal [101] or Ziegler [193].
In Chapter 5 we generalize some of the new results for SLOPE to the above class of estimators. To generalize the notion of pattern to the class of polyhedral gauges, we consider the pattern equivalence class, which allows us to denote its recovery, as the recovery of the subdifferential of pen.
We introduce and characterize the accessibility of the pattern and we give an irrepresentability condition for polyhedral gauges. With tools from polyhedral geometry, especially with the normal cones of faces of considered polyhedra, we discuss the properties of the pattern equality between vectors. We also investigate the relations between sets of vectors having the same pattern with the notion of the model subspace, cf. [182].
Later on, we discuss the pattern recovery by thresholded penalized estimators, which are a generalization of the thresholded LASSO, which considers not only the sign of the estimator, but its entire pattern, Our discussion also extends the characterization of the uniform uniqueness of an estimator from the class of polyhedral norms [156] to polyhedral gauges.

## 1.3 Discrete exponential families

In Chapter 6 we move our focus from Euclidean to discrete spaces, i.e. spaces with a finite number of elements. Here, the major point of our interest are the discrete exponential families, which we understand as exponential families defined on a finite space of states $\mathcal{X}$, $|\mathcal{X}| = K < \infty$. For families on infinite countable sets, see Jacobsen [105]. Then we consider a linear subspace $\mathcal{B}$ of the space of linear functions $\mathbb{R}^{\mathcal{X}}$, which contains a strictly positive constant function. We also introduce a strictly positive weight function $\mu : \mathcal{X} \to (0, \infty)$.We construct the exponential family in a following way:
For a real-valued function $\phi$ we define the partition and log-partition functions,

$$Z(\phi) = \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x), \quad \psi(\phi) = \log Z(\phi),$$

respectively, and *exponential density*

$$p = e(\phi) = e^{\phi - \psi(\phi)} = e^{\phi} / Z(\phi).$$

The exponential family spanned by $\mathcal{B}$ is

$$e(\mathcal{B}) := \{p = e(\phi) : \ \phi \in \mathcal{B}\}.$$

This allows us to derive the likelihood and log-likelihood function. The log-likelihood function is strictly concave. Therefore, if the Maximum Likelihood Estimator (MLE) exists, then it is unique. Despite the boundedness of the likelihood function, MLE may not exist. Our main goal is to establish a new characterization of the existence of MLE and to apply it to specific families, exponential families spanned by Rademacher and Walsh functions, and exponential families of random graphs.
The key tool used in our discussion is the newly introduced notion of set of uniqueness, that is such set $U \subset \mathcal{X}$, that $\phi = 0$ is the only function of a given class of functions that equal zero on $U$. To be more specific, we show that the MLE for $e(\mathcal{B})$ and the i.i.d. sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$ exists if and only if $\{x_1, \ldots, x_n\}$ is the set of uniqueness for the nonnegative cone $\mathcal{B}_+ := \{\phi \in \mathcal{B} : \ \phi \geq 0\}$. We also restate this criterion as a linear programming problem.
Additionally, for proposed applications, we use the probability bounds established by classical tools from random graph theory and from the Coupon Collector's Problem, see [118, pp. 194-195], [68], and [143].
In the family spanned by Walsh functions, we use the properties of the hypercube $\{-1, 1\}^k$ ($k = \log_2 K$), of its subcubes and of the corresponding hypercube graph.

## 1.4 Graphical models

One of those branches of statistics that is known both for using the MLE and for applying penalized estimation is the theory of graphical models. A graphical model is a family of probability distributions of a finite collection of random variables $X_1, X_2, \ldots, X_N$, which are encoded by $N$ vertices of a graph (directed or undirected). The presence (or absence) of an edge between two vertices informs about the conditional dependence (or independence) between the corresponding vertices. When all random variables are Gaussian, we refer to Gaussian graphical models. In this case, the conditional independence structure is fully encoded by the location of zeros in the inverse of the covariance matrix (the precision matrix) of a vector $\boldsymbol{X} = (X_1, \ldots, X_N)'$. We invite the reader interested in graphical models to the existing introductions for this topic [122, 187, 20]. For the discussion on the existence of MLE in discrete hierarchical models, which include graphical models, we refer to the article of Wang, Rauh and Massam [185]. The graphical model version of a LASSO estimator was proposed around fifteen years ago [130, 188, 79] and still gains popularity, cf. e.g. [37]. The graphical SLOPE was proposed independently by Sobczyk in his Ph.D. dissertation [166] and by Mazza-Anthony, Mazoure and Coates [129].
Let us come back to the conditional independence structure. It may be observed that the location of zeros in the precision matrix of $\boldsymbol{X}$ is the same as in the Laplacian matrix of an underlying graph. In our discussion we take a closer look at such graphs and look for those Markov processes, the discretizations of which have a covariance matrix being an inverse of a slightly modified Laplacian matrix. As an example, when an underlying graph is a path of $n$ vertices, then adding 1 to the top-left entry to its Laplacian matrix results in an inverse equal to $(\boldsymbol{\Sigma})_{i,j} = \min\{i, j\}$, which is a discretized version of a covariance matrix of a Wiener process.

## 1.5 Plan of the dissertation

The dissertation is based on five articles, three of which are published, one is currently under the review process, and one is an extended version of a preprint, which can be found on HAL and will be submitted soon.

1. Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal [165] (with P. Graczyk, B. Kołodziejek and M. Wilczyński), Probability and Mathematical Statistics, 42(2):283–302, 2022

2. Pattern recovery by SLOPE [23] (with M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, P. Tardivel and M. Wilczyński), submitted, `https://arxiv.org/pdf/2203.12086.pdf`

3. Pattern Recovery in Penalized and Thresholded Estimation and its Geometry [93] (with P. Graczyk, U. Schneider and P. Tardivel), in preparation, `https://hal.science/hal-03262087v2/document`

4. Maximum likelihood estimation for discrete exponential families and random graphs [22] (with K. Bogdan and M. Bosy), ALEA, 19, 1045–1070 (2022)

5. Remarks on Laplacian of Graphical Models in Various Graphs [164], Proceedings, GSI 2021, Paris, France, July 21–23, 2021

The outline of the dissertation is the following:
Chapter 2 contains preliminaries of the presented research and describes the basic and novel notions used in latter chapters.
In Chapter 3 we focus on the SLOPE estimation in case of design matrix being orthogonal, that is, $\boldsymbol{X}'\boldsymbol{X} = n\boldsymbol{I}_p$. We also present new results on the strong consistency of the SLOPE estimators and on the strong consistency of pattern recovery by SLOPE when the design matrix is orthogonal and illustrate the advantages of the SLOPE clustering in the context of high frequency signal denoising.
Chapter 4 focuses on the SLOPE pattern recovery with no restrictions on the design matrix. Here, we show that recovery of the pattern can be determined by two criteria, called the positivity condition and the subdifferential condition. Later, we introduce a SLOPE irrepresentability condition, which generalizes the well known LASSO irrepresentability condition [82, 41] and then we illustrate it geometrically. Later on, we derive the refined asymptotic conditions on both consistency and strong consistency of the SLOPE estimate and of its pattern.
Chapter 5 concerns the problem of pattern recovery in a general case of a penalty being a polyhedral gauge. We establish the necessary and sufficient condition for the uniform uniqueness of the estimator. Then we derive conditions for the pattern recovery of polyhedral gauge penalized estimators and for their thresholded versions. Finally, we generalize the irrepresentability condition for penalizers being polyhedral gauges and discuss the connections between a pattern and a model from the article by Vaiter et al. [182].
Chapter 6 focuses on the necessary and sufficient conditions for the existence of the Maximum Likelihood Estimator (MLE) in discrete exponential families. The main tool in this article is the notion of sets of uniqueness, i.e. such subsets of a state space on which a function from a given class equals zero if and only if it equals zero on a whole space. We apply our results to the class of Rademacher and Walsh functions and to exponential random graph models (ERGM).
In Chapter 7 we discuss the connection between Gaussian graphical models and the covariance

matrices of discretized Wiener processes. For that reason, we use the discrete Laplacian matrix of an underlying graphical structure.

# Chapter 2

# Preliminaries and basic notions on penalized linear regression

## 2.1 Notation

The content of this section will be completed after unifying the notation in the whole dissertation.

- $\boldsymbol{A}'$ — transpose of a matrix $\boldsymbol{A}$
- $|\boldsymbol{b}|_{\downarrow}$ — nonincreasing permutation of absolute values of coordinates of $\boldsymbol{b}$
- $\partial f(\boldsymbol{x})$ — subdifferential of $f$ at $\boldsymbol{x}$
- $B_{\|\cdot\|}(\boldsymbol{b}, r)$ — ball of center $\boldsymbol{b}$ and radius $r$ in norm $\|\cdot\|$
- $cl(C)$ — closure of $C$
- $int(C)$ — interior of $C$
- $\mathrm{aff}(C)$ — affine hull of $(C)$
- $\overrightarrow{C}$ — unique linear space parallel to $C$
- $cl(C)$ — closure of $C$
- $bd(C)$ — boundary of $C$
- $\mathrm{ri}(C)$ — relative interior of $C$
- $rb(C)$ — relative boundary of $C$
- $\mathrm{conv}(C)$ — convex hull of $C$
- $\mathrm{cone}(C)$ — conic hull of $C$
- $rec(C)$ — recession cone of $C$
- $P_C(\boldsymbol{x})$ — orthogonal projection of $\boldsymbol{x}$ onto $C$
- $C^*$ — polar set of $C$
- $\mathbf{1}_k = (1, \ldots, 1)' \in \mathbb{R}^k$

- $\mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$ — characteristic function of a set $A$
  For the convenience of the reader, when a set $A$ is defined as a set of points satisfying the relation $\phi$ we reduce the notation from $\mathbb{1}_{\{x|\phi(x)\}}(x)$ to $\mathbb{1}_{\{\phi(x)\}}$, e.g. $\mathbb{1}_{\{x>0\}}$.
- $\boldsymbol{J}_n \in \mathbb{R}^{n\times n}$, $(\boldsymbol{J}_n)_{ij} = 1$ for every $1 \leq i, j \leq n$.
- $supp(\boldsymbol{v}) = \{i \in \{1, \ldots, p\} |\ s_i \neq 0\}$ — support of a vector $\boldsymbol{v} \in \mathbb{R}^p$
- $\mathbf{diag}(\boldsymbol{S}) \in \mathbb{R}^{p\times p}$ — diagonal matrix with $\boldsymbol{S} \in \mathbb{R}^p$ on the diagonal
  $(\mathbf{diag}(\boldsymbol{S}))_{i,j} = s_i \mathbb{1}_{\{i=j\}}$
- $\mathcal{M}_p$ — set of all possible SLOPE patterns of $\boldsymbol{b} \in \mathbb{R}^p$
- $\operatorname{sign}(x) := \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ if } x = 0 \\ -1 \text{ if } x < 0 \end{cases}$ — sign function
- $\boldsymbol{sign}(\boldsymbol{x}) = (\operatorname{sign}(x_1), \ldots, \operatorname{sign}(x_p))'$ — sign vector
- $\mathbb{R}^{k+} := \{\boldsymbol{\kappa} \in \mathbb{R}^k :\ \kappa_1 > \kappa_2 > \ldots > \kappa_k > 0\}$
- $\mathcal{S}_p$ — symmetric permutation group on the set $\{1, \ldots, p\}$
- $Q^- := \max\{Q, 0\}$.

## 2.2 Penalized linear regression

### 2.2.1 Linear regression

We consider the following linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ is a response vector with $n$ being a sample size, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a design matrix with $p$ being the number of unknown parameters, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown parameter vector and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a noise term.
The case of $\boldsymbol{\varepsilon} = \mathbf{0}$ ($\boldsymbol{\varepsilon} \neq \mathbf{0}$) is referred later as a noiseless (noisy) case.
In the penalized linear regression we want to find an estimator $\hat{\boldsymbol{\beta}}$ of a vector $\boldsymbol{\beta}$, which is of the form

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min} \left[ \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + pen_{\boldsymbol{\Lambda}}(\boldsymbol{b}) \right],$$

where, depending of the estimation method, $\boldsymbol{\Lambda}$ is a tuning parameter or a vector of tuning parameters (tuning vector). When dealing with a single tuning parameter, we denote it with a small letter $\lambda$, restricting a capital letter $\boldsymbol{\Lambda}$ for a vector $(\lambda_1, \ldots, \lambda_p)'$.
As we are mostly interested with recovering a pattern of a vector $\boldsymbol{\beta}$, we are interested to define it in a general case. For that reason, we use a notion of a subdifferential and define two vectors to have the same pattern if the penalizer has the same subdifferential at those vectors

**Definition 2.2.1** (Subdifferential)**.** *[41, p. 76] Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function. A vector $\boldsymbol{d} \in \mathbb{R}^p$ is called a subgradient of $f$ at point $\boldsymbol{x} \in \mathbb{R}^p$ if*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + (\boldsymbol{y} - \boldsymbol{x})'\boldsymbol{d}, \qquad \forall\ \boldsymbol{y} \in \mathbb{R}^p.$$

*The set of all subgradients of the convex function $f$ at $\boldsymbol{x} \in \mathbb{R}^p$ is called the subdifferential of $f$ at $\boldsymbol{x}$, and it is denoted by $\partial f(\boldsymbol{x})$.*

**Remark 2.2.1.** [17, p. 716] $\boldsymbol{x} \in \mathbb{R}^p$ is a minimum of a convex function $f$ if and only if $\mathbf{0} \in \partial f(\boldsymbol{x})$.

**Example 2.2.2.** [89, Lemma D.5] The subdifferential of the $\ell_1$ norm at $\boldsymbol{x} \in \mathbb{R}^p$ is given by

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

The subdifferential of the $\ell_\infty$ norm at $\mathbf{0}$ is the unit ball of the $\ell_1$ norm and for $\boldsymbol{x} \in \mathbb{R}^p$ where $\boldsymbol{x} \neq \mathbf{0}$ this subdifferential is equal to

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

**Definition 2.2.2** (Equality of patterns)**.** *Let* pen $: \mathbb{R}^p \mapsto \mathbb{R}$ *be a convex penalizer. We say that $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{z} \in \mathbb{R}^p$ have the same pattern with respect to* pen *when* $\partial\text{pen}(\boldsymbol{x}) = \partial\text{pen}(\boldsymbol{z})$*, where* $\partial$pen *represents the subdifferential of* pen*.*

We say that the estimator $\widehat{\boldsymbol{\beta}}$ recovers the pattern of $\boldsymbol{\beta}$ when

$$\partial\text{pen}\left(\widehat{\boldsymbol{\beta}}\right) = \partial\text{pen}(\boldsymbol{\beta}).$$

**Example 2.2.3** (LASSO pattern)**.** In the LASSO regression, i.e., when $\text{pen}(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, the pattern of $\boldsymbol{\beta}$ identifies with its sign vector:

$$\boldsymbol{patt}(\boldsymbol{\beta}) = \boldsymbol{sign}(\boldsymbol{\beta}).$$

The proof of the above example goes straightforwardly from Example 2.2.2.

**Theorem 2.2.4** (Subdifferential description of the SLOPE pattern [156])**.**
*Let $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)$ satisfy $\lambda_1 > \ldots > \lambda_p > 0$. Then*

$$\boldsymbol{patt}(\boldsymbol{b}_1) = \boldsymbol{patt}(\boldsymbol{b}_2) \iff \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}_1) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}_2).$$

The pattern of SLOPE defined in Definition 2.2.2 coincides with the following definition.

**Definition 2.2.3** (SLOPE pattern [23])**.** *The SLOPE pattern is a function* $\boldsymbol{patt} : \mathbb{R}^p \to \mathbb{Z}^p$ *such that*

$$\boldsymbol{patt}(\boldsymbol{b})_i = \text{sign}(b_i) rank(|b_i|),$$

*where $rank(|b_i|) \in \{1, 2, \ldots, k\}$ is defined to be the number of $|c_j|$'s satisfying $|b_i| \geq |c_j|$, where $|c_1|, |c_2|, \ldots, |c_k|$, $k \leq p$, are distinct non-zero values among $|b_1|, \ldots, |b_p|$. We adopt the convention that $rank(0) = 0$.*

**Fact 2.2.5** (Basic properties of SLOPE pattern [156])**.**
(a) for every $1 \leq l \leq \|\boldsymbol{patt}(\boldsymbol{b})\|_\infty$ there exists $j$ such that $|patt(\boldsymbol{b})_j| = l$,

(b) $\text{sign}(\boldsymbol{patt}(\boldsymbol{b})) = \text{sign}(\boldsymbol{b})$ (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 2.2.6.** $\boldsymbol{patt}(4, 0, -1.5, 1.5, -4)' = (2, 0, -1, 1, -2)'$.

As a (nonzero) cluster in a vector $\boldsymbol{M} \in \mathbb{R}^p$ we denote a maximal (in terms of inclusion) set of indices $i$ in $supp(\boldsymbol{M})$ such that the absolute value $\boldsymbol{M}_i$ is the same. For example, the vector $(3, -3, 1, 3, 1, 4)'$ has three clusters: $\{1, 2, 4\}$, $\{3, 5\}$ and $\{6\}$.

**Definition 2.2.4.** *Let $\boldsymbol{M} \neq \boldsymbol{0}$ be a pattern in $\mathcal{M}_p$ with $k = \|\boldsymbol{M}\|_\infty$ nonzero clusters. The pattern matrix $\boldsymbol{U}_{\mathbf{M}} \in \mathbb{R}^{p \times k}$ is defined as follows*

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

**Example 2.2.7.** Let $\boldsymbol{\beta} = (3, -4, -3, 0, 4, 10)'$. Then the pattern of $\boldsymbol{\beta}$ equals $\boldsymbol{M} = (1, -2, -1, 0, 2, 3)'$ and

$$\boldsymbol{U}_{\mathbf{M}} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Note that $\boldsymbol{\beta}$ can be represented by a product of $\boldsymbol{U}_{\mathbf{M}}$ and the vector $(10, 4, 3)' \in \mathbb{R}^{3+}$. For every $\boldsymbol{\beta}$ with pattern $\boldsymbol{M}$ such a vector in $\mathbb{R}^{3+}$ exists.

**Definition 2.2.5.** *Let $\boldsymbol{M} \neq \boldsymbol{0}$ be a pattern in $\mathbb{R}^p$ and $k = \max\{\|\boldsymbol{M}\|_\infty, 1\}$.*
*For $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ we define the clustered design matrix by $\widetilde{\boldsymbol{X}}_{\mathbf{M}} = \boldsymbol{X}\boldsymbol{U}_{\mathbf{M}} \in \mathbb{R}^{n \times k}$ and the clustered parameter by $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = (\boldsymbol{U}_{|\mathbf{M}|_\downarrow})'\boldsymbol{\Lambda}$.*

**Example 2.2.8.** Let $\boldsymbol{X} = (\boldsymbol{X}_1|\boldsymbol{X}_2|\boldsymbol{X}_3|\boldsymbol{X}_4)$, $\boldsymbol{M} = (1, 2, -1, 0)'$
and $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_4)' \in \mathbb{R}^{4+}$. Then $|\mathbf{M}|_\downarrow = (2, 1, 1, 0)'$. The clustered matrix and the clustered parameter equal:

$$\widetilde{\boldsymbol{X}}_{\mathbf{M}} = (\boldsymbol{X}_2|\boldsymbol{X}_1 - \boldsymbol{X}_3) \text{ and } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \begin{pmatrix} \lambda_1 \\ \lambda_2 + \lambda_3 \end{pmatrix}.$$

## 2.3 Convex Polytopes and cones

### 2.3.1 Convex analysis

We consider $\mathbb{R}^n$ with the norm $\|\cdot\|$.

**Definition 2.3.1** (Affine set)**.** *[33, Sec. 2.1.2] A set $C \in \mathbb{R}^n$ is affine if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ and $\theta \in \mathbb{R}$ we have $\theta\boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \in C$.*

**Definition 2.3.2** (Affine hull)**.** *[33, Sec. 2.1.2] The set of all affine combinations of points in a set $C \in \mathbb{R}^p$ is called the affine hull of $C$ (*$\mathrm{aff}(C)$*):*

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

**Remark 2.3.1.** [33, Sec. 2.1.4] $\mathrm{aff}(C)$ is the smallest affine set containing $C$.

**Definition 2.3.3** (Relative interior)**.** *[33, Sec. 2.1.3] Let $C \subset \mathbb{R}^p$. The relative interior of a set $C$ is defined as the interior of $C$ with respect to its affine hull* $\mathrm{aff}(C)$*:*

$$\mathrm{ri}(C) := \{\boldsymbol{x} \in C : \ B_{\|\cdot\|}(\boldsymbol{x}, r) \cap \mathrm{aff}(C) \subset C \text{ for some } r > 0\}.$$

**Definition 2.3.4** (Relative boundary)**.** *[33, Sec. 2.1.3]*

$$rb(C) := \overline{C} \setminus \mathrm{ri}C.$$

**Definition 2.3.5** (Convex set)**.** *[33, Sec. 2.1.4] A set $C \in \mathbb{R}^n$ is convex if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ and $\theta \in [0, 1]$ we have $\theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \in C$.*

**Definition 2.3.6** (Convex hull)**.** *[33, Sec. 2.1.4]*

$$\mathrm{conv}(C) := \{\theta_1 \boldsymbol{x}_1 + \ldots + \theta_k \boldsymbol{x}_k : \ \boldsymbol{x}_i \in C, \theta_i \geq 0, \quad i = 1, \ldots, k, \quad \theta_1 + \ldots + \theta_k = 1\}.$$

**Remark 2.3.2.** [33, Sec. 2.1.4] $\mathrm{conv}(C)$ is the smallest convex set containing $C$.

**Definition 2.3.7** (Convex cone)**.** *[33, Sec. 2.1.5] A set $C \in \mathbb{R}^p$ is a convex cone if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ and $\theta_1, \theta_2 \geq 0$ we have $\theta_1 \boldsymbol{x}_1 + \theta_2 \boldsymbol{x}_2 \in C$.*

**Definition 2.3.8** (Conic hull)**.** *[33, Sec. 2.1.5]*

$$\mathrm{cone}(C) := \{\theta_1 \boldsymbol{x}_1 + \ldots + \theta_k \boldsymbol{x}_k : \ \boldsymbol{x}_i \in C, \theta_i \geq 0, \quad i = 1, \ldots, k\}.$$

**Remark 2.3.3.** [33, Sec. 2.1.5] $\mathrm{cone}(C)$ is the smallest convex cone containing $C$.

**Definition 2.3.9** (Recession cone)**.** *[193, Definition 1.11. 2.1.5]] Let $C \in \mathbb{R}^p$ be a convex set. Then the recession cone of $C$ is defined as*

$$rec(C) := \{\boldsymbol{y} \in \mathbb{R}^p : \ \boldsymbol{x} + t\boldsymbol{y} \in C \ \text{ for all } \boldsymbol{x} \in C, t \geq 0\}.$$

**Remark 2.3.4.** [101, Prop. III. 1.2.1] A non-empty closed and convex set $C$ is compact if and only if $rec(C) = \{\mathbf{0}\}$.

**Definition 2.3.10** (Orthogonal projection)**.** *[33, Sec. 8.1] Let $C \subset \mathbb{R}^p$ be a closed set and let $\boldsymbol{x}_0 \in \mathbb{R}^p$. The orthogonal projection of $\boldsymbol{x}_0$ onto $C$ is defined as*

$$P_C(\boldsymbol{x}_0) := \underset{\boldsymbol{x} \in C}{\arg\min} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$$

**Definition 2.3.11** (Half-space)**.** *[34, p. 9] Let $\boldsymbol{y} \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$. Then*

$$K(\boldsymbol{y}, \alpha) := \{\boldsymbol{x} \in \mathbb{R}^p : \ \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \alpha\}.$$

**Definition 2.3.12** (Hyperplane)**.** *[34, p. 9] Let $\boldsymbol{y} \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$. Then*

$$H(\boldsymbol{y}, \alpha) := \{\boldsymbol{x} \in \mathbb{R}^p : \ \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \alpha\}.$$

**Definition 2.3.13.** *Let $A \subset \mathbb{R}^p$ be an affine space, i.e.*

$$\exists\ \boldsymbol{y}_1, \ldots, \boldsymbol{y}_l \in \mathbb{R}^p,\ \alpha_1, \ldots, \alpha_l \in \mathbb{R} \quad A = \bigcap_{i=1}^{l} H(\boldsymbol{y}_i, \alpha_i).$$

*Then by $\overrightarrow{A}$ we define the unique linear space parallel to $A$*

$$\overrightarrow{A} := \bigcap_{i=1}^{l} H(\boldsymbol{y}_i, 0).$$

**Definition 2.3.14** (Polar set)**.** *[34, p. 37] For any $M \subset \mathbb{R}^p$, the polar set $M^*$ is defined by*

$$M^* := \{\boldsymbol{y} \in \mathbb{R}^p :\ \forall\ \boldsymbol{x} \in M :\ \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq 1\}.$$

*Equivalently, $M^* = \bigcap\limits_{\boldsymbol{x} \in M} K(\boldsymbol{x}, 1)$.*

### 2.3.2 Polytopes

We recall basic definitions and facts about polytopes, which we will use throughout the proofs. The following can be found in textbooks, such as [94] and [193].
A set $P \subseteq \mathbb{R}^p$ is called a polytope if it is the convex hull of a finite set of points $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\} \subseteq \mathbb{R}^p$, that is,

$$P = \operatorname{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}.$$

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

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

for some $\boldsymbol{a} \in \mathbb{R}^p$ and $c \in \mathbb{R}$. Note that $F = \varnothing$ and $F = P$ are faces of $P$ and that any face $F$ is again a polytope. A non-empty face $F$ with $F \neq P$ is called proper. A point $\boldsymbol{x}_0 \in P$ lies in $\operatorname{ri}(P)$ if $\boldsymbol{x}_0$ is not contained in a proper face of $P$. We state two useful properties about faces in the following lemma.

**Lemma 2.3.1.** *Let $P \subseteq \mathbb{R}^p$ be a polytope given by $P = \operatorname{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$, where $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^p$. The following properties hold.*

*(a) If $F$ and $\widetilde{F}$ are faces of $P$, then so is $F \cap \widetilde{F}$.*

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

### 2.3.3 Pattern equivalence class

**Definition 2.3.15** (Pattern equivalence class)**.**

*(a) Let $\boldsymbol{x} \in \mathbb{R}^p$ and let* pen *be a polyhedral gauge. The pattern equivalence class $C_{\boldsymbol{x}}$ is the set of all vectors having the same subdifferential as $\boldsymbol{x}$:*

$$C_{\boldsymbol{x}} := \{\boldsymbol{w} \in \mathbb{R}^p :\ \partial\operatorname{pen}(\boldsymbol{w}) = \partial\operatorname{pen}(\boldsymbol{x})\}.$$

**Definition 2.3.16.** *Let $F \subset \mathbb{R}^p$. As $F^\perp$ we denote the orthogonal complement of $\overrightarrow{\text{aff}(F)}$.*

**Example 2.3.5.** For any $\boldsymbol{x} \in \mathbb{R}^p$, we have $\{\boldsymbol{x}\}^\perp = \{\boldsymbol{0}\}^\perp = \mathbb{R}^p$.

**Lemma 2.3.2.** *Let $\boldsymbol{x}_0 \in F$. Then*

$$\text{lin}(F - \boldsymbol{x}_0) = \overrightarrow{\text{aff}(F)}.$$

*Proof.* By [101, Sec. III.5.3.] we have

$$\text{lin}(\mathbb{R}_+(F - \boldsymbol{x}_0)) = \text{aff}(F) - \boldsymbol{x}_0.$$

The statement is proved after observing that $\text{lin}(F - \boldsymbol{x}_0) = \text{lin}(\mathbb{R}_+(F - \boldsymbol{x}_0))$ and $\text{aff}(F) - \boldsymbol{x}_0 = \overrightarrow{\text{aff}(F)}$. □

### 2.3.4 Normal cones

**Definition 2.3.17.** *[31, p. 15], [101, p.136] The normal cone to a convex set $C$ at a point $\overline{\boldsymbol{x}} \in C$, written $N_C(\overline{\boldsymbol{x}})$ is the convex cone of normal vectors, i.e. vectors $\boldsymbol{d}$ in $\mathbb{R}^p$ such that $\langle \boldsymbol{d}, \boldsymbol{x} - \overline{\boldsymbol{x}} \rangle \leq 0$ for all points $\boldsymbol{x}$ in $C$.*

**Definition 2.3.18.** *[70, Def. 4.12.] If $F$ is a face of a closed convex set $K$ and $\boldsymbol{x} \in \text{ri}F$, then, $N_K(\boldsymbol{x})$ does not depend on $\boldsymbol{x} \in \text{ri } F$ and is denoted by $N_K(F)$ and is called the cone of normals of $K$ in $F$.*

**Lemma 2.3.3.** *[70, Lemma 3.1.] Let $K$ be a closed convex set in $\mathbb{R}^n$. To each $\boldsymbol{x} \in \mathbb{R}^n$ there exists a unique $\boldsymbol{x}' \in K$ such that*

$$\|\boldsymbol{x} - \boldsymbol{x}'\| = \inf_{\boldsymbol{y} \in K} \|\boldsymbol{x} - \boldsymbol{y}\|.$$

**Definition 2.3.19.** *[70, Def. 3.2.] The map*

$$\begin{aligned} p_K \ : \mathbb{R}^n &\longrightarrow K \\ \boldsymbol{x} &\mapsto p_K(\boldsymbol{x}) = \boldsymbol{x}' \end{aligned}$$

*of Lemma 2.3.3 is called the nearest point map relative to $K$.*

**Definition 2.3.20.** *[70, Def. 4.7.] Let $\boldsymbol{x}$ be a point of the closed convex set $K$. We call*

$$N_K(\boldsymbol{x}) := -\boldsymbol{x} + p_K^{-1}(\boldsymbol{x})$$

*the normal cone of $K$ at $\boldsymbol{x}$.*

**Proposition 2.3.1.** *The Definitions 2.3.17 and 2.3.20 are equivalent.*

*Proof.* By [101, Proposition III. 5.3.3.], the point $\boldsymbol{s} \in \mathbb{R}^p$ belongs to $N_K(\boldsymbol{x})$ from Definition 2.3.17 if and only if $\boldsymbol{x} = p_K(\boldsymbol{x} + \boldsymbol{s})$. It implies that $p_K^{-1}(\boldsymbol{x}) = \boldsymbol{x} + N_K(\boldsymbol{x})$, thus

$$N_K(\boldsymbol{x}) = -\boldsymbol{x} + p_K^{-1}(\boldsymbol{x}).$$

□

**Corollary 2.3.1.** *The normal cone $N_K(\boldsymbol{x})$ is invariant under translations of the set $K$.*

**Theorem 2.3.6.** *[70, Theorem 4.13.] Let $K$ be a compact convex set in $\mathbb{R}^p$ and $\boldsymbol{x}$ a relative interior point of a face $F \neq \varnothing$ of $K$. Then*

$$\{\text{ri } N_K(\boldsymbol{x}) : \ F \text{ is a face of } K\} = \{\text{ri } N_K(F) : \ F \text{ is a face of } K\}$$

*is a partition (disjoint covering) of $\mathbb{R}^p$.*

**Definition 2.3.21** (Conjugate face)**.** *[34, p. 40] Let $F$ be a face of a polytope $P$. The conjugate face $F^o$ of a dual polytope $P^*$ is defined as follows*

$$F^o := \{\boldsymbol{y} \in P^* : \ \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 1 \quad \forall \ \boldsymbol{x} \in F\}.$$

Equivalently, we get

$$F^o = \bigcap_{\boldsymbol{x} \in P} K(\boldsymbol{x}, 1) \cap \bigcap_{\boldsymbol{x}' \in F} H(\boldsymbol{x}', 1),$$

where the first and second intersection are equal respectively to $P^*$ and $\text{aff}(F^o)$.
Consider a convex polytope $P = \{\boldsymbol{v} \in \mathbb{R}^p : \ \langle \boldsymbol{s}_j, \boldsymbol{v} \rangle \leq r_j, \ j = 1, \ldots, m\}$, cf. [101, p. 138]. We define its active set for $\boldsymbol{x} \in \mathbb{R}^p$ as $A_P(\boldsymbol{x}) := \{j = 1, \ldots, m : \ \langle \boldsymbol{s}_j, v \rangle = r_j\}$

**Proposition 2.3.2.** *[94, Proposition 14.1] We have*

$$N_P(\boldsymbol{x}) = \text{cone}(\{\boldsymbol{s}_j : \ j \in A_P(\boldsymbol{x})\}).$$

**Theorem 2.3.7.** *Let $B^*$ be the polar set of the polytope $B$. Let $\boldsymbol{x} \in \text{ri}(F)$, where $F$ is a face of $B^*$. Then the normal cone at $\boldsymbol{x}$ to $B^*$ is given by $N_{B^*}(\boldsymbol{x}) = \mathbb{R}_+ F^o$, where $F^o \subset B$ is the conjugate face to $F$.*

*Proof.* ($\mathbb{R}_+ F^o \subset N_{B^*}(\boldsymbol{x})$):
Let $\boldsymbol{y} \in \mathbb{R}_+ F^o$. Then there exists such $\gamma \geq 0$, that $\boldsymbol{y} \in \gamma B$ and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \gamma$ for every $\boldsymbol{x} \in F$. It implies that

$$\langle \boldsymbol{v}, \boldsymbol{y} \rangle \leq \gamma \ \forall \ \boldsymbol{v} \in B^* \text{and} \qquad\qquad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \gamma \ \forall \ \boldsymbol{x} \in F.$$

Therefore $\langle \boldsymbol{v}, \boldsymbol{y} \rangle \leq \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all $\boldsymbol{v} \in B^*$, i.e. $\langle \boldsymbol{v} - \boldsymbol{x}, \boldsymbol{y} \rangle \leq 0$ for all $\boldsymbol{v} \in B^*$, which means that $\boldsymbol{y} \in N_{B^*}(\boldsymbol{x})$.
($N_{B^*}(\boldsymbol{x}) \subset \mathbb{R}_+ F^o$):
Let $\boldsymbol{y} \in N_{B^*}(\boldsymbol{x})$, i.e. $\langle \boldsymbol{y}, \boldsymbol{v} - \boldsymbol{x} \rangle \leq 0$ for every $\boldsymbol{v} \in B^*$. Thus the maximum of the value of $\langle \boldsymbol{y}, \boldsymbol{v} - \boldsymbol{x} \rangle$ over $\boldsymbol{v} \in B^*$ is attained at $\boldsymbol{v} = \boldsymbol{x}$. Moreover, as $\langle \boldsymbol{y}, \boldsymbol{v} \rangle \leq \langle \boldsymbol{y}, \boldsymbol{x} \rangle$ for every $\boldsymbol{v} \in B^*$, we get $\boldsymbol{y} \in \langle \boldsymbol{y}, \boldsymbol{x} \rangle B$. Now it suffices to prove that for every $\boldsymbol{v} \in F$ we have $\langle \boldsymbol{y}, \boldsymbol{v} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$.
Take any $\boldsymbol{v} \in F$. Since $\boldsymbol{x} \in \text{ri}(F) \subset F$, we have $N_{B^*}(\boldsymbol{x}) \subset N_{B^*}(\boldsymbol{v})$. Then $\langle \boldsymbol{y}, \boldsymbol{w} - \boldsymbol{x} \rangle \leq 0$ for every $\boldsymbol{w} \in B^*$, which implies that $\langle \boldsymbol{y}, \boldsymbol{v} - \boldsymbol{x} \rangle \leq 0$, i.e. $\langle \boldsymbol{y}, \boldsymbol{v} \rangle \leq \langle \boldsymbol{y}, \boldsymbol{x} \rangle$.
On the other hand, $\boldsymbol{y} \in N_{B^*}(\boldsymbol{v})$. Therefore $\langle \boldsymbol{y}, \boldsymbol{w} - \boldsymbol{v} \rangle \leq 0$ for every $\boldsymbol{w} \in B^*$, which implies that $\langle \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{v} \rangle \leq 0$, i.e. $\langle \boldsymbol{y}, \boldsymbol{x} \rangle \leq \langle \boldsymbol{y}, \boldsymbol{v} \rangle$.
Thus for every $\boldsymbol{v} \in F$ we have $\langle \boldsymbol{y}, \boldsymbol{v} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$, hence $\boldsymbol{y} \in \langle \boldsymbol{y}, \boldsymbol{x} \rangle F^o \subset \mathbb{R}_+ F^o$. □

### 2.3.5 Polyhedral gauges

**Definition 2.3.22** (Gauge)**.** *[155, p. 53] Let $K \subset \mathbb{R}^p$ be a closed convex set containing $\mathbf{0}$. The gauge function of $K$ is defined as*

$$\inf\{\lambda \geq 0 : \ \boldsymbol{x} \in \lambda K\}$$

The gauge function of $K$ is also known as the Minkowski functional of $K$. When $K$ is a polyhedron, its gauge function is known as a polyhedral gauge pen and can be written as the maximum of linear functions [152, 133]:

$$\forall \boldsymbol{x} \in \mathbb{R}^p, \text{pen}(x) = \max\{0, \boldsymbol{u}_1'\boldsymbol{x}, \ldots, \boldsymbol{u}_l'\boldsymbol{x}\}, \text{ for some } \boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p.$$

Note that a polyhedral gauge, whose unit ball $\{\boldsymbol{x} \in \mathbb{R}^p : \text{pen}(\boldsymbol{x}) \leq 1\}$ is a bounded and symmetric with respect to the origin polyhedron, is a polyhedral norm.

### 2.3.6 Thresholded penalized least squares estimation

**Definition 2.3.23** (Thresholded penalized least squares estimator)**.**
*Let* pen *be a polyhedral gauge,* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$ *and* $\lambda > 0$. *Given* $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{Y})$, *we say that* $\widetilde{\boldsymbol{\beta}}$ *is a thresholded estimator of* $\hat{\boldsymbol{\beta}}$ *if* $\partial\text{pen}(\hat{\boldsymbol{\beta}}) \subseteq \partial\text{pen}(\widetilde{\boldsymbol{\beta}})$.

**Example 2.3.8.** In SLOPE optimization problem, a vector $\widetilde{\boldsymbol{\beta}} = (5, 5, 5, 0, 0)'$ is a thresholded estimator of $\hat{\boldsymbol{\beta}}^{SLOPE} = (4, 4, 2, 0, 0)'$.

### 2.3.7 Permutahedron

**Definition 2.3.24.** *Let* $\boldsymbol{\Lambda} \in \mathbb{R}^p$. *A permutahedron* $P_{\boldsymbol{\Lambda}}$ *is a convex hull of all possible permutations of* $\boldsymbol{\Lambda}$*:*

$$P_{\boldsymbol{\Lambda}} := Conv((\lambda_{\pi(1)}, \ldots, \lambda_{\pi(p)})' : \ \pi \in \mathcal{S}_p). \tag{2.3.1}$$

**Definition 2.3.25** (Signed permutahedron)**.** *Let* $\boldsymbol{\Lambda} \in \mathbb{R}^p$. *A signed permutahedron* $C_{\boldsymbol{\Lambda}}$ *is a convex hull of all possible permutations and sign changes of* $\boldsymbol{\Lambda}$*:*

$$C_{\boldsymbol{\Lambda}} = \{\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_p) \in \mathbb{R}^p : \ \sum_{j\leq i} |\pi|_{(j)} \leq \sum_{j\leq i} \lambda_j : i = 1, 2, \ldots, p\}. \tag{2.3.2}$$

$C_{\boldsymbol{\Lambda}}$ is a unit ball in a norm $J^*_{\boldsymbol{\Lambda}}$ dual to the SLOPE norm $J_{\boldsymbol{\Lambda}}$.

### 2.3.8 Basics on Moore-Penrose inverse

The notion of the Moore-Penrose inverse is crucial in the SLOPE irrepresentability condition and is outlined below, see [91, 16].
If $\boldsymbol{A}$ is an $n \times p$ real matrix then a $p \times n$ matrix $\boldsymbol{A}^+$ is called a Moore-Penrose inverse of $\boldsymbol{A}$ if

$$\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}, \qquad \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$$

and if the matrices $\boldsymbol{A}\boldsymbol{A}^+$ and $\boldsymbol{A}^+\boldsymbol{A}$ are symmetric.
There always exists a unique real Moore-Penrose inverse $\boldsymbol{A}^+$ of a real matrix $\boldsymbol{A}$. In some cases it may be computed quickly to a more convenient form:

- If $\boldsymbol{A}'\boldsymbol{A}$ is an invertible matrix, then $\boldsymbol{A}^+ = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$,
- If $\boldsymbol{A}\boldsymbol{A}'$ is an invertible matrix, then $\boldsymbol{A}^+ = \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^{-1}$,
- $\boldsymbol{A}^+ = \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^+$,
- $\boldsymbol{0}^+ = \boldsymbol{0}$.

In general, the Moore-Penrose inverse $\boldsymbol{A}^+$ is computed using the formula $\boldsymbol{A}^+ = \boldsymbol{A}'(\boldsymbol{A}\boldsymbol{A}')^+$ and the Singular Value Decomposition, which allows to do it numerically.

**Remark 2.3.9.**

- $\boldsymbol{A}\boldsymbol{A}^+$ is an orthogonal projector onto $\mathrm{col}(\boldsymbol{A})$,
- $\mathrm{col}(\boldsymbol{A}^+) = \mathrm{col}(\boldsymbol{A}')$.

### 2.3.9 Functional analysis

**Definition 2.3.26** (Dual norm)**.** *[33, Appendix A.1.6] Let $\|\cdot\|$ be a norm on $\mathbb{R}^p$. The dual norm, denoted $\|\cdot\|_*$, is defined as*

$$\|\boldsymbol{x}\|_* = \sup\{\boldsymbol{x}'\boldsymbol{b} : \ \|\boldsymbol{b}\| \leq 1\}.$$

In particular, the dual sorted $\ell_1$ norm $J_{\boldsymbol{\Lambda}}$ has an explicit expression given in [136]:

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

### 2.3.10 Tools from optimization

**Definition 2.3.27** (Quasi-convexity)**.** *[33, Sec. 3.4.1] A function $f : \mathbb{R}^p \to \mathbb{R}$ is called quasiconvex if its domain and all its sublevel sets*

$$S_\alpha = \{\boldsymbol{x} : \ f(\boldsymbol{x}) \leq \alpha\},$$

*for $\alpha \in \mathbb{R}$, are convex. If a function $-f$ is quasiconvex, then we say that $f$ is quasiconcave.*

**Definition 2.3.28** (Semi-continuity)**.** *[101, Definition 3.2.1] We say that the convex function $f$ is lower semi-continuous, if*

$$\liminf_{\boldsymbol{x}\to\boldsymbol{x}_0} f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$$

*for all $\boldsymbol{x}_0 \in \mathbb{R}^p$. If a function $-f$ is lower semi-continuous, then we say that $f$ is upper semi-continuous.*

**Theorem 2.3.10** (max-min inequality)**.** *[33, Sec. 5.4.1]*

$$\sup_{\boldsymbol{b}\in B} \inf_{\boldsymbol{a}\in A} f(\boldsymbol{a},\boldsymbol{b}) \leq \inf_{\boldsymbol{a}\in A} \sup_{\boldsymbol{b}\in B} f(\boldsymbol{a},\boldsymbol{b})$$

**Definition 2.3.29** (Saddle point)**.** *[33, Sec. 5.4.2] The pair $(\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}) \in A \times B$ is a saddle point for a function $f$ (and $A$ and $B$) if*

$$f(\overline{\boldsymbol{a}}, \boldsymbol{b}) \leq f(\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}) \leq f(\boldsymbol{a}, \overline{\boldsymbol{b}})$$

*for all $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$. In other words, we have*

$$f(\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}) = \inf_{\boldsymbol{a}\in A} f(\boldsymbol{a}, \overline{\boldsymbol{b}}) = \sup_{\boldsymbol{b}\in B} f(\overline{\boldsymbol{a}}, \boldsymbol{b}).$$

The existence of the saddle point implies that it attains the equality in the max-min inequality:

$$\sup_{\boldsymbol{b}\in B} \inf_{\boldsymbol{a}\in A} f(\boldsymbol{a},\boldsymbol{b}) = \inf_{\boldsymbol{a}\in A} \sup_{\boldsymbol{b}\in B} f(\boldsymbol{a},\boldsymbol{b}) = f(\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}).$$

**Theorem 2.3.11** (Sion)**.** *[7, Ch. 7, Theorem 7.] Suppose that*

*(a) $X$ and $Y$ are convex compact subsets,*

*(b) for all $\boldsymbol{y} \in Y$ the function $\boldsymbol{x} \mapsto f(\boldsymbol{x}, \boldsymbol{y})$ is quasiconvex and lower semi-continuous,*

*(c) for all $\boldsymbol{x} \in X$ the function $\boldsymbol{y} \mapsto f(\boldsymbol{x}, \boldsymbol{y})$ is quasiconcave and upper semi-continuous,*

*Then there exists a saddle point $\{\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}\}$.*

**Theorem 2.3.12** (Hardy-Littlewood-Pólya rearrangement inequality)**.** *[98, Theorem 368] For every $x_1 \leq x_2 \leq \ldots \leq x_k \in \mathbb{R}$, $y_1 \leq y_2 \leq \ldots \leq y_k \in \mathbb{R}$ and for every permutation $\sigma \in \mathcal{S}_k$ we have*

$$x_k y_1 + x_{k-1} y_2 + \ldots + x_1 y_k \leq x_{\sigma(1)} y_1 + x_{\sigma(2)} y_2 \ldots x_{\sigma(k)} y_k \leq x_1 y_1 + x_2 y_2 + \ldots + x_k y_k.$$

### 2.3.11 Tools from probability

**Definition 2.3.30** (Convergence in distribution)**.** *[19, p. 329] Let $X_n$ and $X$ be random variables with respective distribution functions $F_n$ and $F$. If $F_n \to F$, then $X_n$ is said to converge in distribution or in law to $X$, written $X_n \xrightarrow{d} X$.*

**Definition 2.3.31** (Convergence in probability)**.** *[19, p. 330]*

$$\lim_n \mathbb{P}[|X_n - X| > \varepsilon] = 0, \qquad \text{for every } \varepsilon > 0.$$

**Definition 2.3.32** (Consistency)**.** *$\hat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}$, if $\boldsymbol{\beta}_n$ converges to $\boldsymbol{\beta}$ in probability.*

**Definition 2.3.33** (Almost sure convergence)**.** *[19, pp. 59-60]*

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) \to X(\omega)\right) = 1.$$

**Definition 2.3.34** (Strong consistency)**.** *$\hat{\boldsymbol{\beta}}_n$ is a strong consistent estimator of $\boldsymbol{\beta}$, if $\boldsymbol{\beta}_n$ converges to $\boldsymbol{\beta}$ almost surely.*

**Theorem 2.3.13** (Borel-Cantelli Lemma)**.** *[19, Theorem 4.3. and 4.4.]*

*(a) If $\sum_n \mathbb{P}(A_n)$ converges, then $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 0$.*

*(b) If $\{A_n\}$ is an independent sequence of events and $\sum_n \mathbb{P}(A_n)$ diverges, then $\mathbb{P}\left(\limsup_{n\to\infty} A_n\right) = 1$.*

**Theorem 2.3.14** (Strong law of large numbers)**.** *[73, VII.8, Theorem 1] Let $X_1, X_2 \ldots$ be i.i.d. random variables with $\mathbb{E}(X) = 0$. For each $n \geq 1$ denote $S_n := X_1 + \ldots X_n$. Then $n^{-1} S_n \to 0$ with probability 1.*

**Theorem 2.3.15** (Lindeberg-Feller Multivariate Central Limit Theorem)**.** *[147, Sec. 4.3.2] Let $\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots, \boldsymbol{X}_n$ be independent random variables with multivariate distribution having mean $\mu$ and covariance matrix $\boldsymbol{\Sigma}$. Then the distribution of $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i - \mu\right)$ tends to $N(\boldsymbol{0}, \boldsymbol{\Sigma})$.*

**Theorem 2.3.16** (CLT for linearly negative quadrant dependent variables)**.** *[113, Corollary 1.1] Let $\{\xi_i\}$ be a centered sequence of linearly negative quadrant dependent random variables*

*such that* $\{\xi_i^2\}$ *is a uniformly integrable family, and let* $\{a_{ni}, 1 \le i \le n\}$ *be a triangular array of nonnegative numbers such that*

$$\sup_n \sum_{i=1}^n \frac{a_{ni}^2}{\sigma_n^2} < \infty$$

*and*

$$\max_{1\le i\le n} \frac{a_{ni}}{\sigma_n} \to 0 \qquad \text{as} \qquad n \to \infty, \tag{2.3.3}$$

*where* $\sigma^2 = Var\left(\sum_{i=1}^n a_{ni}\xi_i\right)$. *If*

$$\sum_{j:|i-j|\ge u} Cov(\xi_i, \xi_j)^- \to 0 \qquad \text{as} \qquad u \to \infty \qquad \text{uniformly in} \qquad i \ge 1,$$

*then, as* $n \to \infty$,

$$\frac{1}{\sigma_n}\sum_{i=1}^n a_{ni}\xi_i \xrightarrow{d} N(0,1).$$

**Theorem 2.3.17** (Tail inequality)**.** *If* $X \sim \mathcal{N}(0,1)$, *then*

$$\mathbb{P}(X > t) \le \frac{1}{t} e^{-t^2/2} \frac{1}{\sqrt{2\pi}}.$$

*Proof.* [43] Note that for $x \ge t > 0$ we have $\frac{x}{t} \ge 1$. Therefore

$$\mathbb{P}(X \ge t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty 1 \cdot e^{-x^2/2} dx \le \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} \cdot e^{-x^2/2} dx = \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}.$$

□

**Theorem 2.3.18** (Law of Iterated Logarithm for triangular array)**.** *[117, Theorem 1 (i)] Let* $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ *be independent random variables such that*

$$\mathbb{E}\varepsilon_n = 0 \qquad \text{and} \qquad \mathbb{E}\varepsilon_n^2 = \sigma^2 \quad \text{for all n} \quad \text{and} \sup_n \mathbb{E}\varepsilon_n^2 < \infty \tag{2.3.4}$$

*for some* $r > 2$. *Let* $\{a_{ni}\}$ *be a double array of constants satisfying* $\sum_{-\infty}^{\infty} a_{ni}^2 < \infty$ *for every* $n$. *Define*

$$S_n = \sum_{-\infty}^{\infty} a_{ni}\varepsilon_n. \tag{2.3.5}$$

*Assume that as* $n \to \infty$,

$$A_n = \sum_{-\infty}^{\infty} a_{ni}^2 \to \infty,$$

*and*

$$\sup_i a_{ni}^2 = o(A_n(\log A_n))^{-\rho} \qquad \text{for all } \rho > 0. \tag{2.3.6}$$

*If there exist constants* $c_i \ge 0$ *and* $d > 2/r$ *such that*

$$\|a_n - a_m\|^2 \le \left(\sum_{i=m+1}^n c_i\right)^d \qquad \text{for} \qquad n > m > m_0 \tag{2.3.7}$$

*and*

$$\left(\sum_{i=m_0}^{n} c_i\right)^d = O(A_n) \quad a.s. \quad n \to \infty, \tag{2.3.8}$$

*then*

$$\limsup_{n\to\infty} \frac{|S_n|}{(2A_n \log\log A_n)^{1/2}} \leq \sigma \quad a.s.$$

# Chapter 3

# Pattern recovery and signal denoising by SLOPE when the design matrix is orthogonal

## 3.1 Introduction

### 3.1.1 Introduction and motivations

The content of this chapter may be found in the recently published article of the author of the dissertation, Graczyk, Kołodziejek and Wilczyński [165]. To start our discussion on the pattern recovery by penalized linear regression methods, we start with the SLOPE estimator and the simplest case of the design matrix $\boldsymbol{X}$ being orthogonal. For that reason let us recall the Linear Multiple Regression. It 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, that is, when the number of predictors $p$ is not greater than the number of explanatory variables $n$ and $\boldsymbol{X}$ is of full rank, the ordinary least squares estimator $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ has an exact formula $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. 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 descriptible by a smaller number of parameters. Several solutions were proposed to deal with such problem. One of them, the Least Absolute Shrinkage and Selection Operator (LASSO [47, 176]) involves penalizing the residual sum of squares $\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$ with an $\ell_1$ norm of $\hat{\boldsymbol{\beta}}$ multiplied by a tuning parameter $\lambda$:

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

We turn our focus on one of the extensions of LASSO, which is the Sorted $\ell_1$ Penalized Estimator (SLOPE [27, 26, 189]). In addition, SLOPE allows one to clusterize the similar coefficients of $\boldsymbol{\beta}$. In SLOPE, the $\ell_1$ norm is replaced by its sorted version $J_{\boldsymbol{\Lambda}}$, which depends on the tuning vector $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$, where $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$:

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

where $\{|\boldsymbol{\beta}|_{(i)}\}_{i=1}^p$ is a decreasing permutation of absolute values of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$:

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

The case of $\mathbf{\Lambda}$ being an arithmetic sequence was studied by Bondell and Reich [29] and called the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR). The special case of SLOPE with $\lambda_1 = \lambda_2 = \ldots = \lambda_p > 0$ is LASSO. For $\mathbf{\Lambda} = (0, \ldots, 0)$ we obtain the OLS estimator.
Clustering the predictors allows for additional dimension reduction by identifying variables with the same absolute values of the regression coefficients. Recently, interest has increased in methods that cluster highly correlated predictors [30, 85, 128, 137, 139, 168]. SLOPE is ideal for this task, since it is capable to identify the low-dimensional structure, which is called the SLOPE pattern, defined by Schneider and Tardivel with the subdifferential of the SLOPE norm $J_{\mathbf{\Lambda}}$, see Theorem 2.2.4.
The clustering properties of SLOPE have been studied before, cf. [29, 76], but the researchers consider strongly correlated predictors, which are used in financial mathematics to group the assets with respect to their partial correlation with hedge fund return times series [116]. We decided to consider the pattern recovery in more general case. We start with supposing the orthogonal design

$$\boldsymbol{X}'\boldsymbol{X} = n\boldsymbol{I}_p. \tag{3.1.1}$$

This is a classical and natural assumption in the case of experimental data; cf. [176]. Moreover, in the asymptotic case, where $n \to \infty$ and $p$ is fixed, it is usually supposed that $\boldsymbol{X}'\boldsymbol{X}/n \to \boldsymbol{C} > 0$, cf. [192, 194]. In (3.1.1) 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. This class of matrices is being widely used in signal analysis, [146, 48]. For general $\boldsymbol{X}$ the problem is considered in Chapter 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, for example, by Zeng and Figueiredo [189]. This dual ball is described explicitely as a signed permutahedron, see, e.g. [136, 156]:

$$C_{\mathbf{\Lambda}} = \left\{ \boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_p) \in \mathbb{R}^p : \sum_{j \le i} |\pi|_{(j)} \le \sum_{j \le i} \lambda_j : i = 1, 2, \ldots, p \right\}. \tag{3.1.2}$$

Here 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 $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$.

### 3.1.2 Outline

In Section 3.2 we derive the connections between $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ in the orthogonal design. We use the minimax theorem of Sion, cf. [7]. In Section 3.3 we focus on the properties of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$. We use the geometric interpretation of SLOPE to explain its ability to identify the SLOPE pattern and provide new theoretical results on support recovery and clustering properties using a representation of SLOPE as a function of the ordinary least squares (OLS) estimator. A similar approach for LASSO was used by Ewald and Schneider, cf. [71].
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 3.4 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 $\boldsymbol{\varepsilon}^{(n)} = (\varepsilon_1^{(n)}, \varepsilon_2^{(n)}, \ldots, \varepsilon_n^{(n)})' \in \mathbb{R}^n$ varies with $n$ and has the normal distribution $N(0, \sigma^2 \boldsymbol{I}_n)$. We make no assumptions about the relations between $\boldsymbol{\varepsilon}^{(n)}$ and $\boldsymbol{\varepsilon}^{(m)}$ for $n \neq m$.

In this chapter we consider the specific, but statistically important, model in which $n \geq p$ and the columns of $\boldsymbol{X}$ are 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 columns of $\boldsymbol{X}$ are not orthogonal. We provide the conditions under which the SLOPE estimator is strongly consistent. Additionally, in case when for each $n$ the design matrix is orthogonal, we provide the conditions on the sequence of tuning parameters such that SLOPE is strongly consistent in the pattern recovery. In Section 3.5 we show the applications of the SLOPE clustering in terms of high frequency signal denoising and illustrate them with simulations. The Appendix covers the proofs of technical results.
Substantially more difficult techniques based on subdifferential calculus are developed in [23]. These techniques are used in [23] to establish the properties of the SLOPE estimator in the general case, where the columns of $\boldsymbol{X}$ are not orthogonal and $p$ may be much larger than $n$. However, the asymptotic results of [23] are derived under stronger assumptions than those of this paper. In [23] the sequence of error terms $\boldsymbol{\varepsilon}^{(n)}$ is incremental and the sequence of tuning parameters has the form $\lambda_n = \alpha_n \Lambda$, where $\alpha_n$ is a given sequence of positive numbers and $\Lambda \in \mathbb{R}^p$ is fixed. In this paper we make no assumptions about the relations between $\boldsymbol{\varepsilon}^{(n)}$ and $\boldsymbol{\varepsilon}^{(m)}$ for $n \neq m$, and the sequence of tuning parameters has a general form.

## 3.2 Approach by minimax theorem

### 3.2.1 Technical results

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a real-valued matrix. Let $r_{\text{SLOPE}}$ denote the minimum value of the SLOPE criterion, attained by $\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$, i.e.

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

Since

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

it follows that

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

We immediately get the following result.

**Corollary 3.2.1.** $\left\| \widehat{\boldsymbol{\beta}}^{\text{SLOPE}} \right\|_2^2 \leq M_0$, *where* $M_0 = \left( \frac{p \|\boldsymbol{Y}\|_2^4}{4\lambda_1^2} \right)$.

From this corollary it is seen that we can clearly limit our search to vectors $\boldsymbol{\beta}$ from the compact set $\mathcal{M} \subset \mathbb{R}^p$ defined by $\mathcal{M} := \{\boldsymbol{b} \in \mathbb{R}^p : \ \|\boldsymbol{b}\|_2^2 \leq M_0\}$. Therefore, we can equivalently define a SLOPE solution by

$$\widehat{\boldsymbol{\beta}}^{\text{SLOPE}} = \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathcal{M}} \left[ \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) \right]. \tag{3.2.1}$$

**Proposition 3.2.1.** *Let* $C_{\boldsymbol{\Lambda}}$ *be the unit ball in the dual SLOPE norm. Then, for each* $\boldsymbol{b} \in \mathbb{R}^p$,

$$J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \max_{\pi \in C_{\boldsymbol{\Lambda}}} \boldsymbol{\pi}' \boldsymbol{b}. \tag{3.2.2}$$

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

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

**Remark 3.2.1.**

(a) A different, longer proof is given in [27, Proposition 1.1]

(b) The formula (3.2.2) holds in much greater generality for Lovász extensions in place of the $J_{\boldsymbol{\Lambda}}$ norm, see [132].

### 3.2.2 Saddle point

In this section we continue the assumption that $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a real-valued matrix. Let the function $r : \mathcal{M} \times C_{\boldsymbol{\Lambda}} \to \mathbb{R}$ be defined by

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

As an immediate consequence of (3.2.1) and Proposition 3.2.1 we obtain

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

It turns out that the order of maximization over $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$ and minimization over $\boldsymbol{b} \in \mathcal{M}$ can be switched without affecting the result. To see this, note that both $C_{\boldsymbol{\Lambda}}$ and $\mathcal{M}$ are convex and compact. Furthermore, for each fixed $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$, $r(\boldsymbol{b}, \boldsymbol{\pi})$ is a convex continuous function with respect to $\boldsymbol{b} \in \mathcal{M}$ and, for each fixed $\boldsymbol{b} \in \mathcal{M}$, $r(\boldsymbol{b}, \boldsymbol{\pi})$ is concave with respect to $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$ (in fact, it is linear). Therefore, all assumptions of the Sion's minimax theorem are fulfilled (see [7, p. 218]) and thus there exists a saddle point $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*) \in \mathcal{M} \times C_{\boldsymbol{\Lambda}}$ such that

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

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

### 3.2.3 SLOPE solution when X has full column rank

Since for each fixed $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$, 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 $\dfrac{\partial r(\boldsymbol{b}, \boldsymbol{\pi})}{\partial \boldsymbol{b}}$ is zero, is a global minimum. If we rewrite $r(\boldsymbol{b}, \boldsymbol{\pi})$ as

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

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 with $\mathbf{0}$ gives the following equation for the optimum point $\boldsymbol{b}_{\boldsymbol{\pi}}$:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{\boldsymbol{\pi}} = \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{\pi}. \tag{3.2.3}$$

Substituting this into the equation for $r(\boldsymbol{b}_{\boldsymbol{\pi}}, \boldsymbol{\pi})$ and using the fact that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, we find that

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

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

**Lemma 3.2.1.** *Assume that $\boldsymbol{X}$ has full column rank. Let $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_p^*)' \in C_{\boldsymbol{\Lambda}}$ be any solution of*

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

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

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

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

*(a)* $\operatorname{sign}(\beta_i^*) \cdot \operatorname{sign}(\pi_i^*) \ge 0$, $i = 1, 2, \ldots, p$,

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

*(c) for any permutation $\tau$ satisfying $|\beta^*_{\tau(1)}| \ge \ldots \ge |\beta^*_{\tau(p)}|$, if there is a $k \in \{2, \ldots, p\}$ such that $\sum_{i=1}^{k-1} \left|\pi^*_{\tau(i)}\right| < \sum_{i=1}^{k-1} \lambda_i$ and $\left|\pi^*_{\tau(k)}\right| > 0$,*
*then $\left|\beta^*_{\tau(k-1)}\right| = \left|\beta^*_{\tau(k)}\right|$.*

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

**Lemma 3.2.2.** *Assume that $\boldsymbol{X}$ has full column rank. The point $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)$ defined as in Lemma 3.2.1 is the 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 3.2.2.** *Assume that $\boldsymbol{X}$ has full column rank. Let the point $\boldsymbol{\beta}^*$ be defined as in Lemma 3.2.1. Then $\boldsymbol{\beta}^*$ is the SLOPE estimator of $\boldsymbol{\beta}$.*

*Proof.* Using the fact that $\max_{\pi \in C_{\mathbf{\Lambda}}} r(\boldsymbol{\beta}^*, \boldsymbol{\pi}) = \min_{\boldsymbol{b} \in \mathcal{M}} \max_{\pi \in C_{\mathbf{\Lambda}}} r(\boldsymbol{b}, \boldsymbol{\pi})$ (see previous lemma) we have

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

□

**Corollary 3.2.2.** *In the linear model satisfying $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ we have*

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}} = \frac{1}{n}\boldsymbol{\pi}^* = \frac{1}{n}\operatorname*{arg\,min}_{\pi \in C_{\mathbf{\Lambda}}} \left\|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\pi}\right\|_2^2 = \operatorname*{arg\,min}_{\pi \in C_{\mathbf{\Lambda}/n}} \left\|\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \boldsymbol{\pi}\right\|_2^2,$$

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

Projections onto $C_{\mathbf{\Lambda}}$ are widely used in [132] in the study of the notion of degrees of freedom. However, the Corollary 3.2.2 is not stated there explicitely.

**Remark 3.2.3.** Assume that $\boldsymbol{X}$ has full column rank. For each $\boldsymbol{\pi} \in C_{\mathbf{\Lambda}}$, the point $\boldsymbol{b}_\pi$ defined in (3.2.3) belongs to

$$\left\{\boldsymbol{b} \in \mathbb{R}^p : \ \|\boldsymbol{b}\|_2^2 \leq M\right\},$$

where $M$ is chosen so that $M > \max\{M_0, M_1\}$ with

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

## 3.3 Properties of SLOPE in the orthogonal design

### 3.3.1 SLOPE vs. OLS

By Theorem 3.2.2 and Corollary 3.2.2, when $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, the orthogonal projection of the ordinary least squares estimator $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{Y}$ onto the unit ball $C_{\mathbf{\Lambda}/n}$ is equal to $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} - \hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$. For $\mathbf{\Lambda} = (200, 100)'$ and $n = 50$ this property is illustrated in Figure 3.1. The figure represents $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ (black arrows) depending on the localization of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ in the orthogonal design. For $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ being the blue point located in the area labeled by $(1, 0)$ the first component of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ is positive and the second is null. For $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ being the yellow point located on the area labeled by $(-1, 1)$ both components of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ have equal absolute value (clusterization), but their signs are opposite. For $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ being the red point located on the area labeled by $(1, 2)$, both components of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ are positive and the first component is smaller than the second one. The blue polytope is the dual SLOPE unit ball $C_{\mathbf{\Lambda}}$ and 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 the Corollary 3.2.2, $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ 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 [26, 175, 61]. This explicit expression gives an analytical way to learn that the SLOPE solution is sparse and built of clusters.

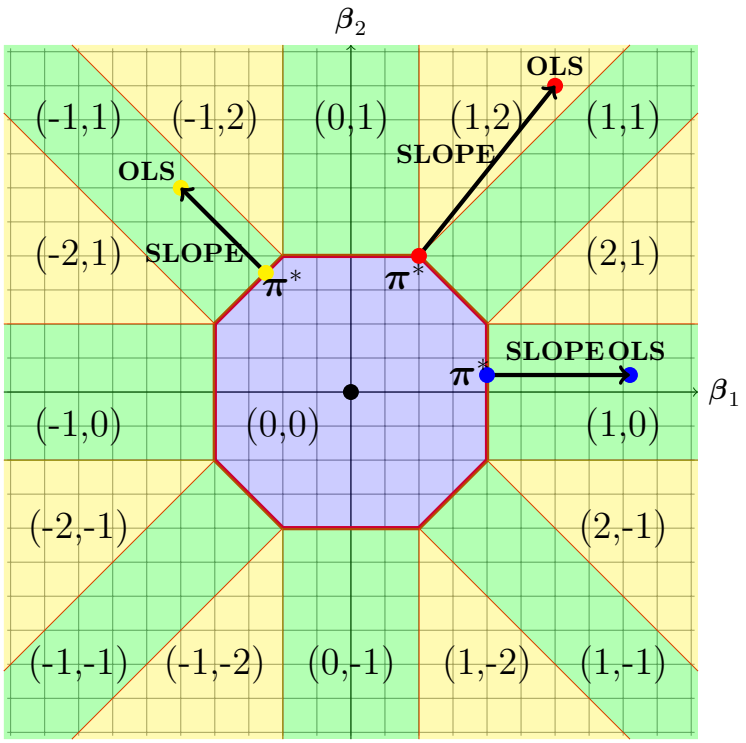



Figure 3.1: The dual unit ball $C_{\mathbf{\Lambda}/n}$ for $\mathbf{\Lambda} = (200, 100)'$ and examples of $\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$ and $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ in the orthogonal design for $n = 50$ and $p = 2$. The labels of each colored set refer to the pattern of $\hat{\boldsymbol{\beta}}^{\text{SLOPE}}$ for $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ lying in this set. The arrows point from $(\hat{\boldsymbol{\beta}}^{\text{OLS}} - \hat{\boldsymbol{\beta}}^{\text{SLOPE}})$ to $\hat{\boldsymbol{\beta}}^{\text{OLS}}$.

**Lemma 3.3.1.** *In the linear model satisfying* $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$ *we have*

$$\underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2n}\left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\right\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b})\right] = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left[\frac{1}{2}\left\|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}-\boldsymbol{b}\right\|_2^2 + J_{\boldsymbol{\Lambda}}(\boldsymbol{b})\right]. \tag{3.3.1}$$

Our proof of this Lemma is given in the Appendix. The next theorem gives a sufficient condition for the clustering effect of the SLOPE estimator in the orthogonal design.

**Theorem 3.3.1.** *Consider a linear model with orthogonal design* $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$. *Let* $\pi$ *be a permutation of* $(1, 2, \ldots, p)$ *such that*

$$\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(1)}\right| \geq \left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(2)}\right| \geq \ldots \geq \left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(p)}\right|.$$

*For* $i \in \{1, 2, \ldots, p-1\}$,
*if* $\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i)}\right| - \left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i+1)}\right| \leq \frac{\lambda_i-\lambda_{i+1}}{n}$, *then* $\left|\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}_{\pi(i)}\right| = \left|\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}_{\pi(i+1)}\right|$.

*Proof.* By Lemma 3.3.1, in the orthogonal design, $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ is the proximal map of $J_{\boldsymbol{\Lambda}/n}(\cdot)$ at $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. The result may be inferred from [26, Lemma 2.3]. □

In the following theorem we derive the necessary and sufficient conditions under which SLOPE in the orthogonal design recovers the support of the vector
$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)'$, i.e. $\quad \widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}_i = 0 \iff \boldsymbol{\beta}_i = 0.$

**Theorem 3.3.2.** *Under orthogonal design* $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$, *let* $\pi$ *be a permutation of* $(1, 2, \ldots, p)$ *that satisfies* $\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(1)}\right| \geq |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(2)}| \geq \ldots \geq |\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(p)}|$. *Without loss of generality suppose that supp*$(\boldsymbol{\beta}) = \{1, 2, \ldots, p_0\}$ *with* $p_0 < p$. *The necessary and sufficient condition for SLOPE to identify the set of relevant covariables is:*

*(a)* $\min\limits_{1\leq i\leq p_0}\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_i\right| > \max\limits_{p_0+1\leq i\leq p}\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_i\right|,$

*(b)* $\sum\limits_{i=k}^{p_0}\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i)}\right| > \frac{1}{n}\sum\limits_{i=k}^{p_0}\lambda_i, \quad \text{for} \quad k = 1, 2, \ldots, p_0,$

*(c)* $\sum\limits_{i=p_0+1}^{k}\left|\widehat{\boldsymbol{\beta}}^{\mathrm{OLS}}_{\pi(i)}\right| \leq \frac{1}{n}\sum\limits_{i=p_0+1}^{k}\lambda_i, \quad \text{for} \quad k = p_0+1, p_0+2, \ldots, p.$

*Proof.* The result may be inferred from the properties of the proximal SLOPE [27, Lemma 2.3 and Lemma 2.4] and from Lemma 3.3.1. □

## 3.4 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 $n \geq 1$ we consider a linear model

$$\boldsymbol{Y}^{(n)} = \boldsymbol{X}^{(n)}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{(n)}, \tag{3.4.1}$$

where $\boldsymbol{Y}^{(n)} = (y_1^{(n)}, y_2^{(n)}, \ldots, y_n^{(n)})' \in \mathbb{R}^n$ is a vector of observations, $\boldsymbol{X}^{(n)} \in \mathbb{R}^{n\times p}$ is a deterministic design matrix with $\mathrm{rank}(\boldsymbol{X}^{(n)}) = p$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)' \in \mathbb{R}^p$ is a vector of unknown

regression coefficients and $\boldsymbol{\varepsilon}^{(n)} = (\varepsilon_1^{(n)}, \varepsilon_2^{(n)}, \ldots, \varepsilon_n^{(n)})' \in \mathbb{R}^n$ is a noise term, which has the normal distribution $N(0, \sigma^2 \boldsymbol{I}_n)$. We make no assumptions about the dependence between $\boldsymbol{\varepsilon}^{(n)}$ and $\boldsymbol{\varepsilon}^{(m)}$ for $n \neq m$. In particular, $\boldsymbol{\varepsilon}^{(n)}$ does not need to 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 $n \geq 1$ its coefficients $\lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \ldots \geq \lambda_p^{(n)} \geq 0$ are fixed and $\lambda_1^{(n)} > 0$. By $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ we denote the SLOPE estimator corresponding to the tuning vector $\boldsymbol{\Lambda}^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})'$:

$$\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^p} \left[ \frac{1}{2} \left\| \boldsymbol{Y}^{(n)} - \boldsymbol{X}^{(n)} \boldsymbol{b} \right\|_2^2 + J_{\boldsymbol{\Lambda}^{(n)}}(\boldsymbol{b}) \right]. \tag{3.4.2}$$

### 3.4.1 Strong consistency of the SLOPE estimator

Let us recall the definition of a strongly consistent estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\mathbf{SLOPE}}$ of $\boldsymbol{\beta}$, i.e. $\forall \boldsymbol{\beta} \in \mathbb{R}^p$ we have $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\mathbf{SLOPE}} \to \boldsymbol{\beta}$ almost surely. Below, we discuss strong consistency of the sequence $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\mathbf{SLOPE}}$ of SLOPE estimators, defined by (3.4.2).

**Theorem 3.4.1.** *Consider the linear regression model* (3.4.1) *and assume that*

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

*where $\boldsymbol{C}$ is a positive definite matrix. Let $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$, $n \geq 1$, be the SLOPE estimator corresponding to the tuning vector $\boldsymbol{\Lambda}^{(n)} = (\lambda_1^{(n)}, \lambda_2^{(n)}, \ldots, \lambda_p^{(n)})'$.*

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

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

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

**Lemma 3.4.1.** *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 a folklore and we prove it in our setting.

**Proposition 3.4.1.** *Consider the linear regression model* (3.4.1).
*If $\lim\limits_n n^{-1} (\boldsymbol{X}^{(n)})' \boldsymbol{X}^{(n)} = \boldsymbol{C}$, where $\boldsymbol{C}$ is positive definite, then $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \xrightarrow{a.s.} \boldsymbol{\beta}$.*

*Proof.* We have

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

Then $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} - \boldsymbol{\beta}\right)$ has the normal distribution $N(0, \sigma^2(n^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)})^{-1})$ and its components satisfy the assumptions of Lemma 3.4.1. Since $\log(n)^{1/2+\delta} = o(\sqrt{n})$, we get the assertion by Lemma 3.4.1. □

*Proof of Theorem 3.4.1.* (a) It follows from Theorem 3.2.1 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 $\|\boldsymbol{\pi}_n^*\|_\infty \leq \lambda_1^{(n)}$. Hence,

$$\frac{\boldsymbol{\pi}_n^*}{n} \xrightarrow{a.s.} \mathbf{0}, \tag{3.4.3}$$

because $\left\|\frac{\boldsymbol{\pi}_n^*}{n}\right\|_\infty \leq \frac{\lambda_1^{(n)}}{n} \to 0$. The assumption that $\mathrm{rank}(\boldsymbol{X}^{(n)}) = p$ implies that the matrix $(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}$ is invertible and therefore 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 with (3.4.3) the fact that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \xrightarrow{a.s.} \boldsymbol{\beta}$, we conclude that

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

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

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

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

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

The last equality follows from the fact that $(\boldsymbol{X}^{(n)})'\boldsymbol{Y}^{(n)} = (\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)}\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}$.
Suppose to the contrary that the true parameter $\boldsymbol{\beta}$ satisfies $\lambda_0\|\boldsymbol{\beta}\|_\infty > \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta}/2$ (which is the case when $\|\boldsymbol{\beta}\|_\infty$ is sufficiently close to 0) and that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}$. Then, using the facts that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} \xrightarrow{a.s.} \boldsymbol{\beta}$ and that $\lim_n n^{-1}(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)} = \boldsymbol{C}$, we have

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

which provides a contradiction. This proves the first part of Theorem 3.4.1 (b). To prove the second part, note that strong consistency requires convergence for any value of the parameter $\boldsymbol{\beta}$. □

**Remark 3.4.2.** The proof of Theorem 3.4.1 (b) does not exclude the possibility that $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}$ when $\boldsymbol{\beta}$ satisfies $\lambda_0\|\boldsymbol{\beta}\|_\infty \leq \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta}/2$.

### 3.4.2 Asymptotic pattern recovery in the orthogonal design

We again consider a sequence of linear models (3.4.1) but this time we assume that for each $n$ the deterministic design matrix $\boldsymbol{X}^{(n)}$ of size $n \times p$ satisfies

$$(\boldsymbol{X}^{(n)})'\boldsymbol{X}^{(n)} = n\boldsymbol{I}_p. \tag{3.4.4}$$

As usual, we assume Gaussian errors $\boldsymbol{\varepsilon}^{(n)} \sim \mathrm{N}(0, \sigma^2\boldsymbol{I}_n)$.
Let $\widehat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = \left(\widehat{\boldsymbol{\beta}}_1^{\mathrm{SLOPE}}(n), \ldots, \widehat{\boldsymbol{\beta}}_p^{\mathrm{SLOPE}}(n)\right)'$ be the SLOPE estimator defined by (3.4.2). With the above notation we present the main result of this section.

**Theorem 3.4.3.** *Assume that*

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

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

$$\liminf_{n\to\infty} \frac{\lambda_i^{(n)} - \lambda_{i+1}^{(n)}}{\sqrt{n}\,(\log(n))^{1/2+\delta}} = m > 0 \quad \textit{for} \quad i = 1, \ldots, p-1. \tag{3.4.5}$$

*Then we have*

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

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

*Proof.* Without loss of generality we may assume that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_p \geq 0$. Indeed, we can always achieve such condition 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 $n$, $\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) = \boldsymbol{patt}(\boldsymbol{\beta})$ a.s. We divide the proof into the following four parts:

(a) $\beta_i = \beta_j > 0 \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) = \widehat{\beta}_j^{\mathrm{SLOPE}}(n)$ a.s. for large $n$,

(b) $\beta_i > \beta_{i+1} \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) > \widehat{\beta}_{i+1}^{\mathrm{SLOPE}}(n)$ a.s. for large $n$,

(c) $\beta_i = 0 \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) = 0$ a.s. for large $n$,

(d) $\beta_i > 0 \implies \widehat{\beta}_i^{\mathrm{SLOPE}}(n) > 0$ a.s. for large $n$.

The points (b) and (d) follow quickly by the strong consistency of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}(n)$. To prove (a) and (c) we observe that for each $n$ we are in the orthogonal design case.
Let $\pi_n$ be a permutation of $(1, 2, \ldots, p)$ satisfying

$$|\widehat{\beta}_{\pi_n(1)}^{\mathrm{OLS}}(n)| \geq |\widehat{\beta}_{\pi_n(2)}^{\mathrm{OLS}}(n)| \geq \ldots \geq |\widehat{\beta}_{\pi_n(p)}^{\mathrm{OLS}}(n)|.$$

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 (a). 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

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

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

$$\widehat{\beta}_{\pi_n(k)}^{\mathrm{SLOPE}}(n) = \widehat{\beta}_{\pi_n(k+1)}^{\mathrm{SLOPE}}(n) \quad \text{a.s.}, \tag{3.4.6}$$

thus $\widehat{\beta}_j^{\mathrm{SLOPE}}(n) = \widehat{\beta}_k^{\mathrm{SLOPE}}(n)$ for $j, k \in S_i$, which finishes the proof of (a).
Now assume that $\pi_n(k), \pi_n(k+1) \in S_i$. Then, by Theorem 3.3.1, the condition (3.4.6) is satisfied if

$$\left|\widehat{\beta}_{\pi_n(k)}^{\mathrm{OLS}}(n)\right| - \left|\widehat{\beta}_{\pi_n(k+1)}^{\mathrm{OLS}}(n)\right| \leq \frac{1}{n}\left(\lambda_k^{(n)} - \lambda_{k+1}^{(n)}\right) \tag{3.4.7}$$

holds for large $n$ and both $\widehat{\beta}_{\pi_n(k)}^{\mathrm{OLS}}(n)$ and $\widehat{\beta}_{\pi_n(k)}^{\mathrm{OLS}}(n)$ 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 following bound

$$\left|\widehat{\beta}_{\pi_n(k)}^{\mathrm{OLS}}(n) - \widehat{\beta}_{\pi_n(k+1)}^{\mathrm{OLS}}(n)\right| \leq \sum_{j \in S_i} \left|\widehat{\beta}_j^{\mathrm{OLS}}(n) - \widehat{\beta}_i^{\mathrm{OLS}}(n)\right|. \tag{3.4.8}$$

Take any $j \in S_i$. Since both $\widehat{\beta}_j^{\mathrm{OLS}}(n)$ and $\widehat{\beta}_i^{\mathrm{OLS}}(n)$ have the normal distribution with the same mean, by Lemma 3.4.1, we have

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

In view of (3.4.8) and (3.4.5), this implies that (3.4.7) holds true for large $n$. Hence, (a) follows.
It remains to establish (c). Assume that $\beta_{p_0} > 0 = \beta_{p_0+1} = \ldots = \beta_p$. Clearly, condition (a) from Theorem 3.3.2 is satisfied thanks to the strong consistency of the OLS estimator. For (b), we have for $k = 1, 2, \ldots, p_0$,

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

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 3.3.2 holds for large $n$. Condition (c) from Theorem 3.3.2 follows from Lemma 3.4.1. Indeed, we have for $\delta > 0$ and $k = p_0 + 1, \ldots, p$,

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

while

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

Thus, all assumptions of Theorem 3.3.2 are verified and the proof is complete. □

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

Table 3.1: Comparison of MSE between different regression methods.

## 3.5 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.
$X_{i,(2*j-1)} = \sin(2\pi ij/n)$ and $X_{i,(2*j)} = \cos(2\pi ij/n)$ for $i = 1, \ldots, 100$
and $j = 1, \ldots, 150$. 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 coordinates of $\boldsymbol{\beta}$ are sorted in a decreasing way. The signs of the nonzero 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 pattern matrix $\boldsymbol{U}_{\mathbf{M}}$ and the clustered design matrix $\widetilde{\boldsymbol{X}}_{\mathbf{M}}$, which is based on the SLOPE pattern.

To perform the debiased SLOPE, we begin with recovering the support and clusters of a true vector $\boldsymbol{\beta}$ with SLOPE. Then, using the obtained SLOPE pattern $M$, we replace the design matrix with its clustered version $\widetilde{\boldsymbol{X}}_M = \boldsymbol{X}\mathbf{U}_M$. 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 $\hat{\boldsymbol{\beta}}^{\mathbf{SLOPE}}$.

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

For LASSO we have the LASSO pattern that is a vector of signs, see Chapter 5. For $\boldsymbol{S} \in \{-1, 0, 1\}^p$, $\|\boldsymbol{S}\|_1$ denotes the number of nonzero coordinates. If $\|bS\|_1 = k \geq 1$, then we define the corresponding pattern matrix $\boldsymbol{U}_S \in \mathbb{R}^{p\times k}$ by

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

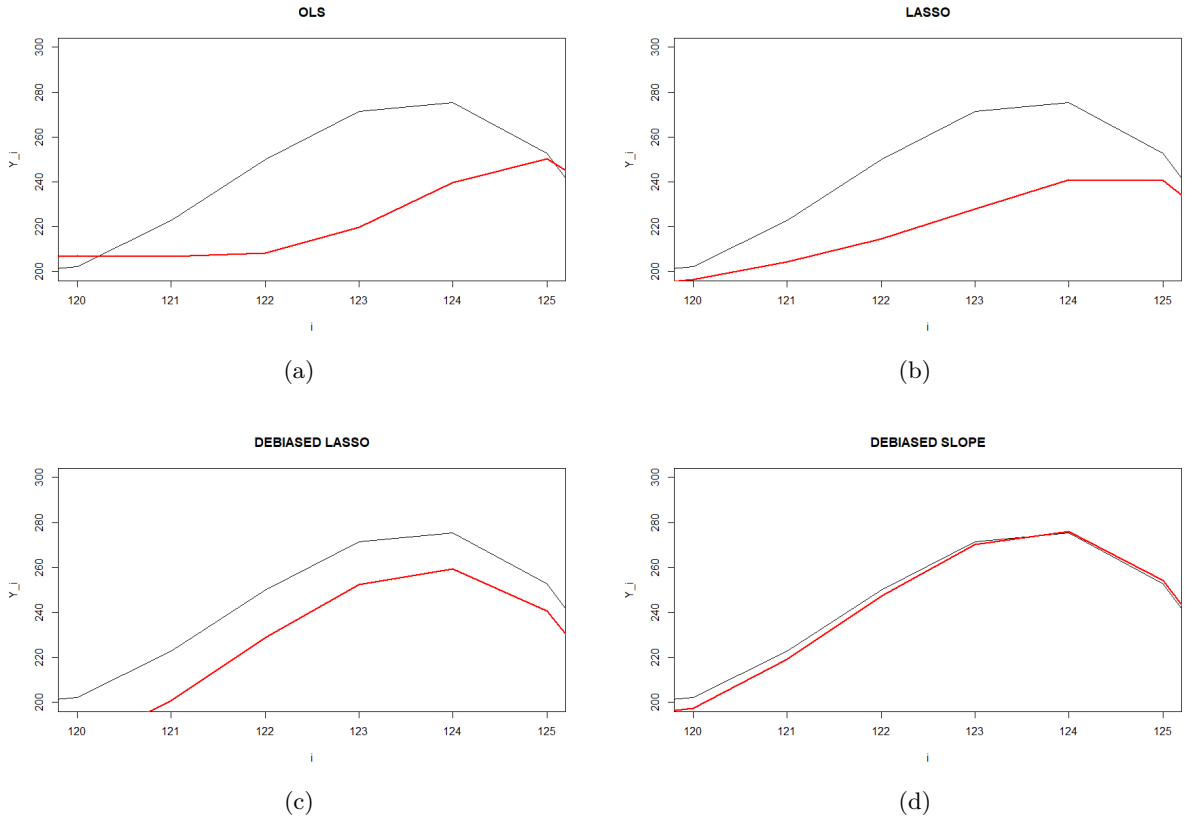
i.e. the submatrix of $\operatorname{diag}(\boldsymbol{S})$ obtained by keeping columns corresponding to indices in $\operatorname{supp}(\boldsymbol{S})$. Then we define the reduced matrix $\widetilde{\boldsymbol{X}}_S$ by

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

Equivalently, we have $\widetilde{\boldsymbol{X}}_S = (S_i X_i)_{i\in\operatorname{supp}(S)}$. For a broader discussion on the pattern matrix, we encourage to see the next Chapter. In our example $\boldsymbol{\varepsilon} \in \mathcal{N}(0, \sigma^2\boldsymbol{I}_n)$ and $\sigma = 30$.

We compare the Mean Square Error and the signal denoising of the classical OLS estimation, the LASSO with the tuning parameter $\lambda_{cv}$ minimizing the cross-validated error, the debiased version of LASSO with $\lambda = 5\lambda_{cv}$ and the debiased version of SLOPE with the tuning vector $\boldsymbol{\Lambda}$ chosen with respect to the sequence proposed below Theorem 3.4.3 ($\lambda_i = 0.1(p + 1 - i)n^{2/3}$).

We also compare debiased SLOPE with debiased LASSO based on a single trial, as shown in Figure 3.4 and Table 3.1. The horizontal lines correspond to the 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.




(a) (b)

(c) (d)

Figure 3.2: 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 $\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

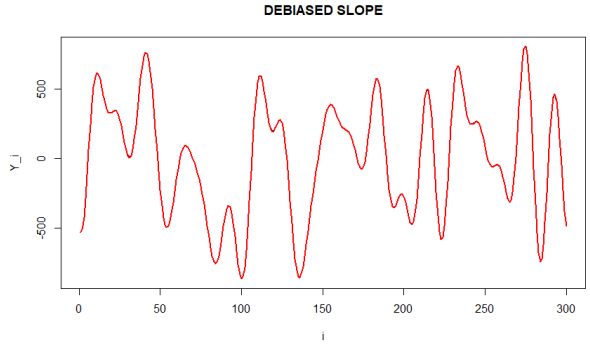
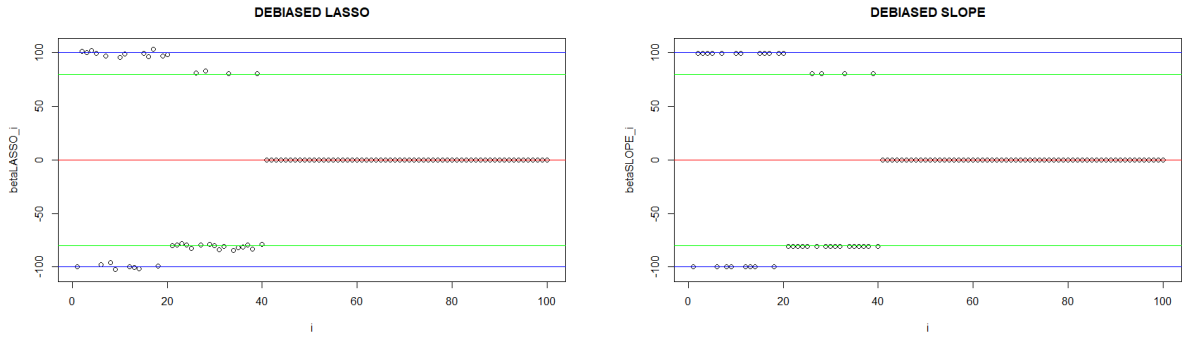



Figure 3.3: Signal denoising by debiased SLOPE on all coordinates of the regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The (almost overlapping) black line and the red line correspond respectively to the true values of $\boldsymbol{X}\boldsymbol{\beta}$ and to $\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$.




(a) (b)

Figure 3.4: Pattern recovery by debiased LASSO (left) and by debiased SLOPE (right) in the same setting as above.

## 3.6 Appendix

*Proof of Lemma 3.2.1.* Since the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is nonnegative 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 $\nabla g$ of $g$ satisfies

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

This implies $(\boldsymbol{\pi} - \boldsymbol{\pi}^*)'\boldsymbol{\beta}^* \leq 0$, 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), (b) and (c), we use the fact that $\boldsymbol{\pi}^*$ maximizes $\boldsymbol{\pi}'\boldsymbol{\beta}^*$ over $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$.
To prove part (a), suppose that $\text{sign}(\beta_i^*) \cdot \text{sign}(\pi_i^*) < 0$ for some $i$ and define

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

Then we have $(\boldsymbol{\pi}^*)'\boldsymbol{\beta}^* < \boldsymbol{\pi}'\boldsymbol{\beta}^*$, which is impossible since $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$.
To prove part (b), consider a permutation $\tau$ of $(1, 2, \ldots, n)$ such that $(|\pi_{\tau(1)}^*|, \ldots, |\pi_{\tau(p)}^*|)$ and $(|\beta_1^*|, \ldots, |\beta_p^*|)$ are similarly sorted. Define the point $\boldsymbol{\pi} = (s_1 \cdot \pi_{\tau(1)}^*, s_2 \cdot \pi_{\tau(2)}^*, \ldots, s_p \cdot \pi_{\tau(p)}^*)$, where $s_i = \text{sign}(\beta_i^*)$, for $i = 1, 2, \ldots, p$. If $(|\pi_{\tau(1)}^*|, \ldots, |\pi_{\tau(p)}^*|) \neq (|\pi_1^*|, \ldots, |\pi_p^*|)$, then, by the Hardy-Littlewood-Pólya rearrangement inequality,

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

which is impossible since $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$.
Finally, to prove part (c), suppose that $\sum_{i=1}^{k-1} |\pi_{\tau(i)}^*| < \sum_{i=1}^{k-1} \lambda_i$, and that $|\pi_{\tau(k)}^*| > 0$. 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 3.2.2.* At first we note that for all $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$

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

where the last inequality follows from the fact that $(\boldsymbol{\pi} - \boldsymbol{\pi}^*)'\boldsymbol{\beta}^* \le 0$, for all $\boldsymbol{\pi} \in C_{\boldsymbol{\Lambda}}$, see the proof of 3.2.1. Therefore, $\max_{\pi \in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta}^*, \boldsymbol{\pi}) = r(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)$. Moreover, from the definition of the point $\boldsymbol{\beta}^*$ it is seen that $r(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*) = \min_{\beta \in \mathcal{M}} r(\boldsymbol{\beta}, \boldsymbol{\pi}^*)$. These two facts imply that

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

Since $\max_{\pi \in C_{\boldsymbol{\Lambda}}} \min_{\beta \in \mathcal{M}} r(\boldsymbol{\beta}, \boldsymbol{\pi}) \le \min_{\beta \in \mathcal{M}} \max_{\pi \in C_{\boldsymbol{\Lambda}}} r(\boldsymbol{\beta}, \boldsymbol{\pi})$ (by the max-min inequality), we have the equality throughout. This completes the proof. □

*Proof of Lemma 3.3.1.* Observe that

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

Therefore, both 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. □

# Chapter 4

# Pattern recovery by SLOPE

## 4.1 Introduction

Most of the content of this chapter may be found in the preprint of Bogdan, Dupuis, Graczyk, Kołodziejek, the author of the dissertation, Tardivel and Wilczyński [23]. We decided to enrich the chapter with geometrical explanation of the SLOPE pattern and its recovery to make it more accessible for the general audience and to inform the reader about the connections between the SLOPE estimator and the convex geometry. Such connections exist also for LASSO and may be useful in its study. The geometrical approach was the one, which made us build the intuition, which led us to establish the theoretical results.
In this chapter we give a deeper and more general focus on recovering the pattern of $\boldsymbol{\beta}$ by SLOPE. Now we abstain from restricting to the case of the orthogonal design matrix and we give novel results on the pattern recovery for a broad generality of situations.
In particular, unlike in the previous chapter, below we concern a model, in which the error is incremental. We also extend the discussion on the properties of the SLOPE estimator to the high-dimensional case, in which $p$ may be much larger than $n$.

### 4.1.1 History of SLOPE

SLOPE estimator was introduced by Bogdan et al. [27, 26]. In their seminal paper, SLOPE was motivated by the control of the expected rate of false discoveries (FDR control) in multiple hypotheses testing. Moreover, they proposed a choice of $\boldsymbol{\Lambda}$ being a Benjamini-Hochberg sequence dependent on a parameter $q \in (0, 1)$, under which, in the orthogonal design, SLOPE controls FDR at a level $q \cdot \frac{p-k}{p}$, where $k$ is the number of non-zero coordinates of $\boldsymbol{\beta}$. They also proposed a fast proximal algorithm computing $\hat{\boldsymbol{\beta}}^{SLOPE}$ in the orthogonal design.
However, the first non-trivial example of SLOPE was introduced a few years earlier by Bondell and Reich [29] under the name OSCAR. Their article deals with the tuning vector $\boldsymbol{\Lambda}$ being an arithmetic sequence. The OSCAR penalty was reformulated in terms of a sorted $\ell_1$ norm by Zeng and Figueiredo [190]. Going with the flow, they independently propose the SLOPE estimator [189] under the name OWL. In [76] they propose its representation as a gauge function, compute its dual and propose another formula for the proximal operator algorithm.
Figueiredo and Nowak [76] prove that if the columns of $\boldsymbol{X}$ are correlated enough, then SLOPE results in their clusterization. Negrinho and Martins [136] connect the SLOPE estimator with the notion of a signed permutahedron $C_{\boldsymbol{\Lambda}}$, noticing that it is a unit ball in the dual of SLOPE norm. Bellec and Tsybakov [13] prove the equality $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - Proj_{C_{\boldsymbol{\Lambda}/n}}(\boldsymbol{Y})$ in the linear regression penalized by norm. They also propose novel oracle inequalities on a prediction error,

given $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ and $\lambda_j = \sigma\sqrt{\log(\frac{2p}{j})/n}$.
Su and Candès [171] prove that with $\boldsymbol{\Lambda}$ being the Benjamini-Hochberg sequence, SLOPE is an asymptotically minimax estimator and show more detailed results on the asymptotics of its squared error for orthogonal ($\boldsymbol{X'X} = \boldsymbol{I}_p$) and gaussian ($X_{ij} \overset{iid}{\sim} N(0, \frac{1}{n})$) designs. Moreover, they point out such sizes of the support of $\boldsymbol{\beta}$ ($\ell_0$-sparsities), for which the risk for SLOPE is much smaller than for LASSO or SURE. Bellec, Lecué and Tsybakov [14] extend the results of Su and Candès [171] on achieving the minimax optimal prediction rate of SLOPE. In particular, they allow the desing matrix $\boldsymbol{X}$ to be deterministic, leaving only mild assumptions on its eigenvalues. They also derive sharp oracle inequalities and prove that the minimaxity holds for an $\ell_q$-estimation error for any $1 \leq q \leq 2$.
Under the gaussian design, Kos and Bogdan [115] use the results of Su and Candès [171] on the asymptotics of SLOPE in order to propose conditions on the cardinality and magnitude of non-zero coordinates of $\boldsymbol{\beta}$, under which the SLOPE with $\boldsymbol{\Lambda}$ being a Benjamini-Hochberg sequence asymptotically achieves $FDR \to 0$ and the power converging to 1. Hu and Lu [103] put their effort to look for an optimal choice of $\boldsymbol{\Lambda}$ in terms of type I error and power of variable selection. They consider an asymptotic scenario, in which $\frac{n}{p} \to \delta > 0$, $\frac{k}{p} \to \rho > 0$ and the vectors $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$ and $\boldsymbol{\Lambda}$ converge to limiting measures.
Sankaran, Bach and Bhattacharrya [154] prove that SLOPE belongs to the family of Lovász extensions. They also propose the smoothed version of SLOPE (called SOWL) and derive sufficient condition for its consistency. Minami [132] studies the projections onto the dual norm ball to study the notion of degrees of freedom. SLOPE is used here as one of examples of submodular norm regularizations.
Schneider and Tardivel [156] propose a geometrical approach to the SLOPE estimator. Firstly, they give full characterization of the uniqueness of $\hat{\boldsymbol{\beta}}^{SLOPE}$ that for a given $\boldsymbol{X}$ holds for any $\boldsymbol{Y} \in \mathbb{R}^n$. They express this condition both in analytical and geometrical way. This is a generalization of an analogous result for LASSO [71]. They also introduce the notion of SLOPE pattern and highlight its connection with faces of the unit ball in the dual $J^*_{\boldsymbol{\Lambda}}$ of the SLOPE norm. In their article, the definition of accessibility of the SLOPE pattern is also introduced and enriched with its geometrical full characterization. Tardivel, Servien and Concordet [175] propose an exact formula for SLOPE in the orthogonal case, using the Cesàro summation.
For gaussian designs, random $\boldsymbol{\beta}$, random $\boldsymbol{\varepsilon}$ and for $\frac{n}{p} \overset{p\to\infty}{\to} \delta > 0$, Zhang and Bu [191] use the projected gradient descent algorithm to find $\boldsymbol{\Lambda}$, which reduces the Mean Squared Error (MSE) and compare obtained MSE with ones achieved by SLOPE with Benjamini-Hochberg $\boldsymbol{\Lambda}$ and by LASSO. To minimize MSE for arbitrary data, they propose finding $\boldsymbol{\Lambda}$ with help of a coordinate descent algorithm, while imposing the restriction of $\boldsymbol{\Lambda}$ to contain no more than $k$ distinct values with $k << p$. In their recent article, Larsson et al. [120] improved solving numerically the SLOPE optimization problem, using modifications of the proximal gradient descent and the proximal coordinate descent algorithm.
Bu et al. [40] proposed an approximate message passing (AMP) algorithm, to quickly obtain a solution, which they prove to converge in $\ell_2$ to $\hat{\boldsymbol{\beta}}^{SLOPE}$. They also refine the asymptotic results of the $\ell_2$-convergence of $\hat{\boldsymbol{\beta}}^{SLOPE}$ to the true $\boldsymbol{\beta}$. Larsson, Bogdan and Wallin [119] introduced the strong screening rule for SLOPE, which generalized the analogous rule for LASSO. The safe screening rules were proposed by Bao, Gu and Huang [8] as well as by Elvira and Herzet [65]. Their safety is understood as no detection any non-zero coordinate of $\boldsymbol{\beta}$ as zero. It is useful in situation, when some prior knowledge of $\boldsymbol{\beta}$ is available.
Brzyski et al. [38] introduce the group SLOPE method. Lee, Sobczyk and Bogdan [123] apply SLOPE to gaussian graphical models, proposing the Neighborhood Selection SLOPE (nsSLOPE)

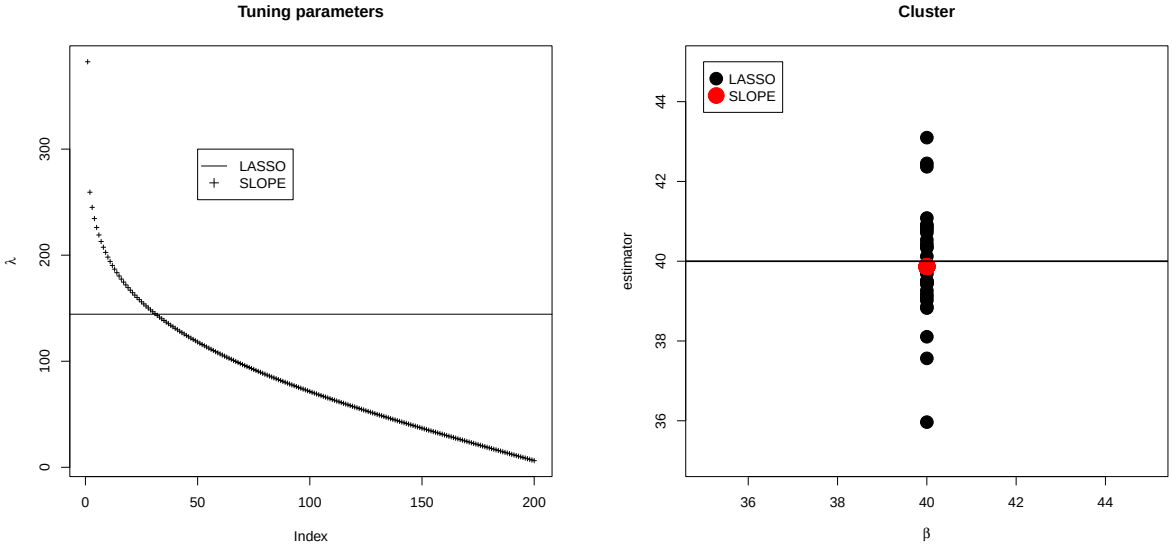
algorithm with the FDR control of edges detection. Mazza-Anthony, Mazoure and Coates [129] independently introduce the graphical SLOPE method (under the name gOWL), which applies the SLOPE norm to estimate the precision matrix by clustering its off-diagonal entries. Recently, Riccobello et al. [148] extended the discussion on graphical SLOPE to t-Student (Tslope) data.
Other applications of SLOPE include e.g. the paper of Sepehri [161], who proposes the bayesian approach in which SLOPE is considered as a maximum a posteriori (MAP) distribution in a bayesian regression problem. Its adaptive version was later proposed by Jiang et al. [109]. Stucky and van de Geer [170] introduce the Square Root SLOPE, in which $\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2^2$ is replaced by $\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2$. It is an analogue of the Square Root LASSO [15].
Recently, Dexheimer and Strauch [56] found a way to apply the results of [14] and used SLOPE to improve the estimation of the drift parameter in Lévy-driven Ornstein-Uhlenbeck processes.

### 4.1.2 Our contribution

In this chapter, we give the necessary and sufficient conditions for recovering the SLOPE pattern. From a mathematical perspective, our main result is Theorem 4.3.1, in which we propose the positivity and the subdifferential condition for the pattern recovery. To ease the interpretation, we also present the dual norm condition and the SLOPE irrepresentability condition, which is 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 [82, 130, 183, 192, 194]. However, we believe that our mathematical perspective is novel, and paves the path for similar analyses of other penalized estimators. For the noisy case, in Theorem 4.5.1 we give the open SLOPE irrepresentability condition, which implies the consistency of the SLOPE pattern for fixed $\boldsymbol{X}$ and the difference between non-equal coordinates of $|\boldsymbol{\beta}|$ diverging to $\infty$. Later on, in Theorems 4.5.3 and 4.5.5, we give conditions for asymptotic pattern recovery, when $p$ is fixed and $n$ diverges to $\infty$. As another main results of this chapter, we provide geometrical interpretations of irrepresentability conditions as well as refined results of the strong consistency of $\hat{\boldsymbol{\beta}}^{SLOPE}$ and of its pattern.

### 4.1.3 Motivation

While the SLOPE ability to identify the pattern of the vector of regression coefficients $\boldsymbol{\beta}$ is interesting by itself, the related reduction of model dimension also brings the advantage in terms of the precision of the estimation of $\boldsymbol{\beta}$. This phenomenon is illustrated in Figure 4.1, which presents the difference in precision of the 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 $\mathbb{P}(X_{i(j+1)} \neq X_{ij}) = 1 - \mathbb{P}(X_{i(j+1)} = X_{ij}) \approx 0.0476$. This value corresponds to the probability of the crossover event between genetic markers spaced every 5 centimorgans. To be more specific, it is close to the inverse of the Haldane mapping function [81, pp. 13-14] evaluated at 0.05, which is equal to $e^{-0.05}\sinh(0.05)$. 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, $\rho(\boldsymbol{X}_i, \boldsymbol{X}_j) \approx 0.9048^{|i-j|}$. 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 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $\beta_1 = \ldots = \beta_{30} = 40$, $\beta_{31} = \ldots = \beta_{200} = 0$ and $\sigma = 5$. In this experiment, the data matrix $\boldsymbol{X}$ and the regression model




(a) (b)

Figure 4.1: Comparison of LASSO and SLOPE when the cluster structure is present in the data. Here $n = 100$, $p = 200$, the rows of the matrix $\boldsymbol{X}$ are simulated as independent binary Markov chains, with the transition probability 0.0476 (corresponding to the genetic distance of 5 centimorgans). The correlation between $i^{th}$ and the $j^{th}$ column of $\boldsymbol{X}$ decays exponentially as $0.9048^{|i-j|}$. The first $k = 30$ columns of $\boldsymbol{X}$ are associated with $\boldsymbol{Y}$ and their nonzero regression coefficients are all equal to 40 (other details are provided in the text). The 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. The right panel represents the LASSO and SLOPE estimates.

are constructed such that the LASSO irrepresentability condition holds. The tuning parameter for LASSO is selected as the smallest value of $\lambda$ for which LASSO can correctly identify the sign of $\boldsymbol{\beta}$. Similarly, the tuning parameter $\boldsymbol{\Lambda}$ is designed such that the SLOPE irrepresentability condition holds and $\boldsymbol{\Lambda}$ 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 Figure 4.1. Both in the case of LASSO and SLOPE, the proposed tuning parameters are close to the values minimizing the mean squared estimation error. Since in this example both LASSO and SLOPE properly estimate null components of $\boldsymbol{\beta}$ at $\mathbf{0}$, the right panel in Figure 4.1 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 vector of regression coefficients $\boldsymbol{\beta}$ virtually without an error, while LASSO estimates are scattered over the interval between 36 and 44. In the result, the squared error of the LASSO estimator is more than 100 times larger than the squared error of SLOPE (63.4 vs 0.53).

## 4.2 Preliminaries on clustering properties by SLOPE

As the central notion of this chapter is the recovery of the SLOPE pattern, we recall its definition:

**Definition 4.2.1.** *The SLOPE pattern is a function* $\boldsymbol{patt} : \mathbb{R}^p \to \mathbb{Z}^p$ *such that*

$$patt(\boldsymbol{b})_i = \operatorname{sign}(b_i) rank(|b_i|),$$

*where* $rank(|b_i|) \in \{1, 2, \ldots, k\}$ *is defined to be the number of* $|c_j|$*'s satisfying* $|b_i| \geq |c_j|$*, where* $|c_1|, |c_2|, \ldots, |c_k|$*,* $k \leq p$*, are distinct non-zero values among* $|b_1|, \ldots, |b_p|$*. We adopt the convention that* $rank(0) = 0$*.*

We also recall that $\mathbb{R}^{k+} = \{\boldsymbol{\kappa} \in \mathbb{R}^k\colon\ \kappa_1 > \ldots > \kappa_k > 0\}$ and that $\mathcal{M}_p$ is the set of SLOPE patterns of $\boldsymbol{b} \in \mathbb{R}^p$. Definition 2.2.4 implies that for $\mathbf{0} \neq \boldsymbol{M} \in \mathcal{M}_p^{\text{SLOPE}}$ and $k = \|\boldsymbol{M}\|_\infty$, for $\boldsymbol{b} \in \mathbb{R}^p$ we have

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

**Example 4.2.1.** Let $\boldsymbol{b} = (2, 5, -2, 0, 8)'$. Then $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{b}) = (1, 2, -1, 0, 3)'$ and for $\boldsymbol{\kappa} = (8, 5, 2)'$ we have

$$\boldsymbol{U}_{\mathbf{M}}\boldsymbol{\kappa} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 8 \\ 5 \\ 2 \end{bmatrix} = \boldsymbol{b}.$$

### 4.2.1 Clustered design matrix and clustered parameter

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

(a) null components $m_i = 0$ lead to the elimination of columns $\boldsymbol{X}_i$ from the design matrix $\boldsymbol{X}$,

(b) a cluster $K \subset \{1, \ldots, p\}$ of $\boldsymbol{M}$ (a subset of $\boldsymbol{M}$ having coordinates equal in absolute value) leads to replace the columns $(\boldsymbol{X}_i)_{i\in K}$ by one column equal to their signed sum: $\sum_{i\in K} \mathrm{sign}(m_i)\boldsymbol{X}_i$.

**Example 4.2.2.** Let $\boldsymbol{b}$ be from Example 4.2.1, $\boldsymbol{X} = (\boldsymbol{X}_1|\boldsymbol{X}_2|\boldsymbol{X}_3|\boldsymbol{X}_4|\boldsymbol{X}_5)$ and $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)' \in \mathbb{R}^{5+}$. Then the clustered matrix and the clustered parameter are:

$$\widetilde{\boldsymbol{X}}_{\mathbf{M}} = (\boldsymbol{X}_5|\boldsymbol{X}_2|\boldsymbol{X}_1 - \boldsymbol{X}_3) \text{ and } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 + \lambda_4 \end{pmatrix}.$$

### 4.2.2 Sorted $\ell_1$ norm, dual sorted $\ell_1$ norm and subdifferential

As SLOPE is based on penalizing the residual sum of squared with the sorted $\ell_1$ norm, we recall its definition:

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

where $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ are the sorted components of $\boldsymbol{b}$ with respect to the absolute value. Given a norm $\|\cdot\|$ on $\mathbb{R}^p$, we recall that the dual norm $\|\cdot\|^*$ is defined by $\|\boldsymbol{b}\|^* = \max\{\boldsymbol{v}'\boldsymbol{b}\colon\ \|\boldsymbol{v}\| \leq 1\}$, for some $\boldsymbol{b} \in \mathbb{R}^p$. In particular, the dual sorted $\ell_1$ norm has an explicit expression given in [136]:

$$J^*_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \max\left\{\frac{|b|_{(1)}}{\lambda_1}, \frac{\sum_{i=1}^2 |b|_{(i)}}{\sum_{i=1}^2 \lambda_i}, \ldots, \frac{\sum_{i=1}^p |b|_{(i)}}{\sum_{i=1}^p \lambda_i}\right\}.$$

We recall the subdifferential of a norm $\|\cdot\|$ at $\boldsymbol{b}$ (see *e.g.* [101, Def. VI.1.2.1]:

$$\begin{aligned} \partial\|\cdot\|(\boldsymbol{b}) &= \{\boldsymbol{v} \in \mathbb{R}^p\colon\ \|\boldsymbol{z}\| \geq \|\boldsymbol{b}\| + \boldsymbol{v}'(\boldsymbol{z} - \boldsymbol{b})\ \forall\, \boldsymbol{z} \in \mathbb{R}^p\}, \\ &= \{\boldsymbol{v} \in \mathbb{R}^p\colon\ \|\boldsymbol{v}\|^* \leq 1 \text{ and } \boldsymbol{v}'\boldsymbol{b} = \|\boldsymbol{b}\|\}. \end{aligned} \tag{4.2.1}$$

For the sorted $\ell_1$ norm, geometrical descriptions of the subdifferential at $\boldsymbol{b} \in \mathbb{R}^p$ have been given in the particular case where $b_1 \geq \ldots \geq b_p \geq 0$ [61, 156, 175]. Below, for an arbitrary $\boldsymbol{b} \in \mathbb{R}^p$, we propose a new and useful formula for the subdifferential of the sorted $\ell_1$ norm. This representation is the crux of the mathematical content of the present chapter.

**Proposition 4.2.1.** *Let $\boldsymbol{b} \in \mathbb{R}^p$ and $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{b})$. Then:*

$$\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \left\{\boldsymbol{v} \in \mathbb{R}^p :\ J^*_{\boldsymbol{\Lambda}}(\boldsymbol{v}) \leq 1 \text{ and } \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\right\}. \tag{4.2.2}$$

In Proposition 4.9.1 we derive a simple characterization of elements of $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$. The notion of SLOPE pattern is related to the subdifferential via the following result.

**Proposition 4.2.2.** *Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$ and $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)'$ where $\lambda_1 > \ldots > \lambda_p > 0$. We have $\boldsymbol{patt}(\boldsymbol{a}) = \boldsymbol{patt}(\boldsymbol{b})$ if and only if $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{a}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$.*

A proof of Proposition 4.2.2 can be found in [156]. In the Appendix, we provide an independent proof, which is based on Proposition 4.2.1.
From now, to comply with Proposition 4.2.2, we assume that the tuning parameter $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)'$ satisfies

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

### 4.2.3 Characterization of SLOPE solutions

SLOPE estimator is a solution of the following optimization problem:

$$S_{\mathbf{X},\mathbf{\Lambda}}(\boldsymbol{Y}) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2^2 + J_{\mathbf{\Lambda}}(\boldsymbol{b})\right\}. \tag{4.2.3}$$

We do not assume that $S_{\mathbf{X},\mathbf{\Lambda}}(\boldsymbol{Y})$ is a singleton. However, note that the cases in which the SLOPE estimator is not unique are very rare. Indeed, the family of matrices $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, for which there exists a $\boldsymbol{Y}\in\mathbb{R}^n$ such that $S_{\mathbf{X},\mathbf{\Lambda}}(\boldsymbol{Y})$ is not a singleton, has a null Lebesgue measure on $\mathbb{R}^{n\times p}$ [156]. If $\ker(\boldsymbol{X})=\{\mathbf{0}\}$, then $S_{\mathbf{X},\mathbf{\Lambda}}(\boldsymbol{Y})$ consists of one element. Recall that a convex function $f$ attains its minimum at a point $\boldsymbol{b}$ if and only if $\mathbf{0}\in\partial f(\boldsymbol{b})$. Since $\partial\left(\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}\|_2^2\right)=\{-\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})\}$, the SLOPE estimator admits the following characterization:

$$\hat{\boldsymbol{\beta}}\in S_{\mathbf{X},\mathbf{\Lambda}}(\boldsymbol{Y}) \quad\Leftrightarrow\quad \boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\in\partial J_{\mathbf{\Lambda}}(\hat{\boldsymbol{\beta}}). \tag{4.2.4}$$

## 4.3 Characterization of pattern recovery by SLOPE

The characterization of pattern recovery by SLOPE given in Theorem 4.3.1 is one of most important results of this chapter. We recall that $\widetilde{\boldsymbol{P}}_{\mathbf{M}} = (\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{X}}'_{\mathbf{M}} = \widetilde{\boldsymbol{X}}_{\mathbf{M}}\widetilde{\boldsymbol{X}}^+_{\mathbf{M}}$ is the orthogonal projection onto $\mathrm{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})$, where $\boldsymbol{A}^+$ denotes the Moore-Penrose inverse of the matrix $\boldsymbol{A}$.

**Theorem 4.3.1.** *Let $\boldsymbol{X}\in\mathbb{R}^{n\times p}$, $\mathbf{0}\neq\boldsymbol{\beta}\in\mathbb{R}^p$, $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}\in\mathbb{R}^n$ and $\mathbf{\Lambda}\in\mathbb{R}^{p+}$. Let $\boldsymbol{M}=\boldsymbol{patt}(\boldsymbol{\beta})\in\mathcal{M}_p^{\mathrm{SLOPE}}$ and $k=\|\boldsymbol{M}\|_\infty$. Denote*

$$\boldsymbol{\pi} = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\mathbf{\Lambda}}_{\mathbf{M}} + \boldsymbol{X}'(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}_{\mathbf{M}})\boldsymbol{Y}. \tag{4.3.1}$$

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

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

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

**Remark 4.3.2.**

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

$$A=\left\{\omega\in\Omega\colon \text{ there exists } \boldsymbol{s}\in\mathbb{R}^{k+} \text{ such that } \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{Y}(\omega)-\widetilde{\mathbf{\Lambda}}_{\mathbf{M}}=\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{s}\right\},$$
$$B=\left\{\omega\in\Omega\colon \ \boldsymbol{\pi}(\omega)\in\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})\right\}.$$

Indeed, since $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}=\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{P}}_{\mathbf{M}}$ then $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{Y}(\omega)$ depends on $\boldsymbol{\varepsilon}_A(\omega)=\widetilde{\boldsymbol{P}}_{\mathbf{M}}\boldsymbol{\varepsilon}(\omega)$. Moreover, $\boldsymbol{\pi}(\omega)$ depends on $\boldsymbol{\varepsilon}_B(\omega)=(\boldsymbol{I}_n-\widetilde{\boldsymbol{P}}_{\mathbf{M}})\boldsymbol{\varepsilon}(\omega)$. Since $\widetilde{\boldsymbol{P}}_{\mathbf{M}}$ is an orthogonal projection, then both $\boldsymbol{\varepsilon}_A$ and $\boldsymbol{\varepsilon}_B$ have null covariance matrices. But $\boldsymbol{\varepsilon}$ is Gaussian and hence $\boldsymbol{\varepsilon}_A$ and $\boldsymbol{\varepsilon}_B$ are independent. Therefore the events $A$ and $B$ are independent.

(b) Under the positivity condition, the subdifferential condition is equivalent to

$$J^*_{\mathbf{\Lambda}}(\boldsymbol{\pi})\leq 1. \qquad \text{(dual norm condition)}$$

Indeed, observe that $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$ (or equivalently, $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$) is necessary for the positivity condition. In view of (4.2.2), using the definition of $\boldsymbol{\pi}$, we see that $\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{\pi} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$ is equivalent to $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$. This follows from the fact that $\widetilde{\boldsymbol{P}}_{\mathbf{M}}$ is the projection matrix onto the vector subspace $\mathrm{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})$, and thus $\mathbf{0}' = [(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}})\widetilde{\boldsymbol{X}}_{\mathbf{M}}]' = \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}})$.

(c) The assertion of Theorem 4.3.1 cannot be strengthened. Indeed, when the SLOPE solution is not unique, the elements of $S_{\mathbf{X},\boldsymbol{\Lambda}}(\boldsymbol{Y})$ may have different SLOPE patterns.

Even if many theoretical properties on sign recovery by LASSO are known (see *e.g.* [183]), we believe that it is relevant to give a characterization of sign recovery by LASSO similar as the characterization of pattern recovery by SLOPE given in Theorem 4.3.1.
To show the similarity between proposed irrepresentability condition for SLOPE and the irrepresentability condition for LASSO [82, 192, 194], we propose analogous definitions to the clustered design matrix and parameter.

**Remark 4.3.3.** Let $\mathbf{0} \neq \boldsymbol{S} \in \{-1, 0, 1\}^p$ and $k = \|\boldsymbol{S}\|_1$ ($k$ is the number of nonzero components of $\boldsymbol{S}$). The signed matrix $\boldsymbol{U}_{\mathbf{S}} \in \mathbb{R}^{p\times k}$ is defined by $\boldsymbol{U}_{\mathbf{S}} = (\mathbf{diag}(\boldsymbol{S}))_{\mathrm{supp}(\mathbf{S})}$,
where $\mathbf{diag}(\boldsymbol{S}) \in \mathbb{R}^{p\times p}$ is a diagonal matrix and $(\mathbf{diag}(\boldsymbol{S}))_{\mathrm{supp}(\mathbf{S})}$ denotes the submatrix of $\mathbf{diag}(\boldsymbol{S})$ obtained by keeping columns corresponding to indices in $\mathrm{supp}(\boldsymbol{S})$. Observe that for any $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^p$ there exists a unique $\boldsymbol{S} \in \{-1, 0, 1\}^p$ and a unique $\boldsymbol{\kappa}_0 \in (0, \infty)^k$ such that $\boldsymbol{\beta} = \boldsymbol{U}_{\mathbf{S}}\boldsymbol{\kappa}_0$. Define the reduced matrix $\widetilde{\boldsymbol{X}}_{\mathbf{S}}$ and the reduced parameter $\widetilde{\lambda}_{\mathbf{S}}$ by

$$\widetilde{\boldsymbol{X}}_{\mathbf{S}} = \boldsymbol{X}\boldsymbol{U}_{\mathbf{S}} \text{ and } \widetilde{\lambda}_{\mathbf{S}} = \lambda\mathbf{1}_k, \text{ where } \mathbf{1}_k = (1, \ldots, 1)' \in \mathbb{R}^k.$$

Similarly as in the proof of Theorem 4.3.1, one may prove that the necessary and sufficient conditions for the LASSO sign recovery (*i.e.* existence of estimator $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ such that $\boldsymbol{sign}(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}) = \boldsymbol{sign}(\boldsymbol{\beta}) = \boldsymbol{S}$) are the following

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

In the noiseless case, when $\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$, the subdifferential condition reduces to $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{S}})^{+}\mathbf{1}_k \in \partial\|\cdot\|_1(\boldsymbol{S})$ (or equivalently, $\|\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{S}})^{+}\mathbf{1}_k\|_\infty \leq 1$ and $\mathbf{1}_k \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{S}})$). Moreover, if we have $\ker(\boldsymbol{X}_{\mathbf{S}}) = \{\mathbf{0}\}$, then $\mathbf{1}_k \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{S}})$ and therefore $\|\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{S}})^{+}\mathbf{1}_k\|_\infty \leq 1$ is equivalent to $\|\boldsymbol{X}'_{\overline{I}}\boldsymbol{X}_I(\boldsymbol{X}'_I\boldsymbol{X}_I)^{-1}\mathbf{S}_I\|_\infty \leq 1$, where $I = \mathrm{supp}(\mathbf{S})$, $\overline{I} = \{1, \ldots, p\} \setminus I$ and $\boldsymbol{X}_I$ (resp. $\boldsymbol{X}_{\overline{I}}$) denotes the submatrix of $\boldsymbol{X}$ obtained by keeping columns corresponding to indices in $I$ (resp $\overline{I}$). This latter expression is known as the irrepresentability condition [82, 192, 194].

From now on, in the definition of SLOPE (4.2.3), we consider the penalty term $J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$ (with a fixed $\boldsymbol{\Lambda} \in \mathbb{R}^{p+}$) to be multiplied by a scaling parameter $\alpha > 0$ and we denote the set of SLOPE solutions by $S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{Y})$.

### 4.3.1 SLOPE irrepresentability condition

As illustrated by Fuchs [82, Theorem 2], Bühlmann and van de Geer [41, Theorem 7.1] and also reminded in Remark 4.3.3, the irrepresentability condition is related to sign recovery by LASSO in the noiseless case. Analogously, studying pattern recovery by SLOPE in the noiseless case allows us to introduce the SLOPE irrepresentability condition. The latter condition will be very useful in the discussed later case when $\boldsymbol{\varepsilon}$ is no longer null. Corollary 4.3.1, which provides a characterization of pattern recovery by SLOPE in the noiseless case, is a consequence of Theorem 4.3.1.

**Definition 4.3.1.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{p+}$ and $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^p$. The noiseless pattern recovery by SLOPE is defined as*

$$\exists\, \alpha > 0 \;\; \exists\, \hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{X}\boldsymbol{\beta}) \text{ such that } \boldsymbol{patt}(\hat{\boldsymbol{\beta}}) = \boldsymbol{patt}(\boldsymbol{\beta}). \tag{4.3.2}$$

**Corollary 4.3.1.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{p+}$ and $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^p$. The noiseless pattern recovery by SLOPE is equivalent to*

$$J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) \leq 1 \text{ and } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \operatorname{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$$

*(or equivalently, to $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$). Moreover, under this condition, there exists $\alpha_0 > 0$ such that for all $\alpha \in (0, \alpha_0)$ there exists $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{X}\boldsymbol{\beta})$ for which $\boldsymbol{patt}(\hat{\boldsymbol{\beta}}) = \boldsymbol{patt}(\boldsymbol{\beta})$.*

Now we are ready to define the SLOPE irrepresentability condition.

**Definition 4.3.2.** *Let $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{\beta})$. We define the SLOPE irrepresentability condition as the following inequality and inclusion:*

$$J^*_{\boldsymbol{\Lambda}}\left(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\right) \leq 1 \text{ and } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \operatorname{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}). \tag{4.3.3}$$

**Remark 4.3.4.** If $\ker(\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \{\mathbf{0}\}$, then $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+ = \boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}})^{-1}$ and consequently the SLOPE irrepresentability condition reads $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) \leq 1$.

**Example 4.3.5.** Let $p = 2$, $\boldsymbol{\Lambda} = (4, 2)'$, $\boldsymbol{\beta} = (5, 0)'$ and $\bar{\boldsymbol{\beta}} = (5, 3)'$. Consider a design matrix $\boldsymbol{X} = (\boldsymbol{X}_1|\boldsymbol{X}_2) \in \mathbb{R}^{n\times 2}$ satisfying

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

- The SLOPE irrepresentability condition does not occur when $\boldsymbol{\beta} = (5, 0)'$. Indeed, $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{\beta}) = (1, 0)'$, $\widetilde{\boldsymbol{X}}_{\mathbf{M}} = \boldsymbol{X}_1$ (thus $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}} = 1$) and $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \lambda_1 = 4$. Therefore

$$J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) = J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) = J^*_{\boldsymbol{\Lambda}}(4\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = 6.4/6 > 1.$$

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

$$J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) = J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\Lambda}) = J^*_{\boldsymbol{\Lambda}}(\boldsymbol{\Lambda}) = 1 \leq 1.$$

Figure 4.2 corroborates graphically that SLOPE irrepresentability condition does not hold for $\boldsymbol{\beta}$, but it is satisfied for $\bar{\boldsymbol{\beta}}$. Note that, in this setup, the SLOPE solution is unique, since $\ker(\boldsymbol{X}) = \{\mathbf{0}\}$. By $\hat{\boldsymbol{\beta}}(\alpha)$ we denote the unique element of $S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{X}\boldsymbol{\beta})$. Then the SLOPE solution path is the function $(0, \infty) \ni \alpha \mapsto \hat{\boldsymbol{\beta}}(\alpha)$.

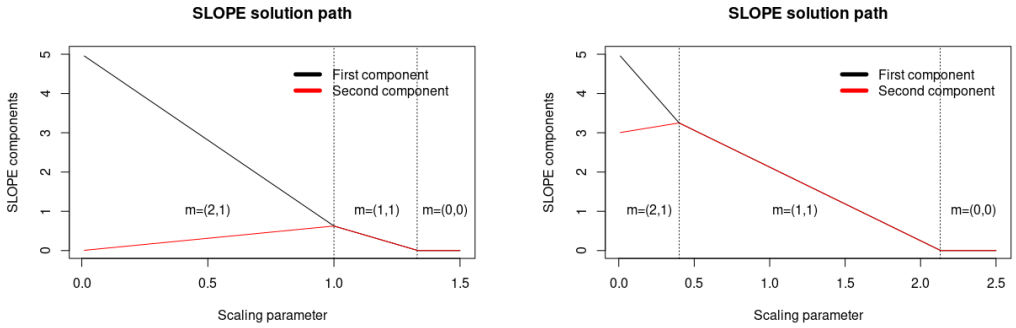



Figure 4.2: SLOPE solution paths for $p = 2$ and $\mathbf{\Lambda} = (4, 2)'$ in the noiseless case. $\boldsymbol{\beta}$ is equal respectively to $(5, 0)'$ (left panel) and $(5, 3)'$ (right panel). As may be seen, the pattern of $\boldsymbol{\beta} = (5, 0)'$ cannot be recovered in the noiseless case for any scaling of $\mathbf{\Lambda}$. The pattern of $\bar{\boldsymbol{\beta}} = (5, 3)'$ is recovered for $\alpha \in (0, 0.4)$.

## 4.4 Geometrical interpretation of Irrepresentability Condition

Let $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^p$. By (4.2.4), for a SLOPE minimizer $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\alpha\mathbf{\Lambda}}(\boldsymbol{X}\boldsymbol{\beta})$ the following occurs:

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

In addition, for $\boldsymbol{M} = \boldsymbol{patt}(\hat{\boldsymbol{\beta}})$, the following facts hold:

- $\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \in \mathrm{col}(\boldsymbol{U}_{\mathbf{M}})$, thus $\frac{1}{\alpha}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \in \boldsymbol{X}'\boldsymbol{X}\,\mathrm{col}(\boldsymbol{U}_{\mathbf{M}})$.
- $\partial J_{\mathbf{\Lambda}}(\hat{\boldsymbol{\beta}}) = \partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$.

Therefore, the noiseless pattern recovery by SLOPE implies that the vector space $\boldsymbol{X}'\boldsymbol{X}\mathrm{col}(\boldsymbol{U}_{\mathbf{M}}) = \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}})$ intersects $\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$. Actually, the vector $\bar{\mathbf{\Pi}} = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\tilde{\mathbf{\Lambda}}_{\mathbf{M}}$ appearing in Corollary 4.3.1 may be interpreted in a geometrical way as we propose below.

**Proposition 4.4.1.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\mathbf{0} \neq \boldsymbol{M} \in \mathcal{M}_p^{\mathrm{SLOPE}}$ and $\mathbf{\Lambda} \in \mathbb{R}^{p+}$. Recall that $\widetilde{\boldsymbol{X}}_{\mathbf{M}} = \boldsymbol{X}\boldsymbol{U}_{\mathbf{M}}$, $\tilde{\mathbf{\Lambda}}_{\mathbf{M}} = \boldsymbol{U}'_{|\mathbf{M}|_\downarrow}\mathbf{\Lambda}$ and $\bar{\mathbf{\Pi}} = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\tilde{\mathbf{\Lambda}}_{\mathbf{M}}$. Then:*

*i) If $\tilde{\mathbf{\Lambda}}_{\mathbf{M}} \notin \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, then $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})) \cap \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \varnothing$.*

*ii) If $\tilde{\mathbf{\Lambda}}_{\mathbf{M}} \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, then $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})) \cap \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \{\bar{\mathbf{\Pi}}\}$.*

*iii) Pattern recovery by SLOPE for $\boldsymbol{\varepsilon} = \mathbf{0}$ is equivalent to*

$$\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) \cap \partial J_{\mathbf{\Lambda}}(\boldsymbol{M}) \neq \varnothing.$$

In other words, the accessibility condition means that the intersection of $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) \cap \mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M}))$ is not empty. Moreover, then it is equal to the vector $\bar{\mathbf{\Pi}} = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\tilde{\mathbf{\Lambda}}_{\mathbf{M}}$. Moreover, when the accessibility condition holds, by Proposition 4.4.1 iii) the noiseless recovery of the SLOPE pattern is equivalent to the subdifferential condition $\mathbf{\Pi} \in \partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$.

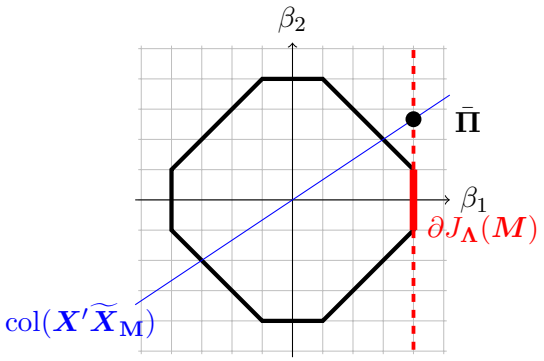



Figure 4.3: Intersection $\bar{\mathbf{\Pi}}$ of $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \mathrm{col}((1, 2/3)')$ and $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M}))$ for $\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 2/3 \\ 2/3 & 1 \end{pmatrix}$, $\mathbf{\Lambda} = (4, 1)'$ and $\boldsymbol{patt}(\boldsymbol{\beta}) = \boldsymbol{M} = (1, 0)'$. Since $\bar{\mathbf{\Pi}} = (4, 8/3)' \notin \partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$, then in the noiseless case, by Proposition 4.4.1, SLOPE does not recover $\boldsymbol{M}$. However, as $\bar{\mathbf{\Pi}}$ exists, the pattern $\boldsymbol{M}$ is accessible, as in the latter examples.

*Proof.* i) Recall that, according to Lemma 4.9.2, $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})) = \{\boldsymbol{v} \in \mathbb{R}^p \colon \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \tilde{\mathbf{\Lambda}}_{\mathbf{M}}\}$. If $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})) \cap \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) \neq \varnothing$, then there exists $\boldsymbol{z} \in \mathbb{R}^k$, where $k = \|\boldsymbol{M}\|_\infty$, such that $\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{z} \in \mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M}))$. Consequently, $\tilde{\mathbf{\Lambda}}_{\mathbf{M}} = \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{z} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{z}$, thus $\tilde{\mathbf{\Lambda}}_{\mathbf{M}} \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$.
ii) If $\tilde{\mathbf{\Lambda}}_{\mathbf{M}} \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, then $\bar{\mathbf{\Pi}} \in \mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M}))$. Indeed, since $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+$ is the projection onto $\mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, we have

$$\boldsymbol{U}'_{\mathbf{M}}\bar{\mathbf{\Pi}} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\tilde{\mathbf{\Lambda}}_{\mathbf{M}} = \tilde{\mathbf{\Lambda}}_{\mathbf{M}}.$$

Moreover, since $\mathrm{col}((\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+) = \mathrm{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})$, we deduce that $\bar{\mathbf{\Pi}} \in \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}})$. To prove that $\bar{\mathbf{\Pi}}$ is the unique point in the intersection, we will show that $\mathrm{col}(\boldsymbol{X}'\widetilde{X}_{\mathbf{M}}) \cap \mathrm{col}(\boldsymbol{U}_{\mathbf{M}})^\perp = \{\mathbf{0}\}$. Indeed, if $\boldsymbol{v} \in \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) \cap \mathrm{col}(\boldsymbol{U}_{\mathbf{M}})^\perp$, then there exists such $\boldsymbol{z} \in \mathbb{R}^k$ that $\boldsymbol{v} = \boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{z} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_{\mathbf{M}}\boldsymbol{z}$ and $(\boldsymbol{U}_{\mathbf{M}}\boldsymbol{z})'\boldsymbol{v} = \mathbf{0}$. Therefore, $(\boldsymbol{U}_{\mathbf{M}}\boldsymbol{z})'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{U}_{\mathbf{M}}\boldsymbol{z} = \mathbf{0}$, which implies that $\boldsymbol{X}\boldsymbol{U}_{\mathbf{M}}\boldsymbol{z} = \widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{z} = \mathbf{0}$ and thus $\boldsymbol{v} = \mathbf{0}$. Finally, if $\mathbf{\Pi} \in \mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M})) \cap \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}})$, then $\mathbf{\Pi} - \bar{\mathbf{\Pi}} \in \mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}})$ and $\boldsymbol{U}'_{\mathbf{M}}(\mathbf{\Pi} - \bar{\mathbf{\Pi}}) = \mathbf{0}$, which implies that $\mathbf{\Pi} = \bar{\mathbf{\Pi}}$.

iii) According to Corollary 4.3.1, pattern recovery by SLOPE in the noiseless case is equivalent to $\bar{\mathbf{\Pi}} \in \partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$ which is equivalent, by i) and ii), to $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) \cap \partial J_{\mathbf{\Lambda}}(\boldsymbol{M}) \neq \varnothing$. □

Proposition 4.4.1 may be interpreted with a notion of the accessibility condition for SLOPE, which was introduced in [156].

**Definition 4.4.1** (Accessible pattern)**.** *[156] Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and* pen *be a polyhedral gauge. We say that the pattern of $\boldsymbol{\beta} \in \mathbb{R}^p$ is accessible with respect to $\boldsymbol{X}$ and* $\lambda$pen*, if there exist $\boldsymbol{y} \in \mathbb{R}^n$ and $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y})$ such that $\widehat{\boldsymbol{\beta}} \overset{\mathrm{pen}}{\sim} \boldsymbol{\beta}$.*

The accessibility of a pattern can be characterized in a geometric and an analytic way.

**Proposition 4.4.2** (Characterization of accessible patterns)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$.*

*(a) Geometric characterization: The SLOPE pattern of $\boldsymbol{\beta} \in \mathbb{R}^p$ is accessible with respect to $\boldsymbol{X}$ if and only if*

$$\mathrm{row}(\boldsymbol{X}) \cap \partial\mathrm{pen}(\boldsymbol{\beta}) \neq \varnothing.$$

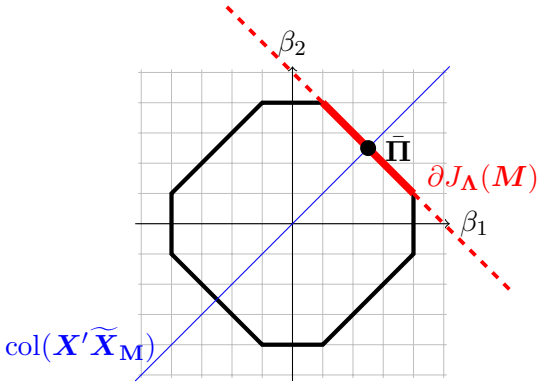



Figure 4.4: Intersection $\bar{\mathbf{\Pi}}$ of $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \mathrm{col}((1,1)')$ and $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M}))$ for $\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 2/3 \\ 2/3 & 1 \end{pmatrix}$, $\mathbf{\Lambda} = (4,1)'$ and $\boldsymbol{patt}(\boldsymbol{\beta}) = \boldsymbol{M} = (1,1)'$. Since $\bar{\mathbf{\Pi}} = (2.5, 2.5)' \in \partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$, then in the noiseless case, by Proposition 4.4.1, SLOPE recovers the pattern $\boldsymbol{M}$ of $\boldsymbol{\beta}$.

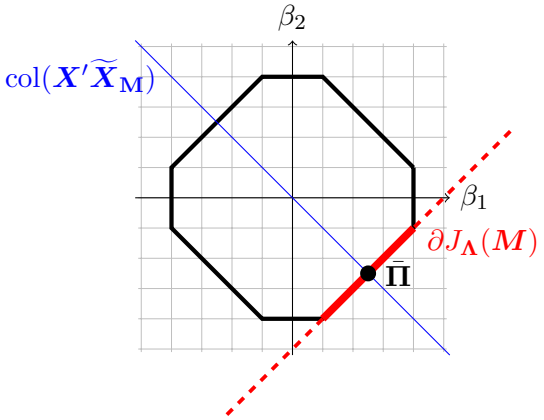



Figure 4.5: Intersection $\bar{\mathbf{\Pi}}$ of $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \mathrm{col}((1,-1)')$ and $\mathrm{aff}(\partial J_{\mathbf{\Lambda}}(\boldsymbol{M}))$ for $\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 2/3 \\ 2/3 & 1 \end{pmatrix}$, $\mathbf{\Lambda} = (4,1)'$ and $\boldsymbol{patt}(\boldsymbol{\beta}) = \boldsymbol{M} = (1,-1)'$. Since $\bar{\mathbf{\Pi}} \notin \partial J_{\mathbf{\Lambda}}(\boldsymbol{M})$, then in the noiseless case, by Proposition 4.4.1, SLOPE recovers the pattern $\boldsymbol{M}$ of $\boldsymbol{\beta}$.

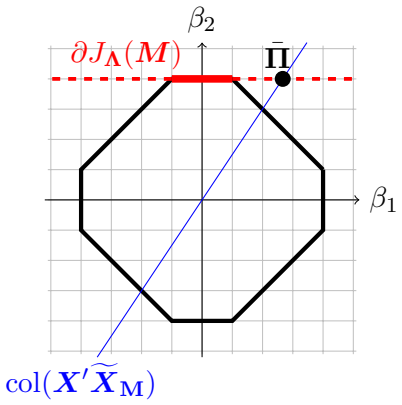



Figure 4.6: Intersection $\bar{\boldsymbol{\Pi}}$ of $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \mathrm{col}((2/3,1)')$ and $\mathrm{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$ for $\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 2/3 \\ 2/3 & 1 \end{pmatrix}$, $\boldsymbol{\Lambda} = (4,1)'$ and $\boldsymbol{patt}(\boldsymbol{\beta}) = \boldsymbol{M} = (0,1)'$. Since $\bar{\boldsymbol{\Pi}} = (8/3,4)' \notin \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$, then in the noiseless case, by Proposition 4.4.1, SLOPE does not recover $\boldsymbol{M}$.

*(b) Analytic characterization: The SLOPE pattern of $\boldsymbol{\beta} \in \mathbb{R}^p$ is accessible with respect to $\boldsymbol{X}$ if and only if for any $\boldsymbol{b} \in \mathbb{R}^p$ we have*

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{b} \implies \mathrm{pen}(\boldsymbol{\beta}) \leq \mathrm{pen}(\boldsymbol{b}).$$

For $p = 2, n \geq 2$, a full rank matrix $\boldsymbol{X}$ and the models with zeros or with 2-element clusters, the Proposition 4.4.1 and Corollary 4.3.1 are illustrated on Figures 4.3 – 4.6. They present the SLOPE dual unit ball $C_{\boldsymbol{\Lambda}}$, linear spaces $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}})$ (blue lines) and, if existing, their intersection points $\bar{\boldsymbol{\Pi}} = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$ (black dots) with the affine spaces $\mathrm{aff}(F_{\mathbf{M}})$ (red dashed lines) of the corresponding pattern faces $F_{\mathbf{M}} = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ (red segments). The existence of $\bar{\boldsymbol{\Pi}}$ is equivalent to the accessibility condition. The SLOPE irrepresentability condition holds if and only if $\bar{\boldsymbol{\Pi}} \in F_{\mathbf{M}}$. It holds true in Figures 4.4 and 4.5 and fails in Figures 4.3 and 4.6.

In the low-dimensional setting, when the design matrix $\boldsymbol{X}$ is of a full column rank, another geometrical interpretation of $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ might be used. As a corollary of Theorem 3.2.2, one may deduce that the difference between $\hat{\boldsymbol{\beta}}^{SLOPE}$ and $\hat{\boldsymbol{\beta}}^{OLS}$ is equal to

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\pi}^* \in (\boldsymbol{X}'\boldsymbol{X})^{-1}C_{\boldsymbol{\Lambda}},$$

where $C_{\boldsymbol{\Lambda}}$ is the unit ball in $J^*_{\boldsymbol{\Lambda}}$ and $\boldsymbol{\pi}^*$ is defined as in Lemma 3.2.1. On Figure 4.7 we present graphically this relation between $\hat{\boldsymbol{\beta}}^{SLOPE}$ and $\hat{\boldsymbol{\beta}}^{OLS}$.

The illustration of the SLOPE dual norm unit ball $C_{\boldsymbol{\Lambda}}$ in $\mathbb{R}^p$ for $p = 3$ may be found in the recently published article of Schneider and Tardivel [156]. We recall it to present that with growing number of explanatory variables $p$ the complexity of $C_{\boldsymbol{\Lambda}}$ rises drastically.

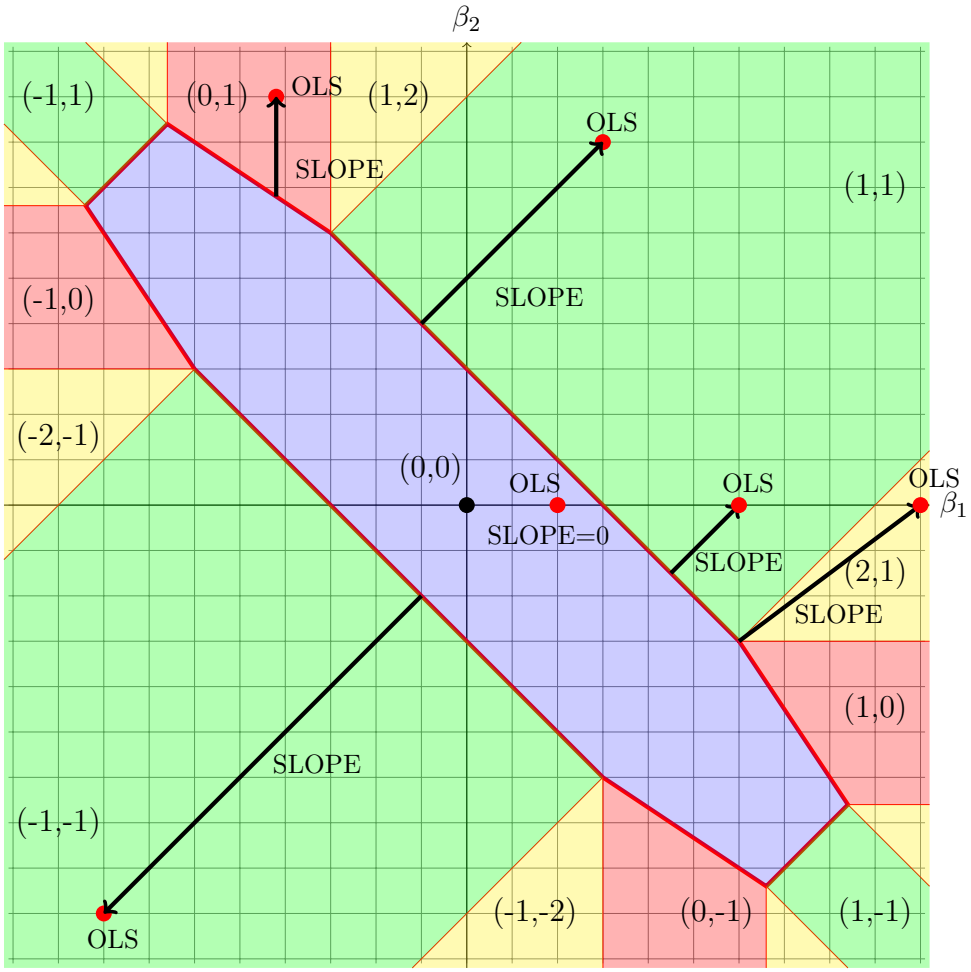



Figure 4.7: Comparison of $\hat{\beta}^{SLOPE}$, $\hat{\beta}^{OLS}$ and their difference $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\varepsilon} \in (\boldsymbol{X}'\boldsymbol{X})^{-1}C_{\boldsymbol{\Lambda}}$ for $p = 2$, $\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 2/3 \\ 2/3 & 1 \end{bmatrix}$ and $\Lambda = (4, 1)'$.

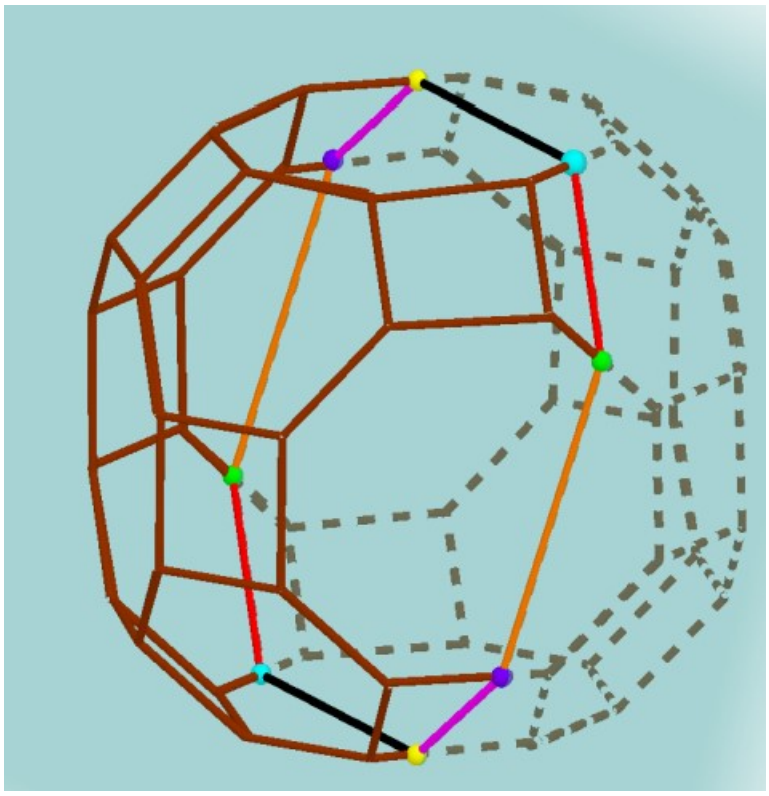


Figure 4.8: $C_{\boldsymbol{\Lambda}}$ in $\mathbb{R}^p$ for $p = 3$ with its two-dimensional subset corresponding to the intersection of $C_{\boldsymbol{\Lambda}}$ with row($\boldsymbol{X}$) from [156, Example 4]. One may observe that in this example there are only 17 accessible patterns out of all 147 patterns for $p = 3$.

While $C_{\boldsymbol{\Lambda}}$ consists of 17 faces for $p = 2$, for $p = 3$ the number of faces increases to 147. For larger values of $p$, the exact number of faces of $C_{\boldsymbol{\Lambda}}$ may be found at the On-line Encyclopedia of Integer Sequences [138, A080253]. The growth of the number of faces is super-exponential with respect to $p$.

## 4.5 Asymptotics of pattern recovery and pattern consistency

In this section we consider two asymptotic scenarios and establish conditions on tuning parameters for which the pattern of $\boldsymbol{\beta}$ is recovered. In Section 4.5.1 we consider the case where gaps between distinct absolute values of $\boldsymbol{\beta}$ diverge and in Section 4.5.2 the case where the sample size $n$ diverges. The proofs rely on Theorem 4.3.1. 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 $\alpha\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{p+}$ is fixed and $\alpha$ varies. We determine the assumptions for the sequence $(\alpha)$ for which both positivity and subdifferential conditions hold true, *i.e.* for which the pattern is recovered.

### 4.5.1 X is a fixed matrix

The subdifferential condition, given in Theorem 4.3.1, says that a vector $\boldsymbol{\pi}$ defined in (4.3.1) belongs to $\partial J_{\alpha\boldsymbol{\Lambda}}(\boldsymbol{M})$, where $\alpha$ is a scaling parameter. This condition is equivalent to requiring

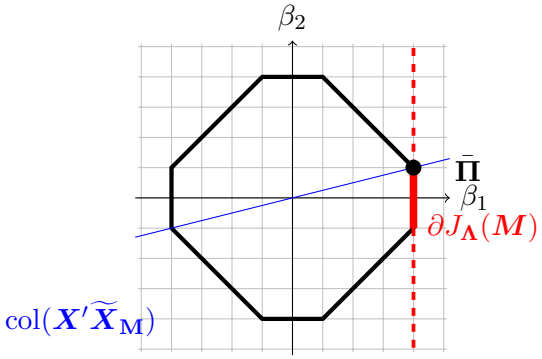



Figure 4.9: Intersection $\bar{\boldsymbol{\Pi}}$ of $\mathrm{col}(\boldsymbol{X}'\widetilde{\boldsymbol{X}}_{\mathbf{M}}) = \mathrm{col}((1,1/4)')$ and $\mathrm{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$ for $\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}$, $\boldsymbol{\Lambda} = (4,1)'$ and $\boldsymbol{patt}(\boldsymbol{\beta}) = \boldsymbol{M} = (1,0)'$. Since $\bar{\boldsymbol{\Pi}} = (4,1)' \in F_{\mathbf{M}} = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$, but it lies on a boundary of $F_{\mathbf{M}}$. Therefore the open SLOPE irrepresentability condition does not hold, even though the irrepresentability condition is satisfied.

that a vector $\boldsymbol{\pi}/\alpha$ belongs to $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$. We denote the vector $\boldsymbol{\pi}/\alpha$ by

$$\boldsymbol{\pi}_\alpha = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\tilde{\boldsymbol{\Lambda}}_{\mathbf{M}} + \frac{1}{\alpha}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}})\boldsymbol{Y} = \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\tilde{\boldsymbol{\Lambda}}_{\mathbf{M}} + \frac{1}{\alpha}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{P}_{\mathbf{M}})\boldsymbol{\varepsilon}, \tag{4.5.1}$$

where in the latter equality we use the fact that $(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}})$ is an orthogonal projection onto $\mathrm{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})^\perp$ and therefore $(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}})\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}})\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{s} = \boldsymbol{0}$, where $\boldsymbol{\beta} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}$ and $\boldsymbol{s} \in \mathbb{R}^{\|\mathbf{M}\|_\infty +}$.
By Theorem 4.3.1, the probability of the pattern recovery by SLOPE is upper bounded by

$$\mathbb{P}\left(\exists \hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{Y}) \text{ such that } \boldsymbol{patt}(\hat{\boldsymbol{\beta}}) = \boldsymbol{patt}(\boldsymbol{\beta})\right) \leq \begin{cases} \mathbb{P}\left(J^*_{\boldsymbol{\Lambda}}(\boldsymbol{\pi}_\alpha) \leq 1\right), \\ 0 \text{ if } \tilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \notin \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}). \end{cases} \tag{4.5.2}$$

The first point in Theorem 4.5.1 shows that the probability of pattern recovery matches with the upper bound (4.5.2) when gaps between different absolute values of terms of $\boldsymbol{\beta}$ are large enough. The last point provides the pattern consistency by SLOPE.
Before stating this Theorem, we introduce the open SLOPE irrepresentability condition.

**Definition 4.5.1** (Open SLOPE irrepresentability condition)**.**

$$\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\tilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in ri\left(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})\right).$$

We illustrate the difference between the SLOPE irrepresentability condition and its open version on Figure 4.9.

**Theorem 4.5.1.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{0} \neq \boldsymbol{M} \in \mathcal{M}_p^{\mathrm{SLOPE}}$, and $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)' \in \mathbb{R}^{p+}$. Consider a sequence of signals $(\boldsymbol{\beta}^{(r)})_{r\geq 1}$ with pattern $\boldsymbol{M}$:*

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

*whose strength is increasing in the following sense:*

$$\Delta_r = \min_{1\leq i<k}\left(s_i^{(r)} - s_{i+1}^{(r)}\right) \overset{r\to\infty}{\longrightarrow} \infty, \textit{ with the convention } s_{k+1}^{(r)} = 0$$

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

(a) *Sharpness of the upper bound: Let $\alpha > 0$. If $\boldsymbol{\varepsilon}$ is random, then the upper bound* (4.5.2) *is asymptotically reached:*

$$\lim_{r\to\infty} \mathbb{P}\left(\exists \hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{Y}^{(r)}) \text{ such that } \boldsymbol{patt}(\hat{\boldsymbol{\beta}}) = \boldsymbol{M}\right) = \begin{cases} \mathbb{P}\left(J^*_{\boldsymbol{\Lambda}}(\boldsymbol{\pi}_\alpha) \le 1\right), \\ 0 \text{ if } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \notin \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}). \end{cases}$$

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

$$\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{ri}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})),$$

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

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

**Remark 4.5.2.**

(a) The open SLOPE irrepresentability condition $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{ri}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$ is equivalent to the following computationally verifiable conditions:

$$\begin{cases} J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) \le 1 \text{ and } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}), \\ \left|\left\{i \in \{1,\dots,p\}\colon \;\; \sum_{j=1}^i |\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}|_{(j)} = \sum_{j=1}^i \lambda_j\right\}\right| = \|\boldsymbol{M}\|_\infty. \end{cases}$$

This equivalence follows from Proposition 4.9.1.

(b) Let us assume that the distribution of $\boldsymbol{\varepsilon}$ and $-\boldsymbol{\varepsilon}$ is equal. Because the unit ball of the dual sorted $\ell_1$ norm is convex, when $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) > 1$ then, independently on $\alpha > 0$, the probability of pattern recovery is smaller than $1/2$, namely

$$\mathbb{P}\left(\exists \hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\alpha\boldsymbol{\Lambda}}(\boldsymbol{Y}) \text{ such that } \boldsymbol{patt}(\hat{\boldsymbol{\beta}}) = \boldsymbol{M}\right) \le 1/2.$$

For LASSO, a similar inequality on the probability of sign recovery is given in [183].

(c) In Section 4.7, 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 [174]).

### 4.5.2 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 $n \ge p$ we consider a linear regression problem

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \tag{4.5.3}$$

where $\boldsymbol{X}_n \in \mathbb{R}^{n\times p}$ is a random design matrix. We will use the following assumptions:

A. $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \dots, \varepsilon_n)'$, where $(\varepsilon_i)_i$ are i.i.d. centered with finite variance.

B1. A sequence of design matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots$ satisfies the condition

$$\frac{1}{n}\boldsymbol{X}'_n\boldsymbol{X}_n \xrightarrow{\mathbb{P}} \boldsymbol{C}, \tag{4.5.4}$$

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

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

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

C. $(\boldsymbol{X}_n)_n$ and $(\varepsilon_n)_n$ are independent.

We will consider a sequence of tuning parameters $(\boldsymbol{\Lambda}_n)_n$ defined by

$$\boldsymbol{\Lambda}_n = \alpha_n \boldsymbol{\Lambda},$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{p+}$ is fixed and $(\alpha_n)_n$ is a sequence of positive numbers.
Let $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ be an element of the set $S_{\mathbf{X}_n, \boldsymbol{\Lambda}_n}(\boldsymbol{Y}_n)$ of SLOPE minimizers. Under Assumption B1, for large $n$ with high probability, the set $S_{\mathbf{X}_n, \boldsymbol{\Lambda}_n}(\boldsymbol{Y}_n)$ consists of one element. Indeed, we have

$$\mathbb{P}\left(\ker(\boldsymbol{X}_n) = \{\mathbf{0}\}\right) = \mathbb{P}\left(\boldsymbol{X}_n' \boldsymbol{X}_n \text{ is positive definite}\right) \overset{n\to\infty}{\longrightarrow} 1$$

and $\ker(\boldsymbol{X}_n) = \{\mathbf{0}\}$ ensures existence of the unique SLOPE minimizer. In a natural setting, the strong consistency of $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ can be characterized in terms of behaviour of the tuning parameter. We use such approach in the Theorem 3.4.1 and Theorem 4.6.2. At this point we note that if (4.6.8) holds almost surely, then the condition $\alpha_n/n \to 0$ ensures that $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}$. Thus, if $\boldsymbol{\beta}$ does not have any non-trivial clusters nor zeros, *i.e.* $\|\boldsymbol{patt}(\boldsymbol{\beta})\|_\infty = p$, then $\alpha_n/n \to 0$ suffices for $\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) \xrightarrow{a.s.} \boldsymbol{patt}(\boldsymbol{\beta})$. However, if $\|\boldsymbol{patt}(\boldsymbol{\beta})\| < p$, 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 Assumption B2 is not necessary for the SLOPE pattern recovery. This assumption was introduced to ensure the existence of a Gaussian vector in the Theorem 4.5.3 (a).

**Theorem 4.5.3.** *Under the assumptions A, B1, C, the following statements hold true.*

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

$$\lim_{n\to\infty} \mathbb{P}\left(\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) = \boldsymbol{patt}(\boldsymbol{\beta})\right) = \mathbb{P}\left(J_{\boldsymbol{\Lambda}}^*(\boldsymbol{V}) \le 1\right),$$

*where $\boldsymbol{V} \sim \mathrm{N}(\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}, \sigma^2[\boldsymbol{C} - \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}])$.*

*(b) Let*

$$\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{ri}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})). \tag{4.5.5}$$

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

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

*if and only if*

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

*(c) The condition*

$$J_{\boldsymbol{\Lambda}}^*\left(\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\right) \le 1 \tag{4.5.6}$$

*is necessary for pattern consistency of SLOPE estimator.*

The random vector $\boldsymbol{V}$ belongs to the smallest affine space containing $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, *i.e.* $\operatorname{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})) = \{\boldsymbol{v} \in \mathbb{R}^p \colon \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\}$, see Lemma 4.9.2.
Condition (4.5.5) is the open SLOPE irrepresentability condition in the $n \to \infty$ regime. The above result should be compared with [192, Theorem 1], where the same conditions on the LASSO tuning parameter ensure consistency of sign recovery by LASSO estimator. Below we make one step further and consider the strong consistency of SLOPE pattern recovery by $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$. Although this was not the main focus of Zhao and Yu, it can be deduced from [192, Theorem 1] that if for $c \in (0,1)$ the LASSO tuning parameter $\lambda_n$ 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 $\operatorname{sign}(\hat{\boldsymbol{\beta}}_n^{\mathrm{LASSO}}) \xrightarrow{a.s.} \operatorname{sign}(\boldsymbol{\beta})$. 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. Below we show that if $\alpha_n/n \to 0$ and $\alpha_n/\sqrt{n} \to \infty$, then $\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})$ is not strongly consistent and one actually has to impose a slightly stronger condition (4.5.7).
For the purpose of the a.s. convergence, we strengthen the assumption on design matrices:

B'. Assume that the rows of $\boldsymbol{X}_n$ are independent and that each row of $\boldsymbol{X}_n$ has the same law as $\boldsymbol{\Xi}$, where $\boldsymbol{\Xi}$ is a random vector whose components are linearly independent almost surely and that $\mathbb{E}[\boldsymbol{\Xi}_i^2] < \infty$ for $i = 1, \dots, p$.

**Remark 4.5.4.** Under B', by the strong law of large numbers, we have $n^{-1}\boldsymbol{X}'_n\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{C}$, where $\boldsymbol{C} = (C_{ij})_{ij}$ with $C_{ij} = \mathbb{E}[\boldsymbol{\Xi}_i\boldsymbol{\Xi}_j]$. Moreover, $\boldsymbol{C}$ is positive definite if and only if the random variables $(\boldsymbol{\Xi}_1, \dots, \boldsymbol{\Xi}_p)$ are linearly independent almost surely. Indeed, for $\boldsymbol{t} \in \mathbb{R}^p$ we have $\boldsymbol{t}'\boldsymbol{C}\boldsymbol{t} = \mathbb{E}[(\sum_{i=1}^p t_i\boldsymbol{\Xi}_i)^2] > 0$ if and only if $\sum_{i=1}^p t_i\boldsymbol{\Xi}_i \neq 0$ a.s. for all $\boldsymbol{t} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$.
Since B' ensures that (4.6.8) holds a.s., it also implies that for large $n$, there almost surely exists a unique SLOPE minimizer. We denote this element by $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$.

**Theorem 4.5.5.** *Under A, B' and C, assume that a sequence $(\alpha_n)_n$ satisfies*

$$\lim_{n\to\infty} \frac{\alpha_n}{n} = 0 \qquad \textit{and} \qquad \lim_{n\to\infty} \frac{\alpha_n}{\sqrt{n \log\log n}} = \infty. \tag{4.5.7}$$

*If (4.5.5) holds, then the sequence $(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})_n$ recovers almost surely the pattern of $\boldsymbol{\beta}$ asymptotically, i.e.*

$$\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) \xrightarrow{a.s.} \boldsymbol{patt}(\boldsymbol{\beta}). \tag{4.5.8}$$

**Remark 4.5.6.** Assume that (4.5.5) is satisfied and set $\alpha_n = c\sqrt{n \log\log n}$ for $c > 0$. Then (4.5.7) is not satisfied and the probability that the correct SLOPE pattern is not recovered is greater than zero. In Section 4.6.1 we present more refined results on the strong consistency of the SLOPE pattern. The $\log\log n$ correction in (4.5.7) comes from the law of iterated logarithm.

## 4.6 Strong consistency of SLOPE and its pattern

### 4.6.1 Refined results on strong consistency of the SLOPE pattern

In this section we aim to give assumptions on the design matrix that are weaker than condition B', but they ensure the almost sure convergence of the pattern of $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$.

A'. $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \dots, \varepsilon_n)'$, where $(\varepsilon_i)_i$ are independent random variables such that

$$\mathbb{E}[\varepsilon_n] = 0 \quad \text{and} \quad \operatorname{Var}(\varepsilon_n) = \sigma^2 \quad \text{for all } n, \quad \text{and } \sup_n \mathbb{E}[|\varepsilon_n|^r] < \infty \tag{4.6.1}$$

for some $r > 2$.

B". A sequence of design matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ satisfies the condition

$$\frac{1}{n}\boldsymbol{X}'_n\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{C}, \tag{4.6.2}$$

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

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

$$\lim_{n\to\infty} \frac{(\log n)^\rho}{\sqrt{n}} \sup_{i,j} \left|X_{ij}^{(n)}\right| = 0 \quad \text{a.s. for all } \rho > 0 \tag{4.6.3}$$

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

$$\sup_j \left[\sum_{i=1}^{m} \left(X_{ij}^{(n)} - X_{ij}^{(m)}\right)^2 + \sum_{i=m+1}^{n} \left(X_{ij}^{(n)}\right)^2\right] \leq \left(\sum_{i=m+1}^{n} c_i\right)^d \quad \text{a.s.,} \tag{4.6.4}$$

$$\left(\sum_{i=m_0}^{n} c_i\right)^d = O(n) \quad \text{a.s.} \tag{4.6.5}$$

C. $(\boldsymbol{X}_n)_n$ and $(\varepsilon_n)_n$ are independent.

We note that conditions (4.6.3) and (4.6.4) are trivially satisfied in the i.i.d. rows setting of Remark 4.5.4 or Assumption B'. The main ingredient of the proof of the strong pattern consistency is the law of iterated logarithm (4.9.9) which holds trivially under B'. Below we establish the same result under more general B". The technical assumption (4.6.4) is a kind of weak continuity assumption on the rows of $\boldsymbol{X}_n$ as it says that the $\ell_2$ distance between $j^{th}$ rows of $\boldsymbol{X}_n$ and $\boldsymbol{X}_m$ should not be too large.

**Lemma 4.6.1.** *Assume A', B" and C. Then*

$$\limsup_{n\to\infty} \frac{\|\boldsymbol{X}'_n\boldsymbol{\varepsilon}_n\|_\infty}{\sqrt{n\log\log n}} < \infty \qquad a.s. \tag{4.6.6}$$

*Proof.* In view of (4.6.8) we have for $j = 1, \ldots, p$,

$$n^{-1}A_n^{(j)} := n^{-1}\sum_{i=1}^{n}\left(X_{ij}^{(n)}\right)^2 = \left(n^{-1}\boldsymbol{X}'_n\boldsymbol{X}_n\right)_{jj} \xrightarrow{a.s.} C_{jj} > 0. \tag{4.6.7}$$

We apply the general law of iterated logarithm for weights forming a triangular array from [117]. The result follows directly from [117, Theorem 1], which we recall in Theorem 2.3.18. Defining $a_{ni}^{(j)} := X_{ij}^{(n)}$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$, $n \geq 1$ and 0 otherwise, we have

$$(\boldsymbol{X}'_n\boldsymbol{\varepsilon}_n)_j = \sum_{i=-\infty}^{\infty} a_{ni}^{(j)}\varepsilon_i$$

and therefore we fall within the framework of (2.3.5). Then, (4.6.1), (4.6.3), (4.6.4) and (4.6.5) coincide with (2.3.4), (2.3.6), (2.3.7) and (2.3.8), respectively. Let $\mathbb{P}(\,\cdot\,|(\boldsymbol{X}_n)_n)$ be a regular

conditional probability. Then, applying Theorem 2.3.18 on the probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\boldsymbol{X}})$ to our sequence, we obtain that for $j = 1, \ldots, n$,

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

Averaging over $(\boldsymbol{X}_n)_n$ and using (4.6.7) again, we obtain the assertion. □

**Theorem 4.6.1.** *Assume A', B" and C. Suppose that $(\alpha_n)_n$ satisfies*

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

*If* (4.5.5) *is satisfied, then* $\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) \xrightarrow{a.s.} \boldsymbol{patt}(\boldsymbol{\beta})$.

Comments:

(a) Under reasonable assumptions (see *e.g.* [117, Theorem 1 (iii)]), one can show that

$$\limsup_{n\to\infty} \frac{\|\boldsymbol{X}'_n\boldsymbol{\varepsilon}_n\|_\infty}{\sqrt{n\log\log n}} > 0 \qquad \text{a.s.}$$

Since $\alpha_n^{-1}\boldsymbol{X}'_n\boldsymbol{\varepsilon}_n \xrightarrow{a.s.} \boldsymbol{0}$ 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 $\boldsymbol{\varepsilon}_n$. For example, if $\boldsymbol{\varepsilon}_n$ is Gaussian, then one can use results from [167]. We note that [167] offers a very similar result as [117], but their assumptions are not quite comparable, see [167, 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 $\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\varepsilon}_{n+1}$, *i.e.* the error need not be incremental. For orthogonal design such approach was taken in [165]. It is proved there that one obtains the a.s. SLOPE pattern consistency with the second limit condition of Theorem 4.6.1 replaced by $\lim_{n\to\infty} \alpha_n/\sqrt{n\log n} = \infty$. This result can be generalized to non-orthogonal designs.

### 4.6.2 Strong consistency of the SLOPE estimator

**Lemma 4.6.2.** *Let $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ with $\varepsilon_i$ i.i.d., centered and having finite variance. Suppose that*

$$\frac{1}{n}\boldsymbol{X}'_n\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{C} > 0. \tag{4.6.8}$$

*and that $(\boldsymbol{\varepsilon}_n)_n$ and $(\boldsymbol{X}_n)_n$ are independent. Then $n^{-1}\boldsymbol{X}'_n\boldsymbol{\varepsilon}_n \xrightarrow{a.s.} 0$.*

*Proof.* Denote by $\mathbb{P}(\,\cdot \mid (\boldsymbol{X}_n)_n)$ the regular conditional probability. By [52, Th. 1.1] applied to the sequence $(n^{-1}\boldsymbol{X}'_n\boldsymbol{\varepsilon}_n)_j$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P}(\,\cdot \mid (\boldsymbol{X}_n)_n))$, we obtain

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

Thus, applying the expectation to both sides above we obtain the assertion. □

**Theorem 4.6.2.** *Assume that* $\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n$, *where* $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ *with* $\varepsilon_i$ *i.i.d., centered and having finite variance. Suppose* (4.6.8) *and that* $(\boldsymbol{\varepsilon}_n)_n$ *and* $(\boldsymbol{X}_n)_n$ *are independent. Let* $\boldsymbol{\Lambda}_n = (\lambda_1^{(n)}, \ldots, \lambda_p^{(n)})'$. *Then, for large* $n$, $S_{\mathbf{X}_n, \boldsymbol{\Lambda}_n}(\boldsymbol{Y}_n) = \{\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}\}$ *almost surely.*
*If* $\boldsymbol{\beta} \neq \mathbf{0}$, *then*

$$\left(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}\right) \Longleftrightarrow \left(\lim_{n\to\infty} \frac{\lambda_1^{(n)}}{n} = 0\right). \tag{4.6.9}$$

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

*Proof of Theorem 4.6.2.* The assumption (4.6.8) implies that the matrix $\boldsymbol{X}_n'\boldsymbol{X}_n$ is positive definite for large $n$ almost surely and hence ensures that $\ker(\boldsymbol{X}_n) = \{\mathbf{0}\}$. It implies the uniqueness of the SLOPE solution.
By Proposition 4.2.1, $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}$ is the SLOPE estimator of $\boldsymbol{\beta}$ in a linear regression model $\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n$ if and only if for $\boldsymbol{\pi}_n = \boldsymbol{X}_n'(\boldsymbol{Y}_n - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})$ we have

$$J_{\boldsymbol{\Lambda}}^*(\boldsymbol{\pi}_n) \leq 1 \tag{4.6.10}$$

and

$$\boldsymbol{U}_{\mathbf{M}_n}'\boldsymbol{\pi}_n = \widetilde{\boldsymbol{\Lambda}}_n, \tag{4.6.11}$$

where $\boldsymbol{M}_n = \boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}})$ and $\widetilde{\boldsymbol{\Lambda}}_n = \boldsymbol{U}'_{|\boldsymbol{M}_n|\downarrow}\boldsymbol{\Lambda}_n$. By the definition of $\boldsymbol{\pi}_n$ we have

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{Y}_n - (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{\pi}_n = \hat{\boldsymbol{\beta}}_n^{\mathrm{OLS}} - \left(\frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\left(\frac{1}{n}\boldsymbol{\pi}_n\right).$$

Since in our setting $\hat{\boldsymbol{\beta}}_n^{\mathrm{OLS}}$ is strongly consistent, $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}$ if and only if $(n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\left(n^{-1}\boldsymbol{\pi}_n\right) \xrightarrow{a.s.} \mathbf{0}$. In view of (4.6.8), we have $(n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\left(n^{-1}\boldsymbol{\pi}_n\right) \xrightarrow{a.s.} \mathbf{0}$ if and only if $n^{-1}\boldsymbol{\pi}_n \xrightarrow{a.s.} \mathbf{0}$.
Assume $n^{-1}\lambda_1^{(n)} \to 0$. By (4.6.10) we have $\|\boldsymbol{\pi}^n\|_\infty \leq \lambda_1^{(n)}$, which gives

$$\left\|\frac{\boldsymbol{\pi}_n}{n}\right\|_\infty \leq \frac{\lambda_1^{(n)}}{n} \to \mathbf{0}.$$

Therefore, (4.6.9) implies that $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} \xrightarrow{a.s.} \boldsymbol{\beta}$.
Now assume that $\boldsymbol{\beta} \neq \mathbf{0}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ is strongly consistent, i.e. $n^{-1}\boldsymbol{\pi}_n \xrightarrow{a.s.} \mathbf{0}$. Then, (4.6.11) gives

$$p\|\boldsymbol{\pi}_n\|_\infty \geq \|\boldsymbol{U}_{\mathbf{M}_n}'\boldsymbol{\pi}_n\|_\infty = \|\widetilde{\boldsymbol{\Lambda}}_n\|_\infty \geq \lambda_1^{(n)} \tag{4.6.12}$$

provided $\boldsymbol{M}_n \neq \mathbf{0}$. Applying (4.6.10) for $\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}} = \mathbf{0}$, we note that $\boldsymbol{M}_n(\omega) = \mathbf{0}$ if and only if

$$J_{n^{-1}\boldsymbol{\Lambda}_n}^*\left(n^{-1}\boldsymbol{X}_n(\omega)'\boldsymbol{Y}_n(\omega)\right) \leq 1.$$

In view of Lemma 4.6.2, it can be easily verified that $n^{-1}\boldsymbol{X}_n'\boldsymbol{Y}_n \xrightarrow{a.s.} \boldsymbol{C}\boldsymbol{\beta}$. Since

$$\left\|\frac{1}{n}\boldsymbol{\pi}_n\right\|_\infty \quad \geq \quad \left\|\frac{1}{n}\boldsymbol{\pi}_n\right\|_\infty \mathbb{1}_{\{\mathbf{M}_n=0\}} \quad = \quad \left\|\frac{1}{n}\boldsymbol{X}_n'\boldsymbol{Y}_n\right\|_\infty \mathbb{1}_{\{\mathbf{M}_n=0\}},$$

we see that for $\boldsymbol{\beta} \neq \mathbf{0}$, we have $\boldsymbol{M}_n \neq 0$ for large $n$ almost surely. Thus, for $\boldsymbol{\beta} \neq \mathbf{0}$ we eventually obtain for large $n$

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

□

## 4.7 Simulation study

Our simulations aim at illustrating Theorems 4.5.1 and 4.5.3 and at showing that the results provided in these theorems are somehow unified. We consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ has i.i.d. $\mathrm{N}(0,1)$ entries. Up to a constant, we choose components of $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)'$ as expected values of ordered standard Gaussian statistics. Let $V_{(1)} \geq \ldots \geq V_{(p)}$ be ordered statistics of i.i.d. $\mathrm{N}(0,1)$ random variables. An approximation of $\mathbb{E}[V_{(i)}]$ for some $i \in \{1, \ldots, p\}$, denoted $E(i,p)$, is given below (see [99] and references therein)

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

where $\Phi$ is the cumulative distribution function of a $\mathrm{N}(0,1)$ random variable. We set $\lambda_i = E(i,p) + E(p-1,p) - 2E(p,p)$ for $i = 1, \ldots, p$. Note that since $E(1,p) > \ldots > E(p,p)$, then $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)' \in \mathbb{R}^{p+}$.

For the design matrix $\boldsymbol{X}$ and the vector of regression coefficients $\boldsymbol{\beta}$ we consider two cases:

- $\boldsymbol{X}$ is orthogonal, and the components of $\boldsymbol{\beta}$ are all equal with a magnitude that tends to infinity.

- $\boldsymbol{X}$ is asymptotically orthogonal, the sample size diverges to infinity and the components of $\boldsymbol{\beta}$ are equal to 1.

### 4.7.1 Sharp upper bound when X is orthogonal

In Figure 4.10 we have $p = 100$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is orthogonal (*i.e* $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_{100}$) and $\beta_1 = \ldots = \beta_p = c > 0$. To compute the value $\alpha_{0.95}$ of the scaling parameter for which the upper bound equals 0.95, we note that $\boldsymbol{\pi}_\alpha$ is a Gaussian vector of $\mathrm{N}\left(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}, \alpha^{-2}\boldsymbol{X}'(\boldsymbol{I} - \widetilde{\boldsymbol{X}}_{\mathbf{M}}\widetilde{\boldsymbol{X}}^+_{\mathbf{M}})\boldsymbol{X}\right)$ distribution. Moreover, since $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{\beta}) = (1, \ldots, 1)'$, we have

$$\begin{aligned} \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} &= \left(\frac{1}{p}\sum_{i=1}^p \lambda_i, \ldots, \frac{1}{p}\sum_{i=1}^p \lambda_i\right) \text{ and} \\ \boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{\boldsymbol{X}}_{\mathbf{M}}\widetilde{\boldsymbol{X}}^+_{\mathbf{M}})\boldsymbol{X} &= \begin{pmatrix} 1-1/p & -1/p & \ldots & -1/p \\ -1/p & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1/p \\ -1/p & \ldots & -1/p & 1-1/p \end{pmatrix}. \end{aligned} \tag{4.7.1}$$

Since the distribution of $\boldsymbol{\pi}_\alpha$ is given and the open SLOPE irrepresentability condition occurs, for arbitrary $q \in (0,1)$ one can choose $\alpha_q$ for which $\mathbb{P}(J^*_{\boldsymbol{\Lambda}}(\boldsymbol{\pi}_{\alpha_q}) \leq 1) = q$.

In the following graph, $q = 0.95$ and $\alpha_{0.95} \approx 1.391$.

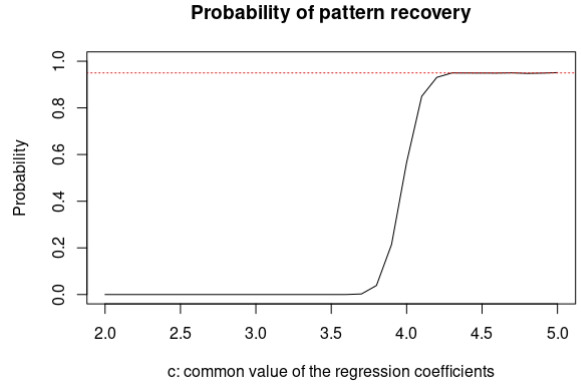



Figure 4.10: The estimates of probability of pattern recovery by SLOPE as a function of $c$ where $c = \beta_1 = \ldots = \beta_{100} > 0$. For each point the probability is computed via $10^5$ Monte-Carlo experiments. The scaling parameter $\alpha_{0.95} \approx 1.391$ is chosen to fix the upper bound at 0.95. Note that when $c$ is large, the probability of pattern recovery is approximately equal to 0.95.

### 4.7.2 Limiting probability when X is asymptotically orthogonal

In Figure 4.11, $\boldsymbol{X} \in \mathbb{R}^{n\times 100}$ has i.i.d. $\mathrm{N}(0,1)$ entries, $\beta_1 = \ldots = \beta_{100} = 1$ and $\alpha_{0.95} \approx 1.391$. Actually, since $n^{-1}\boldsymbol{X}'\boldsymbol{X}$ converges to $\boldsymbol{I}_{100}$, when $\boldsymbol{patt}(\boldsymbol{\beta}) = (1,\ldots,1)'$ the Gaussian vector involved in the limiting probability has the same mean and covariance as (4.7.1).

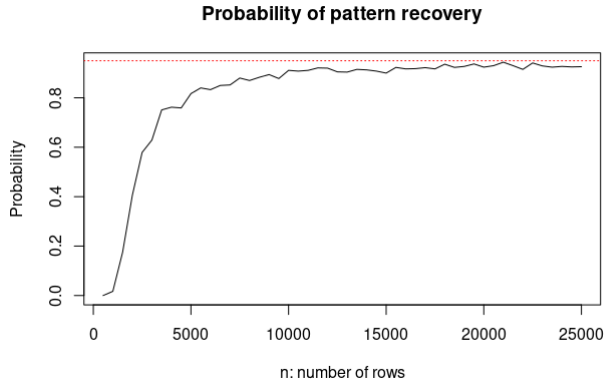



Figure 4.11: The estimates of probability of pattern recovery by SLOPE as a function of $n$. For each point the probability is computed via $10^3$ Monte-Carlo experiments.

## 4.8 Discussion

This chapter makes 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, SLOPE irrepresentability condition is rather stringent and imposes a strict restriction on the number of nonzero clusters in $\boldsymbol{\beta}$. On the other hand, in [174] it is shown that a much weaker condition for LASSO is required to separate

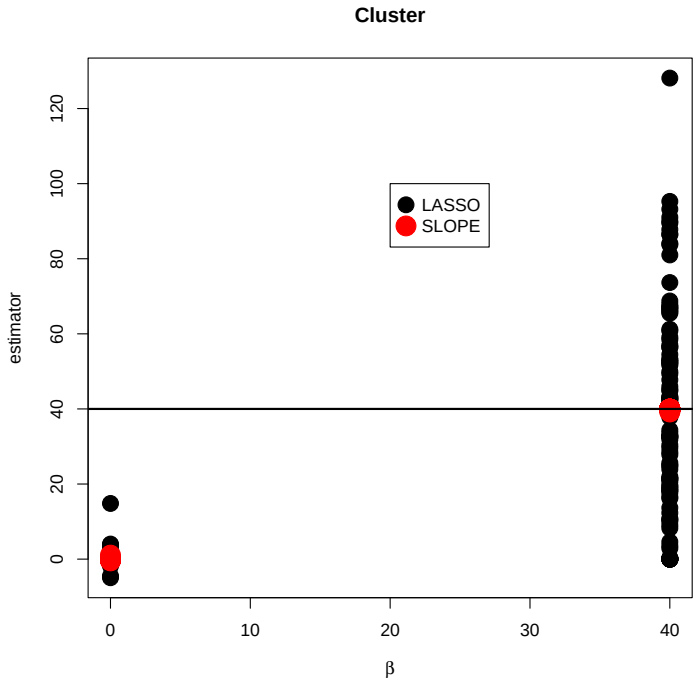



Figure 4.12: Comparison of LASSO and SLOPE when the cluster structure is present in the data. Here $n = 100$, $p = 200$, and the correlation between $i^{th}$ and $j^{th}$ column of $\boldsymbol{X}$ is equal to $0.9048^{|i-j|}$. First $k = 100$ columns of $\boldsymbol{X}$ are associated with $\boldsymbol{Y}$ and their nonzero regression coefficient are all equal to 40. The SLOPE and LASSO irrepresentability conditions are not satisfied, but SLOPE, contrary to LASSO, satisfies the accessibility condition.

the estimators of the null components of $\boldsymbol{\beta}$ from the estimators of nonzero regression coefficients. This condition, called accessibility (also called identifiability), requires that the vector $\boldsymbol{\beta}$ has a minimal $\ell_1$ norm among all vectors $\boldsymbol{\gamma}$ such that $\boldsymbol{X\beta} = \boldsymbol{X\gamma}$. Thus, when the accessibility condition is satisfied, one can recover the sign of $\boldsymbol{\beta}$ by thresholding LASSO estimates. Empirical results from [174] suggest that this weaker condition is also sufficient for the sign recovery by the adaptive LASSO [194]. In Chapter 5 it is shown that a similar result holds for SLOPE. In this case, the accessibility condition is satisfied if the vector $\boldsymbol{\beta}$ has the smallest sorted $\ell_1$ norm among all vectors $\boldsymbol{\gamma}$ such that $\boldsymbol{X\beta} = \boldsymbol{X\gamma}$. In Chapter **??** it is shown that when the accessibility condition is satisfied then SLOPE properly ranks the estimators of regression coefficients and the SLOPE pattern can be recovered by shrinking similar estimates towards the cluster centers. Figure 4.12 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 $\boldsymbol{X}$ and the vector $\boldsymbol{Y}$ are generated as in the example illustrated in Figure 4.1 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 can not 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 can not properly identify the pattern, it estimates $\boldsymbol{\beta}$ with such a good precision that the difference between the estimated and the true

pattern is hardly visible on the graph. These nice ranking and estimation properties of SLOPE bring a promise for efficient pattern recovery by appropriate thresholded and adaptive SLOPE versions.

## 4.9 Appendix — Proofs

### 4.9.1 Proof of Proposition 4.2.1

Note that if $\boldsymbol{M} = \boldsymbol{0}$, then the statement is valid by (4.2.1). Thus, we may later assume that $\boldsymbol{M} \neq \boldsymbol{0}$. To simplify the notation, we write $\widetilde{\boldsymbol{\Lambda}}$ instead of $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$. The elements of $\widetilde{\boldsymbol{\Lambda}}$ are denoted by $\widetilde{\boldsymbol{\Lambda}}_l$, $l = 1, \ldots, k$. Let $k = \|\boldsymbol{M}\|_\infty$. Before proving Proposition 4.2.1, note that, by assumption, there exists $\boldsymbol{s} \in \mathbb{R}^{k+}$ such that $\boldsymbol{b} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}$. Consequently, $|\boldsymbol{b}|_\downarrow = \boldsymbol{U}_{|\mathbf{M}|_\downarrow}\boldsymbol{s}$ and thus

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

Moreover, with $p_l = |\{i\colon\ |m_i| \geq k + 1 - l\}|$, we have $\widetilde{\boldsymbol{\Lambda}}_l = \lambda_{p_{l-1}+1} + \ldots + \lambda_{p_l}$, $l = 1, \ldots, k$.

*Proof of Proposition 4.2.1.*
First, we prove the inclusion $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) \subset \left\{\boldsymbol{v} \in \mathbb{R}^p\colon\ J^*_{\boldsymbol{\Lambda}}(\boldsymbol{v}) \leq 1 \text{ and } \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}\right\}$. Let $\boldsymbol{v} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$. Since $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{v}) \leq 1$ (see (4.2.1)) then, by definition of the dual sorted $\ell_1$ 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 $\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}$.
For all $l \in \{1, \ldots, k\}$ we have the following inequality

$$\sum_{i=1}^{l}[\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v}]_i = \sum_{i\colon\ |m_i|\geq k+1-l} \operatorname{sign}(m_i)v_i \leq \sum_{i\colon\ |m_i|\geq k+1-l} |v_i| \leq \sum_{i=1}^{p_l}|\boldsymbol{v}|_{(i)} \leq \sum_{i=1}^{p_l}\lambda_i = \sum_{i=1}^{l}\widetilde{\boldsymbol{\Lambda}}_i. \quad (4.9.1)$$

Note that

$$\begin{aligned}\boldsymbol{b}'\boldsymbol{v} = (\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s})'\boldsymbol{v} = \sum_{i=1}^{k} s_i[\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v}]_i = \sum_{l=1}^{k-1}(s_l - s_{l+1})\sum_{i=1}^{l}[\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v}]_i + s_k\sum_{i=1}^{k}[\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v}]_i \\ \leq \sum_{l=1}^{k-1}(s_l - s_{l+1})\sum_{i=1}^{l}\widetilde{\boldsymbol{\Lambda}}_i + s_k\sum_{i=1}^{k}\widetilde{\boldsymbol{\Lambda}}_i = \sum_{l=1}^{k} s_l\widetilde{\boldsymbol{\Lambda}}_l = J_{\boldsymbol{\Lambda}}(\boldsymbol{b}).\end{aligned}$$

Moreover, since $\boldsymbol{v} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, we have $\boldsymbol{b}'\boldsymbol{v} = J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$ (see (4.2.1)). Therefore,

$$\sum_{i=1}^{l}[\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v}]_i = \sum_{i=1}^{l}\widetilde{\boldsymbol{\Lambda}}_i \quad \text{for} \quad l = 1, \ldots, k$$

and thus the inequalities given in (4.9.1) are the equalities. Thus

$$[\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v}]_l = \widetilde{\boldsymbol{\Lambda}}_l \quad \text{for} \quad l = 1, \ldots, k$$

and hence that $\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}$.
Now we prove the other inclusion, $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) \supset \left\{\boldsymbol{v} \in \mathbb{R}^p\colon\ J^*_{\boldsymbol{\Lambda}}(\boldsymbol{v}) \leq 1 \text{ and } \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\right\}$. Assume that $\boldsymbol{v} \in \mathbb{R}^p$ satisfies $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{v}) \leq 1$ and $\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}$. To prove that $\boldsymbol{v} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, it remains to establish that $\boldsymbol{b}'\boldsymbol{v} = J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$ (see (4.2.1)). Since $\boldsymbol{b} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}$, we have

$$\boldsymbol{b}'\boldsymbol{v} = (\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s})'\boldsymbol{v} = \boldsymbol{s}'\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \boldsymbol{s}'\widetilde{\boldsymbol{\Lambda}} = J_{\boldsymbol{\Lambda}}(\boldsymbol{b}).$$

□

### 4.9.2 Proof of Proposition 4.2.2

**Lemma 4.9.1.** *Let $\boldsymbol{\Lambda} \in \mathbb{R}^{p+}$ and $\boldsymbol{b} \in \mathbb{R}^p$. If $\boldsymbol{\Lambda} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, 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 $\check{\boldsymbol{\pi}} = (\lambda_1, \ldots, \lambda_{i-1}, -\lambda_i, \lambda_{i+1}, \ldots, \lambda_p)$ we have $J^*_{\boldsymbol{\Lambda}}(\check{\boldsymbol{\pi}}) \leq 1$ and one may deduce that

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

Consequently $\boldsymbol{\Lambda} \notin \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, which leads to a contradiction.
Now, let us assume that $b_i < b_j$ for some $1 \leq i < j \leq p$. We define $\check{\boldsymbol{\pi}}$ as the following:

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

Note that $J^*_{\boldsymbol{\Lambda}}(\check{\boldsymbol{\pi}}) \leq 1$. 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

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

Consequently, $\boldsymbol{\Lambda} \notin \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, which again leads to a contradiction. □

Let $\psi$ be an orthogonal transformation defined by $\psi\colon \boldsymbol{x} \in \mathbb{R}^p \mapsto (v_1 b_{r(1)}, \ldots, v_p b_{r(p)})$ where $v_1, \ldots, v_p \in \{-1, 1\}$ and let $r$ be a permutation on $\{1, \ldots, p\}$. Before proving Proposition 4.2.2, let us recall that for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$ we have $J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = J_{\boldsymbol{\Lambda}}(\psi(\boldsymbol{b}))$, $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = J^*_{\boldsymbol{\Lambda}}(\psi(\boldsymbol{b}))$ and $\boldsymbol{b}'\boldsymbol{a} = \psi(\boldsymbol{b})'\psi(\boldsymbol{a})$, therefore $\partial J_{\boldsymbol{\Lambda}}(\psi(\boldsymbol{b})) = \psi(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}))$.

*Proof of Proposition 4.2.2.* If $\boldsymbol{patt}(\boldsymbol{a}) = \boldsymbol{patt}(\boldsymbol{b})$, then, according to Proposition 4.2.1, $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{a}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$. Set $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{a})$ and $\widetilde{\boldsymbol{M}} = \boldsymbol{patt}(\boldsymbol{b})$. It remains to prove that if $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{a}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$, then $\boldsymbol{M} = \widetilde{\boldsymbol{M}}$. Since the subdifferential $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{a})$ depends on $\boldsymbol{a}$ only through its pattern, then by Proposition 4.2.1 we have $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{a}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ and similarly $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \partial J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}})$.
First let us assume that $\boldsymbol{M} = |\boldsymbol{M}|_\downarrow$ namely $M_1 \geq M_2 \geq \ldots \geq M_p \geq 0$. In this case, $\boldsymbol{M}'\boldsymbol{\Lambda} = J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ and hence $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_p)' \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$. Since $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}) = \partial J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}})$, it follows from Lemma 4.9.1 that $\widetilde{M}_1 \geq \ldots \geq \widetilde{M}_p \geq 0$, because $\boldsymbol{\Lambda} \in \partial J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}})$. To prove that $\boldsymbol{M} = \widetilde{\boldsymbol{M}}$, first let us establish that $M_p = \widetilde{M}_p = 0$ or $M_p = \widetilde{M}_p = 1$. If $M_p = 0$ and $\widetilde{M}_p = 1$, then we set $\check{\boldsymbol{\pi}} = (\lambda_1, \ldots, \lambda_{p-1}, 0)'$, where $J^*_{\boldsymbol{\Lambda}}(\check{\boldsymbol{\pi}}) \leq 1$. Because

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

we have $\check{\boldsymbol{\pi}} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ and $\check{\boldsymbol{\pi}} \notin \partial J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}})$ which provides a contradiction. We proceed analogously for $M_p = 1$ and $\widetilde{M}_p = 0$. To achieve proving that $\boldsymbol{M} = \widetilde{\boldsymbol{M}}$, let us establish that $m_i = m_{i+1}$ and $\widetilde{m}_i = \widetilde{m}_{i+1}$ or $m_i > m_{i+1}$ and $\widetilde{m}_i > \widetilde{m}_{i+1}$. If $m_i = m_{i+1}$ and $\widetilde{m}_i > \widetilde{m}_{i+1}$ then we define $\check{\boldsymbol{\pi}}$ satisfying $J^*_{\boldsymbol{\Lambda}}(\check{\boldsymbol{\pi}}) \leq 1$, as follows

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

Since $\lambda_i m_i + \lambda_{i+1} m_{i+1} = \lambda_{i+1} m_i + \lambda_i m_{i+1}$ and $\lambda_i \widetilde{m}_i + \lambda_{i+1} \widetilde{m}_{i+1} > \lambda_{i+1} \widetilde{m}_i + \lambda_i \widetilde{m}_{i+1}$ then

$$J_{\boldsymbol{\Lambda}}(\boldsymbol{M}) = \boldsymbol{\Lambda}' \boldsymbol{M} = \breve{\boldsymbol{\pi}}' \boldsymbol{M} \text{ and } J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}}) = \boldsymbol{\Lambda}' \widetilde{\boldsymbol{M}} > \breve{\boldsymbol{\pi}}' \widetilde{\boldsymbol{M}}.$$

Consequently $\breve{\boldsymbol{\pi}} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ and $\breve{\boldsymbol{\pi}} \notin \partial J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}})$, which provides a contradiction. We proceed analogously for $m_i > m_{i+1}$ and $\widetilde{m}_i = \widetilde{m}_{i+1}$. Finally, if $\boldsymbol{M} \neq |\boldsymbol{M}|_{\downarrow}$, then we may pick an orthogonal transformation $\psi$ as defined above for which $\psi(\boldsymbol{M}) = |\boldsymbol{M}|_{\downarrow}$. Since $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}) = \partial J_{\boldsymbol{\Lambda}}(\widetilde{\boldsymbol{M}})$ implies that $\partial J_{\boldsymbol{\Lambda}}(\psi(\boldsymbol{M})) = \partial J_{\boldsymbol{\Lambda}}(\psi(\widetilde{\boldsymbol{M}}))$, then the first part of the proof establishes that $\psi(\widetilde{\boldsymbol{M}}) = \psi(\boldsymbol{M})$ and thus $\boldsymbol{M} = \widetilde{\boldsymbol{M}}$.

□

Recall that $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{x}) \leq 1$ if and only if

$$|x|_{(1)} + \ldots |x|_{(j)} \leq \lambda_1 + \ldots + \lambda_j, \qquad j = 1, \ldots, p. \tag{4.9.2}$$

The following result follows from the proof of Proposition 4.2.1.

**Proposition 4.9.1.** *Assume $\boldsymbol{x} \in \mathbb{R}^p$ satisfies $J^*_{\boldsymbol{\Lambda}}(\boldsymbol{x}) \leq 1$ and let $\boldsymbol{b} \in \mathbb{R}^p$. Then $\boldsymbol{x}$ belongs to $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$ if and only if the following three conditions hold true:*

*(a) If $b_i \neq 0$, then $\operatorname{sign}(x_i) = \operatorname{sign}(b_i)$,*

*(b) If $|b_i| > |b_j|$ then $|x_i| \geq |x_j|$,*

*(c) The equalities hold in (4.9.2) for $j \in \{n_1, n_2, \ldots, n_k\}$, where $n_j = |\{i\colon\ |m_i| \geq k + 1 - j\}|$ with $(M_1, \ldots, M_p)' = \boldsymbol{patt}(\boldsymbol{b})$.*

### 4.9.3 Proof of Theorem 4.3.1

*Proof of Theorem 4.3.1. Necessity.* Let us assume that there exists $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\boldsymbol{\Lambda}}(\boldsymbol{Y})$ with $\boldsymbol{patt}(\hat{\boldsymbol{\beta}}) = \boldsymbol{M}$. Consequently, $\hat{\boldsymbol{\beta}} = \boldsymbol{U}_{\mathbf{M}} s$ for some $\boldsymbol{s} \in \mathbb{R}^{k+}$.
By Proposition 4.2.2, $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \partial J_{\boldsymbol{\Lambda}}(\hat{\boldsymbol{\beta}}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$. Multiplying this inclusion by $\boldsymbol{U}'_{\mathbf{M}}$, due to (4.2.2), we get $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$ and so

$$\widetilde{\boldsymbol{X}}'_{\mathbf{M}} \boldsymbol{Y} - \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}} \boldsymbol{X} \hat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}} \widetilde{\boldsymbol{X}}_{\mathbf{M}} \boldsymbol{s}. \tag{4.9.3}$$

The positivity condition is proven.
We apply $(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+$ from the left to (4.9.3) and use the fact that $\widetilde{\boldsymbol{P}}_{\mathbf{M}} = (\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+ \widetilde{\boldsymbol{X}}'_{\mathbf{M}}$ is the projection onto $\operatorname{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})$. Since $\boldsymbol{X}\hat{\boldsymbol{\beta}} \in \operatorname{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})$, we have $\widetilde{\boldsymbol{P}}_{\mathbf{M}} \boldsymbol{X} \hat{\boldsymbol{\beta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Thus,

$$\widetilde{\boldsymbol{P}}_{\mathbf{M}} \boldsymbol{Y} - (\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+ \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

The above equality gives the subdifferential condition:

$$\begin{aligned} \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}) \ni \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \boldsymbol{X}'(\boldsymbol{Y} - (\widetilde{\boldsymbol{P}}_{\mathbf{M}} \boldsymbol{Y} - (\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+ \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}})) \\ &= \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+ \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} + \boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_{\mathbf{M}}) \boldsymbol{Y} = \boldsymbol{\pi}. \end{aligned} \tag{4.9.4}$$

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

$$\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}} \boldsymbol{Y} - \widetilde{\boldsymbol{X}}'_{\mathbf{M}} \widetilde{\boldsymbol{X}}_{\mathbf{M}} \boldsymbol{s}. \tag{4.9.5}$$

Let us show that $\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s} \in S_{\mathbf{X},\boldsymbol{\Lambda}}(\boldsymbol{Y})$. By definition of $\boldsymbol{U}_{\mathbf{M}}$, we have $\boldsymbol{patt}(\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}) = \boldsymbol{M}$, thus $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$. Moreover, using (4.9.4) and (4.9.5) one may deduce that

$$\begin{aligned}\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}) \ni \boldsymbol{\pi} &= \boldsymbol{X}'(\boldsymbol{Y} - \widetilde{\boldsymbol{P}}_{\mathbf{M}}\boldsymbol{Y} + (\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) \\ &= \boldsymbol{X}'(\boldsymbol{Y} - \widetilde{\boldsymbol{P}}_{\mathbf{M}}\boldsymbol{Y} + (\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}(\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{Y} - \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{s})) \\ &= \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}).\end{aligned}$$

Consequently, $\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s} \in S_{\mathbf{X},\boldsymbol{\Lambda}}(\boldsymbol{Y})$. □

### 4.9.4 Proof of Corollary 4.3.1

*Proof of Corollary 4.3.1.* If SLOPE recovers the pattern of $\boldsymbol{\beta}$ in the noiseless case, then, by Theorem 4.3.1, the subdifferential condition reads as: $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$.
Conversely, if $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+}\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ then, by Theorem 4.3.1, it remains to show that the positivity condition occurs for $\alpha > 0$ small enough. Since $\boldsymbol{\beta} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}$ for some $\boldsymbol{s} \in \mathbb{R}^{k+}$, where $k = \|\boldsymbol{M}\|_{\infty}$, we have

$$\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{Y} - \alpha\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{s} - \alpha\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}.$$

Therefore for $\alpha > 0$ small enough, $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{Y} - \alpha\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}\mathbb{R}^{k+}$ and thus, the positivity condition is proven. □

### 4.9.5 Proof of Theorem 4.5.1

**Lemma 4.9.2.** *Let* $\mathbf{0} \neq \boldsymbol{b} \in \mathbb{R}^p$ *and* $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{b})$. *Then the smallest affine space containing* $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})$ *is* $\operatorname{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})) = \{\boldsymbol{v} \in \mathbb{R}^p\colon\ \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\}$.

*Proof.* According to Proposition 4.2.1 we have

$$\operatorname{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b})) \subset \{\boldsymbol{v} \in \mathbb{R}^p\colon\ \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\}.$$

Moreover, according to Theorem 4 in [156] we have

$$\dim(\operatorname{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{b}))) = \|\boldsymbol{M}\|_{\infty} = \dim(\{\boldsymbol{v} \in \mathbb{R}^p\colon\ \boldsymbol{U}'_{\mathbf{M}}\boldsymbol{v} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}\}),$$

which achieves the proof. □

*Proof of Theorem 4.5.1. (i) Sharpness of the upper bound.* According to Theorem 4.3.1, the pattern recovery by SLOPE is equivalent to have simultaneously the positivity condition and the subdifferential condition satisfied. The upper bound (4.5.2) 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 $\infty$. Clearly the upper bound is reached when $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \notin \operatorname{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, thus we may assume that $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \operatorname{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$. Recall that $\boldsymbol{\beta}^{(r)} = \boldsymbol{U}_M\boldsymbol{s}^{(r)}$ for $\boldsymbol{s}^{(r)} \in \mathbb{R}^{k+}$ and thus $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{Y}^{(r)} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}\boldsymbol{s}^{(r)} + \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{\varepsilon}$. As $\widetilde{\boldsymbol{X}}'_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^{+} = \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}})^{+}$ is the projection on $\operatorname{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, we obtain

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

Note that by the assumption on $\Delta_r$:

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

Consequently, for $r$ large enough we have

$$\widetilde{\boldsymbol{X}}'_{\mathbf{M}}\boldsymbol{Y}^{(r)} - \alpha_r\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \widetilde{\boldsymbol{X}}'_{\mathbf{M}}\widetilde{\boldsymbol{X}}_{\mathbf{M}}\mathbb{R}^{k+}.$$

Since this fact is true for any realization of $\boldsymbol{\varepsilon}$, we get

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

*(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 $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{ri}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$. First we observe that

$$\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} + \frac{1}{\alpha_r}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{P}_{\mathbf{M}})\varepsilon \overset{r\to\infty}{\longrightarrow} \boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}. \tag{4.9.6}$$

Note by Lemma 4.9.2 that $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} + \alpha_r^{-1}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{P}_{\mathbf{M}})\varepsilon \in \mathrm{aff}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$. Indeed, since $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{col}(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})$, we have

$$\underbrace{\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}}_{=\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}} + \frac{1}{\alpha_r}\underbrace{\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{P}_{\mathbf{M}})\boldsymbol{\varepsilon}(\omega)}_{=\mathbf{0}} = \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}.$$

The second term above is zero due to the fact that $(\boldsymbol{I}_n - \widetilde{P}_{\mathbf{M}})$ is an orthogonal projection onto $\mathrm{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}})^\perp$. When $\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{ri}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$, due to (4.9.6), one may deduce that for sufficiently large $r$ we have

$$\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} + \frac{1}{\alpha_r}\boldsymbol{X}'(\boldsymbol{I}_n - \widetilde{P}_{\mathbf{M}})\boldsymbol{\varepsilon} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}).$$

Consequently, when $r$ is sufficiently large, both the positivity and the subdifferential conditions occur which, by Theorem 4.3.1, concludes the proof. □

### 4.9.6 Proofs from Section 4.5.2

In this part we give proofs of Theorem 4.5.3 and Theorem 4.5.5. They are preceded by a series of simple lemmas. For reader's convenience we recall the setting of Section 4.5.2.

A. $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \dots, \varepsilon_n)'$, where $(\varepsilon_i)_i$ are i.i.d. centered with finite variance $\sigma^2$.

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

B2. $\frac{\max_{i=1,\dots,n}|X_{ij}^{(n)}|}{\sqrt{\sum_{i=1}^n (X_{ij}^{(n)})^2}} \xrightarrow{\mathbb{P}} 0$, where $\boldsymbol{X}_n = (X_{ij}^{(n)})_{ij}$, for each $j = 1, \dots, p$.

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

C. $(\boldsymbol{X}_n)_n$ and $(\varepsilon_n)_n$ are independent.

We consider a sequence of tuning parameters $(\mathbf{\Lambda}_n)_n$ defined by $\mathbf{\Lambda}_n = \alpha_n \mathbf{\Lambda}$, where $\mathbf{\Lambda} \in \mathbb{R}^{p+}$ 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 pattern $\boldsymbol{M}$, *i.e.* $\widetilde{\mathbf{\Lambda}} := \boldsymbol{U}'_{|\boldsymbol{M}|\downarrow}\mathbf{\Lambda}$, $\widetilde{\mathbf{\Lambda}}_n := \alpha_n \widetilde{\mathbf{\Lambda}}$ and $\widetilde{\boldsymbol{X}}_n := \boldsymbol{X}_n \boldsymbol{U}_{\mathbf{M}}$.

**Lemma 4.9.3.**

*(a) Under A, B1, B2 and C,*

$$\frac{1}{\sqrt{n}} \boldsymbol{X}'_n \boldsymbol{\varepsilon}_n \xrightarrow{d} \boldsymbol{V} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{C}). \tag{4.9.7}$$

*(b) Under A, B1 and C,*

$$\frac{1}{n} \boldsymbol{X}'_n \boldsymbol{\varepsilon}_n \xrightarrow{\mathbb{P}} 0. \tag{4.9.8}$$

*(c) Under A, B' and C,*

$$0 < \limsup_{n\to\infty} \frac{\|\boldsymbol{X}'_n \boldsymbol{\varepsilon}_n\|_\infty}{\sqrt{n \log\log n}} < \infty \qquad a.s. \tag{4.9.9}$$

*Proof of* (4.9.7). It is enough to show that for any Borel subset $A \subset \mathbb{R}^p$ one has

$$\mathbb{P}\left( \frac{1}{\sqrt{n}} \boldsymbol{X}'_n \boldsymbol{\varepsilon}_n \in A \mid (\boldsymbol{X}_n)_n \right) \xrightarrow{\mathbb{P}} \mathbb{P}\left(\boldsymbol{V} \in A\right). \tag{4.9.10}$$

Since both sides above are bounded, the convergence in probability implies convergence in $L^1$ and therefore establishes (4.9.7). To show (4.9.10), we will prove that for any subsequence $(n_k)_k$, there exists a sub-subsequence $(n_{k_l})_l$ for which, as $l \to \infty$,

$$\mathbb{P}\left( \frac{1}{\sqrt{n_{k_l}}} \boldsymbol{X}'_{n_{k_l}} \boldsymbol{\varepsilon}_{n_{k_l}} \in A \mid (X_n)_n \right) \xrightarrow{a.s.} \mathbb{P}\left(\boldsymbol{V} \in A\right). \tag{4.9.11}$$

Let $\mathbb{P}_{\mathbf{X}}$ denote the regular conditional probability $\mathbb{P}(\cdot \mid (\boldsymbol{X}_n)_n)$ 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

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

We have

$$\begin{aligned}
\mathrm{Var}_{\mathbf{X}}\left( \frac{1}{\sqrt{n_{k_l}}} \boldsymbol{X}'_{n_{k_l}} \boldsymbol{\varepsilon}_{n_{k_l}} \right) &= \frac{1}{n_{k_l}} \mathbb{E}\left[ \boldsymbol{X}'_{n_{k_l}} \boldsymbol{\varepsilon}_{n_{k_l}} \boldsymbol{\varepsilon}'_{n_{k_l}} \boldsymbol{X}_{n_{k_l}} \mid (\boldsymbol{X}_n)_n \right] \\
&= \frac{1}{n_{k_l}} \boldsymbol{X}'_{n_{k_l}} \mathbb{E}\left[ \boldsymbol{\varepsilon}_{n_{k_l}} \boldsymbol{\varepsilon}'_{n_{k_l}} \right] \boldsymbol{X}_{n_{k_l}} = \frac{\sigma^2}{n_{k_l}} X'_{n_{k_l}} \boldsymbol{X}_{n_{k_l}} \xrightarrow{a.s.} \sigma^2 \boldsymbol{C} > 0,
\end{aligned}$$

and one can apply multivariate Lindeberg-Feller CLT on the space $(\Omega, \mathcal{F}, \mathbb{P}_{\mathbf{X}})$ to prove (4.9.11). Alternatively, the same result follows from Theorem 2.3.16, 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 (2.3.3)). For our application, the assumption of nonnegative weights in Theorem 2.3.16 is not essential.
For (ii) we observe that previous derivations imply that $\mathrm{Var}_{\mathbf{X}}(n^{-1} \boldsymbol{X}'_n \boldsymbol{\varepsilon}_n) \xrightarrow{\mathbb{P}} 0$. We deduce that $\mathbb{P}_{\mathbf{X}}(n^{-1} \boldsymbol{X}'_n \boldsymbol{\varepsilon}_n) \xrightarrow{\mathbb{P}} 0$ and hence (ii) follows upon averaging over $(\boldsymbol{X}_n)_n$.
Eq. (4.9.9) is the law of iterated logarithm for an i.i.d. sequence $(\Xi_i \varepsilon_i)_i$. □

**Lemma 4.9.4.** *Let $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{\beta})$. Assume $\alpha_n/n \to 0$.*

*(a) Under A, B1 and C, the positivity condition is satisfied for large $n$ with high probability.*

*(b) Under A, B' and C, the positivity condition is almost surely satisfied for large $n$.*

*Proof.* If $\boldsymbol{M} = \boldsymbol{0}$, then the positivity condition is trivially satisfied. Thus, we consider $\boldsymbol{M} \neq \boldsymbol{0}$.
(i) Since $\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n$ is invertible for large $n$ with high probability, the positivity condition is equivalent to

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

Let $\boldsymbol{s}_0 \in \mathbb{R}^{k+}$ be defined through $\boldsymbol{\beta} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}_0$, where $k = \|\boldsymbol{M}\|_\infty$. We will show that if $\alpha_n/n \to 0$, then $s_n \xrightarrow{\mathbb{P}} \boldsymbol{s}_0$. 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 $\boldsymbol{s}_n$ as

$$\boldsymbol{s}_n = (\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\widetilde{\boldsymbol{X}}_n'\boldsymbol{Y}_n - \alpha_n(\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\widetilde{\boldsymbol{\Lambda}}.$$

Since $\boldsymbol{\beta} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}_0$, we conclude $\boldsymbol{X}_n\boldsymbol{\beta} = \boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}_0 = \widetilde{\boldsymbol{X}}_n\boldsymbol{s}_0$, so the linear regression model takes the form $\boldsymbol{Y}_n = \widetilde{\boldsymbol{X}}_n\boldsymbol{s}_0 + \boldsymbol{\varepsilon}_n$. Thus, $(\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\widetilde{\boldsymbol{X}}_n'\boldsymbol{Y}_n$ is the OLS estimator of $\boldsymbol{s}_0$.
By assumption B and Lemma 4.9.3, we deduce that

$$(\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\widetilde{\boldsymbol{X}}_n'\boldsymbol{Y}_n = \boldsymbol{s}_0 + (n^{-1}\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\boldsymbol{U}_{\mathbf{M}}\frac{1}{n}\boldsymbol{X}_n'\boldsymbol{\varepsilon}_n \xrightarrow{\mathbb{P}} \boldsymbol{s}_0 + [(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}U_{\mathbf{M}}]\boldsymbol{0} = \boldsymbol{s}_0.$$

To complete the proof, we note that

$$\alpha_n(\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\widetilde{\boldsymbol{\Lambda}} = \frac{\alpha_n}{n}\left[n(\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1}\widetilde{\boldsymbol{\Lambda}}\right] \xrightarrow{\mathbb{P}} \boldsymbol{0}\left[(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}\right] = \boldsymbol{0}.$$

(ii) If one assumes B' instead of B1, then $n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{C}$ and by (4.9.9), $n^{-1}\boldsymbol{X}_n'\boldsymbol{\varepsilon}_n \xrightarrow{a.s.} \boldsymbol{0}$. The result follows along the same lines as (i). □

For $\boldsymbol{M} \neq \boldsymbol{0}$ we denote

$$\boldsymbol{\pi}_n^{(1)} = \boldsymbol{X}_n'(\widetilde{\boldsymbol{X}}_n')^{+}\widetilde{\boldsymbol{\Lambda}}_n, \qquad \boldsymbol{\pi}_n^{(2)} = \boldsymbol{X}_n'(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_n)\boldsymbol{Y}_n,$$
$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_n^{(1)} + \boldsymbol{\pi}_n^{(2)},$$

which simplifies in the $\boldsymbol{M} = \boldsymbol{0}$ case to $\boldsymbol{\pi}_n = \boldsymbol{\pi}_n^{(2)} = \boldsymbol{X}_n'\boldsymbol{Y}_n$.
Recall that the subdifferential condition is equivalent to $J^*_{\boldsymbol{\Lambda}_n}(\boldsymbol{\pi}_n) \leq 1$ and $\widetilde{\boldsymbol{\Lambda}}_n \in \mathrm{col}(\widetilde{\boldsymbol{X}}_{\mathbf{M}}')$ and the latter is satisfied in our setting. Since $\alpha J_{\boldsymbol{\Lambda}} = J_{\alpha\boldsymbol{\Lambda}}$, the subdifferential condition is satisfied if and only if

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

In view of results shown below, $\alpha_n^{-1}\boldsymbol{\pi}_n^{(1)}$ converges almost surely, while $n^{-1/2}\boldsymbol{\pi}_n^{(2)}$ 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 4.9.5.** *(a)*
*(i) Assume A, B1 and C. If $\boldsymbol{M} \neq \boldsymbol{0}$, then*

$$\frac{1}{\alpha_n}\boldsymbol{\pi}_n^{(1)} \xrightarrow{\mathbb{P}} \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}.$$

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

$$\boldsymbol{V} \sim \mathrm{N}\left(0, \sigma^2\left[\boldsymbol{C} - \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\right]\right).$$

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

*(b) Assume A, B' and C.*

*(i') If* $\boldsymbol{M} \neq \mathbf{0}$*, then*

$$\frac{1}{\alpha_n}\boldsymbol{\pi}_n^{(1)} \xrightarrow{a.s.} \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}}.$$

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

*Proof.* (i) Assumption B implies that

$$\boldsymbol{X}_n'\widetilde{\boldsymbol{X}}_n(\widetilde{\boldsymbol{X}}_n'\widetilde{\boldsymbol{X}}_n)^{-1} = \frac{1}{n}\boldsymbol{X}_n'\boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}})^{-1} \xrightarrow{\mathbb{P}} \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}.$$

(ii) When $\boldsymbol{\beta} = \boldsymbol{U}_{\mathbf{M}}\boldsymbol{s}_0$, then the linear regression model takes the form $\boldsymbol{Y}_n = \widetilde{\boldsymbol{X}}_n\boldsymbol{s}_0 + \boldsymbol{\varepsilon}_n$. Since $\widetilde{\boldsymbol{P}}_n$ is the projection matrix onto $\mathrm{col}(\widetilde{\boldsymbol{X}}_n)$, we have $(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_n)\widetilde{\boldsymbol{X}}_n = 0$. Thus,

$$\begin{aligned} n^{-1/2}\boldsymbol{\pi}_n^{(2)} &= n^{-1/2}\boldsymbol{X}_n'(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_n)\boldsymbol{Y}_n = n^{-1/2}\boldsymbol{X}_n'(\boldsymbol{I}_n - \widetilde{\boldsymbol{P}}_n)\boldsymbol{\varepsilon}_n \\ &= \left[I_p - \boldsymbol{X}_n'\boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{X}_n'\boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}})^{-1}\boldsymbol{U}_{\mathbf{M}}'\right]\left[n^{-1/2}\boldsymbol{X}_n'\boldsymbol{\varepsilon}\right]. \end{aligned}$$

By assumption B we have,

$$n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n\boldsymbol{U}_{\mathbf{M}})^{-1}\boldsymbol{U}_{\mathbf{M}}' \xrightarrow{\mathbb{P}} \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\boldsymbol{U}_{\mathbf{M}}'.$$

Thus, by Lemma 4.9.3 and Slutsky's theorem, we obtain (ii).

(iii) follows directly from (ii). Assumption B' implies that $n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n \xrightarrow{a.s.} \boldsymbol{C}$ and thus (i') is proven in the same way as (i). (ii') follows from (4.9.9).

□

*Proof of Theorem 4.5.3.* (a) is a direct consequence of Lemmas 4.9.4 and 4.9.5. Since positivity condition is satisfied for large $n$ with high probability, for (b) we have with $\boldsymbol{M} = \boldsymbol{patt}(\boldsymbol{\beta})$,

$$\begin{aligned} \lim_{n\to\infty}\mathbb{P}\left(\boldsymbol{patt}(\hat{\boldsymbol{\beta}}_n^{\mathrm{SLOPE}}) = \boldsymbol{M}\right) &= \lim_{n\to\infty}\mathbb{P}\left(\boldsymbol{\pi}_n \in \partial J_{\alpha_n\boldsymbol{\Lambda}}(\boldsymbol{M})\right) = \lim_{n\to\infty}\mathbb{P}\left(\alpha_n^{-1}\boldsymbol{\pi}_n \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})\right) \\ &\geq \lim_{n\to\infty}\mathbb{P}\left(\alpha_n^{-1}\boldsymbol{\pi}_n \in \mathrm{ri}(\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))\right) = 1, \end{aligned} \tag{4.9.12}$$

where in the last inequality we use Portmanteau Theorem, assumption (4.5.5) and the fact that the sequence $(\alpha_n^{-1}\boldsymbol{\pi}_n)_n$ converges in distribution to $\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}$ if and only if $\alpha_n/\sqrt{n} \to \infty$.
Condition (4.5.6) implies that $\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1} \in \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$. Since $(\alpha_n^{-1}\boldsymbol{\pi}_n)_n$ converges in probability to $\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}_{\mathbf{M}}'\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}$, the necessity of this condition is explained by (4.9.12). □

*Proof of Theorem 4.5.5.* By Lemma 4.9.4, the positivity condition is satisfied for large $n$ almost surely. By Lemma 4.9.5 (i) and (iii), we have

$$\boldsymbol{a}_n := \frac{1}{\alpha_n}\boldsymbol{\pi}_n \xrightarrow{a.s.} \boldsymbol{C}\boldsymbol{U}_{\mathbf{M}}(\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{C}\boldsymbol{U}_{\mathbf{M}})^{-1}\widetilde{\boldsymbol{\Lambda}} =: \boldsymbol{a}_0.$$

It is easy to see that $\boldsymbol{U}'_{\mathbf{M}}\boldsymbol{a}_n = \widetilde{\boldsymbol{\Lambda}}$. By the condition $\boldsymbol{a}_0 \in \mathrm{ri}(J_{\boldsymbol{\Lambda}}(\boldsymbol{M}))$ it follows that $\boldsymbol{a}_n \in J_{\boldsymbol{\Lambda}}(\boldsymbol{M})$ almost surely for sufficiently large $n$. Therefore $\boldsymbol{\pi}_n \in J_{\boldsymbol{\Lambda}_n}(\boldsymbol{M})$ for large $n$ almost surely and thus the subdifferential condition is also satisfied. □

# Chapter 5

# Geometry of Pattern Recovery by Penalized and Thresholded Estimators

## 5.1 Introduction

The content of this chapter is based on the preprint of Graczyk, Schneider, the author of the dissertation and Tardivel [93].
As in previous chapters, we consider the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector of regression coefficients and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a random noise term. We assume that the distribution of $\boldsymbol{\varepsilon}$ is symmetric, continuous and its density is strictly positive.
Many penalized estimators of $\boldsymbol{\beta}$ have been broadly studied in literature, e.g., LASSO [47, 176], SLOPE [189, 26, 136], OSCAR [29], fused LASSO [177], fused adaptive LASSO [149], clustered LASSO [163], PACS [162] and generalized LASSO [179]. When the loss function is the residual sum of squares, these estimators minimize, with respect to $\boldsymbol{b} \in \mathbb{R}^p$, a function:

$$f(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \mathrm{pen}(\boldsymbol{b}),$$

where $\lambda > 0$ is the tuning parameter and the penalty term pen is a real-valued polyhedral gauge, i.e. a nonnegative convex function, homogeneous, vanishing at $\mathbf{0}$ and whose unit ball is a polyhedron. Note that in this definition we do not require boundedness of a unit ball.
The literature related to penalized least squares estimators is vast and many of these estimators have interesting and relevant structures as illustrated, e.g., in [182]. For instance, LASSO is sparse, i.e. some components of this estimator are null. The fused LASSO is a sparse and piecewise constant estimator [177], the supremum norm promotes a cluster of components maximal in absolute value [107], and SLOPE and OSCAR estimators promote clusters made of those components of these estimators, which are equal in absolute value [29, 76, 156, 24, 165].

### 5.1.1 Pattern recovery by penalized least squares estimators

We define a polyhedral gauge pen as a nonnegative convex function, homogeneous, vanishing at $\mathbf{0}$ and whose unit ball is a polyhedron. Every polyhedral gauge can be written as the maximum

of a finite number of linear functions [152, 133]

$$\forall \boldsymbol{x} \in \mathbb{R}^p,\ \mathrm{pen}(\boldsymbol{x}) = \max\{0, \boldsymbol{u}_1'\boldsymbol{x}, \ldots, \boldsymbol{u}_l'\boldsymbol{x}\},\ \text{for some } \boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p.$$

Note that a polyhedral gauge with a bounded and symmetric unit ball $\{\boldsymbol{x} \in \mathbb{R}^p :\ \mathrm{pen}(\boldsymbol{x}) \leq 1\}$ is a polyhedral norm.
Given $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$ and $\lambda > 0$, the set $S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$ of solutions of a penalized least squares optimization problem is defined as follows:

$$S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y}) := \underset{\boldsymbol{b}\in\mathbb{R}^p}{\arg\min}\, \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda\mathrm{pen}(\boldsymbol{b}). \tag{5.1.1}$$

It is important to know that the set $S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$ is not empty:

**Theorem 5.1.1.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\lambda > 0$ and $\mathrm{pen}(\boldsymbol{x}) = \max\{0, \boldsymbol{u}_1'\boldsymbol{x}, \ldots, \boldsymbol{u}_l'\boldsymbol{x}\}$, where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p$. Then the function*

$$f(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \mathrm{pen}(\boldsymbol{b})$$

*has at least one minimizer.*

The proof is given in the Appendix.
Note that, potentially, the set $S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$ might be not a singleton, i.e. the penalized least squares estimator might be not unique. Below we recall the definition of the pattern equivalence class, which is one of the most important notions in this chapter.

**Definition 2.2.2** (Equality of patterns)**.** Let $\mathrm{pen} : \mathbb{R}^p \mapsto \mathbb{R}$ be a polyhedral gauge. We say that $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{z} \in \mathbb{R}^p$ have the same pattern with respect to pen, denoted $\boldsymbol{x} \overset{\mathrm{pen}}{\sim} \boldsymbol{z}$, when

$$\partial\mathrm{pen}(\boldsymbol{x}) = \partial\mathrm{pen}(\boldsymbol{z}),$$

where $\partial$pen is the subdifferential of pen. The set of all vectors having the same subdifferential as $\boldsymbol{x}$, denoted $C_{\boldsymbol{x}}$, is called the pattern equivalence class.

In Theorem 5.2.2 we prove that pattern equivalence classes, illustrated in Section 2.2, are given by normal cones of $B^*$, where $B^*$ is the polar set of the unit ball of pen.
For the $\ell_1$ norm penalization, two vectors $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^p$ have the same pattern if and only if $\boldsymbol{sign}(\boldsymbol{x}) = \boldsymbol{sign}(\boldsymbol{z})$. More generally, two vectors having the same pattern with respect to a polyhedral gauge penalty have a specific structure as illustrated on many examples in Section 5.3. Given $\boldsymbol{X}$ and $\boldsymbol{Y}$, we aim at recovering the pattern of $\boldsymbol{\beta}$. For LASSO this means the recovery of $\boldsymbol{sign}(\boldsymbol{\beta})$.
In Theorem 5.4.4 we give a necessary condition for pattern recovery by penalized least squares estimators, called noiseless recovery condition. Later, in Section 5.5, we introduce penalized estimators relaxing this condition. Beforehand, we recall some of the already known conditions for pattern recovery.

### Conditions for pattern recovery — examples

*LASSO.* Below we assume the uniqueness of the LASSO estimator. Then we note $\hat{\beta}^{\mathrm{LASSO}}$ as a unique element of $S_{\mathbf{X},\lambda\|.\|_1}(\boldsymbol{Y})$. As mentioned above, LASSO estimation is a sparse method that nullifies some of the components with positive probability, entailing that the estimator also performs variable selection. Instigated by this sparsity property, an abundant literature

has arisen to deal with the recovery of the location of the non-null components of $\boldsymbol{\beta}$, or, more specifically, the recovery of the sign vector of $\boldsymbol{\beta}$ [83, 130, 183, 192, 194]. An evident necessary condition for sign recovery by LASSO is for $\boldsymbol{sign}(\boldsymbol{\beta})$ to be accessible by LASSO, i.e. for a fixed $\lambda > 0$, there has to exist such $\boldsymbol{Y} \in \mathbb{R}^n$ that $\boldsymbol{sign}(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\boldsymbol{Y})) = \boldsymbol{sign}(\boldsymbol{\beta})$. Otherwise, the sign recovery is impossible. A geometrical characterization of accessible sign vectors is given in [160, 156]. However, the accessibility of $\boldsymbol{sign}(\boldsymbol{\beta})$ by LASSO does not mean that the probability of sign recovery by LASSO is close to 1 even if the non-null components of $\boldsymbol{\beta}$ are extremely large. Actually, for sign recovery with a probability larger than 1/2 a stronger condition is needed, called the irrepresentability condition [183], which is satisfied when

$$\|\boldsymbol{X}'_{\overline{I}}\boldsymbol{X}_I(\boldsymbol{X}'_I\boldsymbol{X}_I)^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty \leq 1,$$

where $I := \{i \in [p] : \ \beta_i \neq 0\}$ and $\overline{I} := \{i \in [p] : \ \beta_i = 0\}$.

*SLOPE.* The notions of accessibility condition and irrepresentability condition for SLOPE have been recently introduced respectively in [156] and in Chapter 4. In particular, in Chapter 4, similarly as for LASSO, when the SLOPE irrepresentability condition does not occur, the probability of pattern recovery is smaller than 1/2.

*Generalized LASSO.* By substituting the $\ell_1$ norm with a polyhedral gauge $\mathrm{pen} = \|\boldsymbol{D}.\|_1$ for a fixed matrix $\boldsymbol{D}$ of a linear map on $\mathbb{R}^p$, one constructs an estimator $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\|\mathbf{D}.\|_1}(\boldsymbol{Y})$, where $\boldsymbol{D}\widehat{\boldsymbol{\beta}}$ has some null components. It is a reason why the generalized LASSO is frequently used for structure recovery. One should be aware that the structure induced by generalized LASSO depends on the matrix $\boldsymbol{D}$.
For instance, when $\boldsymbol{D}$ is a matrix such that $\boldsymbol{Db} = (b_2 - b_1, \ldots, b_p - b_{p-1})'$ (denoted $\boldsymbol{D}^{\mathrm{tv}}$ below), then the penalty term $\|\boldsymbol{D}.\|_1$ promotes equality between neighbouring components of $\widehat{\boldsymbol{\beta}}$. Moreover, this estimator can recover the jump set: $\{i \in [p-1] : \ \beta_i \neq \beta_{i+1}\}$ [104]. Actually, articles [145, 140] provide theoretical properties for the jump set recovery under the irrepresentability condition.
The noiseless recovery condition and the irrepresentability condition can be relaxed using thresholded estimators as explained below.

### 5.1.2 Pattern recovery by a thresholded estimator

Theorem 5.5.1 generalizes results known for LASSO to a wide class of penalized estimators. Specifically, we prove that thresholding penalized least squares estimators allows the recovery of the pattern of $\beta$ with large probability under a weaker condition than before. We recall the definition of thresholded estimator below.

**Definition 2.3.23.**
Let pen be a polyhedral gauge, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$ and $\lambda > 0$. Given $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$, we say that $\widetilde{\boldsymbol{\beta}}$ is a thresholded estimator of $\widehat{\boldsymbol{\beta}}$ if $\partial\mathrm{pen}(\widehat{\boldsymbol{\beta}}) \subseteq \partial\mathrm{pen}(\widetilde{\boldsymbol{\beta}})$.

One of the examples of a thresholded penalized estimator is the thresholded LASSO 5.5.1.
Given a threshold $\tau \geq 0$, the thresholded LASSO $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO},\tau}$ is defined as follows

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

Note that for every $\tau \geq 0$ we have $\partial\|.\|_1(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}) \subseteq \partial\|.\|_1(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO},\tau})$ and thus $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO},\tau}$ is a thresholded estimator of $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ in the sense of Definition 2.3.23.
The thresholded LASSO does not have the same statistical properties as LASSO, cf. [131, 186]. Concerning sign recovery by thresholded LASSO, the accessibility condition is necessary and sufficient, which was proven by Tardivel and Bogdan [174]. Moreover, they also proved that, contrarily to LASSO, thresholded LASSO can recover the sign of $\boldsymbol{\beta}$ with a large probability under the accessibility condition, even if the irrepresentability condition is not satisfied, as soon as non-null components of $\boldsymbol{\beta}$ are sufficiently large. This nice property for sign recovery by thresholded LASSO remains true for thresholded basis pursuit [153, 55, 54].

## 5.2 Geometry of pattern equivalence

Let $F$ be a face of a polytope $P$. We propose the following relation between normal cones $N_P(\boldsymbol{x})$ and $N_F(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathrm{ri}(F)$, that seems unknown.

**Proposition 5.2.1.** *Let $\boldsymbol{x} \in \mathrm{ri}(F) \subset P$. Then*

$$\mathrm{lin}(N_P(\boldsymbol{x})) = N_F(\boldsymbol{x}).$$

*Proof.* A polytope $P$ can be represented as $\{\boldsymbol{v} \in \mathbb{R}^p : \ \langle \boldsymbol{s}_j, \boldsymbol{v}\rangle \leq r_j, \ j = 1, \dots, m\}$, cf. [101, p. 138]. Then one may describe a face $F \subset P$ as

$$F = P \cap \{\boldsymbol{v} \in \mathbb{R}^p : \ \langle -\boldsymbol{s}_j, \boldsymbol{v}\rangle \leq -r_j, \ j \in A_P(\boldsymbol{x})\},$$

where $A_P(\boldsymbol{x}) := \{j = 1, \dots, m : \ \langle \boldsymbol{s}_j, \boldsymbol{v}\rangle = r_j\}$ is the set of active constraints for $\boldsymbol{x}$. Without loss of generality one may rearrange the sequence $\{\boldsymbol{s}_j\}_{j=1}^m$ such that $A_P(\boldsymbol{x}) = \{1, \dots, a\}$, where $a \leq m$. Then $F$ is equal to

$$F = \{\boldsymbol{v} \in \mathbb{R}^p : \ \langle \widetilde{\boldsymbol{s}_j}, \boldsymbol{v}\rangle \leq \widetilde{r}_j, \ j = 1, \dots, m + a\},$$

where

$$\widetilde{\boldsymbol{s}_j} = \begin{cases} \boldsymbol{s}_j, & j \leq m, \\ -\boldsymbol{s}_{j-m}, & j > m \end{cases} \qquad \textit{and} \qquad \widetilde{r}_j = \begin{cases} r_j, & 1 \leq m, \\ -r_j, & j > m. \end{cases}$$

The corresponding set of active constraints $A_F(\boldsymbol{x})$ is equal to

$$A_F(\boldsymbol{x}) = A_P(\boldsymbol{x}) \cup \{m + 1, \dots, m + a\}.$$

By Proposition 2.3.2 we obtain

$$N_F(\boldsymbol{x}) = \mathit{cone}\{\widetilde{\boldsymbol{s}_j} : \ j \in A_F(\boldsymbol{x})\} = \mathit{cone}\{\boldsymbol{s}_j, -\boldsymbol{s}_j : \ j \in A_P(\boldsymbol{x})\} = \mathrm{lin}\{\boldsymbol{s}_j : \ j \in A_P(\boldsymbol{x})\} = \mathrm{lin}(N_P(\boldsymbol{x})).$$

□

We will also use the following property of normal cones.

**Proposition 5.2.2.** *Let $\boldsymbol{x} \in \mathrm{ri}(F) \subset P$. Then the normal cone $N_P(\boldsymbol{x})$ has the property*

$$\mathrm{lin}(N_P(\boldsymbol{x})) = \left(\overrightarrow{\mathrm{aff}(F)}\right)^{\perp}.$$

*Proof.* ($\subset$) We start by noticing that for any $\boldsymbol{d} \in \operatorname{lin}(N_P(\boldsymbol{x})) = N_F(\boldsymbol{x})$ and for any $\boldsymbol{s} \in F$ we have $\langle \boldsymbol{d}, \boldsymbol{s} - \boldsymbol{x} \rangle = 0$. Indeed, since $\boldsymbol{s} = \boldsymbol{x} - (\boldsymbol{x} - \boldsymbol{s})$ and $\boldsymbol{x} \in \operatorname{ri}(F)$, then there exists such $\delta > 0$ that $\boldsymbol{x} + \delta(\boldsymbol{x} - \boldsymbol{s}) \in F$. Therefore,

$$0 = \langle \boldsymbol{d}, \boldsymbol{x} - \boldsymbol{x} \rangle = \frac{\delta}{\delta + 1} \langle \boldsymbol{d}, \boldsymbol{x} - (\boldsymbol{x} - (\boldsymbol{x} - \boldsymbol{s})) \rangle + \frac{1}{\delta + 1} \langle \boldsymbol{d}, \boldsymbol{x} - (\boldsymbol{x} + \delta(\boldsymbol{x} - \boldsymbol{s})) \rangle,$$

while both summands are not larger than zero. Thus, $\langle \boldsymbol{d}, \boldsymbol{x} - \boldsymbol{s} \rangle = \langle \boldsymbol{d}, \boldsymbol{x} - (\boldsymbol{x} - (\boldsymbol{x} - \boldsymbol{s})) \rangle = 0$. It means that for every $\boldsymbol{s} \in \operatorname{aff}(F)$ we have $\langle \boldsymbol{d}, \boldsymbol{s} \rangle = \langle \boldsymbol{d}, \boldsymbol{x} \rangle$. It implies that for every $\boldsymbol{s} \in \overrightarrow{\operatorname{aff}(F)}$ we have $\langle \boldsymbol{d}, \boldsymbol{s} \rangle = 0$, thus $\boldsymbol{d} \in \left(\overrightarrow{\operatorname{aff}(F)}\right)^{\perp}$. By the arbitrariness of $\boldsymbol{d}$, we obtain $N_F(\boldsymbol{x}) \subset \left(\overrightarrow{\operatorname{aff}(F)}\right)^{\perp}$ and then $\operatorname{lin}(N_P(\boldsymbol{x})) \subset \left(\overrightarrow{\operatorname{aff}(F)}\right)^{\perp}$.
($\supset$) Let $\boldsymbol{v} \in \left(\overrightarrow{\operatorname{aff}(F)}\right)^{\perp} \overset{Lemma\ 2.3.2}{=} \operatorname{lin}\{\boldsymbol{w} - \boldsymbol{x} : \ \boldsymbol{w} \in F\}$. Then $\langle \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{w} \rangle = 0$ for every $\boldsymbol{w} \in F$. It implies that $\boldsymbol{v} \in N_F(\boldsymbol{x})$, which by Proposition 5.2.1 equals $\operatorname{lin}(N_P(\boldsymbol{x}))$. □

**Remark 5.2.1.** In addition to the property of normal cones from Proposition 2.3.2, we deduce from Propositions 5.2.2 and 2.3.2 that

$$\operatorname{lin}\left(\{\boldsymbol{s}_j : \ j \in A_P(\boldsymbol{x})\}\right) = \left(\overrightarrow{\operatorname{aff}(F)}\right)^{\perp}.$$

### 5.2.1 Pattern equivalence classes and normal cones

**Definition 2.3.15** (Pattern equivalence class)**.** [Pattern equivalence class] Let $\boldsymbol{x} \in \mathbb{R}^p$ and let *pen* be a polyhedral gauge. The pattern equivalence class $C_{\boldsymbol{x}}$ is the set of all vectors having the same subdifferential as $\boldsymbol{x}$:

$$C_{\boldsymbol{x}} := \{\boldsymbol{w} \in \mathbb{R}^p : \ \partial \operatorname{pen}(\boldsymbol{w}) = \partial \operatorname{pen}(\boldsymbol{x})\}.$$

**Lemma 5.2.1.** *Let pen be a polyhedral gauge and $\boldsymbol{x} \neq \boldsymbol{0}$. Then $C_{\boldsymbol{x}} \subset N(F_{\mathbf{x}})$.*

*Proof.* Let $\boldsymbol{w} \in C_{\boldsymbol{x}}$ and $\boldsymbol{s} \in \operatorname{ri}(F_{\boldsymbol{x}})$. Since $\boldsymbol{s} \in F_{\boldsymbol{x}} = F_{\boldsymbol{w}}$ then, for all $\boldsymbol{z} \in B^*$, we have

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

Consequently, $\boldsymbol{w} \in N_{B^*}(\boldsymbol{s}) = N_{B^*}(F_{\boldsymbol{x}})$. □

**Theorem 5.2.2.** *The pattern cone $C_{\boldsymbol{x}}$ is the relative interior of the normal cone of the face $F_{\boldsymbol{x}}$ of the dual unit ball $B^*$*

$$C_{\boldsymbol{x}} = \operatorname{ri}(N_{B^*}(F_{\boldsymbol{x}})). \tag{5.2.1}$$

*Proof.* If $\partial \operatorname{pen}(\boldsymbol{x}) = F$, we denote $C_{\boldsymbol{x}} = C_F$. $\mathbb{R}^p$ can be partitioned both into pairwise disjoint pattern sets and into pairwise disjoint relative interiors of corresponding normal cones:

$$\bigsqcup_F C_F = \mathbb{R}^p = \bigsqcup_F \operatorname{ri}(N(F)).$$

Thus it is sufficient to prove the inclusion $C_F \subset \operatorname{ri}(N(F))$, or equivalently

$$C_{\boldsymbol{x}} \subset \operatorname{ri}(N_{B^*}(F_{\boldsymbol{x}})) \ \text{ for } \ \boldsymbol{x} \in \mathbb{R}^p. \tag{5.2.2}$$

Let $\boldsymbol{w} \in C_{\boldsymbol{x}}$. We want to show that $\boldsymbol{w} \in ri(N_{B^*}(F_{\boldsymbol{x}}))$, i.e. there exists $\delta > 0$ such that any point $\boldsymbol{z} \in B(\boldsymbol{w}, \delta) \cap \text{aff}(N_{B^*}(F_{\boldsymbol{x}}))$ belongs to $C_{\boldsymbol{x}}$.
According to Lemma 5.8.9, for $\delta > 0$ small enough we obtain $F_{\boldsymbol{z}} \subseteq F_{\boldsymbol{w}}$, which equals $F_{\boldsymbol{x}}$, since $\boldsymbol{w} \in C_{\boldsymbol{x}}$. Moreover, if $F_{\boldsymbol{z}} \neq F_{\boldsymbol{x}}$, then one may pick $\boldsymbol{u} \in F_{\boldsymbol{z}} \subset F_{\boldsymbol{x}}$ and $\boldsymbol{v} \in F_{\boldsymbol{x}} \setminus F_{\boldsymbol{z}}$. Observe that $\boldsymbol{u} - \boldsymbol{v} \in \overrightarrow{\text{aff}(F_{\boldsymbol{x}})}$. Then, by Lemma 2.3.1, we have $\boldsymbol{x} \in C_{\boldsymbol{x}} \subset N_{B^*}(F_{\boldsymbol{x}})$, which by Lemma 5.8.1 belongs to $\left(\overrightarrow{\text{aff}(F_{\boldsymbol{x}})}\right)^{\perp}$. Therefore $\langle \boldsymbol{x}, \boldsymbol{u} - \boldsymbol{v} \rangle = 0$.
We also have $\boldsymbol{z} \in \text{aff}(N_{B^*}(F_{\boldsymbol{x}})) = \text{lin}(N_{B^*}(F_{\boldsymbol{x}}))$, which by Lemma 5.8.1 belongs to $\left(\overrightarrow{\text{aff}(F_{\boldsymbol{x}})}\right)^{\perp}$. Therefore $\langle \boldsymbol{z}, \boldsymbol{u} - \boldsymbol{v} \rangle = 0$, too.
Consequently, $\boldsymbol{v}'\boldsymbol{z} = \boldsymbol{u}'\boldsymbol{z} = \text{pen}(\boldsymbol{z})$ and thus $\boldsymbol{v} \in F_{\boldsymbol{z}}$, which leads to a contradiction. Therefore $F_{\boldsymbol{z}} = F_{\boldsymbol{x}}$, which means that $\boldsymbol{z} \in C_{\boldsymbol{x}}$. □

### 5.2.2 Model subspace recovery

More generally, for a wide class of penalty terms including polyhedral gauges, Vaiter et al. [182] showed that the irrepresentability condition is sufficient for the model subspace recovery by penalized least squares estimators. The notion of model subspace is related to the notion of pattern. Specifically, the model subspace of $\boldsymbol{x} \in \mathbb{R}^p$ is a vector subspace of $\mathbb{R}^p$ perpendicular to $\partial \text{pen}(\boldsymbol{x})$. For the $\ell_1$ norm two vectors $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^p$ have the same model subspace when $\text{supp}(\boldsymbol{x}) = \text{supp}(\boldsymbol{z})$. In the particular case of LASSO, Theorem 6 in [182] shows that $\|\boldsymbol{X}'_{\overline{I}}\boldsymbol{X}_I(\boldsymbol{X}'_I\boldsymbol{X}_I)^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty < 1$ is a sufficient condition for model subspace recovery, *i.e.* the recovery of $\text{supp}(\boldsymbol{\beta})$. Whereas correct, this statement is not optimal. Indeed, when $\|\boldsymbol{X}'_{\overline{I}}\boldsymbol{X}_I(\boldsymbol{X}'_I\boldsymbol{X}_I)^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty < 1$, it is well known that LASSO actually can recover $\boldsymbol{sign}(\boldsymbol{\beta})$ and a fortiori $\text{supp}(\boldsymbol{\beta})$ [183]. Whereas we do not retain the notion of model subspace, in Theorem 5.2.3 we prove that the model subspace coincides with the linear span of a pattern equivalence class.

**Definition 5.2.1** (Model subspace)**.** *([182]) The model subspace $S_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is the orthogonal complement of $\overrightarrow{\text{aff}(F_{\boldsymbol{x}})}$:*

$$S_{\boldsymbol{x}} := \left(\overrightarrow{\text{aff}(F_{\boldsymbol{x}})}\right)^{\perp},$$

Our objective in this section is to prove that $S_{\boldsymbol{x}} = \text{lin}(C_{\boldsymbol{x}})$, i.e. the linear space generated by a pattern equivalence class coincides with the model subspace from [182].

**Theorem 5.2.3.** $S_{\boldsymbol{x}} = \text{lin}(C_{\boldsymbol{x}})$.

*Proof.*

$$S_{\boldsymbol{x}} = \left(\overrightarrow{\text{aff}(F_{\boldsymbol{x}})}\right)^{\perp} \overset{Prop.\ 5.2.2}{=} \text{lin}(N(F_{\boldsymbol{x}})) = \text{lin}(\text{ri}(N(F_{\boldsymbol{x}}))) \overset{Thm.\ 5.2.2}{=} \text{lin}(C_{\boldsymbol{x}}).$$

□

**Remark 5.2.4.** The equality $S_{\boldsymbol{x}} = \text{lin}(C_{\boldsymbol{x}})$ does not hold for general penalizers. Indeed, when pen is a strictly convex function, then $S_{\boldsymbol{x}} = \mathbb{R}^p$, but $\text{lin}(C_{\boldsymbol{x}})$ does not need to be equal to $\mathbb{R}^p$. For example, $pen(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2$ gives $\text{lin}(C_{\boldsymbol{0}}) = \{\boldsymbol{0}\} \neq \mathbb{R}^p$.

## 5.3 Examples of polyhedral gauges and their patterns

**Pattern for the $\ell_1$ norm (LASSO):**

Subdifferentials $\partial\|.\|_1(\boldsymbol{x}) = \partial\|.\|_1(\boldsymbol{z})$ are equal if and only if $\boldsymbol{sign}(\boldsymbol{x}) = \boldsymbol{sign}(\boldsymbol{z})$. For instance, if $\boldsymbol{x} = (1.45, -0.38, 1.56, 0, -2.76)'$, then $\boldsymbol{sign}(\boldsymbol{x}) = (1, -1, 1, 0, -1)'$. Figure 5.1 illustrates the comparison between projections onto $B^*$ and the signs of $\widehat{\boldsymbol{\beta}}^{\text{LASSO}}$ for $p = 2$ and orthogonal $\boldsymbol{X}$.

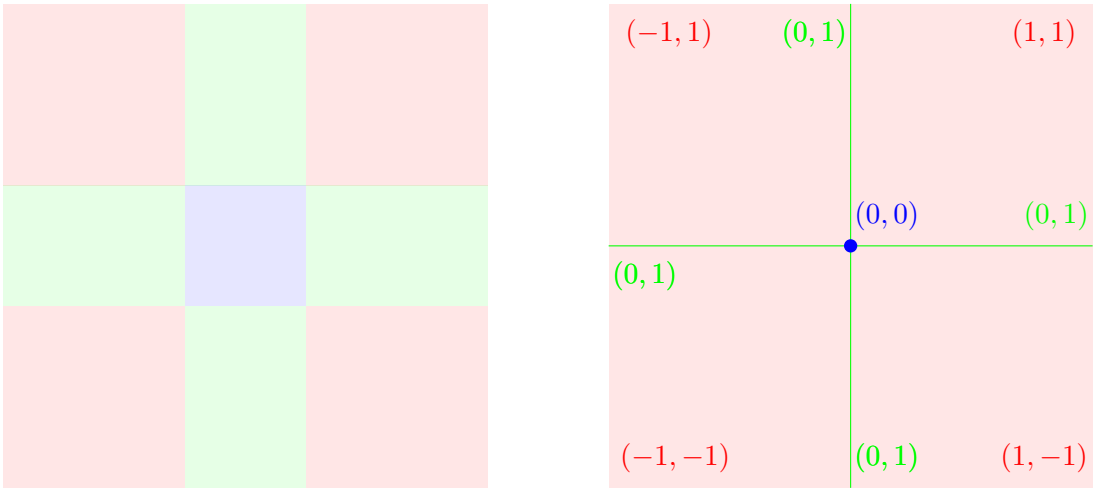



Figure 5.1: Pattern (sign) equivalence classes for LASSO in orthogonal design with $p = 2$: $\text{pen}(\boldsymbol{b}) = \|\boldsymbol{b}\|_1 = |b_1| + |b_2|$. On the left the blue polytope is $B^*$. Red and green (unbounded) sets are the preimages, with respect to the projection onto $B^*$, of its vertices and edges, respectively. The picture on the right presents sign equivalence classes of vectors in $\mathbb{R}^2$.

**Pattern for the $\ell_\infty$ norm:**

The vector $\boldsymbol{sign}^\infty(\boldsymbol{x}) \in \{-1, *, 1\}^p$ is defined as follows

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

The notation $*$ represents components that are not maximal in absolute value. Subdifferentials $\partial\|.\|_\infty(\boldsymbol{x}) = \partial\|.\|_\infty(\boldsymbol{z})$ are equal if and only if $\boldsymbol{sign}^\infty(\boldsymbol{x}) = \boldsymbol{sign}^\infty(\boldsymbol{z})$. For instance, if $\boldsymbol{x} = (1.45, 1.45, 0.56, 0, -1.45)'$ then $\boldsymbol{sign}^\infty(\boldsymbol{x}) = (1, 1, *, *, -1)'$. Figure 5.2 illustrates the comparison between projections onto $B^*$ and the pattern equivalence classes of $\widehat{\boldsymbol{\beta}}$ for $\text{pen}(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$, $p = 2$ and orthogonal $\boldsymbol{X}$.

**Pattern for the SLOPE norm:**

Let $\boldsymbol{x} \in \mathbb{R}^p$. The SLOPE pattern of $\boldsymbol{x}$, $\text{patt}(\boldsymbol{x})$, is defined by

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

where $\text{rank}(|x|)_i \in \{0, 1, \ldots, k\}$, $k$ is the number of non-zero distinct values in $\{|x_1|, \ldots, |x_p|\}$, see 1.2.1. Let $\boldsymbol{\Lambda} \in \mathbb{R}^p$ where $\lambda_1 > \cdots > \lambda_p > 0$. Then the subdifferentials $\partial J_{\boldsymbol{\Lambda}}(\boldsymbol{x}) = \partial J_{\boldsymbol{\Lambda}}(\boldsymbol{z})$

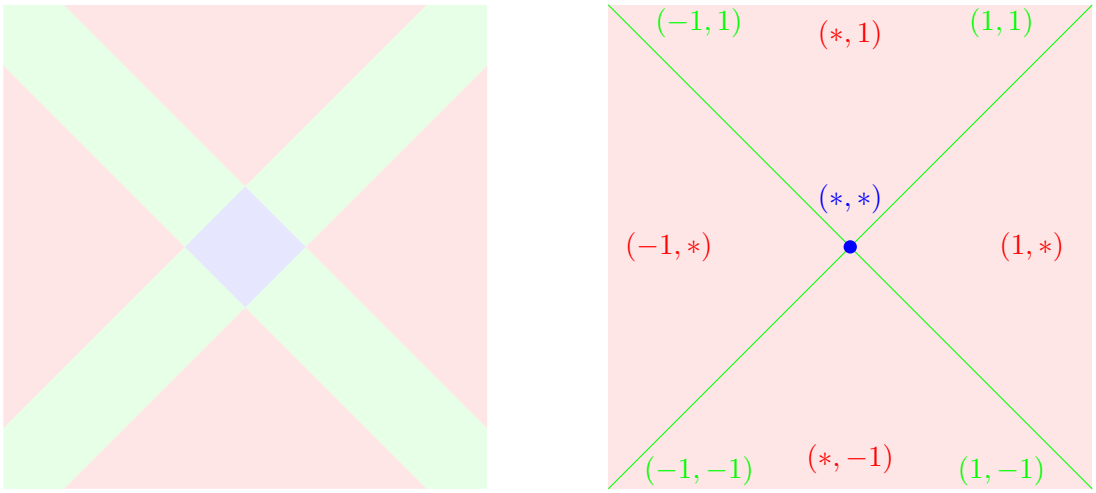



Figure 5.2: Pattern equivalence classes for $\ell_\infty$ norm, orthogonal $\boldsymbol{X}$ and $p = 2$: $\mathrm{pen}(\boldsymbol{b}) = \|\boldsymbol{b}\|_\infty = \max\{|b_1|, |b_2|\}$. On the left the blue polytope is $B^*$. Red and green (unbounded) sets are the preimages, with respect to the projection onto $B^*$, of its vertices and edges. The picture on the right presents pattern equivalence classes of vectors in $\mathbb{R}^2$.

are equal if and only if $\mathrm{patt}(\boldsymbol{x}) = \mathrm{patt}(\boldsymbol{z})$. For instance, if $\boldsymbol{x} = (3.1, -1.2, 0.5, 0, 1.2, -3.1)'$, then $\mathrm{patt}(\boldsymbol{x}) = (3, -2, 1, 0, 2, -3)'$. Figure 5.3 illustrates the comparison between projections onto $B^*$ and the patterns of $\widehat{\boldsymbol{\beta}}^{\mathrm{SLOPE}}$ for $p = 2$ and orthogonal $\boldsymbol{X}$.

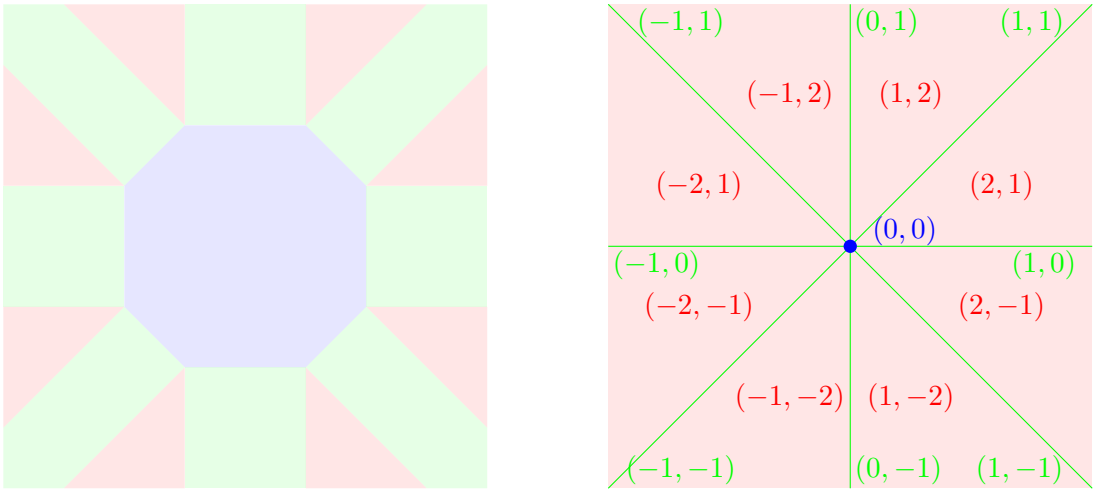



Figure 5.3: Pattern equivalence classes for SLOPE in orthogonal design with $p = 2$ and $\boldsymbol{\Lambda} = (2, 1)'$: $\mathrm{pen}(\boldsymbol{b}) = J_{\boldsymbol{\Lambda}}(\boldsymbol{b}) = \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)}$. On the left the blue polytope is $B^*$. Red and green (unbounded) sets are the preimages, with respect to the projection onto $B^*$, of its vertices and edges. The picture on the right presents pattern equivalence classes of vectors in $\mathbb{R}^2$.

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

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

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

- $\ell_1$ trend filtering:
  Let $p \geq 3$ and let $\boldsymbol{D}^{\text{tf}} \in \mathbb{R}^{(p-2)\times p}$ be the second order difference matrix defined as follows

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

  The total variation and the $\ell_1$ trend filtering [111] are examples of the generalized LASSO with the penalty term $\|\boldsymbol{D}^{\text{tv}}.\|_1$ and $\|\boldsymbol{D}^{\text{tf}}.\|_1$, respectively.

**Pattern for the total variation $\|D^{\text{tv}}.\|_1$:**

Let $p \geq 2$. The vector $\boldsymbol{jump}(\boldsymbol{x}) \in \{\nearrow, \rightarrow, \searrow\}^{p-1}$ is defined as follows

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

Subdifferentials $\partial\|\boldsymbol{D}^{\text{tv}}.\|_1(\boldsymbol{x}) = \partial\|\boldsymbol{D}^{\text{tv}}.\|_1(\boldsymbol{z})$ are equal if and only if $\boldsymbol{jump}(\boldsymbol{x}) = \boldsymbol{jump}(\boldsymbol{z})$. For instance, if $\boldsymbol{x} = (1.45, 1.45, 0.56, 0.56, -0.45, 0.35)'$ then $\boldsymbol{jump}(\boldsymbol{x}) = (\rightarrow, \searrow, \rightarrow, \searrow, \nearrow)'$. Figure 5.4 compares the projections onto $B^*$ with the jump sets of $\widehat{\boldsymbol{\beta}}$ for $p = 2$ and orthogonal $\boldsymbol{X}$.

**Pattern for the trend filtering $\|\boldsymbol{D}^{\text{tf}}.\|_1$:**

Let $p \geq 3$. The vector $\boldsymbol{knot}(\boldsymbol{x}) \in \{l, cx, cv\}^{p-2}$ is defined as follows

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

Consider the piecewise linear curve $L_{\boldsymbol{x}} := \bigcup_{i=1}^{p-1} [(i, x_i), (i+1, x_{i+1})]$. Note that $\text{knot}(x)_i$ is equal to $l$ (resp. $cx$ or $cv$) when, in the neighborhood of $i$, the curve $L_{\boldsymbol{x}}$ is linear (resp. convex or concave). Subdifferentials $\partial\|\boldsymbol{D}^{\text{tf}}.\|_1(\boldsymbol{x}) = \partial\|\boldsymbol{D}^{\text{tf}}.\|_1(\boldsymbol{z})$ are equal if and only if $\boldsymbol{knot}(\boldsymbol{x}) = \boldsymbol{knot}(\boldsymbol{z})$. Figure 5.5 provides an illustration of $\boldsymbol{knot}(\boldsymbol{x})$ for $\boldsymbol{x} = (1, 3, 5, 7, 6, 5, 4, 6, 5)' \in \mathbb{R}^9$.

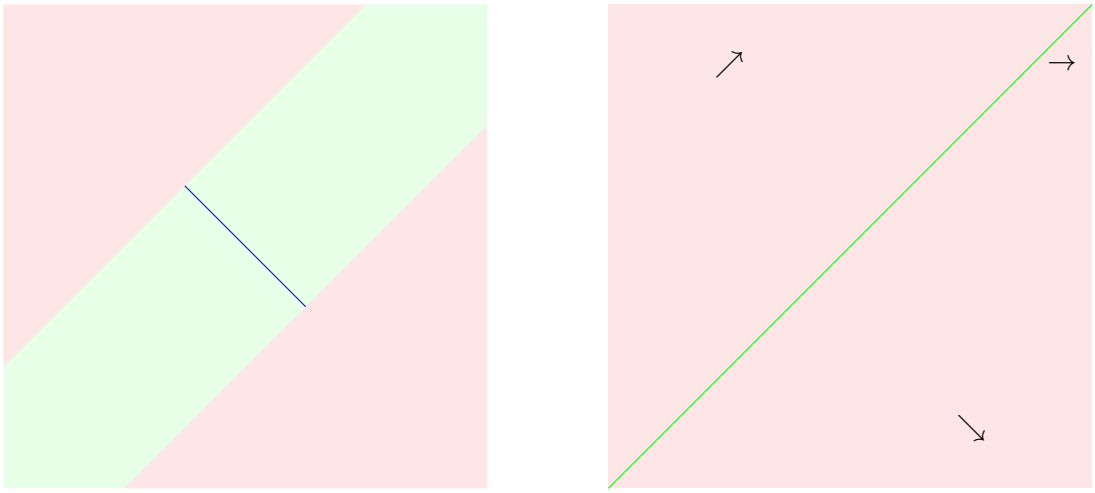


Figure 5.4: Pattern equivalence classes for the total variation in orthogonal design with $p = 2$: $\mathrm{pen}(\boldsymbol{b}) = |b_1 - b_2|$. On the left the blue polytope is $B^*$. Red and green (unbounded) sets are the preimages, with respect to the projection onto $B^*$, of its vertices and edges. The picture on the right presents pattern equivalence classes of vectors in $\mathbb{R}^2$.

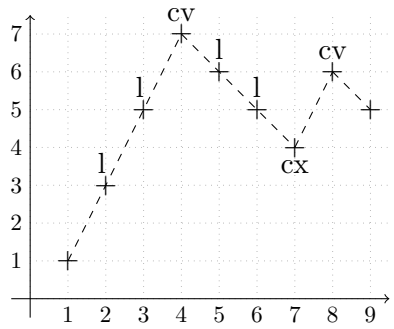



Figure 5.5: In this figure the dotted curve represents the piecewise linear curve $L_{\boldsymbol{x}}$ for $\boldsymbol{x} = (1, 3, 5, 7, 6, 5, 4, 6, 5)'$. Here $\boldsymbol{knot}(\boldsymbol{x}) = (l, l, cv, l, l, cx, cv)'$.

## 5.4 Pattern recovery in penalized estimation

### 5.4.1 Accessibility condition

Below we recall the notion of accessible patterns in the following definition, which generalizes the notion of accessible sign vectors [160, 156] and accessible SLOPE patterns [156] to a broad class of estimators penalized with polyhedral gauges.

**Definition 5.4.1** (Accessible pattern)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and* pen *be a polyhedral gauge. We say that the pattern of $\boldsymbol{\beta} \in \mathbb{R}^p$ is accessible with respect to $\boldsymbol{X}$ and* $\lambda$pen*, if there exist $\boldsymbol{y} \in \mathbb{R}^n$ and $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y})$ such that $\widehat{\boldsymbol{\beta}} \overset{\mathrm{pen}}{\sim} \boldsymbol{\beta}$.*

The accessibility of a pattern can be characterized in a geometric and an analytic way.

**Proposition 5.4.1** (Characterization of accessible patterns)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and* pen $: \mathbb{R}^p \to \mathbb{R}$ *be a real-valued polyhedral gauge.*

*(a) Geometric characterization: The pattern of $\boldsymbol{\beta} \in \mathbb{R}^p$ is accessible with respect to $\boldsymbol{X}$ and*

$\lambda$pen *if and only if*

$$\mathrm{row}(\boldsymbol{X}) \cap \partial \mathrm{pen}(\boldsymbol{\beta}) \neq \varnothing.$$

*(b) Analytic characterization: The pattern of $\boldsymbol{\beta} \in \mathbb{R}^p$ is accessible with respect to $\boldsymbol{X}$ and* $\lambda$pen *if and only if for any $\boldsymbol{b} \in \mathbb{R}^p$ we have*

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{b} \implies \mathrm{pen}(\boldsymbol{\beta}) \leq \mathrm{pen}(\boldsymbol{b}).$$

Based on Proposition 4.4.2, one may see that the notion of accessibility does not depend on the tuning parameter $\lambda$.

### 5.4.2 Noiseless recovery condition

The solution path for a penalized estimator is defined as the curve $0 < \lambda \mapsto \widehat{\boldsymbol{\beta}}(\lambda)$ where $\widehat{\boldsymbol{\beta}}(\lambda)$ is the unique element of $S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$ for fixed $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. The solution paths for the generalized LASSO or OSCAR and the Clustered LASSO are studied in [179] or [173], respectively. Based on this notion, below we define the noiseless recovery condition. Note that the following definition does not require the uniqueness of an estimator.

**Definition 5.4.2** (Noiseless recovery condition)**.** *Let* pen *be a polyhedral gauge, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. We say that the pattern of $\boldsymbol{\beta}$ satisfies the noiseless recovery condition with respect to $\boldsymbol{X}$ and* pen *if*

$$\exists \lambda > 0, \exists \widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{X}\boldsymbol{\beta}) \text{ such that } \widehat{\boldsymbol{\beta}} \overset{\mathrm{pen}}{\sim} \boldsymbol{\beta}.$$

This condition generalizes the noiseless pattern recovery condition for SLOPE (4.3.2).

**Remark 5.4.1.** $\boldsymbol{\beta} = \mathbf{0}$ satisfies the noiseless recovery condition with respect to $\boldsymbol{X}$ and pen, because $\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}$ and $\mathbf{0} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\mathbf{0})$. In other words, the noiseless recovery condition means that in the noiseless case when $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$, in the solution path, one may pick such tuning parameter $\lambda$, that the minimizer has the same pattern as $\boldsymbol{\beta}$.

**Example 5.4.2.** (Solution path for LASSO):
Consider the LASSO optimization problem for

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

Then the LASSO solution path $0 < \lambda \mapsto \hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\lambda)$ is the following curve:

### 5.4.3 Irrepresentability Condition for polyhedral gauges

The following theorem is one of the main results of this chapter. Its proof is based on Theorem 5.2.2.

**Theorem 5.4.3.** *Let pen be a polyhedral norm. Let $\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\beta} \neq \mathbf{0}$. Then the existence of $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{X}\boldsymbol{\beta})$ recovering the pattern of $\boldsymbol{\beta}$ is equivalent to*

$$\boldsymbol{X}'\boldsymbol{X}\mathrm{lin}(C_\beta) \cap F_\beta \neq \varnothing.$$

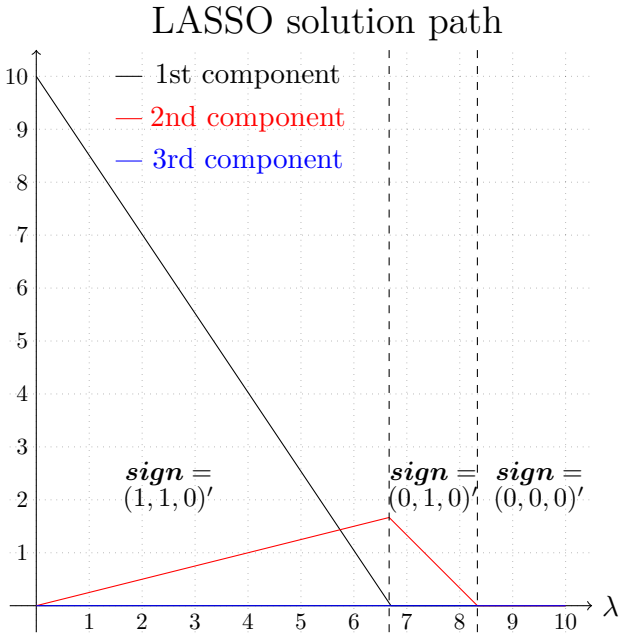



Figure 5.6: Coordinates and the sign of $\hat{\boldsymbol{\beta}}^{\text{LASSO}}$ as the functions of $\lambda > 0$: $\hat{\beta}_1^{\text{LASSO}}(\lambda)$ (black curve), $\hat{\beta}_2^{\text{LASSO}}(\lambda)$ (red) and $\hat{\beta}_3^{\text{LASSO}}(\lambda)$ (blue). Note that $\boldsymbol{sign}(\boldsymbol{\beta}) = (1, 0, 0)'$ is not recovered for any $\lambda > 0$.

*Proof.* ( $\Longrightarrow$ ):
Let $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{X}\boldsymbol{\beta})$ and $F_{\widehat{\beta}} = F_\beta$. Then, by the subdifferential properties, $\frac{1}{\lambda}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}) \in F_\beta$.

$\widehat{\boldsymbol{\beta}}$ recovers the pattern of $\boldsymbol{\beta}$, thus we have $\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}} \in \text{lin}(C_\beta)$. Therefore $\frac{\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}}{\lambda} \in \text{lin}(C_\beta)$ and finally $\frac{1}{\lambda}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \in \boldsymbol{X}'\boldsymbol{X}\text{lin}(C_\beta)$.
( $\Longleftarrow$ ):
By Theorem 5.2.2, the pattern set $C_\beta$ is relatively open.
Let $\boldsymbol{X}'\boldsymbol{X}\text{lin}(C_\beta) \cap F_\beta \neq \varnothing$, i.e. there exists such $\boldsymbol{z} \in V_\beta$ that $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{z} \in F_\beta$. Consider $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} - \lambda\boldsymbol{z}$. Since $\boldsymbol{\beta} \in C_\beta = \text{ri}(C_\beta)$ and $\boldsymbol{z} \in \text{lin}(C_\beta)$, then for $\lambda$ small enough we have $F_{\widehat{\beta}} = F_\beta$. Moreover, we obtain $\frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{z} \in F_\beta$. Therefore $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda pen}(\boldsymbol{X}\boldsymbol{\beta})$, which ends the proof. □

Below we provide a geometrical characterization of the noiseless recovery condition. Neither the above definition nor the geometrical characterization provide an analytic expression for checking the noiseless recovery condition, but for some of the penalized estimators analogous formulas have been already given. For example, when $\text{pen} = \|.\|_1$, the noiseless recovery condition is equivalent to

$$\|\boldsymbol{X}'(\boldsymbol{X}_I')^+\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty \leq 1 \text{ and } \boldsymbol{sign}(\boldsymbol{\beta}_I) \in \text{row}(\boldsymbol{X}_I), \tag{5.4.1}$$

where $I = \{i \in [p] : \ \beta_i \neq 0\}$ and $\boldsymbol{X}_I$ is the matrix whose columns are $(\boldsymbol{X}_j)_{j\in I}$. Note that under the assumption that $\ker(\boldsymbol{X}_I) = \{\boldsymbol{0}\}$ we obtain $\boldsymbol{sign}(\boldsymbol{\beta}_I) \in \text{row}(\boldsymbol{X}_I)$ and the expression (5.4.1) coincides with the irrepresentability condition: $\|\boldsymbol{X}_{\overline{I}}'\boldsymbol{X}_I(\boldsymbol{X}_I'\boldsymbol{X}_I)^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty \leq 1$ where $\boldsymbol{X}_{\overline{I}}$ is a matrix whose columns are $(\boldsymbol{X}_j)_{j\notin I}$ [41, 183, 194, 192]. Thus, the well known irrepresentability condition for LASSO can be thought of as an analytical shortcut for checking the noiseless recovery condition, see Figure 5.6. Indeed, in the above example, we have $\|\boldsymbol{X}_{\overline{I}}'\boldsymbol{X}_I(\boldsymbol{X}_I'\boldsymbol{X}_I)^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty = 30/29 > 1$ and based on Figure 5.6, one may observe that

the noiseless recovery condition does not hold for $\boldsymbol{\beta}$. For SLOPE with $\boldsymbol{M} = \mathrm{patt}(\boldsymbol{\beta})$, the noiseless recovery condition is equivalent to

$$J^*_{\boldsymbol{\Lambda}}(\boldsymbol{X}'(\widetilde{\boldsymbol{X}}'_{\mathbf{M}})^+\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}) \leq 1 \text{ and } \widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}} \in \mathrm{row}(\widetilde{\boldsymbol{X}}_{\mathbf{M}}),$$

where $J_{\boldsymbol{\Lambda}}$ is the sorted $\ell_1$ norm, $\widetilde{\boldsymbol{X}}_{\mathbf{M}}$ is the clustered matrix and $\widetilde{\boldsymbol{\Lambda}}_{\mathbf{M}}$ is the clustered parameter, cf. Definition 2.2.5. In Appendix we also provide an analytic characterization of noiseless pattern recovery when the penalty term is the supremum norm. Below we show that

(a) The noiseless recovery condition is necessary to recover the pattern with a probability larger than 1/2, see Theorem 5.4.4.

(b) Thresholded penalized estimators can recover the pattern of $\boldsymbol{\beta}$ under much weaker condition than the noiseless recovery condition, see Section 5.5.

**Theorem 5.4.4.** *Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a fixed matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\varepsilon}$ follows a symmetric distribution. Let* pen *be a polyhedral gauge. If $\boldsymbol{\beta}$ does not satisfy the noiseless recovery condition with respect to $\boldsymbol{X}$ and* pen*, then*

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

**Corollary 5.4.1.** *If the noiseless recovery condition does not hold for the LASSO (for example, when $\|\boldsymbol{X}'_{\overline{I}}\boldsymbol{X}_I(\boldsymbol{X}'_I\boldsymbol{X}_I)^{-1}\boldsymbol{sign}(\boldsymbol{\beta}_I)\|_\infty > 1$), the following holds*

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

The above result extends the Theorem 2 from [183], which shows that $\mathbb{P}(\boldsymbol{sign}(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}(\lambda)) = \boldsymbol{sign}(\boldsymbol{\beta})) \leq 1/2$ for fixed $\lambda > 0$.
If $\boldsymbol{\beta}$ satisfies the noiseless recovery condition with respect to $\boldsymbol{X}$ and pen, then $\boldsymbol{\beta}$ is accessible with respect to $\boldsymbol{X}$ and pen, by taking $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ in the definition of accessibility. In the following section, we show that thresholded penalized least-squares estimators recover the pattern of $\boldsymbol{\beta}$ under the accessibility condition.

## 5.5 Pattern recovery by thresholded estimators

In practice, some additional information about $\boldsymbol{\beta}$ may be priorly known, e.g. its sparsity. Therefore it is quite natural to threshold small components of $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ and to consider the thresholded LASSO estimator $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO},\tau}$ for some threshold $\tau \geq 0$:

**Definition 5.5.1** (Thresholded LASSO)**.** *The thresholded LASSO [174] is defined in a following way [41, Sec. 2.9]:*

$$\hat{\beta}_i^{\mathrm{LASSO},\tau} = \begin{cases} \hat{\beta}_i^{\mathrm{LASSO}} \text{ if } |\hat{\beta}_i^{\mathrm{LASSO}}| > \tau, \\ 0 \text{ otherwise.} \end{cases} \tag{5.5.1}$$

Moreover, if the threshold is appropriately selected, the estimator allows to recover $\boldsymbol{sign}(\boldsymbol{\beta})$ under weaker conditions than LASSO itself [174]. We aim at generalizing this property to the class of penalized estimators with polyhedral penalty. Before introducing the notion of a thresholded estimator, recall that for any threshold $\tau \geq 0$, the inclusion $\partial\|.\|_1(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}) \subseteq \partial\|.\|_1(\hat{\boldsymbol{\beta}}^{\mathrm{LASSO},\tau})$ occurs. This last inclusion is the keystone concept to introduce the notion of a thresholded estimator as defined in Definition 2.3.23.
The notion of accessibility introduced for penalized estimators in Section 5.4 also covers the thresholded estimators as can be seen below.

**Proposition 5.5.1.** *Let* pen *be a real-valued polyhedral gauge,* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *and* $\boldsymbol{\beta} \in \mathbb{R}^p$*. Then*

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

According to Propositions 4.4.2 and 5.5.1, if the pattern of $\boldsymbol{\beta}$ is not accessible with respect to $\boldsymbol{X}$ and $\lambda$pen, i.e. there exists $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta}$ and $\text{pen}(\boldsymbol{b}) < \text{pen}(\boldsymbol{\beta})$, then for any $\boldsymbol{y} \in \mathbb{R}^n$, $\lambda > 0$, and $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$ we have $\partial\text{pen}(\widehat{\boldsymbol{\beta}}) \not\subseteq \partial\text{pen}(\boldsymbol{\beta})$. Consequently, no penalized nor thresholded penalized estimator can recover the pattern of $\boldsymbol{\beta}$.
On the other hand, if the accessibility condition is satisfied, then both penalized and thresholded penalized estimator can recover the pattern of $\boldsymbol{\beta}$ with different choices of $\boldsymbol{y}$. However, in practice, instead of choosing an appropriate $\boldsymbol{y}$ to recover the pattern of $\boldsymbol{\beta}$, the response of a linear regression model is being used to infer this pattern.
In this direction, by Theorem 5.4.4, if $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, then the recovery of the pattern of $\boldsymbol{\beta}$ with probability larger than $1/2$ requires the noiseless recovery condition, which is stronger than the accessibility condition. This result remains true for any symmetric and continuous noise. In Theorem 5.5.1, we relax the stringent noiseless recovery condition by considering a thresholded estimator. Before stating this theorem, we introduce the following class of thresholded estimators.

**Definition 5.5.2** ($\tau$-thresholded penalized estimator)**.** *Let* pen *be a real-valued polyhedral gauge,* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$*,* $\boldsymbol{Y} \in \mathbb{R}^n$ *and* $\lambda \geq 0$*. Given* $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{Y})$*, we say that* $\widetilde{\boldsymbol{\beta}}^{\tau}$ *is a* $\tau$*-thresholded estimator of* $\widehat{\boldsymbol{\beta}}$ *if*

*(a)* $\partial\text{pen}(\widehat{\boldsymbol{\beta}}) \subseteq \partial\text{pen}(\widetilde{\boldsymbol{\beta}}^{\tau})$*,*

*(a)* $\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^{\tau}\|_\infty \leq \tau$*,*

*(a)* $\dim(\partial\text{pen}(\boldsymbol{b})) \leq \dim(\partial\text{pen}(\widetilde{\boldsymbol{\beta}}^{\tau}))$ *for all* $\boldsymbol{b}$ *with* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{b}\|_\infty \leq \tau$*.*

The thresholded LASSO is, in fact, an example of a $\tau$-thresholded estimator with the threshold $\tau$. Another example of a $\tau$-thresholded estimator when the penalty term is the supremum norm, is given in Algorithm 1. Theorem 5.5.1 shows that a thresholded estimator recovers the pattern of $\boldsymbol{\beta}$ under the accessibility condition and the assumption that the signal is large enough, as formalized in the following theorem.

**Theorem 5.5.1.** *Let* pen *be a real-valued polyhedral gauge,* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$*,* $\boldsymbol{\beta} \in \mathbb{R}^p$*, and* $\lambda > 0$*. Assume that the uniform uniqueness holds, i.e. for any* $\boldsymbol{y} \in \mathbb{R}^n$*, the set* $S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$ *consists of one element* $\widehat{\boldsymbol{\beta}}(\boldsymbol{y})$*. For* $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ *and for* $r \in \mathbb{N}_+$ *set* $\boldsymbol{y}^{(r)} = \boldsymbol{X}(r\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$*. If* $\text{pen}(\boldsymbol{b}) \geq \text{pen}(\boldsymbol{\beta})$ *for any* $\boldsymbol{b} \in \mathbb{R}^p$ *with* $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta}$*, then there exists* $r_0 \in \mathbb{N}_+$ *and* $\tau \geq 0$ *such that for all* $r \geq r_0$

$$\begin{cases}\partial\text{pen}(\boldsymbol{b}) \subseteq \partial\text{pen}(\boldsymbol{\beta}) \text{ for any } \boldsymbol{b} \in \overline{B}_\infty(\widehat{\boldsymbol{\beta}}^{(r)}, \tau)\\ \exists \boldsymbol{b}_0 \in \overline{B}_\infty(\widehat{\boldsymbol{\beta}}^{(r)}, \tau) \text{ such that } \boldsymbol{b}_0 \overset{\text{pen}}{\sim} \boldsymbol{\beta}\end{cases}$$

*Consequently, a* $\tau$*-thresholded penalized estimator* $\widetilde{\boldsymbol{\beta}}^{\tau}(\boldsymbol{y}^{(r)})$ *recovers the pattern of* $\boldsymbol{\beta}$*.*

Similar results in which non-null components are large enough (i.e., $r \geq r_0$ in Theorem 5.5.1) are given in [174] and [54]. In particular, Theorem 5.5.1 corroborates Theorem 1 in [174], which proves that the thresholded LASSO estimator recovers the sign of $\boldsymbol{\beta}$ once the accessibility condition holds and non-null components of $\boldsymbol{\beta}$ are large enough. Similarly as thresholded LASSO, while pen $= \|.\|_\infty$, a $\tau$-estimator can be explicitly computed by Algorithm 1.

**Algorithm 1** Thresholded penalized estimator when the penalty term is the $\ell_\infty$ norm:

**Require:** estimate: $\widehat{\boldsymbol{\beta}}$, threshold $\tau \geq 0$.
**if** $\|\widehat{\boldsymbol{\beta}}\|_\infty \leq \tau$ **then**
$\widetilde{\boldsymbol{\beta}}^\tau \leftarrow \mathbf{0}$.
**else**

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

**end if**
**return** $\widetilde{\boldsymbol{\beta}}^\tau$

## 5.6 Full characterization of the uniform uniqueness

For the pattern recovery by the $\tau$-thresholded penalized estimator, in Theorem 5.5.1 we assume the uniform uniqueness. For that reason, below we provide a a necessary and sufficient condition for uniform uniqueness of the penalized optimization problem (5.1.1) in Theorem 5.6.1. This theorem extends the Theorem 1 from [156] to all polyhedral gauges.

**Theorem 5.6.1** (Necessary and sufficient condition for uniform uniqueness)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\lambda > 0$. Let* pen *be a polyhedral gauge, i.e.,* $\mathrm{pen}(\boldsymbol{x}) = \max\{0, \boldsymbol{u}_1'\boldsymbol{x}, \ldots, \boldsymbol{u}_l'\boldsymbol{x}\}$ *for some $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p$. For $\boldsymbol{y} \in \mathbb{R}^n$ consider the following optimization problem*

$$S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y}) = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda\mathrm{pen}(\boldsymbol{b}). \tag{5.1.1}$$

*Then the solution to the above problem is unique, i.e, $S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y})$ is a singleton for all $\boldsymbol{y} \in \mathbb{R}^n$ if and only if* $\mathrm{row}(\boldsymbol{X})$ *does not intersect a face of the polytope $B^* = \mathrm{conv}\{\mathbf{0}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_l\}$ whose dimension is smaller than* $\dim(\ker(\boldsymbol{X}))$.

Note that every face $F$ of $B^*$ satisfies

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

where $\mathrm{codim}(F) = p - \dim(F)$. For a better explanation of the non-uniqueness of the estimator, below we give an example for the generalized LASSO with $\mathrm{pen}(\boldsymbol{b}) = \|\boldsymbol{D}\boldsymbol{b}\|_1$. Note that if $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}) = \{\mathbf{0}\}$, then for every $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\|\mathbf{D}\cdot\|_1}(\boldsymbol{y})$ and $\boldsymbol{g} \in \ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D})$ we have $\widehat{\boldsymbol{\beta}} + \boldsymbol{g} \in S_{\mathbf{X},\lambda\|\mathbf{D}\cdot\|_1}(\boldsymbol{y})$, too. Thus for every $\boldsymbol{y} \in \mathbb{R}^n$ the minimizer is not unique. Consequently, $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}) = \{\mathbf{0}\}$ is a necessary condition for uniform uniqueness, yet, it is not sufficient, as illustrated in the example given in [9], which we revisit below.

**Example 5.6.2.** [9, p. 19] Consider the following optimization problem

$$\operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \frac{1}{2}\|\boldsymbol{D}\boldsymbol{b}\|_1,$$

where

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & 1 \\ \sqrt{2} & 0 & 0 \end{pmatrix},\ \boldsymbol{D} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \text{ and } \boldsymbol{Y} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

We have $S_{\mathbf{X},\frac{1}{2}\|\mathbf{D}.\|_1}(\boldsymbol{Y}) = \mathrm{conv}\{(0,1/2,0)', (0,0,1/2)'\}$ [9, p. 19]. Since

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

then $B^* = \mathrm{conv}\{\pm(4,2,2)', \pm(2,2,0)', \pm(2,0,2)'\}$. Because the vertex $F = (4,2,2)'$ is a face of $B^*$ which lies in $\mathrm{row}(\boldsymbol{X})$ and satisfies $\dim(F) = 0 < 1 = \dim(\ker(\boldsymbol{X}))$ then, according to Theorem 5.6.1, the uniform uniqueness cannot hold. This complies with the fact that $S_{\mathbf{X},\frac{1}{2}\|\mathbf{D}.\|_1}(\boldsymbol{Y})$ is not a singleton.

When $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}) = \{\mathbf{0}\}$, in broad generality, the set of generalized LASSO solutions is a polytope (a bounded polyhedron) [9, Proposition 4.3.] and its extremal points can be explicitly computed [62]. This description is relevant when this set is not a singleton.

**Lemma 5.6.1.** *Consider the function*

$$f(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Xb}\|_2^2 + \mathrm{pen}(\boldsymbol{b}),$$

*where* $\mathrm{pen}(\boldsymbol{b}) = \max\{0, \boldsymbol{u}_1'\boldsymbol{b}, \ldots, \boldsymbol{u}_l'\boldsymbol{b}\}$ *is a polyhedral gauge. Then the condition* $\ker(\boldsymbol{X}) \cap \ker(\boldsymbol{D}) = \{\mathbf{0}\}$ *is necessary for the uniform uniqueness of the minimizer of* $f$*.*

*Proof.* Let $\mathbf{0} \neq \boldsymbol{h} \in \ker(\boldsymbol{X}) \cap \ker(\mathrm{pen})$ and let $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$. Then

$$f(\widehat{\boldsymbol{\beta}}) \leq f(\widehat{\boldsymbol{\beta}} + \boldsymbol{h}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} + \boldsymbol{h})\|_2^2 + \mathrm{pen}(\widehat{\boldsymbol{\beta}} + \boldsymbol{h}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \max\{0, \boldsymbol{u}_1'(\widehat{\boldsymbol{\beta}} + \boldsymbol{h}), \ldots, \boldsymbol{u}_l'(\widehat{\boldsymbol{\beta}} + \boldsymbol{h})\}.$$

Since $\mathrm{pen}(\boldsymbol{h}) = 0$, we have $\boldsymbol{u}_i'\boldsymbol{h} \leq 0$ for every $i \in [l]$, hence

$$f(\widehat{\boldsymbol{\beta}} + \boldsymbol{h}) \leq \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \max\{0, \boldsymbol{u}_1'\widehat{\boldsymbol{\beta}}, \ldots, \boldsymbol{u}_l'\widehat{\boldsymbol{\beta}}\} = f(\widehat{\boldsymbol{\beta}}).$$

Therefore $\widehat{\boldsymbol{\beta}}$ is not a unique minimizer of $f$. □

## 5.7 Numerical experiments

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

- The matrix $\boldsymbol{X} = (\boldsymbol{X}_1|\ldots|\boldsymbol{X}_{150}) \in \mathbb{R}^{100\times 150}$ has i.i.d. $\mathcal{N}(0, 1/100)$ entries.
- The random noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0,1)$ entries.

### 5.7.1 Numerical experiments for LASSO

For LASSO, the noiseless recovery condition and the accessibility condition depend on $\boldsymbol{\beta}$ through $\boldsymbol{sign}(\boldsymbol{\beta}) \in \{-1, 0, 1\}^p$. Moreover, since the distribution of $\boldsymbol{X}$ is invariant by permutations and sign changes of the columns of $\boldsymbol{X}$, then the probability that a $k-$sparse vector satisfies the noiseless recovery condition is equivalent to the irrepresentability condition:

$$\mathbb{P}_{\mathbf{X}}\left(\|\boldsymbol{X}'(\boldsymbol{X}_I')^+\mathbf{1}_k\|_\infty \leq 1 \text{ and } \mathbf{1}_k \in \mathrm{row}(\boldsymbol{X}_I)\right),$$

where $I = [k]$ and $\mathbf{1}_k = (1, \ldots, 1)' \in \mathbb{R}^k$. Moreover, the accessibility condition is satisfied with probability

$$\mathbb{P}_{\mathbf{X}}(\min\{\|\boldsymbol{\gamma}\|_1 : \ \boldsymbol{X\gamma} = \boldsymbol{X}_I\mathbf{1}_k\} = k).$$

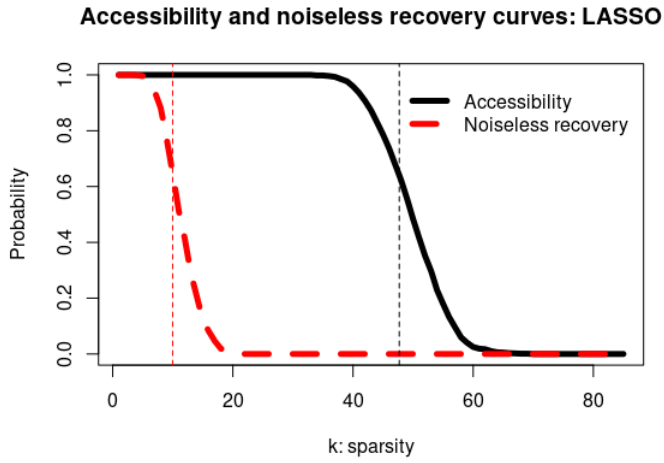



Figure 5.7: Probability of the noiseless recovery condition and the accessibility condition as functions of the support size $k$. The highlited values of $k$ are $k_1 = 50/\log(150) \approx 9.9$ and $k_2 = 100\rho_{DT}(2/3) \approx 47.8$, which are the preimages of 0.5 for the noiseless recovery curve [183] and the accessibility curve [59, 60], respectively.

Figure 5.7 provides these probabilities as functions of $k$.
Figure 5.8 illustrates sign recovery properties by LASSO and thresholded LASSO for a particular observation of $\boldsymbol{X} \in \mathbb{R}^{100\times 150}$, a particular observation of $\boldsymbol{Y} \in \mathbb{R}^{100}$ and for $k$-sparse $\boldsymbol{\beta} \in \mathbb{R}^{150}$ with $\beta_1 = \cdots = \beta_{k/2} = 20$ and $\beta_{k/2+1} = \cdots = \beta_k = -20$. Our examples are given for $k = 4$ and $k = 30$. For the LASSO estimator, we consider the following setting:

- LASSO with a large tuning parameter $\lambda = 2\sqrt{2\log(150)}$ (as suggested by Candès and Plan [42]).
- LASSO with a small tuning parameter; the one provided by SURE formula, which for a given $\boldsymbol{X}$ and $\boldsymbol{Y}$ minimizes the function $0 < \lambda \mapsto \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2 + |\{i \in [p] : \widehat{\beta}_i(\lambda) \neq 0\}|$ where $\widehat{\boldsymbol{\beta}}(\lambda)$ is the LASSO estimator (see *e.g.* [179] or [181]).

### 5.7.2 Numerical experiments when the penalty term is the supremum norm

For the $\ell_\infty$ regularization, the noiseless recovery condition and the accessibility condition depend on $\boldsymbol{\beta}$ through $\boldsymbol{sign}^\infty(\boldsymbol{\beta}) \in \{-1, *, 1\}^p$. Same as for LASSO, since the distribution of $\boldsymbol{X}$ is invariant by permutations and sign changes of the columns of $\boldsymbol{X}$, then the probability that a non-zero vector having $k$ non-maximal components in absolute value satisfies the noiseless recovery condition is given by

$$\mathbb{P}_{\mathbf{X}}(\widetilde{\boldsymbol{X}}'(\widetilde{\boldsymbol{X}}')^+\boldsymbol{e}_1 = \boldsymbol{e}_1) \text{ where } \widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1|\boldsymbol{X}_I) \text{ with } \widetilde{\boldsymbol{X}}_1 = \sum_{i=1}^{p-k} \boldsymbol{X}_i \text{ and } I = \{p-k+1, \dots, p\}.$$

An explicit formula for checking the noiseless recovery condition is given in the Appendix. Moreover, the accessibility condition is satisfied with probability

$$\mathbb{P}_{\mathbf{X}}(\min\{\|\boldsymbol{\gamma}\|_\infty : \ \boldsymbol{X}\boldsymbol{\gamma} = \widetilde{\boldsymbol{X}}_1\} = 1).$$

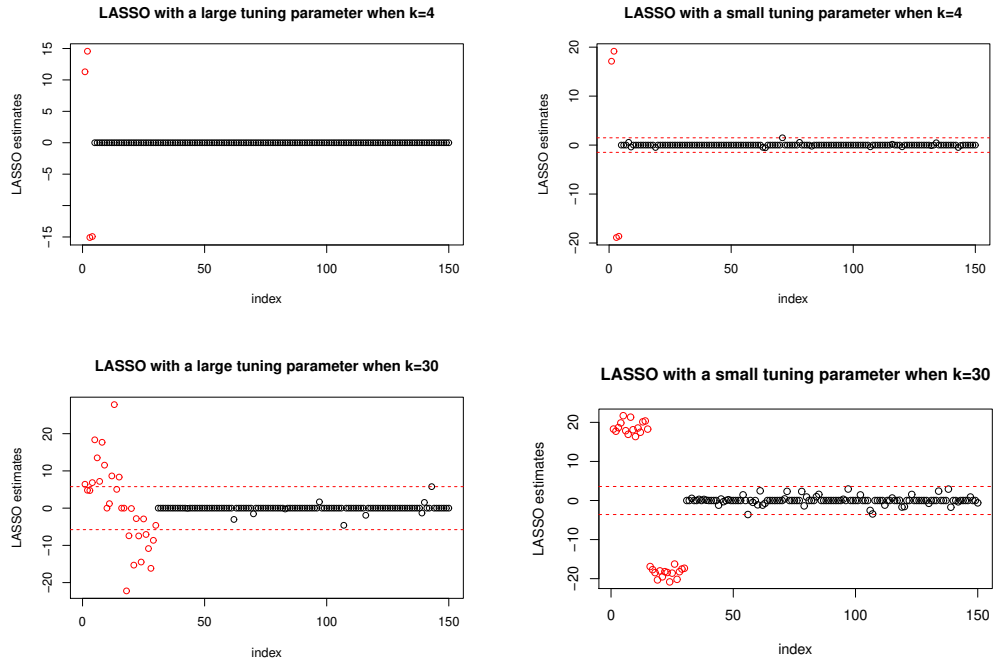



Figure 5.8: Illustrations of sign recovery by LASSO and thresholded LASSO. On the top, when $k = 4$, both the noiseless recovery condition and the accessibility condition hold. Thus, both LASSO and thresholded LASSO can recover the sign of $\boldsymbol{\beta}$. With the large tuning parameter $\lambda = 2\sqrt{2\log(150)}$ the sign of $\boldsymbol{\beta}$ is recovered both by LASSO and thresholded LASSO (top left). When the tuning parameter is small (computed by SURE), some null components of $\boldsymbol{\beta}$ are not correctly estimated at 0 (black points outside the x-axis), but there exists a threshold, for which the thresholded LASSO recovers the sign of $\boldsymbol{\beta}$ (top right). On the bottom, when $k = 30$, the accessibility condition holds but the noiseless recovery condition does not hold, thus thresholded LASSO can recover the sign of $\boldsymbol{\beta}$ but LASSO cannot. When the tuning parameter is large: $\lambda = 2\sqrt{2\log(150)}$, both LASSO and thresholded LASSO fail to recover the sign of $\boldsymbol{\beta}$ (bottom left). When the tuning parameter is small, some null components of $\boldsymbol{\beta}$ are not correctly estimated at 0, but there exists a threshold, for which the thresholded LASSO recovers the sign of $\boldsymbol{\beta}$ (bottom right).

Figure 5.9 provides both the probability of the accessibility condition and the probability of the noiseless recovery condition as functions of $k$.

In Figure 5.10 we illustrate the pattern recovery properties by a penalized estimator and a thresholded penalized estimator for the supremum norm. Specifically, $\boldsymbol{\beta} \in \mathbb{R}^{150}$ satisfies $\beta_1 = \cdots = \beta_{60} = 20$, $\beta_{61} = \cdots = \beta_{120} = -20$ and $\beta_{121} = \cdots = \beta_{150} = 0$. The tuning parameter is given by the SURE formula, which for a given $\boldsymbol{X}$ and $\boldsymbol{Y}$ minimizes the function $0 < \lambda \mapsto \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2 + |\{i \in [p] : \ |\widehat{\beta}_i| < \|\widehat{\boldsymbol{\beta}}(\lambda)\|_\infty\}|$, where $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is the unique element of $S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{Y})$ (see *e.g.* [132] or [181]).

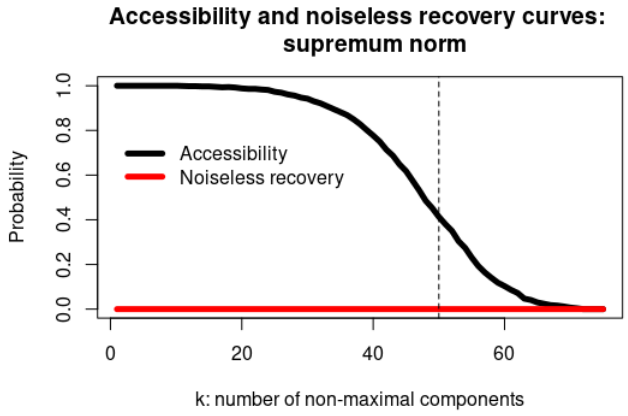



Figure 5.9: Probability of the noiseless recovery and the probability of the accessibility condition as functions of the number $k$ of non-maximal components in absolute value. The value $k = 50$ [4] provides, approximately, the preimage of 0.5 for the accessibility curve.

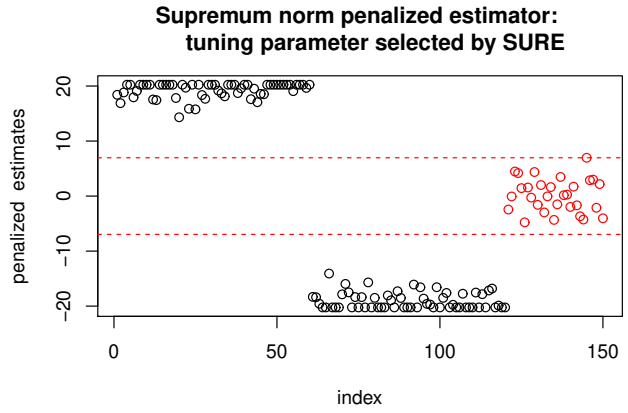



Figure 5.10: Illustrations of the pattern recovery by a penalized estimator and a thresholded penalized for the supremum norm. When $k = 30$, the accessibility condition holds, but the noiseless recovery condition does not hold. Thus, as illustrated on this picture, the recovery of the pattern of $\beta$ requires thresholding the estimator.

## 5.8 Appendix

### 5.8.1 Facts about real-valued polyhedral gauges

Lemma 5.8.1 characterizes the connection between elements of a certain class of convex functions (which encompasses polyhedral gauges) and the faces of a related polytope. It is needed to prove Theorem 5.6.1.

**Lemma 5.8.1.** *Let* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^p$, $P = \mathrm{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ *and*

$$\phi(\boldsymbol{x}) = \max\{\boldsymbol{v}_1'\boldsymbol{x}, \ldots, \boldsymbol{v}_k'\boldsymbol{x}\} \text{ for } \boldsymbol{x} \in \mathbb{R}^p.$$

*Then the subdifferential of* $\phi$ *at* $\boldsymbol{x}$ *is a face of* $P$ *and is given by*

$$\partial\phi(\boldsymbol{x}) = \mathrm{conv}\{\boldsymbol{v}_l : \ l \in I_\phi(\boldsymbol{x})\} = \{\boldsymbol{s} \in P : \ \boldsymbol{s}'\boldsymbol{x} = \phi(\boldsymbol{x})\}, \text{ where } I_\phi(\boldsymbol{x}) = \{l \in [k] : \ \boldsymbol{v}_l'\boldsymbol{x} = \phi(\boldsymbol{x})\}.$$

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

*Proof.* The fact that $\partial\phi(\boldsymbol{x}) = \operatorname{conv}\{\boldsymbol{u}_l : \ l \in I_\phi(\boldsymbol{x})\}$ can be found in [101, p. 259]. To prove the second equality, we consider the following. If $l \in I_\phi(\boldsymbol{x})$, then by definition of $I_\phi(\boldsymbol{x})$, $\boldsymbol{v}_l'\boldsymbol{x} = \phi(\boldsymbol{x})$ and thus $\boldsymbol{v}_l \in \{\boldsymbol{s} \in P : \ \boldsymbol{s}'\boldsymbol{x} = \phi(\boldsymbol{x})\}$. Since the latter set is convex, one may deduce that

$$\operatorname{conv}\{\boldsymbol{v}_l : \ l \in I_\phi(\boldsymbol{x})\} \subseteq \{\boldsymbol{s} \in P : \ \boldsymbol{s}'\boldsymbol{x} = \phi(\boldsymbol{x})\}.$$

For the other inclusion, assume that $\boldsymbol{s} \in P$ is such that $\boldsymbol{s} \notin \operatorname{conv}\{\boldsymbol{v}_l : \ l \in I_\phi(\boldsymbol{x})\}$. Then we have $\boldsymbol{s} = \sum_{l=1}^{k} \alpha_l \boldsymbol{v}_l$ where $\alpha_1, \ldots, \alpha_k \geq 0$, $\sum_{l=1}^{k} \alpha_l = 1$ and $\alpha_{l_0} > 0$ for some $l_0 \notin I_\phi(\boldsymbol{x})$. Since $\boldsymbol{v}_l'\boldsymbol{x} \leq \phi(\boldsymbol{x})$ for all $l \in [k]$ and $\boldsymbol{v}_{l_0}'\boldsymbol{x} < \phi(\boldsymbol{x})$, we also get

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

Consequently, $\boldsymbol{s}'\boldsymbol{x} \neq \phi(\boldsymbol{x})$ and thus

$$\{\boldsymbol{s} \in P : \ \boldsymbol{s}'\boldsymbol{x} = \phi(\boldsymbol{x})\} \subseteq \operatorname{conv}\{\boldsymbol{v}_l : \ l \in I_\phi(\boldsymbol{x})\}.$$

Therefore, $\partial\phi(\boldsymbol{x}) = \operatorname{conv}\{\boldsymbol{v}_l : \ l \in I_\phi(\boldsymbol{x})\} = \{\boldsymbol{s} \in P : \ \boldsymbol{s}'\boldsymbol{x} = \phi(\boldsymbol{x})\}$. Now we show that the subdifferentials of $\phi$ are the non-empty faces of $P$. Let $\boldsymbol{x} \in \mathbb{R}^p$. By definition of $\phi$, $\boldsymbol{v}_l'\boldsymbol{x} \leq \phi(\boldsymbol{x})$ for every $l \in [k]$, so the inequality $\boldsymbol{x}'\boldsymbol{s} \leq \phi(\boldsymbol{x})$ holds for all $\boldsymbol{s} \in P$. This implies that $\partial\phi(\boldsymbol{x})$ is a non-empty face of $P$. Now we show that any non-empty face $F$ of $P$ is equal to $\partial\phi(\boldsymbol{a})$ for some $\boldsymbol{a} \in \mathbb{R}^p$. Let $F = \{\boldsymbol{s} \in P : \ \boldsymbol{a}'\boldsymbol{s} = c\}$ be a non-empty face of $P$ where $\boldsymbol{a} \in \mathbb{R}^p$, $c \in \mathbb{R}$ and $\boldsymbol{a}'\boldsymbol{s} \leq c$ is a valid inequality for all $\boldsymbol{s} \in P$. We prove that $F = \partial\phi(\boldsymbol{a})$. Indeed, take any $\boldsymbol{s} \in F$. We get $\boldsymbol{a}'\boldsymbol{s} = c$ as well as $\boldsymbol{a}'\boldsymbol{s} \leq \phi(\boldsymbol{a})$ as shown above, implying that $c \leq \phi(\boldsymbol{a})$. Analogously, for any $\boldsymbol{s} \in \partial\phi(\boldsymbol{a})$, $\boldsymbol{a}'\boldsymbol{s} = \phi(\boldsymbol{a})$ as well as $\boldsymbol{a}'\boldsymbol{s} \leq c$ since $\partial\phi(\boldsymbol{a}) \subseteq P$, yielding $\phi(\boldsymbol{a}) \leq c$. Therefore one may deduce that $\phi(\boldsymbol{a}) = c$ and thus $F = \partial\phi(\boldsymbol{a})$. □

**Corollary 5.8.1.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\lambda > 0$. Let* pen *be a polyhedral gauge, i.e.,* $\operatorname{pen}(\boldsymbol{x}) = \max\{0, \boldsymbol{u}_1'\boldsymbol{x}, \ldots, \boldsymbol{u}_l'\boldsymbol{x}\}$ *for some $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p$. Then the polar set of the unit ball of* pen *equals*

$$B^* = \operatorname{conv}\{0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_l\}.$$

The following lemma, also needed to show Theorem 5.6.1, states that the fitted values are unique over all non-unique solutions of the penalized problem for a given $y$. It is a generalization of Lemma 1 in [178], which shows this fact for the special case of the LASSO.

**Lemma 5.8.2.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\lambda > 0$ and* pen *be a polyhedral gauge. Then $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ and* $\operatorname{pen}(\widehat{\boldsymbol{\beta}}) = \operatorname{pen}(\widetilde{\boldsymbol{\beta}})$ *for all $\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}} \in S_{\mathbf{X},\mathrm{pen}}(\boldsymbol{Y})$.*

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

$$\|\boldsymbol{Y} - \boldsymbol{X}\check{\boldsymbol{\beta}}\|_2^2 < \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|_2^2.$$

Moreover, by convexity of pen, we have $\operatorname{pen}(\check{\boldsymbol{\beta}}) \leq \frac{1}{2}(\operatorname{pen}(\widehat{\boldsymbol{\beta}}) + \operatorname{pen}(\widetilde{\boldsymbol{\beta}}))$. Consequently,

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\check{\boldsymbol{\beta}}\|_2^2 + \lambda\operatorname{pen}(\check{\boldsymbol{\beta}}) < \frac{1}{2}\left(\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\operatorname{pen}(\widehat{\boldsymbol{\beta}}) + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda\operatorname{pen}(\widetilde{\boldsymbol{\beta}})\right),$$

which contradicts both $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ being minimizers. Finally, $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ implies that $\operatorname{pen}(\widehat{\boldsymbol{\beta}}) = \operatorname{pen}(\widetilde{\boldsymbol{\beta}})$. □

### 5.8.2 Proofs

**Proof of Theorem 5.1.1**

For $\gamma \geq 0$, as $P_\gamma$ we denote the polyhedron:

$$P_\gamma = \{\boldsymbol{b} \in \mathbb{R}^p : \ \mathrm{pen}(\boldsymbol{b}) \leq \gamma\} = \{\boldsymbol{b} \in \mathbb{R}^p : \ \boldsymbol{u}_1'\boldsymbol{b} \leq \gamma, \ldots, \boldsymbol{u}_l'\boldsymbol{b} \leq \gamma\}.$$

**Lemma 5.8.3.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p$ and $R \geq 0$. Then the optimization problem*

$$\min \ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 \text{ subject to the constraint } \underbrace{\boldsymbol{u}_1'\boldsymbol{b} \leq R, \ldots, \boldsymbol{u}_l'\boldsymbol{b} \leq R}_{\boldsymbol{b} \in P_R}. \tag{5.8.1}$$

*has at least one solution.*

*Proof.* Let us set $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{b}$ in the optimization problem (5.8.1). Since the set $\boldsymbol{X}P_R$ is a closed convex set, the problem of minimizing $\|\boldsymbol{Y} - \boldsymbol{z}\|_2^2$ subject to the constraint $\boldsymbol{z} \in \boldsymbol{X}P_R$ has a unique solution $\overline{\boldsymbol{z}} \in \boldsymbol{X}P_R$. Consequently, $\overline{\boldsymbol{z}} = \boldsymbol{X}\overline{\boldsymbol{b}}$ for some $\overline{\boldsymbol{b}} \in P_R$. Finally, $\overline{\boldsymbol{b}}$ is a solution of the optimization problem (5.8.1). □

Let us revisit the polyhedron $P_\gamma$. As every polyhedron, $P_\gamma$ can be decomposed as the sum of its recession cone and a bounded polyhedron [193, Theorem 1.2. and Proposition 1.12.]. Therefore, for $\gamma = 1$ we have:

$$P_1 = \{\boldsymbol{b} \in \mathbb{R}^p : \ \boldsymbol{u}_1'\boldsymbol{b} \leq 0, \ldots, \boldsymbol{u}_l'\boldsymbol{b} \leq 0\} + E,$$

where $E$ is a polytope. Note that, for an arbitrary $\gamma \geq 0$, we have $P_\gamma = P_0 + \gamma E$.

*Proof of 5.1.1.* Let $m := \inf_{\boldsymbol{b} \in \mathbb{R}^p} f(\boldsymbol{b})$. Note that $0 \leq m \leq f(\boldsymbol{0}) = \frac{1}{2}\|\boldsymbol{Y}\|_2^2$. We want to show the existence of such $\boldsymbol{\beta}^* \in \mathbb{R}^p$ that $f(\boldsymbol{\beta}^*) = m$. If $f(\boldsymbol{0}) = m$, then we are done. Assume then that $f(\boldsymbol{0}) = m + \delta$ for some $\delta > 0$. By the definition of infimum, one may consider such sequence $(\boldsymbol{\beta}_k)_{k\geq 1}$ that $f(\boldsymbol{\beta}_k) \leq m + \frac{\delta}{k} \leq f(\boldsymbol{0})$ for every $k \geq 1$.
Step 1: Convergence of $(\boldsymbol{X}\boldsymbol{\beta}_k)_{k\geq 1}$ and $(\mathrm{pen}(\boldsymbol{\beta}_k))_{k\geq 1}$:
At first we show that $(\boldsymbol{X}\boldsymbol{\beta}_k)_{k\geq 1}$ is bounded. Indeed, we have

$$0 \leq \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_k\|_2^2 \leq f(\boldsymbol{\beta}_k) \leq \frac{1}{2}\|\boldsymbol{Y}\|_2^2.$$

Thus $\|\boldsymbol{X}\boldsymbol{\beta}_k\|_2 \leq \|\boldsymbol{Y}\|_2 + \|\boldsymbol{X}\boldsymbol{\beta}_k - \boldsymbol{Y}\|_2 \leq 2\|\boldsymbol{Y}\|_2$ and the sequence $(\boldsymbol{X}\boldsymbol{\beta}_k)_{k\geq 1}$ is bounded. Thus, if it is not convergent, then there exist two subsequences $(\boldsymbol{X}\boldsymbol{\beta}_{r_k})_{k\geq 1} \overset{k\to\infty}{\longrightarrow} \boldsymbol{r}$ and $(\boldsymbol{X}\boldsymbol{\beta}_{s_k})_{k\geq 1} \overset{k\to\infty}{\longrightarrow} \boldsymbol{s}$ converging to two different limits. By the strict convexity of $\varphi(\boldsymbol{t}) := \|\boldsymbol{Y} - \boldsymbol{t}\|_2^2$, one may deduce the following inequality:

$$\lim_{k\to\infty} \|\boldsymbol{Y} - \boldsymbol{X}\frac{\boldsymbol{\beta}_{r_k} + \boldsymbol{\beta}_{s_k}}{2}\|_2^2 = \|\boldsymbol{Y} - \frac{\boldsymbol{r} + \boldsymbol{s}}{2}\|_2^2 < \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{r}\|_2^2 + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{s}\|_2^2.$$

Moreover, the convexity of pen yields

$$\mathrm{pen}\left(\frac{\boldsymbol{\beta}_{r_k} + \boldsymbol{\beta}_{s_k}}{2}\right) \leq \frac{1}{2}\left(\mathrm{pen}(\boldsymbol{\beta}_{r_k}) + \mathrm{pen}(\boldsymbol{\beta}_{s_k})\right).$$

Consequently,

$$\limsup_{k\to\infty} f\left(\frac{\boldsymbol{\beta}_{r_k} + \boldsymbol{\beta}_{s_k}}{2}\right) < \frac{1}{2}\left(f(\boldsymbol{\beta}_{r_k}) + f(\boldsymbol{\beta}_{s_k})\right) = m,$$

which contradicts $m = \inf_{\boldsymbol{b}\in\mathbb{R}^p} f(\boldsymbol{b})$. Therefore the sequence $(\boldsymbol{X}\boldsymbol{\beta}_k)_{k\geq 1}$ is convergent and thus $(\text{pen}(\boldsymbol{\beta}_k))_{k\geq 1}$ converges, too.
Step 2: Existence of the minimizer:
Denote $\boldsymbol{g} := \lim_{k\to\infty} \boldsymbol{X}\boldsymbol{\beta}_k$ and $\gamma := \lim_{k\to\infty} \text{pen}(\boldsymbol{\beta}_k)$. Then we have

$$m = \inf_{\boldsymbol{b}\in\mathbb{R}^p} f(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{g}\|_2^2 + \lambda\gamma.$$

Let $\widehat{\boldsymbol{\beta}}$ be an arbitrary solution of (5.8.1) with $R = \gamma$. We are going to prove that $f(\widehat{\boldsymbol{\beta}}) = \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{g}\|_2^2 + \lambda\gamma = m$.
Case 1:
If $\gamma > 0$, then for $k$ large enough such that $\text{pen}(\boldsymbol{\beta}_k) > 0$ one may set $\boldsymbol{v}_k := \frac{\gamma}{\text{pen}(\boldsymbol{\beta}_k)}\boldsymbol{\beta}_k$. We have $\text{pen}(\boldsymbol{v}_k) = \gamma$ and thus $\boldsymbol{v}_k \in P_\gamma$. Consequently, by definition of $\widehat{\boldsymbol{\beta}}$, we have $\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{v}_k\|_2^2 \geq \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}_k\|_2^2$. Therefore

$$f(\widehat{\boldsymbol{\beta}}) = \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\text{pen}(\widehat{\boldsymbol{\beta}}) \leq \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{v}_k\|_2^2 + \lambda\gamma \overset{k\to\infty}{\longrightarrow} \|\boldsymbol{Y}-\boldsymbol{g}\|_2^2 + \lambda\gamma = m.$$

Case 2:
Let $\gamma = 0$. Because $P_1 = P_0 + E$, where $E$ is a bounded polyhedron, one may write $\boldsymbol{\beta}_k = \boldsymbol{v}_k + \text{pen}(\boldsymbol{\beta}_k)\boldsymbol{w}_k$ where $\boldsymbol{v}_k \in P_0$ and $\boldsymbol{w}_k \in E$. Because $\boldsymbol{X}\boldsymbol{\beta}_k \to \boldsymbol{g}$ and $\text{pen}(\boldsymbol{\beta}_k)\boldsymbol{w}_k \to 0$, one may deduce that $\boldsymbol{X}\boldsymbol{v}_k \to \boldsymbol{g}$. Moreover, since $\text{pen}(\boldsymbol{\beta}_k)\boldsymbol{v}_k \in P_0$, we have the following inequality:

$$f(\widehat{\boldsymbol{\beta}}) = \frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|_2^2 \leq \frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{v}_k\right\|_2^2 \overset{k\to+\infty}{\longrightarrow} \frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{g}\right\|_2^2 = m,$$

which achieves the proof. □

**Proof of Theorem 5.6.1**

*Proof of Theorem 5.6.1.* ($\Longrightarrow$) Assume that there exists a face $F$ of $B^* = \text{conv}\{\boldsymbol{0}, \boldsymbol{u}_1, \ldots, u_l\}$ that intersects $\text{row}(\boldsymbol{X})$ and satisfies $\dim(F) < \dim(\ker(\boldsymbol{X}))$. By Lemma 5.8.1, $F = \partial\text{pen}(\widehat{\boldsymbol{\beta}})$ for some $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$. Let $\boldsymbol{z} \in \mathbb{R}^n$ with $\boldsymbol{X}'\boldsymbol{z} \in F$, which exists by assumption. Now let $\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \lambda\boldsymbol{z}$. Note that $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$ since

$$\boldsymbol{0} \in \partial\left(\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\text{pen}(\widehat{\boldsymbol{\beta}})\right) = \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{y} + \lambda\partial\text{pen}(\widehat{\boldsymbol{\beta}}) \tag{5.8.2}$$

$$\Longleftrightarrow \frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{z} \in \partial\text{pen}(\widehat{\boldsymbol{\beta}}). \tag{5.8.3}$$

Now we construct $\widetilde{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$ with $\widetilde{\boldsymbol{\beta}} \neq \widehat{\boldsymbol{\beta}}$. According to Lemma 5.8.1, $\partial\text{pen}(\widehat{\boldsymbol{\beta}}) = \text{conv}\{\boldsymbol{u}_l : l \in I\}$ where $I = I_{\text{pen}}(\widehat{\boldsymbol{\beta}}) = \{l \in [k] : \boldsymbol{u}_l'\widehat{\boldsymbol{\beta}} = \text{pen}(\widehat{\boldsymbol{\beta}})\}$ and thus $\boldsymbol{u}_l'\widehat{\boldsymbol{\beta}} < \text{pen}(\widehat{\boldsymbol{\beta}})$ whenever $l \notin I$. Now we show that it is possible to pick $\boldsymbol{h} \in \ker(\boldsymbol{X})$ with $\boldsymbol{h} \neq \boldsymbol{0}$, but $\boldsymbol{u}_l'\boldsymbol{h} = 0$ for all $l \in I$. Then we can make $\boldsymbol{h}$ small enough such that $\boldsymbol{u}_l'(\widehat{\boldsymbol{\beta}}+\boldsymbol{h}) \leq \text{pen}(\widehat{\boldsymbol{\beta}})$ still holds for all $l \notin I$, which in turn implies that $\text{pen}(\widehat{\boldsymbol{\beta}}+\boldsymbol{h}) = \max\{\boldsymbol{u}_l'\widehat{\boldsymbol{\beta}} : l \in I\} = \text{pen}(\widehat{\boldsymbol{\beta}})$. This, together with $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\widehat{\boldsymbol{\beta}}+\boldsymbol{h})$, yields $\widehat{\boldsymbol{\beta}} \neq \widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \boldsymbol{h} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$. The existence of $\boldsymbol{0} \neq \boldsymbol{h} \in \ker(\boldsymbol{X})$ such that $\boldsymbol{u}_l'\boldsymbol{h} = 0$ for every $l \in I$ is equivalent to $\ker(\boldsymbol{X}) \cap (\text{col}(\boldsymbol{U}))^\perp \neq \{\boldsymbol{0}\}$, where $\boldsymbol{U} = (\boldsymbol{u}_l)_{l\in I} \in \mathbb{R}^{p\times|I|}$. To prove it, we distinguish two cases:
1) Assume that $\boldsymbol{0} \in \text{aff}\{\boldsymbol{u}_l : l \in I\}$. Then $\text{aff}\{\boldsymbol{u}_l : l \in I\} = \text{col}(\boldsymbol{U})$ and $\text{rk}(\boldsymbol{U}) = \dim(F) < \dim(\ker(\boldsymbol{X}))$. This implies that

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

which proves what was claimed.
2) Assume that $\mathbf{0} \notin \text{aff}\{\boldsymbol{u}_l : \ l \in I\}$. It implies that $\boldsymbol{v} = \boldsymbol{X}'\boldsymbol{z} \in \text{row}(\boldsymbol{X}) \cap \text{conv}\{\boldsymbol{u}_l : \ l \in I\}$ satisfies $\boldsymbol{X}'\boldsymbol{z} \neq 0$. We also have $\text{rk}(\boldsymbol{U}) = \dim(\text{aff}\{\boldsymbol{u}_l : \ l \in I\}) + 1 = \dim(F) + 1 \leq \dim(\ker(\boldsymbol{X}))$, which implies that

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

If $\ker(\boldsymbol{X}) \cap (\text{col}(\boldsymbol{U}))^{\perp} = \{\mathbf{0}\}$, then $\mathbb{R}^p = \ker(\boldsymbol{X}) \oplus (\text{col}(\boldsymbol{U}))^{\perp}$. But, since $\boldsymbol{v} \in \text{row}(\boldsymbol{X}) \cap \text{conv}\{\boldsymbol{u}_l : l \in I\}$, we also have $\ker(\boldsymbol{X}) \subseteq \boldsymbol{v}^{\perp}$ and $(\text{col}(\boldsymbol{U}))^{\perp} \subseteq \boldsymbol{v}^{\perp}$, yielding a contradiction and proving the claim.

($\Longleftarrow$) Assume that there exists $\boldsymbol{y} \in \mathbb{R}^n$ such that $\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$ with $\widehat{\boldsymbol{\beta}} \neq \widetilde{\boldsymbol{\beta}}$. Then

$$\frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \in \partial\text{pen}(\widehat{\boldsymbol{\beta}}) \quad \text{and} \quad \frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \in \partial\text{pen}(\widetilde{\boldsymbol{\beta}}).$$

According to Lemma 5.8.2, $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$, thus $\frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \in \text{row}(\boldsymbol{X})$. Consequently, one may deduce that $\text{row}(\boldsymbol{X})$ intersects the face $\partial\text{pen}(\widehat{\boldsymbol{\beta}}) \cap \partial\text{pen}(\widetilde{\boldsymbol{\beta}})$. Let $F^* = \text{conv}\{\boldsymbol{u}_l : \ l \in I^*\}$ be a face of $\partial\text{pen}(\widehat{\boldsymbol{\beta}}) \cap \partial\text{pen}(\widetilde{\boldsymbol{\beta}})$ of the smallest dimension among all faces of $\partial\text{pen}(\widehat{\boldsymbol{\beta}}) \cap \partial\text{pen}(\widetilde{\boldsymbol{\beta}})$ intersecting $\text{row}(\boldsymbol{X})$. By minimality of $\dim(F^*)$, $\text{row}(\boldsymbol{X})$ intersects the relative interior of $F^*$, namely, there exists $\boldsymbol{z} \in \mathbb{R}^n$ such that $\boldsymbol{v} = \boldsymbol{X}'\boldsymbol{z}$ lies in $F^*$, but not on a proper face of $F^*$. Now we will show that if $\dim(F^*) \geq \dim(\ker(\boldsymbol{X}))$, then $\text{row}(\boldsymbol{X})$ intersects a proper face of $F^*$, yielding a contradiction.
For this, we start with observing that $\dim(F^*) = \dim(\text{aff}\{\boldsymbol{u}_l : \ l \in I^*\})$ and that we can write the affine space $\text{aff}\{\boldsymbol{u}_l : \ l \in I^*\} = \boldsymbol{u}_{l_0} + \text{col}(\widetilde{\boldsymbol{U}}^*)$ where $l_0 \in I^*$ and $\widetilde{U}^* = (\boldsymbol{u}_l - \boldsymbol{u}_{l_0})_{l \in I^* \setminus \{l_0\}} \in \mathbb{R}^{p \times |I^*|-1}$. It implies that $\dim(F^*) = \text{rk}(\widetilde{\boldsymbol{U}}^*)$.
Now, let $\boldsymbol{h} = \widehat{\boldsymbol{\beta}} - \widetilde{\beta} \neq \mathbf{0}$. By Lemma 5.8.2, $\boldsymbol{h} \in \ker(\boldsymbol{X})$ and $\text{pen}(\widehat{\boldsymbol{\beta}}) = \text{pen}(\widetilde{\beta})$ by Lemma 5.8.2, and. Moreover, since $\boldsymbol{u}_l \in \partial\text{pen}(\widehat{\boldsymbol{\beta}}) \cap \partial\text{pen}(\widetilde{\beta})$ for all $l \in I^*$, then by Lemma 5.8.1 we get

$$\boldsymbol{u}_l'\boldsymbol{h} = \boldsymbol{u}_l'\widehat{\boldsymbol{\beta}} - \boldsymbol{u}_l'\widetilde{\boldsymbol{\beta}} = \text{pen}(\widehat{\boldsymbol{\beta}}) - \text{pen}(\widetilde{\boldsymbol{\beta}}) = 0 \quad \forall l \in I^*.$$

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

$$\dim(\text{row}(\boldsymbol{X})) + \dim(\text{col}(\widetilde{\boldsymbol{U}}^*)) \geq \text{rk}(\boldsymbol{X}) + \dim(\ker(\boldsymbol{X})) = p.$$

If $\text{row}(\boldsymbol{X}) \cap \text{col}(\widetilde{\boldsymbol{U}}^*) = \{\mathbf{0}\}$, then $\mathbb{R}^p = \text{row}(\boldsymbol{X}) \oplus \text{col}(\widetilde{\boldsymbol{U}}^*)$. However, the last equation cannot hold since $\text{row}(\boldsymbol{X}) = \ker(\boldsymbol{X})^{\perp} \subseteq \boldsymbol{h}^{\perp}$ and $\text{col}(\widetilde{\boldsymbol{U}}^*) \subseteq \text{col}(U^*) \subseteq \boldsymbol{h}^{\perp}$, where $\boldsymbol{h} \neq \mathbf{0}$. Consequently, there exists $\mathbf{0} \neq \widetilde{\boldsymbol{v}} \in \text{row}(\boldsymbol{X}) \cap \text{col}(\widetilde{\boldsymbol{U}}^*)$. The affine line $L = \{\boldsymbol{X}'\boldsymbol{z} + t\widetilde{\boldsymbol{v}} : \ t \in \mathbb{R}\} \subseteq \text{row}(\boldsymbol{X})$ intersects the relative interior of $F^*$ at $t = 0$. It is contained in $\text{aff}(F^*) = \boldsymbol{u}_{l_0} + \text{col}(\widetilde{\boldsymbol{U}}^*)$, since $\boldsymbol{X}'\boldsymbol{z} \in F^*$ and $\widetilde{\boldsymbol{v}} \in \text{col}(\widetilde{\boldsymbol{U}}^*)$. Therefore, $L$ must intersect a proper face of $F^*$ by Lemma 2.3.1 (ii). But then also $\text{row}(\boldsymbol{X})$ intersects a proper face of $F^*$, which yields the required contradiction. □

**Proof of Proposition 4.4.2**

The following lemma can be seen as a generalization of Proposition 4.1 from [88] from the $\ell_1$ norm to all convex functions.

**Lemma 5.8.4.** *Let $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\phi$ be a convex function on $\mathbb{R}^p$. Then* $\text{row}(\boldsymbol{X})$ *intersects $\partial\phi(\boldsymbol{\beta})$ if and only if, for any $\boldsymbol{b} \in \mathbb{R}^p$, the following implication holds*

$$\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{b} \implies \phi(\boldsymbol{\beta}) \leq \phi(\boldsymbol{b}). \tag{5.8.4}$$

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

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

Then (5.8.4) holds for any $\boldsymbol{b} \in \mathbb{R}^p$ if and only if $\boldsymbol{\beta}$ is a minimizer of the function $\phi(\boldsymbol{b}) + \iota_\beta(\boldsymbol{b})$. By definition of the subdifferential it may be shown that $\partial\iota_\beta(\boldsymbol{\beta}) = \text{row}(\boldsymbol{X})$. Therefore we can deduce that the implication (5.8.4) holds for any $\boldsymbol{b} \in \mathbb{R}^p$ if and only if

$$\boldsymbol{0} \in \text{row}(\boldsymbol{X}) + \partial\phi(\boldsymbol{\beta}) \iff \text{row}(\boldsymbol{X}) \cap \partial\phi(\boldsymbol{\beta}) \neq \varnothing.$$

□

*Proof of Proposition 4.4.2.* By Lemma 5.8.4, the geometric characterization of accessible patterns is equivalent to the analytic one. We show the geometric characterization.
( $\implies$ ) If the pattern of $\boldsymbol{\beta}$ is accessible with respect to $\boldsymbol{X}$ and $\lambda$pen, then there exists $\boldsymbol{y} \in \mathbb{R}^n$ and $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{y})$ such that $\widehat{\boldsymbol{\beta}} \overset{\text{pen}}{\sim} \boldsymbol{\beta}$. Because $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\text{pen}}(\boldsymbol{Y})$, we get $\frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \in \partial\text{pen}(\widehat{\boldsymbol{\beta}}) = \partial\text{pen}(\boldsymbol{\beta})$, thus row$(\boldsymbol{X})$ intersects $\partial\text{pen}(\boldsymbol{\beta})$.

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

**Proof of Theorem 5.4.4**

**Lemma 5.8.5.** *Let $\phi : \mathbb{R}^p \to \mathbb{R}$ be the polyhedral gauge defined as*

$$\phi(\boldsymbol{x}) = \max\{0, \boldsymbol{u}_1'\boldsymbol{x}, \ldots, \boldsymbol{u}_l'\boldsymbol{x}\} \text{ for some } \boldsymbol{u}_1, \ldots, \boldsymbol{u}_l \in \mathbb{R}^p.$$

*If $\partial\phi(\boldsymbol{x}) = \partial\phi(v)$, then $\partial\phi(\boldsymbol{x}) = \partial\phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v}) = \partial\phi(\boldsymbol{v})$ for all $\alpha \in [0, 1]$.*

*Proof.* Let $\boldsymbol{s} \in \partial\phi(\boldsymbol{x}) = \partial\phi(\boldsymbol{v})$. Since $\boldsymbol{s}$ is both a subgradient at $\boldsymbol{x}$ and at $\boldsymbol{v}$, the following inequalities hold

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

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

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

Using the convexity of $\phi$, we arrive at

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

By Lemma 5.8.1 we have $\partial\phi(\boldsymbol{x}) = \text{conv}\{\boldsymbol{u}_l : \ l \in I\}$, where $I_\phi(\boldsymbol{x}) = \{l \in [k] : \ \boldsymbol{u}_l'\boldsymbol{x} = \phi(\boldsymbol{x})\}$. Therefore, if $\boldsymbol{u}_l \in \partial\phi(\boldsymbol{x}) = \partial\phi(\boldsymbol{v})$, then $\boldsymbol{u}_l'\boldsymbol{x} = \phi(\boldsymbol{x})$ and $\boldsymbol{u}_l'\boldsymbol{v} = \phi(\boldsymbol{v})$, thus

$$\boldsymbol{u}_l'(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v}) = \alpha\phi(\boldsymbol{x}) + (1-\alpha)\phi(\boldsymbol{v}) = \phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v}).$$

Consequently, $\boldsymbol{u}_l \in \partial\phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v})$. On the other hand, if $\boldsymbol{u}_l \notin \partial\phi(\boldsymbol{x})$, then $\boldsymbol{u}_l'\boldsymbol{x} < \phi(\boldsymbol{x})$ and $\boldsymbol{u}_l'\boldsymbol{v} < \phi(\boldsymbol{v})$, thus

$$\boldsymbol{u}_l'(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v}) < \alpha\phi(\boldsymbol{x}) + (1-\alpha)\phi(\boldsymbol{v}) = \phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v}).$$

Consequently, $\boldsymbol{u}_l \notin \partial\phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{v})$ and the claim follows. □

**Lemma 5.8.6.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. Then the following set is convex*

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

Note that $V_\beta$ may be an empty set.

*Proof.* Assume that $V_\beta \neq \varnothing$. Let $\boldsymbol{y}$, $\widetilde{\boldsymbol{y}} \in V_\beta$. Then there exist $\lambda > 0$ and $\widetilde{\lambda} > 0$ such that $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y})$ and $\widetilde{\boldsymbol{\beta}} \in S_{\mathbf{X},\widetilde{\lambda}\mathrm{pen}}(\widetilde{\boldsymbol{y}})$ with $\partial\mathrm{pen}(\widehat{\boldsymbol{\beta}}) = \partial\mathrm{pen}(\widetilde{\boldsymbol{\beta}}) = \partial\mathrm{pen}(\boldsymbol{\beta})$. Consequently,

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \in \lambda\partial\mathrm{pen}(\boldsymbol{\beta}) \text{ and } \boldsymbol{X}'(\widetilde{\boldsymbol{y}} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \in \widetilde{\lambda}\partial\mathrm{pen}(\boldsymbol{\beta}).$$

Let $\alpha \in (0,1)$ and $\check{\boldsymbol{y}} = \alpha\boldsymbol{y} + (1-\alpha)\widetilde{\boldsymbol{y}}$. Define $\check{\lambda} = \alpha\lambda + (1-\alpha)\widetilde{\lambda}$ and $\check{\boldsymbol{\beta}} = \alpha\widehat{\boldsymbol{\beta}} + (1-\alpha)\widetilde{\beta}$. Now we show that $\check{\boldsymbol{y}} \in V_\beta$. Indeed, observe that

$$\boldsymbol{X}'\left(\check{\boldsymbol{y}} - \boldsymbol{X}\check{\boldsymbol{\beta}}\right) = \alpha\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})+(1-\alpha)\boldsymbol{X}'(\widetilde{\boldsymbol{y}}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}) \in \alpha\lambda\partial\mathrm{pen}(\boldsymbol{\beta})+(1-\alpha)\widetilde{\lambda}\partial\mathrm{pen}(\boldsymbol{\beta}) = \check{\lambda}\partial\mathrm{pen}(\boldsymbol{\beta}).$$

By Lemma 5.8.5, $\partial\mathrm{pen}(\check{\boldsymbol{\beta}}) = \partial\mathrm{pen}(\alpha\widehat{\boldsymbol{\beta}} + (1-\alpha)\widetilde{\boldsymbol{\beta}}) = \partial\mathrm{pen}(\boldsymbol{\beta})$, then also $\check{\boldsymbol{\beta}} \in S_{\mathbf{X},\check{\lambda}\mathrm{pen}}(\check{\boldsymbol{y}})$, which proves the claim. □

*Proof of Theorem 5.4.4.* Assume that the noiseless recovery condition does not hold for $\beta$. Then $\boldsymbol{X}\boldsymbol{\beta} \notin V_\beta$, where $V_\beta$ is defined as in Lemma 5.8.6. Consequently, by convexity of $V_\beta$, for any realization of $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ we have $\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \notin V_\beta$ or $\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\varepsilon} \notin V_\beta$. Therefore

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

Consequently,

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

□

### Proof of Proposition 5.5.1

*Proof of Proposition 5.5.1.* We only need to prove the implication ( $\Longleftarrow$ ), as the other implication is obvious. Assume that $\partial\mathrm{pen}(\widehat{\boldsymbol{\beta}}) \subseteq \partial\mathrm{pen}(\boldsymbol{\beta})$. Since $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y})$, we have $\frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \in \partial\mathrm{pen}(\widehat{\boldsymbol{\beta}}) \subseteq \partial\mathrm{pen}(\boldsymbol{\beta})$. Consequently, $\mathrm{row}(\boldsymbol{X})$ intersects $\partial\mathrm{pen}(\boldsymbol{\beta})$ which implies that the pattern of $\boldsymbol{\beta}$ is accessible with respect to $\boldsymbol{X}$ and pen by Proposition 4.4.2. Consequently, there exists $\boldsymbol{y} \in \mathbb{R}^n$ and there exists $\widehat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y})$ for which $\widehat{\boldsymbol{\beta}} \overset{\mathrm{pen}}{\sim} \boldsymbol{\beta}$. □

### Proof of Theorem 5.5.1

**Lemma 5.8.7.** *Let* pen *be a polyhedral gauge on $\mathbb{R}^p$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{v} \in \mathrm{col}(\boldsymbol{X})$. Let $K_1$, $K_2 \geq 0$ be large enough such that the set $C = \{\boldsymbol{b} \in \mathbb{R}^p : \ \mathrm{pen}(\boldsymbol{b}) \leq K_1, \|\boldsymbol{X}\boldsymbol{b} - v\|_2 \leq K_2\}$ is non-empty. If $\ker(\boldsymbol{X}) \cap \ker(\mathrm{pen}) = \{\boldsymbol{0}\}$, then $C$ is compact.*

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

**Lemma 5.8.8.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and* pen *be a polyhedral gauge on $\mathbb{R}^p$. Assume that the uniform uniqueness holds for* (5.1.1)*. Let $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ and set $\boldsymbol{y}^{(r)} = \boldsymbol{X}(r\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$.*
*If $\boldsymbol{\beta}$ is accessible with respect to $\boldsymbol{X}$ and* pen*, then*

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

*Proof.* Since $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)}) \in S_{\mathbf{X},\lambda\mathrm{pen}}(\boldsymbol{y}^{(r)})$, the following inequality holds

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

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

$$\frac{1}{2}\|\boldsymbol{\varepsilon} + \boldsymbol{X}(r\boldsymbol{\beta}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})\|_2^2 + \lambda\mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})) \le \frac{1}{2}\|\boldsymbol{\varepsilon}\|_2^2 + \lambda\mathrm{pen}(r\boldsymbol{\beta}) \tag{5.8.5}$$

and in particular

$$\begin{aligned} \lambda\mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})) &\le \frac{1}{2}\|\boldsymbol{\varepsilon}\|_2^2 + \lambda\mathrm{pen}(r\boldsymbol{\beta}) \\ \implies \mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r) &\le \frac{\|\boldsymbol{\varepsilon}\|_2^2}{2\lambda r} + \mathrm{pen}(\boldsymbol{\beta}) \\ \implies \limsup_{r\to\infty} \mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r) &\le \mathrm{pen}(\boldsymbol{\beta}). \end{aligned} \tag{5.8.6}$$

Consequently, the sequence $\left(\mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r)\right)_{r\in\mathbb{N}_+}$ is bounded. By the Cauchy-Schwarz inequality, the inequality (5.8.5) implies that

$$-\|\boldsymbol{\varepsilon}\|_2 \|\boldsymbol{X}(r\boldsymbol{\beta}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})\|_2 + \frac{1}{2}\|\boldsymbol{X}(r\boldsymbol{\beta}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})\|_2^2 \le \lambda\mathrm{pen}(r\boldsymbol{\beta}) - \lambda\mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})) \tag{5.8.7}$$

$$\implies -\frac{\|\boldsymbol{\varepsilon}\|_2}{r}\left\|\boldsymbol{X}\left(\frac{\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})}{r} - \boldsymbol{\beta}\right)\right\|_2 + \frac{1}{2}\left\|\boldsymbol{X}\left(\frac{\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})}{r} - \boldsymbol{\beta}\right)\right\|_2^2 \le \frac{\lambda\mathrm{pen}(\boldsymbol{\beta})}{r} - \frac{\lambda}{r}\mathrm{pen}\left(\frac{\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})}{r}\right). \tag{5.8.8}$$

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

$$\left(\left\|\boldsymbol{X}\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r - \boldsymbol{\beta}\right)\right\|_2\right)_{r\in\mathbb{N}}. \tag{5.8.9}$$

By (5.8.8) we get

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

Moreover, since the sequence $\left(\mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r)\right)_{r\in\mathbb{N}_+}$ is bounded, we obtain

$$\limsup_{r\to\infty} \frac{\lambda\mathrm{pen}(\boldsymbol{\beta}) - \lambda\mathrm{pen}(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r)}{r} = 0.$$

We conclude that $\alpha = 0$ and the sequence (5.8.9) converges to 0.

Due to uniform uniqueness, by Lemma 5.6.1 we have $\ker(\text{pen}) \cap \ker(\boldsymbol{X}) = \{\boldsymbol{0}\}$ and thus, by Lemma 5.8.7, the sequence $((\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r)_{r\in\mathbb{N}}$ is bounded. Therefore, to prove that $\lim_{r\to\infty} \widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r = \boldsymbol{\beta}$, it suffices to show that $\boldsymbol{\beta}$ is the unique accumulation point of this sequence. We extract a subsequence $(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{q(r)})/q(r))_{r\in\mathbb{N}}$ converging to $\boldsymbol{\gamma} \in \mathbb{R}^p$. By (5.8.6), one may deduce that $\text{pen}(\boldsymbol{\gamma}) \leq \text{pen}(\boldsymbol{\beta})$. Moreover, we have

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

Finally, $\boldsymbol{\gamma}$ satisfies

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

Because the pattern of $\boldsymbol{\beta}$ is accessible, by Lemma 5.8.4 we also have $\text{pen}(\boldsymbol{\beta}) \leq \text{pen}(\boldsymbol{\gamma})$. Then, Lemma 5.6.1 implies that $\boldsymbol{\gamma} = \boldsymbol{\beta}$. Therefore,

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

□

**Lemma 5.8.9.** *Let* pen *be a polyhedral gauge on* $\mathbb{R}^p$*. Then there exists* $\tau_0 > 0$ *depending on* $\boldsymbol{\beta}$ *such that*

$$\partial\text{pen}(\boldsymbol{b}) \subseteq \partial\text{pen}(\boldsymbol{\beta}) \text{ for all } \boldsymbol{b} \in \overline{B}_\infty(\boldsymbol{\beta}, \tau_0).$$

*Proof.* Let $I = \{l \in [k] : \ \boldsymbol{u}_l'\boldsymbol{\beta} = \text{pen}(\boldsymbol{\beta})\}$. By Lemma 5.8.1, $\partial\text{pen}(\boldsymbol{\beta}) = \text{conv}\{\boldsymbol{u}_l\}_{l\in I}$. Since

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

and by the continuity of pen, one may pick $\tau_0 > 0$ small enough such that

$$\boldsymbol{u}_l'\boldsymbol{b} < \text{pen}(\boldsymbol{b}) \ \forall l \notin I, \ \forall \boldsymbol{b} \in \overline{B}_\infty(\boldsymbol{\beta}, \tau_0).$$

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

$$\partial\text{pen}(\boldsymbol{b}) = \text{conv}\{\boldsymbol{u}_l\}_{l\in J} \subseteq \text{conv}\{\boldsymbol{u}_l\}_{l\in I} = \partial\text{pen}(\boldsymbol{\beta}).$$

□

*Proof of Theorem 5.5.1.* By Lemma 5.8.9, there exists $\tau_0 > 0$ such that for any $\boldsymbol{b} \in \overline{B}_\infty(\boldsymbol{\beta}, \tau_0)$ we have $\partial\text{pen}(\boldsymbol{b}) \subseteq \partial\text{pen}(\boldsymbol{\beta})$. By Lemma 5.8.8, $\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r$ converges to $\boldsymbol{\beta}$ when $r$ tends to $\infty$. Consequently, we have

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

Consequently, for $r \geq r_0$ we have

$$\begin{cases} \forall \boldsymbol{b} \in \overline{B}_\infty(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r, \tau_0/2), \ \partial\text{pen}(\boldsymbol{b}) \subseteq \partial\text{pen}(\boldsymbol{\beta}) \\ \exists \widetilde{\boldsymbol{b}} \in \overline{B}_\infty(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)})/r, \tau_0/2), \partial\text{pen}(\widetilde{\boldsymbol{b}}) = \partial\text{pen}(\boldsymbol{\beta}). \end{cases}$$

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

$$\begin{cases} \forall \boldsymbol{b} \in \overline{B}_\infty(\widehat{\beta}(\boldsymbol{y}^{(r)}), r\tau_0/2), \partial\text{pen}(\boldsymbol{b}) \subseteq \partial\text{pen}(\boldsymbol{\beta}) \\ \exists \widetilde{\boldsymbol{b}} \in \overline{B}_\infty(\widehat{\boldsymbol{\beta}}(\boldsymbol{y}^{(r)}), r\tau_0/2), \partial\text{pen}(\widetilde{\boldsymbol{b}}) = \partial\text{pen}(\boldsymbol{\beta}). \end{cases}$$

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

### Noiseless recovery condition for the supremum norm

Recall the noiseless pattern recovery for the supremum norm:

$$\exists \lambda > 0 \; \exists \hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\|.\|_\infty}(\boldsymbol{X}\boldsymbol{\beta}) \text{ such that } \hat{\boldsymbol{\beta}} \overset{\|.\|_\infty}{\sim} \boldsymbol{\beta}.$$

**Remark 5.8.1.** Let pen $\geq 0$ be minimized at $\mathbf{0}$. Then $\boldsymbol{\beta} = \mathbf{0}$ minimizes the function

$$f(\boldsymbol{b}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda \text{pen}(\boldsymbol{b}).$$

*Proof.* The proof is a direct consequence of $\mathbf{0}$ minimizing both $\|\boldsymbol{X}\boldsymbol{b}\|_2$ and pen$(\boldsymbol{b})$. □

**Corollary 5.8.2.** *The noiseless pattern recovery is satisfied by $\boldsymbol{\beta} = \mathbf{0}$ for every polyhedral gauge* pen*.*

**Proposition 5.8.1.** *Let $\mathbf{0} \neq \boldsymbol{\beta} \in \mathbb{R}^p$ and $I := \{i \in [p] : |\beta_i| < \|\boldsymbol{\beta}\|_\infty\}$. Let $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1 | \boldsymbol{X}_I)$, where*

$$\tilde{\boldsymbol{X}}_1 = \sum_{i \notin I} \text{sign}(\beta_i)\boldsymbol{X}_i.$$

*Then the noiseless pattern recovery occurs if and only if*

$$\tilde{\boldsymbol{X}}'(\tilde{\boldsymbol{X}}')^+ \boldsymbol{e}_1 = \boldsymbol{e}_1,$$

*where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)'$.*

Before giving the proof, we recall that the subdifferential of the $\ell_\infty$ norm at $\mathbf{0}$ is the unit ball of the $\ell_1$ norm and for $\boldsymbol{x} \neq \mathbf{0}$ this subdifferential is equal to

$$\begin{aligned} \partial\|.\|_\infty(x) &= \{\boldsymbol{s} \in \mathbb{R}^p : \|\boldsymbol{s}\|_1 \leq 1 \text{ and } \boldsymbol{s}'\boldsymbol{x} = \|\boldsymbol{x}\|_\infty\} \\ &= \left\{ \boldsymbol{s} \in \mathbb{R}^p : \|\boldsymbol{s}\|_1 = 1 \text{ and } \begin{cases} s_i x_i \geq 0 & \text{if } |x_i| = \|\boldsymbol{x}\|_\infty \\ s_i = 0 & \text{otherwise} \end{cases} \right\}. \end{aligned} \tag{5.8.10}$$

*Proof.* ($\Longrightarrow$) Let us assume that there exists $\lambda > 0$ and $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\|.\|_\infty}(\boldsymbol{X}\boldsymbol{\beta})$ such that $\hat{\boldsymbol{\beta}} \overset{\|.\|_\infty}{\sim} \boldsymbol{\beta}$. Then the following property holds

$$\frac{1}{\lambda}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \in \partial\|.\|_\infty(\boldsymbol{\beta}) = \partial\|.\|_\infty(\hat{\boldsymbol{\beta}}). \tag{5.8.11}$$

Let us set $\boldsymbol{c} = (\|\boldsymbol{\beta}\|_\infty, \boldsymbol{\beta}_I')'$ and $\hat{\boldsymbol{c}} = (\|\hat{\boldsymbol{\beta}}\|_\infty, \hat{\beta}_I')'$. By construction, $\boldsymbol{X}\boldsymbol{\beta} = \tilde{\boldsymbol{X}}\boldsymbol{c}$ and $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{c}}$. Consequently, according to (5.8.11), we have

$$\frac{1}{\lambda}\boldsymbol{X}'\tilde{\boldsymbol{X}}(\boldsymbol{c} - \hat{\boldsymbol{c}}) \in \partial\|.\|_\infty(\boldsymbol{\beta}). \tag{5.8.12}$$

By the representation (5.8.10) of $\partial\|\cdot\|_\infty(\boldsymbol{\beta})$, the relation above implies that

$$\begin{cases} \forall i \in I, \boldsymbol{X}_i'\tilde{\boldsymbol{X}}(\boldsymbol{c} - \hat{\boldsymbol{c}}) = 0, \\ \frac{1}{\lambda}\sum_{i \notin I} \text{sign}(\beta_i)\boldsymbol{X}_i'\tilde{\boldsymbol{X}}(\boldsymbol{c} - \hat{\boldsymbol{c}}) = 1, \end{cases}$$

therefore

$$\frac{1}{\lambda}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}(\boldsymbol{c} - \hat{\boldsymbol{c}}) = \boldsymbol{e}_1 \Rightarrow \tilde{\boldsymbol{X}}(\boldsymbol{c} - \hat{\boldsymbol{c}}) = \lambda(\tilde{\boldsymbol{X}}')^+\boldsymbol{e}_1.$$

Using the last implication and (5.8.12), one may deduce that

$$\boldsymbol{X}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \frac{1}{\lambda}\boldsymbol{X}'\tilde{\boldsymbol{X}}(\boldsymbol{c}-\hat{\boldsymbol{c}}) \in \partial\|\cdot\|_\infty(\boldsymbol{\beta}).$$

Finally, we prove that $\tilde{\boldsymbol{X}}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \boldsymbol{e}_1$:

$$\boldsymbol{X}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 \in \partial\|.\|_\infty(\boldsymbol{\beta}) \Rightarrow \begin{cases} \forall i \in I\ \boldsymbol{X}_i'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = 0 \\ \sum\limits_{i\notin I} \operatorname{sign}(\beta_i)\boldsymbol{X}_i'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \boldsymbol{X}_1'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = 1 \end{cases} \Rightarrow \tilde{\boldsymbol{X}}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \boldsymbol{e}_1.$$

($\Longleftarrow$) Again, let $\boldsymbol{c} = (\|\boldsymbol{\beta}\|_\infty, \boldsymbol{\beta}_I')'$. Denote $\hat{\boldsymbol{c}} = \boldsymbol{c} - \lambda\tilde{\boldsymbol{X}}^{+}(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1$ and set $\hat{\boldsymbol{\beta}}$ as follows:

$$\forall i \in [p], \hat{\beta}_i := \begin{cases} \hat{c}_i \text{ when } i \in I, \\ \operatorname{sign}(\beta_i)\hat{c}_1 \text{ when } i \notin I. \end{cases}$$

Note that $\|\hat{\boldsymbol{\beta}}\|_\infty = c_1 = \|\boldsymbol{\beta}\|_\infty$. Moreover, by definition of $I$, $c_1$ is the unique coordinate of $\boldsymbol{c}$ having the maximal absolute value. Therefore, for $\lambda$ small enough, we have $\hat{c}_1 = \|\hat{\boldsymbol{c}}\|_\infty$. It implies that $\|\hat{\boldsymbol{\beta}}\|_\infty = \hat{c}_1$ and this value is attained exactly at coordinates $\hat{\beta}_i :\ i \notin I$. Therefore $\{i \in [p] : |\hat{\beta}_i| < \|\hat{\boldsymbol{\beta}}\|_\infty\} = \{i \in [p] : |\beta_i| < \|\boldsymbol{\beta}\|_\infty\} = I$. Moreover, for $i \notin I$ we have $\beta_i\hat{\beta}_i = \beta_i\operatorname{sign}(\beta_i)\hat{c}_1 = \|\boldsymbol{\beta}\|_\infty\hat{c}_1 > 0$. Consequently, for $\lambda$ small enough we have $\operatorname{sign}^\infty(\hat{\boldsymbol{\beta}}) = \operatorname{sign}^\infty(\boldsymbol{\beta})$. To conclude the proof it is enough to show that $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\|.\|_\infty}(\boldsymbol{X}\boldsymbol{\beta})$, i.e. $\frac{1}{\lambda}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \in \partial\|.\|_\infty(\hat{\boldsymbol{\beta}})$. By Remark 2.3.9 $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{+}$ is the projection onto $\operatorname{col}(\tilde{\boldsymbol{X}})$ and $\operatorname{col}((\tilde{\boldsymbol{X}}')^{+}) = \operatorname{col}(\tilde{\boldsymbol{X}})$. Therefore

$$\frac{1}{\lambda}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) = \frac{1}{\lambda}\boldsymbol{X}'(\tilde{\boldsymbol{X}}\boldsymbol{c} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{c}}) = \boldsymbol{X}'\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{+}(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \boldsymbol{X}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1.$$

Then it suffices to prove that the latter term belongs to $\partial\|.\|_\infty(\boldsymbol{\beta})$. As we assumed that $\boldsymbol{X}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \boldsymbol{e}_1$, we have

$$\begin{aligned} \left(\tilde{\boldsymbol{X}}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = \boldsymbol{e}_1\right) &\Rightarrow \begin{cases} \boldsymbol{X}_1'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = 1 \\ \forall i \in I\ \boldsymbol{X}_i'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = 0 \end{cases} \\ &\Rightarrow \begin{cases} \sum\limits_{i\notin I} \operatorname{sign}(\beta_i)\boldsymbol{X}_i'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = 1 \\ \forall i \in I\ \boldsymbol{X}_i'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 = 0 \end{cases} \Rightarrow \left(\boldsymbol{X}'(\tilde{\boldsymbol{X}}')^{+}\boldsymbol{e}_1 \in \partial\|.\|_\infty(\boldsymbol{\beta})\right). \end{aligned}$$

Consequently, for $\lambda > 0$ small enough, $\hat{\boldsymbol{\beta}} \in S_{\mathbf{X},\lambda\|.\|_\infty}(\boldsymbol{X}\boldsymbol{\beta})$ and $\operatorname{sign}^\infty(\hat{\boldsymbol{\beta}}) = \operatorname{sign}^\infty(\boldsymbol{\beta})$.

□

# Chapter 6

# Maximum likelihood estimation for discrete exponential families and random graphs

## 6.1 Introduction and preliminaries

The content of this chapter comes from the published paper of Bogdan, Bosy and the author of the dissertation [22].

Exponential families are of paramount importance in probability and statistics. They were introduced by Fisher, Pitman, Darmois and Koopman in 1934-36 and have many properties that make them indispensable in theory and applications, see [125] (Section 2.7), [11] (Chapter 9), [5], [57] (Chapter 9.E), [58], and [121]. In this paper we study *discrete* exponential families, more specifically, exponential families on *finite* sets, and give a new characterization of the existence of the maximum likelihood estimator (MLE) for exponential family and the data at hand. We also present applications, in particular for specific exponential families we give a threshold of the sample size sufficient for the existence of MLE with high probability for *i.i.d.* samples.

The computation of MLE is in general difficult with the number of variables increasing. On the other hand, for given data and an exponential family, MLE may fail to exist. In particular, [50, 51] pointed out to problems with the maximum likelihood estimation when the number of parameters is too large for the sample size. He also gave a sufficient condition for MLE to exist almost surely – the Haar condition.

A complete characterization of the existence of MLE for rather general exponential families was given by Barndorff-Nielsen. Namely, by [11] (Theorem 9.13), MLE for a sample and an exponential family exists if and only if the vector of the sample means calculated for a basis of the linear space of exponents belongs to the interior of the convex hull of the pointwise range of the basis.

This beautiful criterion is alas cumbersome to apply. Therefore, [105] gives an alternative condition for discrete exponential families, together with applications to Cox regression, logistic regression and multiplicative Poisson models. Similar condition is presented by [3] for log-linear model. [96] gives a characterization of the existence of MLE for hierarchical log-linear models. His conditions can be interpreted in terms of polytope geometry, see also [69], and [75]. [36] characterizes the existence of MLE when the log-partition function is steep and regularly convex, and interprets the problem of finding MLE as the optimization of the Kullback-Leibler

divergence. [53] connect the properties of MLE in decomposable models with graph-theoretical notions, thus starting the theory of graphical models in statistics. Sufficient conditions for the existence of MLE in specific exponential families are also given by [169] and [25]. [86] looks for MLE in the closure of convex exponential families, relates the existence of MLE with the linear programming feasibility problem, and in the case of nonexistent MLE, reduces the considered exponential family until MLE exists for the family. He also applies MCMC algorithms to calculate MLE. A comparison between the conditions of Barndorff-Nielsen and Jacobsen is discussed in [114]. In addition, Konis presents an implementation of Jacobsen's test using linear programming. A broad survey of the history of log-linear models and further motivation for the study of the existence of MLE can be found in [74, 75].

The main inspiration for our work is [21] (Theorem 2.3) on the existence of MLE for exponential families of continuous functions on finite interval. In Theorem 6.2.2 below we propose a similar characterization, which is new in the setting of discrete exponential families. We obtain the result by a straightforward, self-contained approach, which does not depend on the delicate convex analysis of [11].

The paper is composed as follows. In Section 6.2 we state and prove our criterion, using the notion of *set of uniqueness*. The criterion is restated in Section 6.2.2 as a linear programming problem. In Section 6.3 we give applications to exponential families spanned by Rademacher and Walsh functions, and to exponential families of random graphs. In particular we give sharp or plain thresholds for the sample size to secure the existence of MLE with high probability. In Appendix 6.6 we give auxiliary results and reformulations of our criterion and pin down its connections with the criterion of Barndorff-Nielsen.

**Acknowledgments:** We are grateful to Małgorzata Bogdan, Piotr Ciołek, Persi Diaconis, Hélène Massam, Sumit Mukherjee, Krzysztof Oleszkiewicz, Krzysztof Samotij, Maciej Wilczyński and anonymous referees for comments, corrections, references and discussion.

### 6.1.1 Discrete exponential family

Consider a finite set $\mathcal{X} \neq \varnothing$ and weight function $\mu : \mathcal{X} \to (0, \infty)$. As usual, $\mathbb{R}^{\mathcal{X}}$ is the family of all the real-valued functions on $\mathcal{X}$. For $\phi \in \mathbb{R}^{\mathcal{X}}$ we define the *partition* and the *log-partition* functions,

$$Z(\phi) = \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x), \quad \psi(\phi) = \log Z(\phi), \tag{6.1.1}$$

respectively, and the *exponential density*

$$p = e(\phi) = e^{\phi - \psi(\phi)} = e^{\phi} / Z(\phi). \tag{6.1.2}$$

Clearly, $p > 0$ and $\sum_{x \in \mathcal{X}} p(x)\mu(x) = 1$. For arbitrary real number $c$ we have $\psi(\phi + c) = \psi(\phi) + c$, hence

$$e(\phi + c) = e(\phi). \tag{6.1.3}$$

Moreover, for $\phi_1, \phi_2 \in \mathbb{R}^{\mathcal{X}}$ we have $e(\phi_1) = e(\phi_2)$ if and only if $\phi_1 - \phi_2$ is constant. Consider $x_1, \ldots, x_n \in \mathcal{X}$, a *sample*. For $\phi \in \mathbb{R}^{\mathcal{X}}$ we denote, as usual,

$$\bar{\phi} = \frac{1}{n} \sum_{i=1}^{n} \phi\left(x_i\right).$$

The *likelihood function* of $p = e(\phi)$ is defined as

$$L_{e(\phi)}\left(x_1, \ldots, x_n\right) = L_p\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} p(x_i),$$

and the *log-likelihood function* is

$$l_{e(\phi)}(x_1,\dots,x_n) := \log L_{e(\phi)}(x_1,\dots,x_n) = n\left(\bar{\phi} - \psi(\phi)\right). \tag{6.1.4}$$

Of course, for every $c \in \mathbb{R}$ we have

$$l_{e(\phi+c)}(x_1,\dots,x_n) = l_{e(\phi)}(x_1,\dots,x_n). \tag{6.1.5}$$

We note that the likelihood functions are uniformly bounded. Indeed, for every $\phi \in \mathbb{R}^{\mathcal{X}}$,

$$\psi(\phi) = \log \sum_{x\in\mathcal{X}} e^{\phi(x)}\mu(x) \geq \max_{\mathcal{X}} \phi + \min_{\mathcal{X}} \log \mu, \tag{6.1.6}$$

and so by (6.1.4) and (6.1.6),

$$l_{e(\phi)}(x_1,\dots,x_n) \leq -n \min_{\mathcal{X}} \log \mu \quad \text{and} \quad L_{e(\phi)}(x_1,\dots,x_n) \leq (\min_{\mathcal{X}} \mu)^{-n}. \tag{6.1.7}$$

We fix a linear subspace $\mathcal{B} \subset \mathbb{R}^{\mathcal{X}}$. The *exponential family* spanned by $\mathcal{B}$ is

$$e(\mathcal{B}) := \{p = e(\phi) :\ \phi \in \mathcal{B}\}. \tag{6.1.8}$$

Since $\mathcal{X}$ is a finite set, $e(\mathcal{B})$ will be called *discrete exponential family* (we do not consider infinite countable sets, for which see [105]).
We call $\hat{p} \in e(\mathcal{B})$ an MLE for $x_1,\dots,x_n$ and $e(\mathcal{B})$ if

$$L_{\hat{p}}(x_1,\dots,x_n) = \sup_{p\in e(\mathcal{B})} L_p(x_1,\dots,x_n),$$

or, equivalently,

$$l_{\hat{p}}(x_1,\dots,x_n) = \sup_{p\in e(\mathcal{B})} l_p(x_1,\dots,x_n).$$

The following result is well known (see, e.g., [110] (Theorem 2.1) or [57] (p. 177)), but for the reader's convenience we give a proof in Appendix 6.6.1.

**Lemma 6.1.1.** *If MLE exists, then it is unique.*

Despite the boundedness (6.1.7), MLE may fail to exist, as shown by the following example.

**Example 6.1.1.** Let $\mathcal{X} = \{0,1\}$, $\mu \equiv 1$, $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$, $n = 1$ and $x_1 = 1$. Let $a, b \in \mathbb{R}$ and $\phi = a + b\mathbb{1}_{\{1\}}$. Then $Z(\phi) = e^a(1+e^b)$, $e(\phi) = e^{b\mathbb{1}_{\{1\}}}/(1+e^b)$, and $L_{e(\phi)}(x_1) = e(\phi)(1) = e^b/(1+e^b)$. Thus, $\sup L_{e(\phi)}(x_1) = 1$, but the supremum is not attained for any $a, b \in \mathbb{R}$, so MLE does not exist in this case. On the other hand, if $n = 3$, $x_1 = x_2 = 0$, and $x_3 = 1$, then $L_{e(\phi)}(x_1,x_2,x_3) = e^b/(1+e^b)^3$. By calculus, the maximum is attained when $e^b = 1/2$, therefore $\hat{p} = (2 - \mathbb{1}_{\{1\}})/3$ is the MLE in this case.

We note that the first supremum in Example 6.1.1 is approached when $b \to \infty$, or for the density $p = \mathbb{1}_{\{1\}}$, which, however, is not in $e(\mathbb{R}^{\mathcal{X}})$ but rather in $e(\mathbb{R}^{\{1\}})$. Below in Theorem 6.2.2 we characterize the situation when the genuine MLE exists, and in Theorem 6.2.5 we treat, by a suitable reduction of $\mathcal{X}$, the case when the supremum of the likelihood function is ''attained at infinity". Before we proceed, we owe the reader some comments on the notation used in this paper and in the literature.

### 6.1.2 Alternative setting

Let $d$ be a natural number. Consider a nonempty finite set $S \subset \mathbb{R}^d$, weight $m$ on $S$ and the linear space spanned by the coordinate functions on $\mathbb{R}^d$. The corresponding exponential densities have the form

$$\pi_\theta(y) = e^{\theta \cdot y}/\zeta(\theta), \quad y \in S, \tag{6.1.9}$$

where $\theta \in \mathbb{R}^d$, $\cdot$ is the scalar product in $\mathbb{R}^d$ and $\zeta(\theta) = \sum_{y \in \mathcal{S}} e^{\theta \cdot y} m(y)$. Thus, (6.1.9) is a *natural*, or *standard*, exponential family, see [126] or [36]. Since the range of the vector of parameters $\theta$ is the whole of $\mathbb{R}^d$, which is open, the exponential family (6.1.9) is *regular*, see [122] (Appendix D.1). The setting is actually generic, as we explain momentarily. If functions $\phi_1, \ldots, \phi_d$ span the linear space $\mathcal{B}$ in the general discussion above and we let $T(x) = (\phi_1(x), \ldots, \phi_d(x))$ for $x \in \mathcal{X}$, then for every $\phi \in \mathcal{B}$ there is $\theta \in \mathbb{R}^d$ such that $\phi(x) = \theta \cdot T(x)$ for $x \in \mathcal{X}$, and

$$e(\phi) = e^{\theta \cdot T}/Z(\theta \cdot T). \tag{6.1.10}$$

This is the form used by most authors, see [122] or [110], and $T$ is called the *canonical* statistics. Furthermore, we let $S = T(\mathcal{X}) \subset \mathbb{R}^d$ and $m(y) = \sum_{x:T(x)=y} \mu(x)$ for $y \in S$. With the notation of (6.1.9) and (6.1.10) we have

$$\pi_\theta(y) = e(\phi)(x) \quad \text{if} \quad T(x) = y. \tag{6.1.11}$$

If $x_1, \ldots, x_n \in \mathcal{X}$ is the sample and we denote $y_1 = T(x_1), \ldots, y_n = T(x_n)$, then the corresponding likelihoods are equal, too. Therefore $\pi_{\hat\theta}$ is the maximum likelihood estimator for $y_1, \ldots, y_n$ and $\{\pi_\theta : \ \theta \in \mathbb{R}^d\}$ if and only if $e(\hat\theta \cdot T)$ is the maximum likelihood estimator for $x_1, \ldots, x_n$ and $\{e(\phi) : \ \phi \in \mathcal{B}\}$. This makes a complete connection between our setting and the setting of natural exponential families with finite support $S$. The same setting of discrete exponential families on finite set is described, using slightly different language, in [172] (§6.2). We also recall that if $\phi_1, \ldots, \phi_d$ are affinely independent, then the representation (6.1.10) is *minimal*, see [110] (Chapter 1) or [122], where the affine independence means that $\theta \cdot T = const$ implies $\theta = 0$. In general, one allows the representation to be nonminimal because over-parametrization is often natural in applications. We shall return to this discussion again in Section 6.6.6, but for now we get back to the setting of $\mathcal{B}$ and (6.1.8). The latter allows to work without coordinates and benefit from properties of specific linear spaces $\mathcal{B}$, which could otherwise be obscured by an arbitrary choice of $T$ and $S$.

## 6.2 Main results

Let $\mathbb{1}$ denote the function on $\mathcal{X}$ identically equal to 1. Assume that $\mathbb{1} \in \mathcal{B}$. This entails no restriction on the considered exponential families $e(\mathcal{B})$, but allows an elegant formulation of the criterion of existence of MLE in terms of $\mathcal{B}$, in fact in terms of the cone of nonnegative functions in $\mathcal{B}$:

$$\mathcal{B}_+ := \{\phi \in \mathcal{B} : \ \phi \geq 0\}.$$

We note in passing that Appendix 6.6.6 gives a reformulation of our criterion for the existence of MLE without requiring that $\mathbb{1} \in \mathcal{B}$.

Let $U \subset \mathcal{X}$. We say that $U$ is a *set of uniqueness* for $\mathcal{B}$ if $\phi = 0$ is the only function in $\mathcal{B}$ such that $\phi = 0$ on $U$. Similarly, we say that $U$ is a *set of uniqueness* for $\mathcal{B}_+$ if $\phi = 0$ is the only function in $\mathcal{B}_+$ such that $\phi = 0$ on $U$. Put differently, $U$ is of uniqueness for $\mathcal{B}_+$ if the conditions $\phi \in \mathcal{B}_+$ and $\phi = 0$ on $U$ imply that $\phi = 0$ on $\mathcal{X}$. Of course, if $U$ is a set of uniqueness for $\mathcal{B}$, then $U$ is a set of uniqueness for $\mathcal{B}_+$.

**Example 6.2.1.** Let $\mathcal{X} = \{-2, -1, 0, 1, 2\} \subset \mathbb{R}$. Let $\mathcal{B}$ denote the class of all real functions on $\mathcal{X}$ that are of the form $a + bx$ on $\{-2, -1, 0\}$ and $a + cx$ on $\{0, 1, 2\}$ with some $a, b, c \in \mathbb{R}$. Then $\{-1, 2\}$ is a set of uniqueness for $\mathcal{B}_+$ but $\{-2, 2\}$ is not. We also observe that $\{-1, 2\}$ is not a set of uniqueness for $\mathcal{B}$, so the nonnegativity of functions in $\mathcal{B}_+$ plays a role here.

Being a set of uniqueness is a monotone property in the sense that every set larger than a set of uniqueness is also of uniqueness. Furthermore, if $U$ is a set of uniqueness for $\mathcal{B}_+$ and $\mathcal{A}$ is a linear subspace of $\mathcal{B}$, then $U$ is of uniqueness for $\mathcal{A}_+$.
The following is a crucial definition: For $U \subset \mathcal{X}$ and $\phi \in \mathcal{B}$ we let

$$\lambda_U(\phi) = \max_{\mathcal{X}} \phi - \min_{U} \phi.$$

Here is our characterization of the existence of MLE for discrete exponential families.

**Theorem 6.2.2.** *MLE for $e(\mathcal{B})$ and $x_1, \ldots, x_n \in \mathcal{X}$ exists if and only if $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}_+$.*

*Proof.* Let us start with the "only if" part. If $U = \{x_1, \ldots, x_n\}$ is not a set of uniqueness for $\mathcal{B}_+$, then there is a nonzero function $f \in \mathcal{B}_+$ such that $f(x_1) = \ldots = f(x_n) = 0$. Let $\phi \in \mathcal{B}$ be arbitrary. Let $\varphi = \phi - f$. We have $\bar{\varphi} = \bar{\phi}$, but $\psi(\varphi) < \psi(\phi)$, where $\psi$ is defined in (6.1.1). So, by (6.1.4), $l_{e(\phi)}(x_1, \ldots, x_n) < l_{e(\varphi)}(x_1, \ldots, x_n)$. Therefore no $\phi \in \mathcal{B}$ is MLE for $x_1, \ldots, x_n$. To prove the other implication, we let $U$ be a set of uniqueness for $\mathcal{B}_+$. By (6.1.4) for $\varphi \in \mathcal{B}$,

$$l_{e(\varphi)}(x_1, \ldots, x_n) = n\left(\overline{\varphi} - \psi(\varphi)\right) \leq n\left(\frac{1}{n}\left(\min_{U} \varphi + (n-1)\max_{\mathcal{X}} \varphi\right) - \psi(\varphi)\right).$$

Let $C = \min_{x \in \mathcal{X}} \log \mu(x)$. By (6.1.6), (6.1.5),

$$\begin{aligned} l_{e(\varphi)}(x_1, \ldots, x_n) &\leq \min_{U} \varphi + (n-1)\max_{\mathcal{X}} \varphi - n\max_{\mathcal{X}} \varphi - nC \\ &= -\lambda_U(\varphi) - nC \to -\infty, \end{aligned}$$

as $\lambda_U(\varphi) \to \infty$. By Lemma 6.6.1, $\lambda_U(\varphi) \to \infty$ if $\lambda_{\mathcal{X}}(\varphi) \to \infty$. In particular, there exists $M > 0$ such that if $\lambda_{\mathcal{X}}(\varphi) > M$, then

$$l_{e(\varphi)}(x_1, \ldots, x_n) < l_{e(0)}(x_1, \ldots, x_n) = -n \log \mu(\mathcal{X}).$$

By (6.1.5) and continuity, the maximum of $l_{e(\varphi)}(x_1, \ldots, x_n)$ is attained on the compact set $\{\varphi \in \mathcal{B} : \ 0 \leq \varphi \leq M\}$. □

The above proof is different from that of [21] (Theorem 2.3), [11] (Theorem 9.13) and [172] (Theorem 8.2.1); the use of $\lambda_U$ makes our arguments more direct.

**Remark 6.2.3.** By Theorem 6.2.2 we see that the existence of MLE depends on the sequence $(x_1, \ldots, x_n)$ only through the set $\{x_1, \ldots, x_n\}$. Furthermore, the existence of MLE does not depend on $\mu$, i.e., we may take constant $\mu$ without loosing generality. Summarizing, the existence of MLE depends only on $\mathcal{B}$ and the set $\{x_1, \ldots, x_n\}$. Of course, the actual MLE, say $\widehat{p}$, does depend on the sequence $(x_1, \ldots, x_n)$, the weight $\mu$ and $\mathcal{B}$.

### 6.2.1 Nonexistence of MLE

In this section we elaborate on the case of nonexistence of MLE in the spirit of [86]. To this end we fix $x_1, \ldots, x_n \in \mathcal{X}$ and assume that there is a nontrivial $\delta \in \mathcal{B}_+$ such that $\delta(x_1) = \ldots = \delta(x_n) = 0$. By Theorem 6.2.2, $\sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n)$ is not attained at any $p \in e(\mathcal{B})$. However, the supremum is "attained at infinity", in fact for an exponential density on a proper subset of the state space $\mathcal{X}$. Indeed, fix $\delta$ as above. If $\varphi \in \mathcal{B}$ and $k \in (0, \infty)$, then

$$l_{e(\varphi)}(x_1, \ldots, x_n) \leq l_{e(\varphi - k\delta)}(x_1, \ldots, x_n),$$

see the first part of the proof of Theorem 6.2.2. Furthermore,

$$\psi(\varphi - k\delta) \to \log \sum_{x \in \mathcal{X} : \delta(x) = 0} e^{\varphi(x)} \mu(x), \quad \text{as } k \to \infty. \tag{6.2.1}$$

We let $\widetilde{\mathcal{X}} = \{x \in \mathcal{X} : \ \delta(x) = 0\}$ and carrying on with the notation for $\widetilde{\mathcal{X}}$ we obtain measure $\widetilde{\mu}$, linear space $\widetilde{\mathcal{B}}$ with cone $\widetilde{\mathcal{B}}_+$, log-partition function $\widetilde{\psi}$, likelihood function $\widetilde{L}$, log-likelihood function $\widetilde{l}$ and exponential family $e(\widetilde{\mathcal{B}})$. Put simpler, we discard $\{x \in \mathcal{X} : \ \delta(x) > 0\}$ and achieve the following reduction.

**Lemma 6.2.1.** $\sup_{\widetilde{p} \in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\widetilde{p}}(x_1, \ldots, x_n) = \sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n)$.

*Proof.* For $\phi \in \mathcal{B}$ we let $\widetilde{\phi} = \phi|_{\widetilde{\mathcal{X}}}$. Since $\{x_1, \ldots, x_n\} \subset \widetilde{\mathcal{X}}$,

$$\overline{\widetilde{\phi}} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\phi}(x_i) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) = \overline{\phi}. \tag{6.2.2}$$

Furthermore,

$$\psi(\phi) = \log \left( \sum_{x \in \mathcal{X}} e^{\phi(x)} \mu(x) \right) \geq \log \left( \sum_{x \in \widetilde{\mathcal{X}}} e^{\phi(x)} \mu(x) \right) = \widetilde{\psi}(\widetilde{\phi}).$$

Thus $\overline{\phi} - \psi(\phi) \leq \overline{\widetilde{\phi}} - \widetilde{\psi}(\widetilde{\phi})$, and so

$$\sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n) \leq \sup_{\widetilde{p} \in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\widetilde{p}}(x_1, \ldots, x_n).$$

Let $\delta \in \mathcal{B}_+$ and $k$ be as in (6.2.1). Using (6.2.1) and (6.2.2),

$$l_{e(\phi - k\delta)}(x_1, \ldots, x_n) \to \widetilde{l}_{e(\widetilde{\phi})}(x_1, \ldots, x_n), \quad \text{as } k \to \infty.$$

Therefore,

$$\sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n) \geq \sup_{\widetilde{p} \in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\widetilde{p}}(x_1, \ldots, x_n).$$

□

Motivated by Lemma 6.2.1, we define

$$\{x_1, \ldots, x_n\}_{\mathcal{B}_+} = \bigcap \phi^{-1}(\{0\}), \tag{6.2.3}$$

where the intersection is taken over all $\phi \in \mathcal{B}_+$ such that $\phi(x_1) = \ldots = \phi(x_n) = 0$. Thus for all $\phi \in \mathcal{B}_+$, if $\phi$ vanishes on $\{x_1, \ldots, x_n\}$, then it vanishes on $\{x_1, \ldots, x_n\}_{\mathcal{B}_+}$, and the latter is the largest such set. Put differently, if there is $\delta \in \mathcal{B}_+$ such that $\delta(x_1) = \ldots = \delta(x_n) = 0$ but $\delta(x) > 0$, then $x \notin \{x_1, \ldots, x_n\}_{\mathcal{B}_+}$, and conversely. In particular, $U \subset \mathcal{X}$ is set of uniqueness for $\mathcal{B}_+$ if and only if $U_{\mathcal{B}_+} = \mathcal{X}$.

**Example 6.2.4.** In the setting of Example 6.2.1 we have $\{-2\}_{\mathcal{B}_+} = \{-2\}$ and $\{-1\}_{\mathcal{B}_+} = \{-2, -1, 0\}$.

We note that if $x \notin \{x_1, \ldots, x_n\}_{\mathcal{B}_+}$, then there is $\phi \in \mathcal{B}_+$ such that $\phi = 0$ on $\{x_1, \ldots, x_n\}$ but $\phi(x) > 0$. Since $\mathcal{X}$ is finite, by adding such functions we can construct $\delta \in \mathcal{B}_+$ that vanishes precisely on $\{x_1, \ldots, x_n\}_{\mathcal{B}_+}$, i.e., $\delta^{-1}(\{0\}) = \{x_1, \ldots, x_n\}_{\mathcal{B}_+}$. We adopt the setting of Lemma 6.2.1 with this $\delta$, in particular with $\widetilde{\mathcal{X}} = \{x_1, \ldots, x_n\}_{\mathcal{B}_+}$, and we get the following result.

**Theorem 6.2.5.** *There is a unique $\widetilde{p} \in e(\widetilde{\mathcal{B}})$ such that $\widetilde{l}_{\widetilde{p}}(x_1, \ldots, x_n) = \sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n)$.*

*Proof.* By the definition of $\{x_1, \ldots, x_n\}_{\mathcal{B}_+}$ and by Theorem 6.2.2, Lemma 6.1.1 and 6.2.1, there is a unique $\widetilde{p} \in e(\widetilde{\mathcal{B}})$ such that

$$\widetilde{l}_{\widetilde{p}}(x_1, \ldots, x_n) = \sup_{\hat{p} \in e(\widetilde{\mathcal{B}})} \widetilde{l}_{\hat{p}}(x_1, \ldots, x_n) = \sup_{p \in e(\mathcal{B})} l_p(x_1, \ldots, x_n).$$

□

**Example 6.2.6.** For the first sample in Example 6.1.1 we get $\widetilde{\mathcal{X}} = \{x_1\}_{\mathcal{B}_+} = \{1\}$, and $\widetilde{p} = 1$ on $\widetilde{\mathcal{X}}$.

For more substantial applications of Theorem 6.2.5 we refer to Example 6.3.1 and Example 6.3.5.

### 6.2.2 Linear programming

Before we address special spaces $\mathcal{B}$, we offer the reader a down-to-earth perspective. To start with, by a comment at the beginning of Section 6.2, we get the following simple result.

**Corollary 6.2.1.** *If $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}$ then MLE exists for $e(\mathcal{B})$ and $x_1, \ldots, x_n$.*

Notably, the condition in Corollary 6.2.1 may be verified by solving the following linear problem:

$$\begin{aligned} \phi &\in \mathcal{B}, \\ \phi(x_1) = ... &= \phi(x_n) = 0. \end{aligned}$$

Indeed, $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}$ if and only if the homogeneous linear system has only the trivial solution. In contrast, Theorem 6.2.2 is a linear programming problem. Indeed, $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}_+$ if and only if the supremum of the (objective) function $\sum_{x \in \mathcal{X}} \phi(x)$ is zero for the class of functions satisfying

$$\begin{aligned} \phi &\in \mathcal{B}, \\ \phi(x_1) = ... &= \phi(x_n) = 0, \\ \phi &\geq 0. \end{aligned}$$

In this vein [150] (Appendix C) observe that the condition of Barndorff-Nielsen is actually a linear programming problem and make connections to the geometry (of the convex hull of the set $S$ in Section 6.1.2). The linear programming also occurs in the study of the closures of convex exponential families [86] or binary logistic regression models [114]. Furthermore, [185] consider the linear programming in the case when MLE fails to exist. See also [172] for further information on linear programming and cases of nonexistence of MLE for discrete exponential families. Since the linear programming in general runs in polynomial time, see [157], it should be the method of choice when verifying the existence of MLE for discrete exponential families

and data at hand. Having said this, for special linear spaces $\mathcal{B}$ one can come across interesting mathematics, as we demonstrate below. We also remark in passing that the linear problem in Corollary 6.2.1 is the Haar condition of [51] in our setting. Quite generally, the sufficient Haar condition of Crain for the existence of MLE is in the uniqueness of a linear problem while our necessary and sufficient condition is in the uniqueness of a linear-programming problem. The latter is still computationally manageable but more subtle (and optimal); see also the last sentence in Example 6.2.1 for a difference between these two conditions in a very simple setting.

## 6.3 Applications

Maximization of likelihood is fundamental in estimation, model selection and testing. In many procedures it is important to know if MLE actually exists for given data $x_1, \dots, x_n$ and the linear space of exponents $\mathcal{B}$; see [75] (Introduction) for a list of such problems. [75] interpret the existence of MLE by using the geometry of the polyhedral cone spanned by the rows of a specific design matrix. This result is connected with the criterion of [11]. They also inquire which parameters are estimable when MLE is missing.
Below we show that the notion of the set of uniqueness is useful in characterizing the existence of MLE in discrete exponential families for specific spaces $\mathcal{B}$. There are two types of results we propose:

(a) conditions for the existence of MLE for a given sample,

(b) probability bounds for the existence of MLE for independent identically distributed samples.

To this end let $\mathcal{X}$ and $\mathcal{B}$ be as in Section 6.1.1. Let $X_1, X_2, \dots$ be *i.i.d.* random variables with values in $\mathcal{X}$. We define the random (stopping) time

$$\nu_{\text{uniq}} = \inf\{n \geq 1 : \ \{X_1, \dots, X_n\} \text{ is a set of uniqueness for } \mathcal{B}_+\}.$$

We will estimate tails of the distribution of $\nu_{\text{uniq}}$ in terms of $\mathcal{X}$, $\mathcal{B}$ and $n$. Typically we are interested in uniformly distributed $X_i$'s: $\mathbb{P}(X_i = x) = 1/K$, $x \in \mathcal{X}$, $i = 1, 2, \dots$, where $K = |\mathcal{X}|$. In the setting of Theorem 6.2.2 we consider $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$. We fix arbitrary $\mu > 0$ on $\mathcal{X}$, see Remark 6.2.3. Here is a trivial observation.

**Lemma 6.3.1.** *MLE for $e(\mathbb{R}^{\mathcal{X}})$ and $x_1, \dots, x_n$ exists if and only if $\{x_1, \dots, x_n\} = \mathcal{X}$.*

*Proof.* By Theorem 6.2.2 it is enough to verify that $\mathcal{X}$ is the only set of uniqueness for $\mathbb{R}_+^{\mathcal{X}}$. Obviously, $\mathcal{X}$ is a set of uniqueness for $\mathbb{R}_+^{\mathcal{X}}$ (in fact for $\mathbb{R}^{\mathcal{X}}$). On the other hand, if $U \subset \mathcal{X}$ and $x_0 \in \mathcal{X} \setminus U$, then $\mathbb{1}_{x_0}$ vanishes on $U$ but not on $\mathcal{X}$, hence $U$ is not of uniqueness for $\mathbb{R}_+^{\mathcal{X}}$ (neither it is for $\mathbb{R}^{\mathcal{X}}$). □

**Example 6.3.1.** Using notation of Section 6.2.1, we have $U_{\mathcal{B}_+} = U$, for every $U \subset \mathcal{X}$. Clearly, $U \subset U_{\mathcal{B}_+}$. On the other hand, using Equation (6.2.3), one may observe that for every $x \notin U$ the function $\phi(x) = \mathbb{1}_{\{x\}} \in \mathcal{B}_+$ and $\phi = \{0\}$ on $U$, but $x \notin \phi^{-1}(\{0\})$, so $U_{\mathcal{B}_+} \subset U$. In particular, $\{x_1, \dots, x_n\}_{\mathcal{B}_+} = \{x_1, \dots, x_n\}$ is the new state space $\widetilde{\mathcal{X}}$.

Later on we give examples which use the full strength of Theorem 6.2.2 and the nonnegativity of functions in $\mathcal{B}_+$ therein. For now we propose a probabilistic consequence of Lemma 6.3.1.

**Corollary 6.3.1.** *Let $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$ and $K = |\mathcal{X}|$. Let $X_1, X_2, \ldots$ be independent random variables, each with uniform distribution on $\mathcal{X}$. Then, for every $c \in \mathbb{R}$,*

$$\lim_{K\to\infty} \mathbb{P}\left(\nu_{uniq} < K \log K + Kc\right) = e^{-e^{-c}}.$$

*Proof.* Let $\nu_{\mathcal{X}} = \inf\{n \geq 1 : \ \{X_1, \ldots, X_n\} = \mathcal{X}\}$. The random variable $\nu_{\mathcal{X}}$ yields a connection to the classical Coupon Collector Problem, see [68], and [143]. Namely, by [68],

$$\lim_{K\to\infty} \mathbb{P}\left(\nu_{\mathcal{X}} < K \log K + Kc\right) = e^{-e^{-c}}.$$

By Lemma 6.3.1, $\nu_{\mathcal{X}} = \nu_{\text{uniq}}$, and the proof is complete. □

We aim to cover with large probability the whole of $\mathcal{X}$ by a sample of suitable size depending on $K$.

**Corollary 6.3.2.** *Let $\varepsilon \in (0,1)$, $K = |\mathcal{X}|$ and $\mathcal{B} = \mathbb{R}^{\mathcal{X}}$. Let $X_1, X_2, \ldots$ be independent random variables, each with uniform distribution on $\mathcal{X}$. If $K \to \infty$, then*

$$\mathbb{P}\left(\nu_{uniq} < (1-\varepsilon) K \log K\right) \to 0 \quad \text{and} \quad \mathbb{P}\left(\nu_{uniq} < (1+\varepsilon) K \log K\right) \to 1. \tag{6.3.1}$$

*Proof.* By Lemma 6.3.1 and Corollary 6.3.1, for every $c \in \mathbb{R}$ we get

$$\begin{aligned}\limsup_{K\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} < (1-\varepsilon) K \log K\right) &\leq \limsup_{K\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} < K \log K + Kc\right) \\ &= e^{-e^{-c}}.\end{aligned}$$

Thus $\lim_{K\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} < (1-\varepsilon) K \log K\right) = 0$. The second part of (6.3.1) is obtained analogously. □

**Remark 6.3.2.** We summarize (6.3.1) by saying that $K \log K$ is a *sharp threshold* of the sample size for the existence of MLE for $e(\mathbb{R}^{\mathcal{X}})$ and uniform *i.i.d.* samples. Sharp thresholds are widely used in the theory of random graphs, see [67] (Equation 3). It is also convenient to use them here to indicate the minimal size of *i.i.d.* samples that guarantees the existence of MLE with high probability.

### 6.3.1 Rademacher functions

For $k \in \mathbb{N}$, let us consider $\mathcal{X} = Q_k := \{-1, 1\}^k$, the $k$-dimensional discrete cube with, say, the uniform weight $\mu(\chi) = 2^{-k}$, $\chi \in Q_k$ (but see Remark 6.2.3). Thus, $K = |\mathcal{X}| = 2^k$. For $j = 1, \ldots, k$ and $\chi = (\chi_1, \ldots, \chi_k) \in Q_k$ we define the Rademacher functions:

$$r_j(\chi) = \chi_j,$$

and we denote $r_0(\chi) = 1$. Let

$$\mathcal{B}^k = \text{Lin}\{r_0, r_1, \ldots, r_k\}.$$

We define, as usual, the exponential family

$$e(\mathcal{B}^k) = \{e(r) : \ r \in \mathcal{B}^k\}.$$

**Theorem 6.3.3.** *MLE for $e(\mathcal{B}^k)$ and $x_1, \ldots, x_n \in Q_k$ exists if and only if for all $j = 1, \ldots, k$ we have $\{r_j(x_1), \ldots, r_j(x_n)\} = \{-1, 1\}$.*

*Proof.* By Theorem 6.2.2 we only need to prove that the above condition characterizes the sets of uniqueness for $\mathcal{B}_+^k$. If $j \in \{1, \ldots, k\}$ is such that $r_j(x_1) = \ldots = r_j(x_n) = 1$, then we let $r = r_0 - r_j$. Clearly, $r \in \mathcal{B}_+^k$ and $r$ is not identically zero, but $r(x_i) = 0$ for all $i = 1, \ldots, n$. Thus, $\{x_1, \ldots, x_n\}$ is not a set of uniqueness for $\mathcal{B}_+^k$. Similarly, if $r_j(x_1) = \ldots = r_j(x_n) = -1$, then we consider the function $r = r_0 + r_j \in \mathcal{B}_+^k$. For the converse implication we consider arbitrary

$$r = \sum_{j=0}^{k} a_j r_j \in \mathcal{B}_+^k.$$

Let $\chi = -(\text{sign}(a_1), \ldots, \text{sign}(a_k))$, where, say, $\text{sign}(0) = 1$. Obviously, $\chi \in Q_k$, and since $r(\chi) \geq 0$, we get

$$a_0 \geq \sum_{j=1}^{k} |a_j|. \tag{6.3.2}$$

Assume that $r = 0$ on $\{x_1, \ldots, x_n\}$. Let $j \in \{1, \ldots, k\}$. There are $x, x' \in \{x_1, \ldots, x_n\}$ such that $r_j(x) = 1$ and $r_j(x') = -1$. We have

$$0 = r(x) + r(x') = 2a_0 + \sum_{i \neq j} a_i [r_i(x) + r_i(x')].$$

It follows that

$$a_0 \leq \sum_{i \neq j} |a_i|.$$

By (6.3.2), $a_j = 0$, for every $j \geq 1$. Thereby $a_0 = 0$ and $r \equiv 0$. We see that $\{x_1, \ldots, x_n\}$ is a set of uniqueness for $\mathcal{B}_+^k$. □

**Example 6.3.4.** Let $x \in Q_k$ be arbitrary. By Theorem 6.3.3, MLE for $e\left(\mathcal{B}^k\right)$ and $\{x, -x\}$ exists.

We define the *positive* and *negative half-cubes*, respectively:

$$H_j^+ = \{\chi \in Q_k : \ r_j(\chi) = 1\}, \quad H_j^- = \{\chi \in Q_k : \ r_j(\chi) = -1\}, \quad j = 1, \ldots, k. \tag{6.3.3}$$

We note that $\mathcal{B}^k$ is also spanned by the indicator functions of half-cubes, namely $\mathbb{1}_j^+ = (r_0 + r_j)/2$ and $\mathbb{1}_j^- = (r_0 - r_j)/2$, $j = 1, \ldots, k$.

**Corollary 6.3.3.** *MLE for $e(\mathcal{B}^k)$ and $x_1, \ldots, x_n \in Q_k$ exists if and only if $\{x_1, \ldots, x_n\}$ has a nonempty intersection with each half-cube.*

The proof of Corollary 6.3.3 is immediate from Theorem 6.3.3 and the discussion above.

**Example 6.3.5.** If MLE fails to exist for $e(\mathcal{B}^k)$ and $x_1, \ldots, x_n \in Q_k$, then the following analysis may shed some light on Theorem 6.2.5. Let

$$J = \{j \in \{1, \ldots, k\} : \ \{r_j(x_1), \ldots, r_j(x_n)\} = \{-1, 1\}\}, \quad J' = \{1, \ldots, k\} \setminus J.$$

Since we consider the case when MLE does not exist, by Theorem 6.3.3, $J' \neq \varnothing$. For $j \in J'$ we let

$$H_j = \{\chi \in Q_k : \ r_j(\chi) = r_j(x_1) = \ldots = r_j(x_n)\}.$$

Clearly, this is a half-cube, see (6.3.3). We will show that

$$\{x_1,\dots,x_n\}_{\mathcal{B}_+^k} = \bigcap_{j\in J'} H_j. \tag{6.3.4}$$

We note that for $j \in J'$, $r_j$ is constant on the right-hand side of (6.3.4). Accordingly, the right-hand side of (6.3.4) is isomorphic to $\{-1,1\}^{|J|}$ or to $Q_{|J|}$.
Now if $r = \sum_{j=0}^k a_j r_j \in \mathcal{B}_+^k$ and $r(x_1) = \ldots = r(x_n) = 0$, then $r = \sum_{j\in J} a_j r_j + c \geq 0$ on $\{-1,1\}^{|J|}$, where $c = a_0 + \sum_{j\in J'} a_j r_j(x_1)$ is the sum of terms which are constant on $\bigcap_{j\in J'} H_j$. In the case when $J = \varnothing$, it is obvious that $\{x_1,\dots,x_n\}_{\mathcal{B}_+^k} = \bigcap_{j\in J'} H_j = \{x_1\}$, since $x_1 = \ldots = x_n$. However, if $J \neq \varnothing$, then by definition of $J$ and Theorem 6.3.3 with $k = |J|$, $r = 0$ on $\bigcap_{j\in J'} H_j$. Thus $\bigcap_{j\in J'} H_j \subset \{x_1,\dots,x_n\}_{\mathcal{B}_+^k}$. On the other hand, we observe that for each $j \in J'$, $\mathbb{1}_{H_j^c} = 0$ on the sample and $\mathbb{1}_{H_j^c} > 0$ on $H_j^c$, hence $H_j^c \cap \{x_1,\dots,x_n\}_{\mathcal{B}_+^k} = \varnothing$ and $\{x_1,\dots,x_n\}_{\mathcal{B}_+^k} \subset \bigcap_{j\in J'} H_j$. By Theorem 6.2.5, MLE exists for $e(\widetilde{\mathcal{B}}^k)$ and $x_1,\dots,x_n$ with the measure $\widetilde{\mu} := \mu|_{\widetilde{\mathcal{X}}}$ on $\widetilde{\mathcal{X}} := \bigcap_{j\in J'} H_j$. Of course, $\widetilde{X}$ is isomorphic with $Q_{|J|}$, if we ignore the $J'$ coordinates of the points in $\widetilde{X}$. In this way we may also think that $\widetilde{\mu}$ and $x_1,\dots,x_n$ are on $Q_{|J|}$. Thus, one may calculate the supremum of the log-likelihood function for $e(\mathcal{B}^k)$, $x_1,\dots,x_n$ and $\mu$ as the maximum of a log-likelihood function on $Q_{|J|}$. Of course, the total mass of $\widetilde{\mu}$ is a fraction of that of $\mu$. For instance, if $\mu$ is the uniform probability weight on $Q_k$ then $\widetilde{\mu}$ is uniform with the total mass $2^{-|J'|}$, which adds $n|J'|\log 2$ to the log-likelihood that would be obtained for $Q_{|J|}$ with the uniform probability weight, see, e.g., (6.1.2).

Here is a probabilistic application of Theorem 6.3.3.

**Corollary 6.3.4.** *Let $k \in \mathbb{N}$ and $X_1, X_2, \dots, X_n$ be independent random variables, each with uniform distribution on $Q_k$. Then,*

$$\begin{aligned}\mathbb{P}\left(\textit{MLE exists for } e(\mathcal{B}^k) \textit{ and } X_1,\dots,X_n\right) &= \left(1 - \frac{1}{2^{n-1}}\right)^k \\ &\geq 1 - \frac{k}{2^{n-1}} \to 1, \ \textit{as } n\to\infty.\end{aligned}$$

*Proof.* We have $\mathbb{P}(X_i = x) = 2^{-k}$ for all $x \in Q_k$ and $i = 1,\dots,n$. We let $R_{ij} = r_j(X_i)$ for $i = 1,\dots,n$ and $j = 1,\dots,k$. Thus, $\mathbb{P}(R_{ij} = 1) = \mathbb{P}(R_{ij} = -1) = \frac{1}{2}$ and $\{R_{ij}\}_{i,j}$ are independent. By Theorem 6.3.3,

$$\begin{aligned}&\mathbb{P}\left(\text{MLE exists for } e(\mathcal{B}^k) \text{ and } X_1,\dots,X_n\right)\\ =&\mathbb{P}\left(\{R_{ij} :\ i = 1,\dots,n\} = \{-1,1\} \text{ for } j = 1,\dots,k\right) = \left(1 - \frac{2}{2^n}\right)^k.\end{aligned}$$

Applying the Bernoulli inequality finishes the proof. □

**Corollary 6.3.5.** *For $k \in \mathbb{N}$ let $X_1,\dots,X_{n(k)}$ be independent random variables, each with uniform distribution on $Q_k$. If $n(k) = \log_2 k + b + o(1)$ for some $b \in \mathbb{R}$ as $k\to\infty$, then*

$$\lim_{k\to\infty} \mathbb{P}\left(\textit{MLE exists for } e(\mathcal{B}^k) \textit{ and } X_1,\dots,X_{n(k)}\right) = e^{-2^{1-b}}.$$

*Proof.* By Corollary 6.3.4,

$$\mathbb{P}\left(\text{MLE exists for } e(\mathcal{B}^k) \text{ and } X_1,\dots,X_{n(k)}\right) = \left(1 - \frac{1}{k\ 2^{b-1+o(1)}}\right)^k$$

$$\to e^{-2^{1-b}}, \text{ as } k \to \infty. \tag{6.3.5}$$

□

**Corollary 6.3.6.** $\log_2 k$ *is a sharp threshold of the sample size for the existence of MLE for* $e(\mathcal{B}^k)$ *and i.i.d. uniform samples on* $Q_k$.

*Proof.* Let $\varepsilon \in (0,1)$ and (the sample size) $n = n(k) < (1-\varepsilon)\log_2 k$. Then,

$$\mathbb{P}\left(\nu_{\text{uniq}} < n\right) \leq \mathbb{P}\left(\nu_{\text{uniq}} < (1-\varepsilon)\log_2 k\right).$$

For every $b \in \mathbb{R}$ by the equation in (6.3.5) we have

$$\begin{aligned}\limsup_{k\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} < (1-\varepsilon)\log_2 k\right) &\leq \limsup_{k\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} < \log_2 k + b\right) \\ &= e^{-2^{1-b}}.\end{aligned}$$

Since $b$ is arbitrary, we conclude that $\limsup_{k\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} < n(k)\right) = 0$. Analogously, for the sample size $n = n(k) > (1+\varepsilon)\log_2 k$ we get

$$\liminf_{k\to\infty} \mathbb{P}\left(\nu_{\text{uniq}} > n(k)\right) = 1,$$

which ends the proof. □

The above is in stark contrast to Corollary 6.3.2, as summarized in Remark 6.3.2. Indeed, in the present setting we have $K = |Q_k| = 2^k$, so the sharp threshold for the sample size needed for the existence of MLE is $\log_2 \log_2 K$. The following result on the expectation of $\nu_{\text{uniq}}$ agrees well with the sharp threshold.

**Lemma 6.3.2.** *Let* $\nu_{uniq}$ *be as in Corollary 6.3.5. Let* $H_k = \sum_{i=1}^{k} \frac{1}{k}$ *be the* $k$*-th harmonic number. Then,*

$$\frac{H_k}{\log 2} + 1 \leq \mathbb{E}(\nu_{uniq}) < \frac{H_k}{\log 2} + 2, \quad k = 1, 2, \ldots.$$

*Proof.* Observe that $\nu_{\text{uniq}} = \max\{\tau_1, \ldots, \tau_k\}$, where

$$\tau_j = \min\left\{n \geq 1 : \ \{r_j(X_1), \ldots, r_j(X_n)\} = \{-1, 1\}\right\}, \quad j = 1, \ldots, k.$$

From the fact that $X_1, X_2, \ldots$ are independent and uniformly distributed, we deduce that

$$\mathbb{1}_{r_j(X_i) \neq r_j(X_1)}, \quad i = 2, 3, \ldots, \quad j = 1, 2 \ldots,$$

are independent with symmetric Bernoulli distribution. Then $\tau_1, \ldots, \tau_k$ are independent, and

$$\tau_j + 1 \sim \mathit{Geom}\,(1/2)$$

for $j = 1, \ldots, k$. The result follows from [64]. □

In Section 6.5 we return to Rademacher functions, but for now we turn to exponential families of random graphs, a major motivation for this work.

## 6.4 Random graphs

In this section we focus on random graphs. Their various applications can be found in [150], [159] and [134]. What is important for us, many such models are indeed discrete exponential families. As usual, maximum likelihood can be used to select a suitable graph model within the exponential family, see, e.g., [142] (Chapter 1 and 8) and [18]. In this section we characterize the existence of MLE in such context. The theory of random graphs started with probabilistic proofs of the existence or nonexistence of specific graphs by Erdős, see, e.g., [28]. Asymptotic properties of random graphs were developed in the seminal papers of [66, 67] and [87]. [150] discuss geometric interpretations of the existence of MLE for discrete exponential families with applications to random graphs and social networks. [45] give normalizing constants that are crucial for the computation of MLE for exponential random graph models. Furthermore, they include examples when MLE fails to exist. The same authors together with Sly discuss in [46] the asymptotic probability of the existence and uniqueness of MLE for the $\beta$-model of graphs. This allows to connect the $\beta$-model with a random uniform model of graphs with a given degree sequence, which is then explored using graphons (graph limits, see [127]). They also present an algorithm for the computation of MLE in the $\beta$-model.
[141] put nonasymptotic conditions for the existence of MLE in various random graph models parameterized by vertex-specific parameters. [151] characterize the existence of MLE for $\beta$-models. They interpret the Barndorff-Nielsen's criterion using the geometry of multidimensional polytopes of vertex-degree sequences, see also [75]. [185] transfer the criterion into discrete hierarchical models, using the notion of simplicial complices. These models include, e.g., graphical models and Ising models. Wang, Rauh and Massam also improve the approximation of the set of estimable parameters in the case of the nonexistence of MLE, which is discussed in the setting of marginal polytopes.
Let us start with the notation. Graph is a pair $G = (V, E)$, where $V = \{1, \dots, N\}, N \in \mathbb{N}$, is the set of nodes and $E$ is the set of edges, i.e.,

$$E \subset \binom{V}{2} := \{(r, s) : \ 1 \leq r < s \leq N\}.$$

We only consider simple undirected graphs (containing no loops or multiple edges). Let $m = |E|$. If $m = \binom{N}{2}$, then the graph is called complete and is denoted as $K_N$. On the other hand, the empty graph (with $m = 0$) is denoted as $\overline{K_N}$. For graphs $G = (V, E_1)$ and $H = (V, E_2)$ we let, as usual,

$$G \cup H := (V, E_1 \cup E_2), \qquad G \cap H := (V, E_1 \cap E_2).$$

Furthermore, $G \subset H$ means that $E_1 \subset E_2$. Let $\mathcal{G}_N$ be the family of all the graphs with $N$ nodes, i.e., with $V = \{1, \dots, N\}$. By a random graph we understand a random variable $\mathbb{G}$ with values in $\mathcal{G}_N$. The families of distributions of such random variables are called random graph models. We focus on the exponential model of random graphs $\mathcal{G}_{N,c}$ defined as follows.
For $1 \leq r < s \leq N$ and $G \in \mathcal{G}_N$, we let

$$\mathbb{1}_G(r, s) = \begin{cases} 1, & \text{if } (r, s) \in E, \\ 0, & \text{otherwise.} \end{cases}$$

We define $\chi_{r,s} : \mathcal{G}_N \to \{-1, 1\}$ by $\chi_{r,s}(G) = 1 - 2\mathbb{1}_G(r, s)$. We consider the linear space

$$\mathcal{B}^{\mathcal{G}_N} = \mathrm{Lin}\Big\{ \ 1, \chi_{r,s}(G) : \ 1 \leq r < s \leq N \Big\}.$$

Let $c \in \mathbb{R}^{\binom{V}{2}}$ be a corresponding vector of coefficients. Following the setting of Section 6.1.1 we let $\mu(G) = 1$ for each $G \in \mathcal{G}_N$ (but see Remark 6.2.3) and consider the exponential family

$$\mathcal{G}_{N,c} := e(\mathcal{B}^{\mathcal{G}_N}) = \left\{ p_c := e^{\phi_c - \psi(\phi_c)} : \ c \in \mathbb{R}^{\binom{V}{2}} \right\}, \tag{6.4.1}$$

where

$$\phi_c(G) = \sum_{(r,s)\in\binom{V}{2}} c_{r,s}\chi_{r,s}(G), \qquad\qquad \psi(\phi_c) = \log \sum_{G\in\mathcal{G}_N} e^{\phi_c(G)},$$

for $G \in \mathcal{G}_N$, see also (6.1.3). As usual, for $p_c \in \mathcal{G}_{N,c}$ we let $L_{p_c}(G_1, \ldots, G_n) = \prod_{i=1}^n p_c(G_i)$, etc.

**Lemma 6.4.1.** *Let $c \in \mathbb{R}^{\binom{V}{2}}$ and let $\mathbb{G}$ be a random graph with distribution $\mathcal{G}_{N,c}$. Let $1 \le r < s \le N$. Then the probability of the appearance of the edge $(r, s)$ in $\mathbb{G}$ equals*

$$p_{r,s} = \frac{e^{c_{r,s}}}{1 + e^{c_{r,s}}}. \tag{6.4.2}$$

The result is well known but for convenience a proof is given in Appendix 6.6.3.

**Lemma 6.4.2.** *Let $c \in \mathbb{R}^{\binom{V}{2}}$ and let $\mathbb{G}$ be a random graph with distribution $\mathcal{G}_{N,c}$. Let $1 \le r_1, s_1, r_2, s_2 \le N$, $r_1 < s_1, r_2 < s_2$, and $(r_1, s_1) \ne (r_2, s_2)$. Then the appearances of edges $(r_1, s_1)$ and $(r_2, s_2)$ in $\mathbb{G}$ are independent events.*

The proof of the result is similar to that of Lemma 6.4.1, and can be found in Appendix 6.6.4. For instance, if $p_{r,s} = p \in (0, 1)$ for every edge $(r, s)$, then the exponential random graph with distribution $\mathcal{G}_{N,c}$ is the Erdős-Rényi random graph $\mathcal{G}_{N,p}$ in [66, 67]. The latter means that $\mathbb{P}(e \in E(\mathbb{G})) = p$ for every edge $e \in \binom{V}{2}$, and the events $e \in E(\mathbb{G})$ and $f \in E(\mathbb{G})$ are independent for different edges $e$, $f$.

**Theorem 6.4.1.** *MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and $G_1, \ldots, G_n \in \mathcal{G}_N$ exists if and only if*

$$\bigcup_{i=1}^n G_i = K_N \qquad\qquad \text{and} \qquad\qquad \bigcap_{i=1}^n G_i = \overline{K_N}.$$

*Proof.* By Theorem **6.2.2**, MLE exists if and only if $\{G_1, \ldots, G_n\}$ is of uniqueness for $\mathcal{B}_+^{\mathcal{G}_N}$. We first prove the "only if" part of Theorem 6.4.1. Let us assume that there exists an edge $(r_0, s_0) \notin \bigcup_{i=1}^n G_i$. Then the function $\chi_{r_0,s_0} \in \mathcal{B}_+^{\mathcal{G}_N}$ equals zero on $G_1, \ldots, G_n$, but not on the whole $\mathcal{G}_N$. In addition, if there is an edge $(r_0, s_0) \in \bigcap_{i=1}^n G_i$, then the function $(1 + \chi_{r_0,s_0}) \in \mathcal{B}_+^{\mathcal{G}_N}$ vanishes for $G_1, \ldots, G_n$, but it is not equal to zero, e.g., for the graph $\overline{K_N}$. We next prove the 'if' part of the theorem. Let $\phi = k_0 + \sum_{r<s} k_{r,s}\chi_{r,s} \in \mathcal{B}_+^{\mathcal{G}_N}$, where $k_0, k_{r,s} \in \mathbb{R}$ for all $1 \le r < s \le N$. Since $\phi(G) \ge 0$ for every $G \in \mathcal{G}_N$,

$$k_0 \ge \sum_{r<s} |k_{r,s}|. \tag{6.4.3}$$

Let $(r_0, s_0) \in \binom{V}{2}$. Let $\phi(G_1) = \ldots = \phi(G_n) = 0$. Since $\bigcup_{i=1}^n G_i = K_N$ and $\bigcap_{i=1}^n G_i = \overline{K_N}$, there exists a pair of graphs $G', G'' \in \{G_1, \ldots, G_n\}$ such that $\chi_{r_0,s_0}(G') = 1$, $\chi_{r_0,s_0}(G'') = -1$. Therefore,

$$0 = \phi(G') + \phi(G'') = 2k_0 + \sum_{r<s} k_{r,s} \left(\chi_{r,s}(G') + \chi_{r,s}(G'')\right)$$

$$= 2k_0 + \sum_{\substack{r<s \\ (r,s)\neq(r_0,s_0)}} k_{r,s}\left(\chi_{r,s}(G') + \chi_{r,s}(G'')\right).$$

It follows that $k_0 \leq \sum_{(r,s)\neq(r_0,s_0)} |k_{r,s}|$ and eventually we get $k_{r_0,s_0} = 0$, thanks to (6.4.3). Since $(r_0, s_0)$ is arbitrary, $k_{r,s} = 0$ for every $1 \leq r < s \leq N$. Then also $c_0 = 0$, and thus $\phi \equiv 0$. □

In the above random graph model it is possible to compute explicitly the probability of the existence of MLE for *i.i.d.* samples of graphs in $\mathcal{G}_N$. To this end, for $1 \leq r < s \leq N$ we fix $c_{r,s} \in \mathbb{R}$. By Lemma 6.4.1 the probability of the appearance of the edge $(r, s)$ in random graph $\mathbb{G}$ with distribution $\mathcal{G}_{N,c}$ is

$$p_{r,s} = \frac{e^{c_{r,s}}}{1 + e^{c_{r,s}}}.$$

**Lemma 6.4.3.** *Let $\{\mathbb{G}_1, \ldots, \mathbb{G}_n\}$ be i.i.d. with distribution $\mathcal{G}_{N,c}$. Then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ equals*

$$\prod_{1\leq r<s\leq N} \left(1 - p_{r,s}^n - (1 - p_{r,s})^n\right). \tag{6.4.4}$$

*Proof.* By Theorem **6.4.1**, MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ exists if and only if among the random graphs $\mathbb{G}_1, \ldots, \mathbb{G}_n$ every edge $(r, s)$, $1 \leq r < s \leq N$, appears at least once, but not $n$ times. For every edge $(r, s)$ the above condition is satisfied with probability $1 - (1 - p_{r,s})^n - (p_{r,s})^n$. The independence of the occurrences of different edges in $\mathcal{G}_{N,c}$ yields the product (6.4.4). □

In particular, if $c = 0$, then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ equals

$$\left(1 - 2^{1-n}\right)^{\binom{N}{2}},$$

which is an analogue of Corollary 6.3.5. From the above results we can deduce asymptotic bounds for the *i.i.d.* sample size for which MLE exists with high probability. To this end we recall the classical result on $p = p(N) \in (0, 1)$ such that $\mathbb{G}$ from $\mathcal{G}_{N,p}$ has at least one edge with high probability.

**Remark 6.4.2.** [80] (Lemma 1.10) Let $\mathbb{G}_{N,p(N)}$ be a random graph with distribution $\mathcal{G}_{N,p(N)}$. Then

$$\lim_{N\to\infty} \mathbb{P}\left(\mathbb{G}_{N,p(N)} \text{ has at least one edge}\right) = \begin{cases} 0 & \text{if} \quad p(N) = o\left(N^{-2}\right), \\ 1 & \text{if} \quad N^{-2} = o\left(p(N)\right). \end{cases}$$

The above may be summarized by saying that $N^{-2}$ is a *threshold* for the probability $p$ such that $\mathbb{G}$ with distribution $\mathcal{G}_{N,p}$ has at least one edge. For more information on threshold functions in the theory of random graphs see, e.g., [80]. In particular, a sharp threshold (mentioned previously) is a threshold but the converse is not true in general.

**Lemma 6.4.4.** *Let $\mathbb{G}_1, \ldots, \mathbb{G}_n$ be i.i.d. random variables with distribution $\mathcal{G}_{N,c}$. Then $\log N$ is a threshold of the sample size $n$ for the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$.*

*Proof.* According to Lemma 6.4.3, the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and $\mathbb{G}_1, \ldots, \mathbb{G}_n$ equals

$$P_{\text{MLE}} = \prod_{1\leq r<s\leq N} \left(1 - p_{r,s}^n - (1 - p_{r,s})^n\right).$$

We define the function

$$f(x) = 1 - x^w - (1-x)^w, \quad x \in (0,1),\ w \geq 2. \tag{6.4.5}$$

Clearly, $f(x) = f(1-x)$ and for $w \geq 2$ we have $f$ increasing when $0 < x < \frac{1}{2}$ and decreasing when $\frac{1}{2} < x < 1$. Using (6.4.5) we can bound $P_{\text{MLE}}$ from above by

$$P_{\text{BIG}} := \left(1 - 2^{1-n}\right)^{\binom{N}{2}}.$$

Applying Corollary 6.3.4 and the equality in (6.3.5) for $k = \binom{N}{2}$, we observe that for every $b \in \mathbb{R}$ and for $n = n(N) = \log_2 \binom{N}{2} + b + o(1)$ we have $P_{\text{BIG}} \to e^{-2^{1-b}}$, as $N \to \infty$. Therefore, for $n(N) = o(\log N)$ we obtain $P_{\text{MLE}} \leq P_{\text{BIG}} \to 0$, as $N \to \infty$.
We consider the sample size $n = n(N)$ (depending on $N$). We will prove that if $\log N/n \to 0$ as $N \to \infty$, then $P_{\text{MLE}} \to 1$. To this end we bound $P_{\text{MLE}}$ from below by

$$P_{\text{SMALL}} := (1 - p_{\max}^n - (1-p_{\max})^n)^{\binom{N}{2}},$$

where $c_{\max} = \max_{1 \leq r < s \leq N} |c_{r,s}|$ and $p_{\max} = e^{c_{\max}}/(1 + e^{c_{\max}})$.
Take $n$ independent Erdős-Rényi random graphs $\mathbb{H}_1, \ldots, \mathbb{H}_n$ with distribution $\mathcal{G}_{N,p_{\max}}$. Then the probability of the existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and for $\mathbb{H}_1, \ldots, \mathbb{H}_n$ equals exactly $P_{\text{SMALL}}$. Note that intersection and union of the graphs are also Erdős-Rényi random graphs, namely

$$\bigcap_{i=1}^{n} \mathbb{H}_i \sim \mathcal{G}_{N,p_{\max}^n}, \qquad \bigcup_{i=1}^{n} \mathbb{H}_i = \overline{\bigcap_{i=1}^{n} \overline{\mathbb{H}_i}} \sim \mathcal{G}_{N,1-q_{\max}^n},$$

where

$$q_{\max} := 1 - p_{\max} = \frac{e^{-c_{\max}}}{1 + e^{-c_{\max}}}.$$

From Remark 6.4.2, with high probability we have

$$\bigcap_{i=1}^{n} \mathbb{H}_i = \overline{K_N} \qquad \text{and} \qquad \overline{\bigcup_{i=1}^{n} \mathbb{H}_i} = \overline{K_N},$$

provided

$$p_{\max}^n = o(N^{-2}) \qquad \text{and} \qquad q_{\max}^n = o(N^{-2}).$$

By definition, $c_{\max} > 0$, so $p_{\max} > q_{\max}$. In order to get $P_{\text{SMALL}} \to 1$ as $n \to \infty$, it suffices to have $p_{\max}^n = o(N^{-2})$. If $n(N)/\log N \to \infty$ as $N \to \infty$, then the above condition is satisfied. Therefore $\log N$ is a threshold of the sample size for existence of MLE for $e(\mathcal{B}^{\mathcal{G}_N})$ and independent $\mathbb{G}_1, \ldots, \mathbb{G}_n$ from $\mathcal{G}_{N,c}$. □

## 6.5 Applications to Walsh functions

We return to Rademacher functions to discuss the spaces spanned by their products. Let $k \in \mathbb{N}$, $1 \leq q \leq k$, and

$$\mathcal{B}_q^k = \text{Lin}\left\{w_S :\ S \subset \{1, \ldots, k\} \text{ and } |S| \leq q\right\},$$

where

$$w_S(x) = \prod_{i \in S} r_i(x), \quad x \in Q_k, \quad S \subset \{1, \ldots, k\},$$

are the Walsh functions, see, e.g., [108].
The case $\mathcal{B}_1^k = \mathcal{B}^k$ was discussed in Section 6.3.1 and the case $q = 2$ is related to the Ising model of ferromagnetism in statistical mechanics, see [184] (Example 3.1).

**Lemma 6.5.1.** *The dimension of the linear space $\mathcal{B}_q^k$ is $\sum_{j=0}^{q} \binom{k}{j}$.*

The proof of Lemma 6.5.1 is given in Appendix 6.6.5.

**Corollary 6.5.1.** *For $q \leq \frac{k}{2}$ we have*

$$dim\left(\mathcal{B}_q^k\right) \leq 2^{kH_2(\frac{q}{k})} \leq \left(\frac{ek}{q}\right)^q,$$

*where $H_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function.*

The proof follows from Lemma 6.5.1 and the entropy bound for the sum of binomial coefficients, see, e.g., [84] (Theorem 3.1).
Characterization of the existence of MLE for $e(\mathcal{B}_q^k)$ and the related sharp thresholds seem to be hard for general $q$, even for $q = 2$, see Remark 6.5.2. In the next section we discuss the products of $k - q$ Rademacher functions for fixed $q \in \mathbb{N}$ $(q \leq k)$. We especially focus on the products of $k - 1$ and $k$ Rademacher functions. Below we characterize the existence of MLE for $e(\mathcal{B}_{k-1}^k)$. As we will see, we get a qualitatively different result than that in Section 6.3.1. Let $\mathcal{E}$ and $\mathcal{O}$ be the sets of all those points in $Q_k$ that have an even and odd number of positive coordinates, respectively.

**Theorem 6.5.1.** *MLE exists for $e(\mathcal{B}_{k-1}^k)$ and $x_1, \ldots, x_n \in Q_k$ if and only if $\mathcal{E}$ or $\mathcal{O} \subset \{x_1, \ldots, x_n\}$.*

*Proof.* Thanks to Theorem 6.2.2, we only need to characterize the sets of uniqueness for $\left(\mathcal{B}_{k-1}^k\right)_+$. To this end, we consider the hypercube $G_{Q_k}$, defined as the graph with vertices in $Q_k$ and edges between all pairs of points which differ by exactly one coordinate. Thus,

$$V(G_{Q_k}) = Q_k \text{ and } E(G_{Q_k}) = \{\{x, y\} \in Q_k \times Q_k : \ |\{j : r_j(x) \neq r_j(y)\}| = 1\}.$$

Let $U = \{x_1, \ldots, x_n\}$. Assume that $U$ is a set of uniqueness. Let $e \in \mathcal{E}$ and $o \in \mathcal{O}$. The hypercube graph $G_{Q_k}$ is connected, so there exists a path $(e, v_1, v_2, \ldots, v_{2p}, o)$ in $G_{Q_k}$. Then

$$\left(\mathbb{1}_{\{e,v_1\}} + \mathbb{1}_{\{v_2,v_3\}} + \ldots + \mathbb{1}_{\{v_{2p},o\}}\right) - \left(\mathbb{1}_{\{v_1,v_2\}} + \mathbb{1}_{\{v_3,v_4\}} + \ldots + \mathbb{1}_{\{v_{2p-1},v_{2p}\}}\right) \tag{6.5.1}$$

$$= \mathbb{1}_{\{e\}} + \mathbb{1}_{\{o\}} \tag{6.5.2}$$

is a nontrivial nonnegative function on $Q_k$. Therefore, we must have $\{e, o\} \cap U \neq \varnothing$. Then we easily conclude that $\mathcal{E} \subset U$ or $\mathcal{O} \subset U$.
For the converse implication, we consider $q \in \{0, \ldots, k\}$ and $(k-q)$-subcubes defined as follows,

$$\bigcap_{i=1}^{q} H_{j_i}, \tag{6.5.3}$$

where $1 \leq j_1 < j_2 < \ldots < j_q \leq k$ and $H_{j_i} = H_{j_i}^+$ or $H_{j_i}^-$, see (6.3.3). When $q = k - 1$, the intersection, or a 1-cube, is a pair of points in $Q_k$ which differ by exactly one coordinate, so they have a different parity. Moreover, each such pair can be obtained in this way. Using (6.5.3), as in the proof of Lemma 6.5.1 we see that $\mathbb{1}_{\{e,o\}} \in \mathcal{B}_{k-1}^k$ for each $e \in \mathcal{E}$ and $o \in \mathcal{O}$. Furthermore,

each $q$-subcube of $Q_k$ with $q \geq 1$ can be covered by disjoint pairs $\{e, o\}$ as above. Therefore, the functions $\mathbb{1}_{\{e,o\}} \in \mathcal{B}^k_{k-1}$ with $e \in \mathcal{E}$ and $o \in \mathcal{O}$ span the linear space $\mathcal{B}^k_{k-1}$.
We next claim that for every $f \in \mathcal{B}^k_{k-1}$,

$$\sum_{x \in \mathcal{O}} f(x) = \sum_{x \in \mathcal{E}} f(x). \tag{6.5.4}$$

Indeed, if $f = \mathbb{1}_{\{e,o\}}$ with $e \in \mathcal{E}$ and $o \in \mathcal{O}$, then the equality is true because both sides of (6.5.4) are equal to 1. Since such functions span $\mathcal{B}^k_{k-1}$ it follows that (6.5.4) is true for every $f \in \mathcal{B}^k_{k-1}$. Finally, if nonnegative $f \in \mathcal{B}^k_{k-1}$ vanishes on $\mathcal{E}$, then the sum over $\mathcal{O}$ also equals zero, hence $f \equiv 0$, and the same conclusion holds if we assume that $f = 0$ on $\mathcal{O}$. Thus $U$ is the set of uniqueness if $\mathcal{O} \subset U$ or $\mathcal{E} \subset U$. □

**Remark 6.5.2.** A naïve extension of Corollary 6.3.3 fails for $e(\mathcal{B}^k_2)$, if we try to replace the half-cubes with $(k-2)$-subcubes, that is, quarter-cubes. This is seen from Theorem 6.5.1 for $k = 3$. Indeed, the set

$$\{(1,1,-1),(1,-1,1),(-1,1,1),(1,-1,-1),(-1,1,-1),(-1,-1,1)\}$$

is not of uniqueness for $(\mathcal{B}^3_2)_+$, as follows from (6.5.1) with $e = (-1,-1,-1)$ and $o = (1,1,1)$, even though the set has nonmpty intersection with each quarter-cube.

We will briefly treat the case of $e(\mathcal{B}^k_k)$, as follows.

**Corollary 6.5.2.** *$k2^k \log 2$ is a sharp threshold of the sample size for the existence of MLE for $e(\mathcal{B}^k_k)$ and i.i.d. samples uniform on $Q_k$.*

*Proof.* Observe that $e(\mathcal{B}^k_k)$ is isomorphic to $e(\mathbb{R}^{\mathcal{X}})$ for $|\mathcal{X}| = 2^k$. The existence of MLE for $e(\mathcal{B}^k_k)$ is characterized in (more general) Lemma 6.3.1, and the sharp threshold is given after Corollary 6.3.2. □

Corollary 6.5.2 is in stark contrast with the result for the (smaller) space $e(\mathcal{B}^k_1)$ because for $e(\mathcal{B}^k_1)$ the sharp threshold, and so the threshold, equal $\log_2 k$, by Corollary 6.3.6.

**Remark 6.5.3.** Let $1 \leq q_1 \leq q_2 \leq k$. Then every set $U$ of uniqueness for $(\mathcal{B}^k_{q_2})_+$ is of uniqueness for $(\mathcal{B}^k_{q_1})_+$, because $(\mathcal{B}^k_{q_1})_+ \subset (\mathcal{B}^k_{q_2})_+$.

A characterization of the existence of MLE for $e(\mathcal{B}^k_q)$ for arbitrary $q$, even for $q = 2$, turned out to be difficult. Accordingly, we do not give a sharp threshold for the size of the uniform *i.i.d.* sample needed for the existence of MLE for $e(\mathcal{B}^k_q)$. However, the case of $e(\mathcal{B}^k_{k-q})$ seems a little easier in the sense that we are able to give the less precise threshold for the existence of MLE for $e(\mathcal{B}^k_{k-q})$. Moreover, for each fixed $q$ the threshold for $e(\mathcal{B}^k_{k-q})$ is the same as for $e(\mathcal{B}^k_k)$, namely $k2^k$ as $k \to \infty$.

**Lemma 6.5.2.** *Fix $q \in \mathbb{N}$. Then $k2^k$ is a threshold of the sample size for the existence of MLE for $e(\mathcal{B}^k_{k-q})$ and i.i.d. sample uniform on $Q_k$.*

*Proof.* If $\lim_{k\to\infty} n(k)/(k2^k) = \infty$, then by Remark 6.5.3 and Corollary 6.5.2, for $k \to \infty$ we get

$$\mathbb{P}\left(\left\{X_1, \ldots, X_{n(k)}\right\} \text{ is of uniqueness for } \left(\mathcal{B}^k_{k-q}\right)_+\right)$$

$$\geq \mathbb{P}\left(\left\{X_1,\dots,X_{n(k)}\right\} \text{ is of uniqueness for } \mathcal{B}_k^k\right) \to 1,$$

as needed. On the other hand, every set $U$ of uniqueness for $(\mathcal{B}_{k-q}^k)_+$ must intersect with every subcube defined by fixing last $k-q$ coordinates, because each $q$-subcube is the support of a function in $(\mathcal{B}_{k-q}^k)_+$, to wit, of its indicator. There are $2^{k-q}$ such $q$-subcubes, each of which we can suggestively denote by $(*,\dots,*,\varepsilon_{q+1},\dots,\varepsilon_k)$, where $\varepsilon_{q+1},\dots,\varepsilon_k = \pm 1$. Observe that the family of the above subcubes is a partition of $Q_k$. We consider each $q$-subcube as a coupon in the Coupon Collector Problem. If a sample point falls into the $q$-subcube, we consider the coupon as collected. The probability of collecting a given coupon is $2^{q-k}$. Therefore, if $n(k) = o\left(2^k k\right)$, hence $n(k) = o\left(2^{k-q}\,(k-q)\right)$, then

$$\mathbb{P}\left(\left\{X_1,\dots,X_{n(k)}\right\} \text{ is of uniqueness for } (\mathcal{B}_{k-q}^k)_+\right) \to 0, \qquad \text{as } k\to\infty,$$

as needed. □

## 6.6 Appendix

### 6.6.1 Proof of Lemma 6.1.1

Let $\hat{p} = e(\phi_0), \widetilde{p} = e(\phi_1) \in e(\mathcal{B})$ and $\widehat{p} \neq \widetilde{p}$, so that $\phi_1 - \phi_0 \neq const$. Let $\phi_t = \phi_0 + t(\phi_1 - \phi_0)$, $p_t = e(\phi_t)$ for $t \in \mathbb{R}$ and $l(t) = l_{p_t}(x_1,\dots,x_n)$. We claim that $l$ is strictly concave, that is $l'' < 0$. Indeed, since $\overline{\phi_t} = \overline{\phi_0} + t\overline{\phi_1}$ is a linear function, by (6.1.4) we get

$$l''(t) = -n\frac{d^2}{dt^2}\log Z(\phi_t).$$

Let $X$ be a random variable with values in $\mathcal{X}$ such that $\mathbb{P}(X=x) = p(x)\mu(x)$. As usual, for every $f : \mathcal{X} \to \mathbb{R}$ we have

$$\mathbb{E}f(X) = \sum_{x\in\mathcal{X}} f(x)p(x)\mu(x).$$

Clearly, $(\log Z(\phi_t))' = \frac{Z(\phi_t)'}{Z(\phi_t)}$ and $(\log Z(\phi_t))'' = \frac{Z(\phi_t)''}{Z(\phi_t)} - \left(\frac{Z(\phi_t)'}{Z(\phi_t)}\right)^2$. Hence, thanks to (6.1.1),

$$Z(\phi_t)' = \sum_{x\in\mathcal{X}} e^{\phi_t(x)}\mu(x)\,(\phi_1(x)-\phi_0(x))$$
$$Z(\phi_t)'' = \sum_{x\in\mathcal{X}} e^{\phi_t(x)}\mu(x)\,(\phi_1(x)-\phi_0(x))^2.$$

Thus,

$$\frac{Z(\phi_t)'}{Z(\phi_t)} = \mathbb{E}[\phi_1(X)-\phi_0(X)] \qquad\qquad \frac{Z(\phi_t)''}{Z(\phi_t)} = \mathbb{E}[\phi_1(X)-\phi_0(X)]^2$$

and so

$$\frac{d^2}{dt^2}\log Z(\phi_t) = \mathbb{E}\left[\phi_1(X)-\phi_0(X)-\mathbb{E}(\phi_1(X)-\phi_0(X))\right]^2 > 0,$$

since $\phi_1 - \phi_0$ is not constant. Hence, $l$ is strictly concave, in particular $l(1/2) > (l(0) + l(1))/2$. If $\sup_{p\in e(\mathcal{B})} L_p(x_1,\dots,x_n) = L_{\widehat{p}}(x_1,\dots,x_n) = L_{\widetilde{p}}(x_1,\dots,x_n)$, then $l(1/2) > \sup_{p\in e(\mathcal{B})} l_p(x_1,\dots,x_n)$, which is absurd; thus at most one of $\widetilde{p}$ and $\widehat{p}$ can be the MLE.

### 6.6.2 Control by oscillations

$\lambda_U$ defined in Section 6.2 may be thought of as a specific measure of oscillation of $\phi$. Of course, $\lambda_U \geq 0$. Furthermore, for every $c \in \mathbb{R}$,

$$\lambda_U(\phi + c) = \lambda_U(\phi), \quad \phi \in \mathcal{B}, \tag{6.6.1}$$

and for every (positive number) $k > 0$ we have (homogeneity),

$$\lambda_U(k\phi) = k\lambda_U(\phi), \quad \phi \in \mathcal{B}, k \geq 0. \tag{6.6.2}$$

If $U = \mathcal{X}$, then $\lambda_{\mathcal{X}}(-\phi) = \lambda_{\mathcal{X}}(\phi)$ for $\phi \in \mathcal{B}$, and so $\lambda_{\mathcal{X}}$ is a seminorm. Clearly, $\lambda_U \leq \lambda_{\mathcal{X}}$. However, if there is a nontrivial $\phi \in \mathcal{B}_+$ such that $\phi = 0$ on $U$, then $\lambda_U(\phi) = \sup_{\mathcal{X}} \phi > 0$ but $\lambda_U(-\phi) = 0$. The following result is the engine of Theorem 6.2.2.

**Lemma 6.6.1.** *$U \subset \mathcal{X}$ is the set of uniqueness for $\mathcal{B}_+$ if and only if $\lambda_U$ is comparable with $\lambda_{\mathcal{X}}$ on $\mathcal{B}$, i.e., there exist constants $c_1, c_2 > 0$ such that $c_1\lambda_{\mathcal{X}}(\phi) \leq \lambda_U(\phi) \leq \lambda_{\mathcal{X}}(\phi)$ for all $\phi \in \mathcal{B}$.*

*Proof.* We first prove the "if" part. Assume $U$ is not a set of uniqueness for $\mathcal{B}_+$. Then there exists a nonzero function $\phi \in \mathcal{B}_+$ such that $\phi = 0$ on $U$. We have $\lambda_U(-\phi) = 0$ and $\lambda_{\mathcal{X}}(-\phi) > 0$, hence $\lambda_U$ and $\lambda_{\mathcal{X}}$ are not comparable on $\mathcal{B}$.
We now prove the "only if" part, which is delicate. For all $\vartheta, \phi \in \mathcal{B}$ we have

$$\begin{aligned}\lambda_U(\vartheta + \phi) &\leq \max_{\mathcal{X}} \vartheta + \max_{\mathcal{X}} \phi - \min_{U} \vartheta - \min_{U} \phi \\ &= \lambda_U(\vartheta) + \lambda_U(\phi) \leq \lambda_U(\vartheta) + \lambda_{\mathcal{X}}(\phi).\end{aligned}$$

It follows that $\lambda_U(\vartheta) \geq \lambda_U(\vartheta - \phi) - \lambda_{\mathcal{X}}(\phi)$, hence

$$\lambda_U(\vartheta + \phi) \geq \lambda_U(\vartheta) - \lambda_{\mathcal{X}}(\phi).$$

Therefore, vert$\lambda_U(\vartheta + \phi) - \lambda_U(\vartheta)$vert $\leq \lambda_{\mathcal{X}}(\phi)$. As a consequence, $\lambda_U$ is continuous on $\mathcal{B}$.
We will prove that there is a number $h > 0$ such that $\lambda_U(\phi) \geq h\lambda_{\mathcal{X}}(\phi)$ for every $\phi \in \mathcal{B}$. Let $\mathcal{S} = \{\phi \in \mathcal{B} : \min_{\mathcal{X}} \phi = 0 \text{ and } \max_{\mathcal{X}} \phi = 1\}$. Let $\phi \in \mathcal{S}$. If $\lambda_U(\phi) = 0$, then $\phi = 1$ on $U$. Consider $\varphi = 1 - \phi$. Clearly, $\varphi \geq 0$ and $\varphi = 0$ on $U$. It follows that $\varphi = 0$ on $\mathcal{X}$, because $U$ is of uniqueness. Then $\phi \equiv 1$, which contradicts the assumption $\phi \in \mathcal{S}$. Therefore, $\lambda_U(\phi) > 0$. Since $\mathcal{S}$ is compact and $\lambda_U$ is continuous, $h := \min_{\mathcal{S}} \lambda_U > 0$. By (6.6.2) and (6.6.1) we obtain $\lambda_U(\phi) \geq h\lambda_{\mathcal{X}}(\phi)$ for all $\phi \in \mathcal{B}$. □

### 6.6.3 Proof of Lemma 6.4.1

By (6.4.1), each $G \in \mathcal{G}_N$ appears in $\mathcal{G}_{N,c}$ with probability $p_c(G) = e^{\phi_c(G) - \psi(\phi_c)}$. Then,

$$p_{r,s} = \mathbb{P}\left((r,s) \in E\left(\mathbb{G}\right)\right) = \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} \frac{e^{\phi_c(G)}}{\sum_{G \in \mathcal{G}_N} e^{\phi_c(G)}}$$

$$= \frac{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{\phi_c(G)}}{\sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \in E(G)}} e^{\phi_c(G)} + \sum_{\substack{G \in \mathcal{G}_N \\ (r,s) \notin E(G)}} e^{\phi_c(G)}}$$

$$= \frac{\sum_{\substack{G\in\mathcal{G}_N \\ (r,s)\in E(G)}} e^{\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G)}}{\sum_{\substack{G\in\mathcal{G}_N \\ (r,s)\in E(G)}} e^{\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G)} + \sum_{\substack{G\in\mathcal{G}_N \\ (r,s)\notin E(G)}} e^{\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G)}}. \tag{6.6.3}$$

Note that

$$\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G) = c_{r,s}\chi_{r,s}(G) + C(G),$$

where

$$C(G) = \sum_{\substack{(k,l)\in\binom{V}{2} \\ (k,l)\neq(r,s)}} c_{k,l}\chi_{k,l}(G).$$

Therefore

$$e^{\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G)} = e^{c_{r,s}\chi_{r,s}(G)}\ e^{C(G)}.$$

Clearly, $c_{r,s}\chi_{r,s}(G)$ is $c_{r,s}$ if $(r,s)\in E(G)$ and it is 0 if $(r,s)\notin E(G)$. Thus, (6.6.3) equals

$$\frac{e^{c_{r,s}} \sum_{\substack{G\in\mathcal{G}_N \\ (r,s)\in E(G)}} C(G)}{\sum_{\substack{G\in\mathcal{G}_N \\ (r,s)\in E(G)}} e^{C(G)} + e^{c_{r,s}} \sum_{\substack{G\in\mathcal{G}_N \\ (r,s)\notin E(G)}} e^{C(G)}}.$$

Let $S$ be the graph with only one edge $(r,s)$. The map $G \mapsto G \setminus S$ is a bijection between the graphs with the edge $(r,s)$ and graphs without $(r,s)$. In addition, $C(G) = C(G\setminus S)$, and so we get (6.4.2).

### 6.6.4 Proof of Lemma 6.4.2

By (6.4.1), each $G\in\mathcal{G}_N$ appears in $\mathcal{G}_{N,c}$ with probability $p_c(G) = e^{\phi_c(G)-\psi(\phi_c)}$. Then,

$$\mathbb{P}\left((r_1,s_1),(r_2,s_2)\in E\left(\mathbb{G}\right)\right) = \sum_{\substack{G\in\mathcal{G}_N \\ (r_1,s_1),(r_2,s_2)\in E(G)}} \frac{e^{\phi_c(G)}}{\sum_{G\in\mathcal{G}_N} e^{\phi_c(G)}}.$$

As in the proof of Lemma 6.4.1, we observe that

$$\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G) = c_{r_1,s_1}\chi_{r_1,s_1}(G) + c_{r_2,s_2}\chi_{r_2,s_2}(G) + \widetilde{C}(G),$$

where

$$\widetilde{C}(G) = \sum_{\substack{(k,l)\in\binom{V}{2} \\ (k,l)\neq(r_1,s_1) \\ (k,l)\neq(r_2,s_2)}} c_{k,l}\chi_{k,l}(G).$$

Thus,

$$e^{\sum_{(k,l)\in\binom{V}{2}} c_{k,l}\chi_{k,l}(G)} = e^{c_{r_1,s_1}\chi_{r_1,s_1}(G)}\ e^{c_{r_2,s_2}\chi_{r_2,s_2}(G)}\ e^{\widetilde{C}(G)}.$$

Let $S_1$ and $S_2$ be the graphs with only one edge, $(r_1,s_1)$ and $(r_2,s_2)$, respectively. Let

$$\mathcal{G}_{N_{12}} = \{G\in\mathcal{G}_N :\ S_1\subset G, S_2\subset G\},$$

$$\mathcal{G}_{N_{10}} = \{G \in \mathcal{G}_N : \ S_1 \subset G, S_2 \not\subset G\},$$
$$\mathcal{G}_{N_{02}} = \{G \in \mathcal{G}_N : \ S_1 \not\subset G, S_2 \subset G\},$$
$$\mathcal{G}_{N_{00}} = \{G \in \mathcal{G}_N : \ S_1 \not\subset G, S_2 \not\subset G\}.$$

a partition of $\mathcal{G}_N$. We observe that the maps

$$G \mapsto G \setminus S_1, \quad G \mapsto G \setminus S_2, \quad G \mapsto G \setminus (S_1 \cup S_2)$$

are bijections between $\mathcal{G}_{N_{10}}$, $\mathcal{G}_{N_{02}}$, $\mathcal{G}_{N_{12}}$, respectively, and $\mathcal{G}_{N_{00}}$. Also, for every $G \in \mathcal{G}_N$,

$$\widetilde{C}(G) = \widetilde{C}(G \setminus S_1) = \widetilde{C}(G \setminus S_2) = \widetilde{C}(G \setminus (S_1 \cup S_2)).$$

Put differently, $\widetilde{C}(G)$ does not depend on the edges $(r_1, s_1)$ and $(r_2, s_2)$. As in the proof of Lemma 6.4.1, we obtain

$$\begin{aligned}&\mathbb{P}\left((r_1, s_1), (r_2, s_2) \in E(\mathbb{G})\right)\\ &= \frac{e^{c_{r_1,s_1}} e^{c_{r_2,s_2}}}{1 + e^{c_{r_1,s_1}} + e^{c_{r_2,s_2}} + e^{c_{r_1,s_1}} e^{c_{r_2,s_2}}} = p_{r_1,s_1}\ p_{r_2,s_2}.\end{aligned}$$

### 6.6.5 Proof of Lemma 6.5.1

*Proof.* Consider the positive half-cubes $H_1^+, \ldots, H_k^+$. Let

$$\mathcal{B} = \mathrm{Lin}\left\{\prod_{i \in I_q} \mathbb{1}_{H_i^+} : \ I_q \subset \{0, \ldots, k\} \text{ and } |I_q| \le q\right\}.$$

We have $\mathcal{B} = \mathcal{B}_q^k$, because $r_0 = \mathbb{1}_{Q_k}$, $r_i = 2\mathbb{1}_{H_i^+} - \mathbb{1}_{Q_k}$ and by induction it is easy to see that for every $S \subset \{1, \ldots, k\}$ and $|S| < q$, if Walsh function $w_S \in \mathcal{B}$ then their product with Rademacher function $w_S r_i \in \mathcal{B}$, for any $i = 0, \ldots, n$. Note that for any permutation $\sigma$ of $\{1, 2, \ldots, q\}$,

$$\mathbb{1}_{H_{i_1}^+} \mathbb{1}_{H_{i_2}^+} \cdots \mathbb{1}_{H_{i_q}^+} = \mathbb{1}_{H_{i_{\sigma(1)}}^+} \mathbb{1}_{H_{i_{\sigma(2)}}^+} \cdots \mathbb{1}_{H_{i_{\sigma(q)}}^+}.$$

The functions $\mathbb{1}_{Q_k}$ and $\mathbb{1}_{H_{i_1}^+} \cdots \mathbb{1}_{H_{i_q}^+}$, $1 \le i_1 \le \ldots \le i_q \le k$, are linearly independent. Indeed, assume that

$$r := \alpha_0 \mathbb{1}_{Q_k} + \sum_{i_1, \ldots, i_q \in \{1, \ldots, k\}} \alpha_{i_1 \cdots i_q} \mathbb{1}_{H_{i_1}^+} \cdots \mathbb{1}_{H_{i_q}^+} = 0.$$

There are points $x_0 \in \bigcap_{i=1}^k H_i^-$, $x_{i_1} \ldots x_{i_q} \in \bigcap_{l \in \{i_1, \ldots, i_q\}} H_l^- \cap \bigcap_{l \neq i_1, \ldots, i_q} H_l^-$ for each $1 \le i_1 \le i_2 \le \ldots \le i_q \le k$. We obtain $\alpha_0 = r(x_0) = 0$ and $\alpha_{i_1 \cdots i_q} = r(x_{i_1 \cdots i_q}) = 0$ as needed. □

### 6.6.6 Propagation of extrema, relative interior and the criterion of Barndorff-Nielsen

In this section we give auxiliary results, but also explain connections to the criterion of Barndorff-Nielsen. Let $\mathcal{B}$ be an arbitrary linear subspace of $\mathbb{R}^{\mathcal{X}}$. In Corollary 6.6.1 below we adapt the criterion in Theorem 6.2.2 to such $e(\mathcal{B})$. Let $\mathcal{B}'$ be the linear space spanned by $\mathcal{B}$ and $\mathbb{1}$.

**Lemma 6.6.2.** *If $U \subset \mathcal{X}$, then $\phi = \min_{\mathcal{X}} \phi$ on $U$ implies $\phi = \min_{\mathcal{X}} \phi$ on $\mathcal{X}$ for every $\phi \in \mathcal{B}$ if and only if $\phi = \max_{\mathcal{X}} \phi$ on $U$ implies $\phi = \max_{\mathcal{X}} \phi$ on $\mathcal{X}$ for every $\phi \in \mathcal{B}$.*

*Proof.* The property with the minima is equivalent to the one with the maxima because $\mathcal{B}$ is closed upon multiplication by $-1$ and because $\max(-\phi) = -\min\phi$. □

**Definition 6.6.1.** *We say that $U \subset \mathcal{X}$ propagates extrema for $\mathcal{B}$ if $\phi = \inf_{\mathcal{X}}\phi$ on $U$ implies that $\phi = \inf_{\mathcal{X}}\phi$ on $\mathcal{X}$ for every $\phi \in \mathcal{B}$.*

Due to Lemma 6.6.2, the property could be equivalently stated using maxima.

**Lemma 6.6.3.** *A nonempty $U \subset \mathcal{X}$ propagates extrema for $\mathcal{B}$ if and only if $U$ is of uniqueness for $\mathcal{B}'_+$.*

*Proof.* Assume that $U$ is of uniqueness for $\mathcal{B}'_+$. Let $\phi \in \mathcal{B}$ and $\phi = \min_{\mathcal{X}}\phi$ on $U$. Then $\varphi = \phi - \min_{\mathcal{X}}\phi \in \mathcal{B}'_+$ and $\varphi = 0$ on $U$, so $\varphi = 0$ on $\mathcal{X}$ and $\phi = \min_{\mathcal{X}}\phi$ on $\mathcal{X}$. It follows that $U$ propagates extrema for $\mathcal{B}$. Conversely, assume that $U$ propagates extrema for $\mathcal{B}$. Let $\phi \in \mathcal{B}$. Then $\phi = \varphi + c$ for some $\varphi \in \mathcal{B}$ and $c \in \mathbb{R}$. If $\phi \geq 0$ and $\phi = 0$ on $U$, then $\varphi = \min_{\mathcal{X}}\varphi = -c$ on $U$, hence $\varphi = -c$ on $\mathcal{X}$, and so $\phi = 0$ on $\mathcal{X}$. Thus, $U$ is of uniqueness for $\mathcal{B}'_+$. □

Theorem 6.2.2 yields the following.

**Corollary 6.6.1.** *MLE for $e(\mathcal{B})$ and $x_1, ..., x_n \in \mathcal{X}$ exists if and only if $\{x_1, \ldots, x_n\}$ propagates extrema for $\mathcal{B}$.*

*Proof.* The MLE for $e(\mathcal{B})$ and $e(\mathcal{B}')$ must be the same. Indeed, we have $e(\mathcal{B}) = e(\mathcal{B}')$ so the suprema of the likelihood functions are the same, see Section 6.1.1. Of course, if $\phi \in \mathcal{B}$ and $e(\phi)$ is the MLE for $e(\mathcal{B})$ then it is also the MLE for $e(\mathcal{B}')$. Conversely, if $\phi \in \mathcal{B}'$, then $\phi = \varphi + c$ for some $\varphi \in \mathcal{B}$ and $c \in \mathbb{R}$. If $e(\phi)$ is the MLE for $e(\mathcal{X}')$, then $e(\varphi)$ is the MLE for $e(\mathcal{B})$. Considering $\mathcal{B}'$, by Theorem 6.2.2 we see that MLE for $e(\mathcal{B}')$ and $x_1, ..., x_n \in \mathcal{X}$ exists if and only if $\{x_1, \ldots, x_n\}$ is of uniqueness for $\mathcal{B}'_+$, and – by Lemma 6.6.3 – if and only if $\{x_1, \ldots, x_n\}$ propagates extrema for $\mathcal{B}$. □

The next lemma hinges on the trivial observation that if the sample mean equals the minimum, then the sample is constant.

**Lemma 6.6.4.** *$\{x_1, \ldots, x_n\}$ propagates extrema for $\mathcal{B}$ if and only if for every $\phi \in \mathcal{B}$, $\min_{\mathcal{X}}\phi < \max_{\mathcal{X}}\phi$ implies $\min_{\mathcal{X}}\phi < \bar{\phi} < \max_{\mathcal{X}}\phi$.*

*Proof.* Let $\{x_1, \ldots, x_n\}$ propagate extrema for $\mathcal{B}$. If $\min_{\mathcal{X}}\phi = \bar{\phi}$, then $\phi = \min_{\mathcal{X}}\phi$ on $\{x_1, \ldots, x_n\}$, hence $\phi = \min_{\mathcal{X}}\phi$ on $\mathcal{X}$ and so $\min_{\mathcal{X}}\phi = \max_{\mathcal{X}}\phi$. A similar argument works if $\bar{\phi} = \max_{\mathcal{X}}\phi$; see also Lemma 6.6.2. Conversely, if $\{x_1, \ldots, x_n\}$ does not propagate extrema for $\mathcal{B}$ then there is $\phi \in \mathcal{B}$ such that $\phi = \min_{\mathcal{X}}\phi$ on $\{x_1, \ldots, x_n\}$, but $\max_{\mathcal{X}}\phi > \min_{\mathcal{X}}\phi$. Then $\min_{\mathcal{X}}\phi = \bar{\phi} < \max_{\mathcal{X}}\phi$. □

Recall the setting and notation of Section 6.1.2. The following theorem was essentially proved in [11] (Theorem 9.13), except that it was stated for the minimal representation of exponential families. The formulation presented in Theorem 6.6.1 below was given in [110] (Theorem 3.5), which covers the arbitrary canonical representation and does so with a more direct proof. Notably, [110] uses the notion of relative interior of a convex set. Let $C$ be the convex hull of $S$. We say that $t \in \mathbb{R}^d$ is in the relative interior of $C$ if for every $\theta \in \mathbb{R}^d$, $\min_{y\in C}\theta\cdot y < \max_{y\in C}\theta\cdot y$ implies $\min_{y\in C}\theta\cdot y < \theta\cdot t < \max_{y\in C}\theta\cdot y$.

**Theorem 6.6.1.** *[110] (Theorem 3.5.) MLE for $e(\mathcal{B})$ and $x_1, ..., x_n \in \mathcal{X}$ exists, if and only if $\bar{T}$ is in the relative interior of $C$.*

To close the circle of ideas, we give a self-contained proof of Theorem 6.6.1, which may also be used to obtain Theorem 6.2.2 from Theorem 6.6.1.

*Proof of Theorem 6.6.1.* By the discussion in this section we know very well that MLE for $x_1, ..., x_n$ and $e(\mathcal{B})$ exists if and only if for every $\phi \in \mathcal{B}$, $\min_{\mathcal{X}} \phi < \max_{\mathcal{X}} \phi$ implies $\min_{\mathcal{X}} \phi < \bar{\phi} < \max_{\mathcal{X}} \phi$. Recall that $\phi \in \mathcal{B}$ if and only if there is $\theta \in \mathbb{R}^d$ such that $\phi = \theta \cdot T$. Then $\min_{x \in \mathcal{X}} \phi(x) = \min_{y \in \mathcal{S}} \theta \cdot y = \min_{y \in \mathcal{C}} \theta \cdot y$, $\max_{x \in \mathcal{X}} \phi(x) = \max_{y \in \mathcal{C}} \theta \cdot y$, and, of course, $\bar{\phi} = \theta \cdot \bar{T}$. Therefore the existence of MLE for $x_1, ..., x_n$ and $e(\mathcal{B})$ is equivalent to $\bar{T}$ being in the relative interior of $C$. □

For clarity, we recall that we agreed in Example 6.1.2 that the existence of MLE for $x_1, ..., x_n \in \mathcal{X}$ and $e(\mathcal{B})$ is the same as the existence of MLE for $x_1, ..., x_n$ and the exponential family given by the canonical statistics $T$ and (6.1.10), and that it is equivalent to the existence of MLE for the sample $y_1 := T(x_1), ..., y_n = T(x_n) \in \mathbb{R}^d$ and the standard exponential family in (6.1.11). From the above discussion we also see that the convex hull $C$ and the notion of relative interior are merely auxiliary objects to express the property in Lemma 6.6.4, or the propagation of extrema property.

# Chapter 7

# On Laplacian of Graphical Models in Various Graphs

## 7.1 Introduction

The content of this chapter comes from the published article [164] of the author of the dissertation from the conference "Geometric Science of Information 2021" in Sorbonne University, Paris.
Let $G = (V, E, C)$ be a simple undirected graph, where $V = \{1, 2, \ldots, n\}$ is a set of vertices, $E \subset \binom{n}{2}$ is a set of edges and $C \subset V$ is a set of source vertices, which will be called later as a "root set" or a "root". As a degree deg(v) of a vertex $v \in V$ we treat a number of its neighbours. In our convention the graph Laplacian is defined as $\boldsymbol{L}(G) = \{l_{i,j}\}_{1 \leq i,j \leq n}$ with $l_{i,i} = deg(i)$, $l_{i,j} = -1$ if $\{i, j\}$ is an edge in $G$ and zeros otherwise, cf. e.g. [35]. Note that the graph Laplacian is a singular matrix, since its entries in each row (and each column) sum up to 0. Therefore we introduce an augmented graph Laplacian by adding 1 to every entry $l_{c,c}$ corresponding to $c \in C$. In other words,

$$\boldsymbol{L}_C^*(G) := \boldsymbol{L}(G) + \boldsymbol{E}_C$$

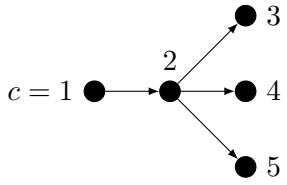
with $\boldsymbol{E}_C$ being a square matrix with 1 in $(c, c)$ for $c \in C$ and zeros everywhere else. We define $\boldsymbol{L}_c^*(G) = \{l_{i,j}^*\}_{1 \leq i,j \leq n}$ with $l_{i,j}^* = l_{i,j} + \mathbb{1}_{\{i=j \in C\}}$.

## 7.2 Trees

Let $T = (V, E)$ be an undirected tree. We may orient it in a following way: Choose one root vertex $C = \{c\}$. Then we orient every edge in a direction from $c$. Following this method we may induce a partial order $\leq$ on the set of vertices such that $v \leq v'$ if and only if there exists a directed path from $v'$ to $v$.
For every vertex $v \in V$ we define its ancestry $\mathrm{AN}(v) := \{w \in V : v \leq w\}$ as a set of vertices in a unique path from $v$ to $c$. Note that both the partial order and AN depend strictly on the choice of $c$.
Now consider an $n$-dimensional Gaussian random variable $(X_1, \ldots, X_n)$ with a covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{i,j}\}_{1 \leq i,j \leq n}$ such that

$$\sigma_{i,j} = |\mathrm{AN}(i) \cap \mathrm{AN}(j)|$$

for every $1 \leq i, j \leq n$.




(a) Underlying graph $G$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 2 & 2 \\ 1 & 2 & 2 & 3 & 2 \\ 1 & 2 & 2 & 2 & 3 \end{bmatrix}$$

(b) Covariance matrix $\boldsymbol{\Sigma}$

$$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & -1 & -1 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

(c) Augmented Laplacian $\boldsymbol{L}_1^*(G)$

**Theorem 7.2.1.** *Let $(X_1, \ldots, X_n)$ be a Gaussian graphical model with an underlying graph $G$ being a tree rooted in $C = \{c\}$. Assume that its covariance matrix $\boldsymbol{\Sigma} = (\sigma_{i,j})$ satisfies*

$$\sigma_{i,j} = |\mathrm{AN}(i) \cap \mathrm{AN}(j)|.$$

*Then the precision matrix $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$ of $(X_1, \ldots, X_n)$ is equal to $\boldsymbol{L}_c^* = \boldsymbol{L}_c^*(G)$.*

*Proof.* It suffices to prove that $L_c^*(G) \cdot \boldsymbol{\Sigma} = \boldsymbol{I}_n$. At first consider the case $i = j = c$. Observe that for any $1 \leq k \leq n$ we have $\sigma_{c,k} = 1$. Therefore:

$$(\boldsymbol{L}_c^* \boldsymbol{\Sigma})_{c,c} = \sum_{k=1}^{n} l_{c,k}^* \sigma_{k,c} = l_{c,c}^* \sigma_{c,j} + \sum_{k \sim c} l_{c,k}^* \sigma_{k,c} = (\deg(c) + 1) + \deg(c) \cdot (-1) \cdot 1 = 1.$$

Now consider $i = j \neq c$:

$$\begin{aligned}(\boldsymbol{L}_c^* \boldsymbol{\Sigma})_{i,i} = \sum_{k=1}^{n} l_{i,k}^* \sigma_{k,i} = l_{i,i}^* \sigma_{i,i} + \sum_{\substack{k \sim i \\ k \in \mathrm{AN}(i)}} l_{i,k}^* \sigma_{k,i} + \sum_{\substack{k \sim i \\ i \in \mathrm{AN}(k)}} l_{i,k}^* \sigma_{k,i} = \\ = \deg(i) \sigma_{i,i} - (\sigma_{i,i} - 1) - (\deg(i) - 1) \sigma_{i,i} = 1.\end{aligned}$$

Now we will prove that the outside of the diagonal of $(\boldsymbol{L}^* \boldsymbol{\Sigma})$ consists only of zeros. Observe that for $i = c, j \neq c$ we have:

$$\begin{aligned}(\boldsymbol{L}_c^* \boldsymbol{\Sigma})_{i,j} = \sum_{k=1}^{n} l_{i,k}^* \sigma_{k,j} = l_{i,i}^* \sigma_{i,j} + \sum_{\substack{k \sim i \\ k \notin \mathrm{AN}(j)}} l_{i,k}^* \sigma_{k,j} + \sum_{\substack{k \sim i \\ k \in \mathrm{AN}(j)}} l_{i,k}^* \sigma_{k,j} = \\ = (\deg(i) + 1) \cdot 1 - (\deg(i) - 1) \cdot 1 - 1 \cdot 2 = 0.\end{aligned}$$

On the other hand, if $j = c, i \neq c$, then:

$$(\boldsymbol{L}_c^* \boldsymbol{\Sigma})_{i,j} = \sum_{k=1}^{n} l_{i,k}^* \sigma_{k,j} = \sum_{k=1}^{n} l_{i,k} \cdot 1 = 0,$$

because each row of the graph Laplacian sums up to 0. Now we let $i, j \neq c, i \neq j$. If $i \in \mathrm{AN}(j)$, then:

$$\begin{aligned}(\boldsymbol{L}_c^* \boldsymbol{\Sigma})_{i,j} = \sum_{k=1}^{n} l_{i,k}^* \sigma_{k,j} = \sum_{k=1}^{n} l_{i,k} \cdot \sigma_{k,j} = \\ = \deg(i) \cdot \sigma_{i,j} - (\deg(i) - 2) \sigma_{i,j} - 1 \cdot (\sigma_{i,j} - 1) - 1 \cdot (\sigma_{i,j} + 1) = 0.\end{aligned}$$

Analogously, if $j \in \mathrm{AN}(i)$, then $\sigma_{k,j} = \sigma_{i,j}$ for every $k \sim i$. Therefore:

$$(\boldsymbol{L}_c^*\boldsymbol{\Sigma})_{i,j} = \sum_{k=1}^{n} l^*_{i,k}\sigma_{k,j} = \sum_{k=1}^{n} l_{i,k} \cdot \sigma_{k,j} = \deg(i) \cdot \sigma_{i,j} - \deg(i) \cdot \sigma_{i,j} = 0.$$

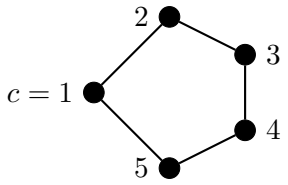
The same argument may be applied in the only case left, when $i \neq j;\ i, j \neq c$, $i \notin \mathrm{AN}(j)$ and $j \notin \mathrm{AN}(i)$. Therefore here also $(\boldsymbol{L}_c^*\boldsymbol{\Sigma})_{i,j} = 0$. □

## 7.3 Discussion and non-tree graphs

### 7.3.1 Non-tree graphs

#### Cycles and complete graphs

The description of the inverse of $\boldsymbol{L}^*(G)$ for general $G$ is much harder for $G$ not being a tree. So far we are not able to present a general formula, thus we consider examples of such inverse for specific classes of graphs $G$. Below we show some examples of $n \times \boldsymbol{\Sigma} = [\boldsymbol{L}_1^*(G)]^{-1}$ for cycles $C_n$ and complete graphs $K_n$ with the root $C = \{1\}$:




(a) Graph $C_5$

$$\begin{bmatrix} 3 & -1 & 0 & 0 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ -1 & 0 & 0 & -1 & 2 \end{bmatrix}$$

(b) $\boldsymbol{L}_1^*(C_5)$

$$\begin{bmatrix} 5 & 5 & 5 & 5 & 5 \\ 5 & 9 & 8 & 7 & 6 \\ 5 & 8 & 11 & 9 & 7 \\ 5 & 7 & 9 & 11 & 8 \\ 5 & 6 & 7 & 8 & 9 \end{bmatrix}$$

(c) $5 \times (\boldsymbol{L}_1^*(C_5))^{-1}$

**Theorem 7.3.1.** *The inverse matrix $\boldsymbol{\Sigma} = (\sigma_{i,j})$ of $\boldsymbol{L}^*(C_n)$ is a symmetric matrix satisfying*

$$\sigma_{i,j} = 1 + \frac{(i-1)(n-j+1)}{n}, \quad \textit{for} \quad i \leq j.$$

*Proof.* Again, we show that $\boldsymbol{L}_1^*\boldsymbol{\Sigma} = \boldsymbol{I}_n$, we assume that the cycle is $1 \to 2 \to 3 \to \ldots \to n \to 1$ and take $C = \{1\}$. Then

$$(\boldsymbol{L}_1^*\boldsymbol{\Sigma})_{c,c} = \sum_{k=1}^{n} l_{c,k}\sigma_{k,c} = \sum_{k=1}^{n} l_{c,k} = 1.$$

If $j \neq c$, then we have

$$\begin{aligned}(\boldsymbol{L}_1^*\boldsymbol{\Sigma})_{c,j} &= \sum_{k=1}^{n} l_{c,k}\sigma_{k,j} = l_{c,c}\sigma_{c,j} + l_{c,2}\sigma_{2,j} + l_{c,n}\sigma_{n,j} = \\ &= 3 \cdot 1 + (-1) \cdot \left(1 + \frac{1 \cdot (n-j+1)}{n}\right) + (-1) \cdot \left(1 + \frac{(j-1) \cdot 1}{n}\right) \\ &= 3 - 1 - \frac{n-j+1}{n} - 1 - \frac{j-1}{n} = 0\end{aligned}$$

and, analogously,

$$(\boldsymbol{L}_1^*\boldsymbol{\Sigma})_{i,c} = \sum_{k=1}^{n} l_{i,k}\sigma_{k,c} = \sum_{k=1}^{n} l_{i,k} = 0.$$

If $i, j \neq c$, then on the main diagonal we have

$$\begin{aligned}(\boldsymbol{L}_1^*\boldsymbol{\Sigma})_{i,i} &= \sum_{k=1}^{n} l_{i,k}\sigma_{k,i} = l_{i,i-1}\sigma_{i-1,i} + l_{i,i}\sigma_{i,i} + l_{i,i+1}\sigma_{i+1,i} = \\ &= -\left(1 + \frac{(i-2)(n-i+1)}{n}\right) + 2\left(1 + \frac{(i-1)(n-i+1)}{n}\right) - \left(1 + \frac{(i-1)(n-i)}{n}\right) = \\ &= \frac{-(i-2)(n-i+1) + 2(i-1)(n-i+1) - (i-1)(n-i)}{n} = \\ &= \frac{(n-i+2) + (i-1)}{n} = 1.\end{aligned}$$

Note that the above calculations are also true for $i = n$ and the $(n+1)^{st}$ row/column of $L$ and $\boldsymbol{\Sigma}$ being treated as $c^{th}$ row/column. Now we only need to consider the outside of the main diagonal. Note that $\sigma_{i,n} = 1 + \frac{i-1}{n}$. Thus

$$(\boldsymbol{L}^*\boldsymbol{\Sigma})_{i,n} = \sum_{k=1}^{n} l_{i,k}\sigma_{k,n} = -\sigma_{i-1,n} + 2\sigma_{i,n} - \sigma_{i+1,n} = 0.$$

Also,

$$\begin{aligned}(\boldsymbol{L}^*\boldsymbol{\Sigma})_{n,j} &= \sum_{k=1}^{n} l_{n,k}\sigma_{k,j} = -\sigma_{n-1,j} + 2\sigma_{n,j} - \sigma_{c,j} = \\ &= -1 - \frac{(j-1)\cdot 2}{n} + 2 + \frac{(j-1)\cdot 1}{n}\cdot 2 - 1 = 0.\end{aligned}$$

Finally, for $i, j \neq 1, n$, $i \neq j$ we have

$$(\boldsymbol{L}^*\boldsymbol{\Sigma})_{i,j} = \sum_{k=1}^{n} l_{i,k}\sigma_{k,j} = -\sigma_{i-1,j} + 2\sigma_{i,j} - \sigma_{i+1,j} = (*)$$

Observe that as $i, j, n$ are pairwise distinct, either $i + 1 \leq j$ or $i - 1 \geq j$. Therefore

$$\begin{cases} i+1 \leq j \Rightarrow (*) = -1 - \frac{(i-2)(n-j+1)}{n} + 2 + 2\cdot\frac{(i-1)(n-j+1)}{n} - 1 - \frac{(i)(n-j+1)}{n} = 0, \\ i-1 \geq j \Rightarrow (*) = -1 - \frac{(j-1)(n-i+2)}{n} + 2 + 2\cdot\frac{(j-1)(n-i+1)}{n} - 1 - \frac{(j-1)(n-i)}{n} = 0, \end{cases}$$

which ends the proof.

□

**Theorem 7.3.2.**

$$[\boldsymbol{L}_1^*(K_n)]_{i,j}^{-1} = \begin{cases} 1 & \text{if } i = 1 \text{ or } j = 1, \\ 1 + \frac{2}{n} & \text{if } 1 < i = j, \\ 1 + \frac{1}{n} & \text{else.} \end{cases}$$

*Proof.* For proving that claim we observe that $\boldsymbol{L}_1^*(K_n) = \boldsymbol{L}(K_n) + \boldsymbol{E}_1 = n\boldsymbol{I}_n - \boldsymbol{J}_n + \boldsymbol{E}_1$. Thus $\boldsymbol{L}_1^*(K_n)$ belongs to an associative algebra being the 5-dimensional matrix space spanned by $\boldsymbol{I}_n, \boldsymbol{E}_1, \boldsymbol{J}_n, \boldsymbol{E}_1\boldsymbol{J}_n$ and $\boldsymbol{J}_n\boldsymbol{E}_1$. Therefore $(\boldsymbol{L}_1^*(K_n))^{-1}$ has to be looked for under the form

$$\frac{1}{n}\boldsymbol{I}_n + a\boldsymbol{E}_1 + b\boldsymbol{J}_n + c\boldsymbol{E}_1\boldsymbol{J}_n + d\boldsymbol{J}_n\boldsymbol{E}_1.$$

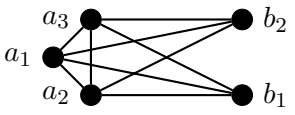
Clearly $(\boldsymbol{L}^*(K_n))^{-1}$ is a symmetric matrix, thus $d = c$. Solving the linear system in $a, b, c$ gives $a = 0, b = (n+1)/n$ and $c = -1/n$. □

### Daisy graphs

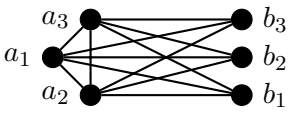
The daisy graphs (cf. e.g. Nakashima and Graczyk [135]) may be understood as a notion between a complete bipartite graph and a complete graph. To be more specific, the daisy graph $D_{a,b}$ is built as a sum of the complete bipartite graph $K_{a,b}$ and a complete subgraph $K_a$, i.e.

$$D_{a,b} := (V = V_A \sqcup V_B, E), \qquad V = V_A \sqcup V_B, \qquad (x,y) \in E \iff \{x,y\} \cap V_A \neq \varnothing.$$

We may interpret $V_A$ and $B_B$ as the internal and the external part of a daisy, respectively. To give an intuition, graphical models based on daisy graphs may be useful for analysis of $a$ internal features of data (without knowledge of any independence among them) and $b$ mutually conditionally independent external factors, which can influence the internal environment. Note that for $b = 1$ we have $D_{a,1}$ being a complete graph $K_{a+1}$ and for $a = 1$ we have $D_{1,a}$ being a star graph, which is a tree (cf. 7.2).




(a) Graph $D_{3,2}$




(b) Graph $D_{3,3}$

The augmented Laplacian (and its inverse) depends on the choice of a root set $C \subset V$. Below we consider four choices of a root in $D_{a,b}$ and their augmented Laplacians:

$$\begin{aligned}
L^*_{in} &:= \boldsymbol{L}(D_{a,b}) + \boldsymbol{E}_c && \text{for } C = \{c\}, \quad c \in V_A,\\
L^{**}_{in} &:= \boldsymbol{L}(D_{a,b}) + \sum_{c \in A} \boldsymbol{E}_c && \text{for } C = V_A,\\
L^*_{ex} &:= \boldsymbol{L}(D_{a,b}) + \boldsymbol{E}_c && \text{for } C = \{c\}, \quad c \in V_B,\\
L^{**}_{ex} &:= \boldsymbol{L}(D_{a,b}) + \sum_{c \in A} \boldsymbol{E}_c && \text{for } C = V_B.
\end{aligned}$$

Without loss of generality we imply such ordering on vertices that the internal vertices precede the external ones. Moreover, concerning cases of one rooted internal (external) vertex we label it as the first (last) vertex.
The exact formulas for $(\boldsymbol{L}^*)^{-1}(D_{a,b})$ are presented below:

**Theorem 4:** Inverses of augmented Laplacians of daisy graphs

$$[\mathrm{L}^{*}_{in}(D_{a,b})]^{-1}_{i,j} = \begin{cases} 1 & \text{if } i=1 \text{ or } j=1, \\ 1+\frac{1}{n} & \text{if } 2 \le i,j \le a, i \ne j, \\ 1+\frac{2}{n} & \text{if } 2 \le i = j \le a, \\ 1+\frac{1}{n} & \text{if } 2 \le i \le a < j, \\ 1+\frac{a-1}{an} & \text{if } a < i,j,\ i \ne j, \\ 1+\frac{a+n-1}{an} & \text{if } a < i = j. \end{cases}$$

$$[\mathrm{L}^{**}_{in}(D_{a,b})]^{-1}_{i,j} = \begin{cases} \frac{n}{a(n+1)} & \text{if } i,j \le a, i \ne j, \\ \frac{n+a}{a(n+1)} & \text{if } i = j \le a, \\ \frac{1}{a} & \text{if } i \le a < j, \\ \frac{1}{a} & \text{if } a < i,j,\ i \ne j, \\ \frac{2}{a} & \text{if } a < i = j. \end{cases}$$

$$[\mathrm{L}^{*}_{ex}(D_{a,b})]^{-1}_{i,j} = \begin{cases} 1 & \text{if } i=n \text{ or } j=n, \\ 1+\frac{1}{a} & \text{if } a < i,j < n, i \ne j, \\ 1+\frac{2}{a} & \text{if } a < i = j < n, \\ 1+\frac{1}{a} & \text{if } i \le a < j < n, \\ 1+\frac{n-1}{an} & \text{if } i,j \le a,\ i \ne j, \\ 1+\frac{a+n-1}{an} & \text{if } i = j \le a. \end{cases}$$

$$[\mathrm{L}^{**}_{ex}(D_{a,b})]^{-1}_{i,j} = \begin{cases} \frac{n+1}{bn} & \text{if } i,j \le a, i \ne j, \\ \frac{n+b+1}{bn} & \text{if } i = j \le a, \\ \frac{1}{b} & \text{if } i \le a < j, \\ \frac{a}{(a+1)b} & \text{if } a < i,j,\ i \ne j, \\ \frac{n}{(a+1)b} & \text{if } a < i = j. \end{cases}$$

*Proof.* Similarly to 7.3.2, the proof relies on the associative algebras being 6-dimensional (for $L^{*}_{ex}$) or 11-dimensional (for $L^{*}_{in}$) matrix spaces. The resulting inverses of augmented Laplacians follow from solving the corresponding systems of equations. □

### 7.3.2 Eigenvalues of augmented Laplacian

Below we show that if the root consists of only one vertex $c$, then the determinant of $\boldsymbol{L}^{*}_{c}(G)$ does not depend on the choice of the root vertex. Moreover, it can be proved that it is equal to the number of spanning trees of $G$.

**Remark 7.3.3.** Let $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ be the eigenvalues of $\boldsymbol{L}^{*}_{c}$. Then

$$\prod_{i=1}^{n} \lambda_i = \det(\boldsymbol{L}^{*}_{c}) = \#\{\text{spanning trees } (G)\}.$$

*Proof.* Recall that $\boldsymbol{L}^{*}_{c}(G)$ differs from $\boldsymbol{L}(G)$ only at $l_{c,c}$, where $c$ is a root vertex of $G$. Let $\boldsymbol{M}$ be a matrix obtained from $\boldsymbol{L}$ by replacing the $c^{th}$ column with a column with 1 at the $c^{th}$ row and zeros elsewhere. Then

$$\det(\boldsymbol{L}^{*}_{c}(G)) = \det(\boldsymbol{L}(G)) + \det(\boldsymbol{M}).$$

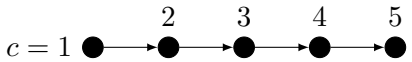
As discussed earlier, the Laplacian matrix $\boldsymbol{L}(G)$ is singular. We can observe that $det(\boldsymbol{M})$ is an $(n-1)\times(n-1)$ cofactor of $\boldsymbol{L}(G)$. By the Kirchhoff's matrix-tree Theorem, any $(n-1)\times(n-1)$ cofactor of $\boldsymbol{L}(G)$ is equal to the number of spanning trees of $G$. □

### 7.3.3 Discussion

#### Interpretations of the inverse of the augmented Laplacian

Below we couple the obtained inverses of $\boldsymbol{L}^{*}$ matrices with the covariance matrices of some classical examples of random walks. At first let us remind that the covariance function of the

Wiener process $(W_t)_{t\geq 0}$ is $\mathrm{Cov}(W_s, W_t) = \min\{s,t\}$. For example, $(W_1, W_2, \ldots, W_n)$ has its covariance matrix equal to $(\mathbf{\Sigma})_{i,j} = \min\{i,j\}$. At Theorem 7.2.1 we proved that this is exactly the inverse of the augmented Laplacian of a path graph with an initial vertex in one of its endpoints. This observation stays consistent with the conditional independence of $W_{t_1}$ and




(a) Path graph $P_n$

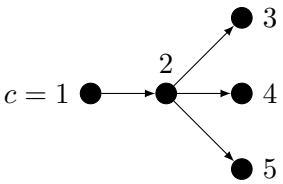
$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 & 4 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

(b) Covariance matrix $\mathbf{\Sigma}$

$$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

(c) Inverse of covariance matrix $\mathbf{\Sigma}^{-1}$

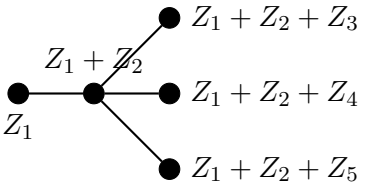
$W_{t_2}$ under $W_t$ for any $t_1 < t < t_2$. Similarly, the broader class of trees and their augmented Laplacians can be connected with sums of the standard Gaussian random variables 'branched' according to the underlying tree graph.




(a) Underlying graph $G$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 3 & 2 & 2 \\ 1 & 2 & 2 & 3 & 2 \\ 1 & 2 & 2 & 2 & 3 \end{bmatrix}$$

(b) Covariance matrix $\mathbf{\Sigma}$




(c) Gaussian random variables with covariance $\mathbf{\Sigma}$

To observe the analogy of the previous examples for the cycle graph $C_n$, we may note that replacing the edge $(n,1)$ with $(n, n+1)$ (with $n+1 \notin V$) gives a path graph

$$[c=1] - [2] - \ldots - [n+1]).$$

Therefore we may consider the model corresponding to the cycle graph as the Wiener model $(W_1, \ldots, W_n, W_{n+1})$ conditioned by $W_{n+1} = W_1$. This gives a sum of a random variable $W_1 \sim \mathcal{N}(0,1)$ and a Brownian bridge "tied down" at 1 and $(n+1)$. Therefore the covariance matrix of $(W_1, \ldots, W_n)$ is equal to

$$\sigma_{i,j} = 1 + \frac{(i-1)(n-j+1)}{n}, \quad i \leq j,$$

cf. Theorem 7.3.1.
In order to find a model with a covariance matrix equal to

$$[\boldsymbol{L}_1^*(K_n)]_{i,j}^{-1} = \begin{cases} 1 & \text{if } i = 1 \text{ or } j = 1, \\ 1 + \frac{2}{n} & \text{if } 1 < i = j, \\ 1 + \frac{1}{n} & \text{else,} \end{cases}$$

observe that all vertices (except of the initial one) are isomorphic and connected, therefore each of their correspondent random variables are mutually equally dependent Let $c = 1$ and let $Z_1, \ldots, Z_n$ be i.i.d. random variables from $\mathcal{N}(0, 1)$. Therefore $X_1, X_2, \ldots, X_n$ are of the form

$$\begin{cases} X_1 = Z_1 \\ X_2 = Z_1 + \alpha Z_2 + \beta(Z_3 + Z_4 + \ldots + Z_n) \\ X_3 = Z_1 + \alpha Z_3 + \beta(Z_2 + Z_4 + \ldots + Z_n) \\ \ldots \\ X_n = Z_1 + \alpha Z_n + \beta(Z_2 + Z_3 + \ldots + Z_{n-1}). \end{cases}$$

The restriction on the covariance matrix of $(X_1, \ldots, X_n)$ induces a system of equations, which is satisfied only for $\beta = \frac{1-\sqrt{\frac{1}{n}}}{n-1}$ and $\alpha = \beta + \sqrt{\frac{1}{n}}$.

# Bibliography


[1] A. Agresti. Foundations of linear and generalized linear models. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2015.

[2] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[3] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71(1):1–10, 1984.

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

[5] E. B. Andersen. Sufficiency and exponential families for discrete sample spaces. J. Amer. Statist. Assoc., 65:1248–1255, 1970.

[6] T. W. Anderson and J. B. Taylor. Strong consistency of least squares estimates in normal linear regression. The Annals of Statistics, 4(4):788–790, 1976.

[7] J.-P. Aubin. Mathematical methods of game and economic theory, volume 7 of Studies in Mathematics and its Applications. North-Holland Publishing Co., Amsterdam-New York, 1979.

[8] R. Bao, B. Gu, and H. Huang. Fast oscar and owl with safe screening rules. In International Conference on Machine Learning, 2020.

[9] A. Barbara, A. Jourani, and S. Vaiter. Maximal solutions of sparse analysis regularization. Journal of Optimization Theory and Applications, 180(2):374–396, 2019.

[10] É. Barbin and J. P. Lamarche. Histoires de probabilités et de statistiques. Ellipses Paris, 2004.

[11] O. Barndorff-Nielsen. Information and exponential families in statistical theory. John Wiley & Sons Ltd., Chichester, 1978. Wiley Series in Probability and Mathematical Statistics.

[12] E. M. L. Beale, M. G. Kendall, and D. W. Mann. The discarding of variables in multivariate analysis. Biometrika, 54:357–366, 1967.

[13] P. Bellec and A. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In Modern problems of stochastic analysis and statistics, volume 208 of Springer Proc. Math. Stat., pages 315–333. Springer, Cham, 2017.

[14] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets Lasso: improved oracle bounds and optimality. Ann. Statist., 46(6B):3603–3642, 2018.

[15] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root Lasso in nonparametric regression. Ann. Statist., 42(2):757–788, 2014.

[16] A. Ben-Israel and T. N. E. Greville. Generalized inverses, volume 15 of CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer-Verlag, New York, second edition, 2003. Theory and applications.

[17] D. P. Bertsekas. Nonlinear programming. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, second edition, 1999.

[18] I. Bezáková, A. Kalai, and R. Santhanam. Graph model selection using maximum likelihood. In Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pages 105–112, New York, NY, USA, 2006. ACM.

[19] P. Billingsley. Probability and measure. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[20] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, New York, 2007.

[21] K. Bogdan and M. Bogdan. On existence of maximum likelihood estimators in exponential families. Statistics, 34(2):137–149, 2000.

[22] K. Bogdan, M. Bosy, and T. Skalski. Maximum likelihood estimation for discrete exponential families and random graphs. ALEA Lat. Am. J. Probab. Math. Stat., 19(1):1045–1070, 2022.

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

[24] M. Bogdan, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, and M. Wilczyński. Pattern recovery by slope. arXiv preprint arXiv:2203.12086, 2022.

[25] M. Bogdan and T. Ledwina. Testing uniformity via log-spline modeling. Statistics, 28(2):131–157, 1996.

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

[27] M. Bogdan, E. van den Berg, W. Su, and E. Candès. Statistical estimation and testing via the sorted l1 norm, 2013.

[28] B. Bollobás. Modern graph theory, volume 184 of Graduate Texts in Mathematics. Springer-Verlag, New York, 1998.

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

[30] H. D. Bondell and B. J. Reich. Simultaneous factor selection and collapsing levels in anova. Biometrics, 65:169–177, 2009.

[31] J. M. Borwein and A. S. Lewis. Convex Analysis and Nonlinear Optimization, Theory and Examples. Springer, 2000.

[32] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. ESAIM Probab. Stat., 9:323–375, 2005.

[33] S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, USA, 2004.

[34] A. Brøndsted. An introduction to convex polytopes, volume 90 of Graduate Texts in Mathematics. Springer-Verlag, New York-Berlin, 1983.

[35] A. E. Brouwer and W. H. Haemers. Spectra of graphs. Universitext. Springer, New York, 2012.

[36] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory, volume 9 of Institute of Mathematical Statistics Lecture Notes—Monograph Series. Institute of Mathematical Statistics, Hayward, CA, 1986.

[37] D. Brzyski. Selecting relevant groups of explanatory variables via convex optimization methods with the false discovery rate control. PhD Thesis, 2015.

[38] D. Brzyski, A. Gossmann, W. Su, and M. Bogdan. Group SLOPE—adaptive selection of groups of predictors. J. Amer. Statist. Assoc., 114(525):419–433, 2019.

[39] D. Brzyski, M. Karas, B. M. Ances, M. Dzemidzic, J. Goñi, T. W. Randolph, and J. Harezlak. Connectivity-informed adaptive regularization for generalized outcomes. Canad. J. Statist., 49(1):203–227, 2021.

[40] Z. Bu, J. M. Klusowski, C. Rush, and W. J. Su. Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. IEEE Trans. Inform. Theory, 67(1):506–537, 2021.

[41] P. Bühlmann and S. van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Heidelberg, 2011.

[42] E. J. Candès and Y. Plan. Near-ideal model selection by $\ell_1$ minimization. The Annals of Statistics, 37(5A):2145–2177, 2009.

[43] cardinal (https://math.stackexchange.com/users/7003/cardinal). Proof of upper-tail inequality for standard normal distribution. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/28754 (version: 2011-03-24).

[44] A. Chatterjee and S. N. Lahiri. Strong consistency of Lasso estimators. Sankhya A, 73(1):55–78, 2011.

[45] S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. Ann. Statist., 41(5):2428–2461, 2013.

[46] S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. Ann. Appl. Probab., 21(4):1400–1435, 2011.

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

[48] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput., 20(1):33–61, 1998.

[49] G. Claeskens. Statistical model choice. In Annual Review of Statistics and Its Application, volume 3, pages 233–256, 2016.

[50] B. R. Crain. Estimation of distributions using orthogonal expansions. Ann. Statist., 2:454–463, 1974. Collection of articles dedicated to Jerzy Neyman on his 80th birthday.

[51] B. R. Crain. Exponential models, maximum likelihood estimation, and the Haar condition. J. Amer. Statist. Assoc., 71(355):737–740, 1976.

[52] J. Cuzick. A strong law for weighted sums of i.i.d. random variables. J. Theoret. Probab., 8(3):625–641, 1995.

[53] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. Ann. Statist., 8(3):522–539, 1980.

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

[55] P. Descloux and S. Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. Journal of Computational and Graphical Statistics, 30(3):530–543, 2021.

[56] N. Dexheimer and C. Strauch. On lasso and slope drift estimators for lévy-driven ornstein–uhlenbeck processes. 2022.

[57] P. Diaconis. Group representations in probability and statistics, volume 11 of Institute of Mathematical Statistics Lecture Notes—Monograph Series. Institute of Mathematical Statistics, Hayward, CA, 1988.

[58] P. Diaconis and D. Freedman. Partial exchangeability and sufficiency. In Statistics: applications and new directions (Calcutta, 1981), pages 205–236. Indian Statist. Inst., Calcutta, 1984.

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

[60] D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906):4273–4293, 2009.

[61] X. Dupuis and P. J. C. Tardivel. Proximal operator for the sorted $\ell_1$ norm: Application to testing procedures based on SLOPE. Journal of Statistical Planning and Inference, 221:1–8, 2022.

[62] X. Dupuis and S. Vaiter. The geometry of sparse analysis regularization. arXiv preprint arXiv:1907.01769, 2019.

[63] J. Dziubański and A. Hejna. Remark on atomic decompositions for the Hardy space $H^1$ in the rational Dunkl setting. Studia Math., 251(1):89–110, 2020.

[64] B. Eisenberg. On the expectation of the maximum of IID geometric random variables. Statist. Probab. Lett., 78(2):135–143, 2008.

[65] C. Elvira and C. Herzet. Safe rules for the identification of zeros in the solutions of the slope problem. ArXiv, abs/2110.11784, 2021.

[66] P. Erdős and A. Rényi. On random graphs. I. Publ. Math. Debrecen, 6:290–297, 1959.

[67] P. Erdős and A. Rényi. On the evolution of random graphs. Magyar Tud. Akad. Mat. Kutató Int. Közl., 5:17–61, 1960.

[68] P. Erdős and A. Rényi. On a classical problem of probability theory. Magyar Tud. Akad. Mat. Kutató Int. Közl., 6:215–220, 1961.

[69] N. Eriksson, S. E. Fienberg, A. Rinaldo, and S. Sullivant. Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. J. Symbolic Comput., 41(2):222–233, 2006.

[70] G. Ewald. Combinatorial convexity and algebraic geometry, volume 168. Springer Science & Business Media, 1996.

[71] K. Ewald and U. Schneider. Uniformly valid confidence sets based on the Lasso. Electron. J. Stat., 12(1):1358–1387, 2018.

[72] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96:1348–1360, 2001.

[73] W. Feller. An introduction to probability theory and its applications. Vol. II. John Wiley & Sons, Inc., New York-London-Sydney, second edition, 1971.

[74] S. E. Fienberg and A. Rinaldo. Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. J. Statist. Plann. Inference, 137(11):3430–3445, 2007.

[75] S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. Ann. Statist., 40(2):996–1023, 2012.

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

[77] R. A. Fisher. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2):179–188, 1936.

[78] L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga. A critical review of LASSO and its derivatives for variable selection under dependence among covariates. Int. Stat. Rev., 90(1):118–145, 2022.

[79] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33, 2010.

[80] A. Frieze and M. Karoński. Introduction to random graphs. Cambridge University Press, Cambridge, 2016.

[81] F. Frommlet, M. g. Bogdan, and D. Ramsey. Phenotypes and genotypes, volume 18 of Computational Biology. Springer-Verlag, London, 2016. The search for influential genes.

[82] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. IEEE Trans. Inform. Theory, 50(6):1341–1344, 2004.

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

[84] D. Galvin. Three tutorial lectures on entropy and counting, 2014.

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

[86] C. J. Geyer. Likelihood and exponential families. ProQuest LLC, Ann Arbor, MI, 1990. Thesis (Ph.D.)–University of Washington.

[87] E. N. Gilbert. Random graphs. Ann. Math. Statist., 30:1141–1144, 1959.

[88] J. C. Gilbert. On the solution uniqueness characterization in the l1 norm and polyhedral gauge recovery. Journal of Optimization Theory and Applications, 172(1):70–101, 2017.

[89] C. Giraud. Introduction to high-dimensional statistics, volume 139 of Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015.

[90] T. Godland and Z. Kabluchko. Projections and angle sums of permutohedra, 2020.

[91] G. H. Golub and C. F. V. Loan. Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2013.

[92] P. Graczyk, T. Luks, and P. Sawyer. Potential kernels for radial Dunkl Laplacians. Canad. J. Math., 74(4):1005–1033, 2022.

[93] P. Graczyk, U. Schneider, T. Skalski, and P. Tardivel. Pattern recovery in penalized and thresholded estimation and its geometry. HAL preprint hal-03262087, 2021.

[94] P. Gruber. Convex and Discrete Geometry. Springer, Heidelberg, 2007.

[95] B. Grünbaum. Convex polytopes, volume 221 of Graduate Texts in Mathematics. Springer-Verlag, New York, second edition, 2003. Prepared and with a preface by Volker Kaibel, Victor Klee and Günter M. Ziegler.

[96] S. J. Haberman. The analysis of frequency data. The University of Chicago Press, Chicago, Ill.-London, 1974. Statistical Research Monographs, Vol. IV.

[97] A. Hald. A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935. Sources and Studies in the History of Mathematics and Physical Sciences. Springer, New York, 2007.

[98] G. H. Hardy, J. E. Littlewood, and G. Pólya. Inequalities. Cambridge, at the University Press, 1952. 2d ed.

[99] H. L. Harter. Expected values of normal order statistics. Biometrika, 48(1/2):151–165, 1961.

[100] S. Helgason. Differential geometry, Lie groups, and symmetric spaces, volume 34 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001. Corrected reprint of the 1978 original.

[101] J.-B. Hiriart-Urruty and C. Lemaréchal. Convex Analysis and Minimization Algorithms I: Fundamentals, volume 305. Springer, Heidelberg, 1993.

[102] R. R. Hocking and R. N. Leslie. Selection of the best subset in regression analysis. Technometrics, 9:531–540, 1967.

[103] H. Hu and Y. M. Lu. SLOPE for sparse linear regression: asymptotics and optimal regularization. IEEE Trans. Inform. Theory, 68(11):7627–7664, 2022.

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

[105] M. Jacobsen. Existence and unicity of MLEs in discrete exponential family distributions. Scand. J. Statist., 16(4):335–349, 1989.

[106] A. Jedrzejewski, J. Lago, G. Marcjasz, and R. Weron. Electricity price forecasting: The dawn of machine learning. IEEE Power and Energy Magazine, 20(3):24–31, 2022.

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

[108] J. Jendrej, K. Oleszkiewicz, and J. O. Wojtaszczyk. On some extensions of the FKN theorem. Theory Comput., 11:445–469, 2015.

[109] W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, V. Rockova, and T. Group. Adaptive bayesian slope: Model selection with incomplete data. Journal of Computational and Graphical Statistics, 31(1):113–137, 2022.

[110] S. Johansen. Introduction to the theory of regular exponential families, volume 3 of Lecture Notes. University of Copenhagen, Institute of Mathematical Statistics, Copenhagen, 1979.

[111] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. $\ell_1$ trend filtering. SIAM review, 51(2):339–360, 2009.

[112] K. Knight and W. Fu. Asymptotics of Lasso-type estimators. Annals of Statistics, 28:1356–1378, 2000.

[113] M.-H. Ko, D.-H. Ryu, T.-S. Kim, and Y.-K. Choi. A central limit theorem for general weighted sums of LNQD random variables and its application. Rocky Mountain J. Math., 37(1):259–268, 2007.

[114] K. Konis. Linear programming algorithms for detecting separated data in binary logistic regression models. PhD thesis, Worcester College, University of Oxford, 2007.

[115] M. Kos and M. Bogdan. On the asymptotic properties of SLOPE. Sankhya A, 82(2):499–532, 2020.

[116] P. J. Kremer, D. Brzyski, M. g. Bogdan, and S. Paterlini. Sparse index clones via the sorted $\ell_1$-norm. Quant. Finance, 22(2):349–366, 2022.

[117] T. L. Lai and 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):320–335, 1982.

[118] P. Laplace. Théorie analytique des probabilités. Courcier, Paris, 1812.

[119] J. Larsson, M. Bogdan, and J. Wallin. The strong screening rule for slope. Proc. 34th Conf. NeurIPS 2020, Vancouver, Canada, pages 1–12, 2020.

[120] J. Larsson, Q. Klopfenstein, M. Massias, and J. Wallin. Coordinate descent for slope. arXiv preprint, arXiv:2210.14780, pages 1–12, 2022.

[121] S. L. Lauritzen. Extreme point models in statistics. Scand. J. Statist., 11(2):65–91, 1984.

[122] S. L. Lauritzen. Graphical models, volume 17 of Oxford Statistical Science Series. The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

[123] S. Lee, P. Sobczyk, and M. Bogdan. Structure learning of gaussian markov random fields with false discovery rate control. Symmetry, 11(10), 2019.

[124] A. Legendre. Nouvelles méthodes pour la détermination des orbites des comètes. Nineteenth Century Collections Online (NCCO): Science, Technology, and Medicine: 1780-1925. F. Didot, 1805.

[125] E. L. Lehmann and G. Casella. Theory of point estimation. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

[126] G. Letac. Lectures on natural exponential families and their variance functions, volume 50 of Monografías de Matemática [Mathematical Monographs]. Instituto de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 1992.

[127] L. Lovász and B. Szegedy. Limits of dense graph sequences. J. Combin. Theory Ser. B, 96(6):933–957, 2006.

[128] A. Maj-Kańska, P. Pokarowski, and A. Prochenka. Delete or merge regressors for linear model selection. Electron. J. Stat., 9(2):1749–1778, 2015.

[129] C. Mazza-Anthony, B. Mazoure, and M. Coates. Learning Gaussian graphical models with ordered weighted $\ell_1$ regularization. IEEE Trans. Signal Process., 69:489–499, 2021.

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

[131] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. The annals of statistics, 37(1):246–270, 2009.

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

[133] S. Mousavi and J. Shen. Solution uniqueness of convex piecewise affine functions based optimization with applications to constrained $\ell_1$ minimization. ESAIM: Control, Optimisation and Calculus of Variations, 25:56, 2019.

[134] R. Mukherjee, S. Mukherjee, and S. Sen. Detection thresholds for the $\beta$-model on sparse graphs. Ann. Statist., 46(3):1288–1317, 2018.

[135] H. Nakashima and P. Graczyk. Wigner and wishart ensembles for graphical models, 2020.

[136] R. Negrinho and A. Martins. Orbit regularization. Advances in neural information processing systems, 27, 2014.

[137] S. Nowakowski, P. Pokarowski, and W. Rejchel. Group lasso merger for sparse prediction with high-dimensional categorical data, 2021.

[138] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2023. Published electronically at `http://oeis.org`.

[139] M.-R. Oelker, J. Gertheiss, and G. Tutz. Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. Stat. Model., 14(2):157–177, 2014.

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

[141] P. O. Perry and P. J. Wolfe. Null models for network data, 2012.

[142] E. J. G. Pitman. Some basic theory for statistical inference. Chapman and Hall, London; A Halsted Press Book, John Wiley & Sons, New York, 1979. Monographs on Applied Probability and Statistics.

[143] A. Pósfai. Approximation theorems related to the coupon collector's problem. PhD Thesis, 6, 2010.

[144] A. Postnikov. Permutohedra, associahedra, and beyond. Int. Math. Res. Not. IMRN, (6):1026–1106, 2009.

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

[146] K. R. Rao, N. Ahmed, and M. A. Narasimhan. Orthogonal transforms for digital signal processing. In Proceedings of the Eighteenth Midwest Symposium on Circuits and Systems (Concordia Univ., Montreal, Que., 1975), pages 1–6. Western Periodicals, North Hollywood, Calif., 1975.

[147] A. Rencher. Methods of Multivariate Analysis. Wiley Series in Probability and Statistics. Wiley, 2003.

[148] R. Riccobello, M. Bogdan, G. Bonaccolto, P. Kremer, S. Paterlini, and P. Sobczyk. Graphical modelling via the sorted l1-norm. arXiv preprint, arXiv:2204.10403, 2022.

[149] A. Rinaldo. Properties and refinements of the fused lasso. Ann. Statist., 37(5B):2922–2952, 2009.

[150] A. Rinaldo, S. E. Fienberg, and Y. Zhou. On the geometry of discrete exponential families with application to exponential random graph models. Electron. J. Stat., 3:446–484, 2009.

[151] A. Rinaldo, S. Petrović, and S. E. Fienberg. Maximum likelihood estimation in the $\beta$-model. Ann. Statist., 41(3):1085–1110, 2013.

[152] R. Rockafellar. Convex Analysis. Princeton University Press, 1997.

[153] V. Saligrama and M. Zhao. Thresholded basis pursuit: Lp algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements. IEEE Transactions on Information Theory, 57(3):1567–1586, 2011.

[154] R. Sankaran, F. Bach, and C. Bhattacharya. Identifying Groups of Strongly Correlated Variables through Smoothed Ordered Weighted $L_1$-norms. In A. Singh and J. Zhu, editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pages 1123–1131. PMLR, 20–22 Apr 2017.

[155] R. Schneider. Convex bodies: the Brunn-Minkowski theory, volume 151 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, expanded edition, 2014.

[156] U. Schneider and P. Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. Journal of Machine Learning Research, 23(331):1–36, 2022.

[157] A. Schrijver. Theory of linear and integer programming. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Ltd., Chichester, 1986. A Wiley-Interscience Publication.

[158] G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464, 1978.

[159] M. Schweinberger, P. N. Krivitsky, C. T. Butts, and J. Stewart. Exponential-family models of random graphs: Inference in finite-, super-, and infinite population scenarios, 2017.

[160] A. Sepehri and N. Harris. The accessible lasso models. Statistics, 51(4):711–721, 2017.

[161] A. A. Sepehri. The bayesian slope. arXiv: Methodology, 2016.

[162] D. B. Sharma, H. D. Bondell, and H. H. Zhang. Consistent group identification and variable selection in regression with correlated predictors. Journal of Computational and Graphical Statistics, 22(2):319–340, 2013.

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

[164] T. Skalski. Remarks on Laplacian of graphical models in various graphs. In Geometric science of information, volume 12829 of Lecture Notes in Comput. Sci., pages 685–692. Springer, Cham, [2021] ©2021.

[165] T. Skalski, P. Graczyk, B. Kołodziejek, and M. Wilczyński. Pattern recovery and signal denoising by slope when the design matrix is orthogonal. Probability and Mathematical Statistics, 42(2):283–302, 2022.

[166] P. Sobczyk. Identifying low-dimensional structures through model selection in high-dimensional data. PhD Thesis, 2019.

[167] U. Stadtmüller. A note on the law of iterated logarithm for weighted sums of random variables. Ann. Probab., 12(1):35–44, 1984.

[168] B. G. Stokell, R. D. Shah, and R. J. Tibshirani. Modelling high-dimensional categorical data using nonconvex fusion penalties. J. R. Stat. Soc. Ser. B. Stat. Methodol., 83(3):579–611, 2021.

[169] C. J. Stone. Large-sample inference for log-spline models. Ann. Statist., 18(2):717–741, 1990.

[170] B. Stucky and S. van de Geer. Sharp oracle inequalities for square root regularization. J. Mach. Learn. Res., 18:Paper No. 67, 29, 2017.

[171] W. Su and E. Candès. Slope is adaptive to unknown sparsity and asymptotically minimax. Annals of Statistics, 44:1038–1068, 2016.

[172] S. Sullivant. Algebraic statistics, volume 194 of Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2018.

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

[174] P. J. Tardivel and M. Bogdan. On the sign recovery by least absolute shrinkage and selection operator, thresholded least absolute shrinkage and selection operator, and thresholded basis pursuit denoising. Scandinavian Journal of Statistics, 49:1636–1668, 2022.

[175] P. J. Tardivel, R. Servien, and D. Concordet. Simple expressions of the lasso and slope estimators in low-dimension. Statistics, 54(2):340–352, 2020.

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

[177] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108, 2005.

[178] R. J. Tibshirani. The lasso problem and uniqueness. Electronic Journal of statistics, 7:1456–1490, 2013.

[179] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. Annals of Statistics, 40:1198–1232, 2012.

[180] B. Uniejewski, J. Nowotarski, and R. Weron. Automated variable selection and shrinkage for day-ahead electricity price forecasting. Energies, 9(8), 2016.

[181] S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal. The degrees of freedom of partly smooth regularizers. Annals of the Institute of Statistical Mathematics, 69(4):791–832, 2017.

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

[183] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). IEEE Transactions on Information Theory, 55:2183–2202, 2009.

[184] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1-2):1–305, 2008.

[185] N. Wang, J. Rauh, and H. Massam. Approximating faces of marginal polytopes in discrete hierarchical models. Ann. Statist., 47(3):1203–1233, 2019.

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

[187] J. Whittaker. Graphical Models in Applied Multivariate Statistics. Wiley, 2009.

[188] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 94:19–35, 2007.

[189] X. Zeng and M. Figueiredo. Decreasing weighted sorted $\ell_1$ regularization. IEEE Signal Processing Letters, 21:1240–1244, 2014.

[190] X. Zeng and M. A. Figueiredo. Solving oscar regularization problems by fast approximate proximal splitting algorithms. Digital Signal Processing, 31:124–135, 2014.

[191] Y. Zhang and Z. Bu. Efficient designs of slope penalty sequences in finite dimension. In A. Banerjee and K. Fukumizu, editors, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 3277–3285. PMLR, 13–15 Apr 2021.

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

[193] G. Ziegler. Lectures on Polytopes, volume 152. Springer, New York, 2012.

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