Genetic factors influencing the response to the therapy

Małgorzata Bogdan

Faculty of Pure and Applied Mathematics
Wroclaw University of Science and Technology

Vienna, 08/11/2016

Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Team

- Wroclaw University of Science and Technology (Poland): Damian Brzyski (Indiana University), Piotr Sobczyk, Piotr Szulc, Malgorzata Bogdan
- Stanford University (USA): Chiara Sabatti, Emmanuel Candès, Hua Tang, Ewout van den Berg (IBM Research Center), Weijie Su (University of Pennsylvania), Christine Peterson (University of Texas)
- Medical University of Vienna (Austria): Florian Frommlet, Franz König
- Tulane University (USA): Alexej Gossmann
- Ecole Polytechnique (France): Julie Josse

Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Schedule

- Why to consider genetic background ?

Schedule

- Why to consider genetic background ?
- Association Studies

Schedule

- Why to consider genetic background ?
- Association Studies
- Multiple Testing

Schedule

- Why to consider genetic background ?
- Association Studies
- Multiple Testing
- Model selection criteria

Schedule

- Why to consider genetic background ?
- Association Studies
- Multiple Testing
- Model selection criteria
- SLOPE (Sorted L-one penalized estimation)

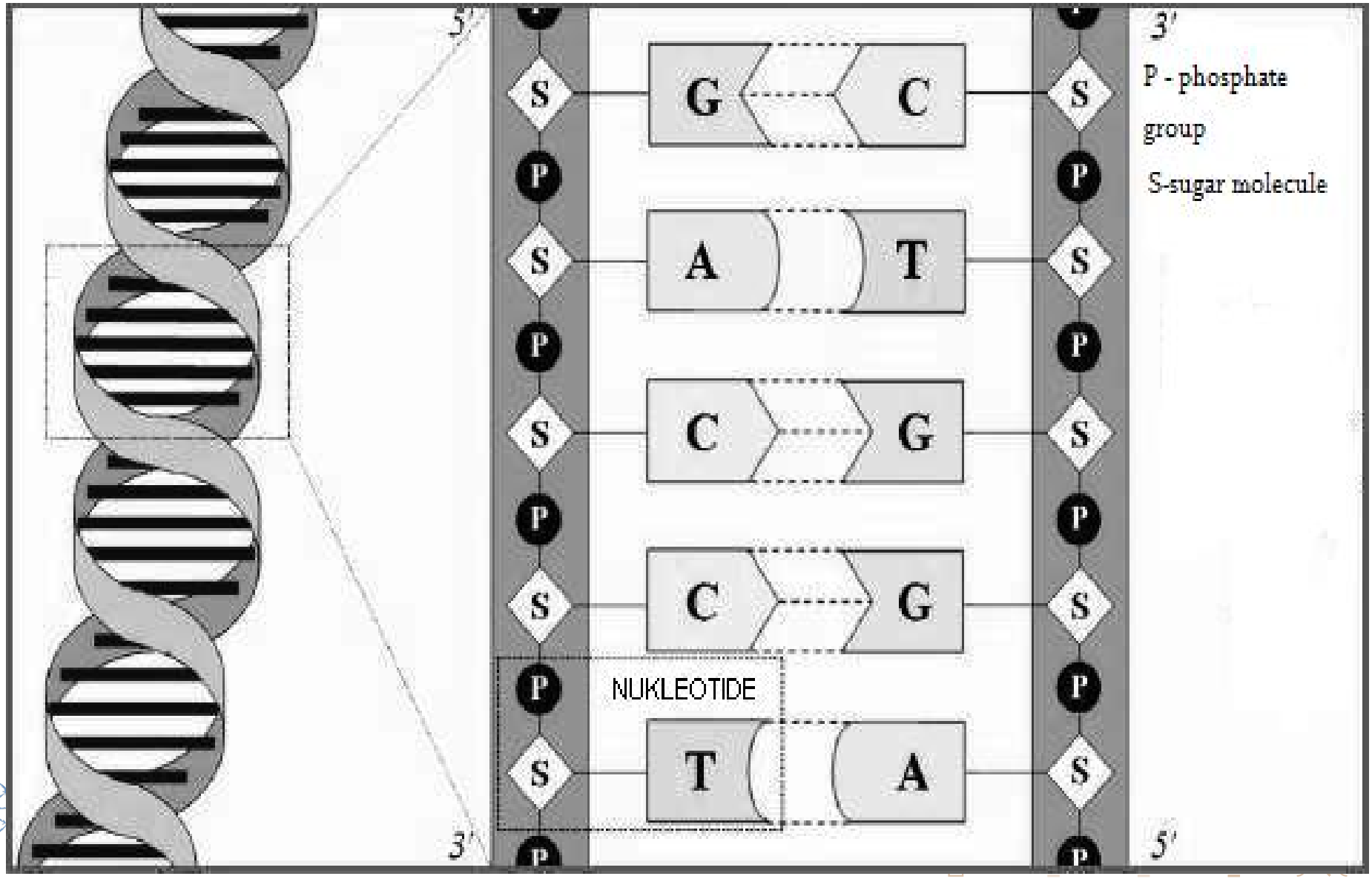
Schedule

- Why to consider genetic background ?
- Association Studies
- Multiple Testing
- Model selection criteria
- SLOPE (Sorted L-one penalized estimation)
- Gene expression data - SLOPE and subspace clustering

Schedule

- Why to consider genetic background ?
- Association Studies
- Multiple Testing
- Model selection criteria
- SLOPE (Sorted L-one penalized estimation)
- Gene expression data - SLOPE and subspace clustering
- Simulation study in the context of clinical trials

DNA Structure



Genetic variability

- About 99,9% of genetic information is the same for all people.
- A ***polymorphism*** is a difference in DNA structure, which is present in at least 1% of population
- A ***Single Nucleotide Polymorphism(SNP)*** is a polymorphism with the difference in the single base:
  - A typical SNP: a position in DNA in which
    - 85% of population has Cytosine(C)
    - 15% has a Thymine(T).
- There are usually two forms of a SNP at a given locus
- three genotypes : AA, Aa, aa.

Main purpose

MAIN PURPOSE: finding mutations in DNA sequence that influence a characteristic of interest.

Main purpose

MAIN PURPOSE: finding mutations in DNA sequence that influence a characteristic of interest.

Example - identification of genes that influence the patient's response to the treatment

1. Increasing the power of detection of treatment effects.
2. Identification of groups of patients for personalized therapies.
3. Larger chance that a medicine will pass clinical trials (at least in some groups of patients)

Main purpose

MAIN PURPOSE: finding mutations in DNA sequence that influence a characteristic of interest.

Example - identification of genes that influence the patient's response to the treatment

1. Increasing the power of detection of treatment effects.
2. Identification of groups of patients for personalized therapies.
3. Larger chance that a medicine will pass clinical trials (at least in some groups of patients)

Y - relevant quantitative characteristic

Main purpose

MAIN PURPOSE: finding mutations in DNA sequence that influence a characteristic of interest.

Example - identification of genes that influence the patient's response to the treatment

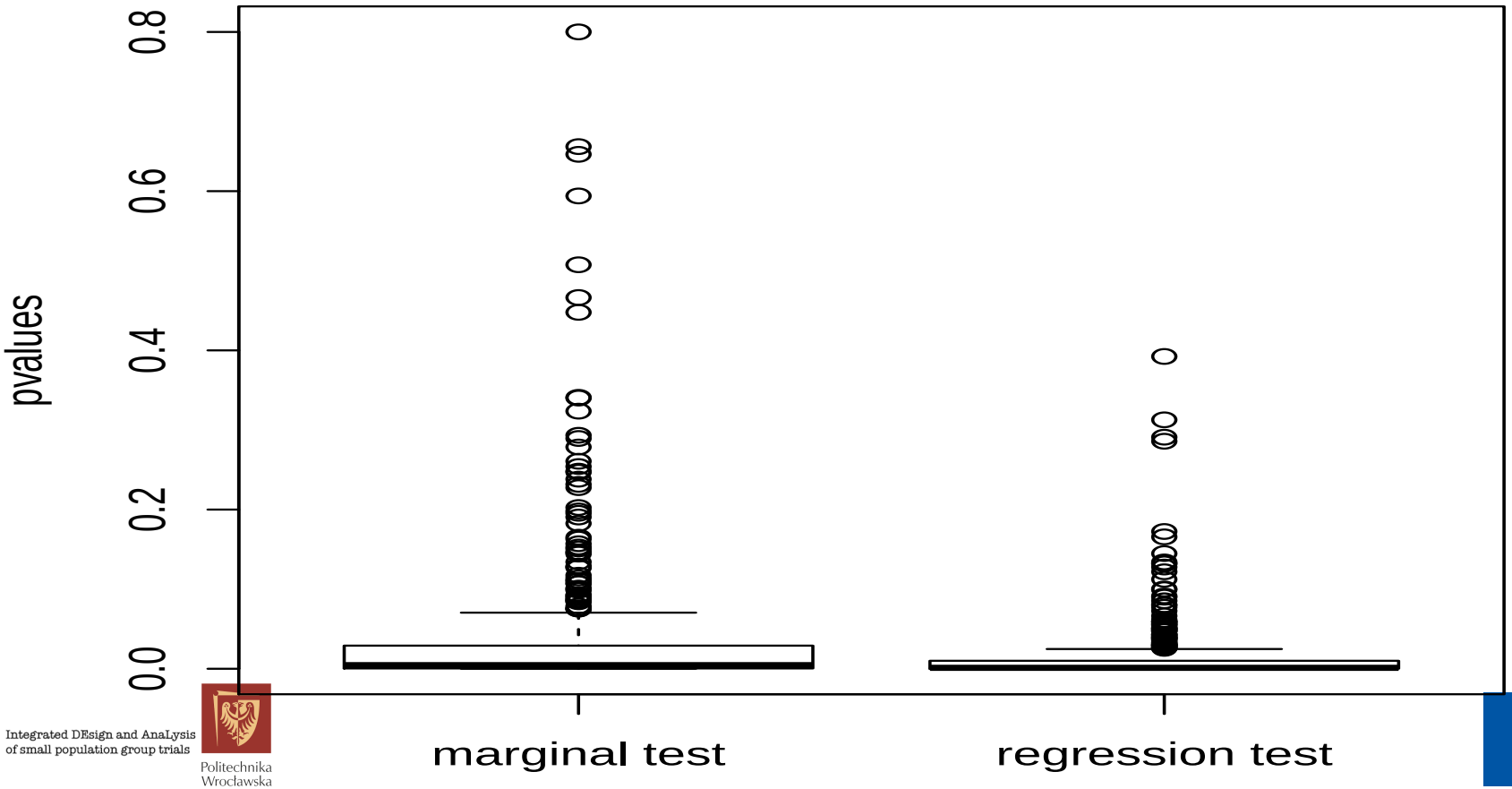
1. Increasing the power of detection of treatment effects.
2. Identification of groups of patients for personalized therapies.
3. Larger chance that a medicine will pass clinical trials (at least in some groups of patients)

Y - relevant quantitative characteristic

Examples: change in blood pressure, cholesterol level, etc.

Simulation example - Increasing the power of detection of treatment effects

20 influential genes, no gene-treatment interaction
gain in power 93% vs 83%



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Data structure

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

Data structure

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

$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - treatment indicators

Data structure

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

$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - treatment indicators

$G_{n \times m}$ - matrix of genotypes

Data structure

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

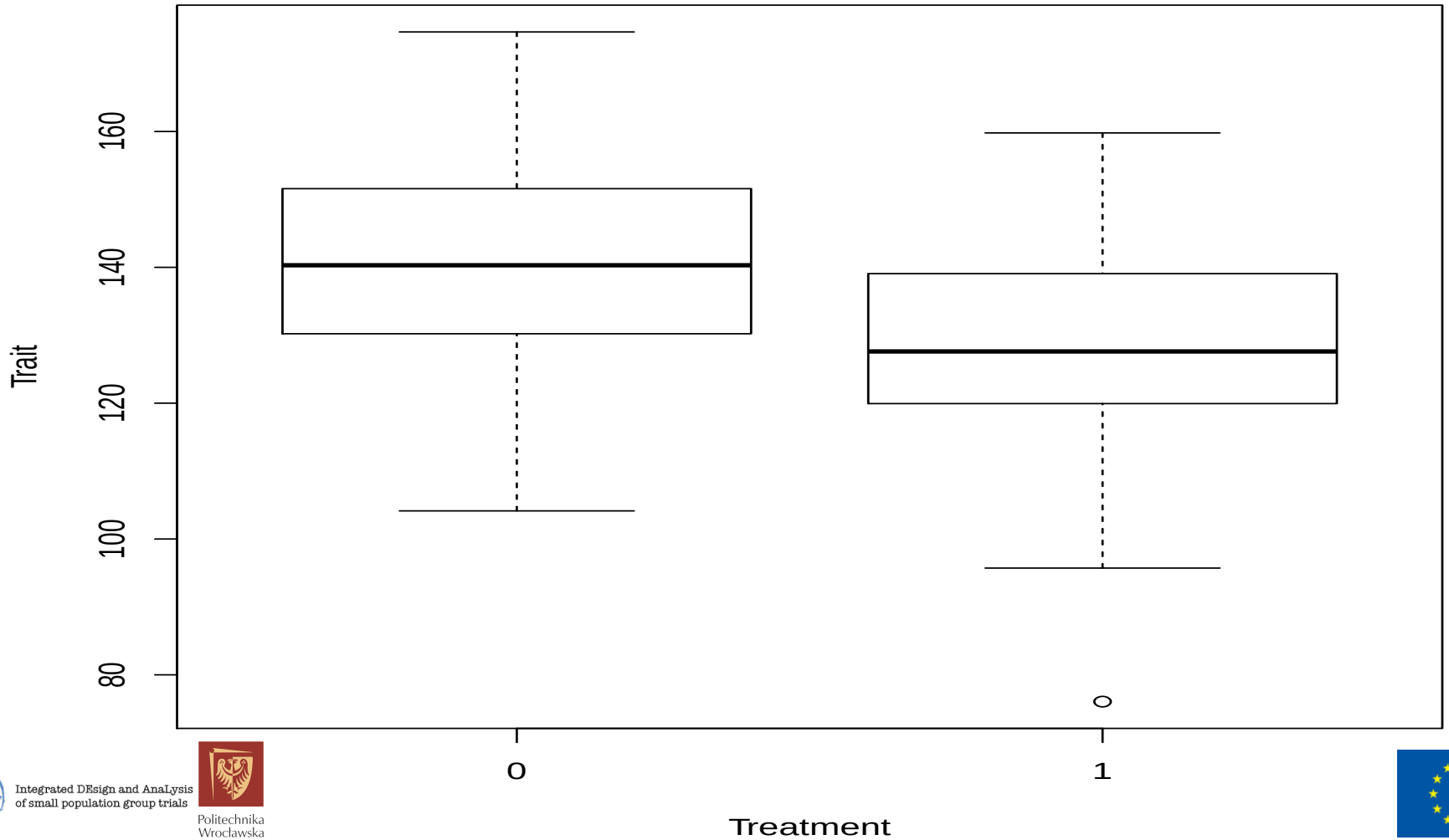
$T = (T_1, \ldots, T_n)^T$, $T_i \in \{0, 1\}$ - treatment indicators

$G_{n \times m}$ - matrix of genotypes

Usual coding

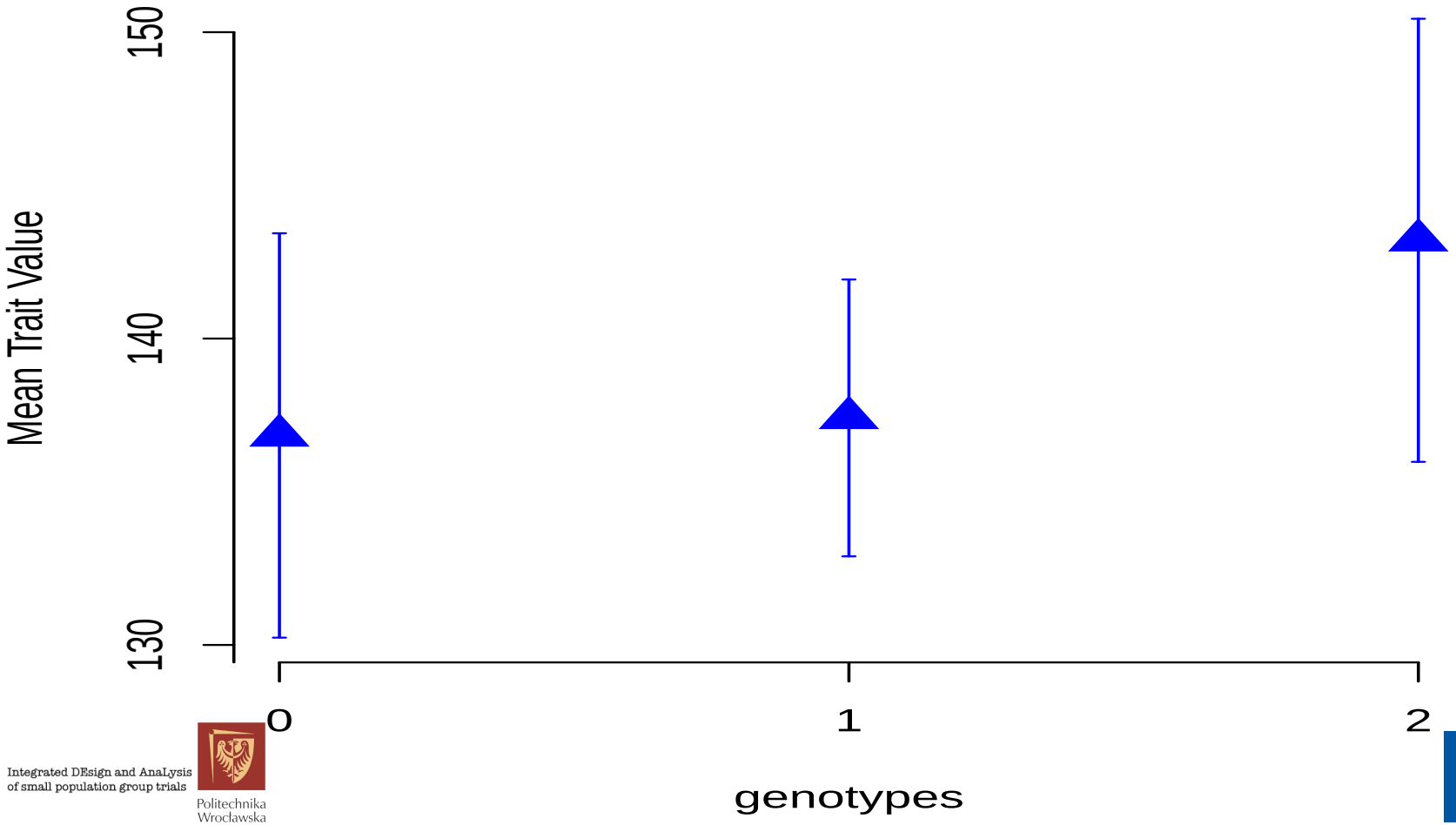
$$Z_{ij} = \begin{cases} 0 & \text{gdy} \quad G_{ij} = AA \\ 1 & \text{gdy} \quad G_{ij} = Aa \\ 2 & \text{gdy} \quad G_{ij} = aa \end{cases}$$

Simulation example - lowering blood pressure

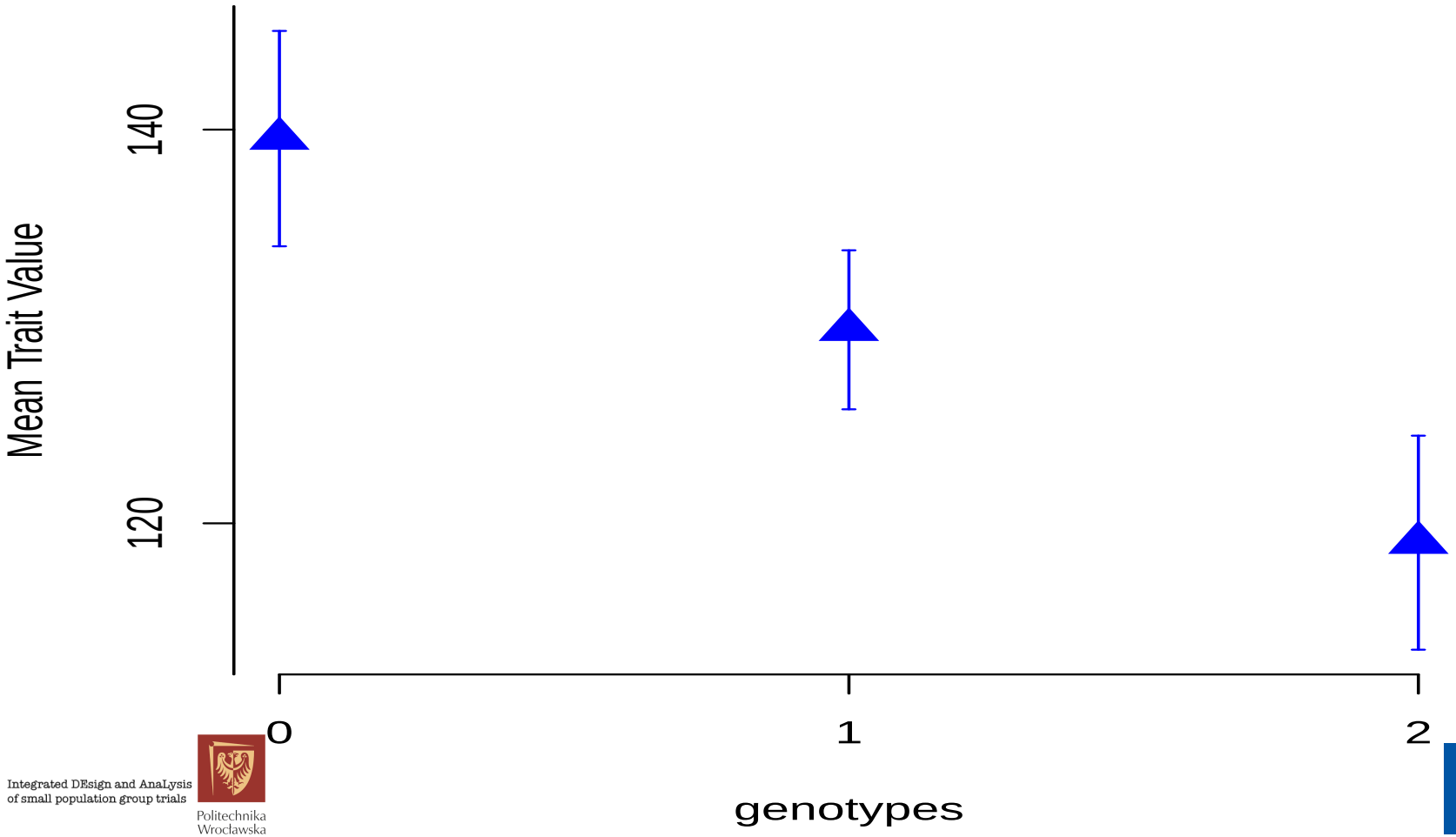


Integrated DEsign and AnaLysis of small population group trials
Politechnika Wrocławska

Control group



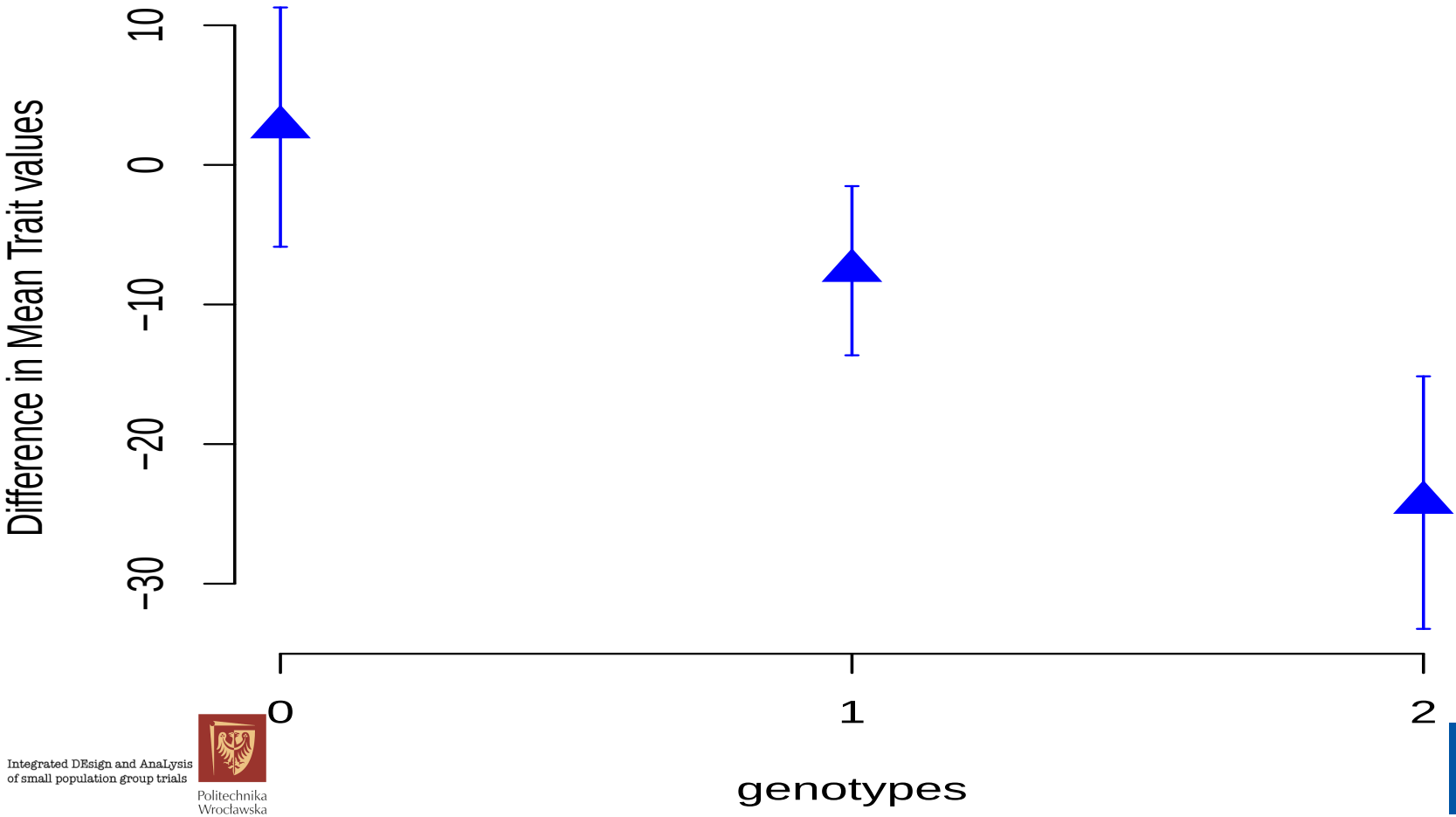
Treatment group



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

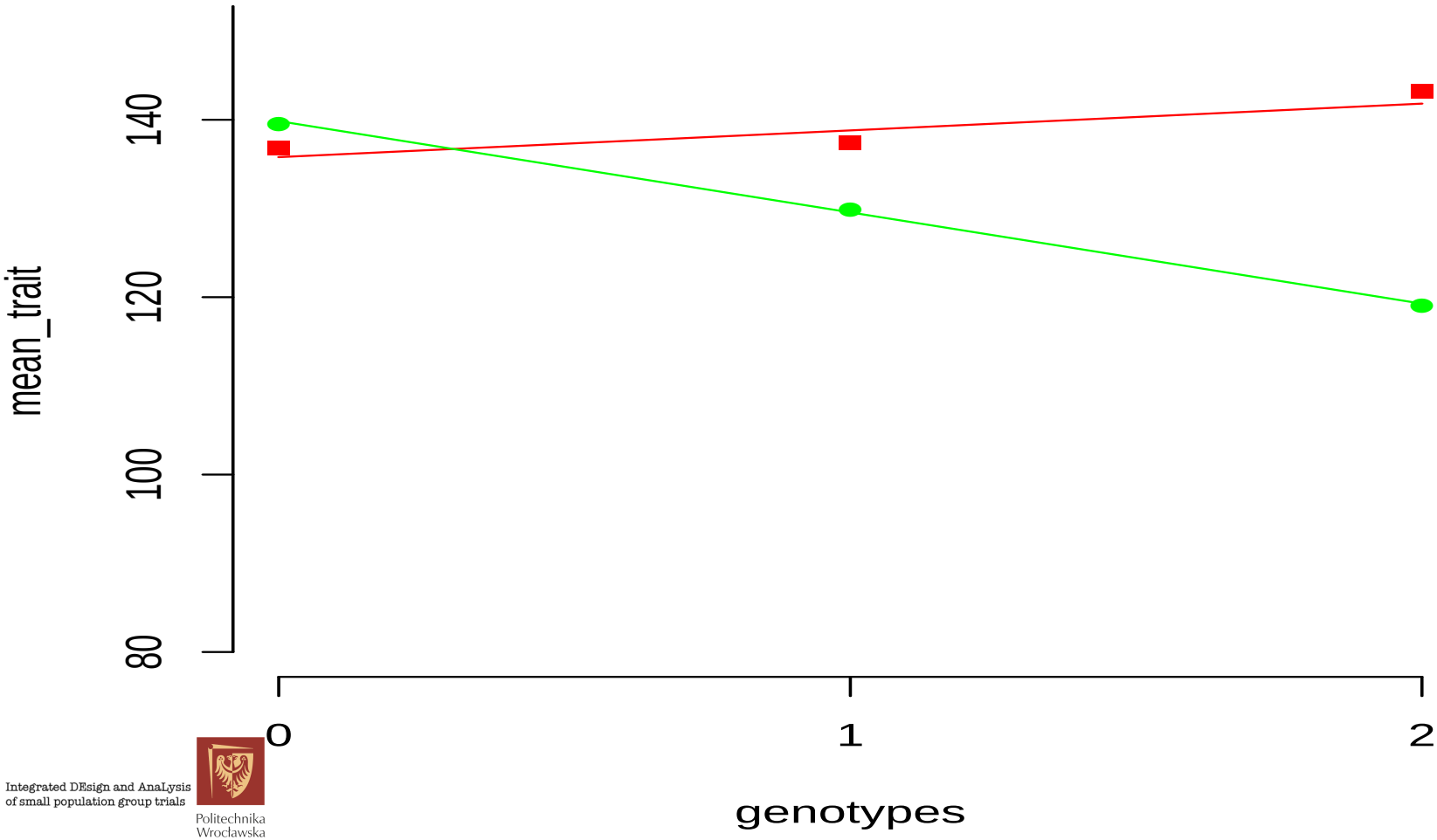
Treatment effect



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Gene-Treatment Interaction



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Statistical model

Identification of genetic background

$Y_i = \beta_0 + \sum_{j=1}^{m} \nu_j Z_{ij} \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

Statistical model

Identification of genetic background

$Y_i = \beta_0 + \sum_{j=1}^{m} \nu_j Z_{ij} \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

Genetic background and gene-treatment interactions

$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

Statistical model

Identification of genetic background

$Y_i = \beta_0 + \sum_{j=1}^{m} \nu_j Z_{ij} \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

Genetic background and gene-treatment interactions

$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

Statistical model

Identification of genetic background

$Y_i = \beta_0 + \sum_{j=1}^{m} \nu_j Z_{ij} \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

Genetic background and gene-treatment interactions

$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

$$p = 2m + 2, \quad X = [1|T|Z|ZT] \quad , \beta = [\beta_0, \beta_1, \nu, \gamma]^T$$

Identification of genetic background

$Y_i = \beta_0 + \sum_{j=1}^{m} \nu_j Z_{ij} \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

Genetic background and gene-treatment interactions

$Y_i = \beta_0 + \beta_1 T_i + \sum_{j=1}^{m} \nu_j Z_{ij} + \sum_{j=1}^{m} \gamma_j Z_{ij} T_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

$$p = 2m + 2, \quad X = [1|T|Z|ZT] \quad , \beta = [\beta_0, \beta_1, \nu, \gamma]^T$$

$$R(Z) = E(Y|T = 1, Z) - E(Y|T = 0, Z) = \beta_1 + \sum_{j=1}^{m} \gamma_j Z_{ij}$$

Simple linear regression

$$Y_i = \beta_0 + \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

Simple linear regression

$$Y_i = \beta_0 + \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad .$$

$\hat{\beta}_j$: least squares estimate β_j

$$\hat{\beta}_j \sim N(\beta_j, \sigma_j^2)$$

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

$$\text{Significance level: } \alpha = P_{H_{0j}}(|z_j| > c)$$

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

$$\text{Significance level: } \alpha = P_{H_{0j}}(|z_j| > c)$$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

$$\text{Significance level: } \alpha = P_{H_{0j}}(|z_j| > c)$$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

$$\text{Significance level: } \alpha = P_{H_{0j}}(|z_j| > c)$$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

Multiple testing (1)

$$H_{0j} : \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

$$\text{Significance level: } \alpha = P_{H_{0j}}(|z_j| > c)$$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

Multiple testing (1)

$$H_{0j}: \beta_j = 0 \quad \text{vs} \quad \beta_j \neq 0$$

$$\text{Reject } H_{0j} \text{ when } z_j = \frac{|\hat{\beta}_j|}{\sigma_j} > c$$

$$\text{Significance level: } \alpha = P_{H_{0j}}(|z_j| > c)$$

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

$$\alpha = 0.05, p_0 = 2000 \rightarrow E(V) = 100$$

Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Multiple testing procedures

Bonferroni correction (FWER control): Apply significance level $\frac{\alpha}{p}$.

Multiple testing procedures

Bonferroni correction (FWER control): Apply significance level $\frac{\alpha}{p}$.

Reject H_{0j} when $|z_j| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2 \log p}(1 + o_p)$

Bonferroni correction (FWER control): Apply significance level $\frac{\alpha}{p}$.

Reject H_{0j} when $|z_j| \geq \Phi^{-1}\left(1-\frac{\alpha}{2p}\right) = \sqrt{2\log p}(1+o_p)$

Benjamini-Hochberg procedure (FDR control)

(1) Sort p-values: $|p|_{(1)} \leq |p|_{(2)} \leq \ldots \leq |p|_{(p)}$

(2) Identify the largest j such that

$$|p|_{(j)} \leq \alpha \frac{j}{p}, \tag{1}$$

Call this index j_{SU}.

(3) Reject $H_{(j)}$ if and only if $j \leq j_{\mathsf{SU}}$

Simulation study (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Sample POPRES of real genomes from dbGaP

- 309790 SNPs for 649 individuals of European ancestry

Simulation study (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Sample POPRES of real genomes from dbGaP

- 309790 SNPs for 649 individuals of European ancestry
- k = 40 independent (unlinked) causal mutations

Simulation study (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Sample POPRES of real genomes from dbGaP

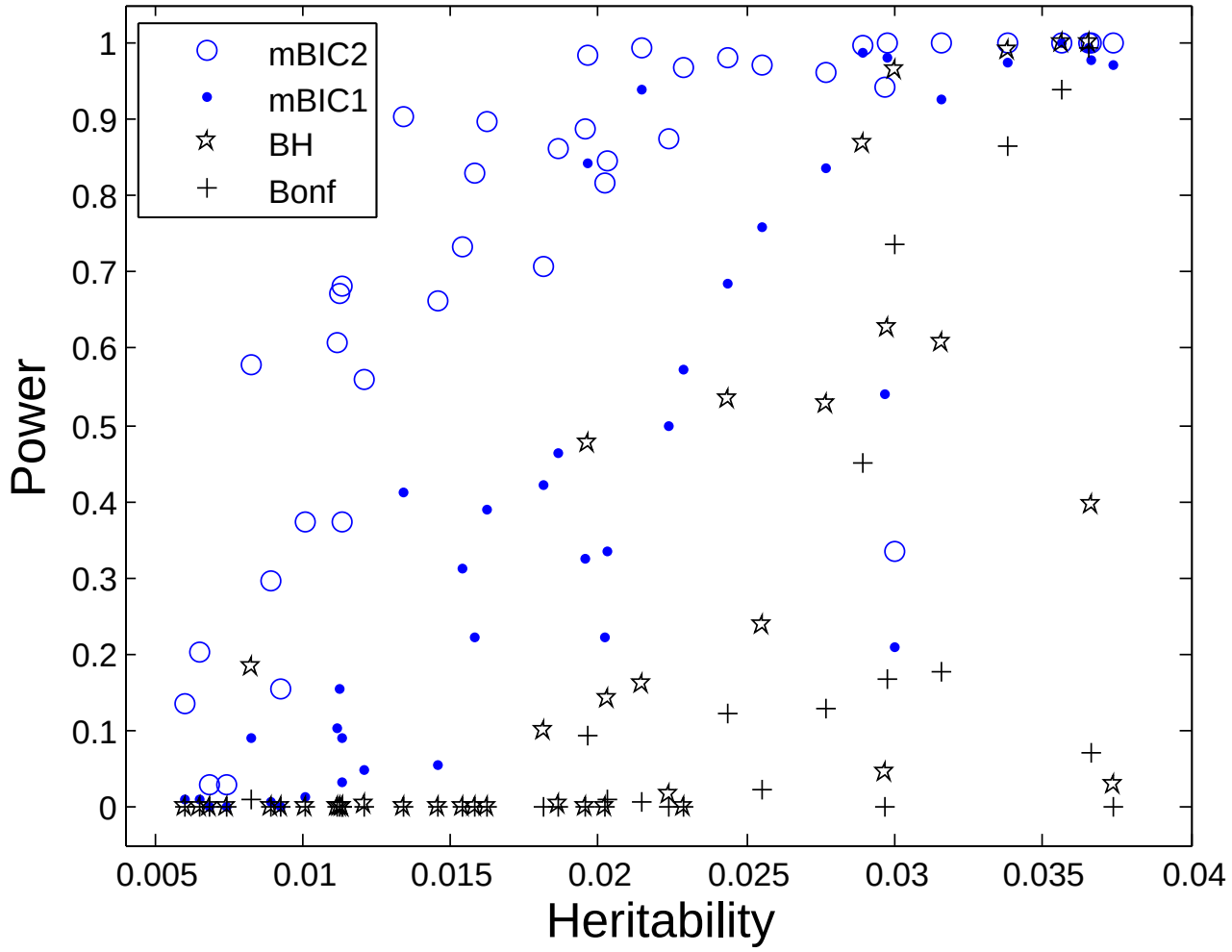
- 309790 SNPs for 649 individuals of European ancestry
- k = 40 independent (unlinked) causal mutations
- 1000 replications from the additive model M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$

Simulation study (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Sample POPRES of real genomes from dbGaP

- 309790 SNPs for 649 individuals of European ancestry
- k = 40 independent (unlinked) causal mutations
- 1000 replications from the additive model M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$
- β_j uniformly distributed over the interval [0.27, 0.66]

Power



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Problem with multiple testing

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

Problem with multiple testing

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

Problem with multiple testing

$$\hat{\beta}_X \approx \frac{\hat{C}ov(Y,X)}{\hat{V}arX}$$

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$$

$$\hat{C}ov(Y, X_1) = \beta_1 \hat{V}arX_1 + \sum_{i=2}^{k} \beta_i \hat{C}ov(X_1, X_i) + \hat{C}ov(X_1, \epsilon)$$

Model selection criteria (1)

Goal: Estimation of β in model

$$Y = X_{n\times p}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n}), \quad p >> n$$

Model selection criteria (1)

Goal: Estimation of β in model

$$Y = X_{n\times p}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n}), \quad p >> n$$

It can be done under the assumption that $||\beta||_0 = k << n$ (sparsity assumption)

Model selection criteria (1)

Goal: Estimation of β in model

$$Y = X_{n\times p}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n}), \quad p >> n$$

It can be done under the assumption that $||\beta||_0 = k << n$ (sparsity assumption)

Model selection criteria: minimize $||Y - X\beta||^2 + pen(k)$

Model selection criteria (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ incur many false discoveries when p is large

Model selection criteria (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ incur many false discoveries when p is large

Risk Inflation Criterion [RIC, Foster and George (1994)]

$pen(k) = 2\sigma^2 k \log p$ - "Bonferroni correction"

Model selection criteria (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ incur many false discoveries when p is large

Risk Inflation Criterion [RIC, Foster and George (1994)]

$pen(k) = 2\sigma^2 k \log p$ - "Bonferroni correction"

BHRIC (Abramovich et al. (2006), Foster and Stine (1999), Birge and Massart (2001))

$pen(k) = 2\sigma^2 \sum_{i=1}^{k} \log(p/i) = 2\sigma^2(k \log p - log(k!))$ - "BH correction"

Model selection criteria (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ incur many false discoveries when p is large

Risk Inflation Criterion [RIC, Foster and George (1994)]

$pen(k) = 2\sigma^2 k \log p$ - "Bonferroni correction"

BHRIC (Abramovich et al. (2006), Foster and Stine (1999), Birge and Massart (2001))

$pen(k) = 2\sigma^2 \sum_{i=1}^{k} \log(p/i) = 2\sigma^2(k \log p - log(k!))$ - "BH correction"

Bogdan et al. (Genetics, 2004), Żak-Szatkowska and Bogdan (CSDA,2011), Frommlet et al. (CSDA, 2012)

combining BIC and RIC and BHRIC penalty

Admixtures (1)

Problem in GWAS - loss of power due to multiple testing, large sample sizes needed

Admixtures (1)

Problem in GWAS - loss of power due to multiple testing, large sample sizes needed

Solution available in admixed populations - use information on ancestry

Admixtures (1)

Problem in GWAS - loss of power due to multiple testing, large sample sizes needed

Solution available in admixed populations - use information on ancestry

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", biorxiv, doi: http://dx.doi.org/10.1101/062554, 2016.

Admixtures (1)

Problem in GWAS - loss of power due to multiple testing, large sample sizes needed

Solution available in admixed populations - use information on ancestry

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", biorxiv, doi: http://dx.doi.org/10.1101/062554, 2016.

P. Szulc, R package *bigstep* available on *CRAN*, can handle data which do not fit into RAM

Admixtures (1)

Problem in GWAS - loss of power due to multiple testing, large sample sizes needed

Solution available in admixed populations - use information on ancestry

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", biorxiv, doi: http://dx.doi.org/10.1101/062554, 2016.

P. Szulc, R package *bigstep* available on *CRAN*, can handle data which do not fit into RAM

Strong correlation - reduce the effective number of tests by a factor of 100

Admixtures (1)

Problem in GWAS - loss of power due to multiple testing, large sample sizes needed

Solution available in admixed populations - use information on ancestry

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", biorxiv, doi: http://dx.doi.org/10.1101/062554, 2016.

P. Szulc, R package *bigstep* available on *CRAN*, can handle data which do not fit into RAM

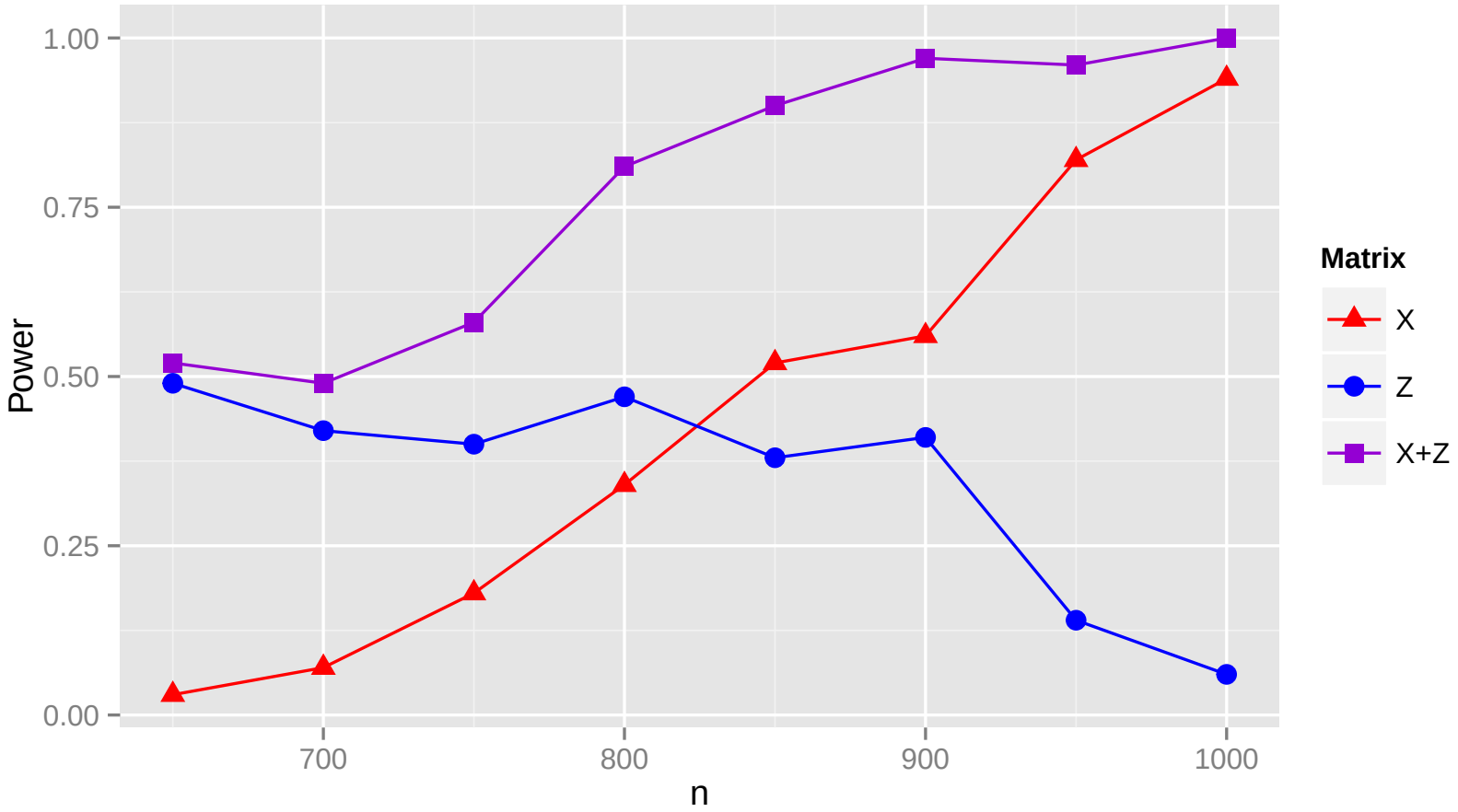
Strong correlation - reduce the effective number of tests by a factor of 100

$$\begin{aligned} \text{mBIC2}(X_M, Z_A) &= n \log \text{RSS} + (k_1 + k_2) \log n + 2k_1 \log(p/4) \\ &\quad + 2k_2 \log(p^{eff}/4) - 2\log(k_1!) - 2\log(k_2!) \end{aligned}$$

Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Admixtures (3)



SLOPE (1)

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. J. Candès, "SLOPE – Adaptive Variable Selection via Convex Optimization", *Annals of Applied Statistics*, **9** (3), 1103–1140, 2015.

SLOPE estimate is given by

$$\min_{b\in\mathbb{R}^p} \quad \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$

SLOPE (1)

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. J. Candès, "SLOPE – Adaptive Variable Selection via Convex Optimization", *Annals of Applied Statistics*, **9** (3), 1103–1140, 2015.

SLOPE estimate is given by

$$\min_{b\in\mathbb{R}^p} \quad \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$

$\lambda_1 = \ldots = \lambda_p = \sqrt{2\log p}$ - LASSO Bonferroni corrected

SLOPE (1)

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. J. Candès, "SLOPE – Adaptive Variable Selection via Convex Optimization", *Annals of Applied Statistics*, **9** (3), 1103–1140, 2015.

SLOPE estimate is given by

$$\min_{b\in\mathbb{R}^p} \quad \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$

$\lambda_1 = \ldots = \lambda_p = \sqrt{2\log p}$ - LASSO Bonferroni corrected

SLOPE is the extension of LASSO in the direction of BH procedure, controls FDR for sparse signals and weakly correlated predictors

SLOPE (1)

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. J. Candès, "SLOPE – Adaptive Variable Selection via Convex Optimization", *Annals of Applied Statistics*, **9** (3), 1103–1140, 2015.

SLOPE estimate is given by

$$\min_{b\in\mathbb{R}^p} \quad \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$

$\lambda_1 = \ldots = \lambda_p = \sqrt{2\log p}$ - LASSO Bonferroni corrected

SLOPE is the extension of LASSO in the direction of BH procedure, controls FDR for sparse signals and weakly correlated predictors

R package *SLOPE* available on CRAN, by E. Patterson

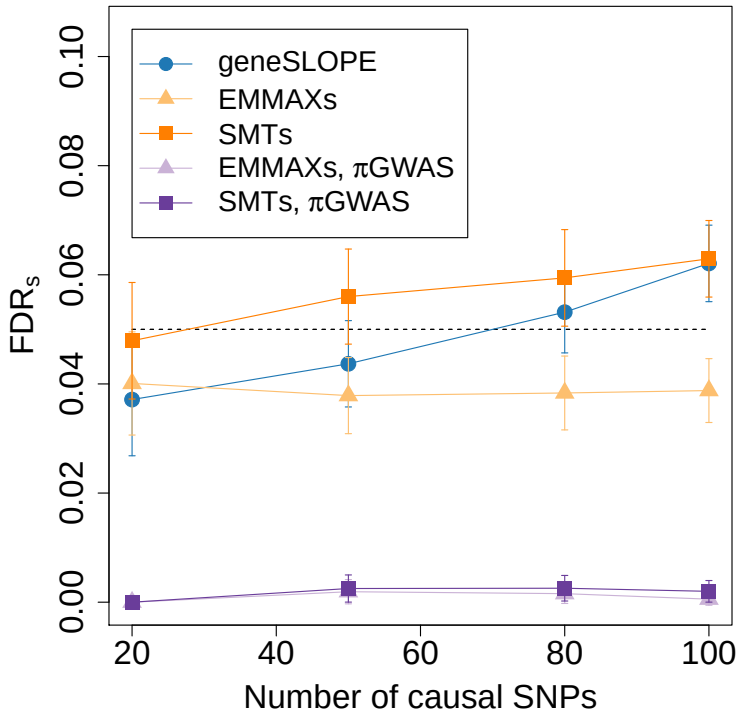
Application for real GWAS

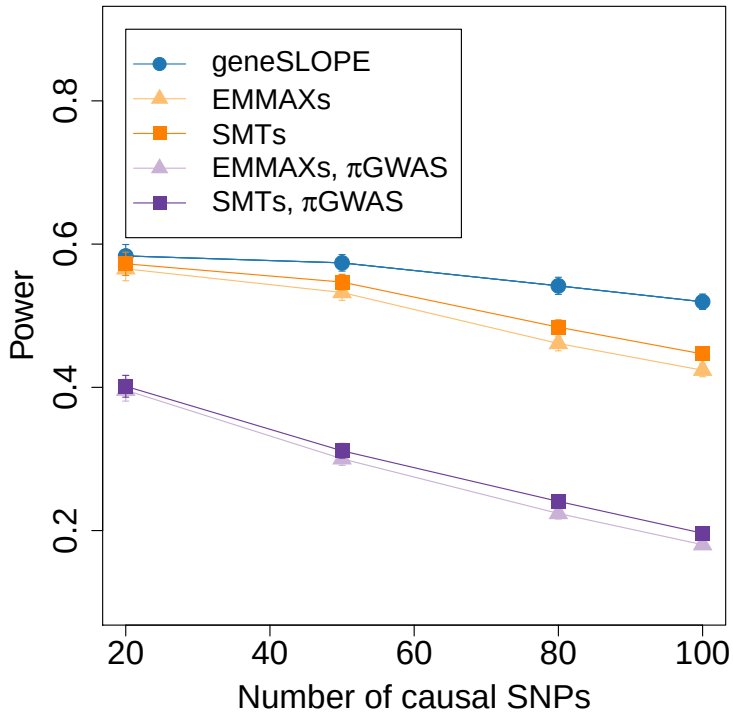
D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries", Genetics, 2016, doi: 10.1534/genetics.116.193987.

Application for real GWAS

D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries", Genetics, 2016, doi: 10.1534/genetics.116.193987.

R package *geneSLOPE* available on CRAN, by P. Sobczyk


geneSLOPE
EMMAXs
SMTs
EMMAXs, πGWAS
SMTs, πGWAS
FDR$_s$
0.00
0.02
0.04
0.06
0.08
0.10
20
40
60
80
100
Number of causal SNPs


geneSLOPE
EMMAXs
SMTs
EMMAXs, πGWAS
SMTs, πGWAS
Power
0.2
0.4
0.6
0.8
20
40
60
80
100
Number of causal SNPs

GroupSLOPE (1)

D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", arXiv: 1610.04960, 2016

GroupSLOPE (1)

D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", arXiv: 1610.04960, 2016

R package *grpSLOPE* available on CRAN by A. Gossmann

GroupSLOPE (1)

D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", arXiv: 1610.04960, 2016

R package *grpSLOPE* available on CRAN by A. Gossmann

A. Gossmann, S. Cao, D. Brzyski, L. Zhao, H. Deng, and Y. Wang, "A sparse regression method for group-wise feature selection with false discovery rate control", invited for *IEEE/ACM Transactions on Computational Biology and Bioinformatics*

Simulations

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

Simulations

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

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

Simulations

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

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

Scenario 2: modeling dominance

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

$$y = [X, Z][\beta'_X, \beta'_Z]' + \epsilon \ .$$

Simulations

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

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

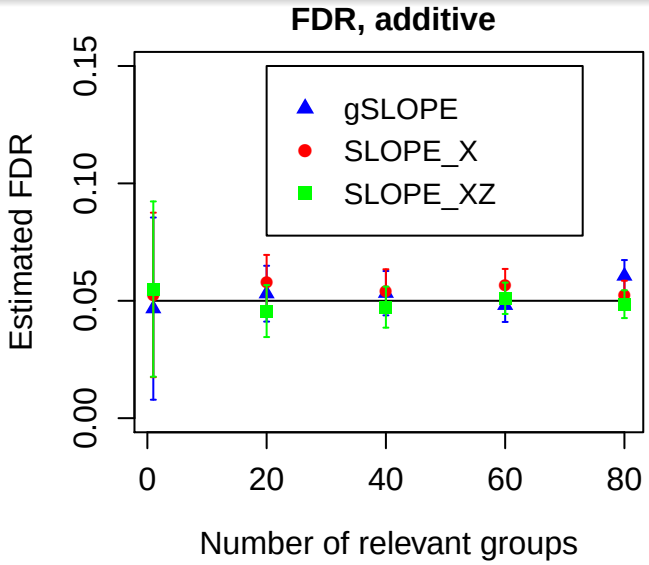
Scenario 2: modeling dominance

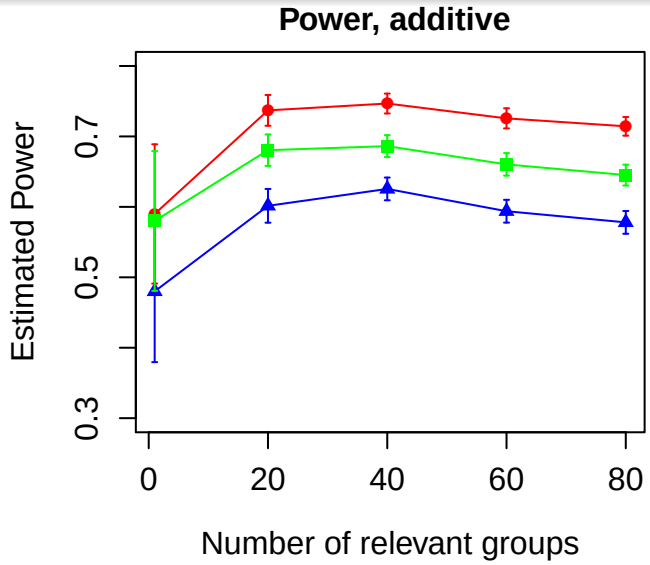
$$\tilde{z}_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases} , \tag{2}$$

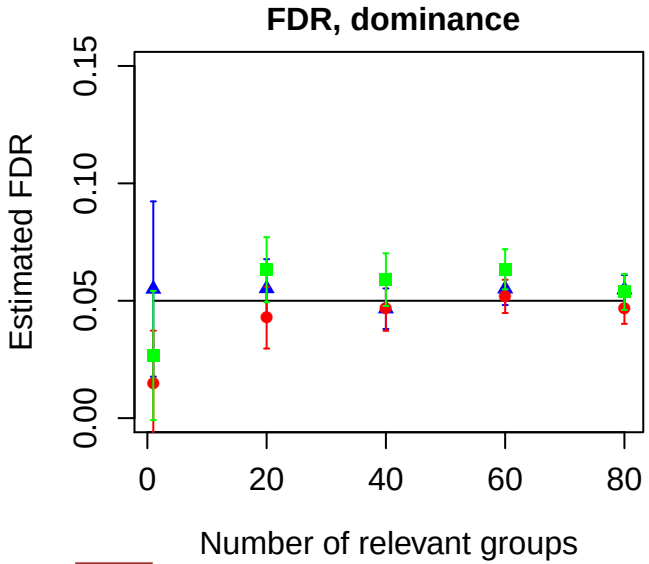
$$y = [X, Z][\beta'_X, \beta'_Z]' + \epsilon \ .$$

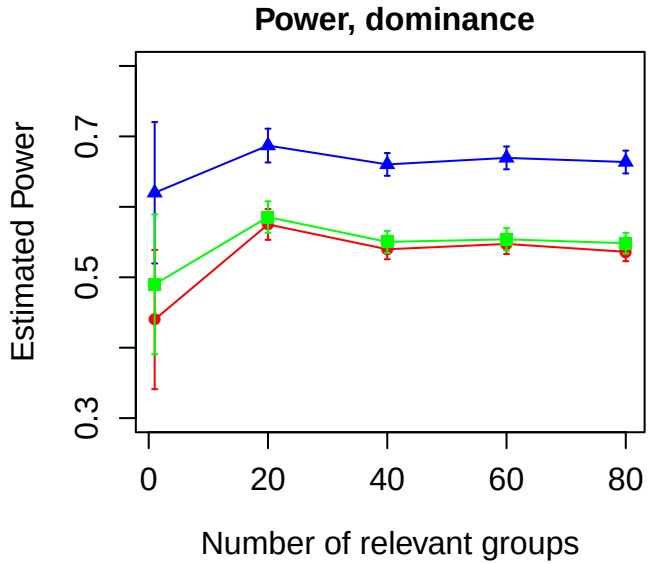
One group contains two columns: additive and dominance dummy variable for a given SNP

Simulation results







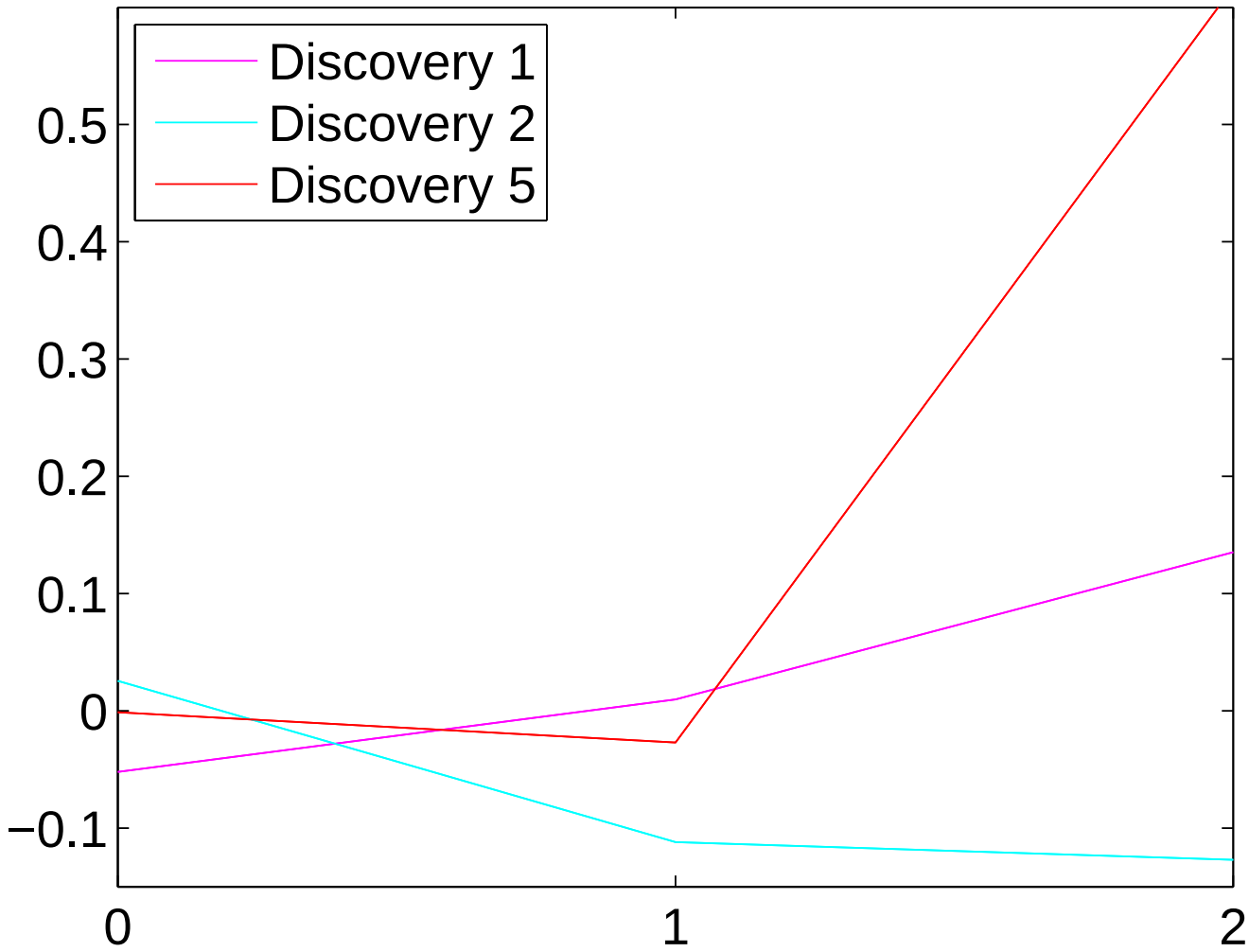


Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Genes Influencing Level of Triglicerides

5 new discoveries with group SLOPE - recessive rare genetic variants. Discovery 5 - 37 rare homozygotes



Personalized therapies (1)

P. Szulc, F.Frommlet, F. König, M. Bogdan, "Selecting predictive biomarkers from genomic data" - in preparation

Personalized therapies (1)

P. Szulc, F.Frommlet, F. König, M. Bogdan, "Selecting predictive biomarkers from genomic data" - in preparation

Identify patients with
$R(Z) = E(Y|T = 1, Z) - E(Y|T = 0, Z) > 0$

Personalized therapies (1)

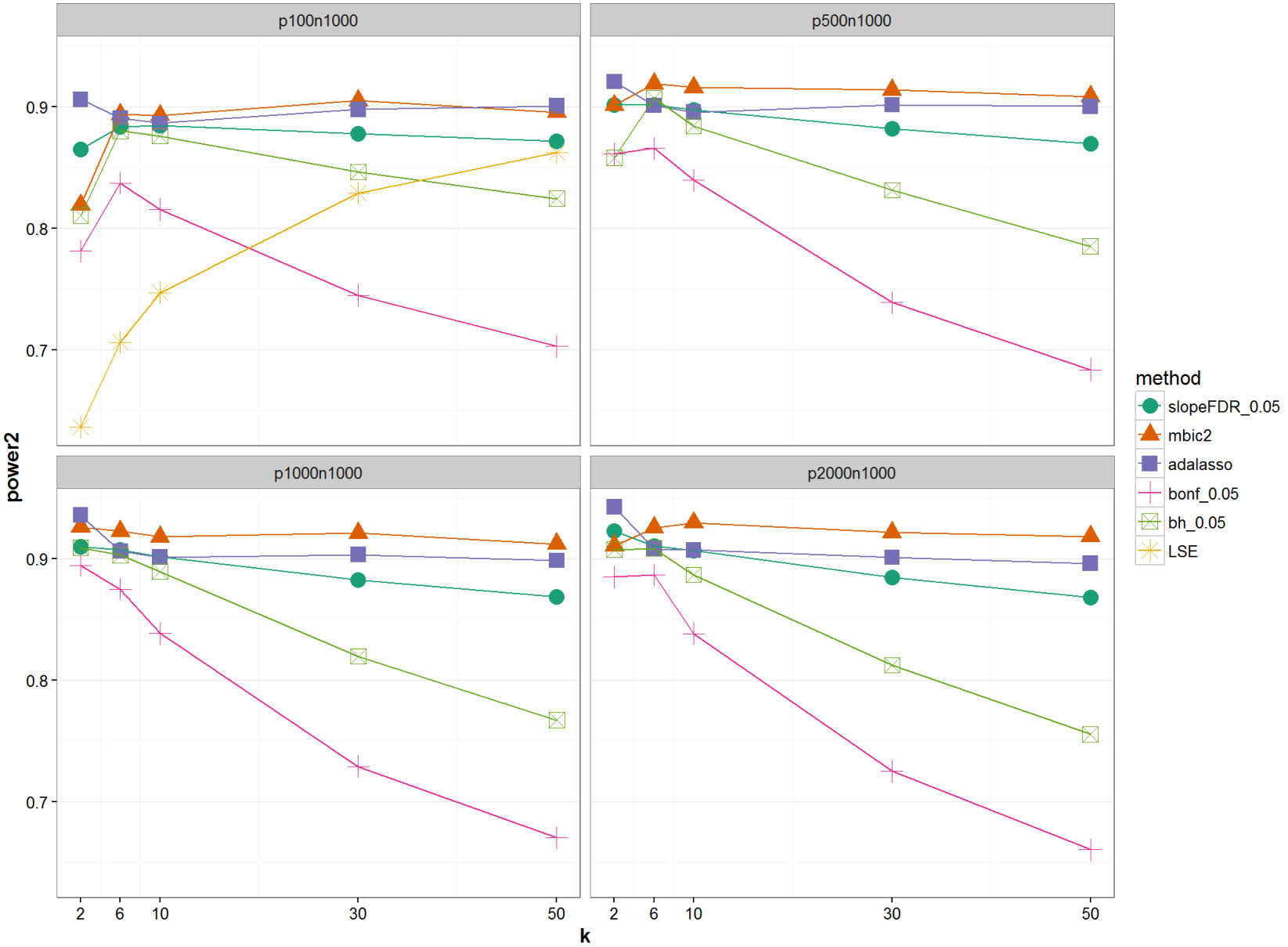
P. Szulc, F.Frommlet, F. König, M. Bogdan, "Selecting predictive biomarkers from genomic data" - in preparation

Identify patients with
$R(Z) = E(Y|T = 1, Z) - E(Y|T = 0, Z) > 0$

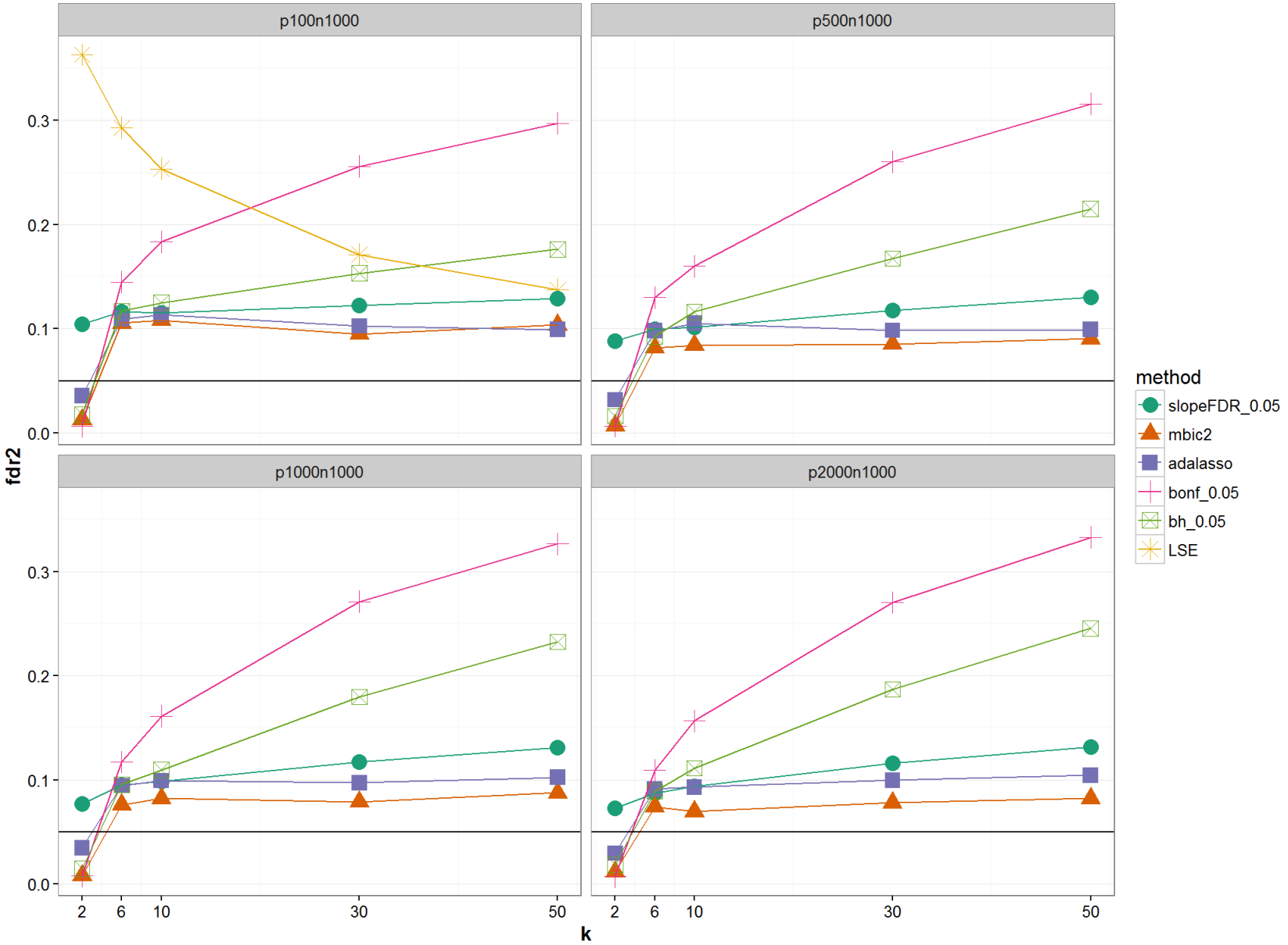
Two new tests:

- Test for the treatment effect within the selected model.
- Use part of the sample to identify the model and test for the treatment effect in selected patients from the second group.

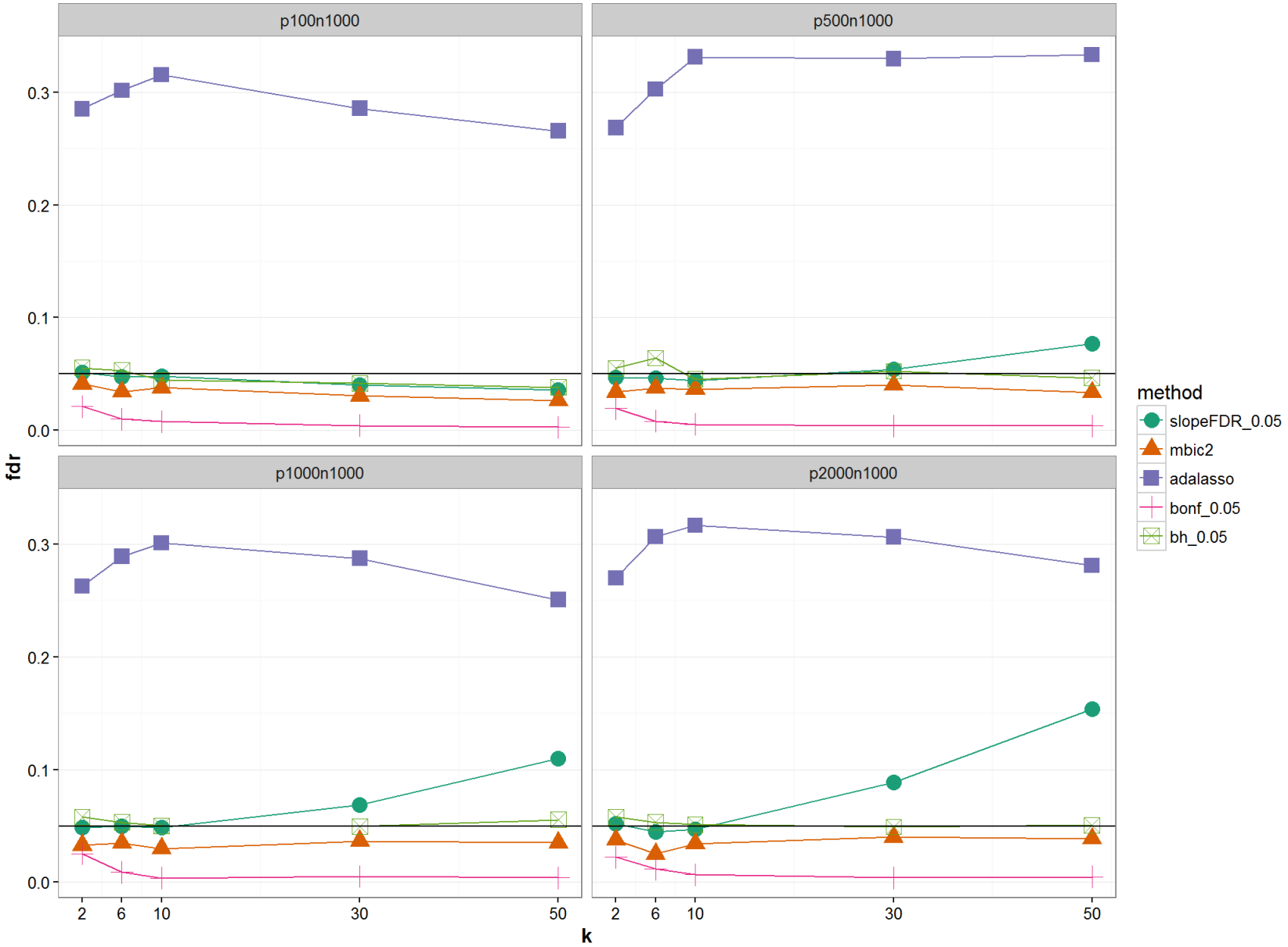
Power at the patient's level



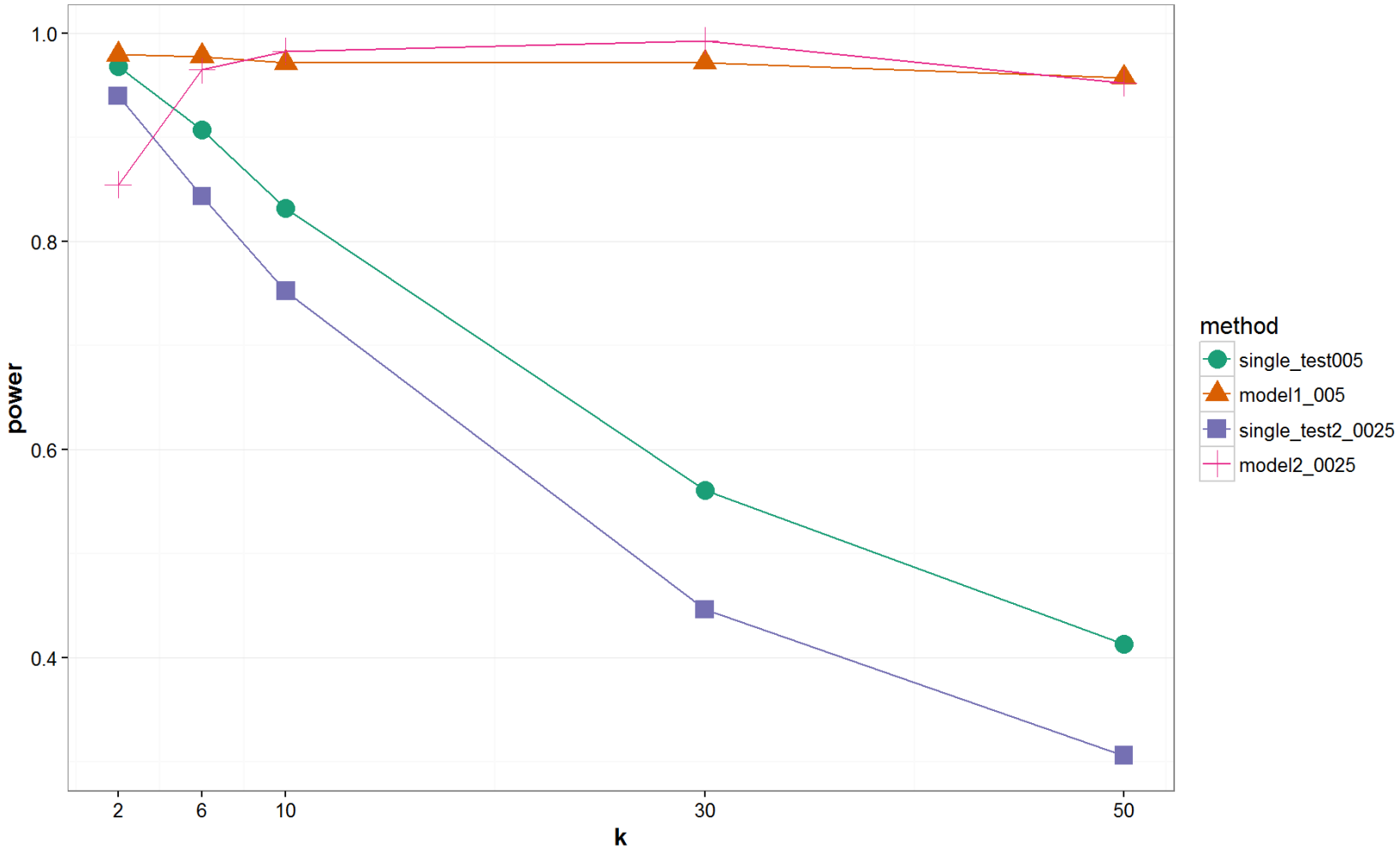
FDR at the patient's level



FDR at the gene level



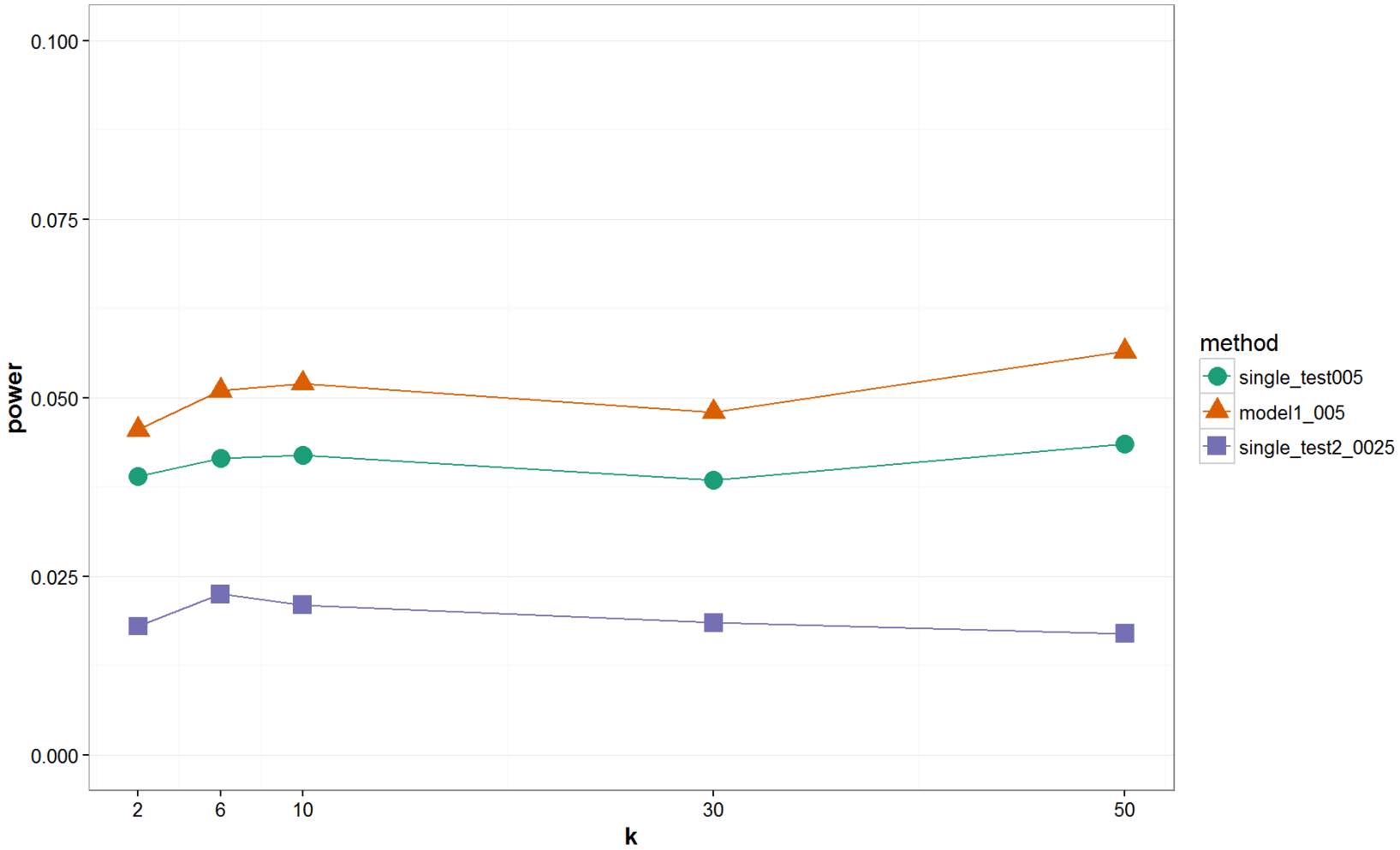
Power for testing the treatment effect



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Type I error



Integrated DEsign and AnaLysis of small population group trials

Politechnika Wrocławska

Transition curve for LASSO

W.Su, M. Bogdan, E.J. Candès, "False Discoveries Occur Early on the Lasso Path", to appear in Annals of Statistics, arxiv: 1511.01957

Transition curve for LASSO

W.Su, M. Bogdan, E.J. Candès, "False Discoveries Occur Early on the Lasso Path", to appear in Annals of Statistics, arxiv: 1511.01957

Solution (to optimally control FDR) combine SLOPE with knockoffs, under investigation

Transition curve for LASSO

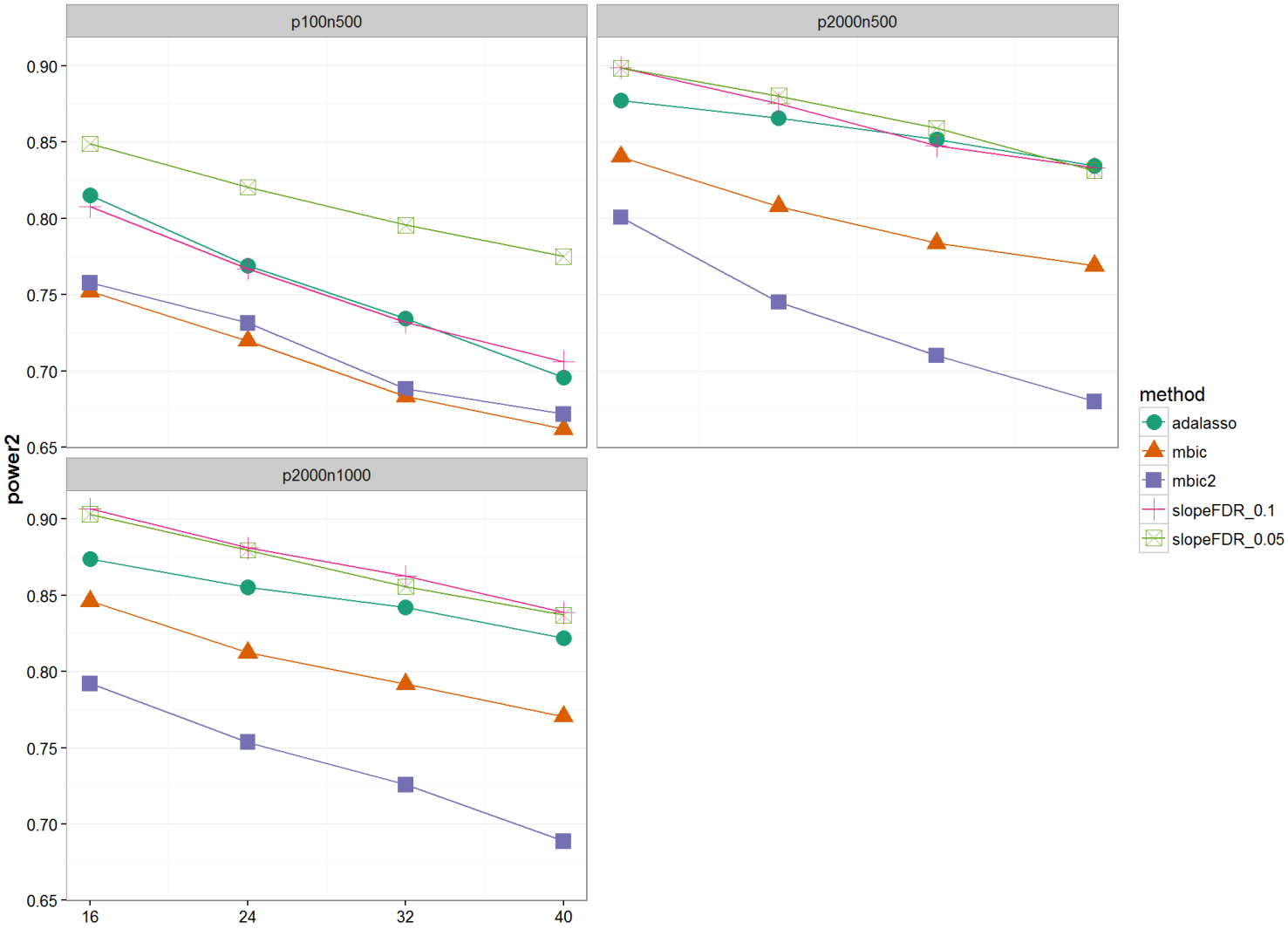
W.Su, M. Bogdan, E.J. Candès, "False Discoveries Occur Early on the Lasso Path", to appear in Annals of Statistics, arxiv: 1511.01957

Solution (to optimally control FDR) combine SLOPE with knockoffs, under investigation

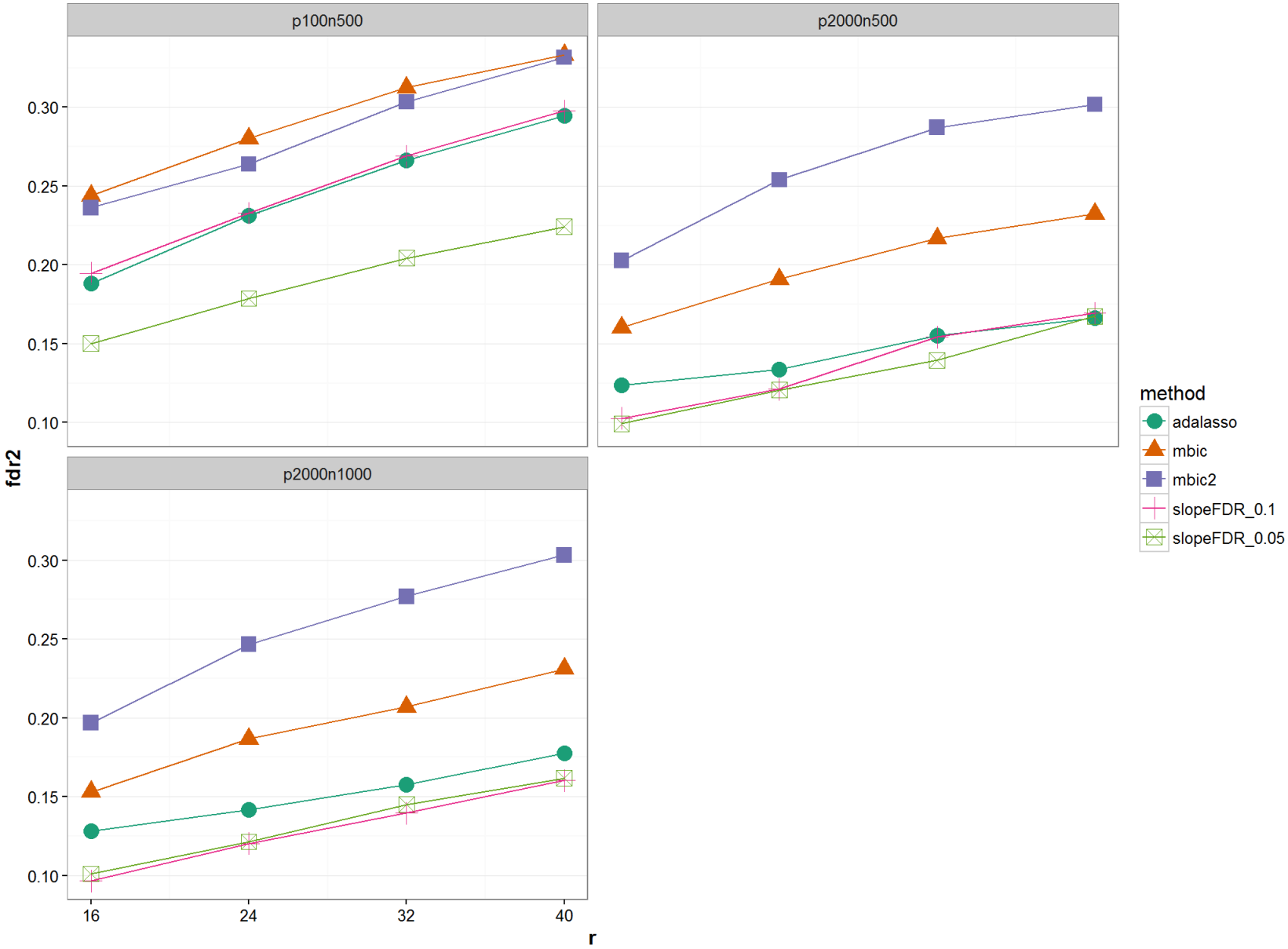
Other direction of current research - concentrate on prediction properties also in correlated designs e.g. gene expression data

Power at the patient's level - hidden factors (proxy for gene expressions)

$$X_{n\times p} = F_{n\times r}C_{r\times p} + \epsilon, \quad Y = F\beta + \epsilon$$



FDR at the patient's level



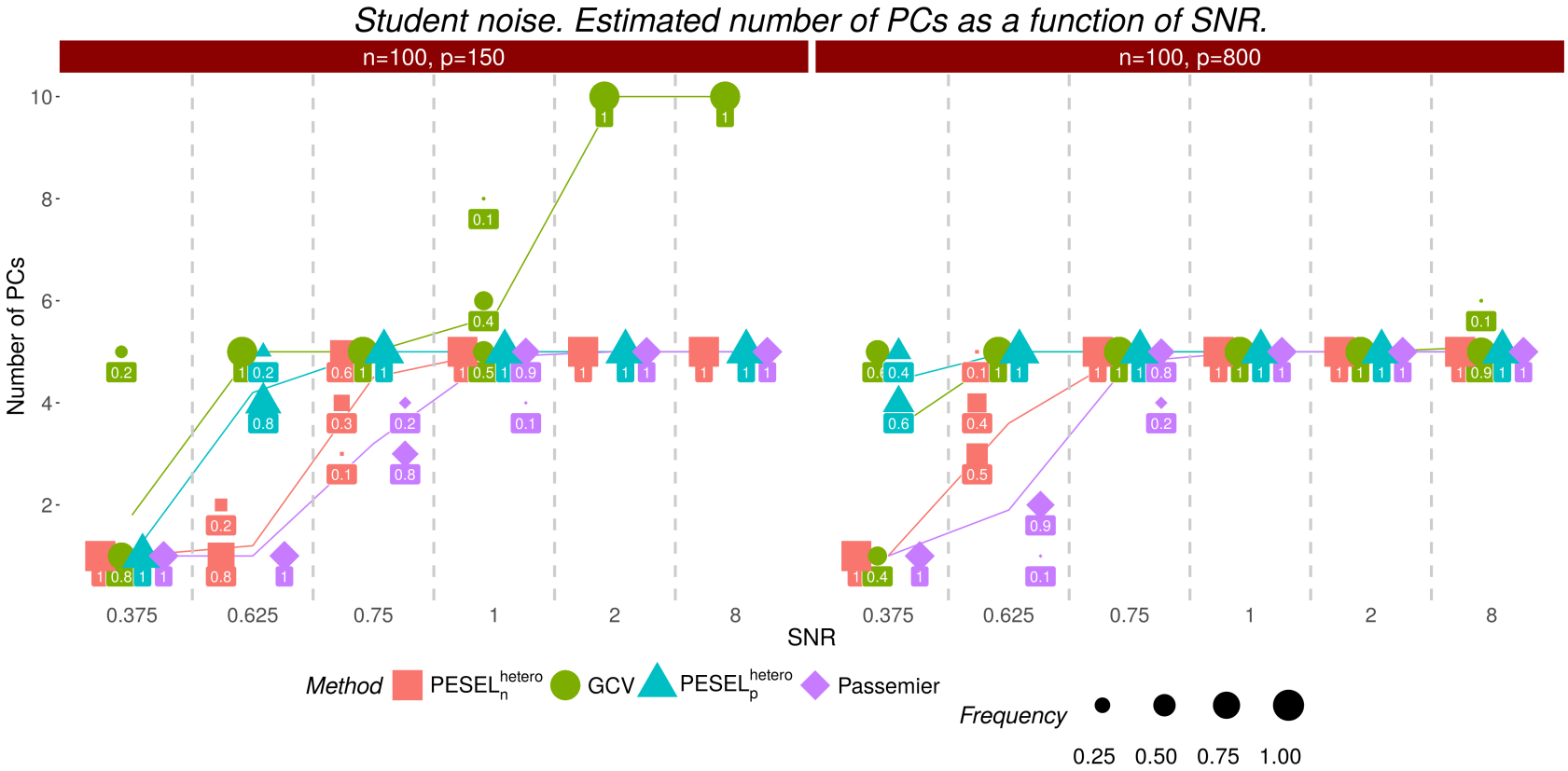
Analyzing Gene Expression Data (2)

R package *ClustofVar* available on *github* by P. Sobczyk - identification of genetic pathways by subspace clustering and modified BIC

Analyzing Gene Expression Data (2)

R package *ClustofVar* available on *github* by P. Sobczyk - identification of genetic pathways by subspace clustering and modified BIC

P. Sobczyk, M. Bogdan, J. Josse, "Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood", arxiv: 1606.05333, 2016


Student noise. Estimated number of PCs as a function of SNR.
n=100, p=150
n=100, p=800
Number of PCs
10
8
6
4
2
0.375
0.625
0.75
1
2
8
SNR
Method
PESEL$_n^{hetero}$
GCV
PESEL$_p^{hetero}$
Passemier
Frequency
0.25
0.50
0.75
1.00

Publications (1)

- M. Bogdan, E. van den Berg, C. Sabatti, W. Su, E. J. Candès, "SLOPE – Adaptive Variable Selection via Convex Optimization", *Annals of Applied Statistics*, **9** (3), 1103–1140, 2015.
- D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries", Genetics, 2016, available on journal web-page.
- D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", arXiv, 2016.
- 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 AISTATS2016, JMLR:W and CP* **vol.51**, 780–789, 2016.

Publications (2)

- W.Su, M. Bogdan, E.J. Candès, “False Discoveries Occur Early on the Lasso Path”, to appear in Annals of Statistics, arxiv: 1511.01957, 2015.
- P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", biorxiv, doi: http://dx.doi.org/10.1101/062554, 2016.
- P. Sobczyk, M. Bogdan, J. Josse, "Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood", arxiv: 1606.05333, 2016.

R packages

- *SLOPE* by E. Patterson
- *geneSLOPE* by P. Sobczyk
- *grpSLOPE* by A. Gossmann
- *bigstep* (mBIC, mBIC2) by P. Szulc
- *ClustofVar* by P. Sobczyk