3.27pt

Regularization methods in multiple regression

Malgorzata Bogdan

University of Wroclaw, Lund University

KAU, 06/03/201

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

$X_{n \times p}$ - matrix of regressors

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression:

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

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression:

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

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

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression:

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

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

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

Ridge regression (1)

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

Ridge regression (1)

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

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \; M = X(X'X + \gamma I)^{-1} X'$$

Ridge regression (1)

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

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \; M = X(X'X + \gamma I)^{-1} X'$$

$$Tr[M] = Tr\left[(X'X + \gamma I)^{-1} X'X\right]$$

Ridge regression (1)

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

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \; M = X(X'X + \gamma I)^{-1} X'$$

$$Tr[M] = Tr\left[(X'X + \gamma I)^{-1} X'X\right]$$

$$Tr[M] = \sum_{i=1}^{n} \lambda_i(M), \quad \text{where} \quad \lambda_1(M), \ldots, \lambda_n(M) \; \text{are eigenvalues of} \; M$$

Eigenvalues of M

$$X'Xu = \lambda u$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \quad (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \quad (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \quad Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \quad (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \quad Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \quad (X'X + \gamma I)^{-1} u = \frac{1}{\lambda + \gamma} u$$

$$(X'X + \gamma I)^{-1} X'Xu = \frac{\lambda}{\lambda + \gamma} u, \quad Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

$$\hat{P}E = RSS + 2\sigma^2 \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

$$= \frac{\gamma^2}{(1+\gamma)^2}\beta_i^2 + \frac{\sigma^2}{(1+\gamma)^2}$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

$$= \frac{\gamma^2}{(1+\gamma)^2}\beta_i^2 + \frac{\sigma^2}{(1+\gamma)^2}$$

$$E||\hat{\beta} - \beta||^2 = \frac{\gamma^2}{(1+\gamma)^2}||\beta||^2 + \frac{p\sigma^2}{(1+\gamma)^2}$$

Ridge regression - orthogonal design (2)

When rigde is better than LS ?

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge regression - orthogonal design (2)

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

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

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

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

Otherwise, when

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

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

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

Otherwise, when

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

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

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit

$$Y = X\beta$$

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

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover 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.

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

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover 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.

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

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

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover 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.

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

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

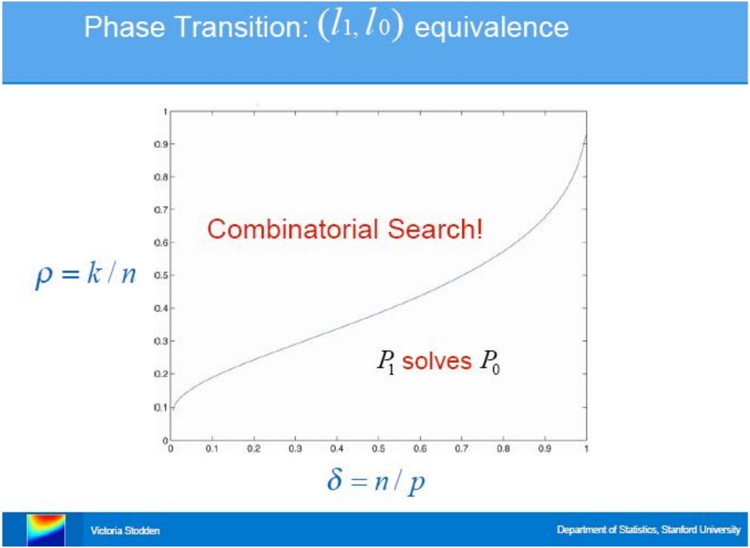
Basis Pursuit can recover β if k is small enough.

Transition curve (Donoho and Tanner, 2005)

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

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

Transition curve (2)



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 $||Y - Xb||_2^2 \leq \epsilon$

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

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 $||Y - Xb||_2^2 \leq \epsilon$

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

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

Selection of the tuning parameter for LASSO

- General rule: the reduction of λ_L results in identification of more elements from the true support (true discoveries) but at the same time it produces more falsely identified variables (false discoveries)
- The choice of λ_L is challenging- e.g. crossvalidation typically leads to many false discoveries
- When $X^TX = I$ Lasso selects X_j iff $|\hat{\beta}_j^{LS}| > \lambda$
- Selection $\lambda = \sigma\Phi^{-1}(1 - \alpha/(2p)) \approx \sigma\sqrt{2\log p}$ corresponds to Bonferroni correction and controls FWER.

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 $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

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 $I := \{i \in \{1, \dots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentable condition:

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

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 $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentable condition:

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

When

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

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

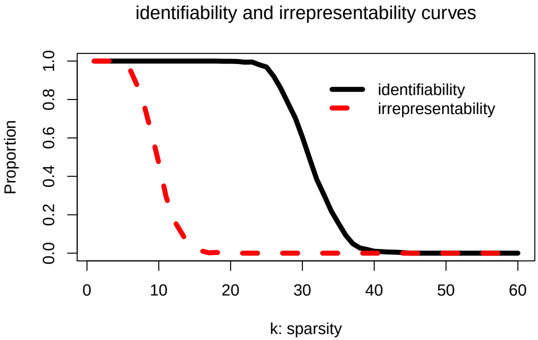
Separation of true and false predictors

Theorem (Tardivel, Bogdan, 2019)

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.

Irrepresentability and identifiability curves

n=100, p=300, elements of X were generated as iid N(0,1)



Modifications of LASSO

Corollary

Appropriately thresholded LASSO can properly identify the sign of sufficiently large β *if and only if* β *is identifiable with respect to* l_1 *norm.*

Conjecture

Adaptive (reweighted) LASSO can properly identify the sign of sufficiently large β *if and only if* β *is identifiable with respect to* l_1 *norm.*

Problem with shrinkage

Intuitive explanation:

$$\hat{\beta} = \eta_\lambda(\beta_i + X_i' z + v_i)$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) \rangle$$

$$\eta_\lambda(t) = sign(t)(|t| - \lambda)_+, \quad \text{applied componentwise}$$

Problem with shrinkage

Intuitive explanation:

$$\hat{\beta} = \eta_\lambda(\beta_i + X_i' z + v_i)$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) \rangle$$

$$\eta_\lambda(t) = sign(t)(|t| - \lambda)_+, \quad \text{applied componentwise}$$

If $X^T X = I$ then $X_i' z = Z_i \sim N(0,1)$, $v_i = 0$ and H_{0i} is rejected if $\beta_i + Z_i > \lambda$

Problem with shrinkage

Intuitive explanation:

$$\hat{\beta} = \eta_\lambda(\beta_i + X_i'z + v_i)$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) \rangle$$

$$\eta_\lambda(t) = sign(t)(|t| - \lambda)_+, \quad \text{applied componentwise}$$

If $X^T X = I$ then $X_i'z = Z_i \sim N(0,1)$, $v_i = 0$ and H_{0i} is rejected if $\beta_i + Z_i > \lambda$

When the design is not orthogonal: $v_i \neq 0$ - additional noise, dependent on λ (level of shrinkage), the level of sparsity and magnitude of true signals

Adaptive LASSO [Zou, *JASA* 2006], [Candès, Wakin and Boyd, *J. Fourier Anal. Appl.* 2008]

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

where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.

Adaptive LASSO [Zou, *JASA* 2006], [Candès, Wakin and Boyd, *J. Fourier Anal. Appl.* 2008]

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

where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.

Reduces bias and improves model selection properties

Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))

Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))

Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}

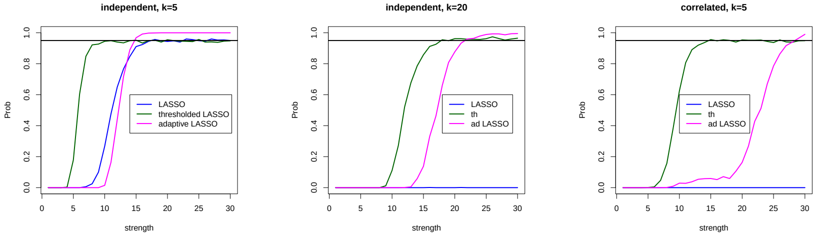
Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}
4. For adaptive LASSO - weights based on LASSO estimator with λ as in 2 and 3, selection based on LASSO with λ as in 1

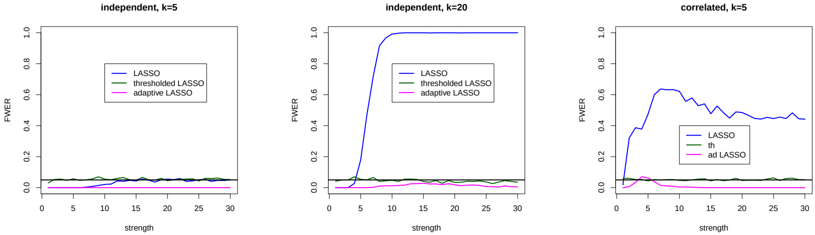
Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}
4. For adaptive LASSO - weights based on LASSO estimator with λ as in 2 and 3, selection based on LASSO with λ as in 1
5. Threshold selected by using knockoff control variables (Foygel-Barber and Candès, 2015; Candès, Fan, Janson, Lv, 2016)

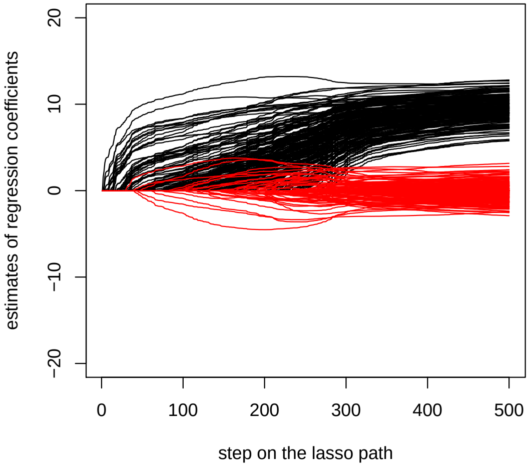
Probability of the sign recovery



Family Wise Error Rate



Thresholded LASSO (1)



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 $X_{aug} = [X, \tilde{X}]$

Function $w : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is *faithful* if it obeys
(I) w is antisymmetric, $w(v, u) = -w(u, v)$
(II) for any fixed c, $w(x, c)$ tends to infinity as $|x| \to \infty$.

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 $X_{aug} = [X, \tilde{X}]$

Function $w : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is *faithful* if it obeys
(I) w is antisymmetric, $w(v, u) = -w(u, v)$
(II) for any fixed c, $w(x, c)$ tends to infinity as $|x| \to \infty$.

$$W_j = w(\widehat{\beta}_j, \widehat{\beta}_{p+j})$$

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

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

and select

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

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

Knockoff filter

Define a random threshold as

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

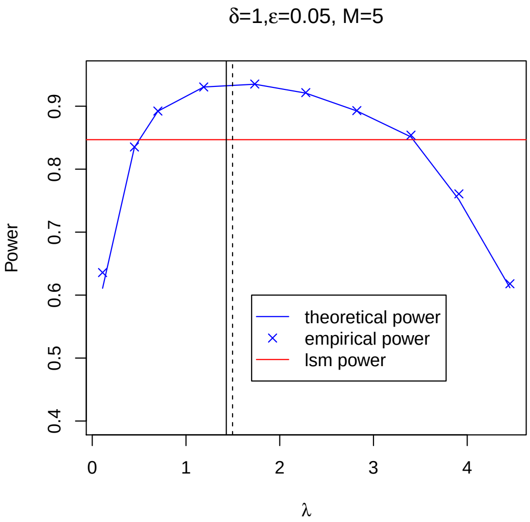
and select

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

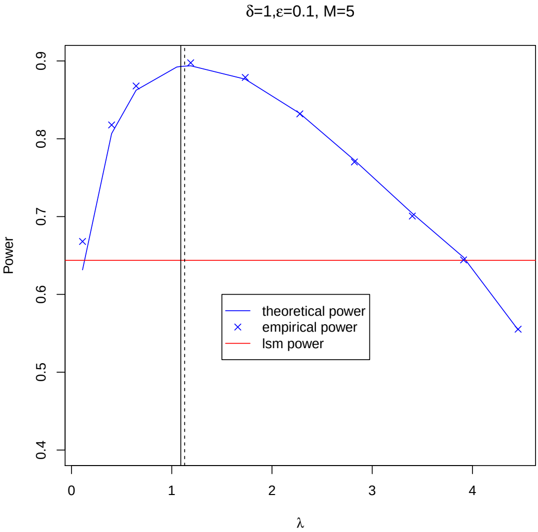
Foygel-Barber and Candès (2015), Candès, Fan, Janson and Lv (2017) - The above knockoff procedure $KN(\lambda, q)$ controls FDR at the level q.

Example: Lasso coefficient difference statistics $LCD(\lambda, q)$

$$W_j(\lambda) = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$$


δ=1,ε=0.05, M=5
Power
0.4
0.5
0.6
0.7
0.8
0.9
0
1
2
3
4
λ
theoretical power
empirical power
lsm power

Gain in power over LSM



Theoretical results using the mean field asymptotics

Su, B., Candès, Ann. Stat. 2017 - FDR-Power Tradeoff Diagram for LASSO

Theoretical results using the mean field asymptotics

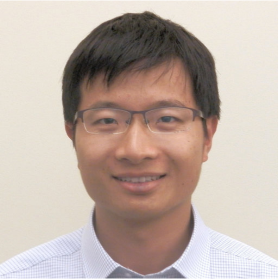
Su, B., Candès, Ann. Stat. 2017 - FDR-Power Tradeoff Diagram for LASSO

Weinstein, Su, Bogdan, Barber, Candés, 2020 - Breaking the tradoff diagram with thresholded LASSO

Sorted L-One Penalized Estimation

M.B., E.van den Berg, C.Sabatti, W.Su, E.J.Candès, AOAS 2015






Sorted L-One Penalized Estimation

$$\hat{\beta} = argmin_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|^2_{\ell_2} + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

where $\left|b\right|_{(1)} \geq \ldots \geq \left|b\right|_{(p)}$ are ordered magnitudes of coefficients of b and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ is the sequence of tuning parameters.

Sorted L-One Penalized Estimation

$$\hat{\beta} = argmin_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

where $\left|b\right|_{(1)} \geq \ldots \geq \left|b\right|_{(p)}$ are ordered magnitudes of coefficients of b and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ is the sequence of tuning parameters.

The above optimization problem is convex and can be efficiently solved even for large design matrices.

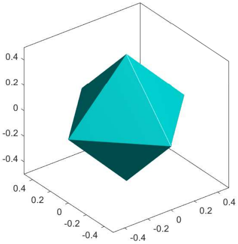
$$\hat{\beta} = argmin_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

where $\left|b\right|_{(1)} \geq \ldots \geq \left|b\right|_{(p)}$ are ordered magnitudes of coefficients of b and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ is the sequence of tuning parameters.

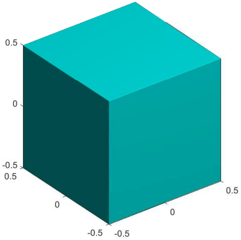
The above optimization problem is convex and can be efficiently solved even for large design matrices.

Sorted L-One Norm: $J_\lambda(b) = \sum_{i=1}^{p} \lambda_i \left|b\right|_{(i)}$ reduces to $||b||_1$ if $\lambda_1 = \ldots = \lambda_p$ and to $||b||_\infty$ if $\lambda_1 > \lambda_2 = \ldots = \lambda_p = 0$.

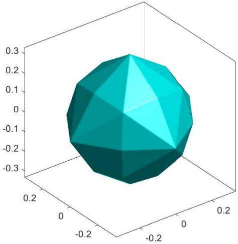
Unit balls for different SLOPE sequences by D.Brzyski



[(2,2,2)]



[(2,0,0)]



[(3,2,1)]

FDR control with SLOPE

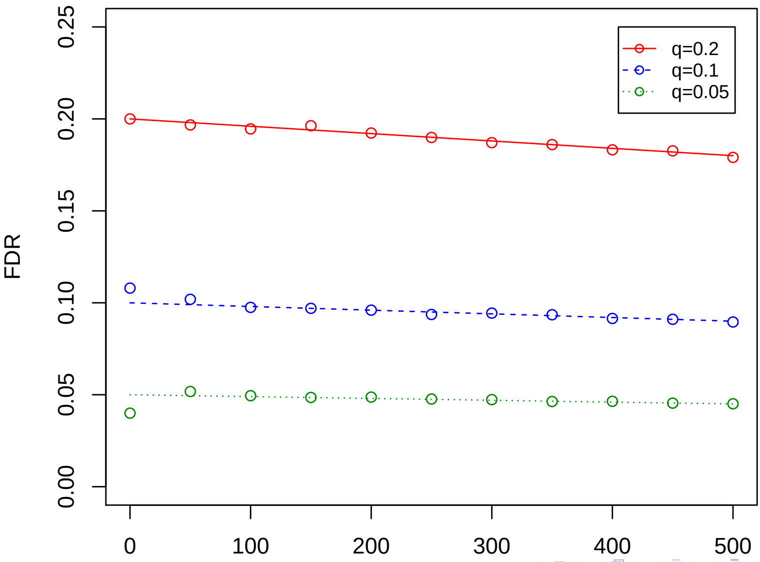
Theorem (B,van den Berg, Sabatti, Su and Candès (2015))

When $X^T X = I$ *SLOPE with*

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

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

Orthogonal design, $n = p = 5000$



Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \rightarrow 0$ and $\frac{k \log p}{n} \rightarrow 0$.

X is standardized so that each column has a unit L_2 norm.

Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (2018, AOS):

SLOPE with the BH related sequence of tuning parameters attains minimax rate for the estimation error $||\hat{\beta} - \beta||^2$.

Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (2018, AOS):

SLOPE with the BH related sequence of tuning parameters attains minimax rate for the estimation error $||\hat{\beta} - \beta||^2$.

SLOPE rate of the estimation error - $k \log(p/k)$

LASSO rate of the estimation error - $k \log p$

Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (2018, AOS):

SLOPE with the BH related sequence of tuning parameters attains minimax rate for the estimation error $||\hat{\beta} - \beta||^2$.

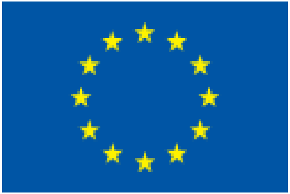
SLOPE rate of the estimation error - $k \log(p/k)$

LASSO rate of the estimation error - $k \log p$

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

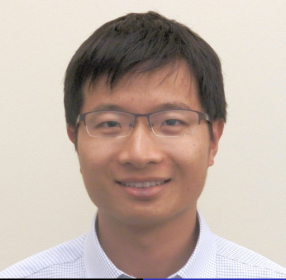
IDEAL

Integrated DEsign and AnaLysis
of small population group trials



Group SLOPE, (D.Brzyski, A.Gossmann, W.Su and MB, JASA, 2019)





Selection of the group of predictors

Identification of groups of predictors:

$$[[\beta]]_I := \left(\|X_{I_1}\beta_{I_1}\|_2, \ldots, \|X_{I_m}\beta_{I_m}\|_2 \right)^{\mathsf{T}} \quad .$$

Selection of the group of predictors

Identification of groups of predictors:

$$[[\beta]]_I := \left(\|X_{I_1}\beta_{I_1}\|_2, \ldots, \|X_{I_m}\beta_{I_m}\|_2\right)^{\mathsf{T}} \quad .$$

$$\beta^{gS} := argmin_b \ \left\{\frac{1}{2}\|y - Xb\|_2^2 + \sigma J_\lambda\left(W[[b]]_I\right)\right\},$$

where W is a diagonal matrix with $W_{i,i} := w_i$, for $i = 1, \ldots, m$.

Selection of the group of predictors

Identification of groups of predictors:

$$[[\beta]]_I := \left(\|X_{I_1}\beta_{I_1}\|_2, \ldots, \|X_{I_m}\beta_{I_m}\|_2\right)^{\mathsf{T}} \quad .$$

$$\beta^{gS} := argmin_b \quad \left\{\frac{1}{2}\|y - Xb\|_2^2 + \sigma J_\lambda\left(W[[b]]_I\right)\right\},$$

where W is a diagonal matrix with $W_{i,i} := w_i$, for $i = 1, \ldots, m$.

Selection of

$$\lambda_i^{\max} := \max_{j=1,\ldots,m}\left\{\frac{1}{w_j}F_{\chi_{l_j}}^{-1}\left(1 - \frac{q \cdot i}{m}\right)\right\}$$

allows to control group FDR and obtain a minimax rate of estimation of $[[\beta]]_I$ if variables in different groups are orthogonal to each other.

Identification of groups of predictors:

$$[[\beta]]_I := \left(\|X_{I_1}\beta_{I_1}\|_2, \ldots, \|X_{I_m}\beta_{I_m}\|_2\right)^{\mathsf{T}} \quad .$$

$$\beta^{gS} := argmin_b \;\; \left\{\frac{1}{2}\|y - Xb\|_2^2 + \sigma J_\lambda\left(W[[b]]_I\right)\right\},$$

where W is a diagonal matrix with $W_{i,i} := w_i$, for $i = 1, \ldots, m$.

Selection of

$$\lambda_i^{\max} := \max_{j=1,\ldots,m}\left\{\frac{1}{w_j} F_{\chi_{l_j}}^{-1}\left(1 - \frac{q \cdot i}{m}\right)\right\}$$

allows to control group FDR and obtain a minimax rate of estimation of $[[\beta]]_I$ if variables in different groups are orthogonal to each other.

Heuristic adjustment for the situation when variables in different groups are independent.

$n = 5402$, $p = 26233$ - roughly independent SNPs

$n = 5402$, $p = 26233$ - roughly independent SNPs

Scenario 1: $Y = X\beta + z$ - additive model

$$X_{ij} = \begin{cases} -1 & \text{for} \quad aa \\ 0 & \text{for} \quad aA \\ 1 & \text{for} \quad AA \end{cases}, \tag{2}$$

Applications for GWAS

$n = 5402$, $p = 26233$ - roughly independent SNPs

Scenario 1: $Y = X\beta + z$ - additive model

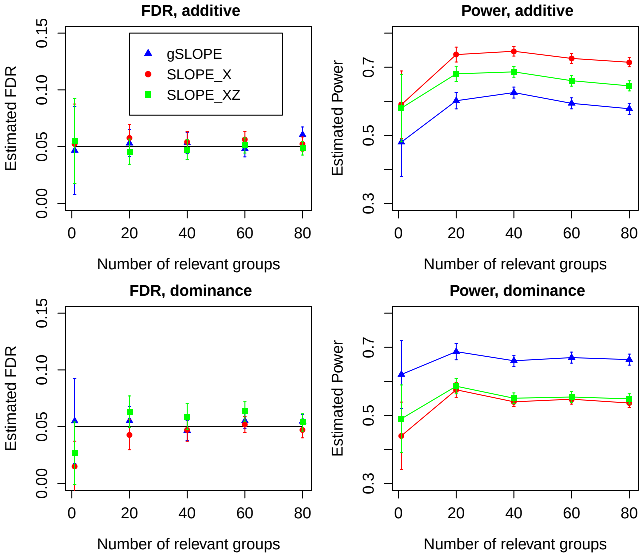
$$X_{ij} = \begin{cases} -1 & \text{for} \quad aa \\ 0 & \text{for} \quad aA \\ 1 & \text{for} \quad AA \end{cases}, \tag{2}$$

Scenario 2: modeling dominance

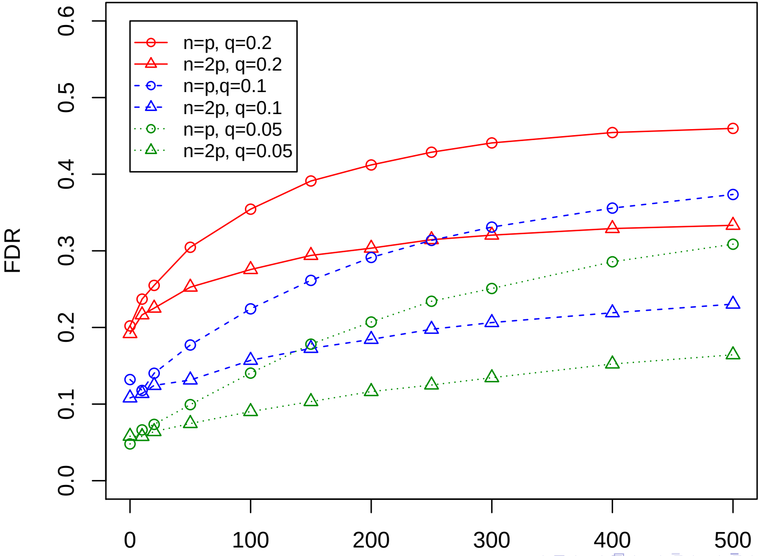
$$Z_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases}, \tag{3}$$

$$y = [X, Z][\beta_X', \beta_Z']' + \epsilon \quad .$$

Simulation results



Gaussian design (1), $n = p = 5000$



Spike and Slab LASSO (Rockova, George, 2018)

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

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

Spike and Slab LASSO (Rockova, George, 2018)

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

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

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"

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

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

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

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

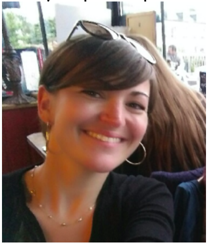
In consecutive iterations γ_i is replaced with

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

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.

Adaptive SLOPE with missing values (1)

W. Jiang, MB, J.Josse, B.Miasojedow, V.Rockova, TraumaBase Group (in progress)
code available at `github.com/sjmMajewski/SLOBE-Rcpp`

traumabase.eu

ABSLOPE (2)

Prior for β is given by

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

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

ABSLOPE (2)

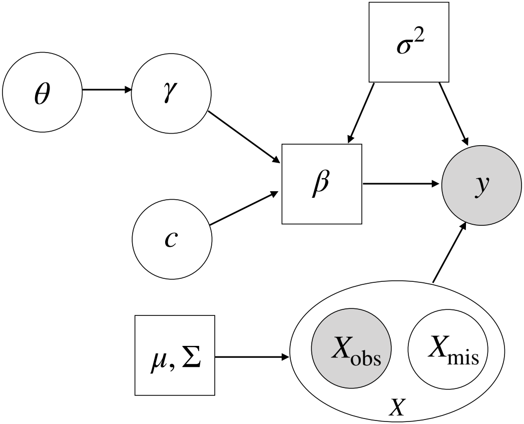
Prior for β is given by

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

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:

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


θ
γ
c
σ^2
β
y
μ, Σ
X_{obs}
X_{mis}
X

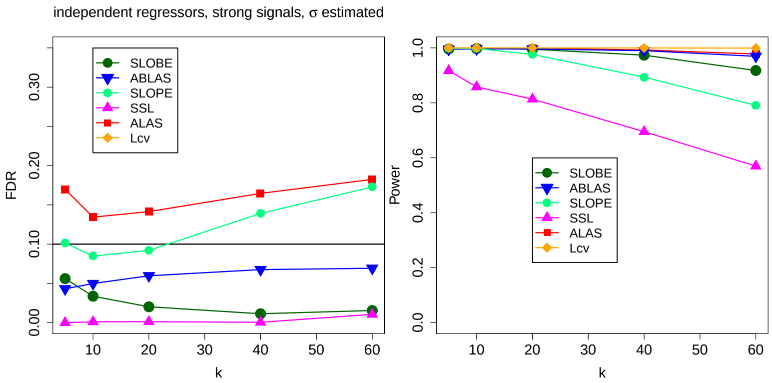
Stochastic approximation EM algorithm

- $\pi(\theta)$ - 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,

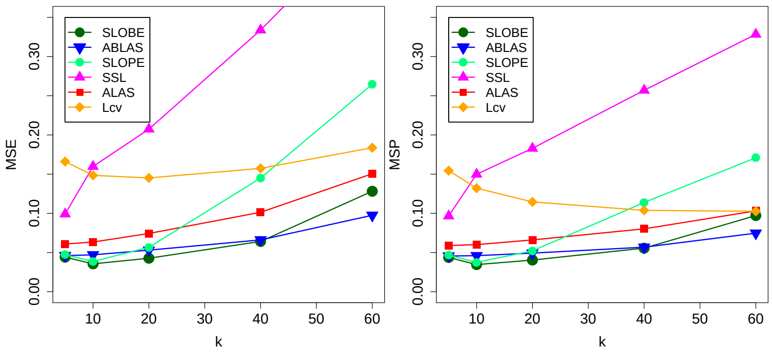
$$\psi^{t+1} = \psi^t + \eta_t \left[\hat{\psi}^t_{MLE} - \psi^t \right],$$

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

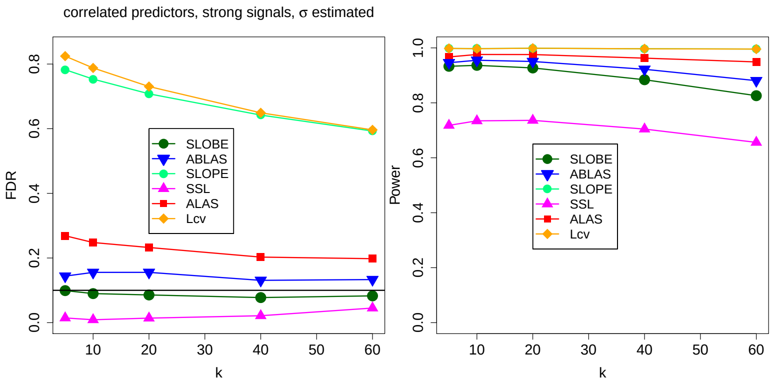
$n = p = 500$, $\rho = 0$, $Na = 10\%$, independent regressors



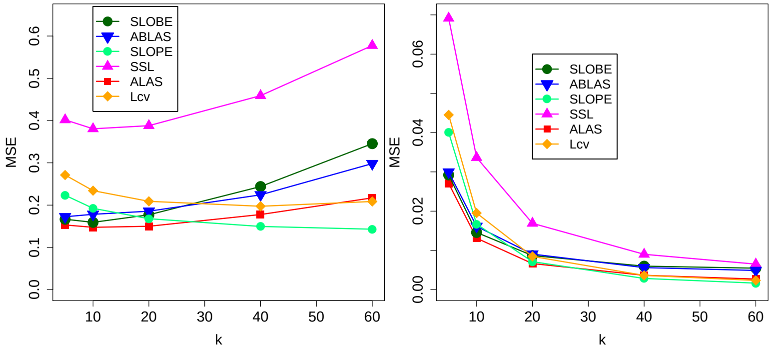
$n = p = 500$, $\rho = 0$, $Na = 10\%$, independent regressors



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



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



LASSO and SLOPE work

- 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.
- M. Kos, M. Bogdan, "On the asymptotic properties of SLOPE", Sankhya A 82 (2), 499-532, 2020.
- W. Jiang, M. Bogdan, J. Josse, B. Miasojedow, V. Rockova, TraumaBase Group, "Adaptive Bayesian SLOPE – High-dimensional Model Selection with Missing Values", arxive 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", arxive 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.

SLOPE packages in R

- *SLOPE* by J.Larsson - also for Generalized Linear Models (logistic, Poisson regression)
- *grpSLOPE* by A. Gossmann
- *geneSLOPE* by P. Sobczyk

- 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", to appear in "Handbook of Multiple Comparisons", Chapman Hall/CR, 2021.

Open SLOPE projects

- PhD position at the Department of Statistics at Lund University (Sweden)

  https://stat-lu.github.io/PhDpos/

  https://lu.varbi.com/en/what:job/jobID:383509/
- Google of Summer Code - creating package for ABSLOPE, mentored by J. Larsson from Lund