Unveiling low-dimensional patterns induced by convex non-differentiable regularizers

Małgorzata Bogdan
University of Wroclaw (Poland), Lund University (Sweden)

Kongres Statystyki Polskiej 2024

2nd of July, 2024

Outline

- Pattern definition
- L_1 norm : support recovery vs separability of signal and noise
- SLOPE: FDR control and clustering properties
- Pattern recovery for general classifiers in the low dimensional setup

Penalization by polyhedral gauges

$$Y = X\beta + \varepsilon,$$

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

Penalization by polyhedral gauges

$$Y = X\beta + \varepsilon,$$

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

$$\hat{\beta} = \mathop{\mathrm{Arg\,min}}_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda \mathrm{pen}(b), \qquad (1)$$

where pen is a real-valued polyhedral gauge, i.e. a non-negative and positively homogeneous convex function that vanishes at 0 and its unit ball is given by a polyhedron.

Some examples of polyhedral gauges

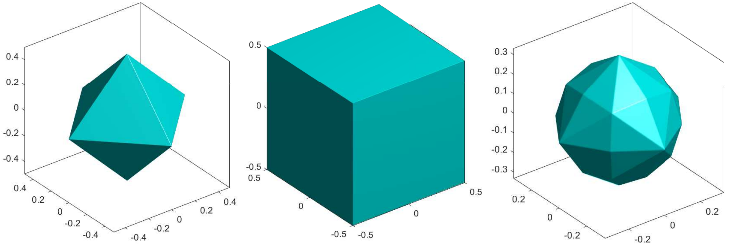
$L_1(b) = \sum_i |b_i|$, $L_\infty = max_i |b_i|$

SLOPE(B., van den Berg, Su, Candès, arxiv 2013, AoAS, 2015)

OWL(Zeng and Figuereido, IEEE Signal Process. Lett., 2014)

$$J_\lambda(b) = \sum_i \lambda_i |\beta|_{(i)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$.



((a)) (2,2,2) ((b)) (2,0,0) ((c)) (3,2,1) for a convex function $\phi : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a subgradient of ϕ at $\beta \in \mathbb{R}^p$ if

$$\phi(b) \geq \phi(\beta) + s'(b - \beta) \ \ \forall b \in \mathbb{R}^p.$$

The convex, non-empty set of all subgradients of ϕ at β is called the subdifferential of ϕ at β, denoted by $\partial_\phi(\beta)$.

Subdifferential

For a convex function $\phi : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a subgradient of ϕ at $\beta \in \mathbb{R}^p$ if

$$\phi(b) \geq \phi(\beta) + s'(b - \beta) \ \ \forall b \in \mathbb{R}^p.$$

Subdifferential

For a convex function $\phi : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a subgradient of ϕ at $\beta \in \mathbb{R}^p$ if

$$\phi(b) \geq \phi(\beta) + s'(b - \beta) \;\; \forall b \in \mathbb{R}^p.$$

The convex, non-empty set of all subgradients of ϕ at β is called the subdifferential of ϕ at β, denoted by $\partial_\phi(\beta)$.

Subdifferential

For a convex function $\phi : \mathbb{R}^p \to \mathbb{R}$, a vector $s \in \mathbb{R}^p$ is a subgradient of ϕ at $\beta \in \mathbb{R}^p$ if

$$\phi(b) \geq \phi(\beta) + s'(b - \beta) \;\; \forall b \in \mathbb{R}^p.$$

The convex, non-empty set of all subgradients of ϕ at β is called the subdifferential of ϕ at β, denoted by $\partial_\phi(\beta)$.

For $\phi(\beta) = \lambda|\beta|$,

$$\partial_\phi(\beta) = \begin{cases} \lambda \text{ for } & \beta > 0 \\ [-\lambda, \lambda] \text{ for } & \beta = 0 \\ -\lambda \text{ for } & \beta < 0 \end{cases}$$

$B^* = \partial_{\mathrm{pen}}(0)$ is the unit ball of the norm which is dual to $\mathrm{pen}(\cdot)$

Pattern definition (Graczyk, Schneider, Skalski, Tardivel, arxiv 2024)

β and $\tilde{\beta} \in \mathbb{R}^p$ have the same pattern with respect to pen if

$$\partial_{\mathrm{pen}}(\beta) = \partial_{\mathrm{pen}}(\tilde{\beta}),$$

Pattern definition (Graczyk, Schneider, Skalski, Tardivel, arxiv 2024)

β and $\tilde{\beta} \in \mathbb{R}^p$ have the same pattern with respect to pen if

$$\partial_{\mathrm{pen}}(\beta) = \partial_{\mathrm{pen}}(\tilde{\beta}),$$

C_β - pattern equivalence class (the set of all elements of $\mathbb{R}^p$ sharing the same pattern as β)

Pattern definition (Graczyk, Schneider, Skalski, Tardivel, arxiv 2024)

β and $\tilde{\beta} \in \mathbb{R}^p$ have the same $\underline{\text{pattern with respect to pen}}$ if

$$\partial_{\mathrm{pen}}(\beta) = \partial_{\mathrm{pen}}(\tilde{\beta}),$$

C_β - pattern equivalence class (the set of all elements of $\mathbb{R}^p$ sharing the same pattern as β)

ℓ_1-norm: The pattern corresponds to the sign vector

$$\mathrm{sign}(\beta) = (\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_p))'.$$

SLOPE:

$$\mathrm{patt}_{\mathrm{slope}}(\beta)_j = \mathrm{sign}(\beta_j)\,\mathrm{rank}(|\beta|)_j$$

For $\beta = (3.1, -1.2, 0.5, 0, 1.2, -3.1)'$,
$\mathrm{patt}_{\mathrm{slope}}(\beta) = (3, -2, 1, 0, 2, -3)'$.

Pattern recovery under L_1 norm: Basis Pursuit

$$Y = X_{n\times p}\beta, \quad p > n$$

Pattern recovery under L_1 norm: Basis Pursuit

$$Y = X_{n \times p}\beta, \quad p > n$$

Goal: Identify the sparsest solution

Pattern recovery under L_1 norm: Basis Pursuit

$$Y = X_{n \times p}\beta, \quad p > n$$

Goal: Identify the sparsest solution

Basis Pursuit (Chen and Donoho, 1994): Estimate β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

Pattern recovery under L_1 norm: Basis Pursuit

$$Y = X_{n\times p}\beta, \quad p > n$$

Goal: Identify the sparsest solution

Basis Pursuit (Chen and Donoho, 1994): Estimate β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \; \text{ then } \; \|\gamma\|_1 > \|\beta\|_1.$$

Pattern recovery under L_1 norm: Basis Pursuit

$$Y = X_{n \times p}\beta, \quad p > n$$

Goal: Identify the sparsest solution

Basis Pursuit (Chen and Donoho, 1994): Estimate β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \text{ then } \|\gamma\|_1 > \|\beta\|_1.$$

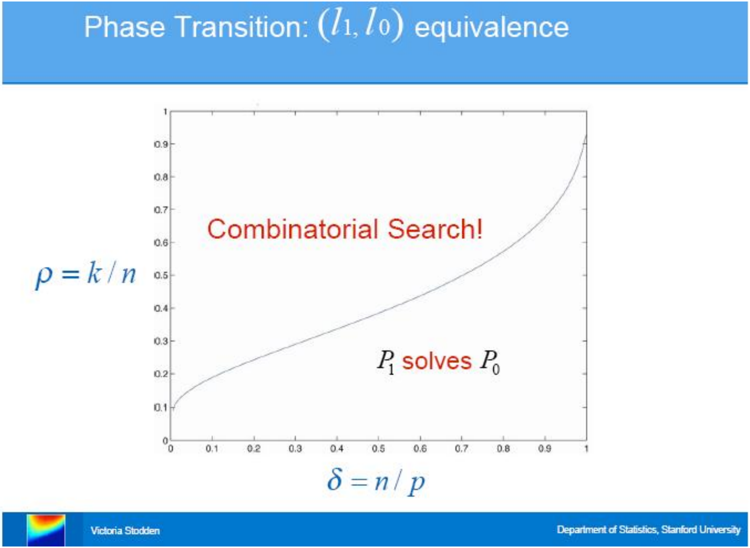
$$k = ||\beta||_0 = \#\{i : \ \beta_i \neq 0\}$$

Transition curve (Donoho and Tanner, 2005)

Let's assume than $p \to \infty$, $n/p \to \delta > 0$ and $k/n \to \varepsilon > 0$.

If X_{ij} are iid $N(0, \tau^2)$ then the probability that BP recovers the sparsest solution converges to 1 if $\varepsilon < \rho(\delta)$ and to 0 if $\varepsilon > \rho(\delta)$, where $\rho(\delta)$ is the *transition curve*.

Transition curve (2)



Phase Transition: (l_1, l_0) equivalence

Victoria Stodden

Department of Statistics, Stanford University

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \varepsilon$

Or : $\min_{b\in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda||b||_1$

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \varepsilon$

Or : $\min_{b\in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda||b||_1$

BPDN (Chen and Donoho, 1994) or LASSO (Tibshirani, 1996)

Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$

Irrepresentability condition

The sign vector of β is defined as $S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$, where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$

Irrepresentability condition:

$$\|X'X_I(X_I'X_I)^{-1}S(\beta_I)\|_\infty \leq 1$$

Irrepresentability condition

The sign vector of β is defined as $S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$, where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$

Irrepresentability condition:

$$\|X'X_I(X_I'X_I)^{-1}S(\beta_I)\|_\infty \leq 1$$

When

$$\|X'X_I(X_I'X_I)^{-1}S(\beta_I)\|_\infty > 1$$

then LASSO can not identify the true support in the noisless case and in the noisy case the probability of the support recovery by LASSO is smaller than 0.5 (Wainwright, 2009).

Irrepresentability condition (2)

If $p \to \infty$, $n/p \to \delta$ then $k < \frac{p\delta}{2 \log p}(1 + o(1))$

Irrepresentability condition (2)

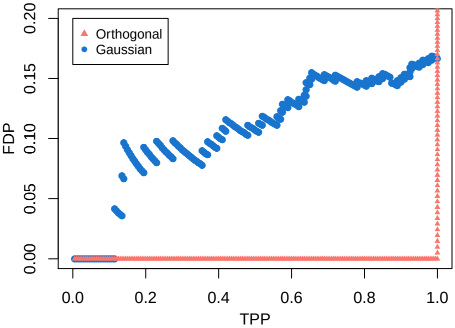
If $p \to \infty$, $n/p \to \delta$ then $k < \frac{p\delta}{2\log p}(1 + o(1))$

Corollary: Even in the noiseless case LASSO can not recover the true support if $\frac{k}{p} \to \varepsilon > 0$

False Discoveries along the lasso path (Su, B. and Candès, (AoS, 2017))

$$R := \left|\{i:\ \hat{\beta}_i \neq 0\}\right| \ , V := \left|\{i:\ \beta_i = 0, \quad \hat{\beta}_i \neq 0\}\right|$$

$$FDP := \frac{V}{\max\{R,1\}} \ , TPP = \frac{R-V}{k}$$



FDP-Power tradeoff

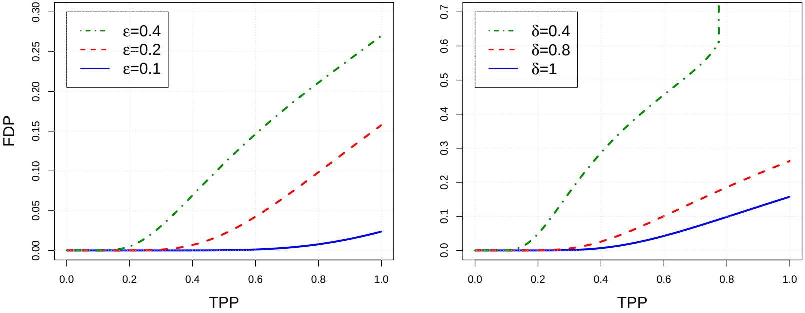
Theorem (Su, Bogdan, Candes, 2017)

Fix $\delta \in (0, \infty)$ and $\varepsilon \in (0, 1)$. Then the event

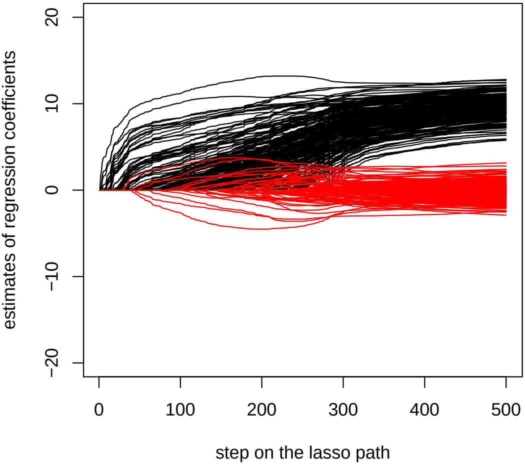
$$\bigcap_{\lambda \geq 0.01} \left\{ FDP(\lambda) \geq q^{\star}\left(TPP(\lambda)\right) - 0.001 \right\} \tag{2}$$

holds with probability tending to one.

FDR-Power trade-off (2)



Thresholded LASSO (1)



Support recovery by thresholded LASSO

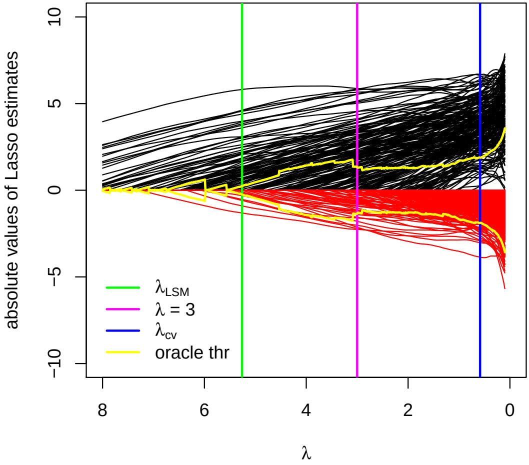
Theorem (Tardivel, B., SJS 2022)

For any $\lambda > 0$ LASSO can separate well the causal and null features if and only if vector β is identifiable with respect to l_1 norm and $\min_{i \in I} |\beta_i|$ is sufficiently large.

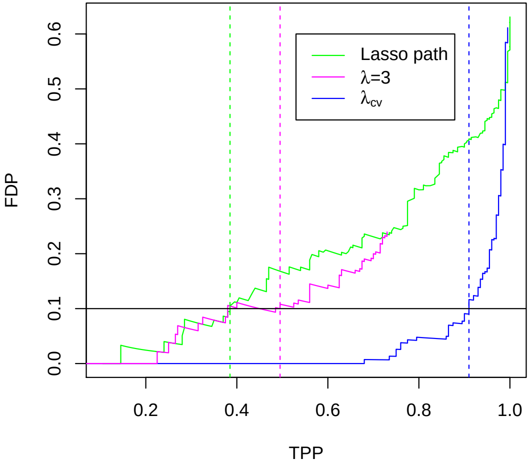
Corollary

Thresholded LASSO can identify sufficiently large signals if $\varepsilon < \phi(\delta)$, where $\phi(\cdot)$ is the transition curve of Donoho and Tanner (2005)

Thresholded LASSO (2)



Thresholded LASSO (3)



Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

$\hat{\beta}(\lambda)$ - vector of $2p$ estimates of regression coefficients by LASSO applied on the augmented design matrix $X_{aug} = [X, \tilde{X}]$

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

$\hat{\beta}(\lambda)$ - vector of $2p$ estimates of regression coefficients by LASSO applied on the augmented design matrix $X_{aug} = [X, \tilde{X}]$

LCD importance statistics:

$$W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{p+j}|$$

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

$\hat{\beta}(\lambda)$ - vector of $2p$ estimates of regression coefficients by LASSO applied on the augmented design matrix $X_{aug} = [X, \tilde{X}]$

LCD importance statistics:

$$W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{p+j}|$$

LSM importance statistics:

$$T_j = max\{\lambda : |\widehat{\beta}_j| > 0\}$$

$$\tilde{W}_j = T_j - T_{p+j}$$

Knockoff filter

Define a random threshold as

$$\hat{t}(\lambda) = \min\left\{ t > 0 : \frac{1 + \#\{j : W_j(\lambda) \leq -t\}}{\#\{j : W_j(\lambda) \geq t\}} \leq q \right\}$$

and select

$$\widehat{\mathcal{S}(\lambda)} = \{j : W_j(\lambda) \geq \hat{t}(\lambda)\},$$

Knockoff filter

Define a random threshold as

$$\hat{t}(\lambda) = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j(\lambda) \le -t\}}{\#\{j : W_j(\lambda) \ge t\}} \le q \right\}$$

and select

$$\widehat{\mathcal{S}(\lambda)} = \{j : W_j(\lambda) \ge \hat{t}(\lambda)\},$$

Foygel-Barber and Candès (2015), Candès, Fan, Janson and Lv (2017) - The above knockoff procedure $KN(\lambda, q)$ controls $FDR = E(FDP)$ at the level q.

Asymptotic Theory for Knockoffs, (Weinstein, Su, B., Barber, Candès, AoS 2023)

$$\frac{k}{p} \to \varepsilon < 2\varepsilon^*(\delta/2), \tag{3}$$

where $\varepsilon^*(\delta)$ is a point on the Donoho–Tanner transition curve.

Definition

A sequence of random variables Π_m is said to be *ε-sparse and growing*, if $\mathbb{P}(\Pi_m \neq 0) = \varepsilon$ for all m, and $\mathbb{P}(|\Pi_m| > M | \Pi_m \neq 0) \to 1$ as $m \to \infty$ for every $M > 0$.

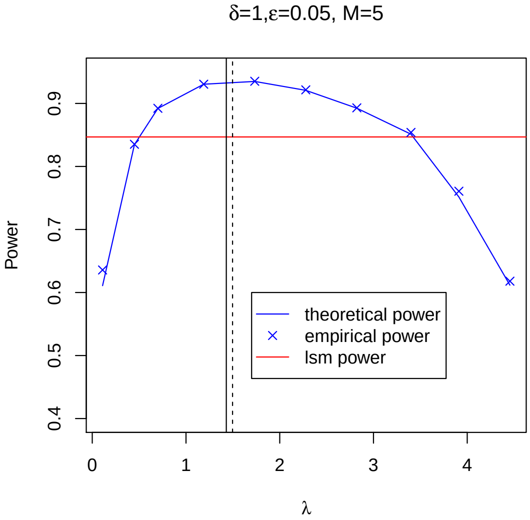
Theorem

For any ε-sparse and growing sequence $\{\Pi_m\}$, it holds that for any fixed $0 < \lambda_1 < \lambda_2$ and any $\nu > 0$, there exist m' and $n'(m)$ s.t.

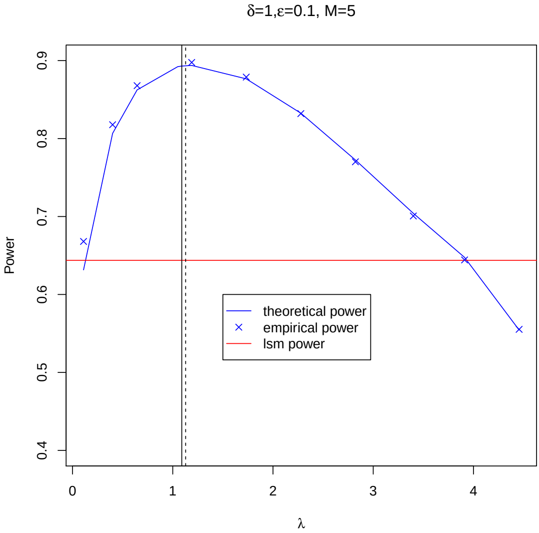
$$\mathbb{P}\left(\inf_{\lambda_1 \leq \lambda \leq \lambda_2} TPP(\lambda, \Pi_m, q) > 1 - \nu\right) \geq 1 - \nu$$

if $m \geq m'$ and $n \geq n'(m)$.

Gain in power over LSM



Gain in power over LSM



Polygedral Gauges, Graczyk et al. 2023

Noisless recovery condition (Irrepresentability) for β and X: there exists such $b \in \mathrm{lin}(C_\beta)$ such that $X'Xb \in \partial_{\mathrm{pen}}(\beta)$, i.e. $X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$

Polygedral Gauges, Graczyk et al. 2023

Noisless recovery condition (Irrepresentability) for β and X: there exists such $b \in \mathrm{lin}(C_\beta)$ such that $X'Xb \in \partial_{\mathrm{pen}}(\beta)$, i.e.
$X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$

Accesibility (Identifiability)
Geometric characterization: The pattern of $\beta \in \mathbb{R}^p$ is accessible with respect to X and $\lambda\mathrm{pen}$ if and only if

$$\mathrm{row}(X) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing,$$

i.e. there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\mathrm{pen}}(\beta)$

Polygedral Gauges, Graczyk et al. 2023

Noisless recovery condition (Irrepresentability) for β and X: there exists such $b \in \mathrm{lin}(C_\beta)$ such that $X'Xb \in \partial_{\mathrm{pen}}(\beta)$, i.e. $X'X\mathrm{lin}(C_\beta) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing$

Accesibility (Identifiability)
Geometric characterization: The pattern of $\beta \in \mathbb{R}^p$ is accessible with respect to X and $\lambda\mathrm{pen}$ if and only if

$$\mathrm{row}(X) \cap \partial_{\mathrm{pen}}(\beta) \neq \varnothing,$$

i.e. there exists $z \in \mathbb{R}^n$ such that $X'z \in \partial_{\mathrm{pen}}(\beta)$

If accessibility is satisfied, for sufficiently large signals the pattern of the true signal is nested within the pattern of the estimator.

SLOPE

- SLOPE (B., van den Berg, Su, Candès, arxiv 2013, B.,van den Berg, Sabatti, Su, Candès, AoAS, 2015) penalizes larger coefficients more stringently

$$\hat{\beta}_{SLOPE} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|y - X\beta\|^2 + \sigma \sum_{j=1}^{p} \lambda_j |\beta|_{(j)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(p)}$.

Theorem (B,van den Berg, Su and Candès (2013))

When $X^T X = I$ SLOPE with

$$\lambda_i^{BH} := \sigma \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

controls FDR at the level $q\frac{p_0}{p}$.

Optimality in prediction and estimation

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (Annals of Statistics, 2018):

SLOPE with the BH related sequence of tuning parameters adapts to the unknown sparsity and attains minimax prediction and estimation rates $\frac{k}{n}\log(p/k)$ for the estimation error $||\hat{\beta}-\beta||^2$.

Optimality in prediction and estimation

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (Annals of Statistics, 2018):

SLOPE with the BH related sequence of tuning parameters adapts to the unknown sparsity and attains minimax prediction and estimation rates $\frac{k}{n}\log(p/k)$ for the estimation error $||\hat{\beta}-\beta||^2$.

Fixed λ LASSO rate of convergence - $\frac{k}{n}\log(p)$

Optimality in prediction and estimation

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (Annals of Statistics, 2018):

SLOPE with the BH related sequence of tuning parameters adapts to the unknown sparsity and attains minimax prediction and estimation rates $\frac{k}{n}\log(p/k)$ for the estimation error $||\hat{\beta}-\beta||^2$.

Fixed λ LASSO rate of convergence - $\frac{k}{n}\log(p)$

Extension to classification by logistic regression by Abramovich and Grinshtein (2018, IEEE Trans. Inf. Theory)

SLOPE pattern (Schneider, Tardivel, JMLR 2022)

Definition

A vector $M \in Z^p$ is a SLOPE model if either $M = 0$ or for all $1 \leq l \leq \|M\|_\infty$ there exists j such that $|M_j| = l$.
Moreover, for $b \in \mathbb{R}^p$ its SLOPE model $\mathrm{mdl}(b)$ is defined in a following way:

- $sign(\mathrm{mdl}(b)) = sign(b)$ (sign preservation),
- $|b_i| = |b_j| |\mathrm{mdl}(b)_i| = |\mathrm{mdl}(b)_j|$ (clustering preservation),
- $|b_i| > |b_j| |\mathrm{mdl}(b)_i| > |\mathrm{mdl}(b)_j|$ (hierarchy preservation).

Example

Let $\beta = (4, 0, -1.5, 1.5, -4)$. Then $\mathrm{mdl}(\beta) = (2, 0, -1, 1, -2)$.

SLOPE model matrix(1)

Definition

Let m be a model for SLOPE in R^p where $\|m\|_\infty = k$ (the number of non-null clusters). The matrix $U_m \in \mathbb{R}^{p\times k}$ is defined as follows

$$\forall i \in \{1,\dots,p\}, \forall j \in \{1,\dots,k\}, (U_m)_{ij} = sign(m_i)\mathbf{1}_{(|m_i|=k+1-j)}.$$

By convention, when $m = 0$ we define the null model matrix as $U_0 := 0$.

Model matrix example

Let $p = 8$ and $m = (3, -3, 2, 1, 2, -1, 0, 3)$. Here $k = 3$ and the model matrix is

$$U_m = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Irrepresentability condition for SLOPE (Skalski, B., Graczyk, Kołodziejek, Tardivel, Wilczyński, arxiv 2022)

$$\tilde{X} = XU_M, \tilde{\Lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k) \quad \text{where} \quad \tilde{\lambda}_j = \sum_{i=k_{j-1}+1}^{k_j} \lambda_i.$$

Irrepresentability condition:

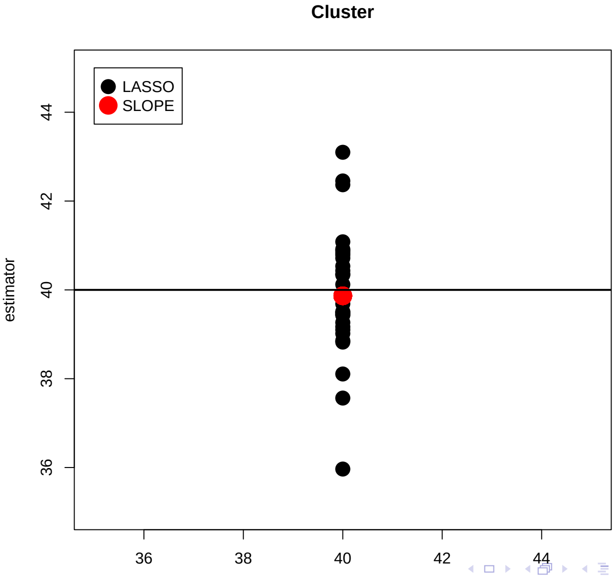
$$J_\lambda^D(X'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{\Lambda}) \leq 1$$

where

$$J_\lambda^D(x) := \max\left\{\frac{|x|_{(1)}}{\lambda_1}, \ldots, \frac{\sum_{i=1}^p |x|_{(i)}}{\sum_{i=1}^p \lambda_i}\right\}, \text{ where} |x|_{(1)} \geq \ldots \geq |x|_{(p)}.$$

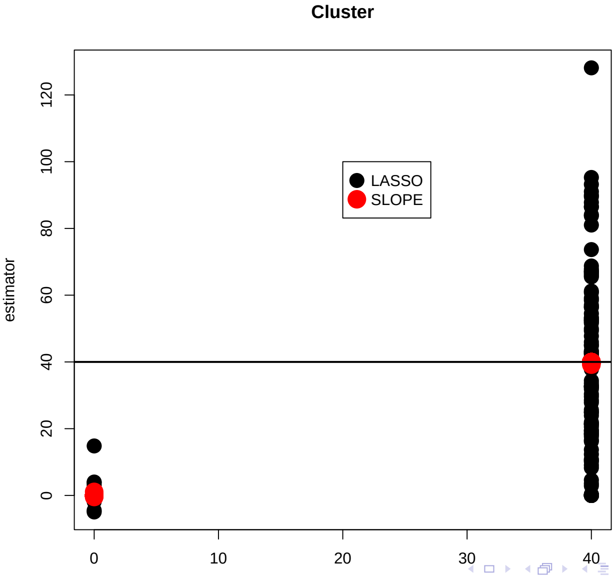
SLOPE vs LASSO (1)

$n = 100$, $p = 200$, $k = 30$, exponentially decaying correlation

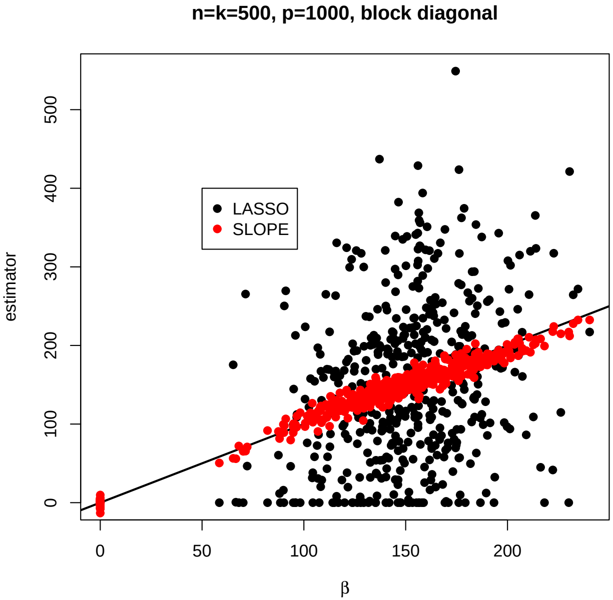


SLOPE vs LASSO (2)

$n = 100,\ p = 200,\ k = 100$, exponentially decaying correlation



SLOPE vs LASSO (2)



Low dimensional asymptotics, Hejny, Wallin, B. arxiv 2024

$$y = X\beta^0 + \varepsilon$$

$X = (X_1, .., X_n)^T$, with $X_1, X_2, \dots$ i.i.d. centered random vectors in $\mathbb{R}^p$ with covariance matrix C and $\varepsilon \sim N(0, \sigma^2 I)$.

Low dimensional asymptotics, Hejny, Wallin, B. arxiv 2024

$$y = X\beta^0 + \varepsilon$$

$X = (X_1, .., X_n)^T$, with $X_1, X_2, \dots$ i.i.d. centered random vectors in $\mathbb{R}^p$ with covariance matrix C and $\varepsilon \sim N(0, \sigma^2 I)$.

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + f_n(\beta)$$

Low dimensional asymptotics, Hejny, Wallin, B. arxiv 2024

$$y = X\beta^0 + \varepsilon$$

$X = (X_1, .., X_n)^T$, with $X_1, X_2, \dots$ i.i.d. centered random vectors in $\mathbb{R}^p$ with covariance matrix C and $\varepsilon \sim N(0, \sigma^2 I)$.

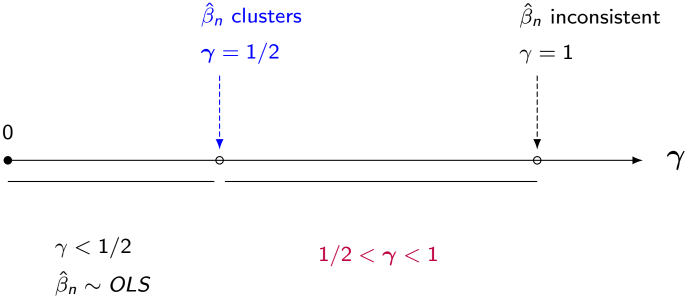
$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + f_n(\beta)$$

$$f(\beta) = \max\{v_1^T\beta, \dots, v_k^T\beta\} + g(\beta),$$

where $v_1, \dots, v_k$ are the regularizer specific vectors in $\mathbb{R}^p$, and $g(\beta)$ is a convex differentiable function.

Penalty scaling

$f_n = n^\gamma f$



$\hat{\beta}_n$ clusters

$\gamma = 1/2$

$\hat{\beta}_n$ inconsistent

$\gamma = 1$

0

γ

$\gamma < 1/2$

$\hat{\beta}_n \sim OLS$

$1/2 < \gamma < 1$

Asymptotic error distribution

Directional derivative of a function $f : \mathbb{R}^p \to \mathbb{R}$ at a point x in direction u:

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

Asymptotic error distribution

Directional derivative of a function $f : \mathbb{R}^p \to \mathbb{R}$ at a point x in direction u:

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

Theorem

Let $f : \mathbb{R}^p \to \mathbb{R}$ be any convex penalty function and $f_n = n^{1/2} f$. Assume C is positive definite. Then $\hat{u}_n := \sqrt{n}(\hat{\beta}_n - \beta^0) \xrightarrow{d} \hat{u}$, where

$$\hat{u} := \mathrm{argmin}_u V(u),$$
$$V(u) = \frac{1}{2} u^T C u - u^T W + f'(\beta^0; u), \tag{4}$$

with $W \sim \mathcal{N}(0, \sigma^2 C)$, and $f'(\beta^0; u)$ the directional derivative of f at β^0 in direction u.

Pattern convergence

Theorem

For every convex set $\mathcal{K} \subset \mathbb{R}^p$*:* $\mathbb{P}[\hat{u}_n \in \mathcal{K}] \longrightarrow \mathbb{P}[\hat{u} \in \mathcal{K}]$ *as* $n \to \infty$*.*
In particular, $I(\hat{u}_n)$ *converges weakly to* $I(\hat{u})$*.*

Pattern recovery

$\langle U_{\beta^0} \rangle$ - pattern space of β^0

Theorem

The limiting probability of pattern recovery is given by

$$\mathbb{P}\big[I(\hat{\beta}_n) = I(\beta^0)\big] \underset{n\to\infty}{\longrightarrow} \mathbb{P}\big[\hat{u} \in \langle U_{\beta^0} \rangle\big] = \mathbb{P}\big[\zeta \in \partial f(\beta^0)\big],$$
$$\zeta = C^{1/2} P C^{-1/2} v_0 + C^{1/2}(I - P) C^{-1/2} W,$$

where P is the projection onto $C^{1/2}\langle U_{\beta^0} \rangle$ and v_0 is any vector in $\partial f(\beta^0)$. In particular, if $W \sim \mathcal{N}(0, \sigma^2 C)$,

$$\zeta \sim \mathcal{N}(C^{1/2} P C^{-1/2} v_0, \sigma^2 C^{1/2}(I - P) C^{1/2}).$$

Pattern recovery (2)

Irrepresentability condition:

$$C\langle U_{\beta^0}\rangle \cap ri(\partial f(\beta^0)) \neq \varnothing. \tag{5}$$

Corollary

Assume that $f_n = \alpha n^{1/2} f_0$. Then under the irrepresentability condition

$$\lim_{n\to\infty} \mathbb{P}[I(\hat{\beta}_n) = I(\beta^0)] = 1 - o(\alpha),$$

where $o(\alpha) \to 0$ as $\alpha \to \infty$

Two step procedure

Theorem

Let $\hat{\xi}_n$ be a random sequence such that $\sqrt{n}(\hat{\xi}_n - \beta^0) \xrightarrow{d} W$ for some subgaussian random vector W. Let $\hat{\beta}_n$ minimize

$$M_n(\beta) := \frac{1}{2}\|\hat{\xi}_n - \beta\|_2^2 + n^{-1/2} f(\beta),$$

Then

$$\lim_{n\to\infty} \mathbb{P}\big[I(\hat{\beta}_n) = I(\beta^0)\big] \geq 1 - exp(-c\alpha^2),$$

for some $c > 0$ as $\alpha \to \infty$, provided the irrepresentability condition is satisfied for $C = \mathbb{I}$. This is especially the case for Lasso or SLOPE but not for the fused LASSO.

Concavification of Fused Lasso

For $C = \mathbb{I}$, the Fused Lasso

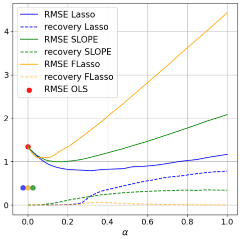
$$f_0(\beta) = \lambda \sum_{i=1}^{p-1} a_i |\beta_{i+1} - \beta_i| + \lambda \sum_{i=1}^{p} a|\beta_i| \ \ ,$$

asymptotically recovers all its patterns, i.e;

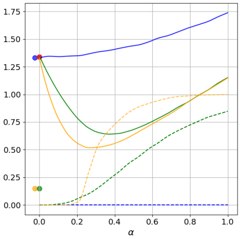
$$\forall \beta^0 \in \mathbb{R}^p; \quad \lim_{n\to\infty} \mathbb{P}[I_A(\hat{\beta}_n) = I_A(\beta^0)] \underset{\lambda\to\infty}{\longrightarrow} 1,$$

if and only if $(0, a_1, \ldots, a_{p-1}, 0)$ forms a strictly concave sequence and the sparsity penalty $a > \max\{a_i + a_{i+1} : 0 \leq i \leq p-1\}$.

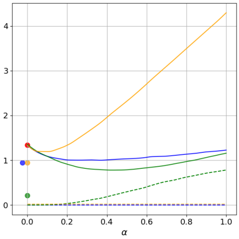
$$f_n = \alpha n^{1/2} f_0$$



$\beta^0 = [0, 0, 1, 0],$



$\beta^0 = [1, 1, 1, 1],$



$\beta^0 = [1, 0, 1, 0]$

Asymptotic FDR control

If $C = \mathbb{I}_{I_0} \oplus \Sigma$, with identity matrix $\mathbb{I}_{I_0}$ on I_0 and an arbitrary positive definite Σ on I_0^c, then

$$\mathrm{FDR}(\hat{\beta}_n) \xrightarrow[n\to\infty]{} C_0 \leq q\frac{p_0}{p}.$$

Clustering in financial applications

- Kremer, Lee, B., Paterlini, *Journal of Banking and Finance* 110, 105687, 2020 - application for portfolio selection.
- Kremer, Brzyski, B., Paterlini, SSRN 3412061, *Quantitative Finance*, 2022.

Different flavor of clustering, Kremer et al, 2022

Figuereido and Nowak (2014) - clustering based on correlations between predictors

Different flavor of clustering, Kremer et al, 2022

Figuereido and Nowak (2014) - clustering based on correlations between predictors

Theorem (Kremer, Brzyski, B., Paterlini, 2021)

Let's assume that columns of X have the same L_2 norm and that the SLOPE solution satisfies $\hat{\beta}_1 \geq \ldots \geq \hat{\beta}_p \geq 0$ (this can always be achieved by permuting columns of X and changing their signs). Then, for any $i \in \{1, \ldots, p-1\}$, it holds

$$\hat{\beta}_i > \hat{\beta}_{i+1} \quad \Longrightarrow \quad X_i^T r_P - X_{i+1}^T r_P \geq \lambda_i - \lambda_{i+1} \ ,$$

where $r_P := Y - X_{\setminus i,i+1}\hat{\beta}_{\setminus i,i+1}$ and $X_{\setminus i,i+1}$ and $\hat{\beta}_{\setminus i,i+1}$ are obtained by removing i^{th} and $i+1^{st}$ columns of and elements of $\hat{\beta}$.

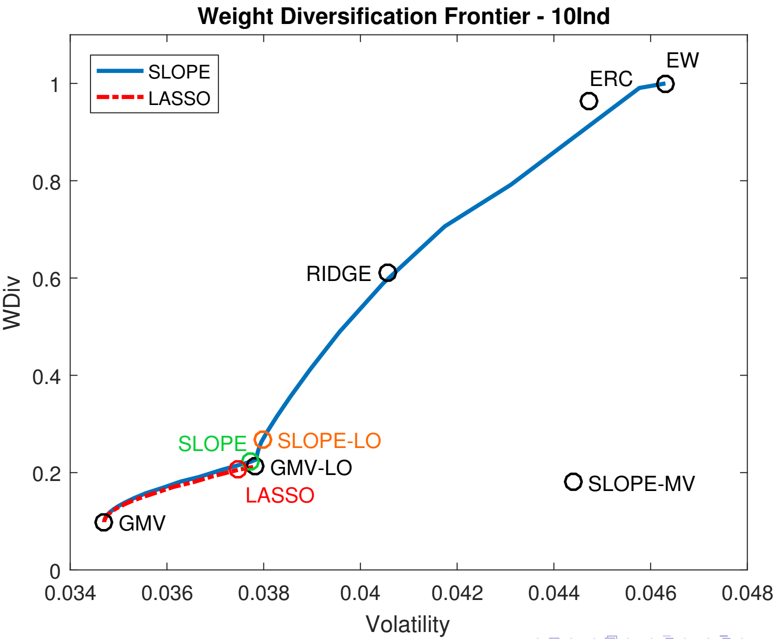
Portfolio Optimization, (Kremmer et al, 2020, JBF)

$R_{t \times k} = (R_1, \ldots, R_k)$ - asset returns, $Cov(R) = \Sigma$

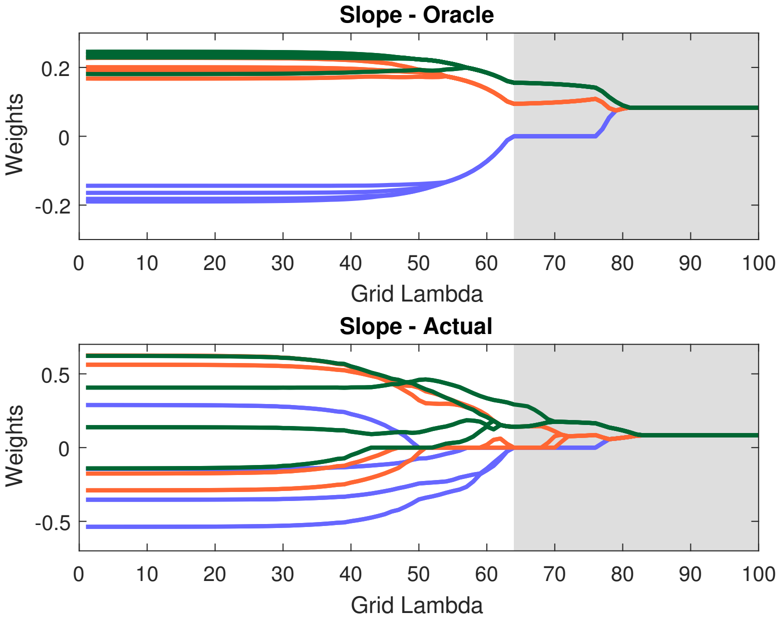
$$\min_{w \in \mathbb{R}^k} w' \Sigma w + J_\lambda(w) \tag{6}$$

$$\text{s.t.} \sum_{i=1}^{k} w_i = 1 \tag{7}$$

Evolution of Portfolio



SLOPE clustering



Current research

- High dimensional asymptotics for general classifiers
- Graphical models - estimating the partial correlation matrix
- Development of adaptive (nonconvex) and thresholded versions with the statistical guarantees [see Jiang, B., Josse, Majewski, Miasojedow, Rockova, JCGS, 2022]
- Efficient implementations for SLOPE - problems with a non-separable penalty.