



LUND UNIVERSITY

Discovering Structural Patterns in Statistical Models via Regularization Asymptotics, Exact Recovery and Estimation

Hejny, Ivan

2026

Document Version:

Publisher's PDF, also known as Version of record

[Link to publication](#)

Citation for published version (APA):

Hejny, I. (2026). *Discovering Structural Patterns in Statistical Models via Regularization: Asymptotics, Exact Recovery and Estimation*. [Doctoral Thesis (compilation), Lund University School of Economics and Management, LUSEM]. Lund University (Media-Tryck).

Total number of authors:

1

General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

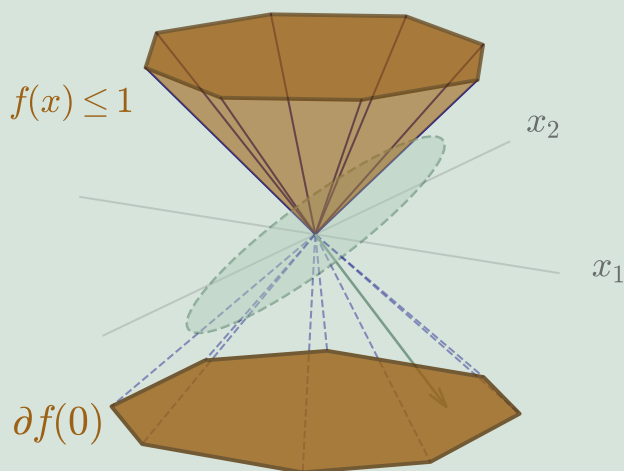
PO Box 117
221 00 Lund
+46 46-222 00 00

Discovering Structural Patterns in Statistical Models via Regularization

Asymptotics, Exact Recovery and Estimation

IVAN HEJNÝ

SCHOOL OF ECONOMICS AND MANAGEMENT | LUND UNIVERSITY



Discovering Structural Patterns in Statistical Models via Regularization

Discovering Structural Patterns in Statistical Models via Regularization

Asymptotics, Exact Recovery and Estimation

Ivan Hejný



LUND
UNIVERSITY

DOCTORAL DISSERTATION

By due permission of the Department of Statistics,
School of Economics and Management, Lund University, Sweden.
To be defended in EC3:207, Holger Crafoord Centre, Lund
on June 9, 2026, at 13:15.

Faculty opponent

Professor Piotr Zwiernik (Universitat Pompeu Fabra, Barcelona, Spain)

Organization LUND UNIVERSITY Department of Statistics Box 7080 SE-220 07 Lund Sweden		Document name DOCTORAL DISSERTATION	
		Date of disputation 2026-06-09	
		Sponsoring organization	
Author(s) Ivan Hejny			
Title and subtitle Discovering Structural Patterns in Statistical Models via Regularization: Asymptotics, Exact Recovery and Estimation			
Abstract In the linear regression model, when the dimension p is fixed and $n \rightarrow \infty$, the asymptotic distribution of the estimation error for Lasso-type M-estimators is well established. In contrast, the convergence of the discrete structures (patterns) induced by these regularizers, such as sparsity or clustering, has remained largely unaddressed, even for the standard Lasso penalty. Because these lower-dimensional structures are sensitive to infinitesimal perturbations, the weak convergence of the continuous estimation error does not guarantee the convergence of the induced patterns. This thesis develops a unified theoretical framework to resolve this discrepancy. We establish that for a range of non-differentiable regularizers with a polyhedral component (including the Lasso, Generalized Lasso, SLOPE, and Elastic Net), the estimated patterns converge toward a limiting pattern determined by an explicit asymptotic formula. This limit is characterized by the Fisher Information matrix of the loss, the score covariance matrix, and the directional derivative of the regularizer. In Paper I, this is achieved in the linear model by utilizing the Hausdorff distance as a suitable mode of set convergence involving subdifferentials. In Paper II, we extend the theory from Paper I to general statistical models via a Stochastic Lipschitz Differentiability (SLD) condition, which controls the fluctuations of the Taylor remainder and, beyond differentiable losses, encompasses robust, non-smooth functions such as the Huber and quantile losses. The asymptotic formula characterizes the specific asymptotic irrepresentability conditions under which the true signal pattern can be recovered with high probability. To bypass instances where these stringent conditions fail (typically in highly correlated predictor structures), we develop adaptive two-step procedures based on proximal operators. Furthermore, the theory identifies a critical degeneracy in the Fused Lasso regarding its inability to recover its own clusters under orthogonal design, motivating the proposal of a Concavified Fused Lasso penalty that resolves this limitation. Papers III and IV apply this asymptotic theory to graphical models and precision matrix estimation. Paper III establishes exact asymptotic limits for the Graphical SLOPE estimator under elliptically distributed data, revealing how it can outperform the Graphical Lasso when clustering structures are present. Paper IV focuses on the scale-invariant PCGLASSO method; we derive an irrepresentability condition under which the true sparsity structure can be recovered, theoretically justifying the method's empirical performance in discovering network hubs. Paper V focuses on the high-dimensional regime ($p \gg n$). By introducing a <i>Surrogate regularizer</i>, a decomposable penalty that locally matches the original polyhedral regularizer around the true signal, we extend established high-dimensional estimation bounds to non-decomposable penalties. Using this surrogate construction, we derive explicit deterministic error bounds and exact pattern recovery conditions. While our primary application focuses on SLOPE, the surrogate approach encompasses penalties such as the Generalized Lasso. Additionally, by linking the linear SLOPE sequence to the Integrated Brownian Bridge, we show that the volume ratio between the dual SLOPE and dual Lasso balls decays at an exact rate of $p^{-1/4}$ for the linear sequence, and we conjecture a $p^{-1/6}$ Gaussian volume ratio decay rate for the Benjamini-Hochberg sequence.			
Key words regularization, pattern convergence, exact recovery, irrepresentability condition, high-dimensional statistics, graphical models, non-decomposable penalization, Lasso, Fused Lasso, SLOPE			
Classification system and/or index terms (if any)			
Supplementary bibliographical information		Language English	
ISSN and key title		ISBN 978-91-90202-00-5 (print) 978-91-90202-01-2 (pdf)	
Recipient's notes		Number of pages 296	Price
		Security classification	

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature _____

Date 22-04-2026

Discovering Structural Patterns in Statistical Models via Regularization

Asymptotics, Exact Recovery and Estimation

Ivan Hejný



LUND
UNIVERSITY

Funding information: This research has been supported by the Swedish Research Council, grant no. 2020-05081.

Cover illustration front: Geometric representation of the primal and dual unit balls of the SLOPE norm.

Introduction: © 2026 Ivan Hejný

Paper I: © 2025 Ivan Hejný, Jonas Wallin, Małgorzata Bogdan, Michał Kos

Paper II: © 2025 Ivan Hejný, Jonas Wallin, Małgorzata Bogdan

Paper III: © 2026 Ivan Hejný, Giovanni Bonaccolto, Philipp Kremer, Sandra Paterlini, Jonas Wallin, Małgorzata Bogdan

Paper IV: © 2026 Małgorzata Bogdan, Adam Chojecki, Ivan Hejný, Bartosz Kołodziejek, Jonas Wallin

Paper V: © 2026 Ivan Hejný, Wojciech Rejchel, Jonas Wallin, Małgorzata Bogdan

Lund University School of Economics and Management

The Department of Statistics

Box 743, SE-220 07

Lund, Sweden

ISBN: 978-91-90202-00-5 (print)

ISBN: 978-91-90202-01-2 (electronic)

Printed in Sweden by Media-Tryck, Lund University

Lund 2026



Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN 

To my family.

Contents

Acknowledgements	iii
Abstract	v
Popular Science Summary	vii
List of papers	ix
Introduction	I
1 Background	I
2 Patterns and Their Asymptotic Convergence	4
3 Irrepresentability, Non-Degeneracy, and Exact Recovery	12
4 Estimation Error	21
5 Open Problems and Future Directions	26
6 Summary of Papers	29
7 Appendix	30
Papers	43
I Unveiling Low-Dimensional Patterns Induced by Convex Non-Differentiable Regularizers	45
II Asymptotic Distribution of Low-Dimensional Patterns Induced by Non-Differentiable Regularizers under General Loss Functions	83
III Asymptotic Theory for Graphical SLOPE: Precision Estimation and Pattern Convergence	119
IV Identifying Network Hubs with the Partial Correlation Graphical Lasso	159
V High-Dimensional Analysis of Non-Decomposable Regularizers	221

Acknowledgements

First, I would like to thank my supervisors, Jonas and Gosia. Your support and trust in me have been immensely helpful. Thank you, Gosia, for opening doors for me to work across different projects. Thank you, Jonas, for having my back and providing the stability and reassurance I needed to produce good research. I've appreciated the many moments when you suddenly appeared in my office to discuss ideas, despite me having to frantically erase the most dispensable spot on my crammed whiteboard for your massive handwriting. I also want to thank my co-authors Wojtek, Michał, Bartosz, Adam, Sandra, Philipp, and Giovanni. I have thoroughly enjoyed working on our many projects together.

Over the last five years, I have learned as much about myself as I have about statistics. Dealing with the uncertainty of research and overcoming the challenges along the way taught me how to navigate the state of not knowing, and how to have faith in the process and in myself.

I am grateful to have spent the last five years at the Department of Statistics: thank you to Krys for leading the way, Jakob for organizing us, Karin, Viktor, Anie, and Paola for doing the necessary work to keep us going, and Timo for initiating our intense table tennis exchanges. A special thank you to my PhD friends Johan, Henrik, Yvette, and Erik, who have accompanied me on this journey. Sharing the ups and downs with you made all the difference. That Henrik and Yvette came to help me with the final checks of the thesis says it all.

Thank you to my friends in Prague, especially Mates, Andrea, Martin, and Kája, who have been there before and throughout. I am very glad we could remain close. A shoutout goes to Georg and Luigi, with whom I went through the studies in Vienna that eventually led me to Lund, and to Signe for making all the dance lessons so much more fun. A special thank you to Zoran for his generosity, and to Gamu for making us all feel at home at her place over the past years. And thank you, Jenny, for your warmth and support in these last weeks.

Finally, this journey would not have been possible without my family. I am incredibly grateful for your support and for helping me retain my balance. I want to thank my grandparents, Eva, Miño, Růženka, and Miki, for all the time we have had together. Most of all, a big thank you to my parents, Marta and Mišo, who have been there for me from the very beginning.

Ivan Hejný
22th of April, 2026

Abstract

In the linear regression model, when the dimension p is fixed and $n \rightarrow \infty$, the asymptotic distribution of the estimation error for Lasso-type M -estimators is well established. In contrast, the convergence of the discrete structures (patterns) induced by these regularizers, such as sparsity or clustering, has remained largely unaddressed, even for the standard Lasso penalty. Because these lower-dimensional structures are sensitive to infinitesimal perturbations, the weak convergence of the continuous estimation error does not guarantee the convergence of the induced patterns.

This thesis develops a unified theoretical framework to resolve this discrepancy. We establish that for a range of non-differentiable regularizers with a polyhedral component (including the Lasso, Generalized Lasso, SLOPE, and Elastic Net), the estimated patterns converge toward a limiting pattern determined by an explicit asymptotic formula. This limit is characterized by the Fisher Information matrix of the loss, the score covariance matrix, and the directional derivative of the regularizer. In Paper I, this is achieved in the linear model by utilizing the Hausdorff distance as a suitable mode of set convergence involving subdifferentials. In Paper II, we extend the theory from Paper I to general statistical models via a Stochastic Lipschitz Differentiability (SLD) condition, which controls the fluctuations of the Taylor remainder and, beyond differentiable losses, encompasses robust, non-smooth functions such as the Huber and quantile losses.

The asymptotic formula characterizes the specific asymptotic irrepresentability conditions under which the true signal pattern can be recovered with high probability. To bypass instances where these stringent conditions fail (typically in highly correlated predictor structures), we develop adaptive two-step procedures based on proximal operators. Furthermore, the theory identifies a critical degeneracy in the Fused Lasso regarding its inability to recover its own clusters under orthogonal design, motivating the proposal of a Concavified Fused Lasso penalty that resolves this limitation.

Papers III and IV apply this asymptotic theory to graphical models and precision matrix estimation. Paper III establishes exact asymptotic limits for the Graphical SLOPE estimator under elliptically distributed data, revealing how it can outperform the Graphical Lasso when clustering structures are present. Paper IV focuses on the scale-invariant PCGLASSO method; we derive an irrepresentability condition under which the true sparsity structure can be recovered, theoretically justifying the method's empirical performance in discovering network hubs.

Paper V focuses on the high-dimensional regime ($p \gg n$). By introducing a *Surrogate regularizer*, a decomposable penalty that locally matches the original polyhedral regularizer around the true signal, we extend established high-dimensional estimation bounds to non-

decomposable penalties. Using this surrogate construction, we derive explicit deterministic error bounds and exact pattern recovery conditions. While our primary application focuses on SLOPE, the surrogate approach encompasses penalties such as the Generalized Lasso. Additionally, by linking the linear SLOPE sequence to the Integrated Brownian Bridge, we show that the volume ratio between the dual SLOPE and dual Lasso balls decays at an exact rate of $p^{-1/4}$ for the linear sequence, and we conjecture a $p^{-1/6}$ Gaussian volume ratio decay rate for the Benjamini-Hochberg sequence.

Popular Science Summary

It can sometimes be difficult to distinguish between what is truly important and what merely seems urgent in the moment, but is simply statistical noise. Especially as the world around us gets more complex, the number of possible explanations can feel overwhelming, and we crave simpler solutions. When researchers model the world, they often rely on a guiding principle: the true rules governing a system are usually relatively simple, relying on only a few key variables or natural groupings.

To uncover these hidden structures, statisticians use mathematical filters known as regularization. These algorithms penalize overly complicated models, forcing the mathematics to favor simpler explanations, or patterns.

For example, the filter might force the model to conclude that only five out of a thousand possible variables are actually relevant, or that certain variables behave identically and should be clustered together. Depending on the context, researchers might choose different regularizers to discover (or at least exploit) the problem-specific lower-dimensional structure, such as sparsity or clusters in the model.

While regularizers like the Lasso or Fused Lasso have become standard tools in modern statistics, a formal mathematical description of how these low-dimensional patterns behave with a fixed model dimension and increasing amounts of data was missing. This thesis fills this gap in our understanding of the behavior of these patterns as the amount of data grows infinitely large. It develops a unifying framework that encapsulates many of the most commonly used regularizers within a broad statistical context, ranging from regression models to graphical models.

Finally, the thesis applies this newly developed language to discover regulatory hubs in gene networks and to explore structures within financial portfolios.

List of papers

This thesis is based on the following papers.

- I Unveiling Low-Dimensional Patterns Induced by Convex Non-Differentiable Regularizers**
Ivan Hejný, Jonas Wallin, Małgorzata Bogdan, Michał Kos (2025). *Annals of the Institute of Statistical Mathematics*.
doi:10.1007/s10463-025-00957-6
- II Asymptotic Distribution of Low-Dimensional Patterns Induced by Non-Differentiable Regularizers under General Loss Functions**
Ivan Hejný, Jonas Wallin, Małgorzata Bogdan (2025). Preprint, *arXiv*,
doi.org/10.48550/arXiv.2506.12621. Under review.
- III Asymptotic Theory for Graphical SLOPE: Precision Estimation and Pattern Convergence**
Ivan Hejný, Giovanni Bonaccolto, Philipp Kremer, Sandra Paterlini, Jonas Wallin, Małgorzata Bogdan (2026). Preprint, *arXiv*,
doi.org/10.48550/arXiv.2604.12771. Under review.
- IV Identifying Network Hubs with the Partial Correlation Graphical Lasso**
Małgorzata Bogdan, Adam Chojewski, Ivan Hejný, Bartosz Kołodziejek, Jonas Wallin (2025). Preprint, *arXiv*,
doi.org/10.48550/arXiv.2508.12258. To be submitted for publication.
- V High-Dimensional Analysis of Non-Decomposable Regularizers**
Ivan Hejný, Wojciech Rejchel, Jonas Wallin, Małgorzata Bogdan (2026). To be submitted for publication.

All papers are reproduced with permission of their respective publishers.

Introduction

Poznatek, ke kterému člověk dospěje vlastním úsilím, má pro něj nesrovnatelně větší hodnotu než poznatek, který je mu pouze předložen.

Knowledge attained through personal effort holds incomparably greater value than knowledge that is simply handed to you.

—Milan Hejný

I Background

At the core of parametric statistics lies the assumption that the observed data $\xi_1, \dots, \xi_n$ are realizations of random variables generated according to a probability law $\mathbb{P}_{\theta_0}$, indexed by an unknown parameter $\theta_0 \in \Theta$. The goal of statistical inference is to estimate this parameter and to quantify the uncertainty associated with the estimation procedure. The overarching theme of this thesis is the analysis of estimators defined by the regularized minimization problem

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}_n(\theta) + f_n(\theta), \quad (1)$$

where $\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta; \xi_1, \dots, \xi_n)$ represents a loss function dependent on the data, and $f_n(\theta)$ is a convex penalty function (regularizer), typically induced by a norm.

This form of regularized $\mathcal{M}$ -estimation has found application across high-dimensional statistics Wainwright (2019), acting as a mathematical tool in fields such as genetics and machine learning Hastie (2009), Hastie et al. (2015), Tibshirani (1996), and Zou and Hastie (2005), signal processing Candès et al. (2006a, 2006b), and quantitative finance Brodie et al. (2009). As these applied domains increasingly rely on regularized estimators to uncover underlying structures and stabilize complex statistical models (such as identifying hub structures in genetic networks or stabilizing precision estimates for financial portfolio optimization, as demonstrated in the appended papers), understanding the mathematical limits of the optimization problem (1) is essential for ensuring their reliability.

Throughout the thesis, it is assumed that the loss $\mathcal{L}_n(\theta)$ is additive:

$$\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\xi_i, \theta), \quad (2)$$

where ℓ is a suitable loss function.

While the regularization term f_n introduces bias, it simultaneously reduces the variance of the estimator. This bias-variance trade-off often yields estimates with lower Mean Squared Error (MSE) than their unpenalized counterparts. This is particularly advantageous when the true signal θ_0 exhibits a lower-dimensional structure, such as sparsity or coefficient clustering. Furthermore, because regularizers tend to force the estimate $\hat{\theta}_n$ into lower-dimensional subspaces, they reveal information about the underlying structural properties, or *pattern*, of the signal θ_0 . The recovery of this pattern is a core theme of this thesis.

Central Questions

Primary: How can we bridge the gap between weak convergence of estimation error and the convergence of discrete structural patterns in non-differentiable regularization?

Secondary: How does this geometric framework allow us to characterize exact recovery, design adaptive procedures, and understand estimation accuracy across general statistical models?

A unifying finding of this thesis is that, across a range of statistical models and regularizers, the estimator $\hat{\theta}_n$ will successfully discover the lower-dimensional structure of the true signal θ_0 with high probability as $n \rightarrow \infty$ if and only if a generalized irrepresentability condition holds:

$$\Phi^*(\Pi u_0) < 1.$$

While the definitions of Φ^* , Π , and u_0 are deferred to Section 3, conceptually, these quantities capture the dual geometry of the regularizer f_n , the projection onto the lower-dimensional subspace of the true signal, and the Fisher Information matrix of the loss, respectively. Specific variations of this *irrepresentability* condition are well understood for the Lasso Zhao and Yu (2006), Elastic Net Jia and Yu (2010) and Yuan and Lin (2007), SLOPE Bogdan et al. (2022), and more broadly Graczyk et al. (2026) and Vaïter et al. (2017) in linear models, and have been extended to graphical model settings Graczyk, Nakashima, et al. (2025) and Ravikumar et al. (2011). The framework developed throughout this thesis unifies pattern recovery for regularizers with a polyhedral component across different statistical contexts. In particular, we demonstrate that the asymptotic limit of the objective function inherently encodes the irrepresentability condition $\Phi^*(\Pi u_0) < 1$, thereby capturing the fundamental mechanism of discrete pattern recovery. This is achieved by bridging the gap between continuous estimation error (Section 4) and discrete structural recovery (Sections 2 and 3).

1.1 The Linear Model and the Need for Regularization

To ground the abstract formulation of (1), it is instructive to consider the canonical example of regression, where one is interested in describing the dependence between a response variable $y \in \mathbb{R}$ and a covariate vector $x \in \mathbb{R}^p$. In the standard linear model, the observations $\xi_i = (y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$ satisfy

$$y_i = x_i^T \theta_0 + \varepsilon_i,$$

where θ_0 is the true parameter and the noise ε_i is centered and independent of the covariates. The standard estimator for θ_0 in this setting is the Ordinary Least Squares (OLS) estimator.

OLS solves the minimization problem (I) using the quadratic loss

$$\ell(y_i, x_i, \theta) = \frac{1}{2}(y_i - x_i^T \theta)^2,$$

with no penalization (i.e., $f_n = 0$). Under standard assumptions on the errors ε_i , the Gauss–Markov theorem establishes that OLS has the minimum variance among all linear unbiased estimators.

The OLS estimator is well-defined only when the design matrix $X = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times p}$ has full column rank p , strictly requiring $n \geq p$. In this classical regime, it admits the explicit solution $\hat{\theta}_n^{\text{OLS}} = (X^T X)^{-1} X^T y$.

When $n < p$, or when X is highly collinear, the unpenalized problem becomes ill-posed. Regularizing the loss function with a convex penalty allows for a well-defined estimation of the unknown parameter. Consider the penalized minimizer

$$\hat{\theta}_n = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^n (y_i - x_i^T \theta)^2 + f_n(\theta),$$

where f_n is a convex penalty function, typically scaled by a tuning parameter λ_n . Penalties based on the ℓ_2 and ℓ_1 norms, such as Ridge, Lasso Tibshirani (1996), or the Elastic Net Zou and Hastie (2005), shrink the estimate toward zero. Even in the classical regime where $n > p$ and OLS exists, introducing this penalty often yields a strictly lower mean squared error ($\mathbb{E}\|\hat{\theta}_n - \theta_0\|_2^2$) by sacrificing unbiasedness to significantly reduce the variance.

Beyond variance stabilization, non-differentiable ℓ_1 -type penalties can induce dimension reduction. Most notably, the Lasso penalty $f_n(\theta) = \lambda_n \|\theta\|_1$ promotes sparsity, forcing certain coefficient estimates to be exactly zero. This variable selection improves the interpretability of the estimator by distinguishing between relevant and irrelevant predictors. Understanding the asymptotic behavior of such regularized estimators, and the limiting structure of the sparsity or clustering patterns they induce, constitutes the primary focus of the thesis.

1.2 Structure of the Thesis

This thesis comprises five appended papers and an introductory essay (*kappa*). As illustrated in Figure 1, the theory developed in Papers I and II is subsequently applied in the context of graphical models in Papers III and IV, and further extended to the high-dimensional regime in Paper V. The *kappa* itself is organized around three themes: Theme I (Section 2) introduces the formal notion of a pattern and establishes the asymptotic framework. Theme II (Section 3) discusses the theoretical conditions required for exact pattern recovery, focusing on the irrepresentability condition and the related concept of non-degeneracy. Theme III (Section 4) focuses on estimation error in both classical and high-dimensional regularization settings. The *kappa* concludes with some open problems and possible future directions in Section 5, and a summary of the appended papers in Section 6. Finally, auxiliary results and mathematical preliminaries are provided in Appendix 7.

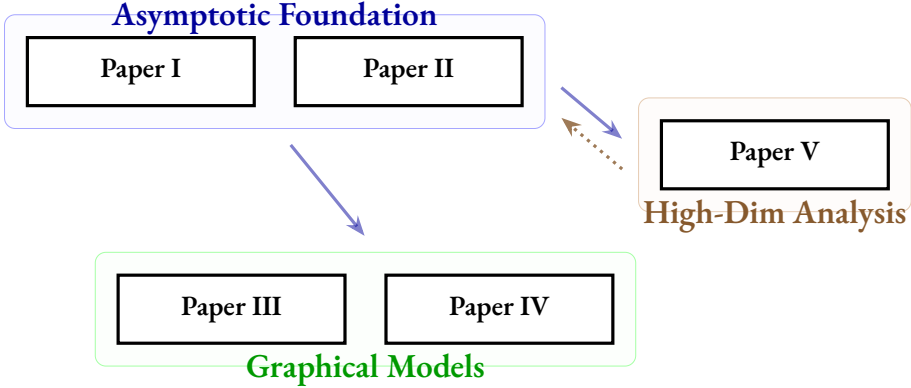


Figure 1: Diagrammatic overview of the thesis structure. The asymptotic foundation established in Papers I and II is applied to graphical models in Papers III and IV. The ideas developed in Papers I and II overlap with the high-dimensional analysis in Paper V.

1.3 Notation

Throughout the text, for any set $A \subseteq \mathbb{R}^p$, we write $\text{con}(A)$ for its convex hull, $\text{span}(A)$ for its linear span, and $\text{ri}(A)$ for its relative interior (i.e., the interior of A in the affine subspace $\text{aff}(A)$ equipped with the subspace topology of $\mathbb{R}^p$). For a vector $x \in \mathbb{R}^p$, we write $\text{sgn}(x)$ for the sign vector $(\text{sgn}(x_1), \dots, \text{sgn}(x_p))^T \in \mathbb{R}^p$. We denote the space of symmetric $p \times p$ real matrices by $\text{Sym}(p)$, and its subset of positive definite matrices by $\text{Sym}_+(p)$. We further denote by U a linear subspace of some Euclidean space equipped with an inner product (usually $\mathbb{R}^p$ for linear models and $\text{Sym}(p)$ for graphical models), and by $U^\perp$ its orthogonal complement. We write P_U and $P_{U^\perp}$ for the orthogonal projections onto U and $U^\perp$, respectively. For a vector v , we write $v_U = P_U v$ and $v_{U^\perp} = P_{U^\perp} v$ for the projected components.

2 Patterns and Their Asymptotic Convergence

While regularization was historically used to stabilize variance in ill-posed problems, non-differentiable penalties induce low-dimensional structures in the estimator $\hat{\theta}_n$, thereby simultaneously performing estimation and model selection. We refer to the specific algebraic structure imposed by a penalty f at a point θ as its *pattern*.

Mathematically, the regularizers considered in this thesis, including Lasso, Fused Lasso, and SLOPE, belong to the class of convex functions with a polyhedral component. These penalties can be expressed in the form:

$$f(\theta) = \max_{i \in \mathcal{S}} \{v_i^T \theta\} + g(\theta), \quad (3)$$

where $\mathcal{S}$ is a finite index set, the fixed vectors $v_i \in \mathbb{R}^p$ for $i \in \mathcal{S}$ determine the geometry of

the polyhedral term of the penalty, and $g(\theta)$ is some differentiable convex function.¹ The smooth term $g(\theta)$ accommodates regularizers like the Elastic Net Zou and Hastie (2005), which combines the polyhedral ℓ_1 norm with a squared ℓ_2 penalty. The *pattern* of f at θ is defined as the set of indices for which the polyhedral term achieves the maximum:

$$I_f(\theta) := \operatorname{argmax}_{i \in \mathcal{S}} \{v_i^T \theta\}. \quad (4)$$

Geometrically, rather than focusing purely on indices, it is helpful to think of regions of constant pattern in $\mathbb{R}^p$. As such, one can associate the pattern at θ_0 with the subset

$$\{\theta \mid I_f(\theta) = I_f(\theta_0)\}.$$

These sets partition the parameter space into finitely many convex regions (see Figure 2). The *pattern space* of f at θ_0 is then defined simply as the span of this set:

$$U_{\theta_0} := \operatorname{span}\{\theta \mid I_f(\theta) = I_f(\theta_0)\}. \quad (5)$$

This notion of an active set appears, for example, in Lewis (2002), and is utilized in standard texts on convex analysis (for details, see Rockafellar and Wets (1998), Example 3.16, or Hiriart-Urruty and Lemaréchal (1993), Example 3.4).

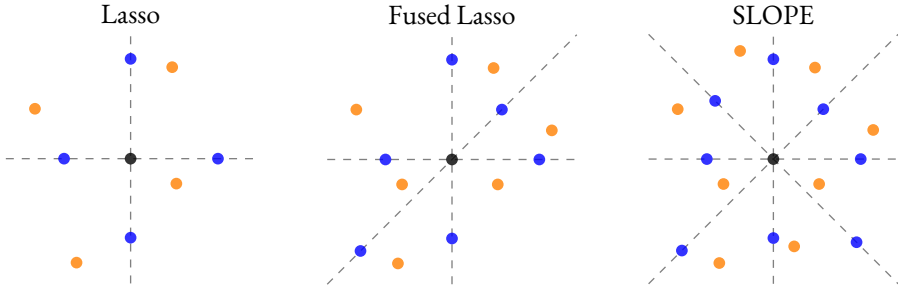


Figure 2: Geometric partition of $\mathbb{R}^2$ by the patterns of Lasso (9 distinct patterns), Fused Lasso (13 distinct patterns), and SLOPE (17 distinct patterns). Each pattern is illustrated by one representative dot. The black dot at the origin denotes the zero-dimensional pattern (full sparsity). Blue dots denote one-dimensional patterns lying on the rays with one degree of freedom. Orange dots are the two-dimensional patterns in the unconstrained sign/rank regions with two degrees of freedom.

Observe that the notion of a model pattern does not depend on the smooth $g(\theta)$ part. Therefore, when discussing patterns, one can often omit $g(\theta)$ entirely. Without the differentiable component, when $g = 0$, (3) becomes a purely polyhedral penalty

$$f(\theta) = \max_{i \in \mathcal{S}} v_i^T \theta \quad (6)$$

¹We assume that 0 is contained in the interior of the convex hull $K = \operatorname{con}\{v_i \mid i \in \mathcal{S}\}$. For essentially all examples, the regularizer f is a norm, i.e., it is symmetric, satisfying $f(\theta) = f(-\theta)$ with $K = -K$. However, symmetry is not strictly required for many general results to hold.

and the pattern is given by

$$I_f(\theta) = \{i \mid v_i^T \theta = f(\theta)\}.$$

An alternative, equivalent characterization of patterns for polyhedral penalties (6) is via subdifferentials. Precisely, for any $\theta_1, \theta_2 \in \mathbb{R}^p$

$$I_f(\theta_1) = I_f(\theta_2) \iff \partial f(\theta_1) = \partial f(\theta_2). \quad (7)$$

This equivalence class definition via subdifferentials is used, for example, in Bogdan et al. (2022), Graczyk, Nakashima, et al. (2025), Graczyk et al. (2026), and Schneider and Tardivel (2022). For more on subdifferentials, we refer to Section 7.1 and standard textbook literature Hiriart-Urruty and Lemaréchal (1993), Rockafellar and Wets (1998), and Rockafellar (2015).

Geometrically, these regions of constant pattern represent the non-differentiable “valleys” or “cusps” within the objective landscape. As long as θ moves strictly within a set of constant pattern, the overall objective $\mathcal{L}_n(\theta) + f_n(\theta)$ changes in a smooth manner. This is because the subdifferential of the polyhedral component $\partial f_n(\theta)$ remains invariant by (7), and the subdifferential of any smooth component g reduces simply to its gradient $\nabla g(\theta)$. The main optimization problem (1) seeks the bottom of one of these valleys, given by the first-order KKT condition:

$$0 \in \nabla \mathcal{L}_n(\theta) + \partial f_n(\theta).$$

The thesis is accompanied by these three examples:

Example 2.1 (Lasso). The most widely used polyhedral regularization is the Least Absolute Shrinkage and Selection Operator (Lasso), proposed by Tibshirani (1996). This uses the ℓ_1 -norm penalty:

$$f(\theta) = \lambda \|\theta\|_1,$$

where $\lambda > 0$ is a hyperparameter of the method. In the polyhedral framework (6), this corresponds to v_i being the 2^p vectors with entries $\pm \lambda$. The pattern $I_f(\theta)$ is uniquely identified by the sign vector $\text{sgn}(\theta)$. The pattern space corresponds to the subspace of vectors sharing the same support as θ_0 :

$$U_{\theta_0} = \{u \in \mathbb{R}^p : \text{supp}(u) \subseteq \text{supp}(\theta_0)\}. \quad (8)$$

Here, the low-dimensional structure is sparsity.

Example 2.2 (Fused Lasso). Beyond sparsity, penalties can enforce equality among coefficients and cluster the parameters. The Fused Lasso, introduced by Tibshirani et al. (2004), penalizes the differences between successive coefficients:

$$f(\theta) = \lambda_1 \sum_{j=1}^{p-1} |\theta_{j+1} - \theta_j| + \lambda_2 \sum_{i=1}^p |\theta_i|. \quad (9)$$

Here, $\lambda_1 > 0$ determines the fusion strength and $\lambda_2 \geq 0$ the sparsity penalty. The induced pattern encodes the information about consecutive clusters, and can be captured by the sign of the differences $\text{sgn}(\theta_1 - \theta_2, \theta_2 - \theta_3, \dots, \theta_{p-1} - \theta_p)$ together with $\text{sgn}(\theta)$.

Example 2.3 (SLOPE). The Sorted L-One Penalized Estimation (SLOPE), introduced by Bogdan et al. (2015), uses a sorted ℓ_1 norm:

$$f(\theta) = \sum_{j=1}^p \lambda_j |\theta|_{(j)}, \quad (10)$$

where $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ is a non-increasing sequence and $|\theta|_{(1)} \geq \dots \geq |\theta|_{(p)}$ are the sorted absolute magnitudes. Unlike the Fused Lasso, SLOPE is invariant under permutations of indices. The SLOPE pattern $I_f(\theta)$ captures three types of information. First, it accounts for the signs $\text{sgn}(\theta)$; second, the partition into clusters of equal magnitude $\{I_0, I_1, \dots, I_p\}$; and lastly, the ordering of these clusters by magnitude.² The SLOPE pattern can be identified with the vector of signed ranks $(\text{rank}(|\theta_i|) \text{sgn}(\theta_i))_{1 \leq i \leq p}$.³ An equivalent formulation of SLOPE is known as OWL Figueiredo and Nowak (2016), and the special case of a linearly decaying sequence was introduced earlier as OSCAR Bondell and Reich (2008).

Although our theory takes a unifying perspective on polyhedral penalization, the methods discussed above fall into completely different complexity classes. The number of distinct patterns for both Lasso and Fused Lasso grows exponentially as p increases, whereas for SLOPE it grows faster than the factorial rate. Even for $p = 6$, the number of distinct SLOPE patterns reaches 423,857: over 500 times more than the 729 Lasso patterns (see Table 1). This combinatorial complexity makes exact pattern recovery for regularizers such as SLOPE challenging, even when the sample size n is much larger than the dimension p .

Table 1: Number of distinct patterns by regularization method

Method	Growth Rate	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$	$p = 6$
Lasso	3^p	3	9	27	81	243	729
Fused Lasso	$O(4.56^p)$	3	13	59	269	1227	5597
SLOPE	$O\left(p! \left(\frac{2}{\ln 2}\right)^p\right)$	3	17	147	1697	24483	423857

2.1 Asymptotic distribution of the error

The asymptotic framework developed in Papers I–IV investigates the optimization problem

$$\mathcal{L}_n(\theta) + n^{-1/2} \alpha f(\theta), \quad (11)$$

where $\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\xi_i, \theta)$ as in (2), $\alpha > 0$ is a fixed scalar factor, f is a polyhedral norm of the form (3), and the observations ξ_i are i.i.d. with $\xi_i \sim \mathbb{P}_{\theta_0}$. The scaling rate $O(n^{-1/2})$ is the exact rate for low-dimensional asymptotics required to obtain non-trivial asymptotic formulas for the error (see Figure 3). The data-dependent loss term $\mathcal{L}_n(\theta) = O_p(1)$ dominates

²A (SLOPE) cluster is a subset of indices $I \subset \{1, \dots, p\}$ such that $|\theta_i| = |\theta_j|$ whenever $i, j \in I$, and $|\theta_i| \neq |\theta_j|$ if $i \in I, j \notin I$.

³When we refer to a SLOPE pattern, we always mean the pattern associated with a strictly decreasing sequence $\lambda_1 > \dots > \lambda_p > 0$.

the $O(n^{-1/2})$ penalization; thus, for typical losses $\ell(\xi, \theta)$ such as the negative log-likelihood, standard M -estimation theory (e.g., van der Vaart (1998)) often guarantees that $\widehat{\theta}_n$ remains consistent for θ_0 .

However, at the level of discrete patterns, the sequence $I_f(\widehat{\theta}_n)$ never truly stabilizes as $n \rightarrow \infty$, exhibiting what could best be described as “pattern-jumping.” For a fixed $\alpha > 0$, the sequence $I_f(\widehat{\theta}_n)$ will persistently jump between different patterns “around” θ_0 . Because the pattern is highly sensitive to infinitesimal perturbations, it never permanently stabilizes at $I_f(\theta_0)$ despite the continuous consistency of the estimator itself. By *asymptotic pattern recovery*, we mean that by tuning α to a sufficiently high value, the probability of an exact pattern match, $\mathbb{P}[I_f(\widehat{\theta}_n) = I_f(\theta_0)]$, can be made arbitrarily close to 1. We explore this asymptotic pattern recovery in closer detail in the next subsection.

For the remainder of this section, we temporarily set aside the discrete pattern structure and focus instead on the asymptotic distribution of the continuous rescaled error:

$$\widehat{u}_n := \sqrt{n}(\widehat{\theta}_n - \theta_0).$$

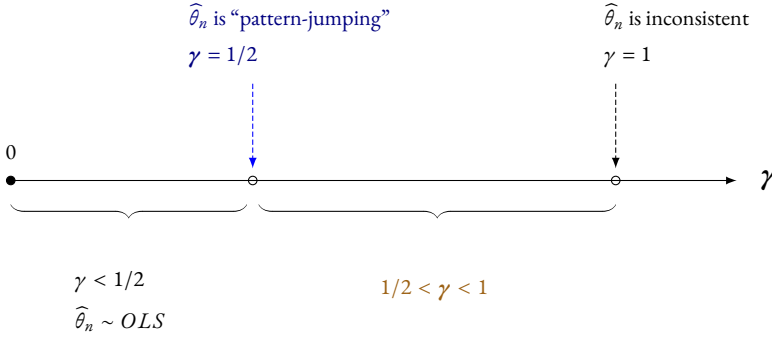


Figure 3: Asymptotic scaling regimes for the estimator $\widehat{\theta}_n$ minimizing $\mathcal{L}_n(\theta) + n^\gamma f(\theta)$ for fixed p as $n \rightarrow \infty$. The penalty exponent γ determines the asymptotic behavior: **(1)** $\gamma < 1/2$: The penalty is negligible, $\widehat{\theta}_n$ is asymptotically equivalent to OLS, and no model selection occurs. **(2)** $\gamma = 1/2$: At this critical threshold, the rescaled error $\widehat{u}_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in distribution, and the probability of exact pattern recovery is strictly in $(0, 1)$. **(3)** $1/2 < \gamma < 1$: The rescaled $\widehat{u}_n$ error diverges, but exact pattern recovery is possible (provided the irrepresentability condition holds).

Substituting $\theta = \theta_0 + \frac{u}{\sqrt{n}}$ into (ii), yields the objective function minimized by the rescaled error $\widehat{u}_n$:

$$V_n(u) = n \left(\mathcal{L}_n \left(\theta_0 + \frac{u}{\sqrt{n}} \right) - \mathcal{L}_n(\theta_0) \right) + \alpha \sqrt{n} \left(f \left(\theta_0 + \frac{u}{\sqrt{n}} \right) - f(\theta_0) \right).$$

If the loss $\ell(\xi, \theta)$ is twice differentiable in θ , a second-order Taylor expansion of $\mathcal{L}_n$ and an approximation of f via its directional derivative yield:

$$V_n(u) = \frac{1}{2} u^T \nabla^2 \mathcal{L}_n(\theta_0) u + u^T \sqrt{n} \nabla \mathcal{L}_n(\theta_0) + \alpha f'(\theta_0; u) + R_n(u), \quad (12)$$

where $R_n(u)$ is a remainder term that, under suitable regularity conditions on the loss, becomes negligible as $n \rightarrow \infty$. In the limit, the rescaled error $\widehat{u}_n$ converges in distribution to the random vector $\widehat{u}$ that minimizes the pointwise limit objective function:

$$V(u) = \frac{1}{2}u^T C u - u^T W + \alpha f'(\theta_0; u), \quad (13)$$

where $W \sim \mathcal{N}(0, C_\Delta)$, and the characteristic matrices are defined as:

$$C = \mathbb{E}[\nabla^2 \ell(\xi, \theta_0)],$$

$$C_\Delta = \text{Cov}(\nabla \ell(\xi, \theta_0)).$$

The asymptotic distribution of the error is thus entirely determined by the law of the minimizer $\widehat{u}$. The expected Hessian C , the score covariance C_Δ , and the regularizer-specific directional derivative $u \mapsto f'(\theta_0; u)$ uniquely characterize the asymptotic behavior of the regularized estimator.

Remark 2.4. Note that in the absence of regularization ($f = 0$), the asymptotic distribution is exact. By solving $0 = \nabla V(u) = C u - W$, we recover the classical M -estimation limit

$$\widehat{u} = C^{-1}W \sim \mathcal{N}(0, C^{-1}C_\Delta C^{-1}).$$

For MLE estimators, where the loss $\ell(x, \theta) = \log(p_\theta(x))$, under standard regularity conditions, the Hessian C and covariance C_Δ both coincide with the Fisher Information matrix $I(\theta_0) = C^{-1}$, and the asymptotic error is $\widehat{u} \sim \mathcal{N}(0, C^{-1})$.

To illustrate the general asymptotic framework, consider the following examples appearing throughout the thesis:

Example 2.5 (Linear regression). Consider the linear model $y = x^T \theta_0 + \varepsilon$, with the regressor vector x such that $\mathbb{E}[x] = 0$, positive definite covariance $\mathbb{E}[xx^T]$, and noise satisfying $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon^2] = \sigma^2$. Assume the regressors x and the noise are independent, $x \perp \varepsilon$. Consider the quadratic loss

$$\ell(y, x, \theta) = \frac{1}{2}(y - x^T \theta)^2.$$

The derivatives are $\nabla \ell(\theta_0) = -\varepsilon x$ and $\nabla^2 \ell(\theta_0) = xx^T$, hence $C = \mathbb{E}[xx^T]$ and $C_\Delta = \sigma^2 C$. The empirical loss is

$$\mathcal{L}_n(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2 = \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i, \theta).$$

The Taylor expressions from (12) converge by the Central Limit Theorem and the Strong Law of Large Numbers, respectively, to:

$$-\sqrt{n} \nabla \mathcal{L}_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i x_i \xrightarrow{d} W \sim \mathcal{N}(0, C_\Delta),$$

$$\nabla^2 \mathcal{L}_n(\theta_0) = \frac{1}{n} \sum_{i=1}^n x_i x_i^T \xrightarrow{a.s.} \mathbb{E}[xx^T] = C.$$

Example 2.6 (Graphical Models). Let $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ be i.i.d. copies of a centered random vector $X \in \mathbb{R}^p$. We are interested in estimating the inverse covariance matrix $\Theta_0 := (\mathbb{E}[XX^T])^{-1}$. Most notably, we consider the Gaussian Graphical Model $X \sim \mathcal{N}(0, \Sigma)$, where Σ is some positive definite covariance. The sparsity structure of $\Theta_0 = \Sigma^{-1}$ encodes the conditional independencies: $(\Theta_0)_{ij} = 0$ if and only if $X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}$ (e.g. Lauritzen (1996)).

Minimizing (11) with the loss given by the negative log-likelihood of the Gaussian yields

$$\mathcal{L}_n(\Theta) = -\frac{1}{2} \log \det(\Theta) + \frac{1}{2} \text{tr}(\Theta S),$$

where $S = \frac{1}{n} \sum_{i=1}^n X^{(i)} (X^{(i)})^T$ is the empirical covariance. We have

$$\ell(x, \Theta) = -\frac{1}{2} \log(\det(\Theta)) + x^T \Theta x.$$

The distribution of the error $\sqrt{n}(\hat{\Theta}_n - \Theta_0)$ converges to the graphical analog of the asymptotic distribution (13) in vectorized form:⁴

$$\hat{U} = \underset{U \in \text{Sym}(p)}{\text{argmin}} \left\{ \frac{1}{2} \text{vec}(U)^T C \text{vec}(U) - \text{vec}(U)^T W + \alpha f'(\Theta_0; U) \right\}, \quad (14)$$

where $W \sim \mathcal{N}(0, C_\Delta)$ with $C = \frac{1}{2}(\Theta_0^{-1} \otimes \Theta_0^{-1})$ and $C_\Delta = \frac{1}{4} \text{Cov}((XX^T))$. Under Gaussianity where the data are distributed as $X \sim \mathcal{N}(0, \Sigma)$, we get $C_\Delta = C$, (see Remark 2.4).

Similar distributional convergence results hold for various statistical models and fall within the classical statistical asymptotic theory of M -estimators van der Vaart and Wellner (1996) and van der Vaart (1998). The addition of the convex penalty f remains well-behaved and does not complicate the asymptotics. This type of asymptotics and the limiting formulas were established for Lasso-type estimators and the Fused Lasso in Fu and Knight (2000) and Tibshirani et al. (2004).

2.2 Pattern convergence

The initial motivation that led to the development of the theory in Paper I and Paper II was the observation that distributional convergence of $\hat{u}_n$ toward $\hat{u}$ does not necessarily guarantee that the pattern $I_f(\hat{u}_n)$ converges weakly to $I_f(\hat{u})$. This was a relevant and non-trivial question. It was not addressed, likely overlooked in the existing literature, and lacking a formal argument even in the seminal works on ℓ_1 -regularization Fu and Knight (2000) and Tibshirani et al. (2004).

Here, pattern convergence refers to standard weak convergence with respect to the discrete topology on the finite set of patterns.

Definition 2.7 (Pattern convergence). Let f be a regularizer of the form (3) and $(\hat{u}_n)_{n \in \mathbb{N}}$ a sequence of random vectors in $\mathbb{R}^p$. We say that $\hat{u}_n$ converges *weakly in f -pattern* to a random vector $\hat{u}$ if

$$\lim_{n \rightarrow \infty} \mathbb{P}[I_f(\hat{u}_n) = \mathbf{p}] = \mathbb{P}[I_f(\hat{u}) = \mathbf{p}], \quad (15)$$

⁴The expressions for C and C_Δ follow directly from the derivatives of the loss with respect to Θ , given by $\nabla \ell(\Theta) = \frac{1}{2}(-\Theta^{-1} + XX^T)$ and $\nabla^2 \ell(\Theta) = \frac{1}{2}(\Theta^{-1} \otimes \Theta^{-1})$.

for every pattern $\mathbf{p} \subset \{1, \dots, p\}$. In this case, we write $I_f(\widehat{u}_n) \xrightarrow{d} I_f(\widehat{u})$.

As demonstrated in Paper I, in the linear model for the non-polyhedral convex penalty $f(\theta) = \max\{\theta_1^2, \theta_2\}$ on $\mathbb{R}^2$, one has weak convergence of $\widehat{u}_n$ toward $\widehat{u}$, but the sparsity pattern $\text{sgn}(\widehat{u}_n)$ fails to converge toward $\text{sgn}(\widehat{u})$. If the true signal is $\theta_0 = (0, 0)^T$, then the penalization with $n^{-1/2}f$ gives $\mathbb{P}(\widehat{\theta}_{n,2} = 0) = 0$ for every n . Yet, in the limit, the directional derivative $f'(0; u) = \max\{0, u_2\}$ enforces sparsity in the second component, yielding $\mathbb{P}(\widehat{u}_2 = 0) > 0$ (see Figure 4). As such, the sparsity structure of the limit $\widehat{u}$ does not reflect the sparsity of the finite-sample estimator $\widehat{\theta}_n$.

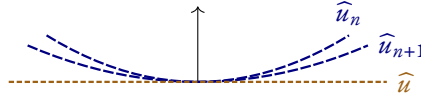


Figure 4: $\widehat{u}_n$ places mass on a sequence of flattening parabolas. The limiting distribution $\widehat{u}$ exhibits sparsity in the second component, whereas any $\widehat{u}_n$ is sparsity-free. The exact same phenomenon occurs for the original estimator sequence $\widehat{\theta}_n = \widehat{u}_n / \sqrt{n}$.

Patterns are discrete objects sensitive to minuscule perturbations, which are completely invisible to the weak convergence of the error $\widehat{u}_n$. Since this discontinuous pattern function renders the Continuous Mapping Theorem inapplicable, it raises the following question:

Under what conditions on f and $\theta \mapsto \ell(\theta)$ does

$$\widehat{u}_n \xrightarrow{d} \widehat{u} \implies I_f(\widehat{u}_n) \xrightarrow{d} I_f(\widehat{u}) \quad ?$$

In Paper I, we establish that for the linear model, it is exactly the sublinear structure of the polyhedral part (3) that stabilizes the pattern spaces and guarantees $I_f(\widehat{u}_n) \xrightarrow{d} I_f(\widehat{u})$. Sublinearity of the regularizer prevents the mass from traveling, which is the underlying cause of the phenomenon shown in Figure 4. In Paper II, we focus on the loss $\ell(\theta)$ and establish sufficient conditions under which polyhedral penalties achieve pattern convergence.

Let f be a penalty of the form (3) and θ_0 be the minimizing value of $G(\theta) = \mathbb{E}[\ell(\xi, \theta)]$. Assume there is an open neighborhood B around θ_0 on which $\theta \mapsto \ell(\xi, \theta)$ is twice continuously differentiable (C^2) for every ξ , and:

- (i) $\|\nabla^2 \ell(\theta_0)\| \leq M$ for $\theta \in B$, where M is a random variable with $\mathbb{E}[M^2] < \infty$.
- (ii) $G(\theta)$ is C^3 on B and has a positive definite Hessian $C = \nabla_{\theta=\theta_0}^2 G(\theta)$ at θ_0 .
- (iii) $\mathbb{E}[\nabla \ell(\theta_0)] = 0$ and the covariance matrix $C_\Delta = \mathbb{E}[\nabla \ell(\theta_0) \nabla \ell(\theta_0)^T] < \infty$. (16)
- (iv) $\widehat{\theta}_n$ is consistent for θ_0 .

Then the rescaled error $\widehat{u}_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ converges in distribution to the minimizer $\widehat{u}$ of (13). Moreover, $\widehat{u}_n$ converges to $\widehat{u}$ in pattern in the sense of (15).

The above conditions⁵ are similar to the classical conditions that ensure the asymptotic normality of M -estimators Pollard (1985) and van der Vaart (1998). Importantly, these conditions are satisfied for the quadratic loss in the linear model (Example 2.5, central to Papers I and II) and the graphical loss (Example 2.6, central to Papers III and IV).

These conditions can be further weakened to incorporate losses that do not satisfy the differentiability assumptions above, such as the robust median or Huber loss. This is done in Paper II using the notion of *Stochastic Lipschitz Differentiability*, which imposes a local uniform control over the remainder term in the first-order Taylor expansion of the loss $\ell(\theta)$. This notion is closely related to stochastic differentiability Pollard (1985), but has been modified to ensure the fundamentally distinct convergence of patterns.

Ultimately, we are interested in the distribution of the pattern for the original estimator $\widehat{\theta}_n$. Intuitively, pattern convergence ensures that the pattern distribution of $\widehat{\theta}_n$ is asymptotically equivalent to that of a tiny perturbation of θ_0 in the direction $\widehat{u}$. More precisely,

$$I_f(\widehat{\theta}_n) = I_f\left(\theta_0 + \frac{\widehat{u}_n}{\sqrt{n}}\right) \xrightarrow[\varepsilon \downarrow 0]{d} \lim_{\varepsilon \downarrow 0} I_f(\theta_0 + \varepsilon \widehat{u}),$$

where $\widehat{u}$ minimizes (13). Asymptotically, the probability of exact pattern recovery approaches:

$$\lim_{n \rightarrow \infty} \mathbb{P}[I_f(\widehat{\theta}_n) = I_f(\theta_0)] = \mathbb{P}(\widehat{u} \in U_{\theta_0}), \quad (17)$$

where U_{θ_0} is the pattern space of f at the point θ_0 . From (13), one can show that the above probability is

$$\mathbb{P}(W \in CU_{\theta_0} + \alpha \partial f(\theta_0)), \quad (18)$$

where $W \sim \mathcal{N}(0, C_\Delta)$. We say that f recovers the pattern θ_0 asymptotically with high probability if the probability in (18) goes to 1 as $\alpha \rightarrow \infty$. This bridges the asymptotic viewpoint to the second central theme of the entire thesis:

3 Irrepresentability, Non-Degeneracy, and Exact Recovery

A natural question in high-dimensional statistics concerns the ability of regularization methods, such as the Lasso, Fused Lasso, or SLOPE, to recover their true patterns. This means that the pattern of the estimator, $I_f(\widehat{\theta}_n)$, exactly matches the pattern of the underlying signal, $I_f(\theta_0)$. Hypothetically, possessing oracle information about the true pattern of the signal θ_0 would reduce a high-dimensional problem to a lower-dimensional one, solvable by standard techniques such as MLE.

In practice, however, the lower-dimensional sparsity or clustering structure of the signal is unknown. Exact recovery of the pattern via regularization requires several stringent assumptions to be satisfied. Consider the linear model $y_i = x_i^T \theta_0 + \varepsilon_i$ with the loss $\mathcal{L}(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2$. Broadly speaking, exact (non-asymptotic) pattern recovery requires three conditions:

⁵In Papers II, III, and IV, the consistency condition (iv) is replaced by uniform tightness and a uniform envelope condition that together imply consistency. Also, as discussed in Paper II, compared to mere distributional convergence, guaranteeing pattern convergence requires strengthening the standard L^1 integrability of M to L^2 integrability, and instead of standard C^2 differentiability of the population risk G , we assume the risk to be locally C^3 .

1. *Irrepresentability Condition*, which requires the correlation of the design matrix to respect the pattern structure of the signal; i.e., zeros cannot be highly correlated with non-zeros (Lasso), and clusters cannot be strongly correlated with each other (SLOPE).
2. *Noise Control*, which imposes a bound on the noise that could otherwise destabilize the pattern of the estimate.
3. *Separation Distance*, which ensures that the signal is well separated from other patterns within its own pattern space; i.e., non-zeros are far away from zero (Lasso), and clusters are not too close to each other (SLOPE).

Whereas both the noise control and separation distance conditions benefit from an increasing sample size and are satisfied with high probability in the low-dimensional limit (fixed p as $n \rightarrow \infty$), the irrepresentability condition remains restrictive even with infinite data. As such, it can disrupt exact pattern recovery when the covariates are correlated, and does not resolve even as $n \rightarrow \infty$. In the asymptotic limit, the most succinct version of the irrepresentability condition can be expressed as

$$CU_{\theta_0} \cap \text{ri}(\partial f(\theta_0)) \neq \emptyset, \quad (19)$$

where C is the Hessian of the expected loss, $\nabla_{\theta}^2 \mathbb{E}[\mathcal{L}(\theta)]$, evaluated at the true signal vector θ_0 . For the quadratic loss in the linear model, the Hessian coincides with the covariance matrix of the predictors, $C = \mathbb{E}[\frac{1}{n} X^T X]$ (see Graczyk et al. (2023, 2026) and Schneider and Tardivel (2022)). It is worth noting that this asymptotic formulation is applicable to general statistical models beyond the linear model. As discussed in the previous section, under mild regularity conditions on the loss, the limiting probability of pattern recovery is given explicitly by (18). This probability goes to 1 as $\alpha \rightarrow \infty$ if and only if (19) holds.

For the linear model under an orthogonal design (i.e., $C = I$), where the covariates x_i are independent, the irrepresentability condition (19) reduces to a purely intrinsic property of the regularizer f in relation to the pattern of the signal. We say that θ_0 is *non-degenerate* with respect to f when the pattern space intersects the relative interior of the subdifferential:

$$U_{\theta_0} \cap \text{ri}(\partial f(\theta_0)) \neq \emptyset.$$

Although this assumption is satisfied for all patterns for both Lasso and SLOPE, it fails for instance for $\theta_0 = (1, 2, 2, 3)$ with respect to the Fused Lasso penalty (9). Degeneracy of these monotonic consecutive clusters is the reason why the Fused Lasso cannot reliably recover its own clusters under orthogonal design ($C = I$), irrespective of the hyperparameter tuning λ_1, λ_2 , even as $n \rightarrow \infty$. To resolve this limitation, we develop the Concavified Fused Lasso (Paper 1), which fixes the degeneracy problem, as illustrated in Figure 5. Furthermore, non-degeneracy is the only requirement for higher-order adaptive procedures (Section 3.5) to successfully circumvent the stringent irrepresentability condition (19) and recover the pattern with high probability in the asymptotic limit.

3.1 Decomposable Surrogate Construction

The notion of pattern non-degeneracy is coined in Paper v in relation to the *surrogate regularizer* construction: a decomposable regularizer Φ that locally approximates a non-decomposable penalty f around θ_0 .

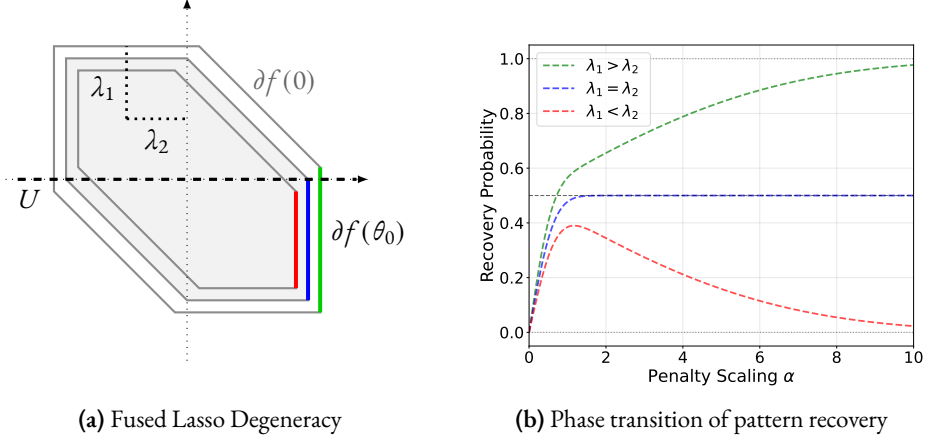


Figure 5: Geometry and Phase Transition of the Fused Lasso. (a) The subdifferential $\partial f(0)$ for the 2D Fused Lasso penalty $f(\theta) = \lambda_1 \|\theta\|_1 + \lambda_2 |\theta_1 - \theta_2|$. For the signal $\theta_0 = (1, 0)^T$, the active subdifferential $\partial f(\theta_0)$ corresponds to the rightmost edges. Exact asymptotic recovery (i.e., $(\hat{\theta}_n)_2 = 0$) under orthogonal design ($C = I$) requires signal non-degeneracy: the pattern space U (dashed line) must intersect the relative interior of $\partial f(\theta_0)$. This holds only for the green configuration ($\lambda_1 > \lambda_2$). (b) The limiting probability of pattern recovery as a function of the penalty scaling α . The green curve satisfies the strict irrepresentability condition, asymptotically reaching probability 1. The red ($\lambda_1 < \lambda_2$) and blue ($\lambda_1 = \lambda_2$) curves represent degenerate configurations where the condition fails, bounding the recovery probability strictly below 1/2 (for detailed derivations, see Appendix 7.2). (Panel (a) adapted from Papers I and v)

The surrogate Φ depends on the true signal $\theta_0 \in U$ and is defined separately by its components on the pattern space U and its orthogonal complement $U^\perp$:

$$\begin{aligned} \Phi_U(u) &:= f(u) & \forall u \in U, \\ \Phi_{U^\perp}(v) &:= f'(\theta_0; v) & \forall v \in U^\perp, \end{aligned} \quad (20)$$

where $f'(\theta_0; v)$ denotes the directional derivative at θ_0 in direction v (see Appendix 7.1). The full surrogate is defined as the sum of the two components

$$\Phi(\theta) := \Phi_U(\theta_U) + \Phi_{U^\perp}(\theta_{U^\perp}).$$

The dual surrogate is given by $\Phi^*(z) = \max\{\Phi_U^*(z_U), \Phi_{U^\perp}^*(z_{U^\perp})\}$, where $\Phi_U^*(z_U) = f^*(z_U)$ and

$$\Phi_{U^\perp}^*(z_{U^\perp}) = \inf\{\gamma > 0 \mid z_{U^\perp} \in \gamma P_{U^\perp} \partial f(\theta_0)\}. \quad (21)$$

As discussed in detail in Paper v, the dual surrogate gauge $\Phi_{U^\perp}^*$ defines an appropriate “distance” on $U^\perp$ that governs both the probability of pattern recovery and the estimation accuracy of the original minimization problem (1) involving f .⁶ We visualize the surrogate and dual surrogate for the SLOPE norm in Figure 6.

⁶The gauge function $\Phi_{U^\perp}^*$ is a proper metric (or norm) on $U^\perp$, whenever the projected face $K = P_{U^\perp} \partial f(\theta_0)$ is symmetric around 0, i.e., $K = -K$. For SLOPE, this is true for all θ_0 .

By construction, Φ is *decomposable* over $(U, U^\perp)$, i.e., $\Phi(v) = \Phi(v_U) + \Phi(v_{U^\perp})$ for any $v \in \mathbb{R}^p$ Negahban et al. (2012). If the original penalty f is itself decomposable over $(U, U^\perp)$:

$$f(u + v) = f(u) + f(v), \quad (22)$$

for any $u \in U$ and $v \in U^\perp$, then the directional derivative coincides with the regularizer on the perpendicular space, $f'(\theta_0; v) = f(v)$. In this decomposable setting, the surrogate simply reduces to the original regularizer, $\Phi = f$. The surrogate Φ extends the theory of decomposable regularizers Negahban et al. (2012) and Wainwright (2019) to non-decomposable ones.

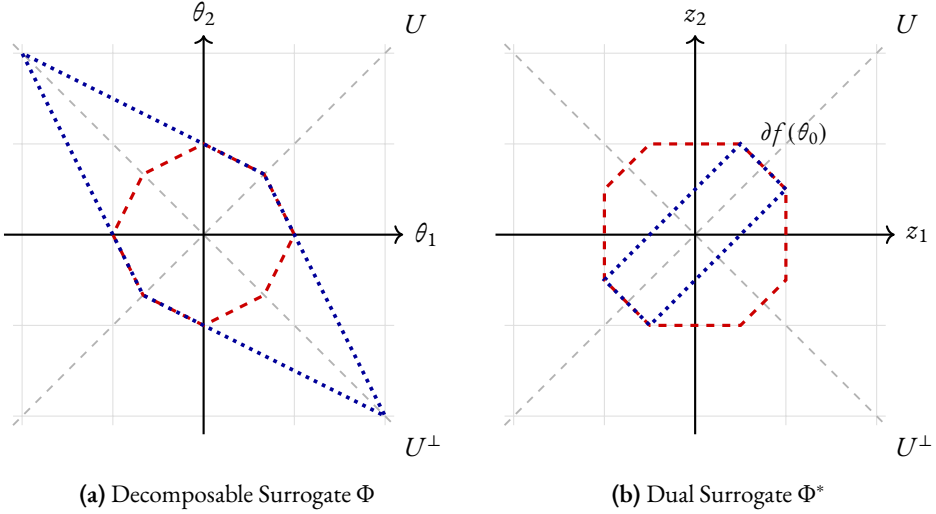


Figure 6: SLOPE norm $f(\theta) = J_\lambda(\theta)$, with $\lambda = (1, \frac{1}{2})$, and its Decomposable Surrogate $\Phi(\theta)$ for $U = \text{span}\{(1, 1)^T\}$. Left: Unit SLOPE ball $\{J_\lambda(\theta) \leq 1\}$ (red, dashed), Decomposable Surrogate unit ball $\{\Phi(\theta) \leq 1\}$ (blue, dotted). Right: corresponding dual norm balls $\{J_\lambda^*(z) \leq 1\}$ and $\{\Phi^*(z) \leq 1\}$. (Adapted from Paper v)

Example 3.1 (Lasso). Let $f(\theta) = \|\theta\|_1$ be the Lasso penalty, and let $S := \text{supp}(\theta_0)$ be the support of the signal. The pattern space is $U = \{\theta \in \mathbb{R}^p \mid \text{supp}(\theta) \subseteq S\}$, and the perpendicular space is $U^\perp = \{\theta \in \mathbb{R}^p \mid \theta_S = 0\}$. By decomposability, $\Phi = f = \|\cdot\|_1$ and $\Phi^* = f^* = \|\cdot\|_\infty$. Explicitly, $\Phi_U(u) = \|u_S\|_1$ and $\Phi_{U^\perp}(v) = f'(\theta_0; v) = \|v_{S^c}\|_1$. For the dual, we have $\Phi_U^*(z_U) = \|z_S\|_\infty$ and $\Phi_{U^\perp}^*(z_{U^\perp}) = \|z_{S^c}\|_\infty$.

3.2 Non-asymptotic view in the linear model

Various formulations and conditions related to the general irrepresentability condition have been developed for the Lasso in the linear model. These include the original irrepresentability condition Zhao and Yu (2006), mutual incoherence Wainwright (2019), and the noiseless recovery condition Graczyk et al. (2023). The surrogate Φ unifies these concepts within a single framework for any polyhedral regularizer.

The irrepresentability condition measures the leakage between the active subspace U and the passive subspace $U^\perp$. This is captured by the cross-operator $\Pi_{U^\perp|U}^n : \mathbb{R}^p \rightarrow \mathbb{R}^p$, defined by:

$$\Pi_{U^\perp|U}^n := (P_{U^\perp} X^T X P_U) (P_U X^T X P_U)^+,$$

where A^+ denotes the Moore-Penrose pseudoinverse of a matrix A , and X is the $n \times p$ design matrix. This operator maps vectors from U to $U^\perp$.⁷

We can now formulate the three exact recovery conditions using the dual surrogate gauge Φ^* . Throughout, let $U := U_{\theta_0}$ be the pattern space of the true signal and assume non-degeneracy of θ_0 w.r.t. f , which gives that $u_0 = P_U \partial f(\theta_0) \in \text{ri}(\partial f(\theta_0))$, and that the restricted Hessian $P_U X^T X P_U$ is invertible on U .⁸ Finally, we let $\|A\|_{\Phi^*} := \sup_{\Phi^*(x) \leq 1} \Phi^*(Ax)$ denote the operator norm w.r.t. Φ^* .

1. **Mutual Incoherence:**

$$\|\Pi_{U^\perp|U}^n\|_{\Phi^*} < 1. \quad (23)$$

This measures the maximum leakage of any pattern-space vector into the orthogonal space strictly through the surrogate geometry.

2. **Irrepresentability:**

$$\Phi_{U^\perp}^* \left(\Pi_{U^\perp|U}^n(u_0) \right) < 1. \quad (24)$$

This is the specific evaluation of the leakage for the actual signal subgradient u_0 .

3. **Noiseless Recovery:**

$$X^T X U \cap \partial f(\theta_0) \neq \emptyset. \quad (25)$$

This is a geometric viewpoint that transforms the pattern space by $X^T X$ Graczyk et al. (2023).

As shown in Paper v, we have the strict implications $(23) \implies (24) \implies (25)$. Moreover, if $\partial f(\theta_0)$ is replaced with $\text{ri}(\partial f(\theta_0))$ in (25), we have $(24) \iff (25)$.

To ground these abstract concepts, we illustrate how they reduce to the well-known conditions for the standard Lasso (as introduced in Example 3.1). Letting $S = \text{supp}(\theta_0)$, the general mutual incoherence (23) can be shown to recover the classical condition Wainwright (2019):

$$\max_{j \in S^c} \|(X_S^T X_S)^{-1} X_S^T X_j\|_1 < 1.$$

The irrepresentability condition (24) evaluates this forward mapping strictly at the sign vector $u_S = \text{sgn}((\theta_0)_S)$, yielding the original formulation by Zhao and Yu Zhao and Yu (2006):

$$\|X_{S^c}^T X_S (X_S^T X_S)^{-1} u_S\|_\infty < 1.$$

Finally, the noiseless recovery condition (25) translates directly to the classical intersection property:

$$X^T X \mathbb{R}^S \cap \partial \|\theta_0\|_1 \neq \emptyset.$$

⁷The map $\Pi_{U^\perp|U}^n$ annihilates $U^\perp$. For $A := P_U X^T X P_U$, we have $U^\perp \subseteq \text{Ker}(A^+)$, because the kernel of the Moore-Penrose pseudoinverse is identical to the kernel of its transpose.

⁸Note that if $A := P_U X^T X P_U$ is invertible when restricted to U , then $A^+|_U = (A|_U)^{-1}$. In this case, the null space is exactly $\text{Ker}(A^+) = U^\perp$, and $AA^+ = A^+A = P_U$.

3.3 Asymptotic pattern recovery

While the non-asymptotic analysis depends on the specific realization of the design matrix, the asymptotic viewpoint isolates the geometry from the finite-sample dependencies. As $n \rightarrow \infty$, the empirical cross-operator $\Pi_{U^\perp|U}^n$ converges to its population counterpart:

$$\Pi_{U^\perp|U} := (P_{U^\perp} C P_U)(P_U C P_U)^+,$$

where C is the expected Hessian of the loss. The general asymptotic irrepresentability condition (19) can be written in terms of the dual surrogate as:

$$\Phi_{U^\perp}^* (\Pi_{U^\perp|U}(u_0)) < 1, \quad (26)$$

where $u_0 = P_U \partial f(\theta_0)$.

Compared to the non-asymptotic condition (24), the asymptotic irrepresentability (26) characterizes pattern recovery for general M -estimators solving (1). Assume that (26) holds, and that $\hat{\theta}_n$ is any sequence of estimators for which the scaled subgradient $\hat{u}_n$ converges in distribution and in pattern to the minimizer $\hat{u}$ of the limiting objective (13). Then, the estimator asymptotically recovers the true pattern in the sense:

$$\lim_{n \rightarrow \infty} \mathbb{P}[I_f(\hat{\theta}_n) = I_f(\theta_0)] \geq 1 - e^{-c\alpha^2}, \quad (27)$$

for some constant $c > 0$. Conversely, if (26) is strictly violated, the limiting probability of recovery is bounded above by $1/2$ for all $\alpha \geq 0$.

We remark that the condition (26) is mathematically equivalent to:⁹

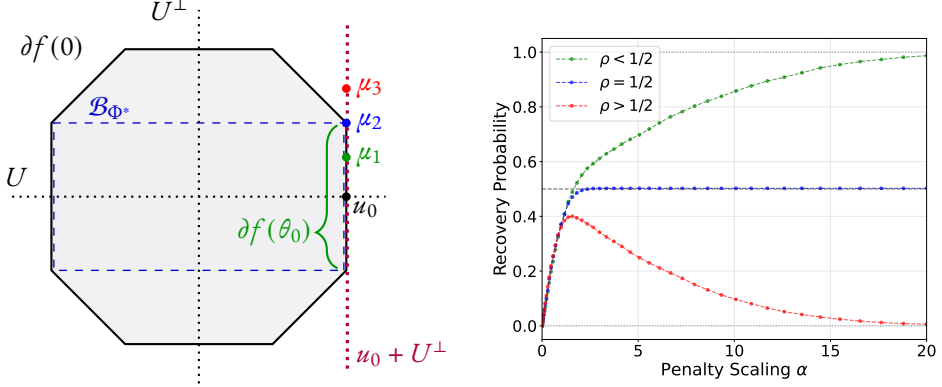
$$\mu := u_0 + \Pi_{U^\perp|U}(u_0) = C P_U (P_U C P_U)^+ u_0 \in \text{ri}(\partial f(\theta_0)). \quad (28)$$

The core ideas of the asymptotic framework are all illustrated in Figure 7. On the left is a depiction of the dual SLOPE ball with $\lambda = (1, \frac{1}{2})$, together with the dual surrogate ball $\mathcal{B}_{\Phi^*} = \{\Phi^*(z) \leq 1\}$ associated with the signal vector $\theta_0 = (1, 0)^T$. The irrepresentability vectors μ as in (28) depend on the Hessian C , given by a correlation matrix with correlation $\rho \in (-1, 1)$. As ρ changes, the vector $\mu = \mu(\rho)$ moves along the affine space of the dual face $\partial f(\theta_0)$. For highly correlated designs, μ escapes the interior of the face, i.e., $\mu \notin \text{ri}(\partial f(\theta_0))$, or equivalently $\Phi^*(P_{U^\perp}\mu) = \Phi^*(\Pi_{U^\perp|U}(u_0)) \geq 1$. In this case, the asymptotic recovery probability (18) remains bounded strictly below $1/2$ for any value of the tuning parameter α . Conversely, if the correlation ρ is small, μ remains close to the pattern space U well-within the face interior, and the recovery probability approaches 1 as $\alpha \rightarrow \infty$.

3.4 Application to graphical models

The asymptotic approach developed in Papers I and II applies directly to Gaussian graphical models, as introduced in Example 2.6. The penalized estimate of the precision matrix Θ_0 is

⁹In Papers I and II, this operator is expressed using the orthogonal projection P onto the skewed subspace $C^{1/2}U$. In that notation, the cross-operator takes the form $\Pi_{U^\perp|U} = P_{U^\perp} C^{1/2} P C^{-1/2}$, and the irrepresentability condition (26) is written as $C^{1/2} P C^{-1/2} u_0 \in \text{ri}(\partial f(\theta_0))$.



(a) Geometric irrepresentability. The surrogate dual ball $\mathcal{B}_{\Phi^*}$ (dashed) locally matches the true dual ball at the active face.

(b) Empirical probability of exact pattern recovery (18) on 5000 Monte Carlo trials for penalty scaling $\alpha \in [0, 20]$.

Figure 7: (a) The geometric representation of the dual ball $\partial f(0)$, the decomposable surrogate ball $\mathcal{B}_{\Phi^*}$, and the subdifferential $\partial f(\theta_0)$ at the active face. The vectors μ_1 , μ_2 , and μ_3 represent the generalized irrepresentability vector (28) for varying correlation levels ρ . (b) Empirical validation demonstrating the sharp phase transition at $\rho = 0.5$. For low correlation ($\rho_1 = 0.4$, green), the vector μ_1 lies strictly in the relative interior of the subdifferential, satisfying the strict irrepresentability condition $\mu_1 \in \text{ri}(\partial f(\theta_0))$ and guaranteeing asymptotic recovery. At the critical threshold ($\rho_2 = 0.5$, blue), μ_2 hits the boundary, capping the recovery probability at 0.5. For high correlation ($\rho_3 = 0.6$, red), μ_3 falls outside the subdifferential, strictly violating the condition and causing exact recovery to fail. (Panel (a) adapted from Paper I)

the minimizer over the set of positive definite matrices:

$$\widehat{\Theta}_n = \underset{\Theta \in \text{Sym}_+(p)}{\text{argmin}} \left\{ n^{-1} \sum_{i=1}^n \ell(X^{(i)}, \Theta) + n^{-1/2} \alpha \text{Pen}(\Theta) \right\}, \quad (29)$$

where $\ell(X, \Theta) = -\frac{1}{2} \log \det(\Theta) + \frac{1}{2} \text{tr}(\Theta X X^T)$ is the negative log-likelihood of the Gaussian model. The asymptotic error of the precision is given in (14). We consider composite penalties of the form:

$$\text{Pen}(\Theta) = f(\rho(\Theta)),$$

where f is a polyhedral norm acting on the off-diagonal entries, and ρ is a smooth transformation. When ρ is the identity map ($\rho(\Theta) = \Theta$), the formulation recovers classical methods. Standard examples include the Graphical Lasso penalty Friedman et al. (2008),

$$f(X) := \lambda \sum_{i < j} |X_{ij}|, \quad (30)$$

and the Graphical SLOPE penalty $f(X) := J_\lambda(X)$, where λ is a penalty sequence acting on the ordered absolute values of the strictly lower-triangular entries (studied in Paper III and Graczyk, Nakashima, et al. (2025)).

The well-known Graphical Lasso irrepresentability condition Ravikumar et al. (2011) is immediately recovered by substituting the Hessian $C = \frac{1}{2}(\Theta_0^{-1} \otimes \Theta_0^{-1})$ from Example 2.6 into the general asymptotic condition (26):

$$\|(P_{S^c} C P_S)(P_S C P_S)^+ \text{vec}(U_0)\|_\infty < 1, \quad (31)$$

where $U_0 = \text{sgn}(\text{oddiag}(\Theta_0))$, and P_S is the (vectorized) projection onto the subspace of symmetric matrices supported on the true edge set $S = \{(i, j) \in \{1, \dots, p\}^2 : (\Theta_0)_{ij} \neq 0\}$. This is succinctly written in terms of the block matrices as $\|C_{S^c S} C_{SS}^{-1} \text{vec}(U_0)_S\|_\infty < 1$.

One limitation of the standard Graphical Lasso is its dependence on data scaling; changing the units of measurement can result in different estimated graphs. Moreover, if a node X_i acts as a hub, being highly connected to other nodes, its diagonal precision entry Θ_{ii} becomes heavily inflated. Even standardizing the marginal variances ($\Sigma_{ii} = 1$) does not resolve this structural inflation. This creates a spillover effect: the inflated diagonal artificially increases the magnitude of the true off-diagonal entries, subjecting them to an excessive penalty and preventing the Graphical Lasso from reliably discovering the true hub structure in the graph (see Appendix 7.4).

To address this, the PCGLASSO method, originally proposed in Carter et al. (2024) and further developed in Paper IV, penalizes the partial correlation matrix directly:

$$\text{Pen}(\Theta) = f(\rho(\Theta)),$$

where f is the ℓ_1 penalty (30), and $\rho(\Theta)_{ij} = \Theta_{ij} / \sqrt{\Theta_{ii} \Theta_{jj}}$, for $i \neq j$, represents the negative partial correlation entries¹⁰. By design, the PCGLASSO is scale-invariant. Although the composite penalty $\text{Pen}(\Theta)$ is no longer polyhedral in Θ , reparametrizing with $Z := \rho(\Theta)$, yields the polyhedral penalty $f(Z)$. Moreover, the estimation error converges weakly and in pattern to an asymptotic distribution given in (48) (Section 7.6). By the chain rule, this yields a modified irrepresentability condition (Section 7.5):

$$\|(P_{S^c} \tilde{C} P_S)(P_S \tilde{C} P_S)^+ \text{vec}(U_0)\|_\infty < 1. \quad (32)$$

Here, $\tilde{C}$ is the Hessian of the reparametrized loss function $Z \mapsto \ell(X, \rho^{-1}(Z))$. By decomposing the true precision as $\Theta_0 = D_0 R_0 D_0$, the transformed Hessian $\tilde{C}$ is shown to depend entirely on the true negative partial correlation R_0 , fully decoupled from the diagonal scaling D_0 . As demonstrated in Paper IV, in networks with hub-like structures, this decoupling allows the PCGLASSO irrepresentability condition (32) to be less restrictive than its standard Graphical Lasso counterpart (31). This theoretical result is consistent with the empirical performance on genetic data; as visualized in Figure 8, PCGLASSO successfully recovers prominent hub structures in genetic network data, whereas the networks estimated by standard GLASSO and Cor-GLASSO remain highly diffuse.

3.5 Higher-order procedures

As discussed in previous sections, the irrepresentability condition remains a stringent requirement for pattern recovery, even in the low-dimensional asymptotic regime with p fixed

¹⁰We set the diagonal entries manually to $\rho(\Theta)_{ii} = 1 + \log(\Theta_{ii})$. This logarithmic augmentation of the diagonal has no effect on the penalty $f(\rho(\Theta))$, but makes ρ a smooth bijection with a smooth inverse.

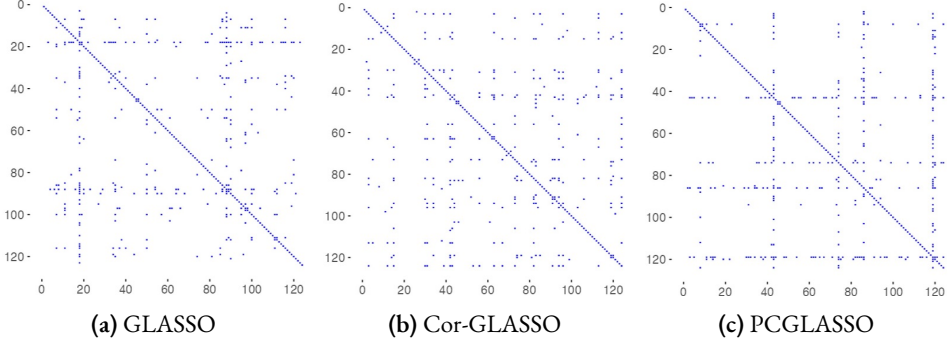


Figure 8: Comparison of estimated sparsity patterns (non-zero entries) of precision matrices on genetic data, tuned to maintain a 2% edge density across all three methods. Standard GLASSO and Cor-GLASSO (which uses $\Sigma_{ii} = 1$) produce highly diffuse networks due to the over-penalization of structurally inflated hub nodes. In contrast, by decoupling the penalty from the diagonal scale, PCGLASSO successfully recovers a structured topology featuring distinct prominent hubs (visible as dense vertical and horizontal alignments). (Adapted from Paper IV)

and $n \rightarrow \infty$. However, any $\sqrt{n}$ -consistent initial estimator $\widehat{\theta}_n^{(1)}$ converges to θ_0 at a rate of $O(n^{-1/2})$ and will therefore remain within a small neighborhood of the true pattern space U_{θ_0} . We can use this initial estimate to construct a second-order estimator $\widehat{\theta}_n^{(2)}$, which bypasses the irrepresentability condition and achieves asymptotic pattern recovery with high probability. In the context of sign recovery, the Adaptive Lasso Zou (2006) utilizes the initial estimate to adaptively reweight the coefficient penalization, achieving exact recovery of zeros. Another approach, first considered for SLOPE in Dupuis and Tardivel (2022) and revisited in greater generality in Graczyk et al. (2026) and our Paper I, is to apply proximal operator thresholding:

$$\widehat{\theta}_n^{(2)} := \text{Prox}_{n^{-1/2}\alpha f} \left(\widehat{\theta}_n^{(1)} \right), \quad (33)$$

where

$$\text{Prox}_f(\beta) := \underset{\xi \in \mathbb{R}^p}{\text{argmin}} \frac{1}{2} \|\beta - \xi\|_2^2 + f(\xi).$$

As shown in Paper I, the second step recovers the true pattern of θ_0 with high probability as $n \rightarrow \infty$ and $\alpha \rightarrow \infty$, provided the signal θ_0 is non-degenerate. This recovery is completely independent of whether the irrepresentability condition holds.

The estimate (33) is designed to recover the pattern; however, it introduces a large bias. To correct for this, one can perform a third step by minimizing the empirical loss $\mathcal{L}_n$ over the estimated pattern space $\widehat{U}$ induced by $\widehat{\theta}_n^{(2)}$. This is well-defined as long as the restricted loss is strictly convex over $\widehat{U}$.

For the linear model, if the restricted Hessian matrix $P_{\widehat{U}} X^T X P_{\widehat{U}}$ is invertible when restricted to $\widehat{U}$, then we can use restricted ordinary least squares (OLS) to estimate:

$$\widehat{\theta}_n^{(3)} := P_{\widehat{U}} (P_{\widehat{U}} X^T X P_{\widehat{U}})^+ X^T y. \quad (34)$$

If $\widehat{U}$ matches the true pattern space and $\dim(U) \ll p$, then (34) acts as an oracle low-dimensional maximum likelihood estimator, reducing the estimation error (see Figure 9).

Higher-order two- or three-step procedures are attractive due to their oracle-like properties. However, their theoretical success relies heavily on the accuracy of the initial estimate $\widehat{\theta}_n^{(1)}$. If the initial estimate is inaccurate, thresholding will fail to recover the underlying structure of the signal. This leads to the question of estimation guarantees for the initial estimate in the first step.

4 Estimation Error

While structural recovery is both theoretically and practically relevant, exact pattern recovery is often difficult to achieve. As established in the previous section, it hinges on the restrictive irrepresentability condition, which can fail even in the low-dimensional limit with p fixed and $n \rightarrow \infty$. This challenge is magnified in high-dimensional settings ($p \gg n$) and with increasingly complex signal structures. Even for a relatively small dimension of $p = 6$, the number of distinct patterns jumps from 729 for the Lasso, to 5,597 for the Fused Lasso, and to 423,857 for SLOPE (see Table 1). Furthermore, while higher-order thresholding methods offer a workaround, their success demands an accurate initial estimate. Ultimately, bounding the estimation error $\|\widehat{\theta}_n - \theta_0\|$ is a prerequisite for any pattern identification.

4.1 Simulation studies in the fixed-dimensional regime

To build intuition for how regularization reduces the estimation error, we use the asymptotic distribution of the scaled error $\widehat{u}$, which minimizes (13). Sampling $\widehat{u}$ provides a natural measure of asymptotic estimation accuracy. As $n \rightarrow \infty$, the root mean squared error $\text{RMSE}_n := (\mathbb{E}\|\sqrt{n}(\widehat{\theta}_n - \theta_0)\|_2^2)^{1/2}$ converges to

$$\text{RMSE} := (\mathbb{E}\|\widehat{u}\|_2^2)^{1/2}. \quad (35)$$

We present two simulation scenarios, one in the linear model (Paper I) and one in the Gaussian Graphical model (Paper III), that illustrate how regularizers exploit low-dimensional structures.

Simulation I:

Figure 9 demonstrates the estimation and pattern recovery performance of the Lasso, Fused Lasso, and SLOPE under a block diagonal correlation structure C (where $C_{13} = C_{24} = 0.8$, and $C_{ii} = 1$ for $1 \leq i \leq 4$). The results highlight that each regularizer is most effective when matched with its corresponding structural assumption: (a) Lasso outperforms when sparsity is the primary structure; (b) Fused Lasso targets consecutive clustering; and (c) SLOPE is the only penalty that captures non-consecutive clusters, which are invisible to the other penalties. The colored dots on the y -axis represent the MLE oracle estimates given the true pattern structure. These correspond to the theoretically attainable asymptotic error of the three-step procedure in (34).

Simulation II:

We consider a Gaussian Graphical Model $X \sim \mathcal{N}(0, \Sigma)$, with a 20×20 covariance Σ consisting of two 10×10 blocks given by a correlation matrix with all off-diagonal entries $\rho = 0.2$. Figure 10a illustrates the convergence of RMSE_n toward the asymptotic RMSE

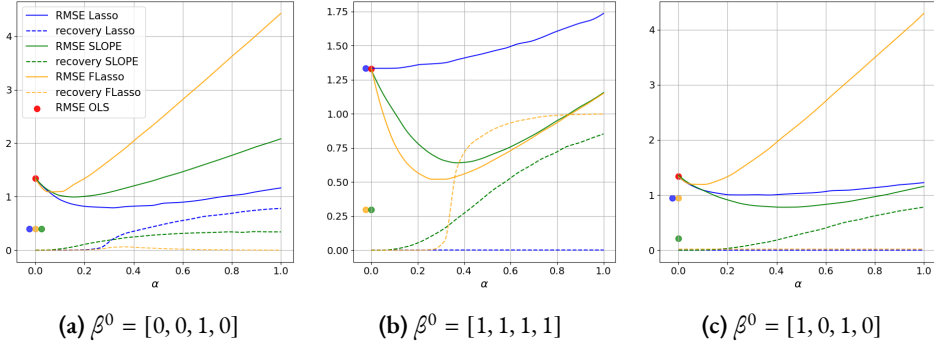


Figure 9: Comparing the asymptotic RMSE (35) for Lasso, Fused Lasso, and SLOPE (full lines), together with the probability of exact SLOPE pattern recovery—that is, correctly identifying all zeros and all clusters (dotted lines). (Adapted from Paper I).

(35) for the graphical Lasso and graphical SLOPE, following the asymptotics in Example 2.6. The penalty for SLOPE follows a normalized Benjamini-Hochberg sequence; for details, we refer to Paper III.

This simulation demonstrates how graphical SLOPE exploits clustering structures in the precision matrix to reduce estimation error compared to graphical Lasso penalization. This advantage for SLOPE becomes apparent even at a finite-sample size of $n = 1000$.

Figure 10b breaks this down further by showing the asymptotic graphical RMSE alongside the clustering error (RMSE-CL), measured on the strictly sub-diagonal entries. The RMSE-CL metric quantifies the distance from the true pattern space of the precision matrix Θ_0 ; it measures the error component introduced when the asymptotic estimation error $\hat{U}$ falsely breaks the clusters.¹¹ We observe that graphical SLOPE achieves a significantly lower RMSE-CL compared to the graphical Lasso. As $\alpha \rightarrow \infty$, the RMSE-CL for SLOPE drops toward 0, indicating perfect pattern recovery of the clustering structure (see Paper III for details).¹² In contrast, the relative underperformance of the graphical Lasso can be entirely explained by its inability to take advantage of the clusters in the true precision matrix.

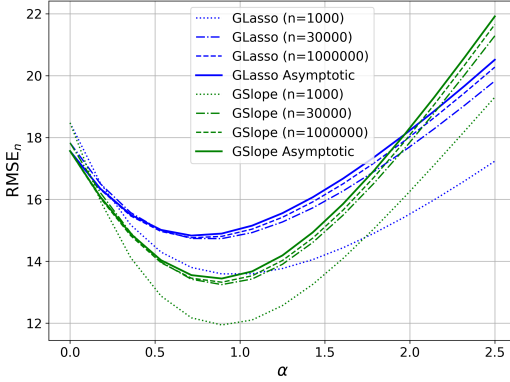
4.2 High-Dimensional estimation

While in the low-dimensional regime ($n \gg p$), a suitable regularizer can improve estimation accuracy by exploiting signal structure, in the high-dimensional regime ($p > n$), regularization becomes strictly necessary. Indeed, when $p > n$, the (convex) empirical loss $\mathcal{L}_n(\theta)$ inevitably contains directions along which it is perfectly flat. For example, under the quadratic loss in the linear model $\mathcal{L}_n(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2$, moving from θ along any nullspace direction $\Delta \in \text{Ker}(X)$ leaves the loss unchanged.

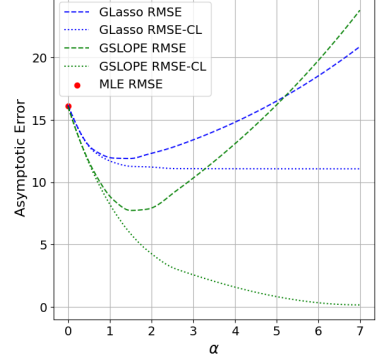
A suitably chosen regularizer f_n can stabilize this ill-posed problem by forcing the estimation error $\hat{\Delta} = \hat{\theta}_n - \theta_0$ into a restrictive cone of “well-behaved” directions. To successfully

¹¹The clustering structure of $\Theta_0 = \Sigma^{-1}$ is preserved under the inversion due to the block-diagonal design.

¹²We remark that RMSE-CL is exactly zero whenever $\hat{U}$ hits the true pattern space of the precision Θ_0 . This corresponds to exact pattern recovery as in (18).



(a) Convergence of RMSE_n toward the asymptotic RMSE given by (35) and (14), as a function of the penalty scaling α . Error is measured on lower-diagonal entries.



(b) Asymptotic RMSE (35) and the clustering error RMSE-CL as a function of the penalty scaling α . Error is measured on strictly sub-diagonal entries.

Figure 10: Asymptotic comparison of graphical Lasso and graphical SLOPE (Example 2.6) under a block diagonal covariance structure with clustered correlation values. (Adapted from Paper III)

control the estimation error, the loss function must exhibit strictly positive curvature along these permissible directions. In the linear model, this requirement is captured by the *restricted eigenvalue* condition Bickel et al. (2009) and van de Geer and Bühlmann (2009), which ensures the empirical covariance matrix $\frac{1}{n}X^T X$ acts as a strongly convex quadratic form over this specific restricted cone.

This geometric intuition was subsequently generalized to arbitrary convex loss functions through the *restricted strong convexity* (RSC) framework Negahban et al. (2012). At the core of this theory is the *decomposability* assumption on the regularizer (see (22)), a property held by penalties such as the Lasso, Group Lasso, and Nuclear norms Wainwright (2019).

In Paper v, we extend the theory to encompass *non-decomposable* regularizers. We achieve this by linking the polyhedral penalty f_n to its decomposable surrogate Φ , defined in (20). Conceptually, our main result (see (38)) establishes that the estimation error can be bounded by a combination of: the geometric complexity of the target subspace, the RSC curvature of the loss, and an approximation term quantifying how far the true signal deviates from perfect lower-dimensional structure. Other works on non-decomposable penalization Bach (2010), Bellec et al. (2018), Lee et al. (2015), and Obozinski and Bach (2012) operate under strict sparsity assumptions, whereas our surrogate approach allows regularizers to exploit more general lower-dimensional structures.

To state the general result, we first establish the necessary notation, visually summarized in Figure 11. To formulate our results in their full generality, we allow the true signal to be “imperfectly” lower-dimensional (i.e., not perfectly sparse or clustered), but simply “close enough” to a lower-dimensional subspace U . We denote this vaguely clustered true signal by θ^* , distinguishing it from a reference vector $\theta_0 \in U$. Conceptually, θ_0 is a vector of our choice that best captures the underlying low-dimensional structure of θ^* .

To keep the notation consistent with Paper v, we drop the subscript n in the mini-

mization formulation (1), writing $\widehat{\theta}$, $\mathcal{L}$, and f . We assume θ_0 is non-degenerate w.r.t. f , and we maintain the notation for the decomposable surrogate Φ as in (22). Note that Φ is defined relative to the reference vector θ_0 , not the true signal θ^* . Finally, we define $\Phi^s(\theta) := \frac{1}{2}(\Phi(\theta) + \Phi(-\theta))$ as the symmetrization of Φ . For SLOPE, $\Phi^s = \Phi$ due to the symmetry of the penalty (Figure 6); however, for penalties such as the Fused Lasso, $\Phi^s \neq \Phi$.¹³

In Paper v, we show that if θ^* lies in a permissible domain $\mathcal{D}_f(\theta_0)$ around the reference vector $\theta_0 \in U$ (see Figure 11), the regularization forces the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ into the star-shaped set:

$$C_\Phi(U; \theta^*) = \{\Delta \in \mathbb{R}^p \mid \Phi(\Delta_{U^\perp}) \leq 3\Phi(\Delta_U) + 4\Phi^s(\theta_{U^\perp}^*)\}, \quad (36)$$

provided that the dual norm of the gradient satisfies the bound $\Phi^*(\nabla \mathcal{L}(\theta^*)) \leq 1/2$. This implies that the penalty f prevents the error from escaping too far into the perpendicular space $U^\perp$. When the true signal lies exactly in the structured subspace ($\theta^* \in U$), the orthogonal projection $\theta_{U^\perp}^*$ vanishes, and this restricted set simplifies to a cone.

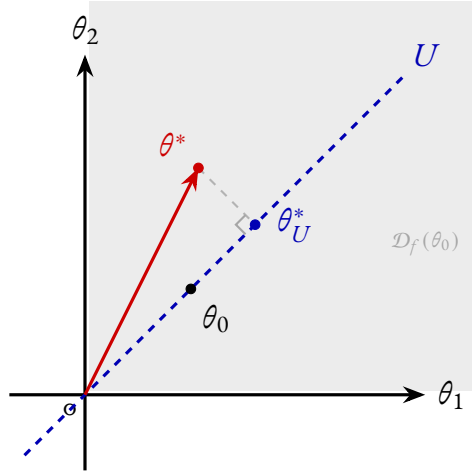


Figure 11: Geometric illustration of the vaguely clustered signal θ^* , its projection θ_U^* onto the SLOPE pattern space $U = \text{span}\{(1, 1)^T\}$, and the reference vector $\theta_0 \in U$. The shaded region indicates the permissible domain $\mathcal{D}_f(\theta_0)$ (the strictly positive quadrant) around the reference vector θ_0 , on which the regularizer f coincides with its decomposable surrogate Φ . (Adapted from Paper v)

To guarantee successful recovery, the loss $\mathcal{L}$ must satisfy the RSC condition with respect to this star-shaped set. Specifically, there must exist constants $\kappa > 0$ and $\tau \geq 0$ such that for every $\Delta \in C_\Phi(U; \theta^*)$:

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq \kappa \|\Delta\|^2 - \tau^2. \quad (37)$$

¹³In general, $\Phi^s = \Phi$ whenever $d_f(\theta_0)$ is symmetric around its projection onto $u_0 = P_U d_f(\theta_0)$.

If this RSC condition holds and the dual gradient is suitably bounded, we establish that the deterministic estimation error is controlled by the following high-dimensional bound:

$$\|\hat{\theta} - \theta^*\|^2 \leq \frac{9}{4\kappa^2} \Psi(U)^2 + \frac{1}{\kappa} \left(2\tau^2 + 4\Phi^s(\theta_{U^\perp}^*) \right). \quad (38)$$

Here, $\Psi(U) = \sup_{u \in U \setminus \{0\}} \frac{\Phi(u)}{\|u\|}$ is the *subspace compatibility constant* linking the regularizer to the error norm.¹⁴

This bound illustrates the mechanics of the surrogate framework: the error decreases with higher curvature κ and smaller tolerance τ , the surrogate Φ governs the bound through the compatibility constant $\Psi(U)$, and the term $\Phi^s(\theta_{U^\perp}^*)$ precisely quantifies the penalty paid for the signal θ^* deviating from the perfect low-dimensional structure U .

To see how our general theory recovers well-studied bounds Hastie et al. (2015) and Wainwright (2019), consider the following examples:

Example 4.1 (The Lasso bound). For $f(\theta) = \lambda\|\theta\|_1$, decomposability ensures the surrogate matches the penalty itself, $\Phi(\theta) = f(\theta)$. The star-shaped set in (36) reduces to $\{\Delta \mid \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 + 4\|\theta_{S^c}^*\|_1\}$, where S indicates the support of the reference vector θ_0 .

The bound (38) then becomes:

$$\|\hat{\theta} - \theta^*\|^2 \leq \frac{9}{4\kappa^2} (\sqrt{s}\lambda)^2 + \frac{1}{\kappa} \left(2\tau^2 + 4\lambda\|\theta_{S^c}^*\|_1 \right),$$

where s is the size of the support set (number of nonzeros in θ_0), and the subspace compatibility constant, given by the maximum value of the ratio $\lambda(\|u_S\|_1/\|u_S\|_2)$, is exactly $\sqrt{s}\lambda$.

Example 4.2 (SLOPE). The analysis for SLOPE is more technically involved, but our theory reduces it to the study of the surrogate regularizer Φ and its dual Φ^* (detailed fully in Paper v). The compatibility constant evaluates to $\sqrt{\sum_{j=1}^m s_j (\bar{\lambda}_{I_j})^2}$, where s_j denotes the sizes of the non-zero clusters, and $\bar{\lambda}_{I_j}$ is the average associated penalty according to the order of the clusters $(I_1, \dots, I_m)$ with respect to the reference vector θ_0 .¹⁵

4.3 The geometric bottleneck

As explained in the previous section, for the RSC condition to yield the high-dimensional bound (38), the error must be constrained to the star-shaped set $C_\Phi(U; \theta^*)$ (36). This assumes that the noise gradient is bounded in terms of the dual surrogate:

$$\Phi^*(\nabla \mathcal{L}(\theta^*)) \leq \frac{1}{2}. \quad (39)$$

In the standard theory for decomposable regularizers, the surrogate and its dual match the original regularizer, meaning $\Phi = f$ and $\Phi^* = f^*$. However, for non-decomposable

¹⁴The displayed bound differs from that in Theorem 1 of Negahban et al. (2012) by a factor of 4 in the denominator and the absence of the regularization parameter in the tolerance term τ^2 . We believe this discrepancy in the original text stems from a minor algebraic oversight in their proof. We elaborate on this, and demonstrate the sharpness of the corrected bound, in the Appendix of Paper v.

¹⁵Since the assortment of the penalty coefficients depends on the cluster orderings, this constant depends on the exact reference vector $\theta_0 \in U$, not merely on the subspace U .

regularizers such as SLOPE, the dual surrogate Φ^* can be significantly larger than f^* , making the gradient condition (39) more difficult to satisfy. Intuitively, in the presence of large clusters, the dual surrogate ball $\{\Phi^*(z) \leq 1\}$ is a heavily trimmed version of the original dual ball $\{f^*(z) \leq 1\}$ (see Figure 6b). This discrepancy does not necessarily point to a statistical limitation of non-decomposable regularizers such as SLOPE, but rather reveals a geometric challenge inherent to the theoretical surrogate framework.

More concretely, consider the linear model where the gradient is $z := \nabla \mathcal{L}(\theta^*) = X^T \varepsilon / n$. The gradient condition decomposes into the pattern space and its orthogonal complement: $\Phi^*(z) = \max\{\Phi_U^*(z_U), \Phi_{U^\perp}^*(z_{U^\perp})\} \leq 1/2$. The first component, $\Phi_U^*(z_U)$, behaves similarly to the ℓ_∞ -norm on each signal cluster and can be readily controlled by standard high-dimensional bounds. It is the perpendicular component, $\Phi_{U^\perp}^*(z_{U^\perp})$ (see (21)), that suffers from the trimming effect and creates the bottleneck.

However, this geometric bottleneck is mitigated when signals within clusters are highly correlated. For instance, if we consider a single positive cluster $\theta_0 = (1, \dots, 1)^T$ and assume the corresponding columns of X are identical, the perpendicular noise component strictly vanishes ($z_{U^\perp} = 0$). Under such structural conditions, this orthogonal noise component is eliminated, and this mitigating effect persists even under strong, albeit imperfect, correlations. This mechanism mathematically formalizes the empirical observation that SLOPE is highly effective in the presence of highly correlated features [Figueiredo and Nowak \(2016\)](#), bypassing the surrogate bottleneck. For the precise mathematical derivations of these surrogate components, we refer the reader to Paper v.

5 Open Problems and Future Directions

5.1 Partial Graphical Methods

The developed theory, expressed through the language of the decomposable surrogate Φ , relies on only mild assumptions regarding the statistical model, the loss $\mathcal{L}_n$, and the regularizer f_n . Consequently, it establishes a foundation for future methodological development.

For example, while Paper IV focuses strictly on sparsity and the PCGLASSO method [Carter et al. \(2024\)](#), one extension is to employ other penalties to exploit clustering structures within the partial correlation matrix. The irrepresentability condition (32) derived for the ℓ_1 penalty can be naturally generalized to:

$$\text{IRR}_f(R_0) = \Phi^* \left((P_{U^\perp} \tilde{C} P_U) (P_U \tilde{C} P_U)^+ \text{vec}(U_0) \right) < 1.$$

This generalized condition (further detailed in Appendix 7.5) directly accommodates penalties beyond the ℓ_1 -norm, and ensures asymptotic pattern recovery with high probability. For instance, applying the SLOPE penalty yields the asymptotic irrepresentability condition for a theoretical “PCGSLOPE” estimator designed to recover the clustering structure of the partial correlation matrix. While our surrogate provides this geometric groundwork, translating this condition into a fully realized methodology, including algorithmic implementation, finite-sample bounds, and empirical validation, remains an active area for future research.

5.2 Higher-Order Methods

As discussed in Section 3.5, thresholding an initial estimate via a proximal operator filter provides low-dimensional pattern recovery guarantees that bypass the stringent irrepresentability condition. While adaptive multi-step procedures for the Fused Lasso have been previously explored in the context of one-dimensional signal denoising Rinaldo (2009), and some existing work Dupuis and Tardivel (2022) and Graczyk et al. (2026) employs the SLOPE penalty as a proximal thresholding mechanism in the linear model, the broader use of clustering penalties remains largely unexplored. Specifically, proximal thresholding with the Fused Lasso or the more refined, non-degenerate Concavified Fused Lasso (Paper 1) has yet to be fully developed. Moreover, a possible future direction is to expand this fixed-dimensional theory into high-dimensional ($p \gg n$) settings, with the goal of obtaining probabilistic guarantees for exact pattern recovery and estimation bounds.

5.3 Phase transitions for SLOPE sequences in High Dimensions

Building on the high-dimensional estimation frameworks discussed in Theme III, a future direction involves understanding the geometric differences between SLOPE and the Lasso through the lens of their dual volumes. Although the gradient condition (39) in our formulation is in terms of the dual surrogate Φ^* , understanding bounds on the dual regularizer f^* is of independent interest. For instance, in Bellec et al. (2018), a bound of the form $f^*(\nabla \mathcal{L}(\theta^*)) < 1/2$ is obtained for SLOPE.

In Paper v, we considered the relative volume of the dual SLOPE ball and the dual Lasso ball given by the ℓ_∞ norm:

$$R(\lambda, p) := \frac{V(\{v : J_\lambda^*(v) \leq 1\})}{V(\{v : \|v\|_\infty \leq 1\})}.$$

For a linearly decaying OSCAR sequence Bondell and Reich (2008),

$$\lambda_i = 1 - \frac{i-1}{p}, \quad \text{for } i = 1, \dots, p,$$

we establish an exact polynomial decay:

$$R(\lambda, p) \asymp p^{-1/4}.$$

The proof follows from linking the SLOPE sequence to a quantile process, and relating the relative volume to the persistence of the Integrated Brownian Motion $\mathbb{P}(\int_0^T B(t)dt > 0)$, which is known to decay at the rate $T^{-1/4}$ Sinai (1992).

We further consider the Benjamini-Hochberg sequence Benjamini and Hochberg (1995):

$$\lambda_i = \Psi^{-1} \left(1 - q \cdot \frac{i}{2p} \right),$$

where Ψ denotes the standard normal CDF, with parameter q , commonly used in the literature and known for controlling the False Discovery Rate under orthogonal design Bogdan et al. (2015). We conjecture that for $q = 1$ the Gaussian volume ratio

$$R_N(\lambda, p) := \frac{\mathbb{P}(J_\lambda^*(Z) \leq 1)}{\mathbb{P}(\|Z\|_\infty \leq \lambda_1)}$$

exhibits the polynomial rate:

$$R_N(\lambda, p) \asymp p^{-1/6}.$$

Although we have strong numerical evidence (e.g., Figure 12) and propose a possible proof direction in Paper v, a formal proof remains an open problem.

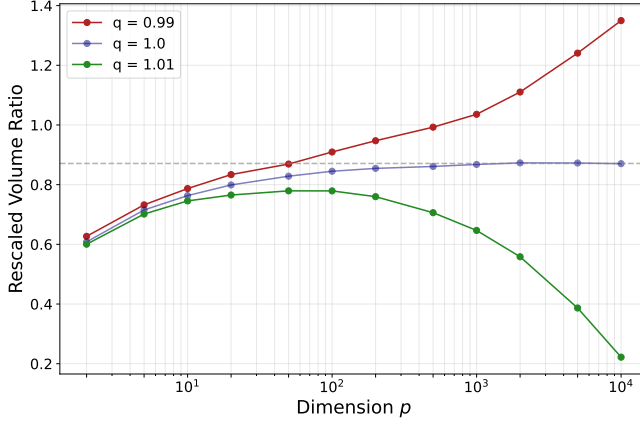


Figure 12: Phase transition of the Gaussian dual SLOPE volume for the Benjamini-Hochberg sequence. The y-axis shows the volume ratio $R_N(\lambda, p)$ rescaled by the conjectured $p^{1/6}$ rate. For the exact BH sequence ($q = 1$, blue), the rescaled ratio stabilizes at a constant, supporting the $R_N \asymp p^{-1/6}$ conjecture. Minor scaling deviations ($q = 0.99, 1.01$) trigger a phase transition into sub-critical stabilization or super-critical exponential collapse. (For $q > 1$, we take $\max\{\lambda_i, 0\}$). (Figure adapted from Paper v)

6 Summary of Papers

6.1 Paper I

While the asymptotic distribution of estimation error for Lasso-type estimators in the linear model is well-established when the dimension p is fixed and $n \rightarrow \infty$ Fu and Knight (2000) and Tibshirani et al. (2004), the convergence of the discrete structures (patterns) induced by these regularizers, such as sparsity or clustering, has remained unaddressed in the literature. Because patterns are highly sensitive to infinitesimal perturbations, weak convergence of the estimator does not guarantee that the underlying pattern will converge. We resolve this discrepancy by proving that the pattern of the estimator converges in distribution to a limiting pattern determined by a precise asymptotic formula. Using the Hausdorff distance as an appropriate mode of set convergence involving subdifferentials, we establish a framework that encompasses a broad class of non-differentiable regularizers including the Lasso, Generalized Lasso, SLOPE, and Elastic Net. Furthermore, the asymptotic formula fully characterizes the probability of exact pattern recovery, proving that recovery approaches certainty if and only if a regularizer-specific irrepresentability condition holds. Finally, we identify a critical degeneracy in the standard Fused Lasso under independent design, propose a Concavified Fused Lasso that resolves this limitation, and develop proximal thresholding procedures to bypass the irrepresentability bottleneck entirely.

6.2 Paper II

Paper II extends the pattern convergence of Paper I beyond the linear model to encompass generalized linear models and robust regression using Huber and quantile losses. Classical conditions for distributional convergence, such as Pollard's stochastic differentiability Pollard (1985), are insufficient to guarantee discrete pattern convergence because they fail to adequately control the fluctuations of the Taylor remainder. To solve this, we introduce Stochastic Lipschitz Differentiability (SLD), requiring that the remainder term of the loss expansion satisfies a tighter Lipschitz-type condition locally around the true parameter. We prove that SLD guarantees asymptotic convergence of patterns for general regularized M-estimators to an explicit asymptotic limit formula expressed in terms of the Hessian (Fisher Information) of the loss, the score covariance matrix, and the directional derivative of the regularizer. We show that SLD is satisfied under standard regularity assumptions on the loss. This enables the full characterization of asymptotic pattern recovery via the irrepresentability condition in general statistical models. Moreover, we verify that the SLD condition holds even for non-smooth loss functions, establishing pattern convergence for regularized Huber and quantile regression.

6.3 Paper III

Paper III applies the asymptotic theory developed in Papers I and II to the estimation of precision matrices in graphical models. Specifically, we analyze the graphical SLOPE (GSLOPE) estimator, which generalizes the graphical Lasso Friedman et al. (2008) by penalizing the off-diagonal entries of the precision matrix to induce simultaneous sparsity and clustering. Our asymptotic error and pattern recovery results extend beyond Gaussianity to elliptically distributed data, allowing for systematic comparison of GSLOPE and GLasso

(using Gaussian loss) and TSLOPE and TLasso (using multivariate t-distribution loss). Simulations demonstrate that SLOPE-based methods can in certain scenarios outperform Lasso-based methods by exploiting underlying clustering structures within the precision matrix. Furthermore, our simulations suggest that TSLOPE generally outperforms GSLOPE under heavy-tailed distributions. Finally, we apply these SLOPE-based graphical methods to identify empirical clustering structures in financial data.

6.4 Paper iv

Paper IV focuses on the Partial Correlation Graphical LASSO (PCGLASSO) Carter et al. (2024), a scale-invariant graphical method that penalizes the partial correlation matrix entries rather than the precision matrix itself. While this direct penalization can yield a more faithful representation of network topology, it induces a non-convex optimization problem that requires careful algorithmic development. Alongside addressing these computational challenges, we derive the exact irrepresentability condition for this estimator. Using the asymptotic formula, we demonstrate that compared to the standard graphical Lasso, the PCGLASSO irrepresentability condition depends exclusively on the true partial correlation matrix, decoupled from the diagonal precision scaling. We prove that PCGLASSO irrepresentability is significantly easier to satisfy when recovering networks with prominent hub structures, providing a rigorous theoretical justification for the method's superior empirical performance.

6.5 Paper v

Paper V extends the well-established high-dimensional theory of decomposable regularization Negahban et al. (2012) to accommodate non-decomposable polyhedral penalties, such as SLOPE and the Generalized Lasso. We achieve this by introducing the concept of the *Surrogate* regularizer, an associated decomposable penalty that locally matches the original regularizer around the true signal. Using this surrogate construction, we establish deterministic high-dimensional bounds on the estimation error. The theory relies on adapting both the restricted strong convexity (RSC) of the loss and the dual gradient bound to the specific geometry of the surrogate. Furthermore, it formally supports the empirically observed fact that SLOPE benefits heavily from strong correlation within signal clusters. The surrogate regularizer simultaneously captures exact structural pattern recovery conditions (including the irrepresentability condition) in polyhedral penalization. Additionally, by linking the linear SLOPE sequence to the Integrated Brownian Bridge, we establish an exact volume ratio decay rate between dual SLOPE and dual Lasso balls at $p^{-1/4}$ for the linear sequence, and conjecture a $p^{-1/6}$ Gaussian volume ratio decay rate for the Benjamini-Hochberg (BH) sequence.

7 Appendix

7.1 Convex Analysis Toolbox

For completeness, we recollect some definitions and facts from convex analysis, which appear throughout the thesis. These are standard and can be found in Hiriart-Urruty and

Lemaréchal (1993), Rockafellar and Wets (1998), and Rockafellar (2015). We shall denote by $\langle \cdot, \cdot \rangle$ the standard Euclidean inner product on $\mathbb{R}^p$. A function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is called *convex* on $\mathbb{R}^p$ if for any pair of points $(x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^p$ and all $\alpha \in (0, 1)$, it holds:

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2).$$

For the rest of the section let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a convex function. The *directional derivative* of f at a point x in direction u is defined as

$$f'(x; u) := \lim_{\varepsilon \downarrow 0} \frac{f(x + \varepsilon u) - f(x)}{\varepsilon}.$$

The limit always exists and $u \mapsto f'(x; u)$ is itself convex. The *subdifferential* of f at the point x is defined as the set:

$$\partial f(x) = \{v \in \mathbb{R}^p : f(y) \geq f(x) + \langle v, y - x \rangle \quad \forall y \in \mathbb{R}^p\}. \quad (40)$$

Geometrically, the subdifferential $\partial f(x)$ can be identified with the set of all slopes of the hyperplanes supporting f at the point x . The subdifferential $\partial f(x)$ is a convex, compact subset in $\mathbb{R}^p$.

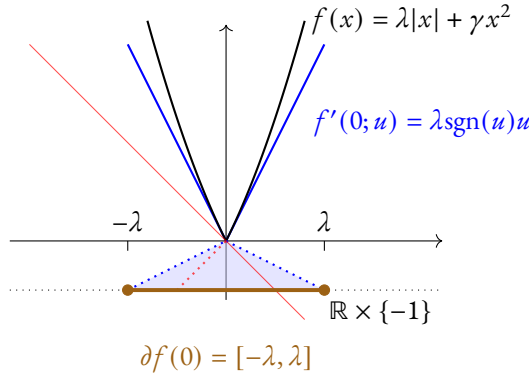


Figure 13: Directional derivative (blue) and subdifferential (bronze) of the Elastic Net at the origin. The subdifferential corresponds to the slopes of the supporting hyperplanes (red).

There is a close link between the directional derivative $f'(x; \cdot)$ and the subdifferential $\partial f(x)$. An equivalent way of defining the subdifferential is via the directional derivative

$$\partial f(x) = \{v \in \mathbb{R}^p : \langle v, u \rangle \leq f'(x; u) \quad \forall u \in \mathbb{R}^p\},$$

see Theorem 1.2.2 in Hiriart-Urruty and Lemaréchal (1993).

A convex function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is called *sublinear* if it is *positively homogeneous*, i.e.,

$$f(tx) = tf(x) \quad \forall x \in \mathbb{R}^p, t > 0.$$

This class includes all norms, hence in particular the Lasso, Generalized Lasso, SLOPE.

Any lower semicontinuous sublinear function admits the representation

$$f(x) = \sup_{v \in \partial f(0)} \langle v, x \rangle, \quad (41)$$

see Theorem 3.3.1 in Hiriart-Urruty and Lemaréchal (1993). Furthermore, the directional derivative $u \mapsto f'(x; u)$ is sublinear and can be written as

$$f'(x; u) = \sup_{v \in \partial f(x)} \langle v, u \rangle. \quad (42)$$

Equation (41) is a special case of (42), since for sublinear f , we have $f(u) = f'(0; u)$. For a closed convex set $C \subset \mathbb{R}^p$ containing the origin, the *gauge* of C is

$$\gamma_C(x) = \inf \{ \lambda > 0 : x \in \lambda C \},$$

and the *support function* of C is

$$\sigma_C(y) = \sup_{x \in C} \langle x, y \rangle.$$

The *polar set* of C is defined by

$$C^\circ = \{ y \in \mathbb{R}^p : \langle y, x \rangle \leq 1 \ \forall x \in C \} = \{ y : \sigma_C(y) \leq 1 \}.$$

Polarity is involutive, i.e., $(C^\circ)^\circ = C$ and the gauge and support functions are dual in the sense that

$$\gamma_C = \sigma_{C^\circ}, \quad \gamma_{C^\circ} = \sigma_C, \quad (43)$$

see Hiriart-Urruty and Lemaréchal (1993, Prop. 3.2.4–Cor. 3.2.5). Their subdifferentials admit the explicit forms

$$\partial \sigma_C(x) = \{ v \in C : \langle v, x \rangle = \sigma_C(x) \}, \quad (44)$$

and, by (41), for any sublinear function f ,

$$\partial f(x) = \{ v \in \partial f(0) : \langle v, x \rangle = f(x) \}. \quad (45)$$

Moreover, setting $C = \partial f(x)$, the directional derivative is $f'(x; u) = \sigma_C(u)$ by (42). Using (44) gives:

$$\partial_u f'(x; u) = \{ v \in \partial f(x) : \langle v, u \rangle = f'(x; u) \}.$$

7.2 Asymptotic Pattern Recovery of the 2D Fused Lasso

Here we illustrate the asymptotic pattern recovery for the Fused Lasso, as shown in Figure 5. We consider the 2D Fused Lasso penalty

$$f(\theta) = \lambda_1 |\theta_1| + \lambda_1 |\theta_2| + \lambda_2 |\theta_1 - \theta_2|,$$

under orthogonal design such that $C = C_\Delta = I_2$, and standard normal noise $W = (W_1, W_2)^T \sim \mathcal{N}(0, I_2)$. We seek to calculate the limiting pattern recovery probability for the true signal $\theta_0 = (1, 0)^T$.

The pattern space U_{θ_0} is the subspace spanned by θ_0 , which is simply the horizontal axis:

$$U_{\theta_0} = \text{span}\{(1, 0)^T\} = \{(u_1, 0)^T \mid u_1 \in \mathbb{R}\}.$$

Next, we characterize the active subdifferential $\partial f(\theta_0)$. At $\theta_0 = (1, 0)^T$, we have $\theta_1 > 0$, $\theta_2 = 0$, and $\theta_1 - \theta_2 > 0$. Evaluating the subdifferential components yields:

$$\partial f(\theta_0) = \lambda_1 \begin{pmatrix} 1 \\ [-1, 1] \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} \lambda_1 + \lambda_2 \\ \lambda_1[-1, 1] - \lambda_2 \end{pmatrix}.$$

Geometrically, this is a vertical line segment centered at $y = -\lambda_2$ with length $2\lambda_1$.

By the general result of equation (18), the asymptotic pattern recovery probability is given by:

$$\lim_{n \rightarrow \infty} \mathbb{P}[I_f(\hat{\theta}_n) = I_f(\theta_0)] = \mathbb{P}(W \in CU_{\theta_0} + \alpha \partial f(\theta_0)).$$

Substituting our geometric definitions into the set addition, we obtain:

$$U_{\theta_0} + \alpha \partial f(\theta_0) = \left\{ \begin{pmatrix} u_1 + \alpha(\lambda_1 + \lambda_2) \\ \alpha(\lambda_1 v - \lambda_2) \end{pmatrix} : u_1 \in \mathbb{R}, v \in [-1, 1] \right\}.$$

Since u_1 can take any value in $\mathbb{R}$, the first coordinate is entirely unconstrained. The event therefore depends exclusively on the second coordinate:

$$\begin{aligned} \mathbb{P}(W \in U_{\theta_0} + \alpha \partial f(\theta_0)) &= \mathbb{P}(W_2 \in \alpha[\lambda_1(-1) - \lambda_2, \lambda_1(1) - \lambda_2]) \\ &= \mathbb{P}(-\alpha(\lambda_1 + \lambda_2) \leq W_2 \leq \alpha(\lambda_1 - \lambda_2)). \end{aligned}$$

Let Ψ denote the cumulative distribution function of the standard normal distribution. Because $W_2 \sim \mathcal{N}(0, 1)$, the probability of exact pattern recovery reduces precisely to:

$$\Psi(\alpha(\lambda_1 - \lambda_2)) - \Psi(-\alpha(\lambda_1 + \lambda_2)).$$

From this explicit formula, the phase transition behavior described in Figure 5 becomes immediately apparent. As $\alpha \rightarrow \infty$, the rightmost term $\Psi(-\alpha(\lambda_1 + \lambda_2))$ always vanishes. The limit is entirely determined by the sign of $(\lambda_1 - \lambda_2)$. As illustrated in Figure 5b:

- Non-degenerate (Green, $\lambda_1 > \lambda_2$): The argument diverges to $+\infty$, and the recovery probability converges to 1.
- Degenerate Boundary (Blue, $\lambda_1 = \lambda_2$): The argument is exactly 0, and the recovery probability is capped at $\Psi(0) = 0.5$.
- Strictly Degenerate (Red, $\lambda_1 < \lambda_2$): The argument diverges to $-\infty$, and the recovery probability collapses to 0.

7.3 Complexity of Regularization

As we have summarized in Table 1, the Lasso, Fused Lasso, and SLOPE regularizers exhibit patterns of increasing complexity. As the dimension p increases, the Lasso penalty on $\mathbb{R}^p$ has $N_p = 3^p$ distinct patterns.

The number of distinct patterns for the Fused Lasso (9) grows as $N_p = O(a^p)$, where $a = \frac{5+\sqrt{17}}{2} \approx 4.56$. To see this, write $N_p = Z_p + W_p$, where Z_p and W_p denote the number of distinct Fused Lasso patterns where the last coefficient satisfies $\theta_p = 0$ and $\theta_p \neq 0$, respectively. This gives the recursive relationship

$$\begin{aligned} Z_p &= Z_{p-1} + W_{p-1} = N_{p-1} \\ W_p &= 2Z_{p-1} + 4W_{p-1}. \end{aligned}$$

Summing the two equations yields the recursion

$$N_p = 5N_{p-1} - 2N_{p-2}.$$

The initial values $N_1 = 3$ and $N_2 = 13$ determine the sequence, see OEIS Foundation Inc. (2026a). Solving this linear recurrence explicitly gives

$$N_p = \frac{1}{2\sqrt{17}} \left[(1 + \sqrt{17}) \left(\frac{5 + \sqrt{17}}{2} \right)^p + (\sqrt{17} - 1) \left(\frac{5 - \sqrt{17}}{2} \right)^p \right].$$

The number of SLOPE patterns is the number of the lower-dimensional faces of a signed permutahedron. This is given by the elements in the Coxeter Complex B_p OEIS Foundation Inc. (2026b), which grows as

$$N_p \sim p! \frac{\sqrt{2}}{4} \left(\frac{2}{\ln 2} \right)^{p+1}.$$

This can be obtained by considering the exponential generating function $B(x) = \frac{e^x}{2-e^{2x}}$ and the simple unique pole $x_0 = \frac{\ln 2}{2}$ on the boundary of the convergence radius of $B(x)$. By standard results in analytic combinatorics (e.g., Theorem IV.10 Flajolet and Sedgewick (2009), Section 5.2 Wilf (2005)), this gives the rate $N_p \sim C p! (x_0)^{-p}$, with $C = -\text{Res}(B, x_0)/x_0$. The residue is simply $\text{Res}(B, x_0) = \frac{g(x_0)}{b'(x_0)} = \frac{\sqrt{2}}{-4}$, where the generating function is written as $B = g/b$.

7.4 Precision Matrix Inflation

Let $X = (X_1, \dots, X_p) \sim \mathcal{N}(0, \Sigma)$, and let $\Theta = \Sigma^{-1}$. For a subset of indices $\mathcal{A}$, and $\mathcal{B}$ being the complement indices, denote

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} \quad \Theta = \begin{pmatrix} \Theta_{AA} & \Theta_{AB} \\ \Theta_{BA} & \Theta_{BB} \end{pmatrix}$$

Considering the residual vector $\varepsilon_A = X_A - \mathbb{E}[X_A | X_B]$, we have by a standard Schur complement argument

$$\text{Cov}(\varepsilon_A) = \text{Var}(X_A | X_B) = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA} = \Theta_{AA}^{-1}.$$

For $\mathcal{A} = \{i\}$ this gives

$$\text{Var}(\varepsilon_i) = \frac{1}{\Theta_{ii}}.$$

Hence, Θ_{ii} is the inverse of the residual variance after regressing away the influence of the other variables.

For $\mathcal{A} = \{i, j\}$ we obtain the partial correlation

$$\text{Corr}(\varepsilon_i, \varepsilon_j) = \frac{-\Theta_{ij}}{\sqrt{\Theta_{ii}}\sqrt{\Theta_{jj}}},$$

which measures the association between the two nodes after stripping away the linear influence of all other nodes.

Decomposing the total variance $\text{Var}(X_i) = \Sigma_{ii}$ into explained variance $\text{Var}(\mathbb{E}[X_i | X_{-i}])$ and the residual variance $\text{Var}(\varepsilon_i)$, and rearranging gives $\text{Var}(\varepsilon_i) = \Sigma_{ii}(1 - R_i^2)$, where R_i^2 is the relative explained variance.¹⁶ Thus, the diagonal precision entry is

$$\Theta_{ii} = \frac{1}{\Sigma_{ii}(1 - R_i^2)}.$$

If X_i is heavily connected to other nodes X_j (i.e., it is a hub), the relative explained variance R_i^2 approaches 1. Consequently, the Θ_{ii} value is inflated by the structural influence of its neighbors. By rearranging the partial correlation formula, we see that the off-diagonal entry is defined as

$$\Theta_{ij} = -\text{Corr}(\varepsilon_i, \varepsilon_j)\sqrt{\Theta_{ii}\Theta_{jj}}.$$

The inflation of the precision diagonal Θ_{ii} increases the magnitude of Θ_{ij} . Because standard GLASSO applies an ℓ_1 penalty to the absolute value $|\Theta_{ij}|$, these inflated hub entries incur a disproportionate penalty in the objective function. To minimize the overall objective, the optimization procedure tends to shrink these specific entries, which can incorrectly force them to 0, and destroy the hub. Even after standardizing to $\Sigma_{ii} = 1$, the inflation persists because it is driven by the conditional structure through R_i^2 , not just by marginal scale.

In contrast, PCGLASSO penalizes the partial correlation $\text{Corr}(\varepsilon_i, \varepsilon_j)$ directly. By standardizing the penalty against the inflated diagonals, it accounts for this artifact and can discover hub structures more reliably than GLASSO.

7.5 Partial Correlation asymptotic pattern recovery

Let f be a polyhedral norm on the off-diagonal entries of a symmetric matrix, such as the Lasso penalty $f(X) := \lambda \sum_{i>j} |X_{ij}|$, which is the ℓ_1 norm on the lower triangular indices, or the SLOPE penalty $f(X) = J_\lambda(X)$. Consider the penalty

$$\begin{aligned} \text{Pen}(\Theta) &= f(\text{Corr}(\Theta)) \\ &= f(\text{Corr}(\Theta) + \log(\text{diag}(\Theta))) \\ &= f(\rho(\Theta)), \end{aligned}$$

¹⁶Note that the decomposition of the variance follows from the fact that $X_i \rightarrow \mathbb{E}[X_i | X_{-i}]$ is a projection in the Hilbert space of centered square-integrable random variables $\mathcal{H} := \{X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) \mid \mathbb{E}[X] = 0\}$, equipped with the inner product $\text{Cov}(X, Y) = \mathbb{E}[XY]$. The decomposition $\text{Var}(X_i) = \text{Var}(X_i | X_{-i}) + \text{Var}(\varepsilon_i)$ is the Pythagorean Theorem in $\mathcal{H}$.

where we have augmented the penalty with a diagonal term. This does not change the penalty since f acts only on off-diagonal elements, but makes the inner transformation

$$\rho(\Theta) = \text{Corr}(\Theta) + \log(\text{diag}(\Theta)) = R + \log(D^2) \quad (46)$$

invertible, with derivative $\text{vec}(\rho'(\Theta_0; U)) = \rho'_{\Theta_0} \text{vec}(U)$. The estimators for $\widehat{R}_n$ and $\widehat{D}_n$ satisfy the decomposition $\widehat{\Theta}_n = \widehat{D}_n \widehat{R}_n \widehat{D}_n$. The information is jointly encapsulated in

$$\rho(\widehat{\Theta}_n) = \widehat{R}_n + \log(\widehat{D}_n^2),$$

where $\widehat{\Theta}_n$ minimizes (29). Since the error $\widehat{U}_n = \sqrt{n}(\widehat{\Theta}_n - \Theta_0)$ converges in distribution to

$$\widehat{U} = \underset{U \in \text{Sym}(p)}{\text{argmin}} \left\{ \frac{1}{2} \text{vec}(U)^T C \text{vec}(U) - \text{vec}(U)^T W + \alpha(f \circ \rho)'(\Theta_0; U) \right\}, \quad (47)$$

where $C = \frac{1}{2}(\Theta_0^{-1} \otimes \Theta_0^{-1})$, $W \sim \mathcal{N}(0, C)$ and the directional derivative simplifies by the chain rule as

$$\begin{aligned} (f \circ \rho)'(\Theta_0; U) &= f'(\rho(\Theta_0); \rho'(\Theta_0; U)) \\ &= f'(R_0; \rho'(\Theta_0; U)), \end{aligned}$$

where the $\log(D_0^2)$ term in $\rho(\Theta_0)$ falls out since the penalty is purely off-diagonal and thus invariant to diagonal shifts. By the delta method

$$\text{vec}(\sqrt{n}(\rho(\widehat{\Theta}_n) - \rho(\Theta_0))) \xrightarrow{d} \rho'_{\Theta_0} \text{vec}(\widehat{U}).$$

Setting $\widehat{V} := \rho'(\Theta_0; \widehat{U})$, we have $\text{vec}(\widehat{V}) = \rho'_{\Theta_0} \text{vec}(\widehat{U})$. Using the inverse substitution $\text{vec}(U) = (\rho'_{\Theta_0})^{-1} \text{vec}(V)$ in (47), we get the asymptotic distribution

$$\widehat{V} = \underset{V \in \text{Sym}(p)}{\text{argmin}} \left\{ \frac{1}{2} \text{vec}(V)^T \widetilde{C} \text{vec}(V) - \text{vec}(V)^T \widetilde{W} + \alpha f'(R_0; V) \right\}, \quad (48)$$

where $\widetilde{W} \sim \mathcal{N}(0, \widetilde{C})$. An explicit expression for the Hessian $\widetilde{C}$ is obtained by Lemma 7.2:

$$\begin{aligned} \widetilde{C} &= ((\rho'_{\Theta_0})^{-1})^T C (\rho'_{\Theta_0})^{-1} \\ &= \frac{1}{2} (T_0^{-1})^T (R_0^{-1} \otimes R_0^{-1}) T_0^{-1}, \end{aligned}$$

with $\rho'_{\Theta_0} = T_0(D_0^{-1} \otimes D_0^{-1})$ and

$$T_0^{-1} = I_{p^2} + \frac{1}{2} P_{od} (I_p \otimes R_0 + R_0 \otimes I_p) P_d,$$

where P_d and P_{od} are the diagonal and off-diagonal projection operators satisfying $P_d \text{vec}(X) = \text{vec}(\text{diag}(X))$, and $P_{od} \text{vec}(X) = \text{vec}(\text{oddiag}(X))$ for any $X \in \mathbb{R}^{p \times p}$.

In particular, the rescaled error of sequence of estimators $\widehat{Z}_n := \rho(\widehat{\Theta}_n) = \widehat{R}_n + \log(\widehat{D}_n^2)$ converges in distribution and in pattern to the minimizer $\widehat{V}$ of (48), for details see Subsection 7.6. Importantly, the objective of the asymptotic error (48) does not depend on the diagonal scaling D_0 , and is entirely determined by the negative partial correlation R_0 and the penalty f .

From (48) one obtains the general irrepresentability condition for graphical methods:

$$\text{IRR}_f(R_0) = \Phi^* \left((P_{U^\perp} \widetilde{C} P_U) (P_U \widetilde{C} P_U)^+ (U_0) \right) < 1, \quad (49)$$

where $U_0 = P_U \partial f(R_0)$, and P_U is the (vectorized) projection onto the pattern space:

$$U_{R_0} = \text{span}\{U \in \text{Sym}(p) : I_f(U) = I_f(R_0)\}. \quad (50)$$

The Φ^* is the dual surrogate regularizer associated to the polyhedral penalty f .¹⁷ If (49) is satisfied, then the pattern is recovered with high probability, i.e.:

$$\lim_{n \rightarrow \infty} \mathbb{P}[I_f(\widehat{R}_n) = I_f(R_0)] \geq 1 - e^{-c\alpha^2},$$

for some $c > 0$. If (49) is violated, then the above probability is bounded from above by $\frac{1}{2}$.

Example 7.1. For PCGLASSO, when $f(X) = \lambda \sum_{i>j} |X_{ij}|$, the estimator of the partial correlation matrix $\widehat{R}_n$ recovers the sign pattern of the true partial correlation R_0 if

$$\text{IRR}_{\text{PCGL}}(R_0) = \|(P_{S^c} \widetilde{C} P_S) (P_S \widetilde{C} P_S)^+ (U_0)\|_\infty < 1,$$

where $U_0 = \text{oddiag}(\text{sgn}(R_0))$, and P_S is the vectorized projection onto the pattern space¹⁸, given by the subspace symmetric matrices with support in $S = \{(i, j) \in \{1, \dots, p\}^2 : (R_0)_{ij} \neq 0\}$:

$$U_{R_0} = \{U \in \text{Sym}(p) : U_{ij} = 0 \quad \forall (i, j) \notin S\}.$$

We remark that in Paper IV, instead of $\widetilde{C}$, we use $\widetilde{\Gamma} = (T_0^{-1})^T (R_0^{-1} \otimes R_0^{-1})$. This does not change the irrepresentability condition. In fact, the cross-operator $\Pi_{S^c|S}$ does not change when $\widetilde{C}$ is replaced with $\widetilde{\Gamma}$.

Lemma 7.2. Let $\Theta_0 = D_0 R_0 D_0$. The directional derivative $\text{vec}(\rho'(\Theta_0; U)) = \rho'_{\Theta_0} \text{vec}(U)$ of $\rho(\Theta) := \text{Corr}(\Theta) + \log(\text{diag}(\Theta)) = R + \log(D^2)$, is given by

$$\rho'_{\Theta_0} = T_0 (D_0^{-1} \otimes D_0^{-1}),$$

where the operator T_0 and its inverse T_0^{-1} are

$$\begin{aligned} T_0 &= I_{p^2} - \Delta_0 \\ T_0^{-1} &= I_{p^2} + \Delta_0, \end{aligned}$$

¹⁷The regularizer f is assumed to be non-degenerate w.r.t. R_0 . This is the case for any R_0 for both Lasso and SLOPE penalization.

¹⁸The pattern is identified with $I_f(X) = \text{sgn}(\text{oddiag}(X))$, and by (50) the pattern can also be written as $U_{R_0} = \text{span}\{U \in \text{Sym}(p) : \text{sgn}(\text{oddiag}(U)) = \text{sgn}(\text{oddiag}(R_0))\}$.

with

$$\begin{aligned}\Delta_0 &= \frac{1}{2}(I_p \otimes R_0 + R_0 \otimes I_p)P_d - P_d \\ &= \frac{1}{2}P_{od}(I_p \otimes R_0 + R_0 \otimes I_p)P_d.\end{aligned}$$

Here P_d is the diagonal projection operator satisfying $P_d \text{vec}(X) = \text{vec}(\text{diag}(X))$ for any $X \in \mathbb{R}^{p \times p}$, and P_{od} is the off-diagonal projection operator satisfying $P_{od} \text{vec}(X) = \text{vec}(\text{oddiag}(X))$. The matrix Δ_0 is nilpotent of degree 2, i.e., $\Delta_0^2 = 0$.

Proof. We briefly sketch the proof here. The vectorization of the directional derivative $\text{Corr}'(\Theta_0; U)$ is shown in Paper IV to be $(T_0 - P_d)(D_0^{-1} \otimes D_0^{-1}) \text{vec}(U)$. The diagonal derivative $(\log \circ \text{diag})'(\Theta_0; U) = \log'(D_0^2; \text{diag}(U))$ is in vectorized form $P_d(D_0^{-1} \otimes D_0^{-1}) \text{vec}(U)$, which follows similarly by expanding the logarithm $\log(D_0^2 + \varepsilon \text{diag}(U)) = \log(D_0^2) + \varepsilon D_0^{-2} \text{diag}(U) + o(\varepsilon)$. Adding the two components gives the derivative: $\rho'_{\Theta_0} = T_0(D_0^{-1} \otimes D_0^{-1})$.

To derive the formula for the inverse T_0^{-1} , we use the identity

$$\frac{1}{2}P_d(I_p \otimes R_0 + R_0 \otimes I_p)P_d = P_d.$$

Applying this, one can expand Δ_0^2 to show that $\Delta_0^2 = 0$. It follows that $(I_{p^2} - \Delta_0)(I_{p^2} + \Delta_0) = I_{p^2} - \Delta_0^2 = I_{p^2}$. $\square$

7.6 Pattern convergence for PCG methods

The map $\rho(\Theta) = \text{Corr}(\Theta) + \log(\text{diag}(\Theta)) = R + \log(D^2)$ in (46) is a C^∞ diffeomorphism between two $p(p+1)/2$ -dimensional open smooth manifolds in $\text{Sym}(p)$. The domain consists of the positive definite matrices $\text{Sym}_+(p)$, and the target space consists of all matrices where the off-diagonal part is induced by a correlation matrix. Precisely:

$$\mathcal{Z} = \rho(\text{Sym}_+(p)) = \{Z \in \text{Sym}(p) \mid Z - \text{diag}(Z) + I_p \in \text{Sym}_+(p)\}.$$

The inverse transformation $\rho^{-1} : \mathcal{Z} \rightarrow \text{Sym}_+(p)$ is C^∞ and is given by

$$\rho^{-1}(Z) = \underbrace{\exp\left(\frac{1}{2}(\text{diag}(Z) - I_p)\right)}_D \underbrace{\left(Z - \text{diag}(Z) + I_p\right)}_R \underbrace{\exp\left(\frac{1}{2}(\text{diag}(Z) - I_p)\right)}_D.$$

The minimizer of (29) is $\hat{\Theta}_n = \rho^{-1}(\hat{Z}_n)$, where $\hat{Z}_n := \hat{R}_n + \log(\hat{D}_n^2)$. The augmented partial correlation estimator minimizes

$$\hat{Z}_n = \hat{R}_n + \log(\hat{D}_n^2) = \underset{Z \in \mathcal{Z}}{\text{argmin}} \left\{ n^{-1} \sum_{i=1}^n \ell(X^{(i)}, \rho^{-1}(Z)) + n^{-1/2} \alpha f(Z) \right\}, \quad (51)$$

where $\ell(X, \Theta) = -\frac{1}{2} \log \det(\Theta) + \frac{1}{2} \text{tr}(\Theta X X^T)$. It is shown in Bogdan et al. (2025) that the standard loss $\Theta \mapsto \ell(X, \Theta)$ satisfies the sufficient conditions for pattern convergence (16)

(formally stated in Corollary 3.4 Paper II). Since $\rho : \Theta \mapsto Z$ is a smooth diffeomorphism, these conditions remain valid for the transformed loss $Z \mapsto \ell(X, \rho^{-1}(Z))$.

In particular, $\widehat{Z}_n$ is consistent for $Z_0 = R_0 + \log(D_0^2)$ ¹⁹. This means that $\sqrt{n}(\widehat{Z}_n - Z_0)$ converges in distribution and in pattern to the minimizer $\widehat{V}$ in (48).

References

- Bach, F. (2010). Structured sparsity-inducing norms through submodular functions. *Advances in Neural Information Processing Systems (NIPS)* 23, 1171–1179.
- Bellec, P. C., Lecué, G., & Tsybakov, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B), 3603–3642. <https://doi.org/10.1214/17-AOS1670>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.
- Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics*, 37, 1705–1732.
- Bogdan, M., Dupuis, X., Graczyk, P., Kołodziejek, T., Stankiewicz, J., Tardivel, P., & Wilczyński, M. (2022). Pattern recovery by SLOPE. *arXiv preprint arXiv:2203.12086*.
- Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., & Candès, E. (2015). Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3), 1103.
- Bogdan, M., Chojecki, A., Hejný, I., Kołodziejek, B., & Wallin, J. (2025). Identifying network hubs with the partial correlation graphical lasso. *arXiv preprint arXiv:2508.12258*.
- Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1), 115–123.
- Brodie, J., Daubechies, I., De Mol, C., Giannone, D., & Loris, I. (2009). Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30), 12267–12272.
- Candès, E. J., Romberg, J., & Tao, T. (2006a). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2), 489–509.
- Candès, E. J., Romberg, J. K., & Tao, T. (2006b). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8), 1207–1223.
- Carter, J. S., Rossell, D., & Smith, J. Q. (2024). Partial correlation graphical LASSO. *Scandinavian Journal of Statistics*, 51(1), 32–63.
- Dupuis, X., & Tardivel, P. J. (2022). Proximal operator for the sorted l1 norm: Application to testing procedures based on slope. *Journal of Statistical Planning and Inference*, 221, 1–8.

¹⁹Consistency follows by the classical argmax argument for uniformly tight sequences (e.g., Corollary 3.2.3 in van der Vaart and Wellner (1996)). The uniform tightness of $\widehat{Z}_n$ is guaranteed by uniform coercivity of the sequence of transformed empirical log-likelihoods over Z . The uniform convergence of the objective functions on compacta to the population risk $Z \mapsto \mathbb{E}[\ell(X, \rho^{-1}(Z))]$, which has a unique minimum at Z_0 , is preserved under the continuous mapping ρ^{-1} .

- Figueiredo, M., & Nowak, R. (2016, September). Ordered weighted ℓ_1 -regularized regression with strongly correlated covariates: Theoretical aspects. In A. Gretton & C. C. Robert (Eds.), *Proceedings of the 19th international conference on artificial intelligence and statistics* (pp. 930–938, Vol. 51). PMLR.
- Flajolet, P., & Sedgewick, R. (2009). *Analytic combinatorics*. Cambridge University Press.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), 432–441.
- Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5), 1356–1378.
- Graczyk, P., Nakashima, H., et al. (2025). From graphical lasso to atomic norms: High-dimensional pattern recovery. *arXiv preprint arXiv:2506.13353*.
- Graczyk, P., Schneider, U., Skalski, T., & Tardivel, P. (2023). Pattern recovery in penalized and thresholded estimation and its geometry. *arXiv preprint arXiv:2307.10158*.
- Graczyk, P., Schneider, U., Skalski, T., & Tardivel, P. (2026). A unified framework for pattern recovery in penalized and thresholded estimation and its geometry. *Journal of Optimization Theory and Applications*, 208(1), 49.
- Hastie, T. (2009). The elements of statistical learning: Data mining, inference, and prediction.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143), 8.
- Hiriart-Urruty, J., & Lemaréchal, C. (1993). *Convex analysis and minimization algorithms I: Fundamentals* (Vol. 305). Springer-Verlag.
- Jia, J., & Yu, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20(2), 595–611.
- Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.
- Lee, J. D., Sun, Y., & Taylor, J. E. (2015). On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics*, 9, 608–642. <https://doi.org/10.1214/15-EJS1013>
- Lewis, A. S. (2002). Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3), 702–725. <https://doi.org/10.1137/S1052623401387623>
- Negahban, S. N., Ravikumar, P., Wainwright, M. J., & Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4), 538–557.
- Obozinski, G., & Bach, F. (2012). *Convex relaxation for combinatorial penalties* (tech. rep.) (arXiv:1205.1240). HAL / INRIA.
- OEIS Foundation Inc. (2026a). Entry A052984 in The On-Line Encyclopedia of Integer Sequences [Accessed: 2026-03-25].
- OEIS Foundation Inc. (2026b). Entry A080253 in The On-Line Encyclopedia of Integer Sequences [Accessed: 2026-03-24].
- Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory*, 1(3), 295–313. <https://doi.org/10.1017/S0266466600011233>
- Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935–980. <https://doi.org/10.1214/11-EJS631>
- Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B), 2922–2952.

- Rockafellar, R. T., & Wets, R. J. (1998). *Variational analysis*. Springer.
- Rockafellar, R. T. (2015). *Convex analysis: (pms-28)*. Princeton University Press.
- Schneider, U., & Tardivel, P. (2022). The geometry of uniqueness, sparsity and clustering in penalized estimation. *Journal of Machine Learning Research*, 23(331), 1–36.
- Sinai, Y. G. (1992). Distribution of some functionals of the integral of a random walk. *Theoretical and Mathematical Physics*, 90(3), 219–241.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2004). Sparsity and Smoothness Via the Fused Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1), 91–108. <https://doi.org/10.1111/j.1467-9868.2005.00490.x>
- Vaïter, S., Peyré, G., & Fadili, J. (2017). Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory*, 64(3), 1725–1737.
- van de Geer, S. A., & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3, 1360–1392. <https://doi.org/10.1214/09-EJS506>
- van der Vaart, A. W., & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes: With applications to statistics* (pp. 16–28). Springer New York. https://doi.org/10.1007/978-1-4757-2545-2_3
- van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.
- Wilf, H. S. (2005). *Generatingfunctionology*. CRC press.
- Yuan, M., & Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1), 143–159. <https://doi.org/10.1111/j.1467-9868.2007.00543.x>
- Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90), 2541–2563.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429. <https://doi.org/10.1198/016214506000000735>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2), 301–320.

Papers

About the author

Ivan Hejný conducted his doctoral studies at the Department of Statistics from 2021 to 2026 under the supervision of Jonas Wallin and Małgorzata Bogdan. His research focuses on regularization methods and their theoretical capacity to discover low-dimensional structures in statistical models. Comprising five research papers, this thesis contributes to the general mathematical foundations of pattern recovery, the asymptotic behavior of structural patterns, and their application to graphical models and extensions into high-dimensional estimation.

