Regularization methods in analysis of large data sets

Malgorzata Bogdan

University of Wroclaw

Baby steps beyond the horizon, 29/08/2022

Outline

- Basics of Linear Regression
- Ridge Regression
- LASSO (Least Absolute Shrinkage and Selection Operator)
- SLOPE (Sorted L-One Penalized Estimator)

Motivation: Paris Hospital, TraumaBase Group Data

- *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: Paris Hospital, TraumaBase Group Data

- *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 $\overset{\text{Predict}}{\longrightarrow}$ Platelet $\Rightarrow$ $X = (X_1, \ldots, X_p) \overset{\text{Regression}}{\longrightarrow} Y$

Motivation: Paris Hospital, TraumaBase Group Data

- *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 $\overset{\text{Predict}}{\longrightarrow}$ Platelet $\Rightarrow$ $X = (X_1, \ldots, X_p) \overset{\text{Regression}}{\longrightarrow} Y$
- **Challenge :**
  How to **select** relevant measurements ?

Linear regression model:

- $y = (y_i)$: vector of response of length n (platelets' counts)
- $X = (X_{ij})$: a design matrix of dimension $n \times p$ (values of explanatory variables)
- $\beta = (\beta_j)$: vector of regression coefficients of length p
- $\varepsilon \sim (0, \sigma^2 I_n)$

$$\text{for } \; i \in \{1, \ldots, n\}, \;\; y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$$

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

Model selection in high-dimension

Linear regression model:

- $y = (y_i)$: vector of response of length n (platelets' counts)
- $X = (X_{ij})$: a design matrix of dimension $n \times p$ (values of explanatory variables)
- $\beta = (\beta_j)$: vector of regression coefficients of length p
- $\varepsilon \sim (0, \sigma^2 I_n)$

$$\text{for } \; i \in \{1, \ldots, n\}, \;\; y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i$$

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

Assumptions:

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

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

$$\hat{Y} = X\hat{\beta}_{LS}: \quad \text{orthogonal projection of } Y \text{ on } colsp(X)$$

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

$$\hat{Y} = X\hat{\beta}_{LS} : \quad \text{orthogonal projection of } Y \text{ on } colsp(X)$$

$$\text{If } \; rank(X) = p \; \text{ then } \; \hat{Y} = X(X'X)^{-1}X'Y$$

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

$$\hat{Y} = X\hat{\beta}_{LS} : \quad \text{orthogonal projection of } Y \text{ on } colsp(X)$$

$$\text{If } \; rank(X) = p \; \text{ then } \; \hat{Y} = X(X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y$$

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

$$\hat{Y} = X\hat{\beta}_{LS} : \quad \text{orthogonal projection of } Y \text{ on } colsp(X)$$

$$\text{If } \; rank(X) = p \; \text{ then } \; \hat{Y} = X(X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y$$

$$Y \sim N(X\beta, \sigma^2 I_n)$$

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

$$\hat{Y} = X\hat{\beta}_{LS} : \quad \text{orthogonal projection of } Y \text{ on } colsp(X)$$

$$\text{If } \; rank(X) = p \; \text{ then } \; \hat{Y} = X(X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y$$

$$Y \sim N(X\beta, \sigma^2 I_n)$$

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Multiple regression model when $n > p$

$$\hat{\beta}_{LS} = argmin_{\beta \in R^p} ||Y - X\beta||^2$$

$$\hat{Y} = X\hat{\beta}_{LS} : \quad \text{orthogonal projection of } Y \text{ on } colsp(X)$$

$$\text{If } \; rank(X) = p \; \text{ then } \; \hat{Y} = X(X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} = (X'X)^{-1}X'Y$$

$$Y \sim N(X\beta, \sigma^2 I_n)$$

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

$\hat{\beta}_{LS}$ minimizes $MSE = E||\hat{\beta} - \beta||^2$ among all unbiased linear estimators.

Selection of important variables

Z-tests,

$$Z_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X'X)^{-1}[i,i]}}$$

Selection of important variables

Z-tests,

$$Z_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X'X)^{-1}[i,i]}}$$

When $\beta_i = 0$ then $Z_i \sim N(0,1)$.

Selection of important variables

Z-tests,

$$Z_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X'X)^{-1}[i,i]}}$$

When $\beta_i = 0$ then $Z_i \sim N(0,1)$.

At the significance level 0.05 we conclude that $\beta_i \neq 0$ if $|Z_i| > \Phi^{-1}(0.975) = 1.96$.

Selection of important variables

Z-tests,

$$Z_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X'X)^{-1}[i,i]}}$$

When $\beta_i = 0$ then $Z_i \sim N(0,1)$.

At the significance level 0.05 we conclude that $\beta_i \neq 0$ if $|Z_i| > \Phi^{-1}(0.975) = 1.96$.

Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

Selection of important variables

Z-tests,

$$Z_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X'X)^{-1}[i,i]}}$$

When $\beta_i = 0$ then $Z_i \sim N(0,1)$.

At the significance level 0.05 we conclude that $\beta_i \neq 0$ if $|Z_i| > \Phi^{-1}(0.975) = 1.96$.

Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

If elements of X are iid from $N(0,1)$ then $X'X$ has a Wishart distribution and the elements on its diagonal have the expected value equal to n.

Selection of important variables

Z-tests,

$$Z_i = \frac{\hat{\beta}_i}{\sigma\sqrt{(X'X)^{-1}[i,i]}}$$

When $\beta_i = 0$ then $Z_i \sim N(0,1)$.

At the significance level 0.05 we conclude that $\beta_i \neq 0$ if $|Z_i| > \Phi^{-1}(0.975) = 1.96$.

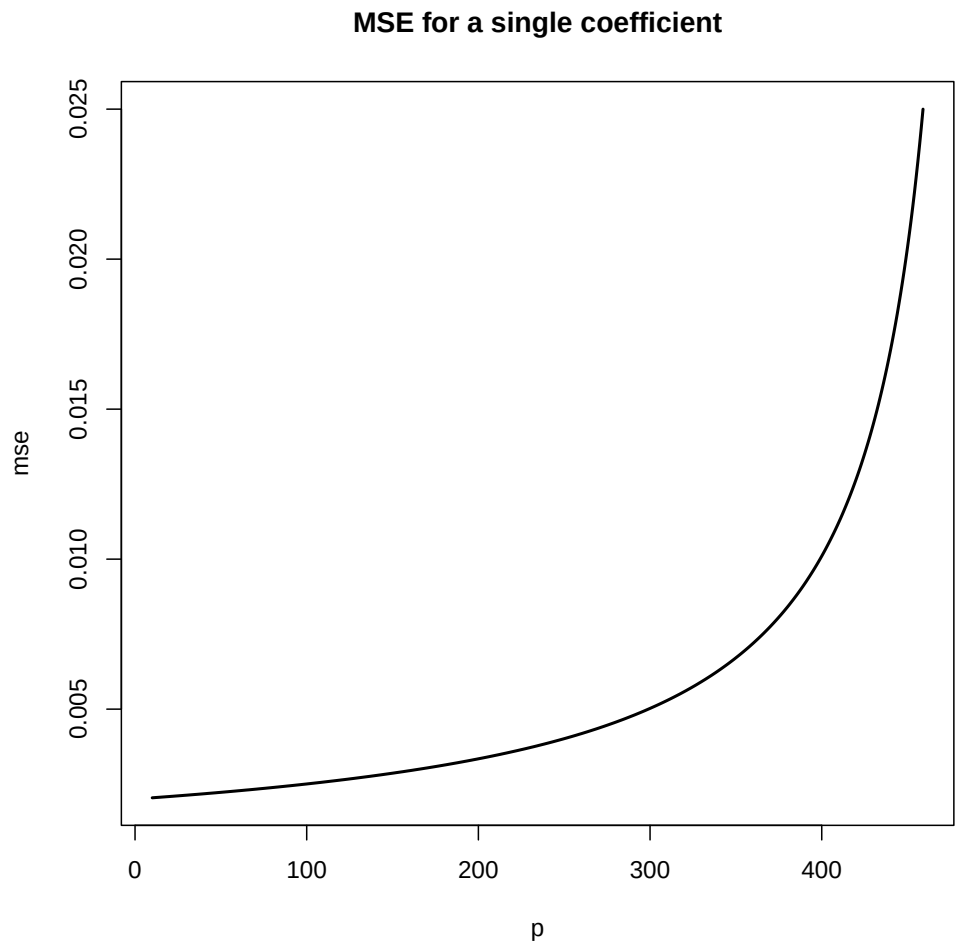
Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

If elements of X are iid from $N(0,1)$ then $X'X$ has a Wishart distribution and the elements on its diagonal have the expected value equal to n.

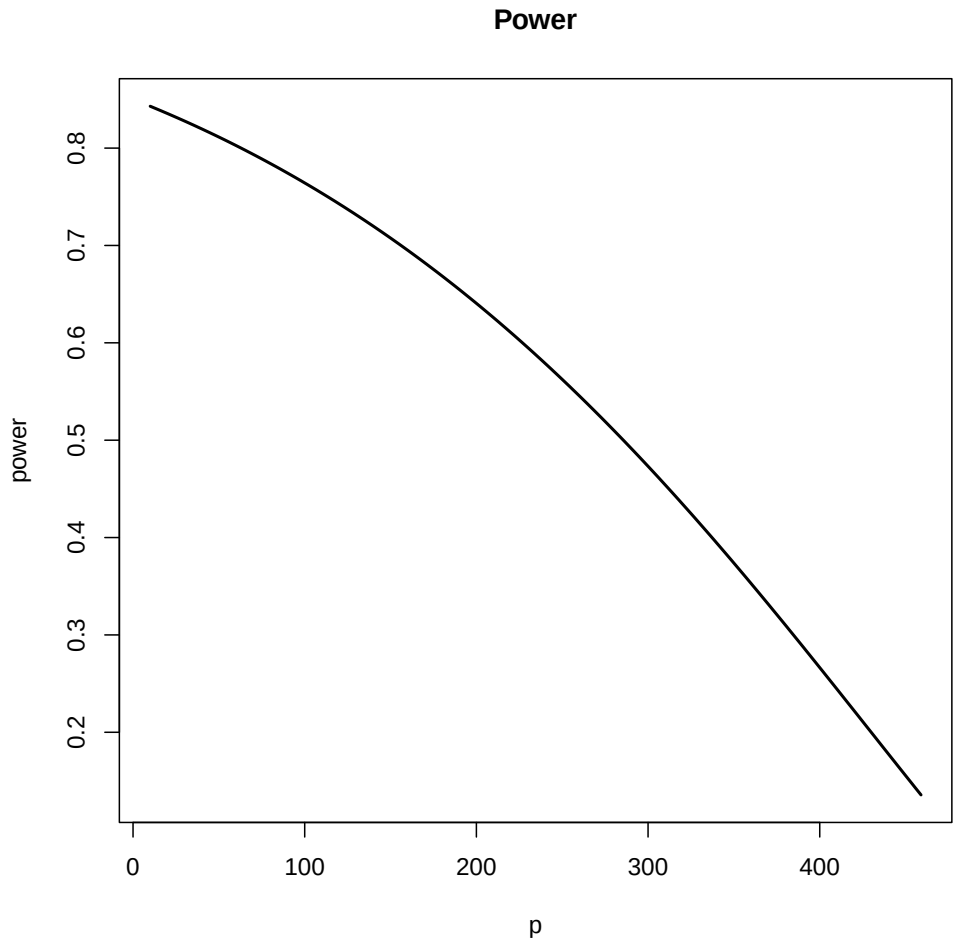
But $(X'X)^{-1}$ has the inverse Wishart distribution and the expected values of the elements on the diagonal are equal to $\frac{1}{n-p-1}$ and increase as p approaches n.

Inflation of MSE

$$n = 500, \quad MSE = E(\hat{\beta}_i - \beta_i)^2$$



Loss of Power



Model selection

Model selection in multiple regression - identification of important variables

Model selection

Model selection in multiple regression - identification of important variables

The residual error $RSS = ||Y - \hat{Y}||^2$ never increases when new variables are added into the model. Thus, minimization of RSS is not a good criterion for model selection.

Model selection

Model selection in multiple regression - identification of important variables

The residual error $RSS = ||Y - \hat{Y}||^2$ never increases when new variables are added into the model. Thus, minimization of RSS is not a good criterion for model selection.

Also, RSS is not a good measure of the prediction error.

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \ ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \ ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \ ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

If $\mu = E(Y) = X\beta$, then

$$PE = E||\mu + \epsilon^* - \hat{Y}||^2 = E||\mu - \hat{Y}||^2 + n\sigma^2$$

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \ ,$$

where ϵ^* is independent on the noise term ϵ in the training sample

We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

If $\mu = E(Y) = X\beta$, then

$$PE = E||\mu + \epsilon^* - \hat{Y}||^2 = E||\mu - \hat{Y}||^2 + n\sigma^2$$

$$RSS = ||Y - \hat{Y}||^2$$

Prediction error of linear operators

If $\hat{Y} = \hat{\mu} = M_{n\times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$

Prediction error of linear operators

If $\hat{Y} = \hat{\mu} = M_{n\times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

Prediction error of linear operators

If $\hat{Y} = \hat{\mu} = M_{n\times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

If $rank(X) = p$ then the unbiased estimator of the prediction error is equal to

$$\hat{PE} = RSS + 2\sigma^2 p \ .$$

Prediction error of linear operators

If $\hat{Y} = \hat{\mu} = M_{n\times n} Y$ then
$PE = E(RSS) + 2\sigma^2 Tr(M)$
In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

If $rank(X) = p$ then the unbiased estimator of the prediction error is equal to

$$\hat{PE} = RSS + 2\sigma^2 p \ .$$

Minimizing $\hat{PE}$ coincides with Akaike Information Criterion (AIC, 1974) which suggests selecting the model for which $RSS + 2\sigma^2 p$ is minimal.

Ridge regression (1)

Number of all possible regression models - 2^p

Ridge regression (1)

Number of all possible regression models - 2^p

Identifying the model which optimizes AIC in NP-hard.

Ridge regression (1)

Number of all possible regression models - 2^p

Identifying the model which optimizes AIC in NP-hard.

Solution - use a convex penalty function

Ridge regression (1)

Number of all possible regression models - 2^p

Identifying the model which optimizes AIC in NP-hard.

Solution - use a convex penalty function

Ridge regression:

$$\hat{\beta} = argmin_{b \in R^p} L(b) \quad \text{, where} \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

Ridge regression (1)

Number of all possible regression models - 2^p

Identifying the model which optimizes AIC in NP-hard.

Solution - use a convex penalty function

Ridge regression:

$$\hat{\beta} = argmin_{b \in R^p} L(b) \quad \text{, where} \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

$$\frac{\partial L(b)}{\partial b} = -2X'(Y - Xb) + 2\gamma b = 0$$

Ridge regression (1)

Number of all possible regression models - 2^p

Identifying the model which optimizes AIC in NP-hard.

Solution - use a convex penalty function

Ridge regression:

$$\hat{\beta} = argmin_{b \in R^p} L(b) \quad \text{, where} \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

$$\frac{\partial L(b)}{\partial b} = -2X'(Y - Xb) + 2\gamma b = 0$$

$$-X'Y + (X'X + \gamma I)b = 0 \quad \Leftrightarrow \quad b = (X'X + \gamma I)^{-1} X'Y$$

Ridge regression (2)

$$\hat{\beta}_R = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

Ridge regression (2)

$$\hat{\beta}_R = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$E(\hat{\beta}_R) = (X'X + \gamma I)^{-1} X'X\beta$$

Ridge regression (2)

$$\hat{\beta}_R = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$E(\hat{\beta}_R) = (X'X + \gamma I)^{-1} X'X\beta$$

When $E||\hat{\beta}_R - \beta||^2 < E||\hat{\beta}_{LS} - \beta||^2$?

Ridge regression (2)

$$\hat{\beta}_R = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$E(\hat{\beta}_R) = (X'X + \gamma I)^{-1} X'X\beta$$

When $E||\hat{\beta}_R - \beta||^2 < E||\hat{\beta}_{LS} - \beta||^2$?

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}\hat{\beta}_{LS}$$

Ridge regression (2)

$$\hat{\beta}_R = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$E(\hat{\beta}_R) = (X'X + \gamma I)^{-1} X'X\beta$$

When $E||\hat{\beta}_R - \beta||^2 < E||\hat{\beta}_{LS} - \beta||^2$?

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}\hat{\beta}_{LS}$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Ridge regression (2)

$$\hat{\beta}_R = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$E(\hat{\beta}_R) = (X'X + \gamma I)^{-1} X'X\beta$$

When $E||\hat{\beta}_R - \beta||^2 < E||\hat{\beta}_{LS} - \beta||^2$?

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}\hat{\beta}_{LS}$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Otherwise, when

$$\gamma < \frac{2p\sigma^2}{||\beta||^2 - p\sigma^2}$$

Least Absolute Shrinkage and Selection Operator (LASSO)

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

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

For a convex function $f : R^p \rightarrow R$ we define the subdifferential as

$$\partial_f(b) = \{v \in \mathbb{R}^p : f(z) - f(b) \geq v'(z - b)\ \forall z \in \mathbb{R}^p\}.$$

.

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

For a convex function $f : R^p \to R$ we define the subdifferential as

$$\partial_f(b) = \{v \in \mathbb{R}^p : f(z) - f(b) \geq v'(z - b)\ \forall z \in \mathbb{R}^p\}.$$

$$\partial_{|x|}(x_0) = \begin{cases} 1 & \text{for} \quad x_0 > 0 \\ -1 & \text{for} \quad x_0 < 0 \\ < -1, 1 > & \text{for} \quad x_0 = 0 \end{cases}$$

.

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

For a convex function $f : R^p \to R$ we define the subdifferential as

$$\partial_f(b) = \{v \in \mathbb{R}^p : f(z) - f(b) \geq v'(z - b)\ \forall z \in \mathbb{R}^p\}.$$

$$\partial_{|x|}(x_0) = \begin{cases} 1 & \text{for} \quad x_0 > 0 \\ -1 & \text{for} \quad x_0 < 0 \\ < -1, 1 > & \text{for} \quad x_0 = 0 \end{cases}$$

The convex function $f(x)$ attains a minimum at x_0 if and only if $0 \in \partial_f(x_0)$.

LASSO for the orthogonal design $X'X = I$

$$\beta^{LS} = Y'X, \;\; ||Y - Xb||^2 + \lambda||b||_1 = Y'Y + \sum_{i=1}^{p} f_i(b_i)$$

$$f_i(x) = x^2 - 2\beta_i^{LS}x + \lambda|x|$$

,

LASSO for the orthogonal design $X'X = I$

$$\beta^{LS} = Y'X, \;\; ||Y - Xb||^2 + \lambda||b||_1 = Y'Y + \sum_{i=1}^{p} f_i(b_i)$$

$$f_i(x) = x^2 - 2\beta_i^{LS}x + \lambda|x|$$

$$\partial_{f_i}(x_0) = 2x_0 - 2\beta_i^{LS} + \lambda\partial_{|x|}(x_0)$$

,

LASSO for the orthogonal design $X'X = I$

$$\beta^{LS} = Y'X, \;\; ||Y - Xb||^2 + \lambda||b||_1 = Y'Y + \sum_{i=1}^{p} f_i(b_i)$$

$$f_i(x) = x^2 - 2\beta_i^{LS}x + \lambda|x|$$

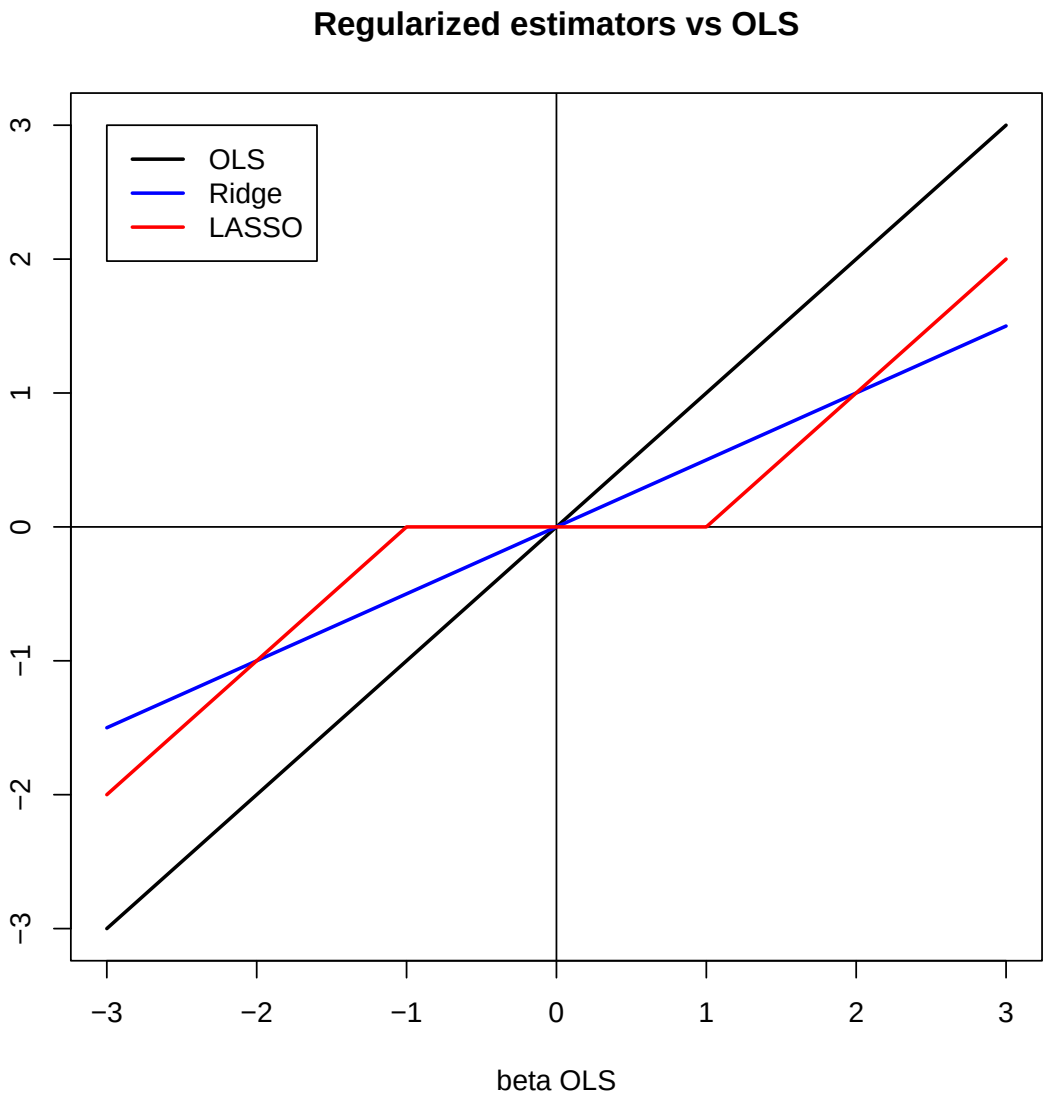
$$\partial_{f_i}(x_0) = 2x_0 - 2\beta_i^{LS} + \lambda\partial_{|x|}(x_0)$$

$$\partial_{f_i}(0) =< -2\beta_i^{LS} - \lambda, -2\beta_i^{LS} + \lambda >$$

,

$$\hat{\beta}_i^{L} = \begin{cases} \beta_i^{LS} - \lambda/2 & \text{when} \quad \beta_i^{LS} > \lambda/2 \\ -\beta_i^{LS} + \lambda/2 & \text{when} \quad \beta_i^{LS} < -\lambda/2 \\ 0 & \text{when} \quad |\beta_i^{LS}| < \lambda/2 \end{cases}$$

Regularized estimators vs OLS



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

Let $\bar{I} = \{1, \ldots, p\} \setminus I$

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

Let $\bar{I} = \{1, \ldots, p\} \setminus I$

Irrepresentability condition:

$$\ker(X_I) = \{0\} \quad \text{and} \quad \|X_{\bar{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

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

Let $\bar{I} = \{1, \ldots, p\} \setminus I$

Irrepresentability condition:

$$\ker(X_I) = \{0\} \quad \text{and} \quad \|X_{\bar{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

In the noisless case (i.e. when $Y = X\beta$) IR is sufficient and necessary for the sign recovery of the sufficiently strong signal.

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

Let $\bar{I} = \{1, \ldots, p\} \setminus I$

Irrepresentability condition:

$$\ker(X_I) = \{0\} \quad \text{and} \quad \|X_{\bar{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

In the noisless case (i.e. when $Y = X\beta$) IR is sufficient and necessary for the sign recovery of the sufficiently strong signal.

IR with a sharp inequality is sufficient and necessary for the sign recovery for the sufficiently large signal to noise ratio $\frac{min_{i \in I}|\beta_i|}{\sigma}$ (see e.g. Wainwright, 2009).

Irrepresentablity vs identifiability

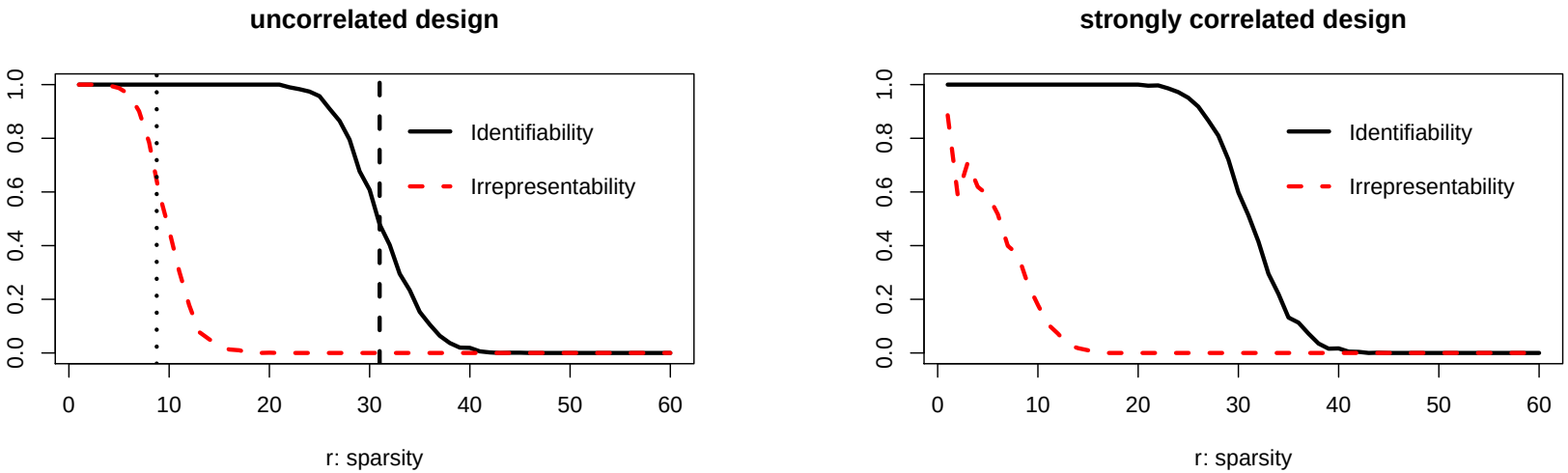


Figure: $n = 100$, $p = 300$, in the right panel $\rho(X_i, X_j) = 0.9$, vertical lines correspond to $n/(2 \log p)$ and the transition curve of Donoho and Tanner (2009).

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

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

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.

SLOPE

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

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

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, $V := \left|\left\{i : \ \beta_i = 0, \quad \widetilde{\beta}_i \neq 0\right\}\right|$
  - $$FDR := \mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right]$$

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

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

Theorem

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

$$\sup_{||\beta||_0 \leq k} P\left(\frac{n||\hat{\beta}_{SL} - \beta||^2}{2\sigma^2 k \log(p/k)} > 1 + 3\epsilon \right) \to 0$$

$$\inf_{\hat{\beta}} sup_{||\beta||_0 \leq k} P\left(\frac{n||\hat{\beta} - \beta||^2}{2\sigma^2 k \log(p/k)} > 1 - \epsilon \right) \to 1$$

(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 $[\frac{k \log(p/k)}{n}]$ under weighted restricted eigenvalue condition (large collection of random matrices)

Asymptotic optimality (2)

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

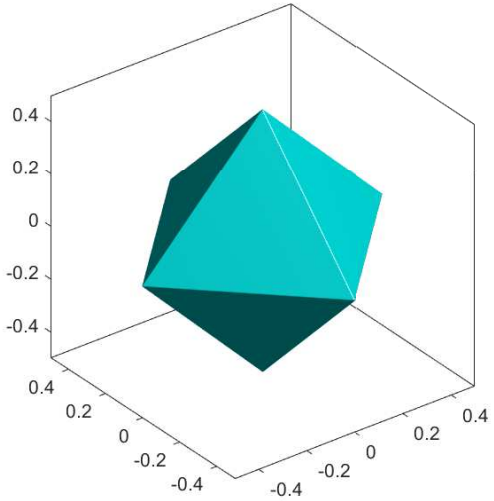
$\lambda_i = \rho\sqrt{2 \log(p/i)}$, ρ is larger than one

Bellec, Lecué, Tsybakov (2016,2017)

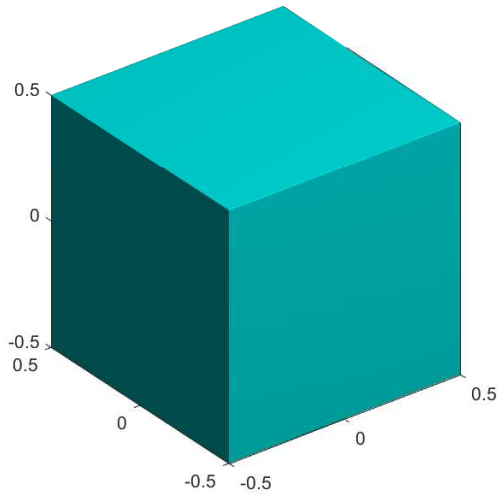
Extension to GLM by Abramovich and Grinshtein (2017)

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

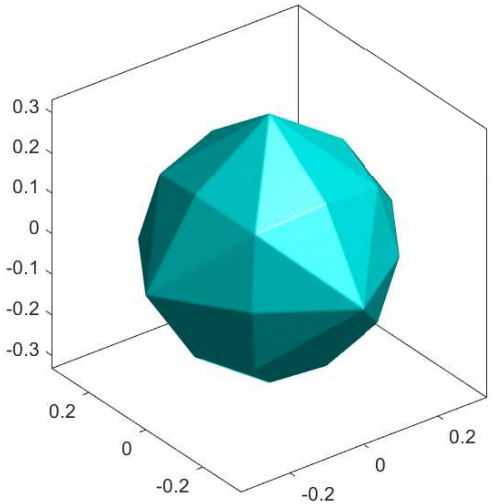
Unit balls for different SLOPE sequences by D.Brzyski



[(2,2,2)]



[(2,0,0)]



[(3,2,1)]

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

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:

$$\mathrm{patt}(b_1) = \mathrm{patt}(b_2) \quad \Leftrightarrow \quad \partial J_\lambda(b_1) = \partial J_\lambda(b_2)$$

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

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

$$\beta_1 X_1 + \beta_2 X_2 + \beta_8 X_8 = \beta_1 (X_1 - X_2 + X_8)$$

$$\tilde{X}_M = X U_M \text{ - pattern-reduced } X$$

$$\tilde{\lambda}_M \in R^k : \quad \tilde{\lambda}_M(j) = \sum_{i=k_{j-1}+1}^{k_j} \lambda_i$$

IR for SLOPE

SLOPE dual norm: $J^*_\lambda(x) = sup\{x'z | J_\lambda(z) \leq 1\}$

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

IR for SLOPE

SLOPE dual norm: $J^*_\lambda(x) = sup\{x'z | J_\lambda(z) \leq 1\}$

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

When $\ker(\tilde{X}_M) = \{0\}$, the SLOPE IR condition takes a form

$$J^*_\lambda \left(X' \tilde{X}_M (\tilde{X}'_M \tilde{X}_M)^{-1} \tilde{\lambda}_M \right) \leq 1.$$

Noisless pattern recovery

Theorem (B.,Dupuis,Graczyk, Kołodziejek, Skalski, Tardivel, Wilczyński (2022))

When $Y = X\beta$ then SLOPE can properly identify a given SLOPE pattern if and only if the irrepresentability condition is satisfied and the signal is strong enough.

Noisless pattern recovery

Theorem (B.,Dupuis,Graczyk, Kołodziejek, Skalski, Tardivel, Wilczyński (2022))

When $Y = X\beta$ then SLOPE can properly identify a given SLOPE pattern if and only if the irrepresentability condition is satisfied and the signal is strong enough.

In the presence of noise we need an additional condition:

$$\left|\left\{i \in \{1,\dots,p\}\colon \sum_{j=1}^{i} |\Pi|_{(j)} = \sum_{j=1}^{i} \lambda_j\right\}\right| = \|M\|_\infty,$$

where $\Pi = X'\tilde{X}_M(\tilde{X}_M'\tilde{X}_M)^{-1}\tilde{\lambda}_M$.

Asymptotic results

p - fixed, $n \to \infty$

Asymptotic results

p - fixed, $n \to \infty$

$$\frac{1}{n} X_n' X_n \xrightarrow{a.s.} C$$

Asymptotic results

p - fixed, $n \to \infty$

$$\frac{1}{n} X_n' X_n \xrightarrow{a.s.} C$$

In IR replace $X' \tilde{X}_M (\tilde{X}_M' \tilde{X}_M)^{-1}$ with $C U_M (U_M' C U_M)^{-1}$

Asymptotic results

p - fixed, $n \to \infty$

$$\frac{1}{n} X_n' X_n \xrightarrow{a.s.} C$$

In IR replace $X' \tilde{X}_M (\tilde{X}_M' \tilde{X}_M)^{-1}$ with $C U_M (U_M' C U_M)^{-1}$

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

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

if and only if $\Lambda = \alpha_n \Lambda_0$ and

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

Identifiability condition for SLOPE

Definition (Identifiability)

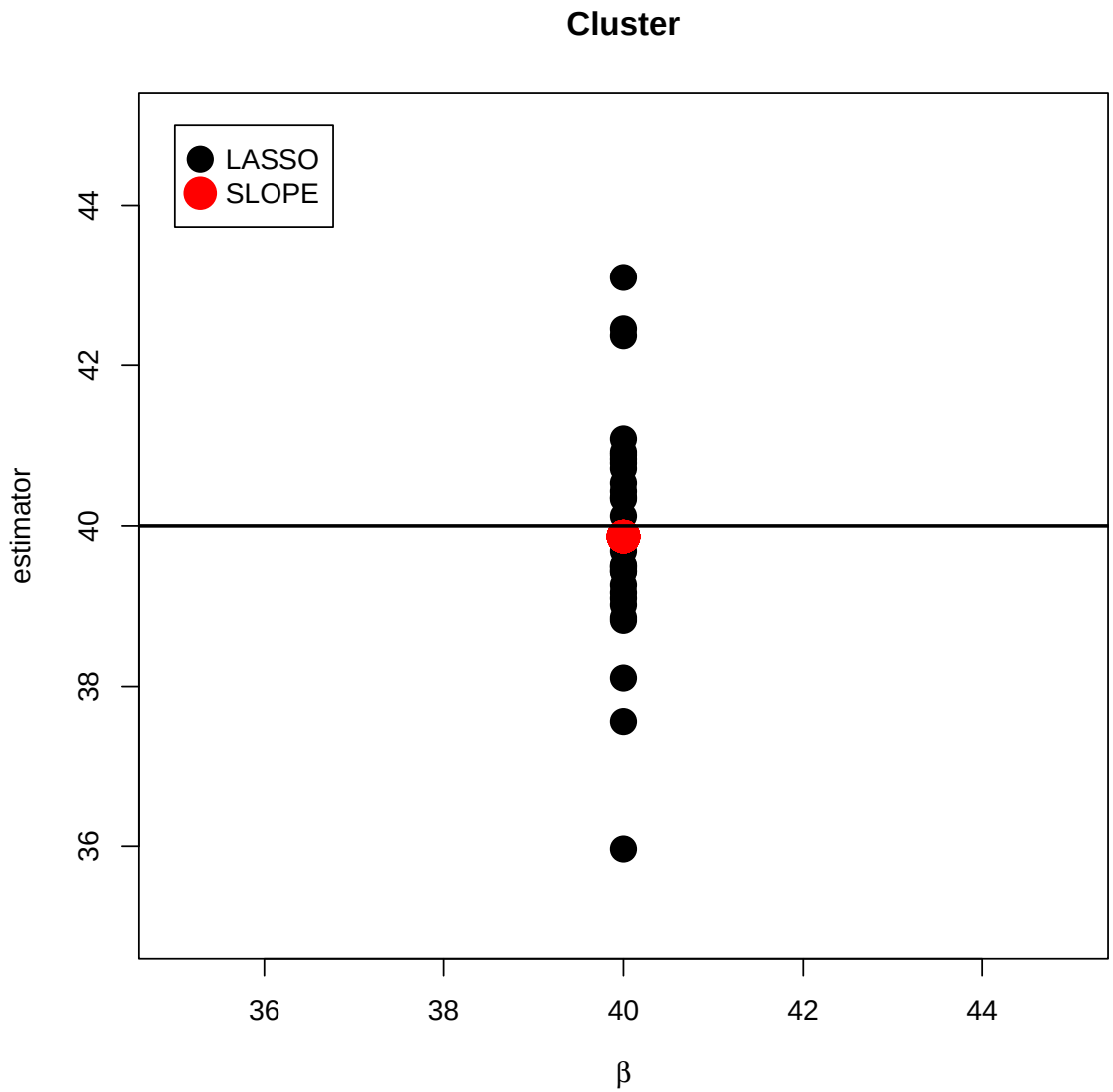
Let X be a $n \times p$ matrix. The vector $\beta \in R^p$ is said to be identifiable with respect to the SLOPE J_λ norm if the following implication holds

$$X\gamma = X\beta \text{ and } \gamma \neq \beta \Rightarrow J_\lambda(\gamma) > J_\lambda(\beta). \tag{2}$$

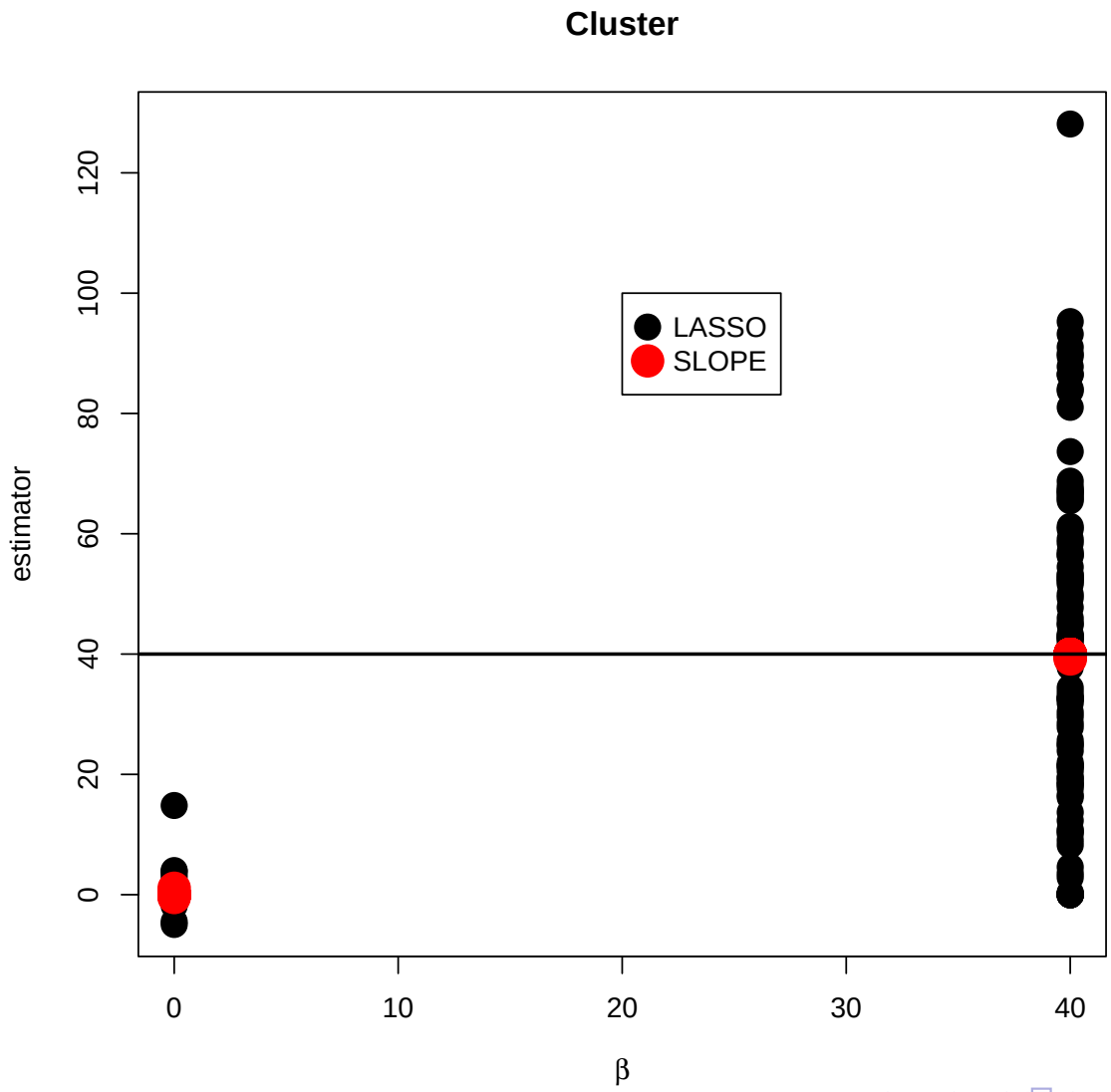
Theorem (Tardivel, Skalski, Graczyk, Schneider (2022))

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



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



Clustering in financial applications

- Kremer, Lee, B., Paterlini, *Journal of Banking and Finance* 110, 105687, 2020 - application for portfolio selection.
- Kremer, Brzyski, B., Paterlini, *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, $Cov(R) = \Sigma$

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

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

$$P = \sum w_i R_i, \sum w_i = 1$$

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

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

$$P = \sum w_i R_i, \sum w_i = 1$$

Portfolio Risk: $Var(P) = w'\Sigma w$

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

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

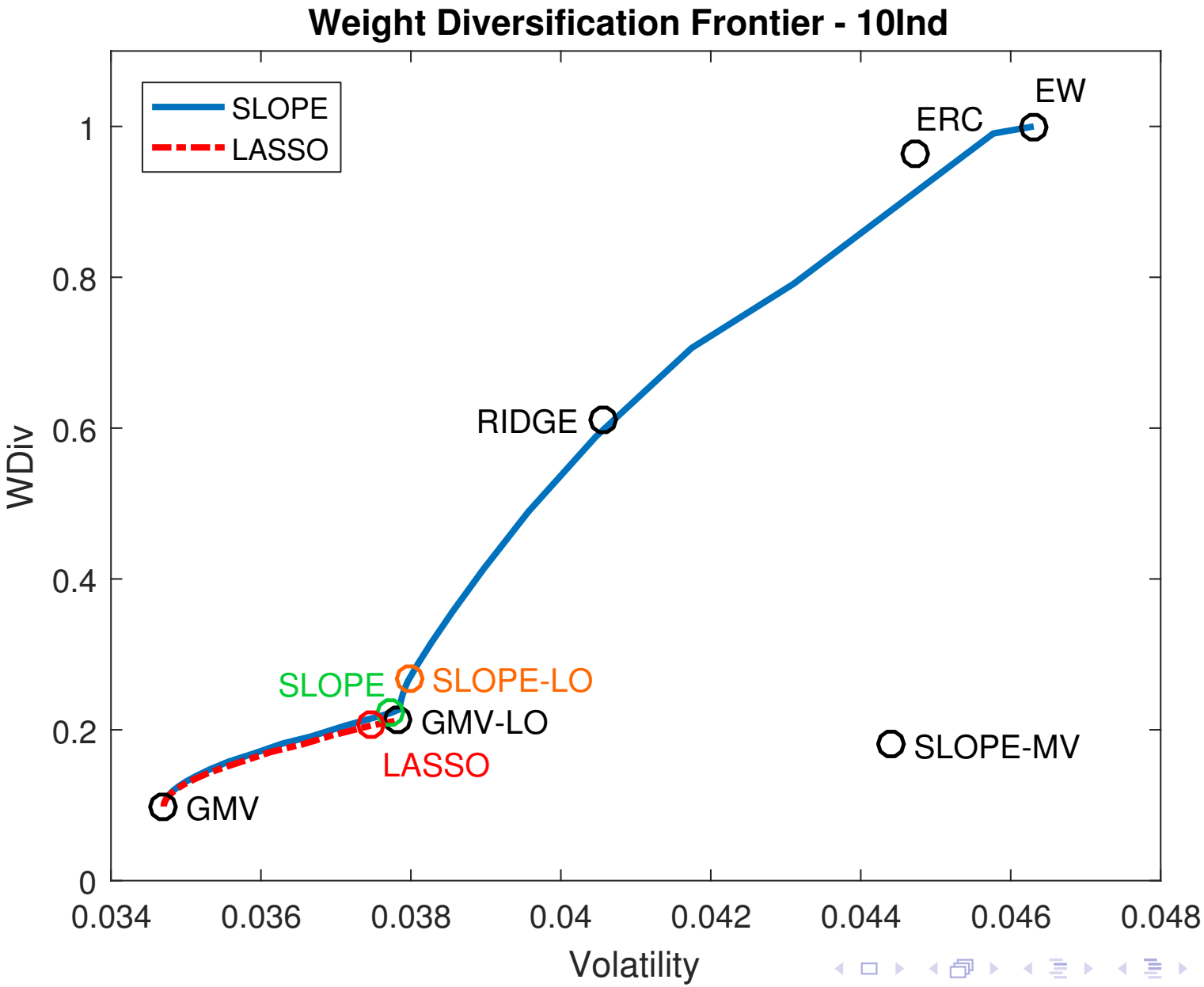
$$P = \sum w_i R_i, \sum w_i = 1$$

Portfolio Risk: $Var(P) = w'\Sigma w$

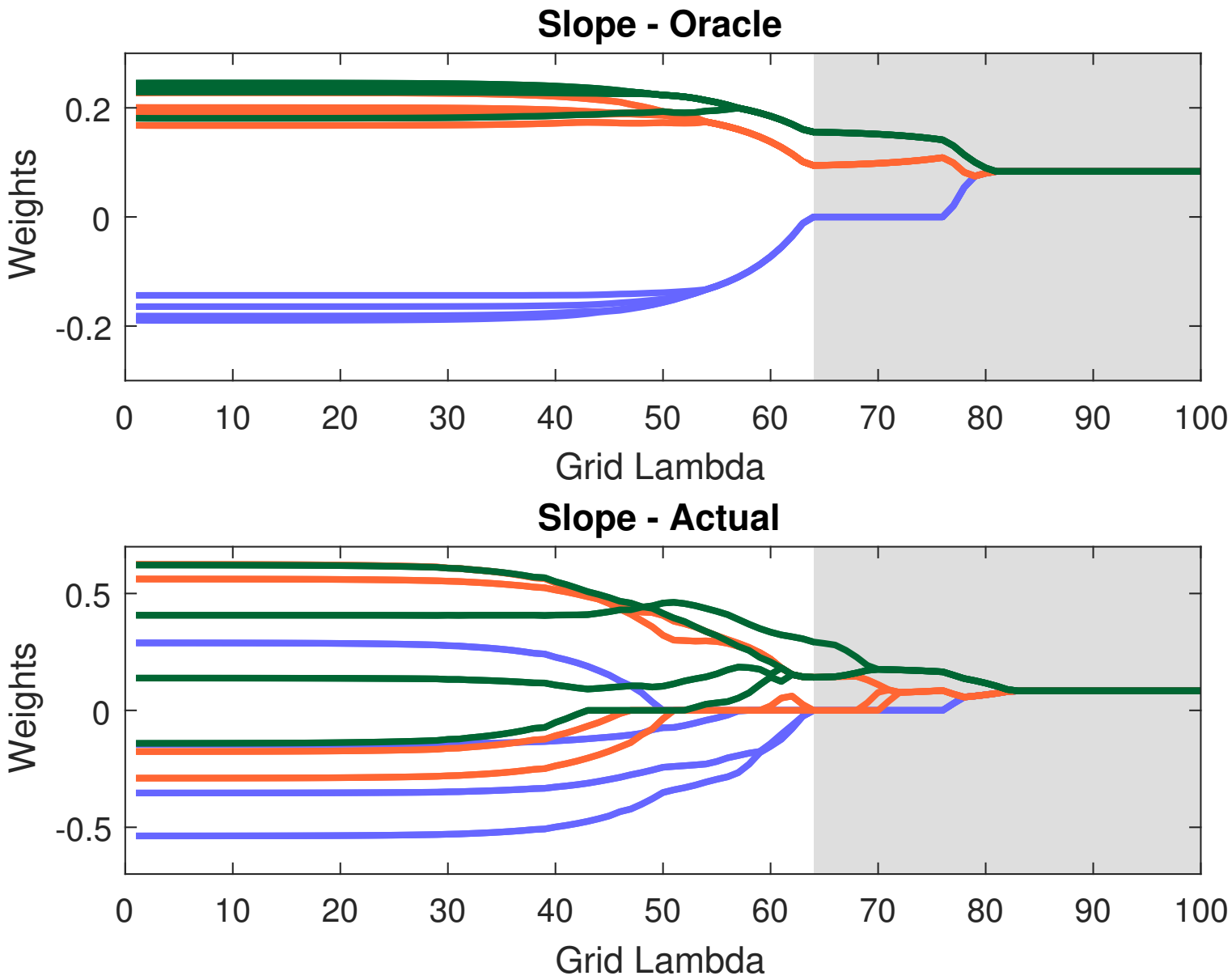
$$\min_{w\in\mathbb{R}^k} w'\Sigma w + J_\lambda(w) \tag{3}$$

$$\text{s.t.} \sum_{i=1}^{k} w_i = 1 \tag{4}$$

Evolution of Portfolio



SLOPE clustering



Applications in Genetics

- Goal - identification of genes influencing some important characteristics (cholesterol level, daily number of drinks)
- Explanatory variables - appropriately coded genotypes of genetic markers
- n in hundreds/thousands, p in hundred thousands
- D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS (Genome Wide Association Studies) false discoveries"', *Genetics*, **205**, 61–75, 2017
- D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", *Journal of the American Statistical Association*, 114(525), 419–433, 2019.

Summaries

- F. Frommlet, M. Bogdan and D. Ramsey, "Phenotypes and genotypes: The Search for Influential Genes", Springer-Verlag, London, 2016
- M. Bogdan and F. Frommlet, "Identifying important predictors in large data bases–multiple testing and model selection", in "Handbook of Multiple Comparisons", Chapman Hall/CR, 2022.

SLOPE packages in R

- *SLOPE* by J.Larsson - also for Generalized Linear Models (logistic, Poisson regression)
- *grpSLOPE* by A. Gossmann
- *geneSLOPE* by P. Sobczyk
- *SLOBE* -adaptive SLOPE by S. Majewski and B. Miasojedow
  - W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, V. Rockova, TraumaBase Group, "Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values", *Journal of Computational and Graphical Statistics*, 31 (1), 113-137, 2022

Motivating example

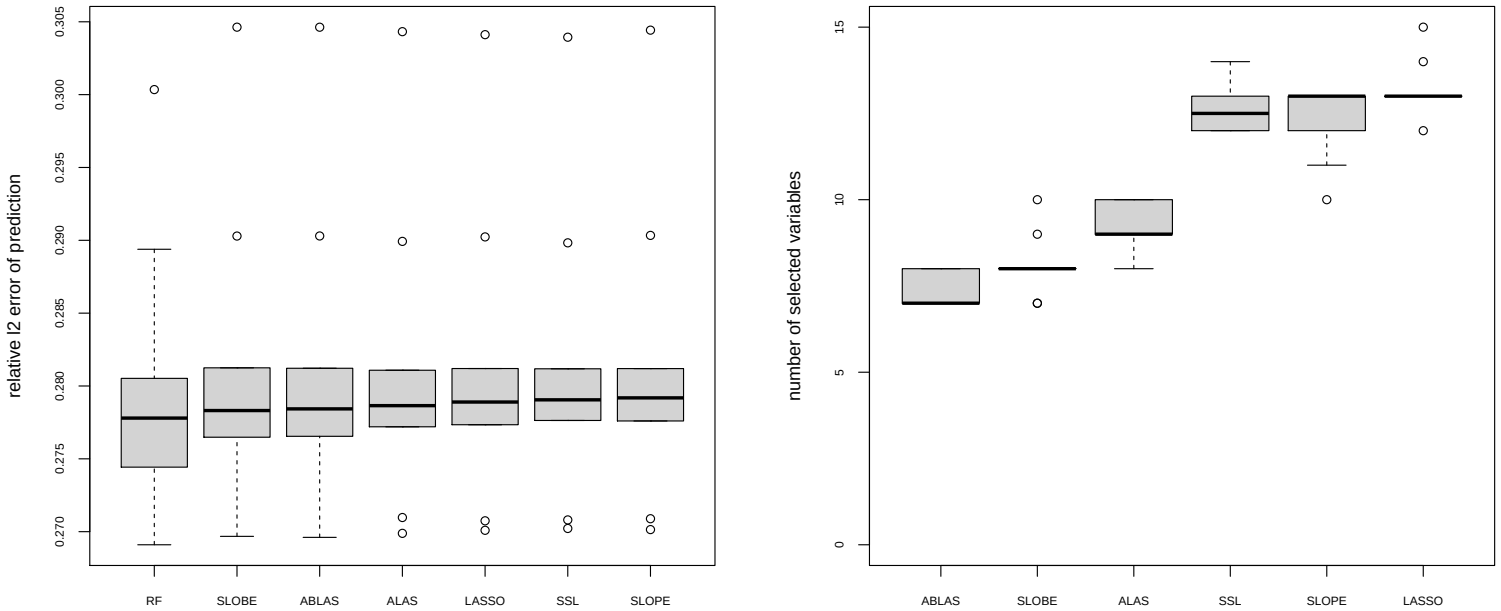


Figure: Empirical distribution of prediction errors and of the number of variables selected by different methods.

$$100\mathit{Platelets} = -8.71\mathrm{Age} - 10.52\mathrm{SI} + 9.16\mathrm{Delta.hemo} - 14.7\mathrm{Lactate} + 14.2\mathrm{HR} - 6.54\mathrm{VE} - 11\mathrm{RBC}.$$

LASSO and SLOPE work

- R. Riccobello, G. Bonaccolto, P. Kremer, S. Paterlini, M. Bogdan, "Sparse Graphical Modelling for Minimum Variance Portfolios", SSRN 4099586, 2022.
- R. Riccobello, M. Bogdan, G. Bonaccolto, P.J. Kremer, S. Paterlini, P. Sobczyk, "Sparse Graphical Modelling via the Sorted L_1 Norm", arXiv preprint arXiv:2204.10403, 2022.
- M. Bogdan, X. Dupuis, P. Graczyk, B. Kołodziejek, T. Skalski, P. Tardivel, M. Wilczyński, "Pattern recovery by SLOPE", arXiv:2203.12086, 2022.
- P.J. Kremer, D. Brzyski, M. Bogdan, S. Paterlini, "Sparse index clones via the sorted L_1-Norm", *Quantitative Finance* 22 (2), 349-366, 2022.
- W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, V. Rockova, TraumaBase Group, "Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values", *Journal of Computational and Graphical Statistics*, 31 (1), 113-137, 2022.
- P.Tardivel, M. Bogdan, "On the sign recovery by least absolute shrinkage and selection operator, thresholded least absolute shrinkage and selection operator, and thresholded basis pursuit denoising", *Scandinavian Journal of Statistics*, 2022.
- F. Frommlet, M. Bogdan, "Identifying important predictors in large data basesâ"Multiple testing and model selection" in *Handbook of Multiple Comparisons*, pp. 139-182, 2022.
- J. Larsson, M. Bogdan, J. Wallin, "The strong screening for SLOPE", NeurIPS 2020.
- P.J. Kremer, S. Lee, M. Bogdan, S. Paterlini, "Sparse portfolio selection via the sorted L1-Norm", *Journal of Banking and Finance* 110, 105687, 2020.
- W. Rejchel, M. Bogdan, "Rank-based Lasso-efficient methods for high-dimensional robust model selection", *Journal of Machine Learning Research* 21 (244), 1-47.

LASSO and SLOPE work

- M. Kos, M. Bogdan, "On the asymptotic properties of SLOPE", Sankhya A 82 (2), 499-532, 2020.
- A. Weinstein, W.J. Su, M. Bogdan, R.F. Barber, E.J. Candès, "A power analysis for knockoffs with the lasso coefficient-difference statistic", arXiv 2020.
- S.Lee, P.Sobczyk, M.Bogdan, "Structure Learning of Gaussian Markov Random Fields with False Discovery Rate Control", *Symmetry* 11 (10), 1311, 2019.
- D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", *Journal of the American Statistical Association*, 114(525), 419–433, 2019.
- W.Su, M. Bogdan, E.J. Candès, "False Discoveries Occur Early on the Lasso Path", *Annals of Statistics*, 45 (5), 2133 – 2150, 2017.
- D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries"', *Genetics*, 205, 61–75, 2017.
- S. Lee, D. Brzyski, M. Bogdan, "Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l_1-Norm", *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:W and CP* vol.51, 780–789, 2016.
- A. Virouleau, A. Guilloux, S. Gaiffas, M. Bogdan, "High-dimensional robust regression and outliers detection with slope", arXiv:1712.02640, 2017.
- W.Su, M. Bogdan, E.J.Candes, "False discoveries occur early on the lasso path", *Annals of Statistics*, 2133-2150, 2017.
- D. Brzyski, C.B. Peterson, P. Sobczyk, E.J. Candes, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries" *Genetics* 205 (1), 61-75, 2017.
- S. Lee, D. Brzyski, M. Bogdan, "Fast saddle-point algorithm for generalized dantzig selector and fdr control with ordered L1-norm", Artificial Intelligence and Statistics, 780-789, 2016.
- M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E.J. Candes, "SLOPE - adaptive variable selection via convex optimization", *Annals of applied statistics* 9 (3), 1103, 2015.
- M. Bogdan, E. van den Berg, W. Su, E. J. Candes, "Statistical estimation and testing via the sorted L1 norm", arXiv:1310.1969, 2013.