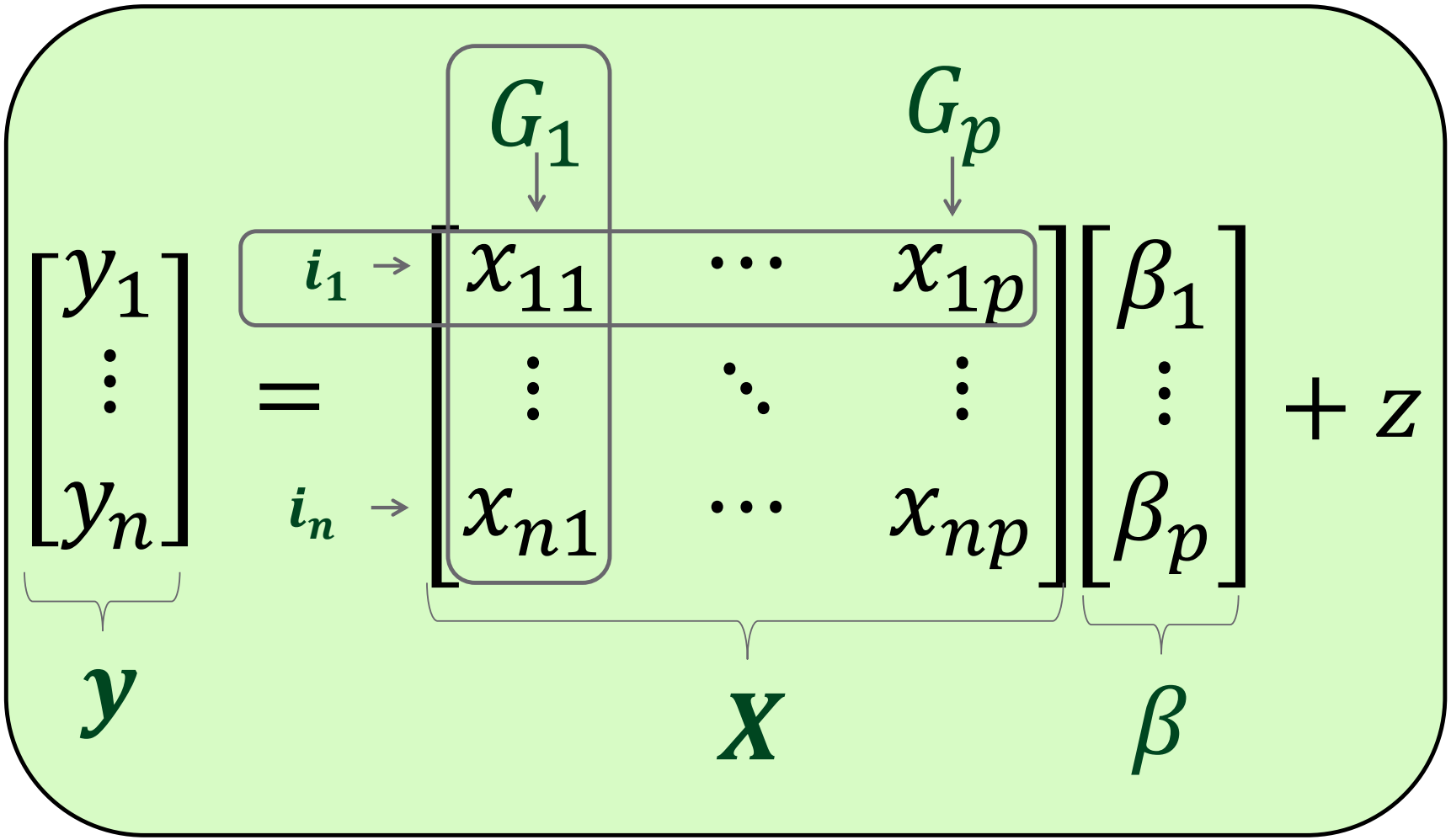
The selection of relevant groups of explanatory variables in GWA studies

Damian Brzyski

Department of Mathematics and Computer Science, Wroclaw University of Technology, Poland
Institute of Mathematics, Jagiellonian University, Cracow, Poland

31.08.2015

Genome-wide association studies (GWAS)



Model selection problem

- Consider the linear regression model of form $y = X\beta + z$, with experiment matrix $X \in M(n, p)$ (with centered, ℓ_2 normalized columns), observation vector y and $z \sim N(0, \sigma^2 I_n)$

Model selection problem

- Consider the linear regression model of form $y = X\beta + z$, with experiment matrix $X \in M(n, p)$ (with centered, ℓ_2 normalized columns), observation vector y and $z \sim N(0, \sigma^2 I_n)$
- We assume that the number of nonzero coefficients in β is small comparing to p (i.e. β is sparse)

Model selection problem

- Consider the linear regression model of form $y = X\beta + z$, with experiment matrix $X \in M(n, p)$ (with centered, ℓ_2 normalized columns), observation vector y and $z \sim N(0, \sigma^2 I_n)$
- We assume that the number of nonzero coefficients in β is small comparing to p (i.e. β is sparse)
- The task is to find the support of β, which corresponds to finding relevant explanatory variables

Existing penalized methods: LASSO

- LASSO is defined as solution to

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

with $\lambda_L > 0$ being tuning parameter

Existing penalized methods: LASSO

- LASSO is defined as solution to

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

  with $\lambda_L > 0$ being tuning parameter

- 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)

Existing penalized methods: LASSO

- LASSO is defined as solution to

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

  with $\lambda_L > 0$ being tuning parameter

- 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)
- Choosing of λ_L is challenging - it is not obvious which sparsity level could be perceived as proper

Existing penalized methods: SLOPE

- SLOPE is defined as solution to

$$\arg\min_b \frac{1}{2}\|y - Xb\|_2^2 + \sigma \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \qquad \text{(SLOPE)}$$

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

Existing penalized methods: SLOPE

- SLOPE is defined as solution to

$$\arg\min_b \frac{1}{2}\|y - Xb\|_2^2 + \sigma \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \qquad \text{(SLOPE)}$$

  where $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ 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 could be efficiently solved even for large design matrices

Existing penalized methods: SLOPE

- SLOPE is defined as solution to

$$\arg\min_{b} \frac{1}{2}\|y - Xb\|_2^2 + \sigma \sum_{i=1}^{p} \lambda_i |b|_{(i)}, \qquad \text{(SLOPE)}$$

  where $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ 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 could be efficiently solved even for large design matrices
- The method reduces to LASSO, when $\lambda_1 = \ldots = \lambda_p$

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β
- We define:

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β
- We define:
  - the number of all discoveries, $R := \left|\left\{i : \ \widetilde{\beta}_i \neq 0\right\}\right|$

False discovery rate (FDR) control

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

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β
- We define:
  - the number of all discoveries, $R := \left|\left\{i : \ \widetilde{\beta}_i \neq 0\right\}\right|$
  - the number of false discoveries, $V := \left|\left\{i : \ \beta_i = 0, \quad \widetilde{\beta}_i \neq 0\right\}\right|$
  - false discovery rate, $FDR := \mathbb{E}\left[\frac{V}{\max\{R,1\}}\right]$

False discovery rate (FDR) control

- Let $\widetilde{\beta}$ be estimate of β
- We define:
  - the number of all discoveries, $R := \left|\left\{i : \ \widetilde{\beta}_i \neq 0\right\}\right|$
  - the number of false discoveries, $V := \left|\left\{i : \ \beta_i = 0, \quad \widetilde{\beta}_i \neq 0\right\}\right|$
  - false discovery rate, $FDR := \mathbb{E}\left[\frac{V}{\max\{R,1\}}\right]$
- The goal is to construct the method for which tuning parameters could be chosen (in explicit way) such as the condition $FDR \leq q$ is met, for predefined $q \in (0, 1)$

FDR control with SLOPE

- In orthogonal situation, i.e. when $X_i^\top X_j = 0$ for $i \neq j$, the condition $FDR \leq q$ is theoretically provided when λ sequence is defined as

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

FDR control with SLOPE

- In orthogonal situation, i.e. when $X_i^\top X_j = 0$ for $i \neq j$, the condition $FDR \leq q$ is theoretically provided when λ sequence is defined as

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

- The heuristic procedure for choosing smoothing parameters was derived for the near orthogonal situation, which was modeled by assuming that entries of X are realizations of independent, zero-mean normal distributions

FDR control with SLOPE

- In orthogonal situation, i.e. when $X_i^\top X_j = 0$ for $i \neq j$, the condition $FDR \leq q$ is theoretically provided when λ sequence is defined as

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

- The heuristic procedure for choosing smoothing parameters was derived for the near orthogonal situation, which was modeled by assuming that entries of X are realizations of independent, zero-mean normal distributions
- It turns out that the mentioned heuristic works well for many other zero-mean distributions

The structure of GWAS data

- Tendency of strong correlation between nearly located columns while columns of distant indices are generally weakly correlated

The structure of GWAS data

- Tendency of strong correlation between nearly located columns while columns of distant indices are generally weakly correlated

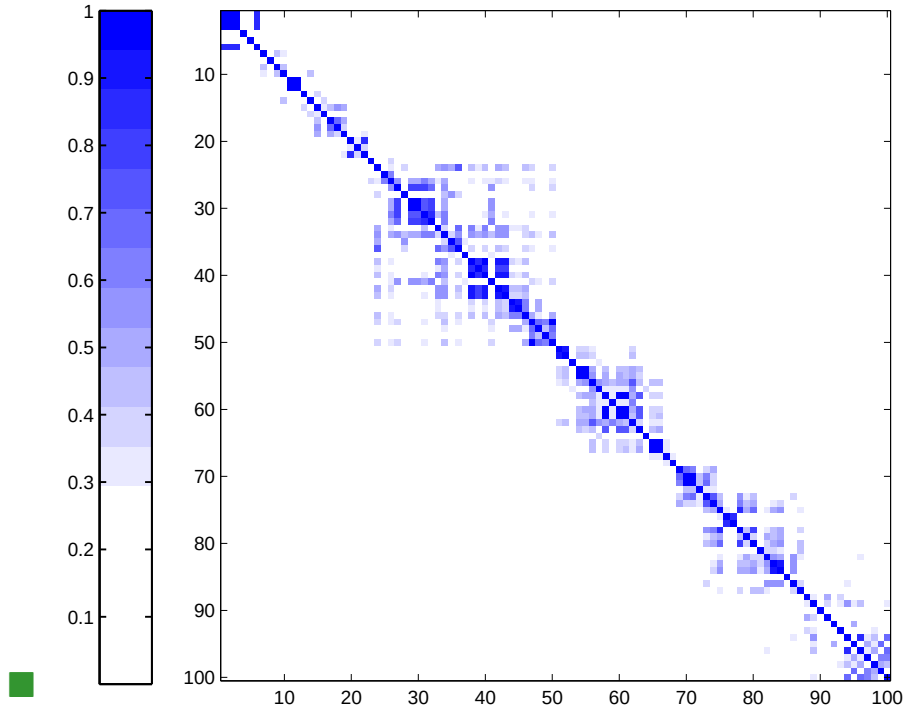
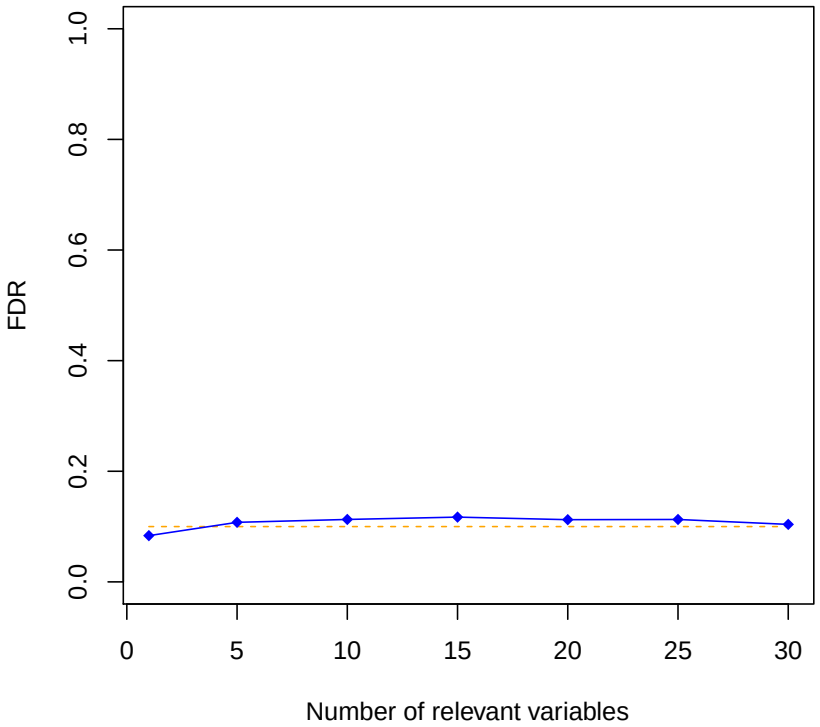
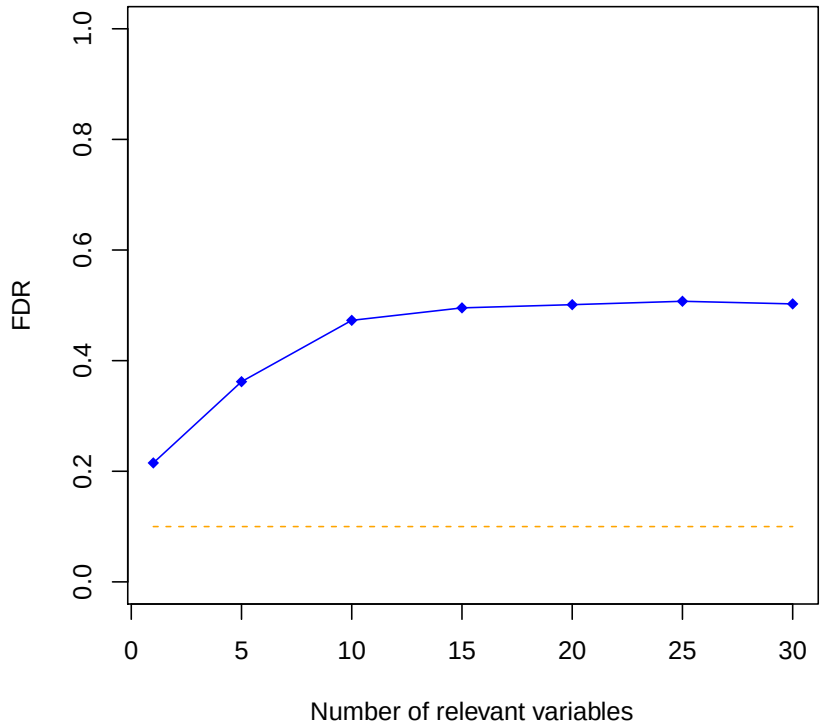


Figure: Histogram of correlation matrix (absolute values) for 100 predictors

SLOPE in GWA studies



(a) Norm. dist. entries, $n = p = 1000$



(b) GWAS data, $n = p = 1000$

New strategy: selecting groups

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$ and denote $l_i := |I_i|$ for $i = 1, \ldots, m$

New strategy: selecting groups

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$ and denote $l_i := |I_i|$ for $i = 1, \ldots, m$
- Consider the linear regression model with m groups of form

$$y = \sum_{j=1}^{m} X_{I_j} \beta_{I_j} + z$$

New strategy: selecting groups

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$ and denote $l_i := |I_i|$ for $i = 1, \ldots, m$
- Consider the linear regression model with m groups of form

$$y = \sum_{j=1}^{m} X_{I_j} \beta_{I_j} + z$$

- We define truly relevant group by condition $\|X_{I_j} \beta_{I_j}\|_2 > 0$

New strategy: selecting groups

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$ and denote $l_i := |I_i|$ for $i = 1, \ldots, m$
- Consider the linear regression model with m groups of form

$$y = \sum_{j=1}^{m} X_{I_j} \beta_{I_j} + z$$

- We define truly relevant group by condition $\|X_{I_j} \beta_{I_j}\|_2 > 0$
- The task is to identify the relevant group instead of individual predictors

New strategy: selecting groups

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$ and denote $l_i := |I_i|$ for $i = 1, \ldots, m$
- Consider the linear regression model with m groups of form

$$y = \sum_{j=1}^{m} X_{I_j} \beta_{I_j} + z$$

- We define truly relevant group by condition $\|X_{I_j} \beta_{I_j}\|_2 > 0$
- The task is to identify the relevant group instead of individual predictors
- STANDARDIZATION: X_{I_j} could be decomposed as $X_{I_j} = U_j R_j$, where $U_j^\top U_j = \mathbf{I}$, we can define $\widetilde{\beta} := \left((R_1 \beta_{I_1})^\top, \ldots, (R_m \beta_{I_m})^\top\right)^\top$, $\widetilde{\beta}_{\widetilde{I}_j} := R_j \beta_{I_j}$. Then

$$\|X_{I_j} \beta_{I_j}\|_2 > 0 \iff \|U_j \widetilde{\beta}_{\widetilde{I}_j}\|_2 > 0 \iff \|\widetilde{\beta}_{\widetilde{I}_j}\|_2 > 0$$

Group false discovery rate (gFDR) control

- Let $\widetilde{\beta}$ be estimate of β and $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$

Group false discovery rate (gFDR) control

- Let $\widetilde{\beta}$ be estimate of β and $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$
- We define:

Group false discovery rate (gFDR) control

- Let $\widetilde{\beta}$ be estimate of β and $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$
- We define:
  - the number of all discovered groups, $gR := \left|\left\{i : \ \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$

Group false discovery rate (gFDR) control

- Let $\widetilde{\beta}$ be estimate of β and $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$
- We define:
  - the number of all discovered groups,
    $gR := \left|\left\{i : \ \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$
  - the number of falsely discovered groups,
    $gV := \left|\left\{i : \ \|X_{I_i}\beta_{I_i}\|_2 = 0, \quad \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$

Group false discovery rate (gFDR) control

- Let $\widetilde{\beta}$ be estimate of β and $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$
- We define:
  - the number of all discovered groups, $gR := \left|\left\{i : \ \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$
  - the number of falsely discovered groups, $gV := \left|\left\{i : \ \|X_{I_i}\beta_{I_i}\|_2 = 0, \quad \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$
  - group false discovery rate, $gFDR := \mathbb{E}\left[\frac{gV}{\max\{gR,\ 1\}}\right]$

Group false discovery rate (gFDR) control

- Let $\widetilde{\beta}$ be estimate of β and $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$
- We define:
  - the number of all discovered groups, $gR := \left|\left\{i : \ \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$
  - the number of falsely discovered groups, $gV := \left|\left\{i : \ \|X_{I_i}\beta_{I_i}\|_2 = 0, \quad \|X_{I_i}\widetilde{\beta}_{I_i}\|_2 > 0\right\}\right|$
  - group false discovery rate, $gFDR := \mathbb{E}\left[\frac{gV}{\max\{gR,\ 1\}}\right]$
- The goal is to control $gFDR$ at assumed level $q \in (0, 1)$

Group SLOPE (gFDR)

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$, $l_i = |I_i|$ and $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$

Group SLOPE (gFDR)

- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$, $l_i = |I_i|$ and $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$
- We introduce the group SLOPE estimate, defined as a solution to

$$\arg\min_b \left\{ \frac{1}{2} \left\| y - Xb \right\|_2^2 + \sigma \sum_{i=1}^{m} \lambda_i \sqrt{l_{(i)}} \left\| b_{I_{(i)}} \right\|_2 \right\}, \qquad \text{(gSLOPE)}$$

where $\sqrt{l_{(i)}} \left\| b_{I_{(i)}} \right\|_2$ is the ith largest coefficient of the vector

$$\left(\sqrt{l_1} \left\| b_{I_1} \right\|_2, \ldots, \sqrt{l_m} \left\| b_{I_m} \right\|_2 \right)^\top$$

Group SLOPE (gFDR)

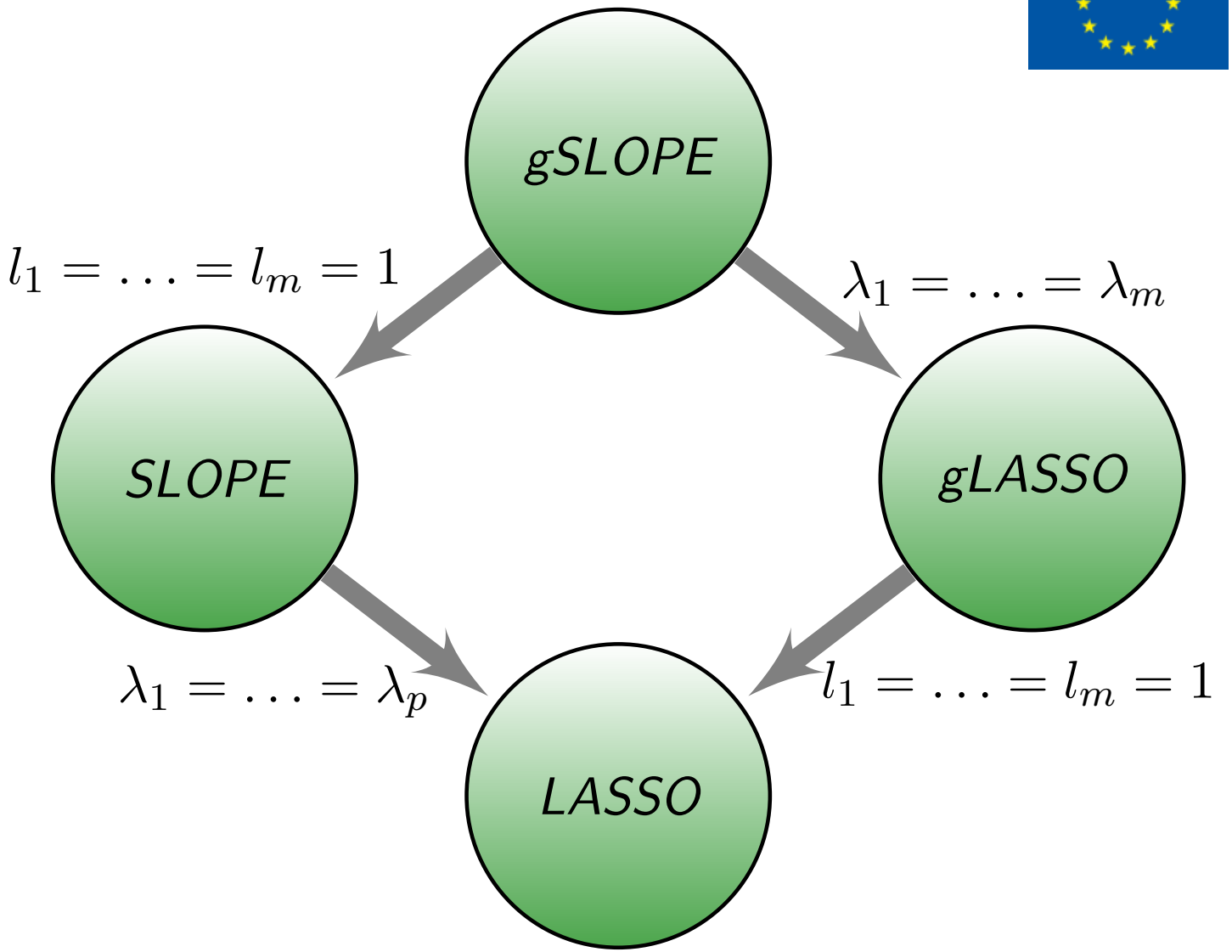
- Let $I = \{I_1, \ldots, I_m\}$ be partition of $\{1, \ldots, p\}$, $l_i = |I_i|$ and $\lambda_1 \geq \ldots \geq \lambda_m \geq 0$
- We introduce the group SLOPE estimate, defined as a solution to

$$\arg\min_b \quad \left\{ \frac{1}{2} \left\| y - Xb \right\|_2^2 + \sigma \sum_{i=1}^{m} \lambda_i \sqrt{l_{(i)}} \left\| b_{I_{(i)}} \right\|_2 \right\}, \qquad \text{(gSLOPE)}$$

where $\sqrt{l_{(i)}} \left\| b_{I_{(i)}} \right\|_2$ is the ith largest coefficient of the vector

$$\left(\sqrt{l_1} \left\| b_{I_1} \right\|_2, \ldots, \sqrt{l_m} \left\| b_{I_m} \right\|_2 \right)^\top$$

- gSLOPE is solution to convex optimization problem which could be efficiently solved (by proximal gradient method)


gSLOPE
$l_1 = \ldots = l_m = 1$
$\lambda_1 = \ldots = \lambda_m$
SLOPE
gLASSO
$\lambda_1 = \ldots = \lambda_p$
$l_1 = \ldots = l_m = 1$
LASSO

Theorem (gFDR control under orthogonal case)

Consider linear regression model with m groups, in which X is experiment matrix satisfying $X_{I_i}^\top X_{I_j} = 0$, for any $i \neq j$. Let m_0 denote the number of truly irrelevant groups. Apply following steps:

- *redefine X, I, $\{l_i\}_{i=1}^m$, p by applying the standardization*
- *fix $q \in (0,1)$*
- *define $\lambda = [\lambda_1, \ldots, \lambda_m]^\top$, for $\lambda_i := \max\limits_{j=1,\ldots,m} \left\{ \frac{1}{\sqrt{l_j}} F_{\chi_{l_j}}^{-1}\left(1 - \frac{q \cdot i}{m}\right) \right\}$, where $F_{\chi_{l_i}}$ is cumulative of chi distribution with l_i degrees of freedom*
- $\widetilde{\beta} := \arg\min\limits_b \left\{ \frac{1}{2} \left\| y - Xb \right\|_2^2 + \sigma \sum_{i=1}^m \lambda_i \sqrt{l_{(i)}} \left\| b_{I_{(i)}} \right\|_2 \right\}$;

Theorem (gFDR control under orthogonal case)

Consider linear regression model with m groups, in which X is experiment matrix satisfying $X_{I_i}^\top X_{I_j} = 0$, for any $i \neq j$. Let m_0 denote the number of truly irrelevant groups. Apply following steps:

- *redefine X, I, $\{l_i\}_{i=1}^m$, p by applying the standardization*
- *fix $q \in (0,1)$*
- *define $\lambda = [\lambda_1, \ldots, \lambda_m]^\top$, for $\lambda_i := \max\limits_{j=1,\ldots,m} \left\{ \frac{1}{\sqrt{l_j}} F^{-1}_{\chi_{l_j}} \left(1 - \frac{q \cdot i}{m}\right) \right\}$, where $F_{\chi_{l_i}}$ is cumulative of chi distribution with l_i degrees of freedom*
- $\widetilde{\beta} := \arg\min\limits_b \left\{ \frac{1}{2} \left\| y - Xb \right\|_2^2 + \sigma \sum_{i=1}^m \lambda_i \sqrt{l_{(i)}} \left\| b_{I_{(i)}} \right\|_2 \right\}$;

Then, it holds

$$gFDR \leq q \cdot \frac{m_0}{m}.$$

Results for near orthogonal situation

Our main goal was to find procedure for generating tuning parameters for gSLOPE in situation when explanatory variables included to different groups are weakly correlated (as in GWAS case).

Results for near orthogonal situation

Our main goal was to find procedure for generating tuning parameters for gSLOPE in situation when explanatory variables included to different groups are weakly correlated (as in GWAS case).

- Basing on heuristics for SLOPE, we have derived the procedure for gSLOPE in situation when entries of design matrix are realizations of independent, zero mean, normal distributions and all groups have the same sizes

Results for near orthogonal situation

Our main goal was to find procedure for generating tuning parameters for gSLOPE in situation when explanatory variables included to different groups are weakly correlated (as in GWAS case).

- Basing on heuristics for SLOPE, we have derived the procedure for gSLOPE in situation when entries of design matrix are realizations of independent, zero mean, normal distributions and all groups have the same sizes
- For arbitrary group sizes we considered two approaches: the conservative (giving gFDR significantly lower than assumed) and the liberal (giving gFDR slightly above the target level but identifying more truly relevant groups than conservative)

Results for near orthogonal situation

Our main goal was to find procedure for generating tuning parameters for gSLOPE in situation when explanatory variables included to different groups are weakly correlated (as in GWAS case).

- Basing on heuristics for SLOPE, we have derived the procedure for gSLOPE in situation when entries of design matrix are realizations of independent, zero mean, normal distributions and all groups have the same sizes
- For arbitrary group sizes we considered two approaches: the conservative (giving gFDR significantly lower than assumed) and the liberal (giving gFDR slightly above the target level but identifying more truly relevant groups than conservative)
- For the fixed sequence of tuning parameters, we have developed the iterative version of gSLOPE, allowing the estimation of σ^2

GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1

GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1
- Columns of the experiment matrix were divided into groups by using the hierarchical clustering algorithm (HCA)

GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1
- Columns of the experiment matrix were divided into groups by using the hierarchical clustering algorithm (HCA)
- As a results we achieved 1358 groups, with average group size close to 12

GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1
- Columns of the experiment matrix were divided into groups by using the hierarchical clustering algorithm (HCA)
- As a results we achieved 1358 groups, with average group size close to 12
- We have performed 200 iterations for each sparsity level from the set $[1, 4, 8, 13, 18, 24]$, in each iteration we generated observations using linear regression model with $\sigma = 1$

GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1
- Columns of the experiment matrix were divided into groups by using the hierarchical clustering algorithm (HCA)
- As a results we achieved 1358 groups, with average group size close to 12
- We have performed 200 iterations for each sparsity level from the set $[1, 4, 8, 13, 18, 24]$, in each iteration we generated observations using linear regression model with $\sigma = 1$
- Tuning parameters were obtain by applying the conservative and the liberal strategies for target gFDR level 0.1

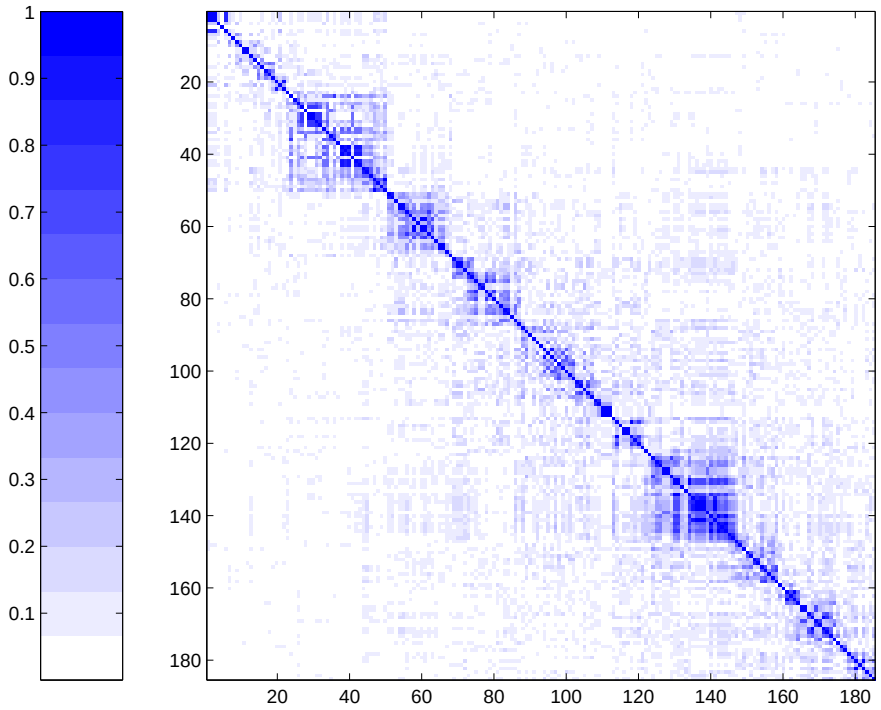
GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1
- Columns of the experiment matrix were divided into groups by using the hierarchical clustering algorithm (HCA)
- As a results we achieved 1358 groups, with average group size close to 12
- We have performed 200 iterations for each sparsity level from the set $[1, 4, 8, 13, 18, 24]$, in each iteration we generated observations using linear regression model with $\sigma = 1$
- Tuning parameters were obtain by applying the conservative and the liberal strategies for target gFDR level 0.1
- Coefficients in β_{I_j} , for truly relevant group j, were generated such as $\|X_{I_j}\beta_{I_j}\|_2 = 8$

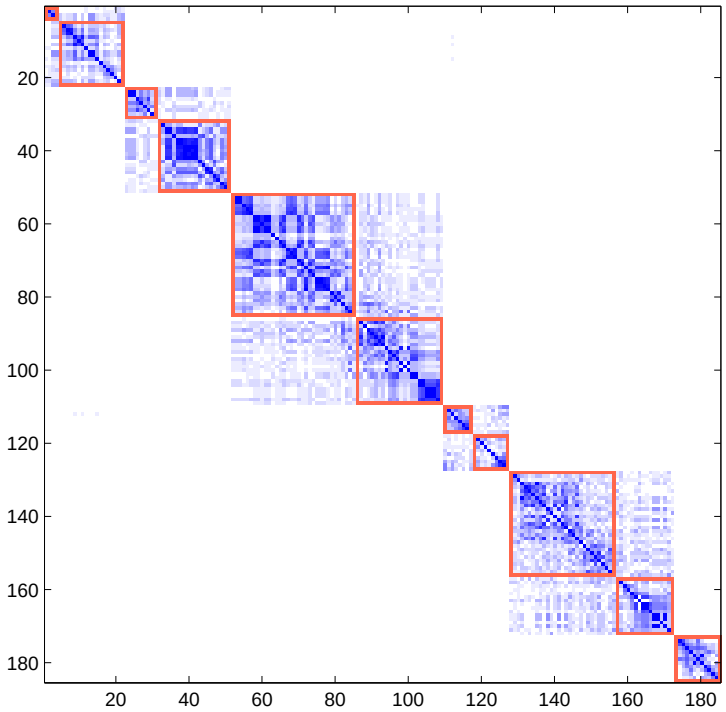
GWAS experiment

- We have analyzed the genetic data collected for 5402 individuals and 16427 genetic regions located in chromosome 1
- Columns of the experiment matrix were divided into groups by using the hierarchical clustering algorithm (HCA)
- As a results we achieved 1358 groups, with average group size close to 12
- We have performed 200 iterations for each sparsity level from the set $[1, 4, 8, 13, 18, 24]$, in each iteration we generated observations using linear regression model with $\sigma = 1$
- Tuning parameters were obtain by applying the conservative and the liberal strategies for target gFDR level 0.1
- Coefficients in β_{I_j} , for truly relevant group j, were generated such as $\|X_{I_j}\beta_{I_j}\|_2 = 8$
- Estimates were obtained by using the iterative version of gSLOPE with σ estimation

Defining groups by HCA

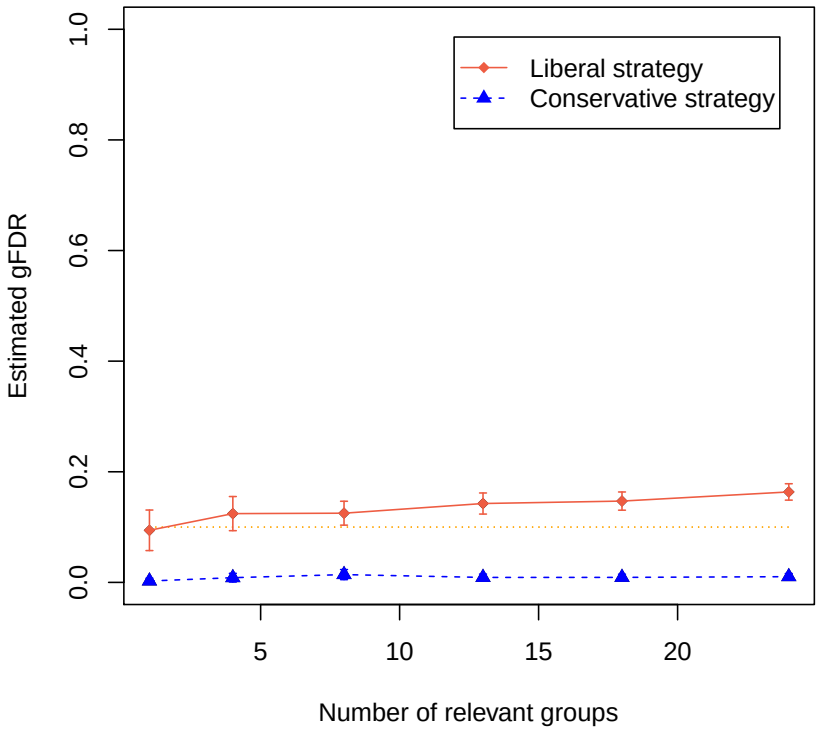


(a) Original

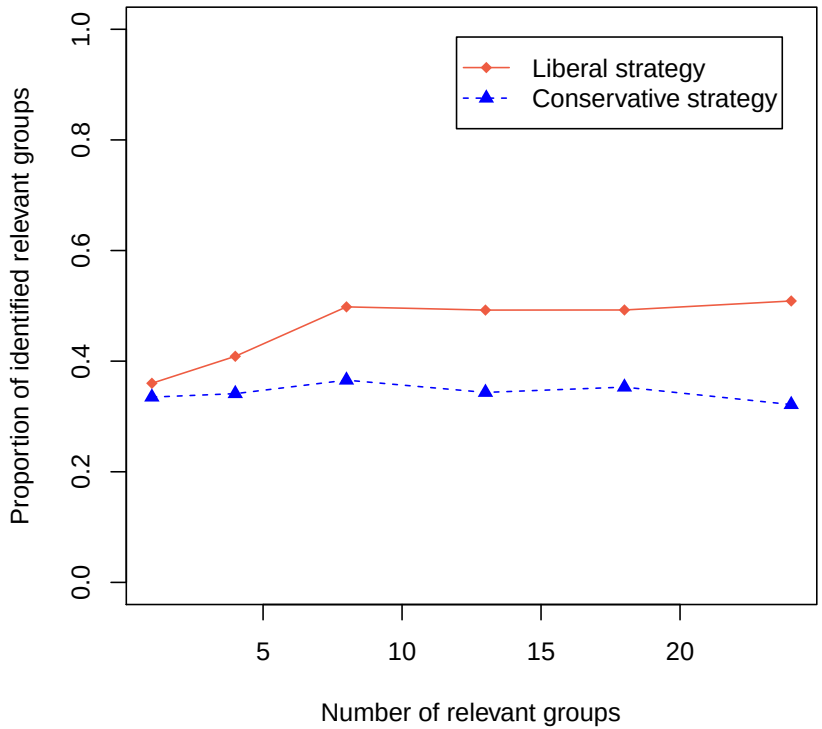


(b) After clustering with HCA

Results



(a) gFDR



(b) Power

[1] M. Bogdan, R. van den Berg, W. Su, E. J. Candes. Statistical Estimation and Testing via the Ordered ℓ_1 Norm. arXiv:1310.1969

[2] Yuan M., Lin Y. Consistent group selection in high-dimensional linear regression. *Bernoulli*, 16(4):1369-1384, 2007.

[3] Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshirani A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 2013.

[4] Gossmann, Alexej and Cao, Shaolong and Wang, Yu-Ping Identification of significant genetic variants via SLOPE, and its extension to Group SLOPE. *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, 2015

[5] Johnson S. C. Hierarchical clustering schemes. *Psychometrika*, Vol. 32, 241–254, 1967.

[6] Yuan M., Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, Series B, 68(1):49-67, 2007.

[7] By Peng Zhao, Guilherme Rocha, Bin Yu The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468-3497, 2009.