Recent results on the Sorted L-One Penalized Estimator

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

27th of April, 2022

Outline

- Sorted L-One Penalized Estimator for multiple regression
- Graphical SLOPE

Model selection in high-dimension

Linear regression model: $$y = X\beta + \varepsilon,$$

- $y = (y_i)$: vector of response of length n
- $X = (X_{ij})$: a standardized design matrix of dimension $n \times p$
- $\beta = (\beta_j)$: regression coefficient of length p
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

Model selection in high-dimension

Linear regression model: $$y = X\beta + \varepsilon,$$

- $y = (y_i)$: vector of response of length n
- $X = (X_{ij})$: a standardized design matrix of dimension $n \times p$
- $\beta = (\beta_j)$: regression coefficient of length p
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

Assumptions:

- high-dimension: p large (comparable or larger than n)

LASSO, (Tibshirani, JRSSB 1994)

- LASSO - solution to the convex optimization problem

$$\mathit{argmin}_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

where $\lambda_L > 0$ is a tuning parameter

LASSO, (Tibshirani, JRSSB 1994)

- LASSO - solution to the convex optimization problem

$$argmin_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

  where $\lambda_L > 0$ is a tuning parameter
- Difficulty - selection of λ_L

LASSO, (Tibshirani, JRSSB 1994)

- LASSO - solution to the convex optimization problem

$$argmin_b \quad \left\{\frac{1}{2}\|y - Xb\|_2^2 + \lambda_L\|b\|_1\right\}, \qquad \text{(LASSO)}$$

  where $\lambda_L > 0$ is a tuning parameter
- Difficulty - selection of λ_L
- $\lambda_L = \sqrt{\frac{2\log p}{n}}$ - related to the Bonferroni criterion, allows for consistency when the signal is sufficiently sparse

LASSO, (Tibshirani, JRSSB 1994)

- LASSO - solution to the convex optimization problem

$$argmin_b \quad \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda_L \|b\|_1 \right\}, \qquad \text{(LASSO)}$$

  where $\lambda_L > 0$ is a tuning parameter
- Difficulty - selection of λ_L
- $\lambda_L = \sqrt{\frac{2 \log p}{n}}$ - related to the Bonferroni criterion, allows for consistency when the signal is sufficiently sparse
- cross-validation - optimization of predictive properties, many false discoveries

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
$S(\beta) = (S(\beta_1), \dots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

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

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 probability of the support recovery by LASSO is smaller than 0.5 (Wainwright, 2009).

Identifiability condition

Definition (Identifiability)

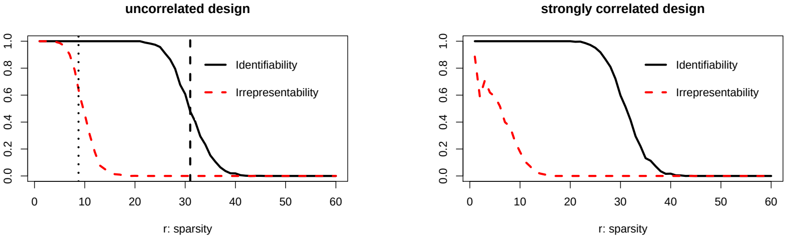
Let X be a $n \times p$ matrix. The vector $\beta \in R^p$ is said to be identifiable with respect to the l norm if the following implication holds

$$X\gamma = X\beta \text{ and } \gamma \neq \beta \Rightarrow \|\gamma\|_1 > \|\beta\|_1. \quad (1)$$

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.

Irrepresentablity vs identifiability



Rysunek: $n = 100$, $p = 300$, in the right panel $\rho(X_i, X_j) = 0.9$, vertical lines correspond to $n/(2 \log p)$ and the transition curve of Donoho and Tanner (2009).

SLOPE

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

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

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

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β
- We define:
  - the number of all discoveries, $R := \left|\left\{i : \ \widetilde{\beta}_i \neq 0\right\}\right|$
  - the number of false discoveries, $V := \left|\left\{i : \ \beta_i = 0, \quad \widetilde{\beta}_i \neq 0\right\}\right|$
  - false discovery rate - expected proportion of false discoveries among all discoveries

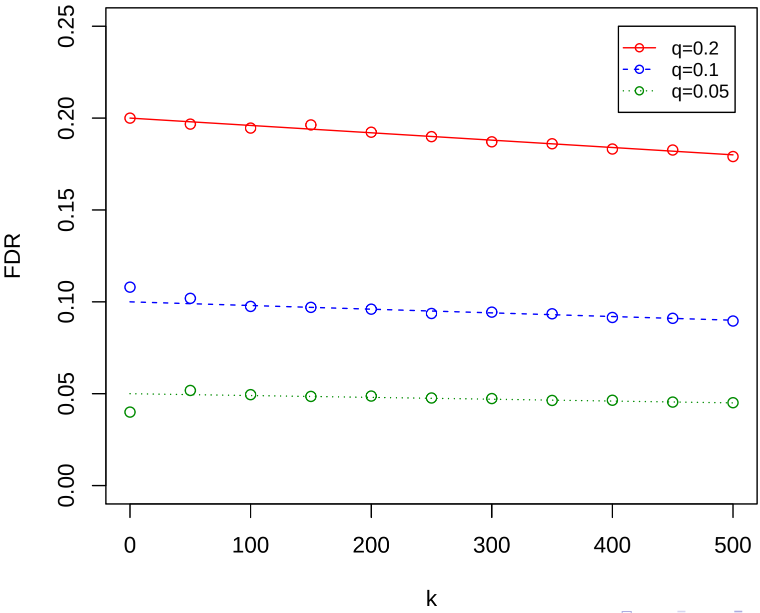
$$FDR := \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right]$$

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 $q\frac{p_0}{p}$.

Orthogonal design, $n = p = 5000$



Asymptotic optimality, Su and Candès (Annals of Statistics, 2016) and FDR control, Kos (2018)

Theorem

Let $X_{ij} \sim N(0, 1/\sqrt{n})$. Fix $0 < q < 1$ and choose $\lambda = \sigma(1+\varepsilon)\lambda^{BH}(q)$ for some arbitrary constant $0 < \varepsilon < 1$. Suppose $k/p \to 0$ and $\frac{k \log p}{n} \to 0$. Then

$$\sup_{||\beta_0|| \leq k} P\left(\frac{||\hat{\beta}_{SL} - \beta||^2}{2\sigma^2 k \log(p/k)} > 1 + 3\varepsilon\right) \to 0$$

$$\inf_{\hat{\beta}} sup_{||\beta_0|| \leq k} P\left(\frac{||\hat{\beta} - \beta||^2}{2\sigma^2 k \log(p/k)} > 1 - \varepsilon\right) \to 1$$

(M. Kos, 2018) If additionally $k^2/n \to 0$ then

$$FDR_n \leq \Delta_n \to \quad q$$

Asymptotic optimality (2)

Minimax estimation/prediction rate $[k \log(p/k)]$ under weighted restricted eigenvalue condition (large collection of random matrices)

Asymptotic optimality (2)

Minimax estimation/prediction rate $[k \log(p/k)]$ under weighted restricted eigenvalue condition (large collection of random matrices)

$\lambda_i = \rho\sqrt{2\log(p/i)}$, ρ is larger than one

Bellec, Lecué, Tsybakov (2016,2017)

Asymptotic optimality (2)

Minimax estimation/prediction rate $[k \log(p/k)]$ under weighted restricted eigenvalue condition (large collection of random matrices)

$\lambda_i = \rho\sqrt{2\log(p/i)}$, ρ is larger than one

Bellec, Lecué, Tsybakov (2016,2017)

Extension to GLM by Abramovich and Grinshtein (2017)

Asymptotic optimality (2)

Minimax estimation/prediction rate $[k \log(p/k)]$ under weighted restricted eigenvalue condition (large collection of random matrices)

$\lambda_i = \rho\sqrt{2\log(p/i)}$, ρ is larger than one

Bellec, Lecué, Tsybakov (2016,2017)

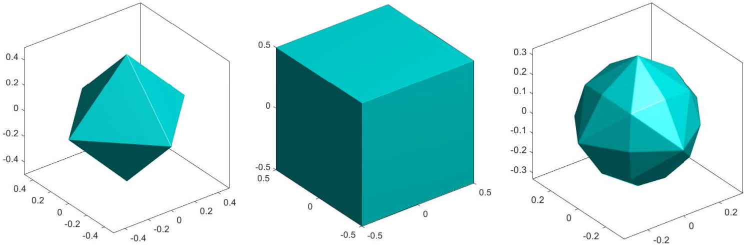
Extension to GLM by Abramovich and Grinshtein (2017)

LASSO rate of convergence - $k \log(p)$

Clustering properties of SLOPE (2)

- Schneider and Tardivel, arxive 2020 - class of models attainable by SLOPE
- B., Dupuis, Graczyk, Kołodziejek, Skalski, Tardivel, Wilczyński, arxiv 2022: Necessary and sufficient condition for SLOPE pattern recovery

Unit balls for different SLOPE sequences by D.Brzyski



SLOPE pattern (Schneider, Tardivel, 2020)

Definition

For $b \in \mathbb{R}^p$ its SLOPE pattern $\mathrm{patt}(b)$ is defined in a following way:

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

Example

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

Fact:

$$\mathrm{patt}(b_1) = \mathrm{patt}(b_2) \quad \Leftrightarrow \quad \partial_{\mathrm{slope}}(b_1) = \partial_{\mathrm{slope}}(b_2)$$

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

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

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 (Bogdan et al. (2022)

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

Irrepresentability condition:

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

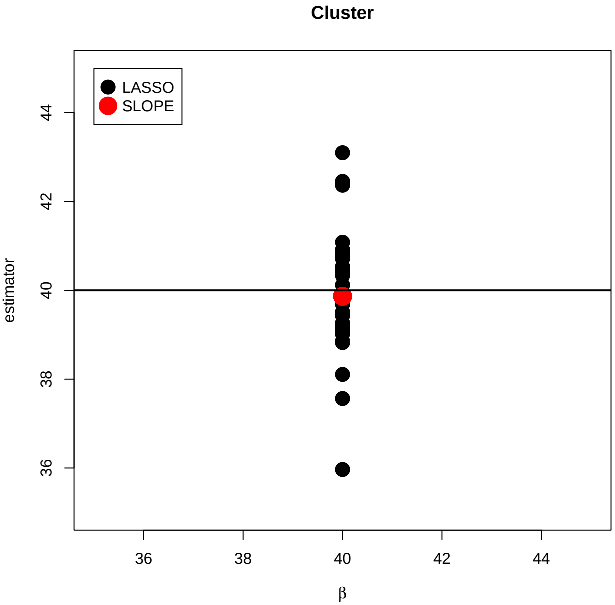
where

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

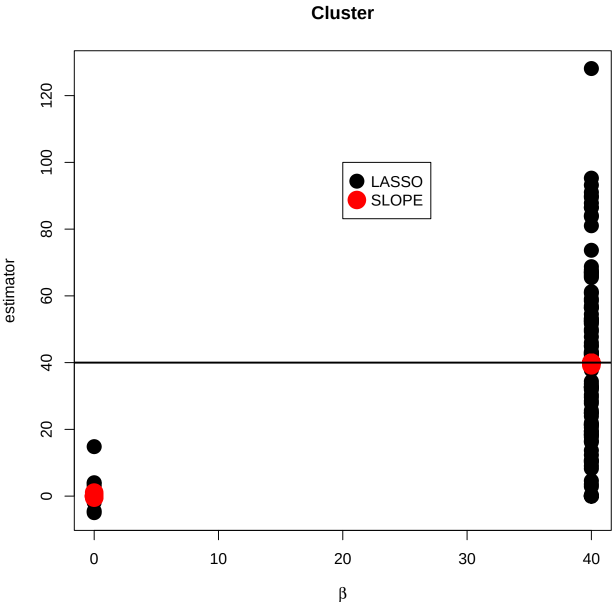
Theorem (Bogdan et al. (2022))

SLOPE can properly identify a given SLOPE model if and only if the irrepresentability condition is satisfied and the signal is strong enough.

LASSO vs SLOPE, $\rho_{ij} = 0.9^{|i-j|}$, $n = 100$, $p = 200$, $k = 30$



LASSO vs SLOPE, $\rho_{ij} = 0.9^{|i-j|}$, $n = 100$, $p = 200$, $k = 30$



Clustering in financial applications

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

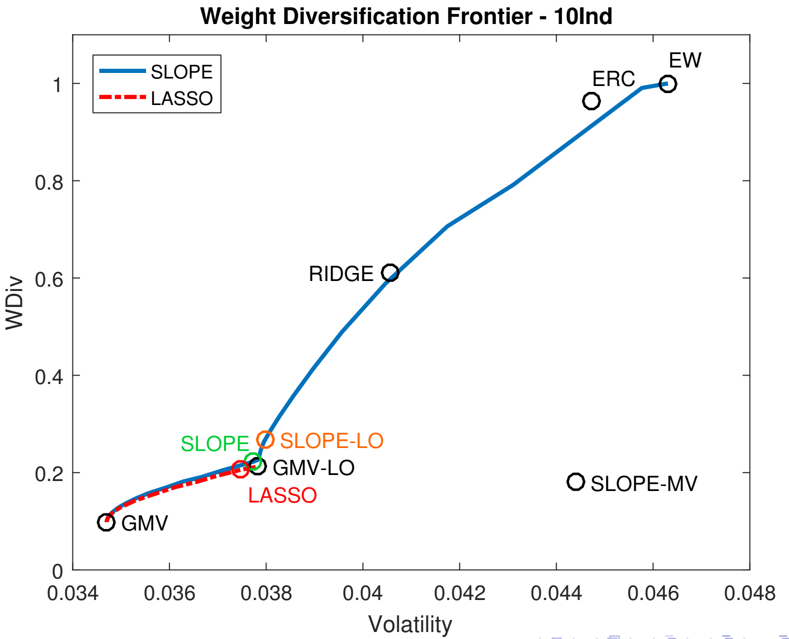
Portfolio Optimization, (Kremmer et al, 2020, JBF)

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

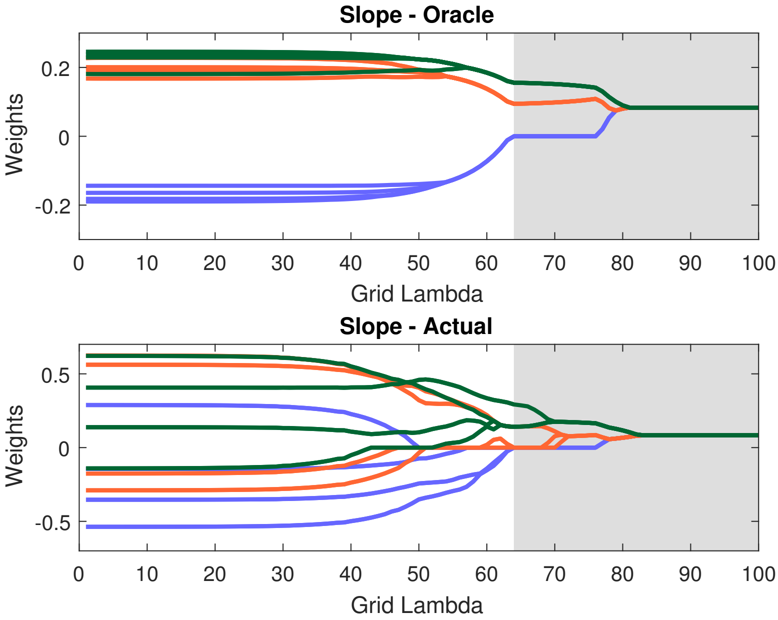
$$\min_{w \in \mathbb{R}^k} w' \Sigma w + J_\lambda(w) \tag{2}$$

$$\text{s.t.} \sum_{i=1}^{k} w_i = 1 \tag{3}$$

Evolution of Portfolio



SLOPE clustering



Gaussian Graphical Model

$$Y = (Y_1, \ldots, Y_p) \sim N(0, \Sigma)$$

Gaussian Graphical Model

$$Y = (Y_1, \ldots, Y_p) \sim N(0, \Sigma)$$

$$\Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

Gaussian Graphical Model

$$Y = (Y_1, \ldots, Y_p) \sim N(0, \Sigma)$$

$$\Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

Y_i is conditionally independent of Y_j if and only if $\Omega_{ij} = 0$

Gaussian Graphical Model

$$Y = (Y_1, \ldots, Y_p) \sim N(0, \Sigma)$$

$$\Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

Y_i is conditionally independent of Y_j if and only if $\Omega_{ij} = 0$

Goal - identification of nonzero elements of Ω

Graphical SLOPE

Riccobello, B., Bonaccolto, Kremer, Paterlini, Sobczyk, arxiv 2022

Graphical SLOPE

Riccobello, B., Bonaccolto, Kremer, Paterlini, Sobczyk, arxiv 2022

$$X_{n\times p} : X_{1\cdot}, \ldots, X_{n\cdot} \;\; \text{iid} \;\; N(0, \Sigma)$$

Graphical SLOPE

Riccobello, B., Bonaccolto, Kremer, Paterlini, Sobczyk, arxiv 2022

$$X_{n\times p} : X_{1\cdot}, \ldots, X_{n\cdot} \text{ iid } N(0, \Sigma)$$

$$S = \frac{1}{n}X'X\text{- sample covariance matrix}$$

Graphical SLOPE

Riccobello, B., Bonaccolto, Kremer, Paterlini, Sobczyk, arxiv 2022

$$X_{n\times p} : X_{1\cdot}, \dots, X_{n\cdot} \text{ iid } N(0,\Sigma)$$

$$S = \frac{1}{n}X'X\text{- sample covariance matrix}$$

$$L(\Omega, X) = C + \frac{n}{2}\log\det\Omega - \frac{n}{2}tr(S\Omega). \tag{4}$$

Graphical SLOPE

Riccobello, B., Bonaccolto, Kremer, Paterlini, Sobczyk, arxiv 2022

$$X_{n\times p} : X_{1\cdot}, \dots, X_{n\cdot} \;\; \text{iid} \;\; N(0, \Sigma)$$

$$S = \frac{1}{n} X' X \text{- sample covariance matrix}$$

$$L(\Omega, X) = C + \frac{n}{2} \log\det\Omega - \frac{n}{2} tr(S\Omega). \tag{4}$$

LASSO:

$$\widehat{\Omega}_L = \operatorname*{arg\,max}_{\Omega\in Sym_+^p} \left[\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1\right] \quad ,$$

Graphical SLOPE

Riccobello, B., Bonaccolto, Kremer, Paterlini, Sobczyk, arxiv 2022

$$X_{n\times p} : X_{1\cdot}, \ldots, X_{n\cdot} \;\text{ iid }\; N(0, \Sigma)$$

$$S = \frac{1}{n} X'X\text{- sample covariance matrix}$$

$$L(\Omega, X) = C + \frac{n}{2}\log\det\Omega - \frac{n}{2}tr(S\Omega). \tag{4}$$

LASSO:

$$\widehat{\Omega}_L = \underset{\Omega\in Sym_+^p}{\arg\max}\left[\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1\right] \quad ,$$

SLOPE:

$$\widehat{\Omega}_{SL} = \underset{\Omega\in Sym_+^p}{\arg\max}\left[\log\det\Omega - tr(S\Omega) - J_\lambda(\Omega)\right] \quad ,$$

Selection of the tuning parameter for Glasso

Banerjee and d'Aspremont (2008), FWER control for block diagonal matrices:

$$\lambda = \frac{t_{n-2}(1 - \frac{\alpha}{2p^2})}{\sqrt{n - 2 + t^2_{n-2}(1 - \frac{\alpha}{2p^2})}}$$

Selection of the tuning parameter for Glasso

Banerjee and d'Aspremont (2008), FWER control for block diagonal matrices:

$$\lambda = \frac{t_{n-2}(1 - \frac{\alpha}{2p^2})}{\sqrt{n - 2 + t^2_{n-2}(1 - \frac{\alpha}{2p^2})}}$$

Ricobello et al. (2022),

$$\lambda^{Bon} = \frac{t_{n-2}(1 - \frac{\alpha}{2m})}{\sqrt{n - 2 + t^2_{n-2}(1 - \frac{\alpha}{2m})}} \quad ,$$

with $m = \frac{p(p-1)}{2}$.

Selection of the tuning parameter for Gslope

Ricobello et al. (2022)

$$m = \frac{p(p-1)}{2}$$

$$\lambda_k^{\text{Holm}} = \frac{t_{n-2}(1 - \frac{\alpha}{2(m+1-k)})}{\sqrt{n - 2 + t_{n-2}^2(1 - \frac{\alpha}{2(m+1-k)})}}$$

Selection of the tuning parameter for Gslope

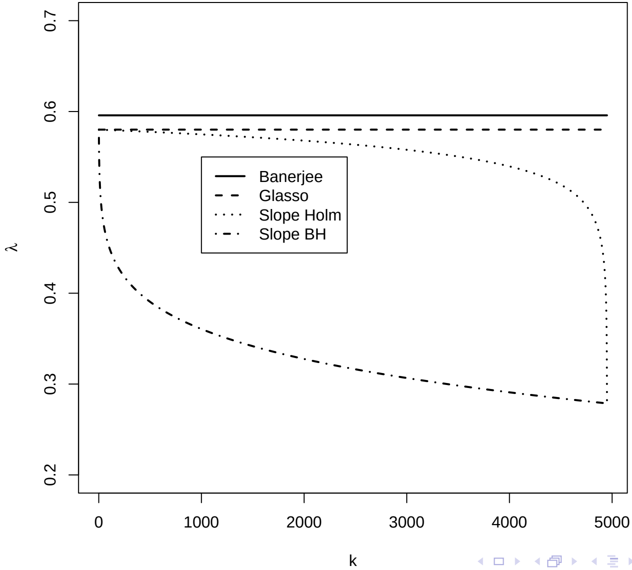
Ricobello et al. (2022)

$$m = \frac{p(p-1)}{2}$$

$$\lambda_k^{\mathrm{Holm}} = \frac{t_{n-2}(1 - \frac{\alpha}{2(m+1-k)})}{\sqrt{n - 2 + t_{n-2}^2(1 - \frac{\alpha}{2(m+1-k)})}}$$

$$\lambda_k^{\mathrm{BH}} = \frac{t_{n-2}(1 - \frac{\alpha k}{2m})}{\sqrt{n - 2 + t_{n-2}^2(1 - \frac{\alpha k}{2m})}}$$

Different tuning sequences, $p = 100$ $(m = 4950)$, $n = 50$



FWER control by Glasso

C_k - the connectivity component of k^{th} node

Theorem

If the tuning parameter for Glasso is equal to λ^{Bon} then

$$P\left(\forall k \in \{1, \ldots, p\} : \hat{C}_k^{\lambda} \subset C_k\right) \geq 1 - \alpha \quad .$$

FWER control by Gslope

C_k - the connectivity component of k^{th} node

Theorem

If the tuning sequence for Gslope is equal to λ^{Holm} and the sample correlation coefficients are such that the Hochberg multiple testing procedure controls FWER then

$$P\left(\forall k \in \{1, \ldots, p\} : \hat{C}_k^{\lambda} \subset C_k\right) \geq 1 - \alpha \quad .$$

Motivation

- SLOPE dual problem

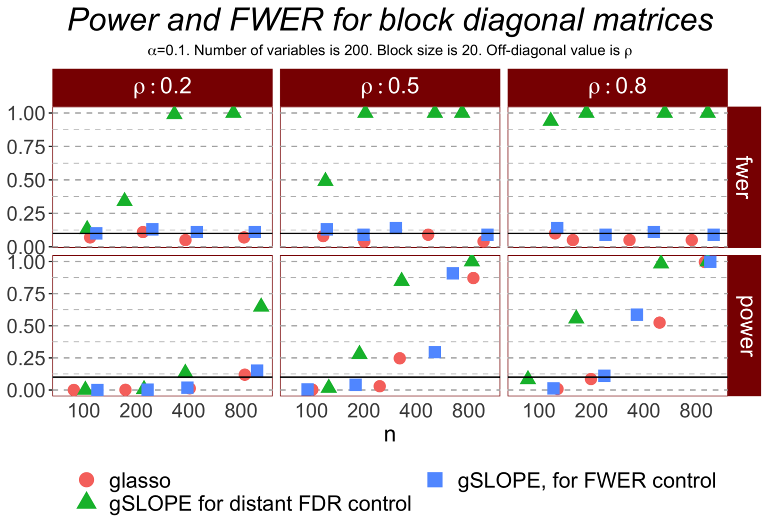
$$\hat{W} = \underset{J_\lambda^D(W-S)\leq 1}{\arg\max} \log det(W) \quad . \tag{5}$$

- Distribution of the sample correlation coefficient when $\rho_{ij} = 0$

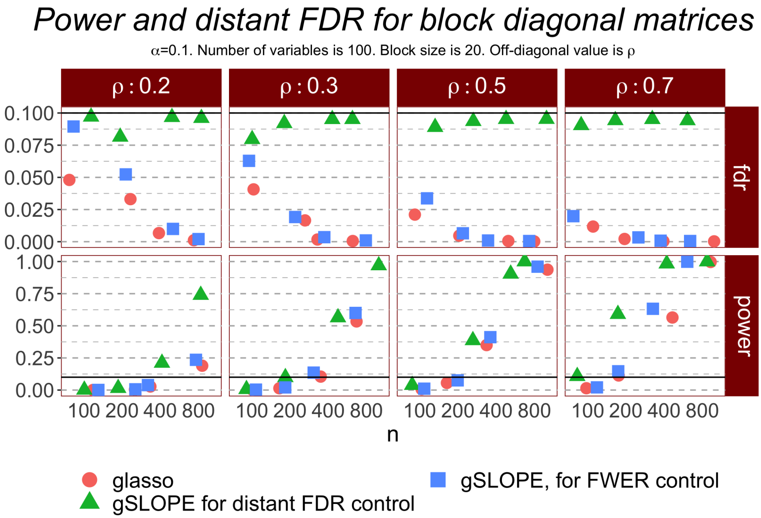
$$\sqrt{n-2}\frac{S_{ij}}{\sqrt{1-S_{ij}^2}} \sim t(n-2)$$

where $t(n-2)$ is Student distribution with $n-2$ degrees of freedom.

FWER control



Distant FDR Control



TSlope, Riccobello

The rows of X are independent from the multivariate t-distribution with the covariance matrix Φ

TSlope, Riccobello

The rows of X are independent from the multivariate t-distribution with the covariance matrix Φ

$$\Omega = \Phi^{-1}$$

TSlope, Riccobello

The rows of X are independent from the multivariate t-distribution with the covariance matrix Φ

$$\Omega = \Phi^{-1}$$

Finegold and Drton (2011) - represent t-distribution as a scale mixture of normals, estimate parameters using EM algorithm with scales as hidden latent variables

TSlope, Riccobello

The rows of X are independent from the multivariate t-distribution with the covariance matrix Φ

$$\Omega = \Phi^{-1}$$

Finegold and Drton (2011) - represent t-distribution as a scale mixture of normals, estimate parameters using EM algorithm with scales as hidden latent variables

Riccobello (2021): Use gSLOPE in the M-step

TSlope, Riccobello

The rows of X are independent from the multivariate t-distribution with the covariance matrix Φ

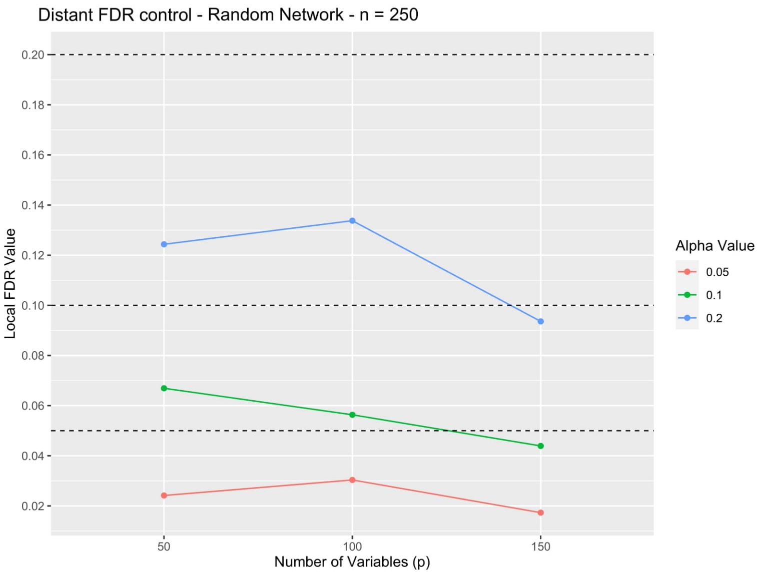
$$\Omega = \Phi^{-1}$$

Finegold and Drton (2011) - represent t-distribution as a scale mixture of normals, estimate parameters using EM algorithm with scales as hidden latent variables

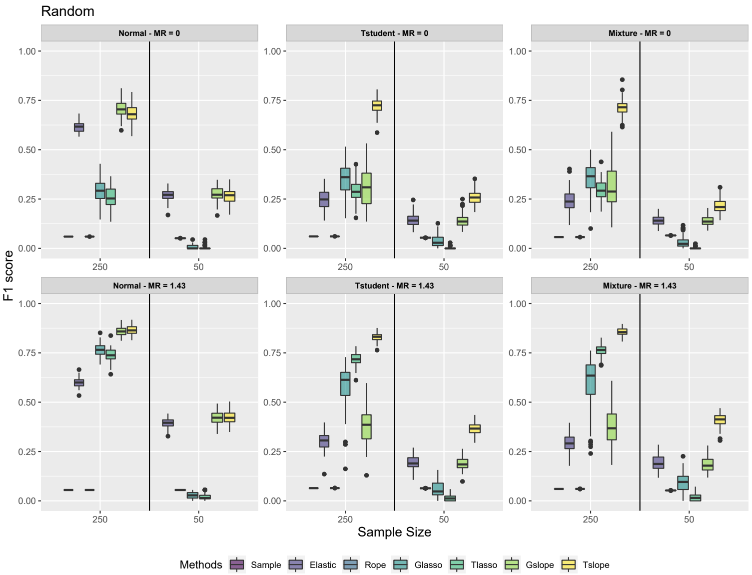
Riccobello (2021): Use gSLOPE in the M-step

Shiny application comparing different methods for estimation of the covariance matrix illustrates very good properties of tSLOPE as compared to other methods

distant FDR control

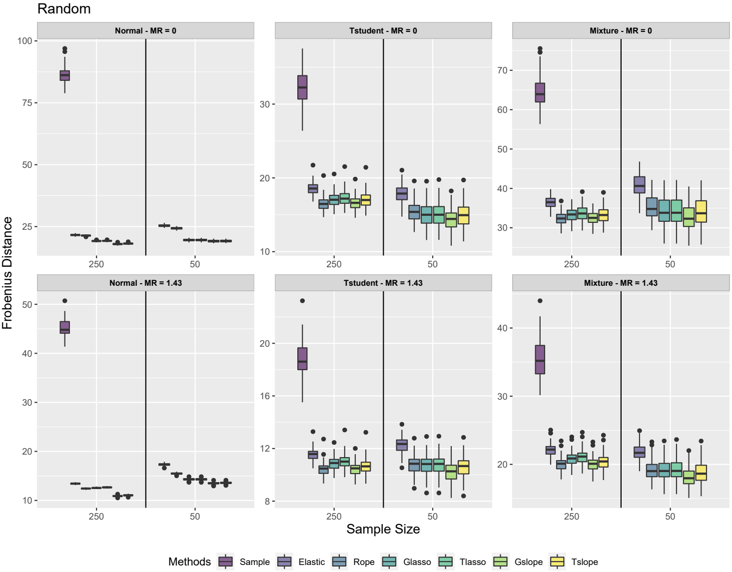


F1 scores



Random
Normal - MR = 0
Tstudent - MR = 0
Mixture - MR = 0
Normal - MR = 1.43
Tstudent - MR = 1.43
Mixture - MR = 1.43
F1 score
1.00
0.75
0.50
0.25
0.00
250
50
Sample Size
Methods
Sample
Elastic
Rope
Glasso
Tlasso
Gslope
Tslope

Frobenius distance



Gene expression network