3.27pt

SLOPE

Malgorzata Bogdan[1,2]

[1], Department of Mathematics, University of Wroclaw

[2]Department of Statistics, Lund University

May, 2020

Outline

- Research with J.K.Ghosh - mBIC and asymptotic optimality of the Benjamini-Hochberg procedure

Outline

- Research with J.K.Ghosh - mBIC and asymptotic optimality of the Benjamini-Hochberg procedure
- SLOPE (Sorted L-One Penalized Estimation)

Outline

- Research with J.K.Ghosh - mBIC and asymptotic optimality of the Benjamini-Hochberg procedure
- SLOPE (Sorted L-One Penalized Estimation)
- Adaptive Bayesian version of SLOPE

Outline

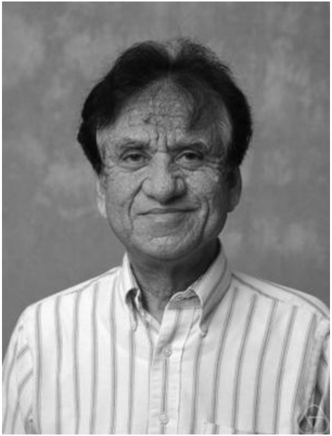
- Research with J.K.Ghosh - mBIC and asymptotic optimality of the Benjamini-Hochberg procedure
- SLOPE (Sorted L-One Penalized Estimation)
- Adaptive Bayesian version of SLOPE
- Screening Rules for SLOPE

Outline

- Research with J.K.Ghosh - mBIC and asymptotic optimality of the Benjamini-Hochberg procedure
- SLOPE (Sorted L-One Penalized Estimation)
- Adaptive Bayesian version of SLOPE
- Screening Rules for SLOPE
- SLOPE for graphical models

Modified version of BIC (1)

M. B, J.K. Ghosh, R.W. Doerge, *Genetics* (2004)



Statistical problem

Selecting important genetic markers based on the multiple regression model:

$X_{n\times p}$ - matrix of genotypes of genetic markers,

$Y_{n\times 1}$ - vector of trait values.

Goal: Identify the best model of the form

$$Y = \beta_0 1 + X_I \beta_I + \varepsilon,$$

I - a subset of $\{1, \ldots, p\}$, $k = |I|$, $\varepsilon \sim N(0, \sigma^2 I)$.

mBIC

mBIC: Select the model minimizing

$$\log RSS + k\log(n) + 2k\log\left(\frac{p}{C}\right),$$

where RSS is the residual sum of squares,

C is the prior expected value of the number of genetic effects.

mBIC

mBIC: Select the model minimizing

$$\log RSS + k\log(n) + 2k\log\left(\frac{p}{C}\right),$$

where RSS is the residual sum of squares,

C is the prior expected value of the number of genetic effects.

mBIC results from supplementing BIC with the Binomial prior $B(p, C/p)$ on the number of genetic effects.

mBIC

mBIC: Select the model minimizing

$$\log RSS + k\log(n) + 2k\log\left(\frac{p}{C}\right),$$

where *RSS* is the residual sum of squares,

C is the prior expected value of the number of genetic effects.

mBIC results from supplementing BIC with the Binomial prior $B(p, C/p)$ on the number of genetic effects.

$2\log p$ term plays a role of the Bonferroni for multiple testing (see e.g. Bogdan, Ghosh, Żak-Szatkowska, QREI, 2008).

Benjamini-Hochberg correction is better (1)

M.B, J.K.G., S.T.Tokdar, *IMS Collections, 2008*

M.B, J.K.G, A.Ochman, S.T.Tokdar,*QREI*, 2007

Multiple testing procedures

$X_1, \ldots, X_p$ - independent, $X_i \sim N(\mu_i, \sigma^2)$

Multiple testing procedures

$X_1, \ldots, X_p$ - independent, $X_i \sim N(\mu_i, \sigma^2)$

$$H_{0i} : \mu_i = 0$$

Multiple testing procedures

$X_1, \ldots, X_p$ - independent, $X_i \sim N(\mu_i, \sigma^2)$

$$H_{0i} : \mu_i = 0$$

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Multiple testing procedures

$X_1, \ldots, X_p$ - independent, $X_i \sim N(\mu_i, \sigma^2)$

$$H_{0i} : \mu_i = 0$$

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq \sigma \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sigma\sqrt{2\log p}(1 + o(1)$.

Multiple testing procedures

$X_1, \ldots, X_p$ - independent, $X_i \sim N(\mu_i, \sigma^2)$

$$H_{0i} : \mu_i = 0$$

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sigma\sqrt{2\log p}(1 + o(1)$.

Benjamini-Hochberg procedure:

(1) $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$

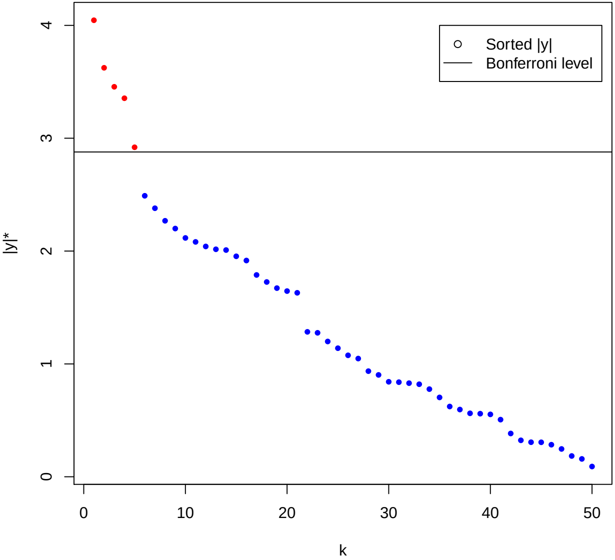
(2) Find the largest index i such that

$$|X|_{(i)} \geq \sigma\Phi^{-1}(1 - \alpha_i), \quad \alpha_i = \alpha\frac{i}{2p}. \tag{1}$$

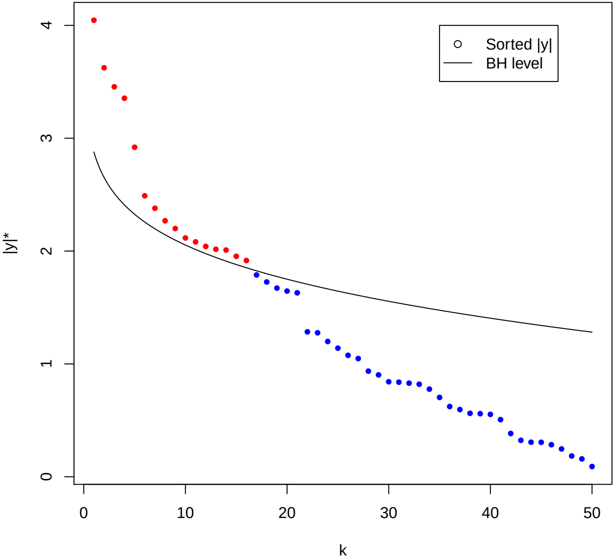
Call this index i_{SU}.

(3) Reject all $H_{(i)}$'s for which $i \leq i_{\mathrm{SU}}$.

Bonferroni correction



Benjamini and Hochberg correction



Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

γ_0 - loss for type I error, γ_A - loss for type II error

Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\theta)\delta_0 + \theta N(0,\tau^2)$

Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\theta)\delta_0 + \theta N(0,\tau^2)$

Bayes oracle → Bayes classifier

Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\theta)\delta_0 + \theta N(0,\tau^2)$
Bayes oracle → Bayes classifier

M.B, A.Chakrabarti, F.Frommlet, JKG, Ann.Statist. 2011: The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if lim $\frac{R}{R_{opt}} \to 1$ (as $p \to \infty$)

Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\theta)\delta_0 + \theta N(0,\tau^2)$

Bayes oracle → Bayes classifier

M.B, A.Chakrabarti, F.Frommlet, JKG, Ann.Statist. 2011: The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if lim $\frac{R}{R_{opt}} \to 1$ (as $p \to \infty$)
BH is ABOS if $\theta \propto p^{-\beta}$, $\beta \in (0,1]$, $\tau \propto \sqrt{2\beta \log p}$

Benjamini-Hochberg correction is better (2)

[B, Ghosh, Tokdar, *IMS Collections* 2008] and [B, Ghosh, Ochman, Tokdar *QREI*, 2007]: empirical comparison of BH with several Bayesian multiple testing procedures with respect to minimizing the Bayes classification risk.

γ_0 - loss for type I error, γ_A - loss for type II error

$\mu_i \sim (1-\theta)\delta_0 + \theta N(0,\tau^2)$

Bayes oracle → Bayes classifier

M.B, A.Chakrabarti, F.Frommlet, JKG, Ann.Statist. 2011: The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if lim $\frac{R}{R_{opt}} \to 1$ (as $p \to \infty$)
BH is ABOS if $\theta \propto p^{-\beta}$, $\beta \in (0,1]$, $\tau \propto \sqrt{2\beta \log p}$
Bonferroni correction is ABOS if $\beta = 1$

mBIC2

$$mBIC2 := \log RSS + k \log n + 2k \log(p/4) - 2 \log(k!)$$

mBIC2

$$mBIC2 := \log RSS + k \log n + 2k \log(p/4) - 2 \log(k!)$$

F.Frommlet, F.Ruhaltinger, P.Twaróg, MB (2011, CSDA)

M. Żak-Szatkowska and MB (CSDA, 2011)

P.Szulc, F. Frommlet, MB, H. Tang (Gen. Epi. 2017)

For similar criteria see also Foster and George (Biometrika 2000) and Abramovich, Benjamini, Donoho and Johnstone (Ann. Statist. 2006).

mBIC2

$$mBIC2 := \log RSS + k \log n + 2k \log(p/4) - 2 \log(k!)$$

F.Frommlet, F.Ruhaltinger, P.Twaróg, MB (2011, CSDA)

M. Żak-Szatkowska and MB (CSDA, 2011)

P.Szulc, F. Frommlet, MB, H. Tang (Gen. Epi. 2017)

For similar criteria see also Foster and George (Biometrika 2000) and Abramovich, Benjamini, Donoho and Johnstone (Ann. Statist. 2006).

mBIC2 is in some sense asymptotically equivalent to the Bayes rule based on the uniform prior on $\{0, \ldots, k_{max}\}$, where $\frac{k_{max}}{p} \rightarrow 0$.

mBIC2

$$mBIC2 := \log RSS + k \log n + 2k \log(p/4) - 2 \log(k!)$$

F.Frommlet, F.Ruhaltinger, P.Twaróg, MB (2011, CSDA)

M. Żak-Szatkowska and MB (CSDA, 2011)

P.Szulc, F. Frommlet, MB, H. Tang (Gen. Epi. 2017)

For similar criteria see also Foster and George (Biometrika 2000) and Abramovich, Benjamini, Donoho and Johnstone (Ann. Statist. 2006).

mBIC2 is in some sense asymptotically equivalent to the Bayes rule based on the uniform prior on $\{0, \ldots, k_{max}\}$, where $\frac{k_{max}}{p} \to 0$.

Problem - numerical complexity of identifying the model minimizing mBIC2.

Different search strategies implemented in the package *bigstep by P.Szulc.*

LASSO

$$\hat{\beta} = argmin_{b \in {}^{p}} \left(\|y - Xb\|_{\ell_2}^{2} + \lambda \|b\|_{\ell_1} \right)$$

LASSO

$$\hat{\beta} = argmin_{b \in {}^{p}} \left(\|y - Xb\|_{\ell_2}^2 + \lambda \, \|b\|_{\ell_1} \right)$$

If $X'X = I$ then LASSO selects X_i if and only if

$$X_i'Y > \lambda$$

LASSO

$$\hat{\beta} = argmin_{b \in ^p} \left(\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1} \right)$$

If $X'X = I$ then LASSO selects X_i if and only if

$$X_i' Y > \lambda$$

LASSO

$$\hat{\beta} = argmin_{b \in ^p} \left(\| y - Xb \|_{\ell_2}^2 + \lambda \, \| b \|_{\ell_1} \right)$$

If $X'X = I$ then LASSO selects X_i if and only if

$$X_i' Y > \lambda$$

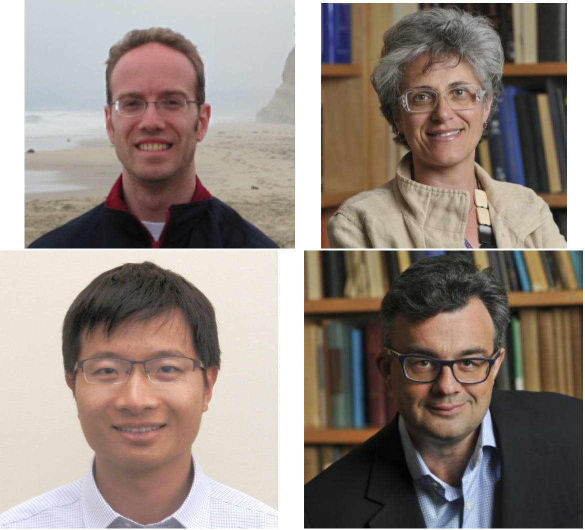
When $\beta_i = 0$ then $X_i' Y \sim N\left(0, \sigma^2\right)$ and the control of FWER is provided by the Bonferroni correction

$$\lambda = \sigma \Phi^{-1} \left(1 - \frac{\alpha}{2p} \right) \approx \sqrt{2 \log p}$$

.

Sorted L-One Penalized Estimation

M.B., E.van den Berg, W.Su, E.J.Candès, arxiv 2013
M.B., E.van den Berg, C.Sabatti, W.Su, E.J.Candès, AOAS 2015



Sorted L-One Penalized Estimation

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

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.

Sorted L-One Penalized Estimation

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

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 Penalized Estimation

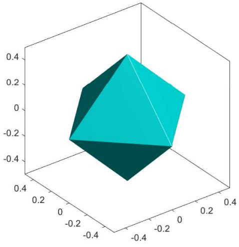
$$\hat{\beta} = argmin_{b \in {}^{p}} \|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

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.

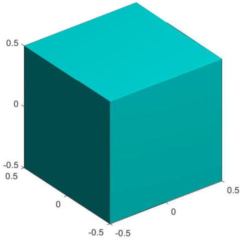
The above optimization problem is convex and can be efficiently solved even for large design matrices.

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 $\lambda_1 > \lambda_2 = \ldots = \lambda_p = 0$.

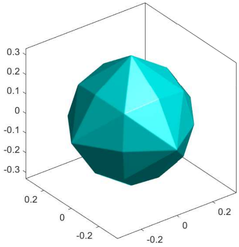
Unit balls for different SLOPE sequences by D.Brzyski



(a) (2,2,2)



(b) (2,0,0)



(c) (3,2,1)

Clustering in the context of portfolio optimization - P. Kremmer, S. Lee, MB and S. Paterlini "Journal of Banking and Finance", 2019

The class of models attainable by SLOPE - U.Schneider and P.Tardivel, arxiv 2020

FDR control with SLOPE

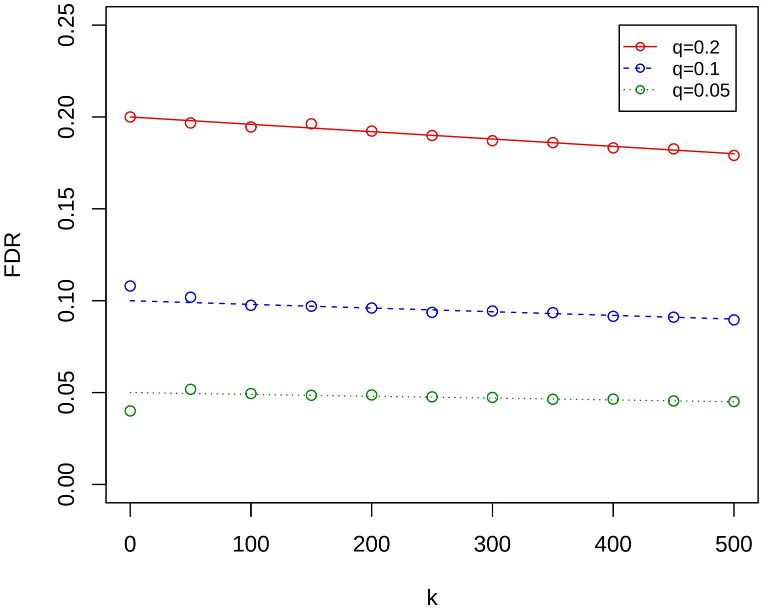
Theorem (B,van den Berg, Su and Candès (2013))

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 minimaxity 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.

Asymptotic minimaxity 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 minimaxity 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 minimaxity 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 - $k \log p$

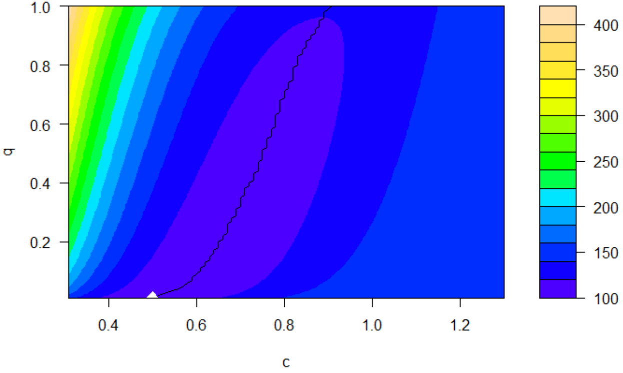
Extension to logistic regression by Abramovich and Grinshtein (2018, IEEE Trans. Inf. Theory)

Predictive properties of SLOPE, Independent predictors

Heat map of $MSE(X\hat{\beta})$

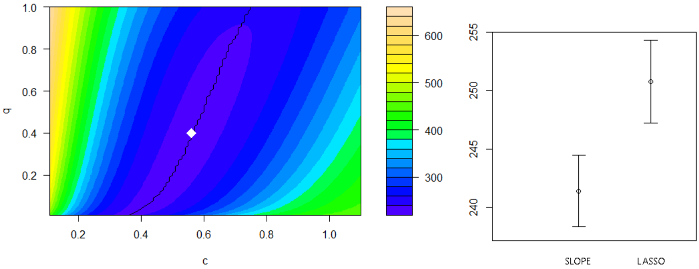
$$\lambda_i = c\Phi\left(1 - \frac{iq}{2p}\right), \quad n = p = 1000, k = 20$$

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



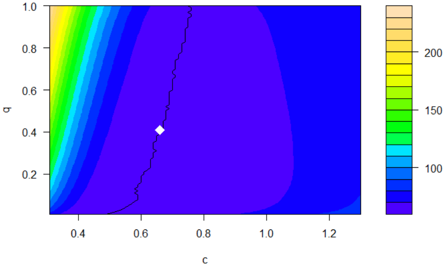
Independent predictors

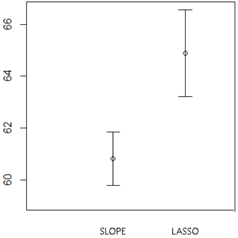
$$n = p = 1000, k = 100$$



Correlated predictors

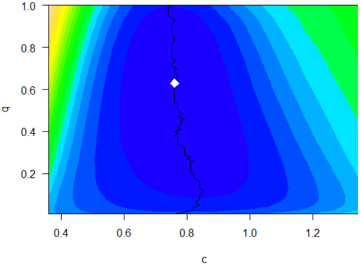
$$n = p = 1000, k = 20, \rho(X_i, X_j) = 0.5 \text{ for } i \neq j$$

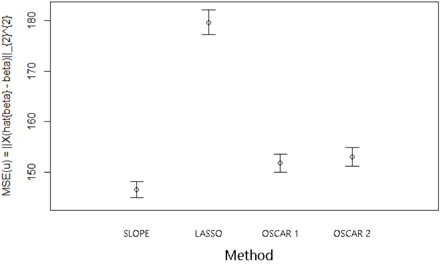




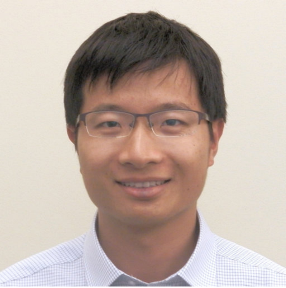
Correlated predictors

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





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:

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

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

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:

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

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

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

Selection of

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

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.

Selection of the group of predictors

Identification of groups of predictors:

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

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

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.

Applications for GWAS

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

Applications for GWAS

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

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

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

Applications for GWAS

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

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

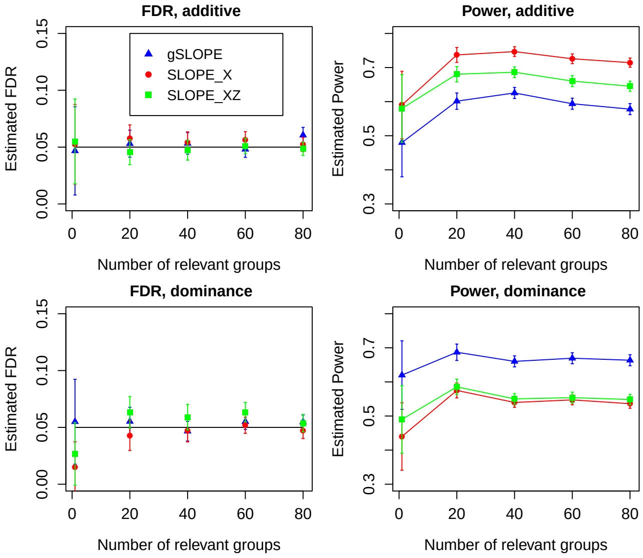
$$X_{ij} = \begin{cases} -1 & \text{for} \quad aa \\ 0 & \text{for} \quad aA \\ 1 & \text{for} \quad AA \end{cases} , \tag{2}$$

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



Robust regression with SLOPE

A.Virouleau, A.Guilloux, S.Gaiffas, MB (arxiv, 2017)

Mean-shift model for robust regression

Candes and Randall (2006), Gannaz (2006) and McCann and Welsch (CSDA, 2007) ,

$$y = X\beta + I\mu + \varepsilon \tag{4}$$

$\mu \in R^n$ is the sparse vector of outliers' effects and $\varepsilon \sim N(0, \sigma^2 I)$

Mean-shift model for robust regression

Candes and Randall (2006), Gannaz (2006) and McCann and Welsch (CSDA, 2007) ,

$$y = X\beta + I\mu + \varepsilon \tag{4}$$

$\mu \in R^n$ is the sparse vector of outliers' effects and $\varepsilon \sim N(0, \sigma^2 I)$

She and Owen (IPOD, JASA, 2012) and Nguyen and Tran (E-lasso, IEEE Trans. Inf. Th., 2013) use L_1 penalty for μ and β

Mean-shift model for robust regression

Candes and Randall (2006), Gannaz (2006) and McCann and Welsch (CSDA, 2007) ,

$$y = X\beta + I\mu + \varepsilon \tag{4}$$

$\mu \in R^n$ is the sparse vector of outliers' effects and $\varepsilon \sim N(0, \sigma^2 I)$

She and Owen (IPOD, JASA, 2012) and Nguyen and Tran (E-lasso, IEEE Trans. Inf. Th., 2013) use L_1 penalty for μ and β

Virouleau, Guilloux, Gaiffas, B (2017) use SLOPE penalties:

$$\min_{\beta \in {}^p, \mu \in {}^n} \left\{ \|y - X\beta - \mu\|_2^2 + 2\rho_1 J_{\tilde{\lambda}}(\beta) + 2\rho_2 J_{\lambda}(\mu) \right\}$$

Mean-shift model for robust regression

Candes and Randall (2006), Gannaz (2006) and McCann and Welsch (CSDA, 2007) ,

$$y = X\beta + I\mu + \varepsilon \tag{4}$$

$\mu \in R^n$ is the sparse vector of outliers' effects and $\varepsilon \sim N(0, \sigma^2 I)$

She and Owen (IPOD, JASA, 2012) and Nguyen and Tran (E-lasso, IEEE Trans. Inf. Th., 2013) use L_1 penalty for μ and β

Virouleau, Guilloux, Gaiffas, B (2017) use SLOPE penalties:

$$\min_{\beta \in^p, \mu \in^n} \left\{ \|y - X\beta - \mu\|_2^2 + 2\rho_1 J_{\tilde{\lambda}}(\beta) + 2\rho_2 J_\lambda(\mu) \right\}$$

$$\lambda_i(\beta) = \sigma \sqrt{\log\left(\frac{2p}{i}\right)}, \ \ \lambda_i(\mu) = \sigma \sqrt{\log\left(\frac{2n}{i}\right)}$$

Estimation properties

Assumptions - restricted eigenvalue condition on X

Satisfied with large probability e.g. if the rows of X are iid gaussian with a positive definite covariance matrix and the numbers of nonzero elements in β and μ are sufficiently small

Estimation properties

Assumptions - restricted eigenvalue condition on X

Satisfied with large probability e.g. if the rows of X are iid gaussian with a positive definite covariance matrix and the numbers of nonzero elements in β and μ are sufficiently small

Rates of convergence for

$$\|\hat{\beta} - \beta\|_2^2 + \|\hat{\mu} - \mu\|_2^2$$

Estimation properties

Assumptions - restricted eigenvalue condition on X

Satisfied with large probability e.g. if the rows of X are iid gaussian with a positive definite covariance matrix and the numbers of nonzero elements in β and μ are sufficiently small

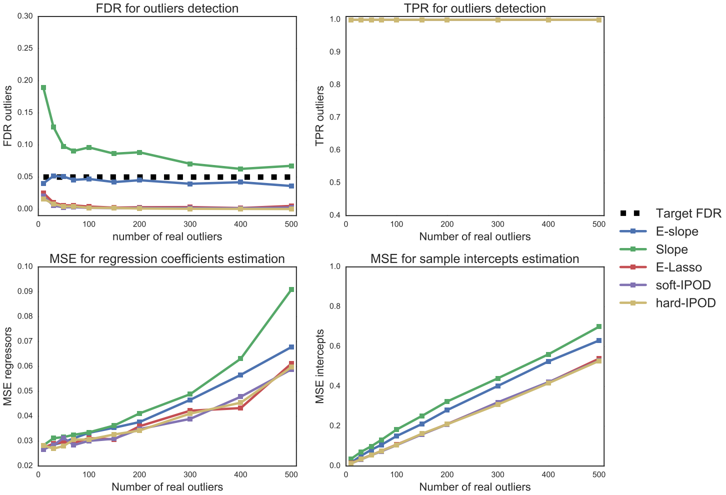
Rates of convergence for

$$\|\hat{\beta} - \beta\|_2^2 + \|\hat{\mu} - \mu\|_2^2$$

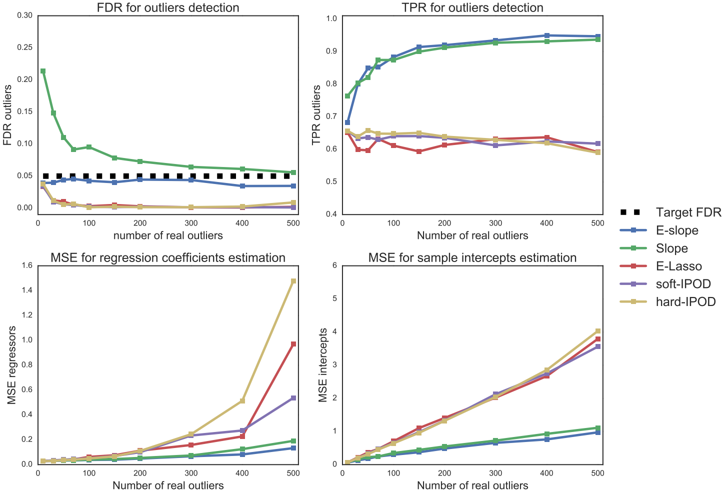
$s = \#\{i : \mu_i \neq 0\}$ - number of outliers,

NO/SL1	$(p \vee s \log(n/s))/n$
L1/L1	$(k \log p \vee s \log n)/n$
L1/SL1	$(k \log p \vee s \log(n/s))/n$
SL1/SL1	$(k \log (p/k) \vee s \log (n/s))/n$

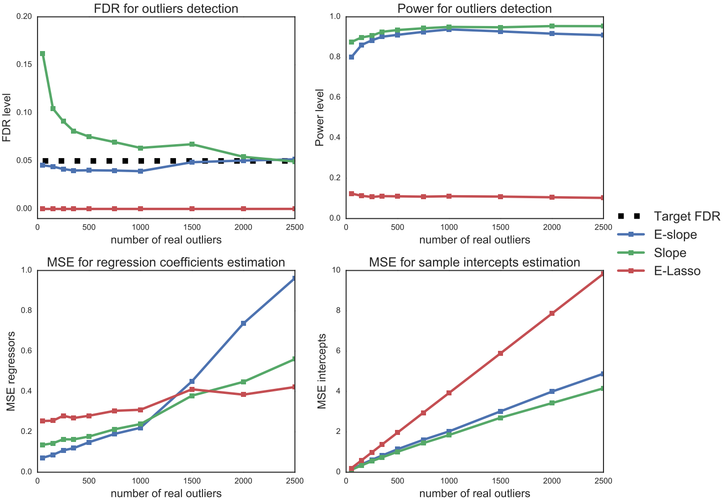
Low dimensional set-up; large outliers



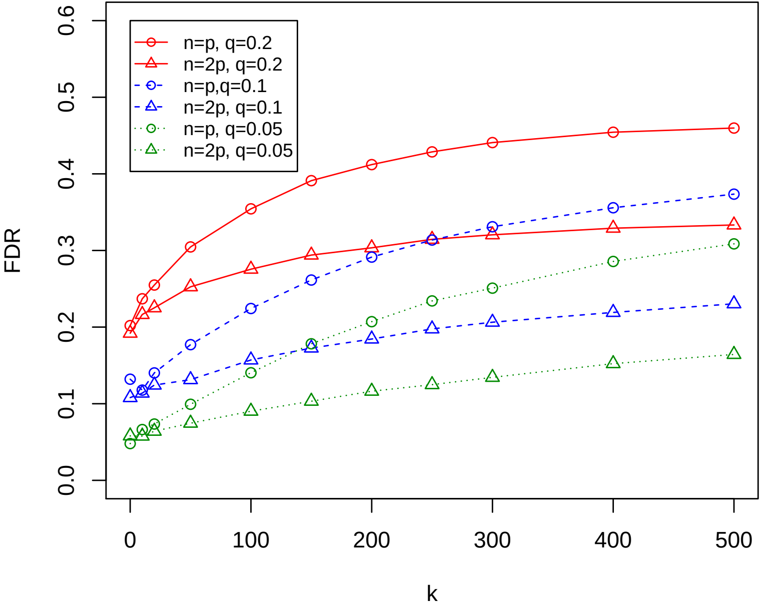
Low dimensional set-up; small outliers



High dimensional set-up; small outliers



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



Problems with FDR control

Similar problems occur for LASSO.

Intuitive explanation:

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

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

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

.

Problems with FDR control

Similar problems occur for LASSO.

Intuitive explanation:

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

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

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

The magnitude of v_i depends on λ (level of shrinkage), the level of sparsity and magnitude of true signals.

Problems with FDR control

Similar problems occur for LASSO.

Intuitive explanation:

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

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

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

The magnitude of v_i depends on λ (level of shrinkage), the level of sparsity and magnitude of true signals.

LASSO can identify the true model only if a very stringent *irrepresentability* condition is satisfied.

Problems with FDR control

Similar problems occur for LASSO.

Intuitive explanation:

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

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

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

The magnitude of v_i depends on λ (level of shrinkage), the level of sparsity and magnitude of true signals.

LASSO can identify the true model only if a very stringent *irrepresentability* condition is satisfied.

Precise FDR-Power Tradeoff under asymptotic assumptions of AMP theory is provided in (Su, B, Candès, AOS 2017).

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^1 norm if the following implication holds

$$X\gamma = X\beta \text{ and } \gamma \neq \beta \Rightarrow \|\gamma\|_1 > \|\beta\|_1. \tag{5}$$

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.

Solutions:

- threshold LASSO estimates (see e.g. LCD knockoffs)
- use adaptive LASSO

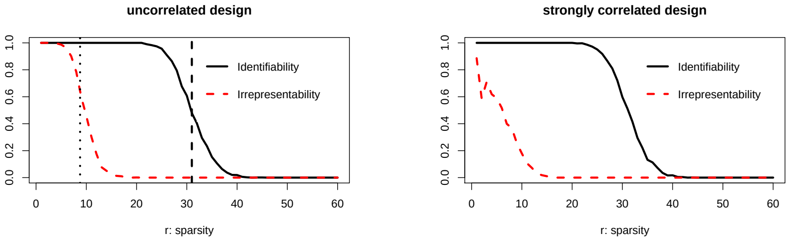


Figure: $n = 100$, $p = 300$, in the right panel $\rho(X_i, X_j) = 0.9$, vertical lines correspond to $n/(2 \log p)$ and the transition curve of Donoho and Tanner (2009).

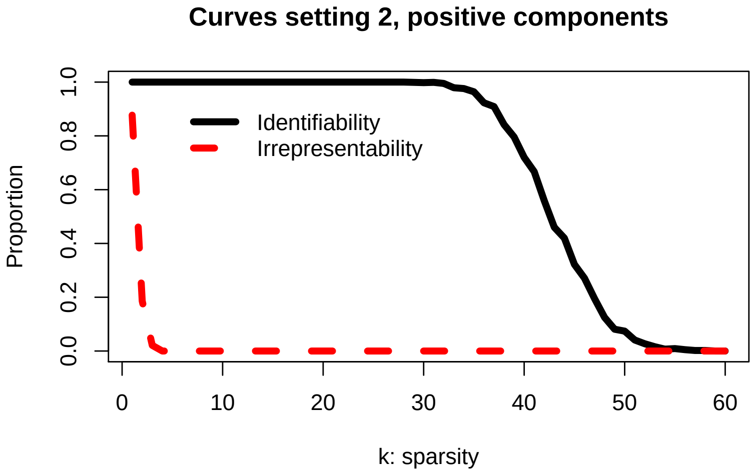


Figure: $n = 100$, $p = 300$, in the right panel $\rho(X_i, X_j) = 0.9$ and all signs of nonzero elements of β are the same.

Adaptive LASSO

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

$$\beta_{aL} = argmin_b \left\{ \frac{1}{2} \left\| y - Xb \right\|_2^2 + \lambda \sum_{i=1}^{p} w_i |b|_i \right\}, \tag{6}$$

where $w_i = \frac{1}{f(|\hat{\beta}_i|)}$, $\hat{\beta}_i$ is some consistent estimator of β_i and f is an increasing function.

Spike and Slab LASSO

V.Rockova, E. George, JASA 2018

Spike and Slab LASSO

V.Rockova, E. George, JASA 2018

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

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

Spike and Slab LASSO

V.Rockova, E. George, JASA 2018

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

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

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"

$$\pi(\beta|\lambda, \gamma) \propto c^{\sum_{i=1}^{p} 1(\gamma_i = 1)} \prod_{i=1}^{p} e^{-w_i|\beta_i|\lambda_0},$$

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

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda_0 \sum_{i=1}^{p} w_i |b_i|$$

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

Spike and Slab LASSO (2)

The maximum aposteriori rule is given by reweighted LASSO

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2} ||y - Xb||_2^2 + \lambda_0 \sum_{i=1}^{p} w_i |b_i|$$

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

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

$$\hat{\beta}(\gamma) = argmin_{b \in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda_0 \sum_{i=1}^{p} w_i |b_i|$$

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

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.

Borrowing information

When updating i^{th} variable θ is replaced by $E(\theta|\beta_{-i})$

Borrowing information

When updating i^{th} variable θ is replaced by $E(\theta|\beta_{-i})$

$\lambda_1 = c\lambda_0$ - fixed at some small value

Borrowing information

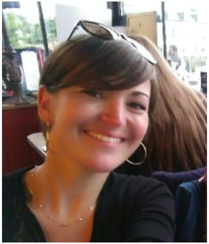
When updating i^{th} variable θ is replaced by $E(\theta|\beta_{-i})$

$\lambda_1 = c\lambda_0$ - fixed at some small value

SSL package creates the path of SSL solutions for the sequence of 100 λ_0 values

Adaptive SLOPE with missing values (1)

W. Jiang, MB, J.Josse, B.Miasojedow, V.Rockova, TraumaBase Group, arxiv 2019



traumabase.eu

Motivation: Paris Hospital

- *Traumabase*$^{\circledR}$ data:
  20000 major trauma patients × 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*® data:
  20000 major trauma patients × 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: Paris Hospital

- *Traumabase*$^{®}$ data:
  20000 major trauma patients × 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**?

Adaptive Bayesian SLOPE

We propose an adaptive version of Bayesian SLOPE (ABSLOPE). After standardizing X so each column has a unit L_2 norm, the prior for β is

$$\mathrm{p}(\beta \mid \gamma, c, \sigma^2; \lambda) \propto c^{\sum_{j=1}^{p} \mathbb{I}(\gamma_j=1)} \prod_j \exp\left\{-w_j|\beta_j|\frac{1}{\sigma}\lambda_{r(W\beta,j)}\right\},$$

Interpretation of the model:

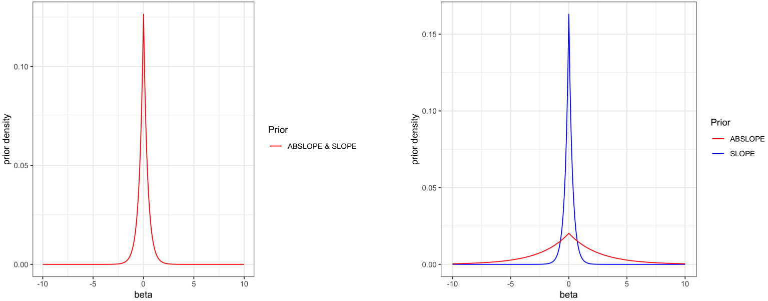
- β_j is large enough $\Rightarrow$ **true signal**; 0 $\Rightarrow$ noise.
- $\gamma_j \in \{0,1\}$ signal indicator. $\gamma_j|\theta \sim \mathit{Bernoulli}(\theta)$ and θ the **sparsity**.
- $1/c \in [1,\infty)$: proportional to the **average signal magnitude**.
- $W = \mathrm{diag}(w_1, w_2, \cdots, w_p)$ and its diagonal element:

$$w_j = c\gamma_j + (1-\gamma_j) = \begin{cases} c, & \gamma_j = 1 \\ 1, & \gamma_j = 0 \end{cases}.$$

Adaptive Bayesian SLOPE

Advantage of introducing W:

- when $\gamma_j = 0$, $w_j = 1$, i.e., the null variables are treated with the regular SLOPE penalty
- when $\gamma_j = 1$, $w_j = c < 1$, i.e, **smaller penalty** $\lambda_{r(W\beta,j)}$ for true predictors than the regular SLOPE one



(a) Null β

(b) Non-null β

Figure: comparison of SLOPE prior and ABSLOPE prior

Major difference between SSL and ABSLOPE

ABSLOPE spike prior is "fixed" and frequentist motivated, with the aim of FDR control

Major difference between SSL and ABSLOPE

ABSLOPE spike prior is ”fixed” and frequentist motivated, with the aim of FDR control

Slab component is ”estimated” via the estimation of the average signal magnitude

Model selection with missing values

Decomposition: $X = (X_{\text{obs}}, X_{\text{mis}})$

Pattern: matrix M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$

Assumption 1: Missing at random (MAR)

$\mathrm{p}(M \mid X_{\text{obs}}, X_{\text{mis}}) = \mathrm{p}(M \mid X_{\text{obs}}) \quad \Rightarrow \quad$ ignorable missing patterns
e.g. People at older age didn't tell his income at larger probability.

Assumption 2: Distribution of covariates

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

Model selection with missing values

Decomposition: $X = (X_{\mathrm{obs}}, X_{\mathrm{mis}})$

Pattern: matrix M with $M_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is observed} \\ 0, & \text{otherwise} \end{cases}$

Assumption 1: Missing at random (MAR)

$\mathrm{p}(M \mid X_{\mathrm{obs}}, X_{\mathrm{mis}}) = \mathrm{p}(M \mid X_{\mathrm{obs}}) \quad \Rightarrow \quad$ ignorable missing patterns
e.g. People at older age didn't tell his income at larger probability.

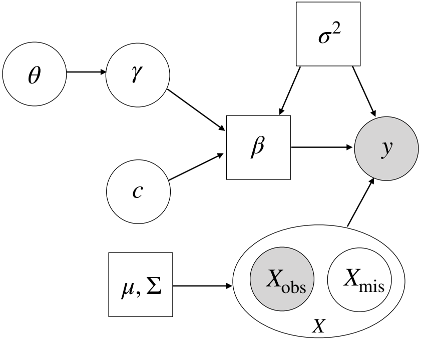
Assumption 2: Distribution of covariates

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

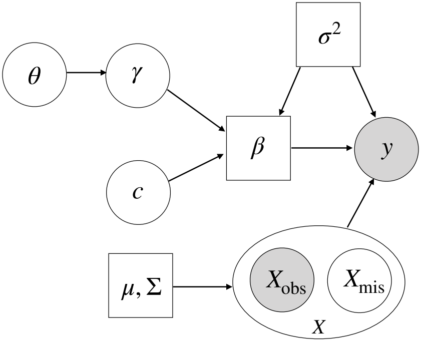
Problem: With NA, only a few methods are available to select a model, and their performances are limited. For example,

- (Claeskens and Consentino, 2008) adapts AIC to missing values ⇒ Impossible to deal with high dimensional analysis.
- (Loh and Wainwright, 2012) LASSO with NA
  ⇒ Non-convex optimization; requires to know bound of $\|\beta\|_1$
  ⇒ difficult in practice

ABSLOPE with missingness: Summary



ABSLOPE with missingness: Summary



$$\begin{aligned}\ell_{\text{comp}} &= \log \mathrm{p}(y, X, \gamma, c;\ \beta, \theta, \sigma^2) \\ &= \log\left\{\mathrm{p}(X;\ \mu, \Sigma)\,\mathrm{p}(y \mid X;\ \beta, \sigma^2)\,\mathrm{p}(\beta;\ \gamma, c)\,\mathrm{p}(\gamma;\ \theta)\,\mathrm{p}(c)\right\}\end{aligned}$$

Objective: Maximize $\ell_{\text{obs}} = \iiint \ell_{\text{comp}}\, dX_{\text{mis}}\, dc\, d\theta\, d\gamma$.

EM algorithm

- *E step:* evaluate

$$Q^t = \mathbb{E}(\ell_{\text{comp}}) \quad \text{wrt} \quad \mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^t, \sigma^t, \mu^t, \Sigma^t).$$

- *M step:* update

$$\beta^t, \sigma^t, \mu^t, \Sigma^t = \arg\max Q^t$$

Problem: The function Q is not tractable. ⇒

1. Monte Carlo EM ? (Wei and Tanner 1990)

EM algorithm

- *E step:* evaluate

$$Q^t = \mathbb{E}(\ell_{\text{comp}}) \quad \text{wrt} \quad \text{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^t, \sigma^t, \mu^t, \Sigma^t).$$

- *M step:* update

$$\beta^t, \sigma^t, \mu^t, \Sigma^t = \arg\max Q^t$$

Problem: The function Q is not tractable. ⇒

1. ~~Monte Carlo EM ?~~
   Expensive to generate a large number of samples.
2. Stochastic Approximation EM (book, Lavielle 2014)
   - One sample in each iteration;

Adapted SAEM algorithm

- *E step:*
  $Q^t = \mathbb{E}(\ell_{\text{comp}}) \quad \text{wrt} \quad \mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^t, \sigma^t, \mu^t, \Sigma^t)$.
  - *Simulation:* draw one sample $(X^t_{\text{mis}}, \gamma^t, c^t, \theta^t)$ from

    $$\mathrm{p}(X_{\text{mis}}, \gamma, c, \theta \mid y, X_{\text{obs}}, \beta^{t-1}, \sigma^{t-1}, \mu^{t-1}, \Sigma^{t-1});$$

    [**Gibbs sampling**]
  - *Stochastic approximation:* update function Q with

    $$Q^t = Q^{t-1} + \eta_t \left(\ell_{\text{comp}}(X^t_{\text{mis}}, \gamma^t, c^t, \theta^t) - Q^{t-1} \right).$$

- *M step:* $\beta^{t+1}, \sigma^{t+1}, \mu^{t+1}, \Sigma^{t+1} = \arg\max Q^{t+1}$.
  [**When $\eta_t = 1$: Reweighted SLOPE, Shrinkage of covariance**]

Details of initialization, generating samples and optimization are in the draft (arXiv:1909.06631)

SLOBE

Instead of using Gibbs sampling γ and c are replaced with the approximation to their conditional expectations given data, β and σ

R package: ABSLOPE

Install package:

```
library(devtools)
install_github("wjiang94/ABSLOPE")
```

R package: ABSLOPE

Install package:

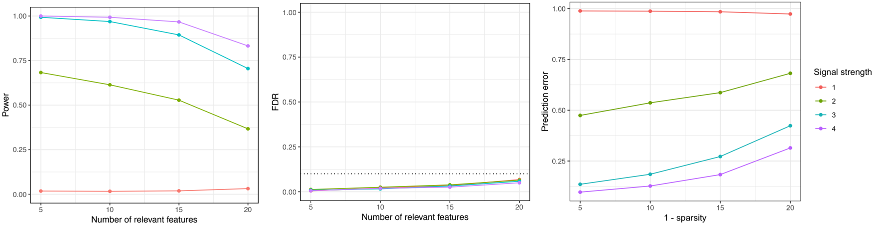
```
library(devtools)
install_github("wjiang94/ABSLOPE")
```

Main algorithm:

```
lambda = create_lambda_bhq(ncol(X),fdr=0.10)
list.res = ABSLOPE(X, y, lambda)
```

Simulation study (200 rep. ⇒ average)

$n = p = 100$, no correlation and 10% missingness



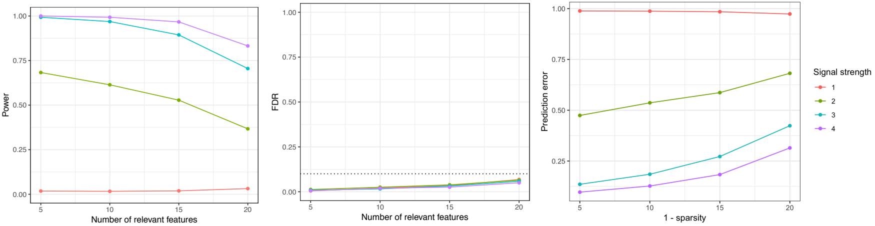
(a) Power

(b) FDR

(c) Prediction error

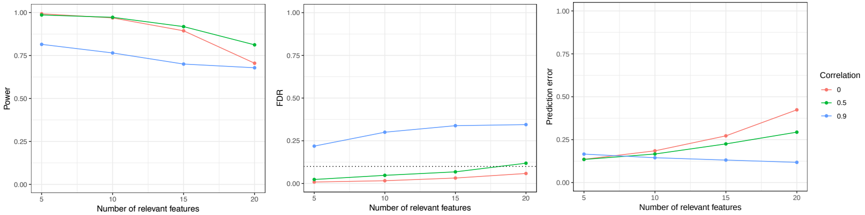
Simulation study (200 rep. ⇒ average)

$n = p = 100$, no correlation and 10% missingness



(g) Power (h) FDR (i) Prediction error

$n = p = 100$, with 10% missingness and strong signal



(j) Power (k) FDR (l) Prediction error

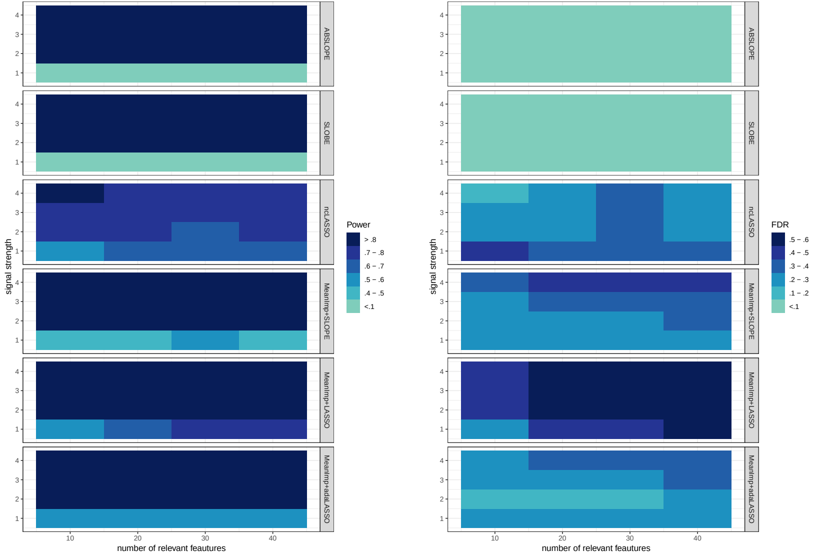
Method comparison

- **ABSLOPE** and **SLOBE**
- **ncLASSO:** non convex LASSO (Loh and Wainwright, 2012)
- **MeanImp + SLOPE:** Mean imputation followed by SLOPE with known σ
- **MeanImp + LASSO:** Mean imputation followed by LASSO, with λ tuned by cross validation
- **MeanImp + adaLASSO:** Mean imputation followed by adaptive LASSO (Zou, 2006)

In the SLOPE type methods, $\lambda =$ BH sequence which controls the FDR at level **0.1**

Method comparison (200 rep. ⇒ average)

500×500 dataset, 10% missingness, $Sigma_{i,j} = 0.5^{|i-j|}$



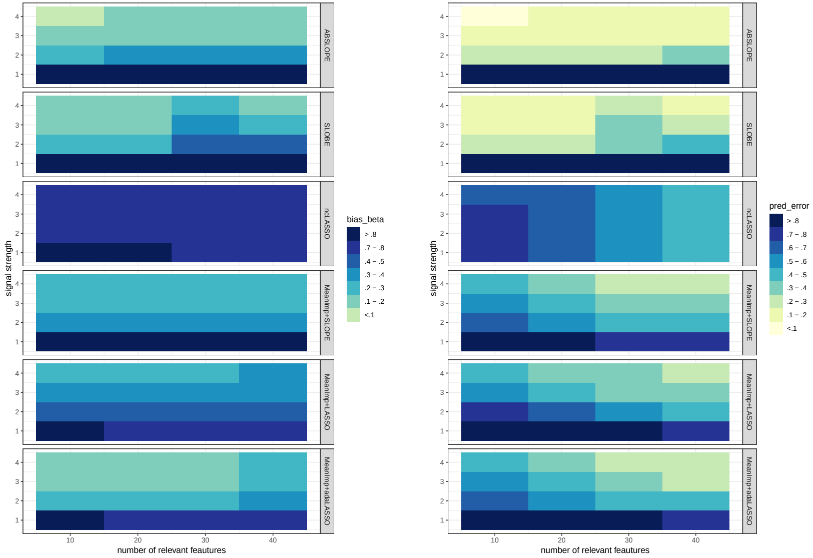
(m) Power

(n) FDR

Figure: Comparison of power (a), FDR (b), bias of β (c) and prediction error

Method comparison (200 rep. ⇒ average)

500×500 dataset, 10% missingness, with correlation

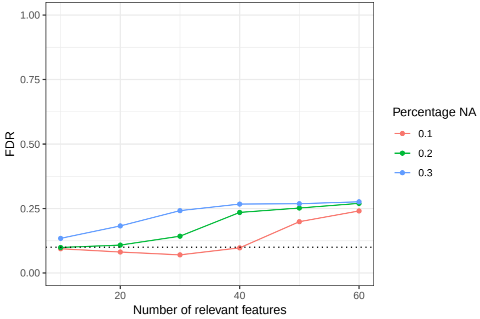


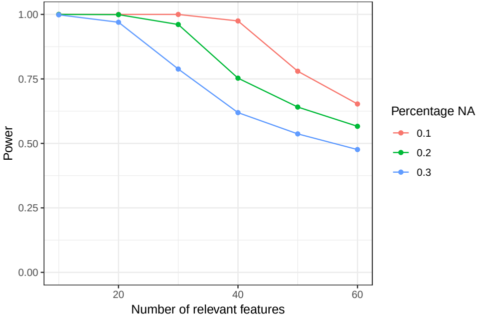
(a) Bias of β

(b) Prediction error

Figure: Comparison of power (a), FDR (b), bias of β (c) and prediction error

$n = p = 500$, $\Sigma_{ij} = 0.5^{|i-j|}$, FDR and Power





Variables in the TraumaBase data set (APHP)

Goal - quick prediction of the level of platelets

- *Age:* Age
- *SI:* Shock index indicates level of occult shock based on heart rate (FC) and systolic blood pressure (PAS). $SI = \frac{FC}{PAS}$. Evaluated on arrival of hospital.
- *PAM:* Mean arterial pressure is an average blood pressure in an individual during a single cardiac cycle, based on systolic blood pressure (PAS) and diastolic blood pressure (PAD). $PAM = \frac{2PAD+PAS}{3}$. Evaluated on arrival of hospital.
- *delta_Hemocue:* The difference between the hemoglobin on arrival at hospital and that in the ambulance.
- *Temps.lieux.hop:* Time spent in hospital *i.e.*, medicalization time, in minutes.
- *Lactates:* The conjugate base of lactic acid.
- *Temperature:* Patient's body temperature.

Variables

- *FC:* heart rate measured on arrival of hospital.
- *Remplissage:* A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *CGR.dechoc:* A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI.SMUR:* Shock index measured on ambulance.
- *PAM.SMUR:* Mean arterial pressure measured in the ambulance.
- *FC.max:* Maximum value of measured heart rate in the ambulance.
- *PAS.min:* Minimum value of measured systolic blood pressure in the ambulance.
- *PAD.min:* Minimum value of measured diastolic blood pressure in the ambulance.

Percentage of missing values

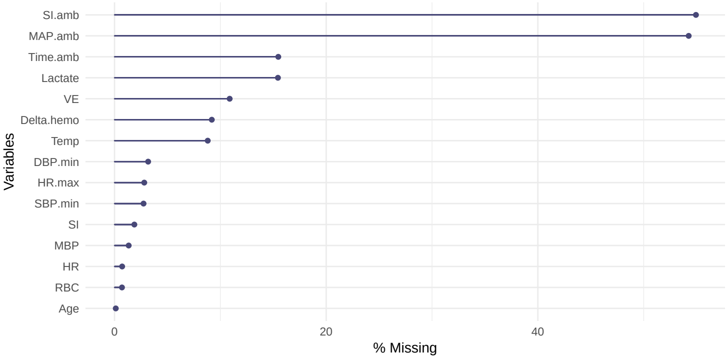
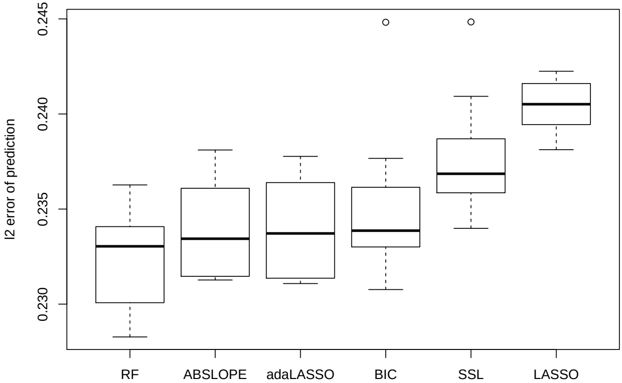


Figure: Percentage of missing values in each pre-selected variable from TraumaBase.

Results

TraumaBase: Measurements $\xrightarrow{\text{Predict}}$ Platelet
Cross-validation: random splits to training and test sets × 10



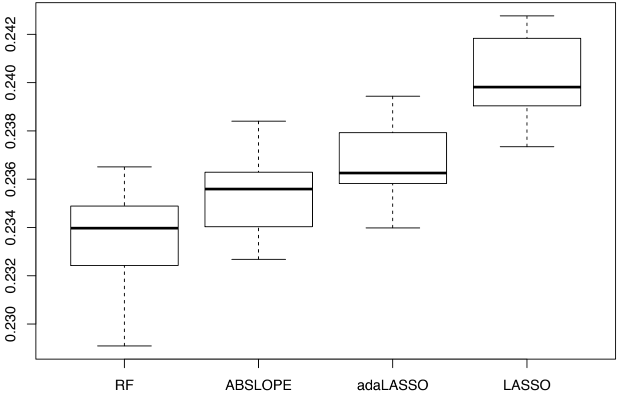
- Comparable to random forest
- Interpretable model selection and estimation results

Selected variables

Figure: Number of times that each variable is selected over 10 replications. Bold numbers indicate which variables are included in the model selected by ABSLOPE.

Variable	ABSLOPE	SLOPE	LASSO	adaLASSO	BIC
Age	**10**	10	4	10	10
SI	**10**	2	0	0	9
MBP	1	10	1	10	1
Delta.hemo	**10**	10	8	10	10
Time.amb	2	6	0	4	0
Lactate	**10**	10	10	10	10
Temp	2	10	0	0	0
HR	**10**	10	1	10	10
VE	**10**	10	2	10	10
RBC	**10**	10	10	10	10
SI.amb	0	0	0	0	0
MBP.amb	0	0	0	0	0
HR.max	3	9	0	1	0
SBP.min	5	10	10	10	[illegible]

With interactions



Selected variables

Method	Variables selected
ABSLOPE	Age * MBP.amb, Delta.hemo * Lactate Lactate * RBC, HR * SBP.min
LASSO	RBC, SBP.min Age * Lactate, Age * VE Delta.hemo * Lactate, Delta.hemo * VE Lactate * VE, Lactate * RBC
adaLASSO	Age * Time.amb, Age * HR Age * MBP.amb, Age * SBP.min MBP * HR, Delta.hemo * VE Lactate * VE,HR * HR.max HR * SBP.min, VE * RBC

Conclusion & Future research

Conclusion:

- ABSLOPE reduces the estimation bias of large regression coefficients.
- This allows for
  1. Improved estimation and prediction properties
  2. FDR control under much wider range of scenarios than for regular SLOPE
- Modeling in a Bayesian framework allows for the estimation of the structure of predictors such as the **signal sparsity** and the **signal strength**;

Future research:

- Deal with other missing mechanisms
- Application for other statistical models (e.g. GLM or Gaussian Graphical Models)
- Theoretical analysis of statistical properties (asymptotic FDR control, minimaxity)
- **Speeding the SLOPE algorithm**

Strong screening rule for SLOPE

J. Larsson, MB, J. Wallin (2020)

Predictor screening rules

Goal

Constructing the SLOPE solution path corresponding to the sequence λ^j, $j \in \{1, \ldots, m\}$ such that for all $i \in \{1, \ldots, p\}$, $\lambda_i^j > \lambda_i^{j+1}$

Predictor screening rules

Goal

Constructing the SLOPE solution path corresponding to the sequence λ^j, $j \in \{1, \ldots, m\}$ such that for all $i \in \{1, \ldots, p\}$, $\lambda_i^j > \lambda_i^{j+1}$

Basic idea

Use the solution at the step j to construct a relatively cheap test to determine which predictors will be inactive before fitting the model for the step $j + 1$.

Predictor screening rules

Goal

Constructing the SLOPE solution path corresponding to the sequence λ^j, $j \in \{1, \ldots, m\}$ such that for all $i \in \{1, \ldots, p\}$, $\lambda_i^j > \lambda_i^{j+1}$

Basic idea

Use the solution at the step j to construct a relatively cheap test to determine which predictors will be inactive before fitting the model for the step $j+1$.

Safe and Heuristic Rules

safe rules	certifies that discarded predictors are not in model
heuristic rules	may incorrectly discard some predictors, which means problem must sometimes be solved several times (in practice never more than twice)

Motivation for lasso strong rule

Assume we are solving the lasso, i.e. minimizing

$$g(\beta) + h(\beta), \qquad h(\beta) := \lambda \sum_{i=1}^{p} |\beta_i|.$$

KKT stationarity condition is

$$\mathbf{0} \in \nabla g(\hat{\beta}) + \partial h(\hat{\beta}),$$

where $\partial h(\hat{\beta})$ is the subdifferential for the ℓ_1 norm with elements given by

$$\partial h(\hat{\beta})_i = \begin{cases} \text{sign}(\hat{\beta}_i)\lambda & \hat{\beta}_i \neq 0 \\ [-\lambda, \lambda] & \hat{\beta}_i = 0, \end{cases}$$

which means that $|\nabla g(\hat{\beta})_i| < \lambda \implies \hat{\beta}_i = 0$.

Gradient estimate

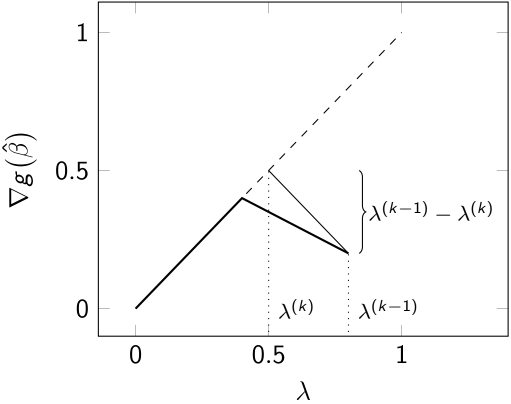
Assume that we are fitting a regularization path and have $\hat{\beta}(\lambda^{(k-1)})$—the solution for $\lambda^{(k-1)}$—and want to discard predictors corresponding to the problem for $\lambda^{(k)}$.

Basic idea: replace $\nabla g(\hat{\beta})$ with an estimate and apply the KKT stationarity criterion, discarding predictors that are estimated to be zero.

What estimate should we use?

The unit slope bound

A simple (and usually conservative) estimate turns out to be $\lambda^{(k-1)} - \lambda^{(k)}$, i.e. assume that the gradient is piece-wise linear function with slope bounded by 1.



The strong rule for the lasso

Discard the *j*th predictor if

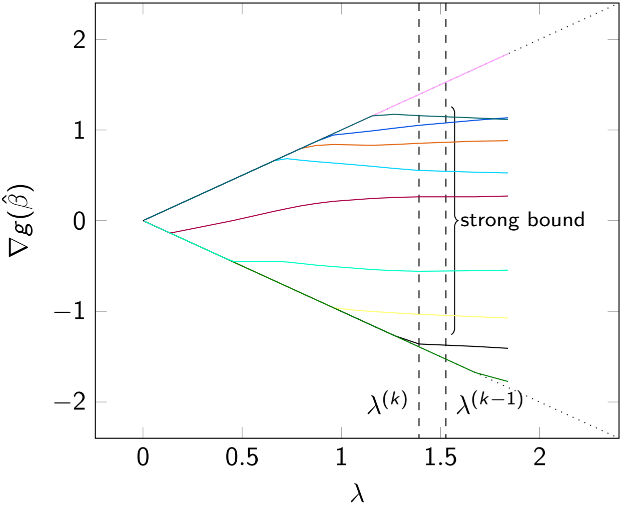
$$\underbrace{\underbrace{\left|\nabla g\left(\hat{\beta}(\lambda^{(k-1)})\right)\right|}_{\text{previous gradient}} + \underbrace{\lambda^{(k-1)} - \lambda^{(k)}}_{\text{unit slope bound}}}_{\text{gradient prediction for } k} < \lambda^{(k)}$$

$$\iff$$

$$\left|\nabla g\left(\hat{\beta}(\lambda^{(k-1)})\right)\right| < 2\lambda^{(k)} - \lambda^{(k-1)}$$

Empirical results show that the strong rule leads to remarkable performance improvements in $p \gg n$ regime (and no penalty otherwise) (**tibshirani2012**).

Strong rule for lasso in action



Strong rule for SLOPE

Exactly the same idea as for lasso strong rule.

The subdifferential for SLOPE is is the set of all $g \in \mathbb{R}^p$ such that

$$g_{\mathcal{A}_i} = \left\{ s \in \mathbb{R}^{\mathrm{card}\,\mathcal{A}_i} \;\middle|\; \begin{cases} \mathrm{cumsum}(|s|_\downarrow - \lambda_{R(s)_{\mathcal{A}_i}}) \preceq \mathbf{0} & \text{if } \beta_{\mathcal{A}_i} = \mathbf{0}, \\ \mathrm{cumsum}(|s|_\downarrow - \lambda_{R(s)_{\mathcal{A}_i}}) \preceq \mathbf{0} & \\ \quad \land \sum_{j \in \mathcal{A}_i} \left(|s_j| - \lambda_{R(s)_j}\right) = 0 & \text{otherwise.} \end{cases} \right\}$$

$\mathcal{A}_i$ defines a cluster containing indices of coefficients equal in absolute value.

$R(x)$ is an operator that returns the ranks of elements in $|x|$.

$|x|_\downarrow$ returns the absolute values of x sorted in non-increasing order.

Strong rule algorithm for SLOPE

```
Input: $c \in \mathbb{R}^p$, $\lambda \in \mathbb{R}^p$, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.
$\mathcal{S}, \mathcal{B} \leftarrow \varnothing$
for $i \leftarrow 1, \ldots, p$ do
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{i\}$
    if $\sum_{j \in \mathcal{B}} \big(c_j - \lambda_j\big) \geq 0$ then
        $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{B}$
        $\mathcal{B} \leftarrow \varnothing$
    end if
end for
Return $\mathcal{S}$ =0
```

Set

$$c := |\nabla g(\hat{\beta}) + \lambda^{(k-1)} - \lambda^{(k)}|_{\downarrow} \qquad \lambda := \lambda^{(k)},$$

and run the algorithm above; the result is the predicted support for $\hat{\beta}(\lambda^{(k)})$ (subject to a permutation).

Efficiency for simulated data

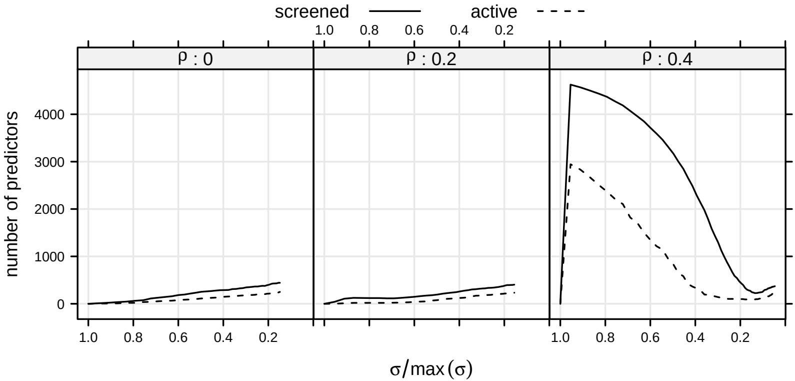


Figure: Gaussian design, $X \in \mathbb{R}^{200 \times 5000}$, predictors pairwise correlated with correlation ρ. There were no violations of the strong rule here.

Efficiency for real data

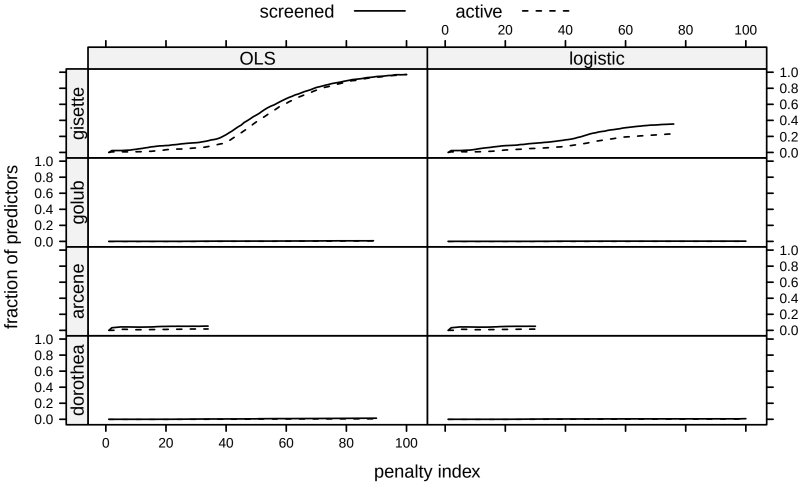


Figure: Efficiency for real data sets. The dimensions of the predictor matrices are 100 × 9920 (arcene), 800 × 88119 (dorothea), 6000 × 4955 (gisette), and 38 × 7129 (golub).

Violations

Violations may occur if the unit slope bound fails, which can occur if ordering permutation of absolute gradient changes, or any predictor becomes active between $\lambda^{(k-1)}$ and $\lambda^{(k)}$.

Thankfully, such violations turn out to be rare.

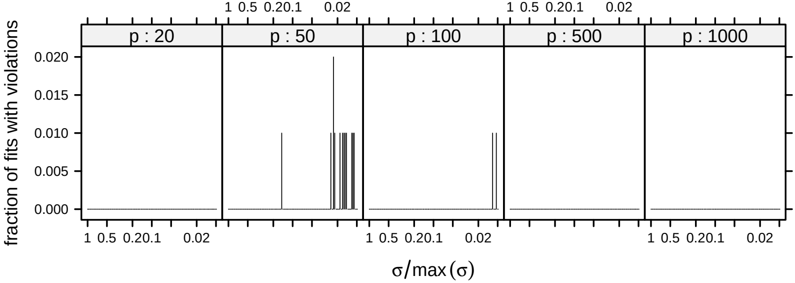


Figure: Violations for sorted ℓ_1 regularized least squares regression with predictors pairwise correlated with $\rho = 0.5$. $X \in \mathbb{R}^{100\times p}$.

Performance

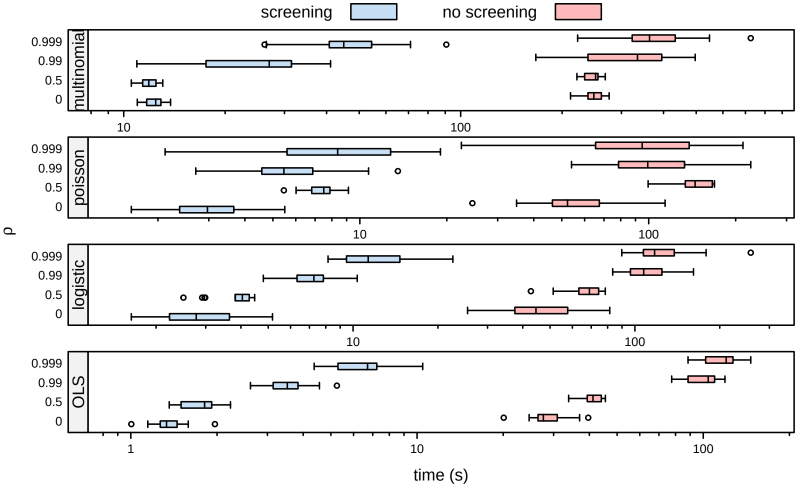


Figure: Performance benchmarks for various generalized linear models with $X \in \mathbb{R}^{200 \times 20000}$. Predictors are autocorrelated through an AR(1) process with correlation ρ.

Algorithms

The original strong rule paper (**tibshirani2012**) presents two strategies for using the screening rule. For SLOPE, we have two slightly modified versions of these algorithms

strong set algorithm

initialize $\mathcal{E}$ with strong rule set

1. fit SLOPE to predictors in $\mathcal{E}$
2. check KKT criteria against $\mathcal{E}^C$; if there are any failures, add predictors that fail the check to $\mathcal{E}$ and go back to 1

Algorithms

The original strong rule paper (**tibshirani2012**) presents two strategies for using the screening rule. For SLOPE, we have two slightly modified versions of these algorithms

strong set algorithm

initialize $\mathcal{E}$ with strong rule set

1. fit SLOPE to predictors in $\mathcal{E}$
2. check KKT criteria against $\mathcal{E}^C$; if there are any failures, add predictors that fail the check to $\mathcal{E}$ and go back to 1

previous set algorithm

initialize $\mathcal{E}$ with ever-active predictors

1. fit SLOPE to predictors in $\mathcal{E}$
2. check KKT criteria against predictors in strong set
   - if there are any failures, include these predictors in $\mathcal{E}$ and go back to 1
   - if there are no failures, check KKT criteria against remaining predictors; if there are any failures, add these to $\mathcal{E}$ and go back to 1

Comparing algorithms

Strong set strategy marginally better for low–medium correlation

Previous set strategy starts to become useful for high correlation

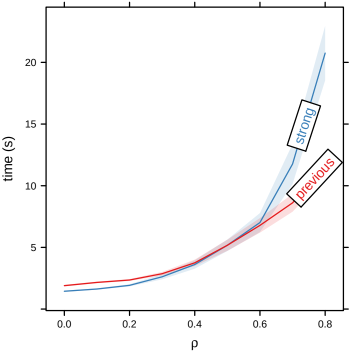


Figure: Performance of strong and previous set strategies for OLS problems with varying correlation between predictors.

Limitations

- the unit slope bound is generally very conservative
- does not use second-order structure in any way
- current methods for solving SLOPE (FISTA, ADMM) do not make as good use of screening rules as coordinate descent does (for the lasso)

The SLOPE package for R

Strong screening rule for SLOPE has been implemented in the R package SLOPE (`https://CRAN.R-project.org/package=SLOPE`).

Features include

- OLS, logistic, Poisson, and multinomial models
- support for sparse and dense predictors
- cross-validation
- efficient codebase in C++

Proximal Newton solver for SLOPE will be implemented this summer with the Google Summer of Code program.