Recent developments on the Sorted L-One Penalized Estimator

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

Montpellier Statistics Seminar

25th of October, 2021

Outline

- Support recovery by LASSO
- Sorted L-One Penalize Estimator
  - Model recovery and estimation properties
  - Applications in finance
  - Adaptive Bayesian SLOPE

Motivation (Jiang, B., Josse, Majewski, Miasojedow, Rockova, TraumaBase Group, JCGS, 2021)

- *Traumabase*® data:
  20000 major trauma patients $\times$ 250 measurements..

Accident type	**Age**	**Sex**	**Blood pressure**	**Lactate**	**Temperature**	**Platelet** (G/L)
Falling	50	M	140		35.6	150
Fire	28	F		4.8	36.7	250
Knife	30	M	120	1.2		270
Traffic accident	23	M	110	3.6	35.8	170
Knife	33	M	106		36.3	230
Traffic accident	58	F	150		38.2	400

Motivation (Jiang, B., Josse, Majewski, Miasojedow, Rockova, TraumaBase Group, JCGS, 2021)

- *Traumabase*® data:
  20000 major trauma patients $\times$ 250 measurements..

Accident type	**Age**	**Sex**	**Blood pressure**	**Lactate**	**Temperature**	**Platelet** (G/L)
Falling	50	M	140		35.6	150
Fire	28	F		4.8	36.7	250
Knife	30	M	120	1.2		270
Traffic accident	23	M	110	3.6	35.8	170
Knife	33	M	106		36.3	230
Traffic accident	58	F	150		38.2	400

- **Objective:**
  Develop models to help emergency doctors make decisions.
  Measurements $\xrightarrow{\text{Predict}}$ Platelet $\Rightarrow$ $X \xrightarrow{\text{Regression}} y$

Motivation (Jiang, B., Josse, Majewski, Miasojedow, Rockova, TraumaBase Group, JCGS, 2021)

- *Traumabase*® data:
  20000 major trauma patients $\times$ 250 measurements..

Accident type	Age	Sex	Blood pressure	Lactate	Temperature	Platelet (G/L)
Falling	50	M	140		35.6	150
Fire	28	F		4.8	36.7	250
Knife	30	M	120	1.2		270
Traffic accident	23	M	110	3.6	35.8	170
Knife	33	M	106		36.3	230
Traffic accident	58	F	150		38.2	400

- **Objective:**
  Develop models to help emergency doctors make decisions.
  Measurements $\xrightarrow{\text{Predict}}$ Platelet $\Rightarrow$ $X \xrightarrow{\text{Regression}} y$
- **Challenge :**
  How to **select** relevant measurements with **missing values**?

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
- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

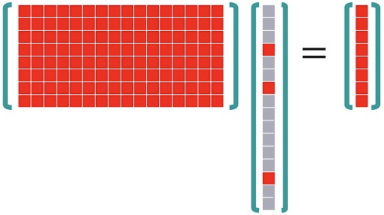
Model selection in high-dimension

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 (including $p \geq n$)
- β is **sparse** with $k < n$ nonzero coefficients



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

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
- 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
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
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:

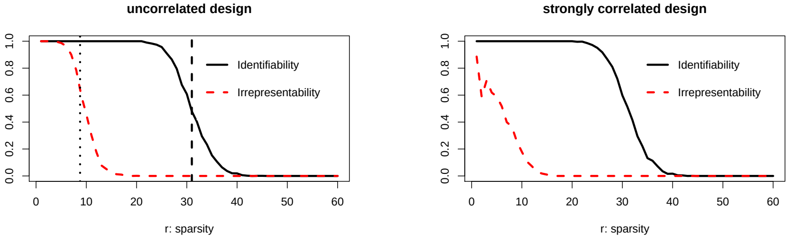
$$\|X' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

When

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

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

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).

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

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

Theorem (Tardivel, B., to appear in SJS)

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.

Modifications of LASSO

Threshold LASSO estimates: e.g. by knockoffs (Barber and Candés (*AOS*, 2015), Candés,Fan, Janson, Lv (*JRSSB, 2018), Weinstein, Su, B., Barber, Candés (arxiv, 2020) or GIC (Pokarowski, Mielniczuk (JMLR, 2015))*

Modifications of LASSO

Threshold LASSO estimates: e.g. by knockoffs (Barber and Candés (*AOS*, 2015), Candés,Fan, Janson, Lv (*JRSSB, 2018), Weinstein, Su, B., Barber, Candés (arxiv, 2020) or GIC (Pokarowski, Mielniczuk (JMLR, 2015))*

Use LASSO to obtain weights for adaptive LASSO [Zou, JASA 2006], [Candès, Wakin and Boyd, J. Fourier Anal. Appl. 2008]

$$\beta_{aL} = argmin_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda \sum_{i=1}^{p} w_i |b|_i \right\}, \tag{2}$$

where $w_i = \frac{1}{f(|\hat{\beta}_i|)}$, $\hat{\beta}_i$ *is the preliminary LASSO estimator.*

Related work

Mean field asymptotics for Approximate Message Algorithms (Bayatti, Montanari, 2012, IEEE Trans. Inf. Th.)

Related work

Mean field asymptotics for Approximate Message Algorithms (Bayatti, Montanari, 2012, IEEE Trans. Inf. Th.)

- Su, B., Candès (AOS, 2017) - LASSO can not identify the true model in the linear sparsity regime for Gaussian designs ($\frac{n}{p} \to \delta \in (0,1)$, $\frac{k}{p} \to \varepsilon \in (0,1), X_{ij} \sim N(0,\tau^2)$), precise (Power, FDP) tradeoff diagram.

.

Related work

Mean field asymptotics for Approximate Message Algorithms (Bayatti, Montanari, 2012, IEEE Trans. Inf. Th.)

- Su, B., Candès (AOS, 2017) - LASSO can not identify the true model in the linear sparsity regime for Gaussian designs ($\frac{n}{p} \to \delta \in (0,1)$, $\frac{k}{p} \to \varepsilon \in (0,1), X_{ij} \sim N(0,\tau^2)$), precise (Power, FDP) tradeoff diagram.
- Tardivel, B. (to appear in SJS) - thresholded LASSO can identify sufficiently large signals if $\varepsilon < \phi(\delta)$, where $\phi(\cdot)$ is the transition curve of Donoho and Tanner (2005)

.

Related work

Mean field asymptotics for Approximate Message Algorithms (Bayatti, Montanari, 2012, IEEE Trans. Inf. Th.)

- Su, B., Candès (AOS, 2017) - LASSO can not identify the true model in the linear sparsity regime for Gaussian designs ($\frac{n}{p} \to \delta \in (0,1)$, $\frac{k}{p} \to \varepsilon \in (0,1), X_{ij} \sim N(0,\tau^2)$), precise (Power, FDP) tradeoff diagram.
- Tardivel, B. (to appear in SJS) - thresholded LASSO can identify sufficiently large signals if $\varepsilon < \phi(\delta)$, where $\phi(\cdot)$ is the transition curve of Donoho and Tanner (2005)
- Weinstein, Su, B., Barber, Candès (arxive, 2020) - LASSO with knockoffs controls FDR at a given level and can identify sufficiently large signals if $\frac{\varepsilon}{2} < \phi\left(\frac{\delta}{2}\right)$.

Related work

Mean field asymptotics for Approximate Message Algorithms (Bayatti, Montanari, 2012, IEEE Trans. Inf. Th.)

- Su, B., Candès (AOS, 2017) - LASSO can not identify the true model in the linear sparsity regime for Gaussian designs ($\frac{n}{p} \to \delta \in (0,1)$, $\frac{k}{p} \to \varepsilon \in (0,1), X_{ij} \sim N(0,\tau^2)$), precise (Power, FDP) tradeoff diagram.
- Tardivel, B. (to appear in SJS) - thresholded LASSO can identify sufficiently large signals if $\varepsilon < \phi(\delta)$, where $\phi(\cdot)$ is the transition curve of Donoho and Tanner (2005)
- Weinstein, Su, B., Barber, Candès (arxive, 2020) - LASSO with knockoffs controls FDR at a given level and can identify sufficiently large signals if $\frac{\varepsilon}{2} < \phi\left(\frac{\delta}{2}\right)$.

Rejchel, B. (JMLR, 2020) - rank LASSO for the single index model ($Y = g(X\beta) + \varepsilon$), condition for consistency for thresholded and adaptive versions

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

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),

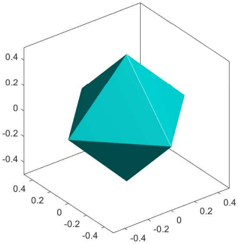
Bellec, Lecué, Tsybakov (Annals of Statistics, 2018):

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 $||\hat{\beta}-\beta||^2$.

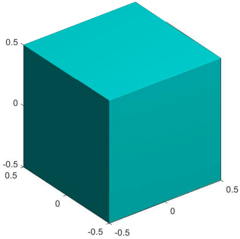
Fixed λ LASSO rate of convergence - $\frac{k}{n}\log(p)$

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

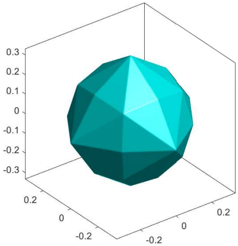
Unit balls for different SLOPE sequences by D.Brzyski



((a)) (2,2,2)



((b)) (2,0,0)



((c)) (3,2,1)

Clustering properties of SLOPE

- Schneider and Tardivel, arxive 2020 - class of models attainable by SLOPE
- Skalski, B., Graczyk, Kołodziejek, Tardivel, Wilczyński, in preparation - Irrepresentability condition for SLOPE

SLOPE model (Schneider, Tardivel, 2020)

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 et al. (2021)

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

Irrepresentability condition:

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

where

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

Theorem (Skalski et al. (2021))

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

Identifiability condition for SLOPE

Definition (Identifiability)

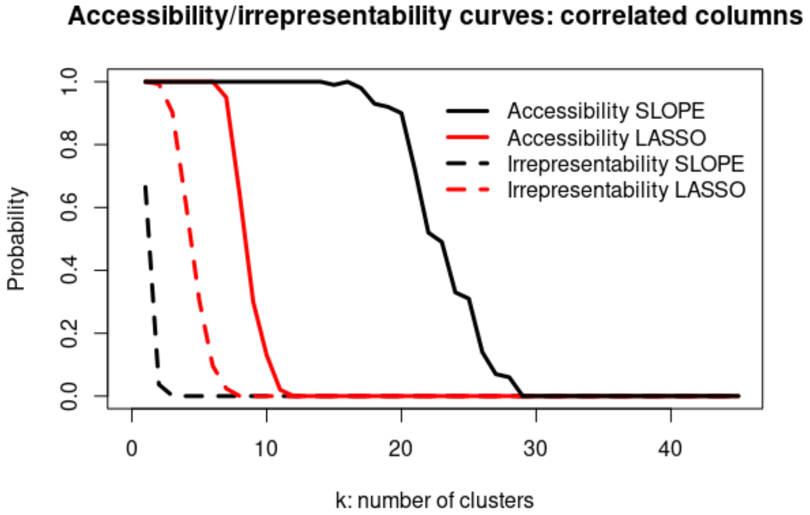
Let X be a $n \times p$ matrix. The vector $\beta \in R^p$ is said to be identifiable with respect to the SLOPE J_λ norm if the following implication holds

$$X\gamma = X\beta \text{ and } \gamma \neq \beta \Rightarrow J_\lambda(\gamma) > J_\lambda(\beta). \tag{3}$$

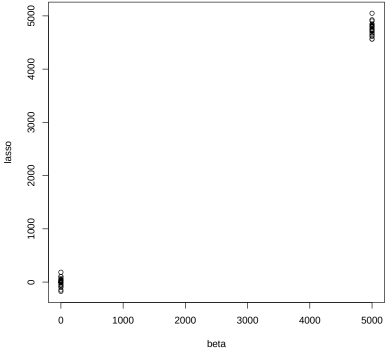
Theorem (Skalski et al. (2021))

For any sequence strictly decreasing positive sequence λ SLOPE can properly order the elements of $\hat{\beta}$ if and only if vector β is identifiable with respect to J_λ norm and $\min_{i \in I} |\beta_i|$ is sufficiently large.

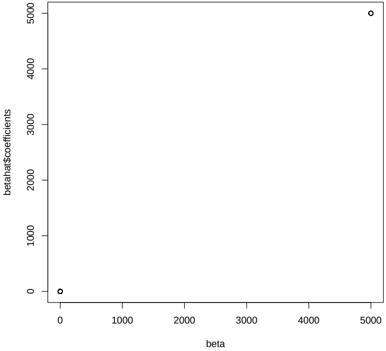
LASSO vs SLOPE, $\rho = 0.9$, $n = 50$, $p = 150$



$n = 100, p = 300, k = 25$, independent

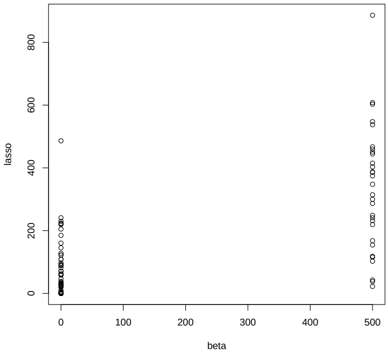


((d)) LASSO

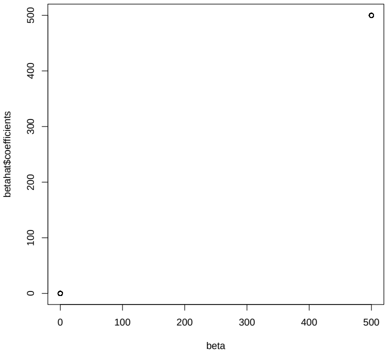


((e)) SLOPE

$n = 100, p = 300, k = 30, \rho = 0.7$

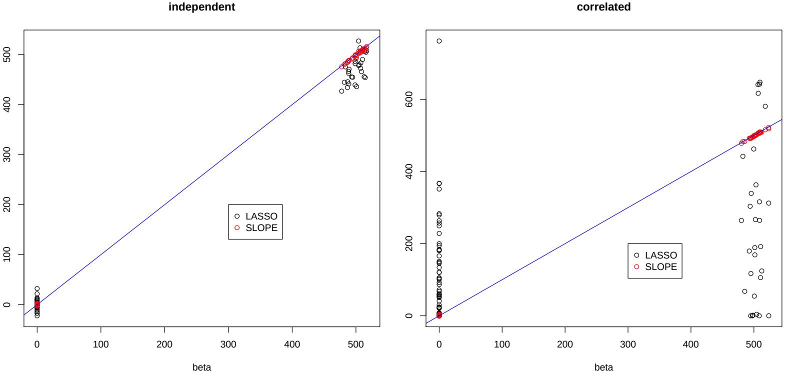


((f)) LASSO



((g)) SLOPE

$n = 100, p = 300, k = 30, \rho = 0.7$

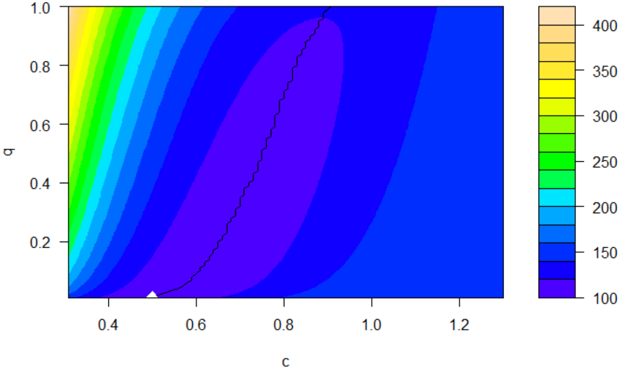


Heat Maps of $MSE(X\hat{\beta})$ by D. Nowakowski

Independent predictors

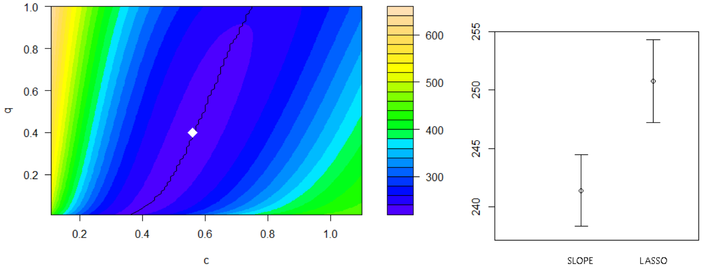
$$\lambda_i = c\Phi\left(1 - \frac{iq}{2p}\right), \quad n = p = 1000, k = 20$$

$$\text{for } \; i \in S, \;\; \beta_i = \sqrt{2\log\frac{p}{k}}$$



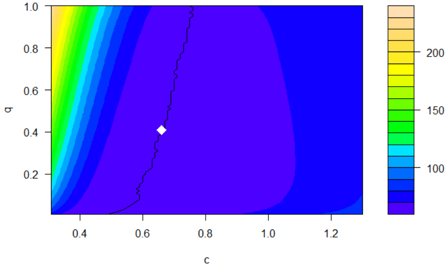
Independent predictors

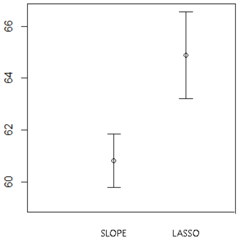
$$n = p = 1000, k = 100$$



Correlated predictors

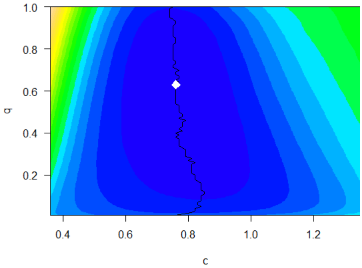
$$n = p = 1000, k = 20, \rho(X_i, X_j) = 0.5 \text{ for } i \neq j$$

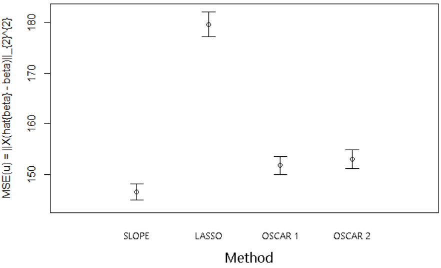




Correlated predictors

$$n = p = 1000, k = 100$$





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, 2021, to appear in *Quantitative Finance* - application for index tracking.

Different flavor of clustering, Kremer et al, 2021

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

Different flavor of clustering, Kremer et al, 2021

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 $\hat{\beta}$.

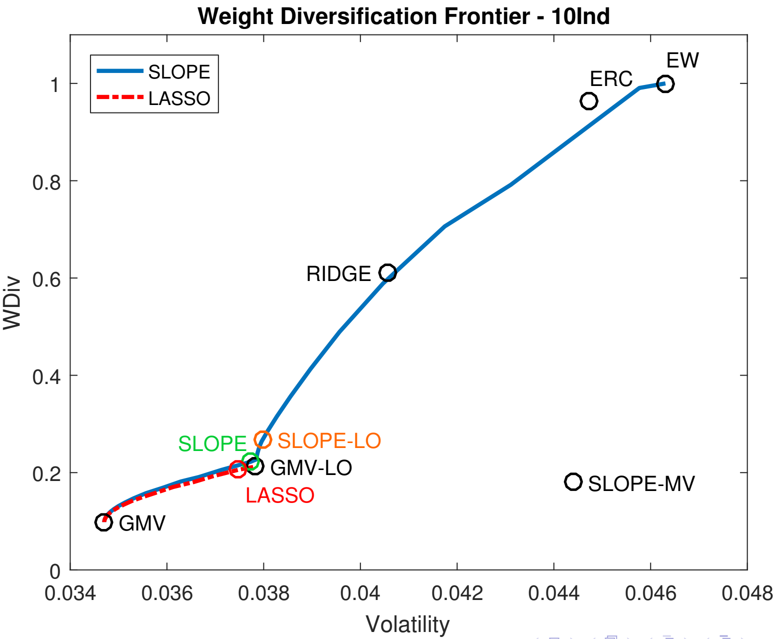
Portfolio Optimization, (Kremmer et al, 2020, JBF)

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

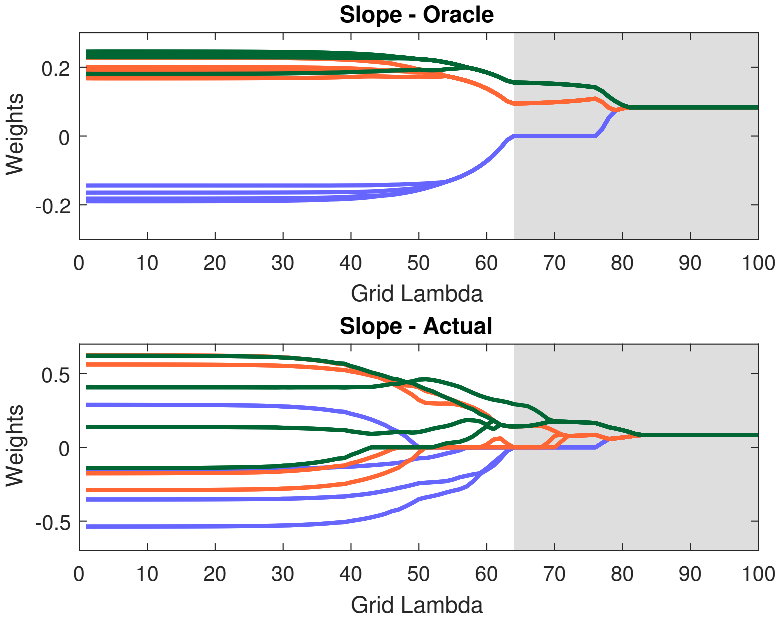
$$\min_{w\in\mathbb{R}^k} w'\Sigma w + J_\lambda(w) \tag{4}$$

$$\text{s.t.} \sum_{i=1}^{k} w_i = 1 \tag{5}$$

Evolution of Portfolio

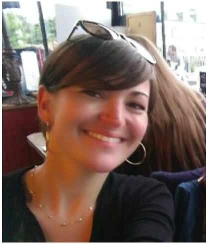


SLOPE clustering



Adaptive SLOPE with missing values (1)

W. Jiang, MB, J.Josse, S. Majewski, B.Miasojedow, V.Rockova, TraumaBase Group (to appear in JCGS)


traumabase.eu

Spike and Slab LASSO (Rockova, George, 2018)

LASSO has a Bayesian interpretation as a posterior mode under the Laplace prior

$$\pi(\beta) = C(\lambda) \prod_{i=1}^{n} {}^{-|\beta_i|\lambda}$$

Spike and Slab LASSO (Rockova, George, 2018)

LASSO has a Bayesian interpretation as a posterior mode under the Laplace prior

$$\pi(\beta) = C(\lambda) \prod_{i=1}^{n} {}^{-|\beta_i|\lambda}$$

Spike and Slab LASSO uses a spike and slab Laplace prior:

$$\gamma = (\gamma_1, \ldots, \gamma_p)$$

$\gamma_i = 1$ if β_i is "large" and $\gamma_i = 0$ if β_i is "small"

$$\pi(\beta|\lambda, \gamma) \propto c^{\sum_{i=1}^{p} 1(\gamma_i=1)} \prod_{i=1}^{p} {}^{-w_i|\beta_i|\lambda},$$

where $w_i = 1$ if $\gamma_i = 0$ and $w_i = c \in (0, 1)$ if $\gamma_i = 1$.

Spike and Slab LASSO (2)

The maximum aposteriori rule is given by reweighted LASSO

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda \sum_{i=1}^{p} w_i |b_i|$$

$$w_i = c\gamma_i + (1 - \gamma_i)$$

Spike and Slab LASSO (2)

The maximum aposteriori rule is given by reweighted LASSO

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda \sum_{i=1}^{p} w_i |b_i|$$

$$w_i = c\gamma_i + (1 - \gamma_i)$$

Prior for γ: $\gamma_1, \ldots, \gamma_p$ are iid such that

$$P(\gamma_i = 1) = \theta = 1 - P(\gamma_i = 0)$$

Spike and Slab LASSO (2)

The maximum aposteriori rule is given by reweighted LASSO

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2} ||y - Xb||_2^2 + \lambda \sum_{i=1}^{p} w_i |b_i|$$

$$w_i = c\gamma_i + (1 - \gamma_i)$$

Prior for γ: $\gamma_1, \ldots, \gamma_p$ are iid such that

$$P(\gamma_i = 1) = \theta = 1 - P(\gamma_i = 0)$$

In consecutive iterations γ_i is replaced with

$$\pi_i^t = P(\gamma_i = 1 | \beta^t, c) = \frac{c\theta e^{-c|\beta_i^t|\lambda_0}}{c\theta e^{-c|\beta_i^t|\lambda_0} + (1-\theta) e^{-|\beta_i^t|\lambda_0}}$$

and then a new estimate $\hat{\beta}^{t+1}$ is calculated by solving reweighted LASSO with the vector γ replaced with the vector π^t.

ABSLOPE (2)

Prior for β is given by

$$\pi(\beta|\gamma, c, \sigma^2) \propto c^{\sum_{i=1}^{n} 1(\gamma_i=1)} \prod_{i=1}^{n} e^{-w_i|\beta_i|\lambda_{r(W\beta,i)}},$$

where W is the diagonal matrix with $W_{ii} = w_i$ and $\lambda = \lambda^{BH}$

ABSLOPE (2)

Prior for β is given by

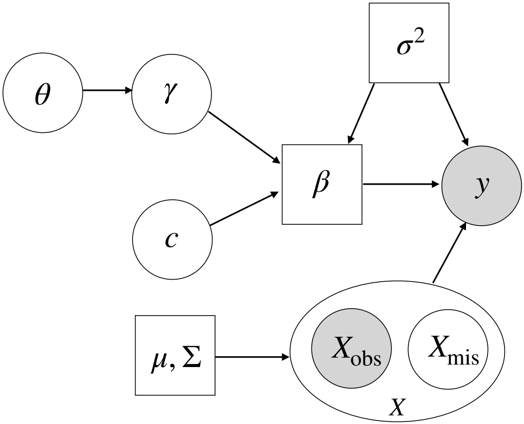
$$\pi(\beta|\gamma, c, \sigma^2) \propto c^{\sum_{i=1}^{n} 1(\gamma_i=1)} \prod_{i=1}^{n} e^{-w_i|\beta_i|\lambda_{r(W\beta,i)}},$$

where W is the diagonal matrix with $W_{ii} = w_i$ and $\lambda = \lambda^{BH}$

Missing at Random (MAR) mechanism under assumption $X_i = (X_{i1}, \ldots, X_{ip})$ is normally distributed:

$$X_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma), \quad i = 1, \cdots, n.$$

Graphical model of ABSLOPE



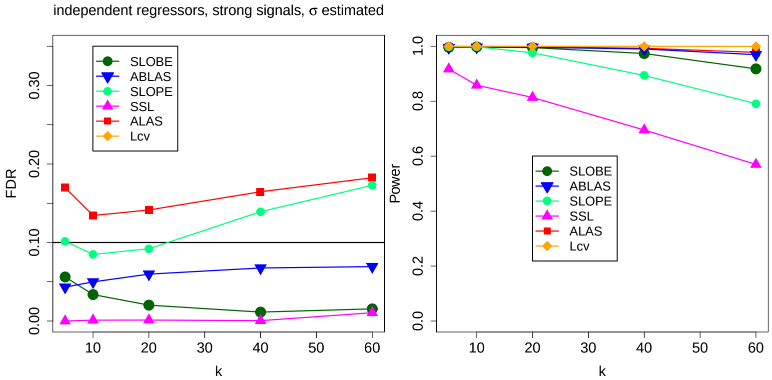
Stochastic approximation EM algorithm

- $\pi(\theta)$ - B(a,b), $\pi(c)$ - $U(0,1)$
- Gibbs sampling of latent variables : $\theta, c, \gamma, c, X_{mis}$
- Estimate parameters $\beta, \sigma, \mu, \Sigma$ by maximizing the complete-data likelihood with sampled values for the latent variables
- When $p > n$, Σ is estimated using the shrinkage estimator of Ledoit and Wolf (2004)
- Approximation of SAEM: ψ,

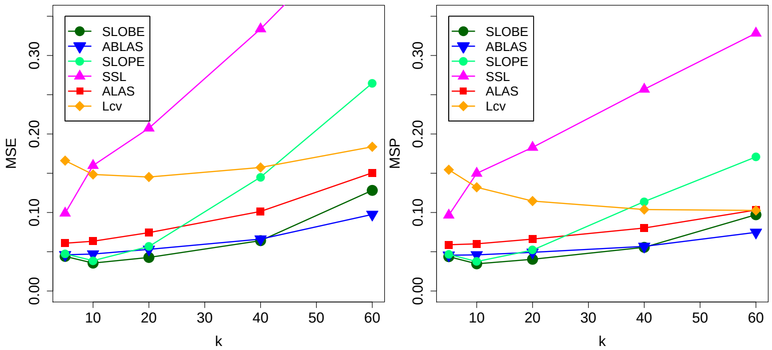
$$\psi^{t+1} = \psi^t + \eta_t \left[\hat{\psi}^t_{MLE} - \psi^t\right],$$

$\eta^t = 1$ for $t \in \{1, \ldots, t_0\}$ and $\eta^t = \frac{1}{t-t_0}$ for $t > t_0$

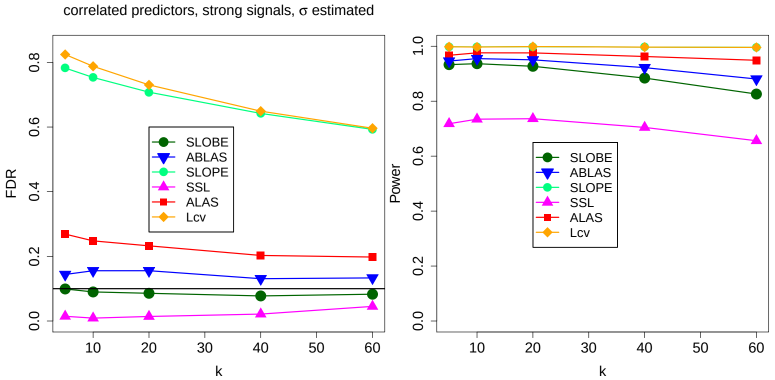
$n = p = 500$, $\rho = 0$, $Na = 10\%$, independent regressors



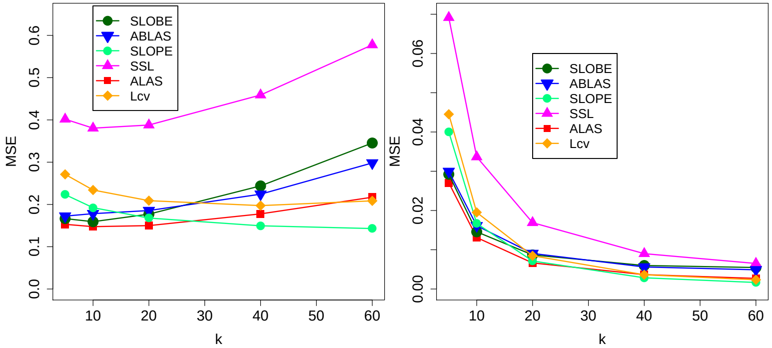
$n = p = 500$, $\rho = 0$, $Na = 10\%$, independent regressors



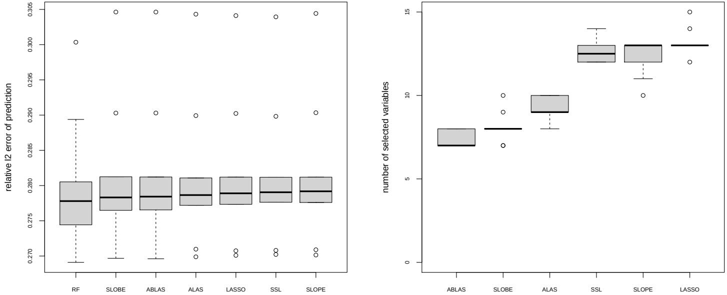
$n = p = 500$, $\rho = 0$, $Na = 10\%$, correlated regressors



$n = p = 500$, $\rho = 0$, $Na = 10\%$, correlated predictors



Motivating example



Rysunek: Empirical distribution of prediction errors of different methods over 10 replications for the TraumaBase data and of the number of variables selected by different methods.