3.27pt

Statystyczna Analiza "Grubych" Zbiorów Danych

Małgorzata Bogdan

Instytut Matematyczny
Uniwersytet Wrocławski

Wrocław, 28/09/2017

Baza danych (1)

Inventory List

Highlight items to reorder? Yes

Inventory ID	Name	Description	Unit Price	Quantity in Stock	Inventory Value	Reorder Level	Reorder Time in Days	Quantity in Reorder	Discontinued?
IN0001	Item 1	Desc 1	$51.00	25	$1,275.00	29	13	50	
IN0002	Item 2	Desc 2	$93.00	132	$12,276.00	231	4	50	
IN0003	Item 3	Desc 3	$57.00	151	$8,607.00	114	11	150	
IN0004	Item 4	Desc 4	$19.00	186	$3,534.00	158	6	50	
IN0005	Item 5	Desc 5	$75.00	62	$4,650.00	39	12	50	
IN0006	Item 6	Desc 6	$11.00	5	$55.00	9	13	150	
~~IN0007~~	~~Item 7~~	~~Desc 7~~	~~$56.00~~	~~58~~	~~$3,248.00~~	~~109~~	~~7~~	~~100~~	~~yes~~
IN0008	Item 8	Desc 8	$38.00	101	$3,838.00	162	3	100	
IN0009	Item 9	Desc 9	$59.00	122	$7,198.00	82	3	150	
IN0010	Item 10	Desc 10	$50.00	175	$8,750.00	283	8	150	
IN0011	Item 11	Desc 11	$59.00	176	$10,384.00	229	1	100	
IN0012	Item 12	Desc 12	$18.00	22	$396.00	36	12	50	
IN0013	Item 13	Desc 13	$26.00	72	$1,872.00	102	9	100	
IN0014	Item 14	Desc 14	$42.00	62	$2,604.00	83	2	100	
IN0015	Item 15	Desc 15	$32.00	46	$1,472.00	23	15	50	
IN0016	Item 16	Desc 16	$90.00	96	$8,640.00	180	3	50	
~~IN0017~~	~~Item 17~~	~~Desc 17~~	~~$97.00~~	~~57~~	~~$5,529.00~~	~~98~~	~~12~~	~~50~~	~~Yes~~
IN0018	Item 18	Desc 18	$12.00	6	$72.00	7	13	50	
IN0019	Item 19	Desc 19	$82.00	143	$11,726.00	164	12	150	
IN0020	Item 20	Desc 20	$16.00	124	$1,984.00	113	14	50	
IN0021	Item 21	Desc 21	$19.00	112	$2,128.00	75	11	50	
IN0022	Item 22	Desc 22	$24.00	182	$4,368.00	132	15	150	
~~IN0023~~	~~Item 23~~	~~Desc 23~~	~~$29.00~~	~~106~~	~~$3,074.00~~	~~142~~	~~1~~	~~150~~	~~Yes~~
IN0024	Item 24	Desc 24	$75.00	173	$12,975.00	127	9	100	
IN0025	Item 25	Desc 25	$14.00	28	$392.00	21	8	50	

Genetyczna Baza Danych

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]
[1,]	1	2	1	1	1	1	0	1	1	1	2	1	1
[2,]	0	1	1	2	2	2	0	1	1	0	0	0	2
[3,]	2	1	2	2	2	1	1	1	1	2	2	1	1
[4,]	1	1	1	1	0	1	1	2	2	0	0	0	1
[5,]	1	1	1	2	0	0	2	1	1	2	1	0	2
[6,]	2	0	1	2	1	1	1	2	1	1	0	2	1
[7,]	2	0	0	0	0	1	0	0	2	2	0	0	2
[8,]	1	2	0	2	2	1	1	1	1	2	2	1	0
[9,]	2	2	1	1	1	1	0	0	1	0	1	0	1
[10,]	0	2	1	1	1	0	1	0	2	2	0	1	1
[11,]	2	0	0	2	1	1	0	1	1	2	2	0	1
[12,]	0	0	2	2	1	1	1	0	1	1	2	0	1
[13,]	1	1	1	0	0	0	2	1	0	0	1	1	1
[14,]	2	0	1	2	0	2	2	1	1	1	1	1	0
[15,]	2	1	2	2	2	2	2	2	0	2	1	1	0
[16,]	1	2	2	0	2	1	1	0	1	2	2	2	0
[17,]	1	1	1	1	1	1	2	1	2	0	1	1	1
[18,]	0	1	1	0	1	2	1	0	0	0	0	0	1
[19,]	0	2	0	0	0	0	2	1	2	1	2	0	0
[20,]	1	0	2	0	2	1	1	1	2	1	1	1	1
[21,]	0	1	0	0	2	1	0	2	0	0	1	1	0
[22,]	1	0	1	1	1	1	2	0	1	0	0	1	2

"Gruba" baza danych

n - liczba rekordów (osobników)

"Gruba" baza danych

n - liczba rekordów (osobników)

p - liczba kolumn (zmiennych)

"Gruba" baza danych

n - liczba rekordów (osobników)

p - liczba kolumn (zmiennych)

$X_{n \times p}$ - baza danych w postaci macierzowej

"Gruba" baza danych

n - liczba rekordów (osobników)

p - liczba kolumn (zmiennych)

$X_{n \times p}$ - baza danych w postaci macierzowej

"'Długa baza danych": $n >> p$ - "klasyczna" statystyka, problemy informatyczne, obliczeniowe

"Gruba" baza danych

n - liczba rekordów (osobników)

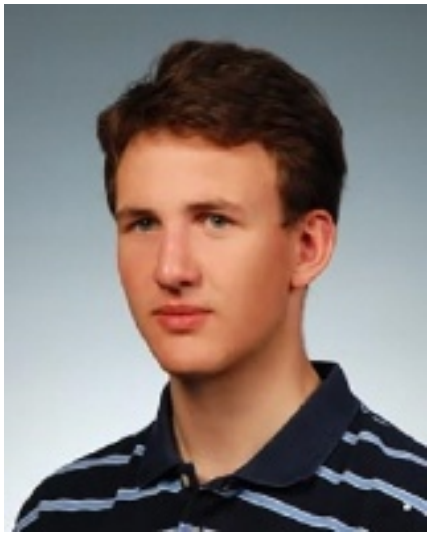
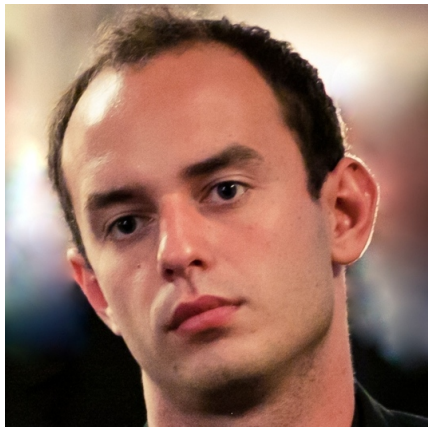
p - liczba kolumn (zmiennych)

$X_{n \times p}$ - baza danych w postaci macierzowej

"'Długa baza danych": $n >> p$ - "klasyczna" statystyka, problemy informatyczne, obliczeniowe

"Gruba baza danych": $p >> n$ - istotne problemy współczesnej statystyki

IDEAL

Przykładowe problemy

Identyfikacja powiązań między zmiennymi

Przykładowe problemy

Identyfikacja powiązań między zmiennymi

Przewidywanie wartości jednej ze zmiennych w oparciu o pozostałe zmienne:

Przykładowe problemy

Identyfikacja powiązań między zmiennymi

Przewidywanie wartości jednej ze zmiennych w oparciu o pozostałe zmienne:

przewidywanie ryzyka niespłacenia kredytu

Przykładowe problemy

Identyfikacja powiązań między zmiennymi

Przewidywanie wartości jednej ze zmiennych w oparciu o pozostałe zmienne:

przewidywanie ryzyka niespłacenia kredytu

identyfikacja genów wpływających na pewną cechę

Przykładowe problemy

Identyfikacja powiązań między zmiennymi

Przewidywanie wartości jednej ze zmiennych w oparciu o pozostałe zmienne:

przewidywanie ryzyka niespłacenia kredytu

identyfikacja genów wpływających na pewną cechę

przewidywanie reakcji pacjentów na terapię

Przykładowe problemy

Identyfikacja powiązań między zmiennymi

Przewidywanie wartości jednej ze zmiennych w oparciu o pozostałe zmienne:

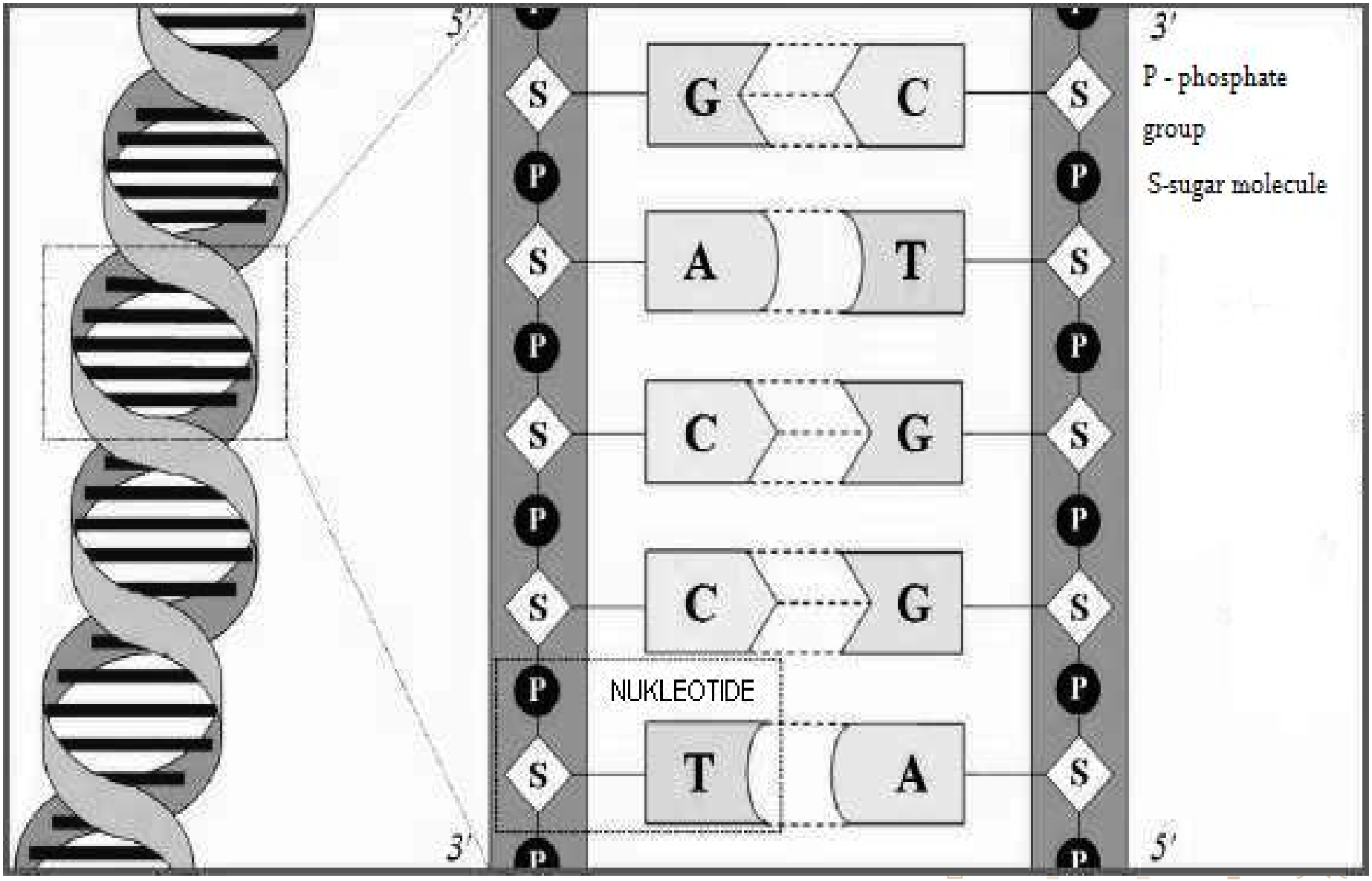
przewidywanie ryzyka niespłacenia kredytu

identyfikacja genów wpływających na pewną cechę

przewidywanie reakcji pacjentów na terapię

wybór zwierząt do hodowli

Struktura DNA



Zmienność genetyczna w populacjach ludzkich

- Około 99,9% informacji genetycznej jest dokładnie taka sama u wszystkich ludzi.
- ***Polimorfizm*** to różnica w strukturze DNA, która występuje u co najmniej 1% populacji.
- ***Polimorfizm Pojedynczego Nukelotydu (Single Nucleotide Polymorphism, SNP)*** - polimorfizm w pojedynczej bazie nukleotydowej:
  - Typowy SNP: pozycja w której
    - 85% populacji ma Cytozynę (C)
    - 15% ma Tyminę (T).
- Zwykle w danym lokusie występuję tylko dwie formy SNPa
- trzy genotypy : AA, Aa, aa.

Główny cel

Cel: Lokalizacja mutacji wpływających na zadane cechy.

Główny cel

Cel: Lokalizacja mutacji wpływających na zadane cechy.

Y - cecha ilościowa

Główny cel

Cel: Lokalizacja mutacji wpływających na zadane cechy.

Y - cecha ilościowa

Przykłady: ciśnienie krwi, poziom cholesterolu, poziom ekspresji genu

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cech dla n osobników

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cech dla n osobników

$G_{n \times M}$ - macierz genotypów dla M SNP-ów

Struktura danych

$Y = (Y_1, \ldots, Y_n)^T$ - wektor wartości cech dla n osobników

$G_{n \times M}$ - macierz genotypów dla M SNP-ów

Zwykle $n \approx k \times 100$ lub $k \times 1000$, $m \approx k \times 10,000$ lub $m \approx k \times 100,000$

Wielokrotne testowanie (1)

Wielokrotne testowanie : osobne testy w każdym SNP-ie

Wielokrotne testowanie (1)

Wielokrotne testowanie : osobne testy w każdym SNP-ie

Analiza wariancji lub prosta regresja liniowa - zwykłe kodowanie efektów addytywnych

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

Wielokrotne testowanie (1)

Wielokrotne testowanie : osobne testy w każdym SNP-ie

Analiza wariancji lub prosta regresja liniowa - zwykłe kodowanie efektów addytywnych

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

Model regresji

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

Wielokrotne testowanie (1)

Wielokrotne testowanie : osobne testy w każdym SNP-ie

Analiza wariancji lub prosta regresja liniowa - zwykłe kodowanie efektów addytywnych

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

Model regresji

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

$\hat{\beta}_i$: klasyczny estymator β_i liczony metodą najmniejszych kwadratów

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2)$$

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Odrzucamy H_{0i} kiedy $|\hat{\beta}_i| > c$

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Odrzucamy H_{0i} kiedy $|\hat{\beta}_i| > c$

Poziom istotności: $\alpha = P_{H_{0i}}(|\hat{\beta}_i| > c)$

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Odrzucamy H_{0i} kiedy $|\hat{\beta}_i| > c$

Poziom istotności: $\alpha = P_{H_{0i}}(|\hat{\beta}_i| > c)$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	M_0
H_0 fałszywa	T	S	M_1
	W	R	M

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Odrzucamy H_{0i} kiedy $|\hat{\beta}_i| > c$

Poziom istotności: $\alpha = P_{H_{0i}}(|\hat{\beta}_i| > c)$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	M_0
H_0 fałszywa	T	S	M_1
	W	R	M

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

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Odrzucamy H_{0i} kiedy $|\hat{\beta}_i| > c$

Poziom istotności: $\alpha = P_{H_{0i}}(|\hat{\beta}_i| > c)$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	M_0
H_0 fałszywa	T	S	M_1
	W	R	M

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

$$E(V) = \alpha M_0$$

Wielokrotne testowanie (2)

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2), \quad i = 1, \ldots, m$$

$$H_{0i} : \beta_i = 0 \quad \text{vs} \quad \beta_i \neq 0$$

Odrzucamy H_{0i} kiedy $|\hat{\beta}_i| > c$

Poziom istotności: $\alpha = P_{H_{0i}}(|\hat{\beta}_i| > c)$

	H_0 przyjęta	H_0 odrzucona	
H_0 prawdziwa	U	V	M_0
H_0 fałszywa	T	S	M_1
	W	R	M

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

$$E(V) = \alpha M_0$$

$$\alpha = 0.05, M_0 = 5000 \rightarrow E(V) = 250$$

Procedury wielokrotnego testowania

Korekta Bonferroniego: Stosujemy poziom istotności $\frac{\alpha}{M}$.

W GWAS typowo $M \approx 10^6$ i $\alpha = 5 \times 10^{-8}$

Procedury wielokrotnego testowania

Korekta Bonferroniego: Stosujemy poziom istotności $\frac{\alpha}{M}$.

W GWAS typowo $M \approx 10^6$ i $\alpha = 5 \times 10^{-8}$

Procedura Benjaminiego-Hochberga:

(1) $|\hat{\beta}|_{(1)} \geq |\hat{\beta}|_{(2)} \geq \ldots \geq |\hat{\beta}|_{(M)}$

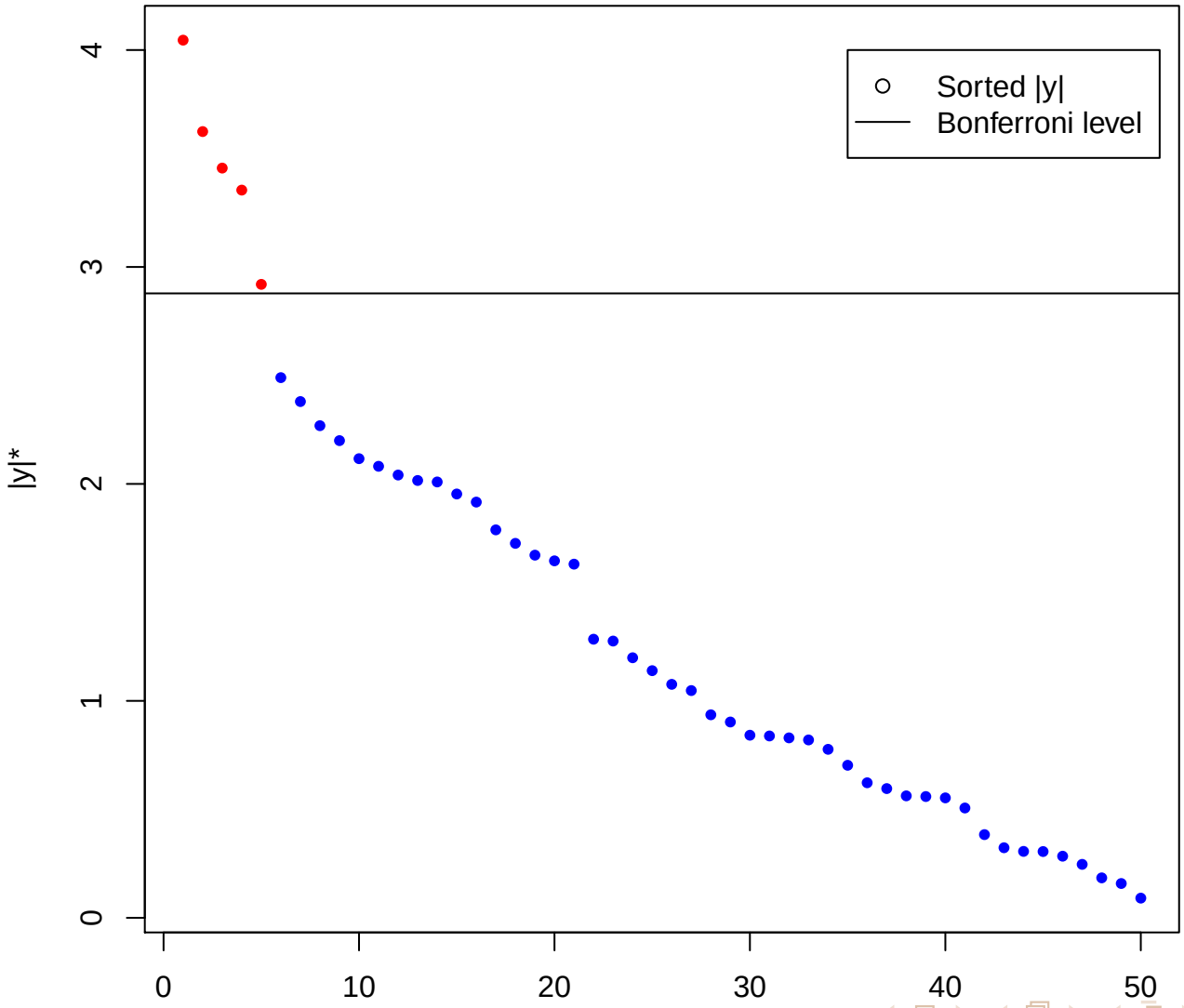
(2) Szukamy największego indeksu i takiego, że

$$|\hat{\beta}|_{(i)} \geq \sigma \Phi^{-1}(1 - \alpha_i), \quad \alpha_i = \alpha \frac{i}{2M}, \tag{1}$$

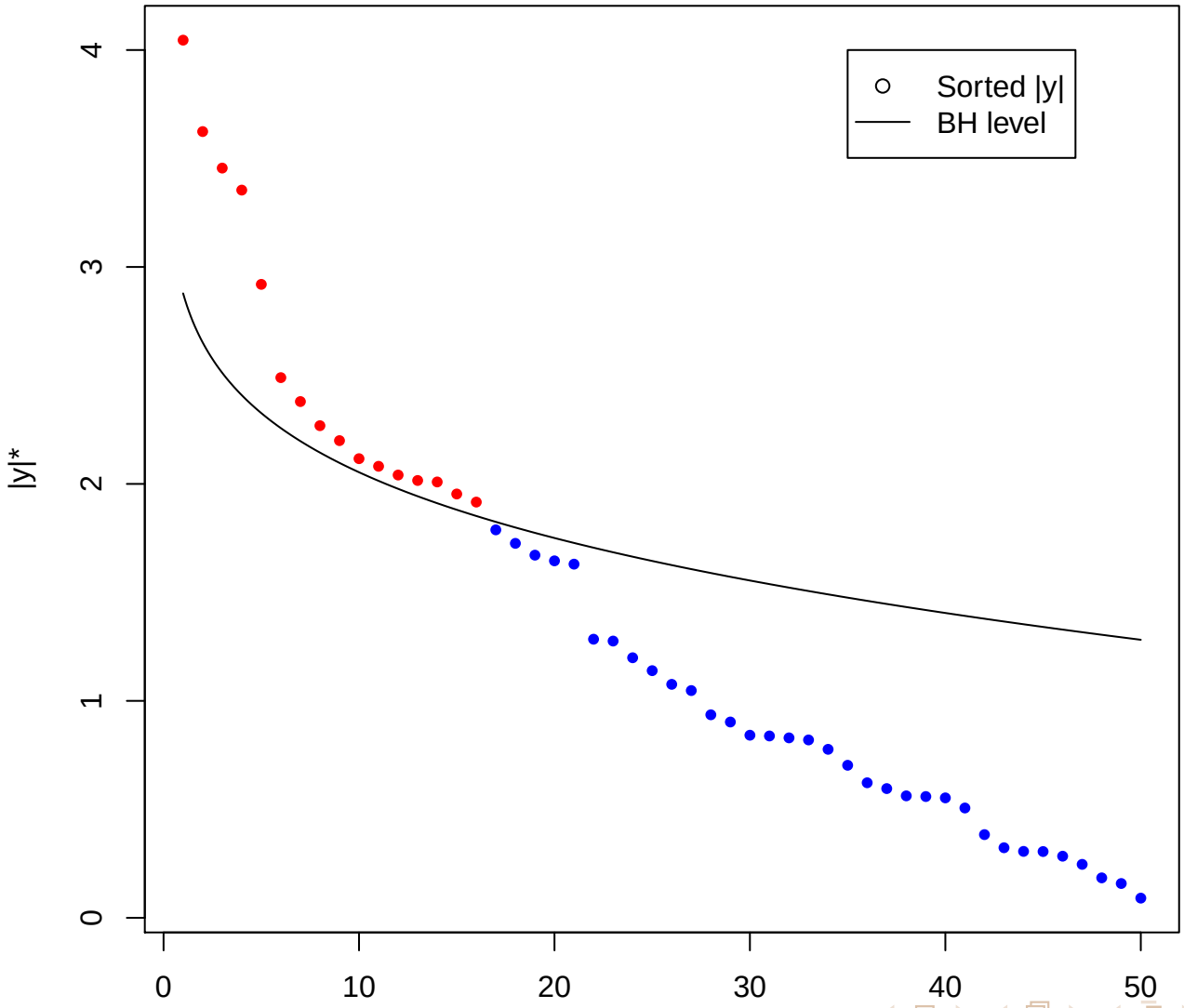
Nazywamy ten indeks i_{SU}.

(3) Odrzucamy wszystkie $H_{(i)}$ dla których $i \leq i_{\mathsf{SU}}$

Korekta Bonferroniego



Korekta Benjaminiego-Hochberga



Badanie symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

- 309790 SNPów dla 649 osób pochodzenia europejskiego

Badanie symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

- 309790 SNPów dla 649 osób pochodzenia europejskiego
- k = 40 odległych (niezależnych) mutacji

Badanie symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

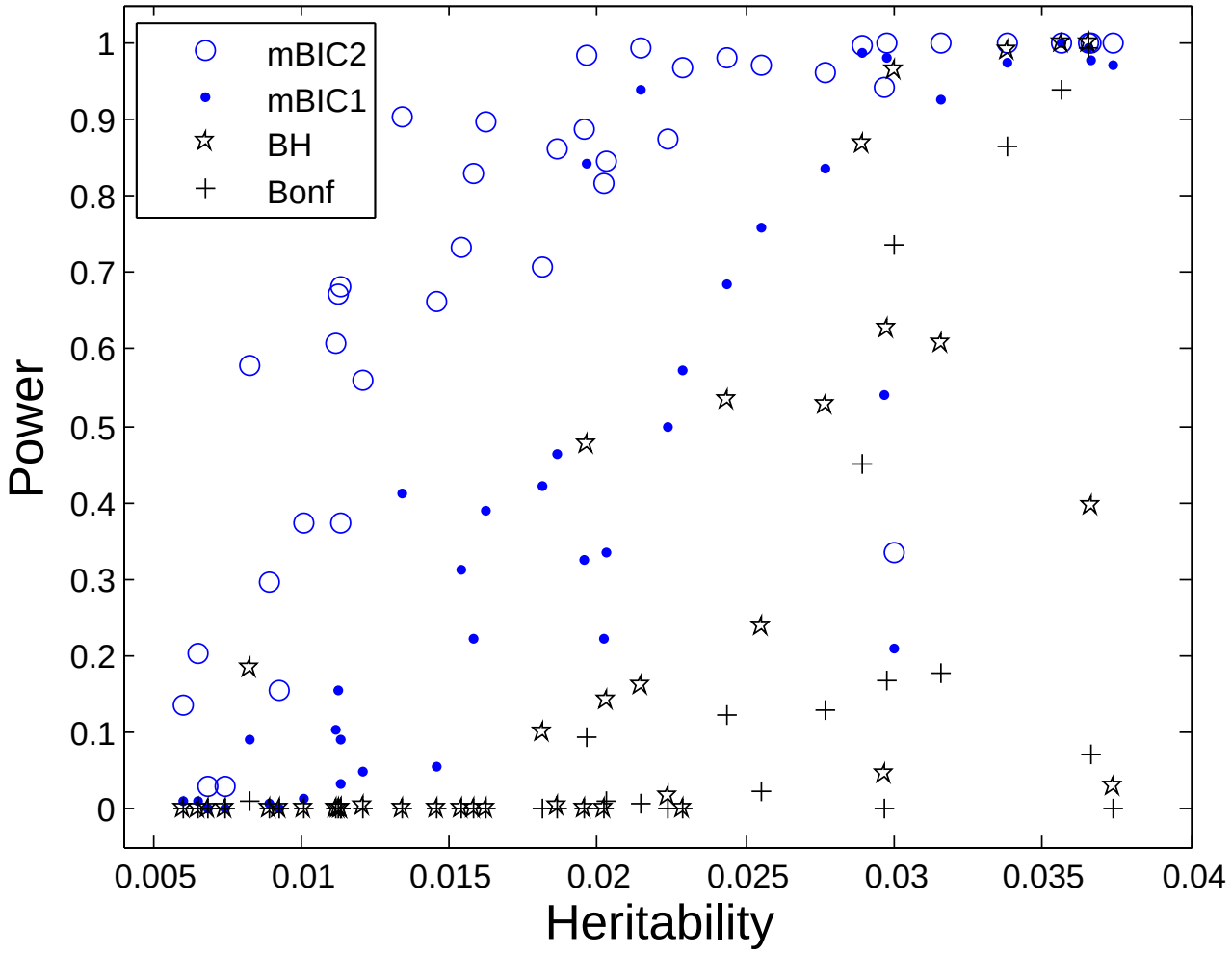
- 309790 SNPów dla 649 osób pochodzenia europejskiego
- k = 40 odległych (niezależnych) mutacji
- 1000 replikacji z modelu addytywnego M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$

Badanie symulacyjne (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

Próba POPRES rzeczywistych genomów z dbGaP

- 309790 SNPów dla 649 osób pochodzenia europejskiego
- k = 40 odległych (niezależnych) mutacji
- 1000 replikacji z modelu addytywnego M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$
- β_j równomiernie rozłożone na odcinku [0.27, 0.66]

Power


mBIC2
mBIC1
BH
Bonf
Power
Heritability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04

Kryteria wyboru modelu (1)

Cel: Estymacja β w modelu

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

Kryteria wyboru modelu (1)

Cel: Estymacja β w modelu

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

Zadanie wykonalne przy założeniu $||\beta||_0 = k << n$ (założenie o rzadkości)

Kryteria wyboru modelu (1)

Cel: Estymacja β w modelu

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

Zadanie wykonalne przy założeniu $||\beta||_0 = k << n$ (założenie o rzadkości)

Kryteria wyboru modelu: minimalizujemy $||Y - X\beta||^2 + pen(k)$

Kryteria wyboru modelu (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ generują dużo fałszywych odkryć gdy p jest duże

Kryteria wyboru modelu (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ generują dużo fałszywych odkryć gdy p jest duże

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

$pen(k) = 2\sigma^2 k \log p$ - "korekta Bonferroniego"

Kryteria wyboru modelu (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ generują dużo fałszywych odkryć gdy p jest duże

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

$pen(k) = 2\sigma^2 k \log p$ - "korekta Bonferroniego"

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

$pen(k) = 2\sigma^2 \sum_{i=1}^{k} \log(p/i)$ - "korekta Benjaminiego-Hochberga"

Kryteria wyboru modelu (2)

AIC $pen(k) = 2k$, BIC $pen(k) = k \log n$ generują dużo fałszywych odkryć gdy p jest duże

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

$pen(k) = 2\sigma^2 k \log p$ - "korekta Bonferroniego"

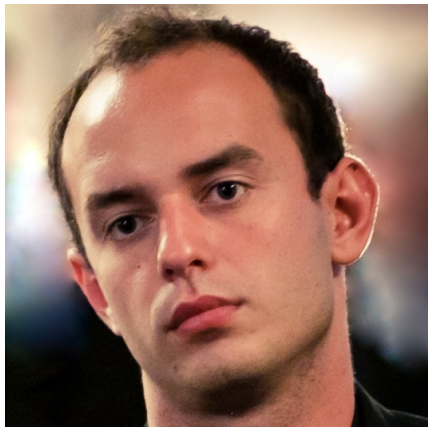
BHRIC (Abramovich et al. (2006), Foster and Stine (1999), Birge and Massart (2001))

$pen(k) = 2\sigma^2 \sum_{i=1}^{k} \log(p/i)$ - "korekta Benjaminiego-Hochberga"

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

mBIC, mBIC2 - połączenie kary BIC i BHRIC - motywacja Bayesowska

Populacje mieszane (1)



Populacje mieszane (2)

Problem - duża utrata mocy w związku z wielokrotnym testowaniem, potrzebne duże próby

Populacje mieszane (2)

Problem - duża utrata mocy w związku z wielokrotnym testowaniem, potrzebne duże próby

Rozwiązanie dostępne w populacjach mieszanych - wykorzystanie informacji o pochodzeniu fragmentów genomu

Populacje mieszane (2)

Problem - duża utrata mocy w związku z wielokrotnym testowaniem, potrzebne duże próby

Rozwiązanie dostępne w populacjach mieszanych - wykorzystanie informacji o pochodzeniu fragmentów genomu

Silne korelacje - 100-krotna redukcja liczby testów

Populacje mieszane (2)

Problem - duża utrata mocy w związku z wielokrotnym testowaniem, potrzebne duże próby

Rozwiązanie dostępne w populacjach mieszanych - wykorzystanie informacji o pochodzeniu fragmentów genomu

Silne korelacje - 100-krotna redukcja liczby testów

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", *Genetic Epidemiology, 2017*.

Populacje mieszane (2)

Problem - duża utrata mocy w związku z wielokrotnym testowaniem, potrzebne duże próby

Rozwiązanie dostępne w populacjach mieszanych - wykorzystanie informacji o pochodzeniu fragmentów genomu

Silne korelacje - 100-krotna redukcja liczby testów

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", *Genetic Epidemiology, 2017*.

P. Szulc, pakiet *bigstep*, przetwarza dane które nie mieszczą się w RAM

Populacje mieszane (2)

Problem - duża utrata mocy w związku z wielokrotnym testowaniem, potrzebne duże próby

Rozwiązanie dostępne w populacjach mieszanych - wykorzystanie informacji o pochodzeniu fragmentów genomu

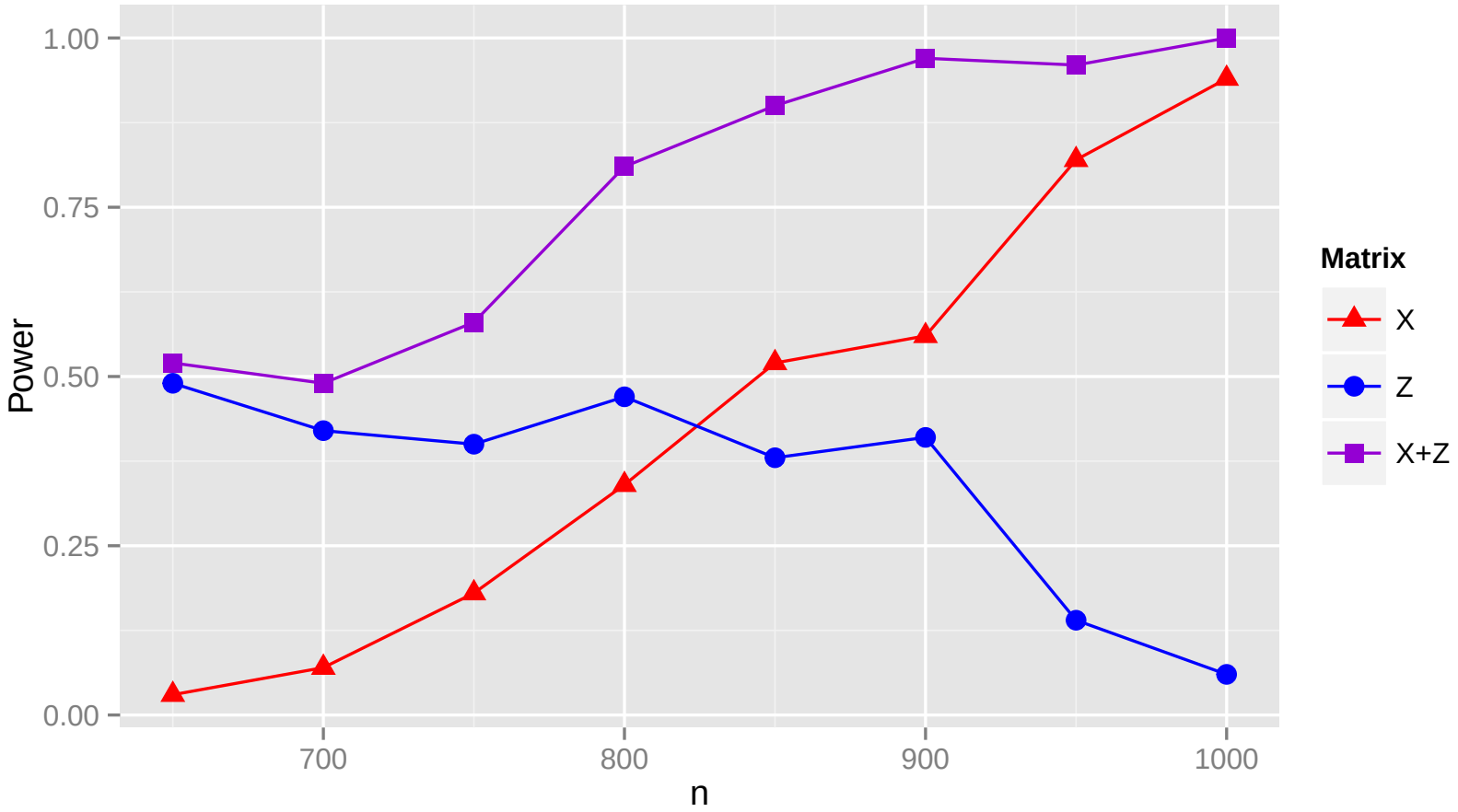
Silne korelacje - 100-krotna redukcja liczby testów

P. Szulc, M. Bogdan, F. Frommlet, H. Tang, "Joint Genotype- and Ancestry-based Genome-wide Association Studies in Admixed Populations", *Genetic Epidemiology, 2017*.

P. Szulc, pakiet *bigstep*, przetwarza dane które nie mieszczą się w RAM

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

Populacje mieszane (3)



SLOPE - wypukła relaksacja mBIC2





Stanford

SLOPE (2)

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 jest rozwiązaniem problemu optymalizacyjnego

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

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 (2)

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.

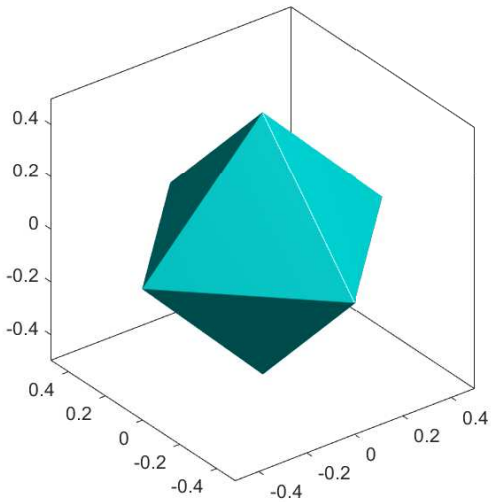
SLOPE jest rozwiązaniem problemu optymalizacyjnego

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

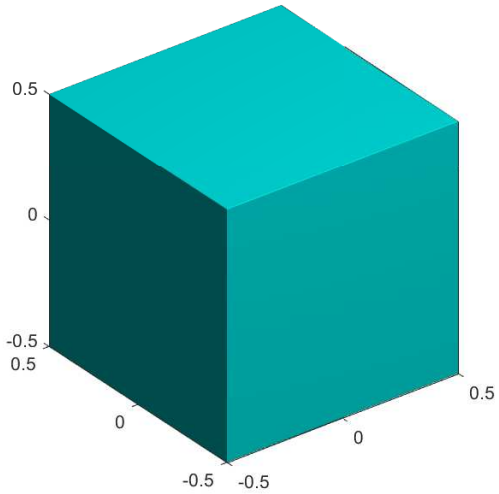
where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$

Pakiet *SLOPE* autorstwa E. Pattersona

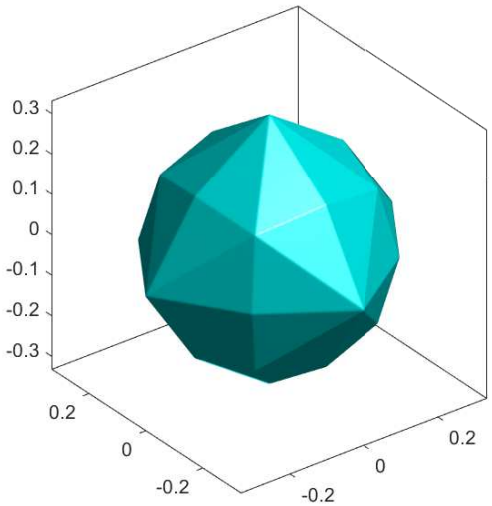
Kule jednostkowe dla różnych wersji SLOPE autorstwa D.Brzyskiego



[(2,2,2)]



[(2,0,0)]



[(3,2,1)]

Asymptotyczna optymalność

SLOPE w przeciwieństwie do LASSO umożliwia uzyskanie optymalnego rzędu błędu estymacji

Asymptotyczna optymalność

SLOPE w przeciwieństwie do LASSO umożliwia uzyskanie optymalnego rzędu błędu estymacji

Su i Candès, *Ann. Statist. 2016*, z dokładnością do stałej, Gaussowski plan eksperymentu

Asymptotyczna optymalność

SLOPE w przeciwieństwie do LASSO umożliwia uzyskanie optymalnego rzędu błędu estymacji

Su i Candès, *Ann. Statist. 2016*, z dokładnością do stałej, Gaussowski plan eksperymentu

Bellec, Lecué, Tsybakov, arxive, 2016, ogólne macierze spełniające "Restricted Eigenvalue Condition"

Asymptotyczna optymalność

SLOPE w przeciwieństwie do LASSO umożliwia uzyskanie optymalnego rzędu błędu estymacji

Su i Candès, *Ann. Statist. 2016*, z dokładnością do stałej, Gaussowski plan eksperymentu

Bellec, Lecué, Tsybakov, arxive, 2016, ogólne macierze spełniające "Restricted Eigenvalue Condition"

Bellec i Tsybakov, arxive 2017

Asymptotyczna optymalność

SLOPE w przeciwieństwie do LASSO umożliwia uzyskanie optymalnego rzędu błędu estymacji

Su i Candès, *Ann. Statist. 2016*, z dokładnością do stałej, Gaussowski plan eksperymentu

Bellec, Lecué, Tsybakov, arxive, 2016, ogólne macierze spełniające "Restricted Eigenvalue Condition"

Bellec i Tsybakov, arxive 2017

Rozszerzenia do GLM: Abramovicha i Grinshtein, arxive, 2017
Alquier, Cottet i Lecué, arxive, 2017

Inne artykuły o SLOPE

D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries", Genetics, 2017.

Inne artykuły o SLOPE

D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries", Genetics, 2017.

Pakiet *geneSLOPE* autorstwa P. Sobczyka

Inne artykuły o SLOPE

D. Brzyski, C.B. Peterson, P.Sobczyk, E.J. Candès, M. Bogdan, C. Sabatti, "Controlling the rate of GWAS false discoveries", Genetics, 2017.

Pakiet *geneSLOPE* autorstwa P. Sobczyka

S. Lee, D. Brzyski, M. Bogdan, "Fast Saddle-Point Algorithm for Generalized Dantzig Selector and FDR Control with the Ordered l_1-Norm", *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:W and CP* **vol.51**, 780–789, 2016

Grupowe SLOPE (1)

GrupoweSLOPE (2)

D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", w recenzji do *Journal of the American Statistical Association*, Nagroda ENAR (East North American Region of International Biometric Society) dla Damiana Brzyskiego, Marzec 2017

GrupoweSLOPE (2)

D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", w recenzji do *Journal of the American Statistical Association*, Nagroda ENAR (East North American Region of International Biometric Society) dla Damiana Brzyskiego, Marzec 2017

Pakiet *grpSLOPE* autorstwa A. Gossmanna

GrupoweSLOPE (2)

D. Brzyski, A. Gossmann, W.Su, M. Bogdan, "Group SLOPE - adaptive selection of groups of predictors", w recenzji do *Journal of the American Statistical Association*, Nagroda ENAR (East North American Region of International Biometric Society) dla Damiana Brzyskiego, Marzec 2017

Pakiet *grpSLOPE* autorstwa A. Gossmanna

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", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017

Symulacje w kontekście genetycznym

$n = 5402$, $p = 26233$ - słabo skorelowanych SNPów

Symulacje w kontekście genetycznym

$n = 5402$, $p = 26233$ - słabo skorelowanych SNPów

Scenariusz 1: $Y = X\beta + z$ - model addytywny

$n = 5402$, $p = 26233$ - słabo skorelowanych SNPów

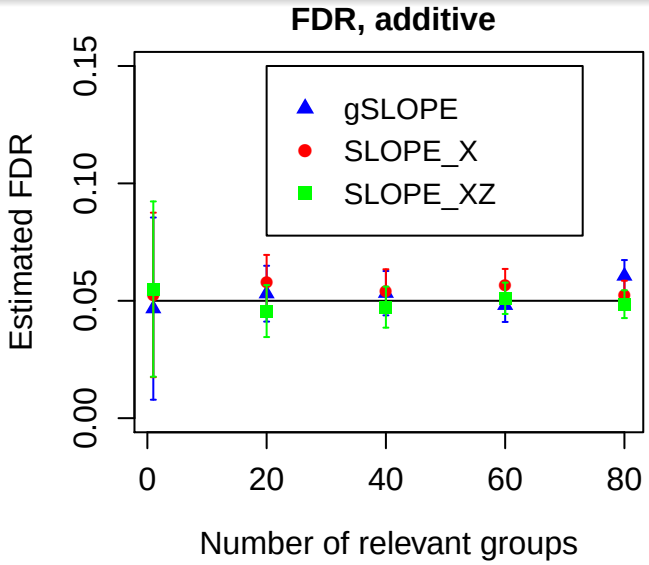
Scenariusz 1: $Y = X\beta + z$ - model addytywny

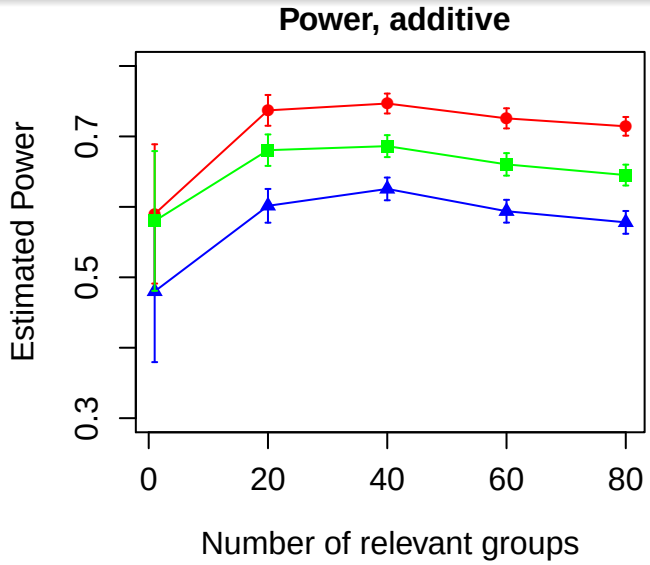
Scenariusz 2: modelowanie dominancji

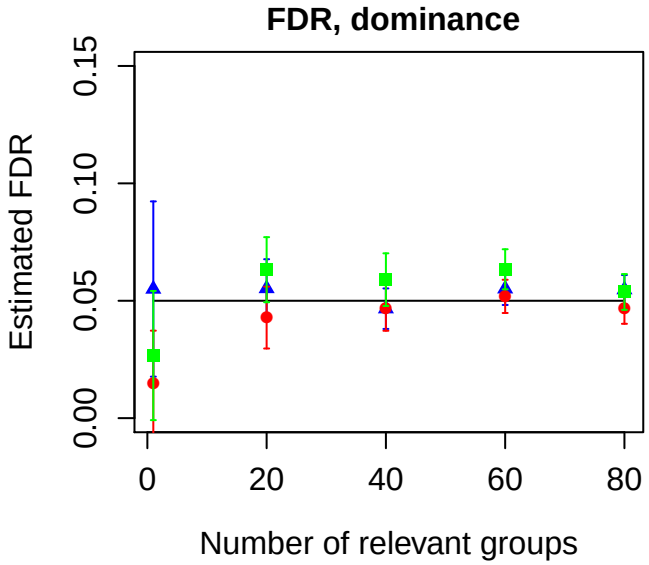
$$\tilde{z}_{ij} = \begin{cases} -1 & \text{for} \quad aa, AA \\ 1 & \text{for} \quad aA \end{cases} , \tag{2}$$

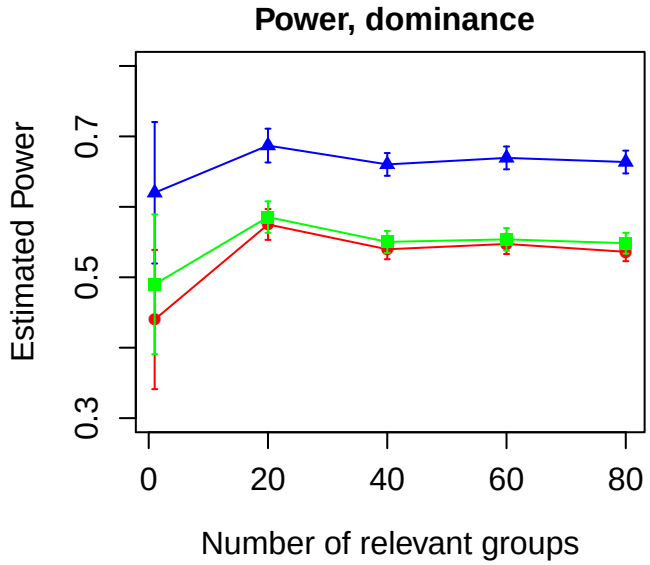
$$y = [X, Z][\beta'_X, \beta'_Z]' + \epsilon \ .$$

Wyniki symulacji



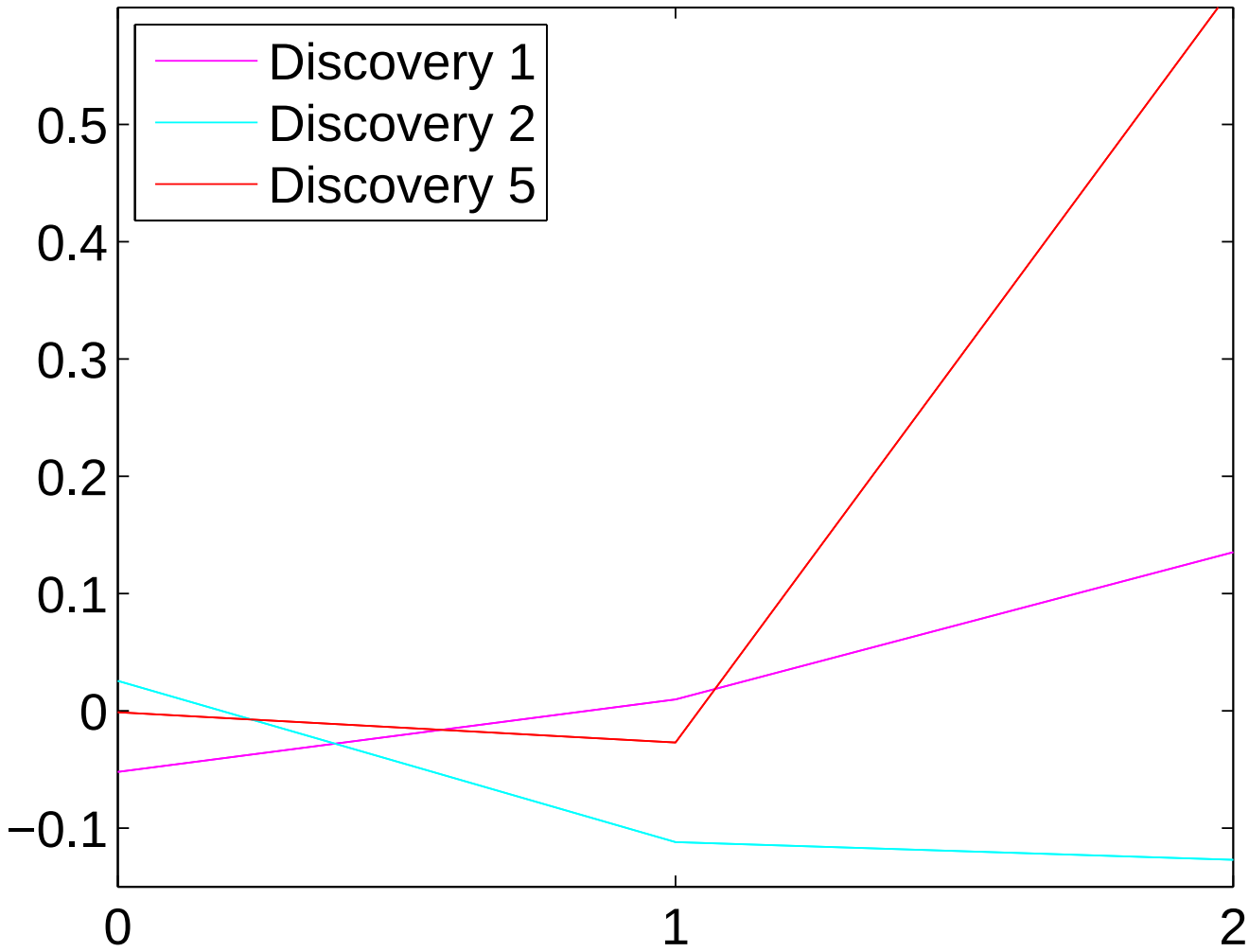






Geny wpływające na poziom trójglicerydów

5 nowych odkryć za pomocą grupowego SLOPE - rzadkie recesywne warianty.



Analiza Składowych Głównych - redukcja wymiaru w danych ómicznych"

$X_{n\times p}$ - macierz danych (np. tzw. ekspresje genów), $n = k \times 100$, $p \approx 20000$ - liczba genów

Analiza Składowych Głównych - redukcja wymiaru w danych ómicznych"

$X_{n\times p}$ - macierz danych (np. tzw. ekspresje genów), $n = k \times 100$, $p \approx 20000$ - liczba genów

Analiza Składowych Głównych - redukcja wymiaru w danych ómicznych"

$X_{n\times p}$ - macierz danych (np. tzw. ekspresje genów), $n = k \times 100$, $p \approx 20000$ - liczba genów

Założenie : $X = M + E$, gdzie M jest niskiego rzędu a E jest szumem losowym

Analiza Składowych Głównych - redukcja wymiaru w danych ómicznych"

$X_{n\times p}$ - macierz danych (np. tzw. ekspresje genów), $n = k \times 100$, $p \approx 20000$ - liczba genów

Założenie : $X = M + E$, gdzie M jest niskiego rzędu a E jest szumem losowym

Zwykle zakłada się, że $e_{ij} \sim N(0, \sigma)$

Analiza Składowych Głównych - redukcja wymiaru w danych ómicznych"

$X_{n\times p}$ - macierz danych (np. tzw. ekspresje genów), $n = k \times 100$, $p \approx 20000$ - liczba genów

Założenie : $X = M + E$, gdzie M jest niskiego rzędu a E jest szumem losowym

Zwykle zakłada się, że $e_{ij} \sim N(0, \sigma)$

Cel matematyczny - odzyskanie M, separacja sygnału od szumu

Analiza Składowych Głównych - redukcja wymiaru w danych ómicznych"

$X_{n\times p}$ - macierz danych (np. tzw. ekspresje genów), $n = k \times 100$, $p \approx 20000$ - liczba genów

Założenie : $X = M + E$, gdzie M jest niskiego rzędu a E jest szumem losowym

Zwykle zakłada się, że $e_{ij} \sim N(0, \sigma)$

Cel matematyczny - odzyskanie M, separacja sygnału od szumu

Cel praktyczny - kompresja danych,kilka wektorów bazowych [składowych głównych] może zawierać większość informacji danych i być użytych do predykcji

Analiza składowych głównych (2)

Metoda - rozkład X według wartości osobliwych:

$$X = U_{n\times l} D_{l\times l} V^T_{l\times p} \ ,$$

$$U^T U = I_{l\times l}, V^T V = I_{l\times l}, \ \ l = min\{n, p\}$$

Analiza składowych głównych (2)

Metoda - rozkład X według wartości osobliwych:

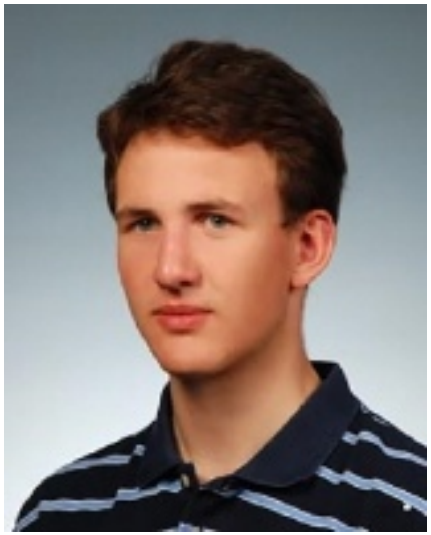
$$X = U_{n\times l} D_{l\times l} V^T_{l\times p} \ ,$$

$$U^T U = I_{l\times l}, V^T V = I_{l\times l}, \ \ l = min\{n, p\}$$

Cel statystyczny - ustalenie rzędu k macierzy M

PESEL (PEnalized SEmi-integrated Likelihood)

Sobczyk, Bogdan, Josse, Journal of Computational Graphical Statistics, 2017



Bayesowskie Kryterium Informacyjne (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - zagnieżdżony ciąg modeli statystycznych

Bayesowskie Kryterium Informacyjne (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - zagnieżdżony ciąg modeli statystycznych

Bayesowskie Kryterium Informacyjne (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - zagnieżdżony ciąg modeli statystycznych

W naszym przypadku A_k - model zakładający, że liczba niezerowych wartości singularnych jest niewiększa od k

Bayesowskie Kryterium Informacyjne (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - zagnieżdżony ciąg modeli statystycznych

W naszym przypadku A_k - model zakładający, że liczba niezerowych wartości singularnych jest niewiększa od k

θ - wektor parametrów modelu A_k:

elementy macierzy $U_k \in S_{k,n}$, $V_k \in S_{k,p}$, D_k, i σ

$S_{k,n}$ - rozmaitość Stiefela macierzy ortonormalnych wymiaru $n \times k$

Bayesowskie Kryterium Informacyjne (BIC) (1)

$A_1 \in A_2 \in A_3 \ldots$ - zagnieżdżony ciąg modeli statystycznych

W naszym przypadku A_k - model zakładający, że liczba niezerowych wartości singularnych jest niewiększa od k

θ - wektor parametrów modelu A_k:

elementy macierzy $U_k \in S_{k,n}$, $V_k \in S_{k,p}$, D_k, i σ

$S_{k,n}$ - rozmaitość Stiefela macierzy ortonormalnych wymiaru $n \times k$

$l(X, \theta)$ - funkcja wiarogodności (gęstość rozkładu łącznego opisującego dane)

Bayesowskie Kryterium Informacyjne (BIC) (2)

W ogólnej sytuacji Bayesowskie Kryterium Informacyjne zaleca wybór modelu maksymalizującego

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

gdzie N jest liczbą niezależnych obserwacji

Bayesowskie Kryterium Informacyjne (BIC) (2)

W ogólnej sytuacji Bayesowskie Kryterium Informacyjne zaleca wybór modelu maksymalizującego

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

gdzie N jest liczbą niezależnych obserwacji

Uzasadnione (zgodne) w przypadkach gdy $dim(A_k) = const$ gdy $N \to \infty$

Bayesowskie Kryterium Informacyjne (BIC) (2)

W ogólnej sytuacji Bayesowskie Kryterium Informacyjne zaleca wybór modelu maksymalizującego

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

gdzie N jest liczbą niezależnych obserwacji

Uzasadnione (zgodne) w przypadkach gdy $dim(A_k) = const$ gdy $N \to \infty$

U nas $N = np$ a $dim(A_k)$ rośnie liniowo ze względu na n i na p

Bayesowskie Kryterium Informacyjne (BIC) (2)

W ogólnej sytuacji Bayesowskie Kryterium Informacyjne zaleca wybór modelu maksymalizującego

$$max_{\theta \in A_k} \log l(X, \theta) - 1/2 dim(A_k) \log N$$

gdzie N jest liczbą niezależnych obserwacji

Uzasadnione (zgodne) w przypadkach gdy $dim(A_k) = const$ gdy $N \to \infty$

U nas $N = np$ a $dim(A_k)$ rośnie liniowo ze względu na n i na p

Pomysł - redukcja liczby parametrów poprzez wycałkowanie względem pewnego rozkładu a priori

PESEL dla dużego p

Zakładamy, że $M = TW^T$, gdzie
$T = [t_{i,l}]_{n \times k}$ jest macierzą zawierającą wartości zmiennych ukrytych rozpinających dane,
$W = [w_{i,l}]_{p \times k}$ jest macierzą współczynników.

PESEL dla dużego p

Zakładamy, że $M = TW^T$, gdzie
$T = [t_{i,l}]_{n\times k}$ jest macierzą zawierającą wartości zmiennych ukrytych rozpinających dane,
$W = [w_{i,l}]_{p\times k}$ jest macierzą współczynników.

rozkład apriori -

$$w_{j\cdot} \sim N(0, I_k)$$

co implikuje, że $x_{\cdot 1}, \ldots, x_{\cdot p}$ są niezależnymi zmiennymi losowymi z rozkładu

$$x_{\cdot j} \sim N(0; TT^T + \sigma^2 I_n) \quad .$$

PESEL dla dużego p

Zakładamy, że $M = TW^T$, gdzie
$T = [t_{i,l}]_{n \times k}$ jest macierzą zawierającą wartości zmiennych ukrytych rozpinających dane,
$W = [w_{i,l}]_{p \times k}$ jest macierzą współczynników.

rozkład apriori -

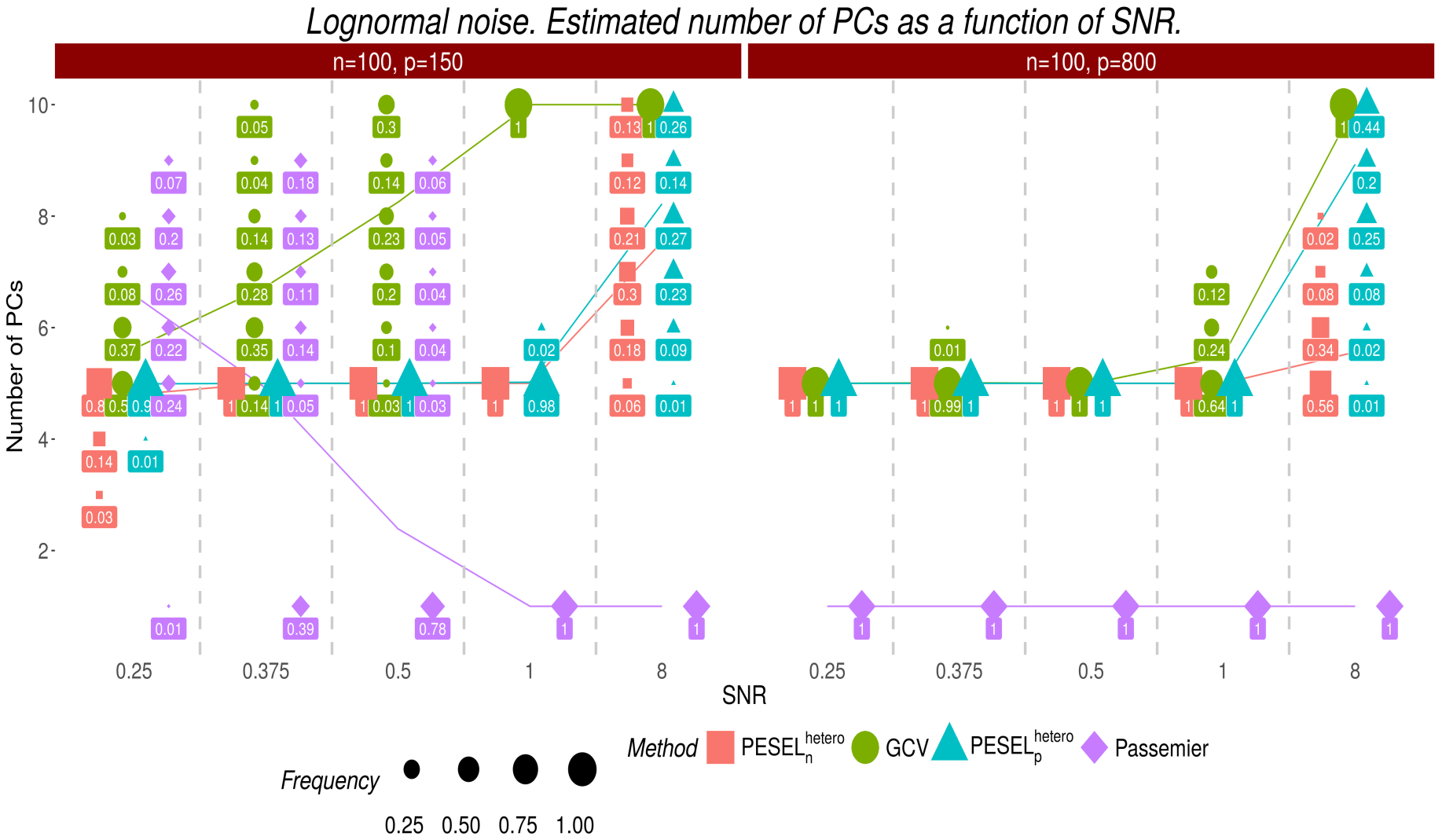
$$w_{j\cdot} \sim N(0, I_k)$$

co implikuje, że $x_{\cdot 1}, \ldots, x_{\cdot p}$ są niezależnymi zmiennymi losowymi z rozkładu

$$x_{\cdot j} \sim N(0; TT^T + \sigma^2 I_n) \quad .$$

Teraz mamy p niezależnych wektorów a liczba parametrów nie zależy od p - możemy pryncypialnie stosować BIC o ile $p >> n$

Zakłócenia z rozkładu logarytmiczno-normalnego



Klastrowanie w podprzestrzeniach

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Josse

Klastrowanie w podprzestrzeniach

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Josse

Nagrody dla mlodych naukowców dla P. Sobczyka (Vienna workshop on simulation, 2015) i S. Wilczyńskiego (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Klastrowanie w podprzestrzeniach

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Josse

Nagrody dla mlodych naukowców dla P. Sobczyka (Vienna workshop on simulation, 2015) i S. Wilczyńskiego (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Cel: Identyfikacja grup powiązanych zmiennych (ścieżek sygnałowych) i wybór odpowiednich składowych głównych.

.

Klastrowanie w podprzestrzeniach

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Josse

Nagrody dla mlodych naukowców dla P. Sobczyka (Vienna workshop on simulation, 2015) i S. Wilczyńskiego (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Cel: Identyfikacja grup powiązanych zmiennych (ścieżek sygnałowych) i wybór odpowiednich składowych głównych.

Matematycznie: podział zmiennych na grupy z których każda rozpięta jest przez kilka zmiennych ukrytych.

Klastrowanie w podprzestrzeniach

Multiple Latent Component Clustering, Sobczyk, Wilczyński, Bogdan, Josse

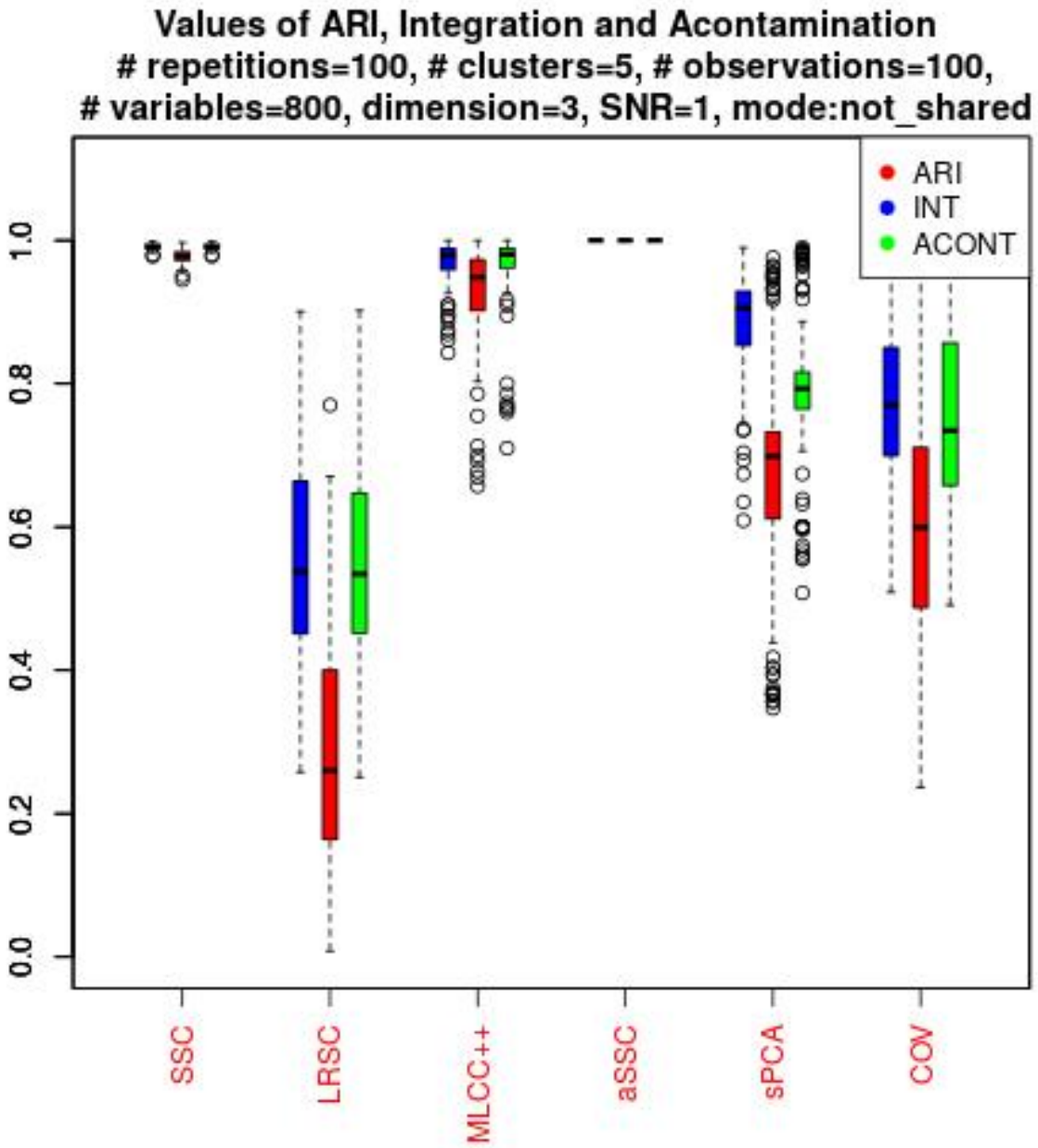
Nagrody dla mlodych naukowców dla P. Sobczyka (Vienna workshop on simulation, 2015) i S. Wilczyńskiego (International Conference on Biometrics and Bio-Pharmaceutical Statistics, Wiedeń 2017)

Cel: Identyfikacja grup powiązanych zmiennych (ścieżek sygnałowych) i wybór odpowiednich składowych głównych.

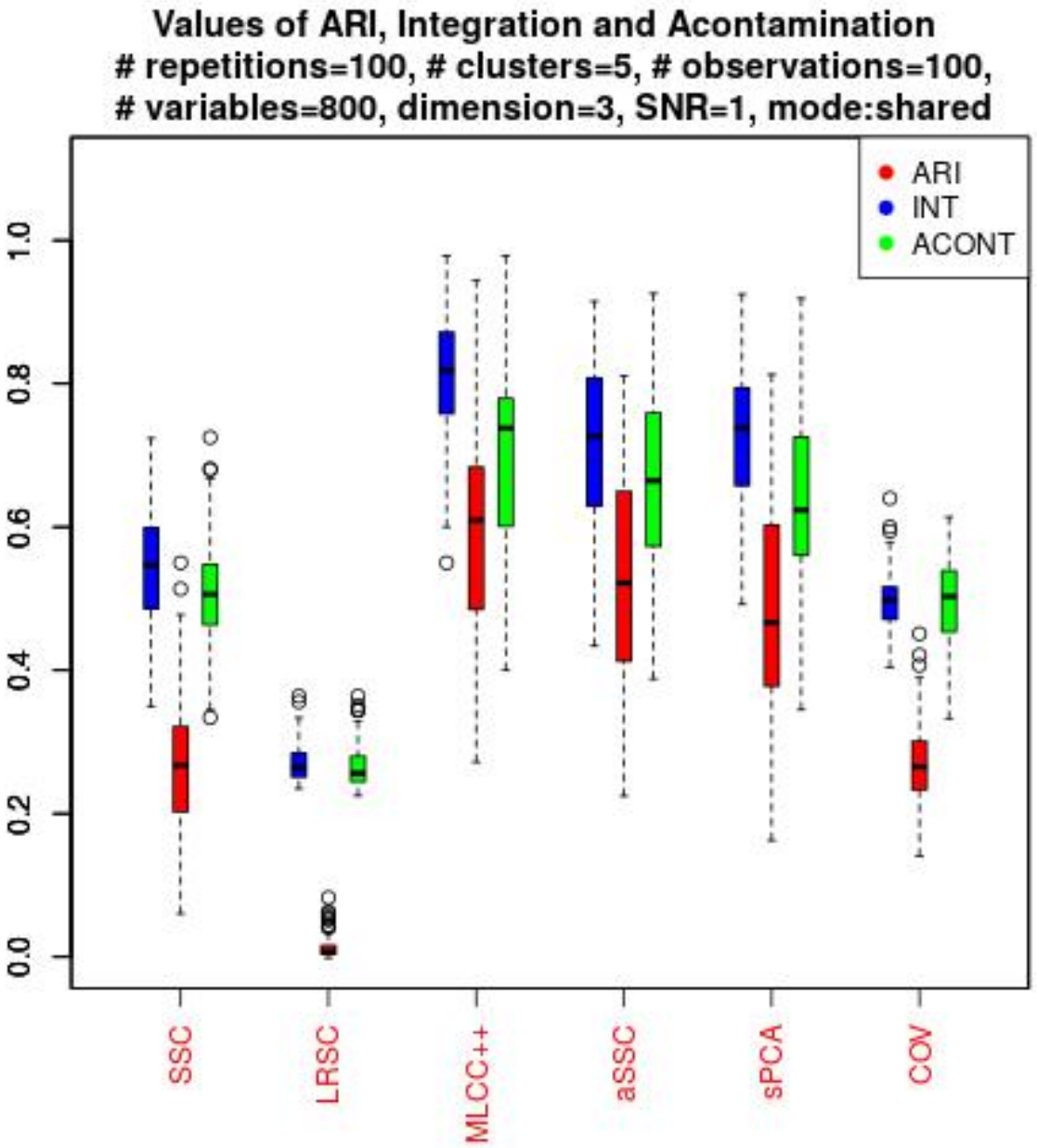
Matematycznie: podział zmiennych na grupy z których każda rozpięta jest przez kilka zmiennych ukrytych.

Pakiet *varclust* autorstwa P. Sobczyka- Algorytm K-średnich wokół składowych głównych. Estymacja liczby grup i ich wymiarowości za pomocą modyfikacji kryterium Bayesowskiego.

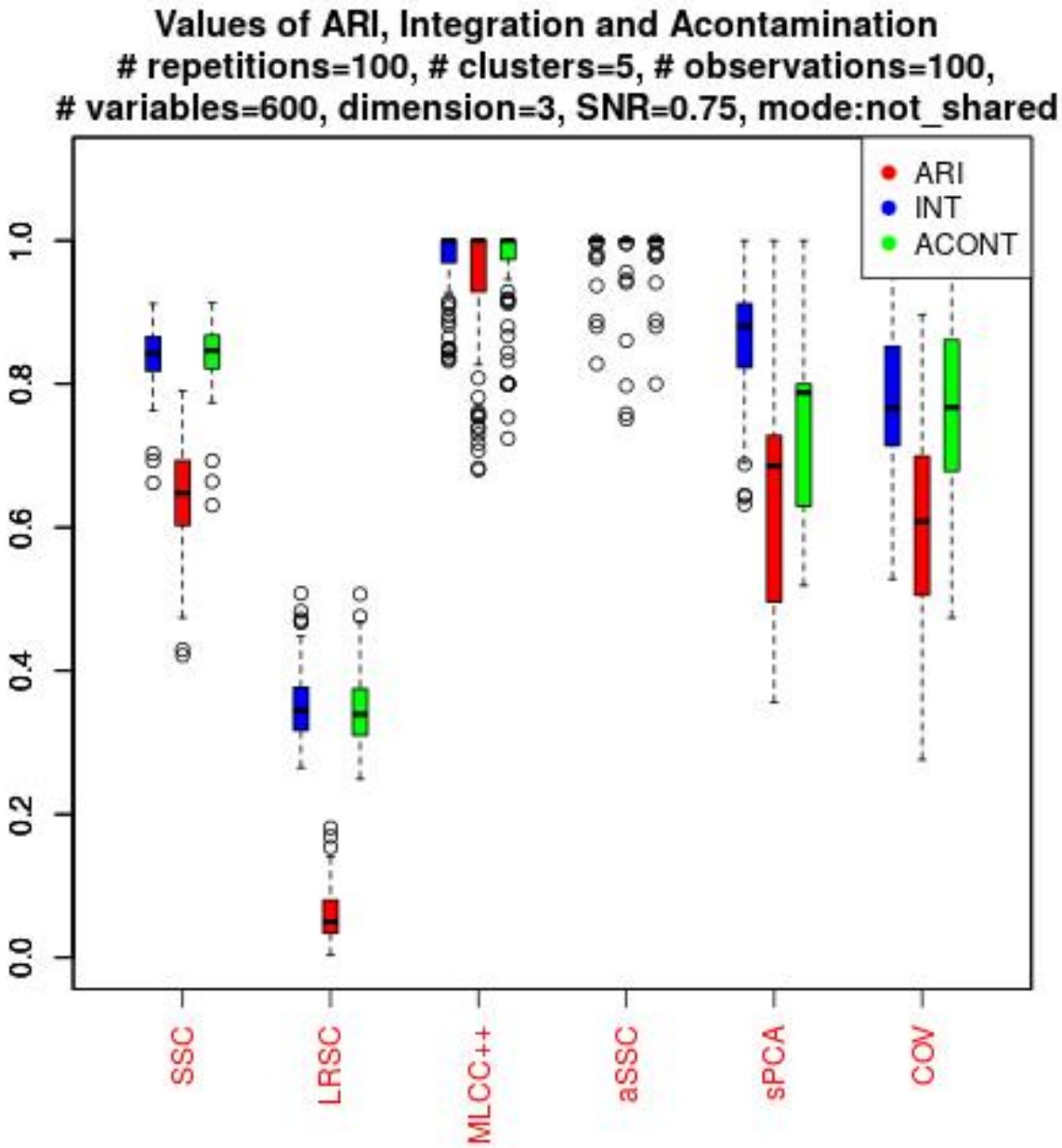
Varclust (1)


Values of ARI, Integration and Acontamination
repetitions=100, # clusters=5, # observations=100,
variables=800, dimension=3, SNR=1, mode:not_shared
ARI
INT
ACONT
0.0
0.2
0.4
0.6
0.8
1.0
SSC
LRSC
MLCC++
aSSC
sPCA
COV

Varclust (2)


Values of ARI, Integration and Acontamination
repetitions=100, # clusters=5, # observations=100,
variables=800, dimension=3, SNR=1, mode:shared
ARI
INT
ACONT
0.0
0.2
0.4
0.6
0.8
1.0
SSC
LRSC
MLCC++
aSSC
sPCA
COV

Varclust (3)


Values of ARI, Integration and Acontamination
repetitions=100, # clusters=5, # observations=100,
variables=600, dimension=3, SNR=0.75, mode:not_shared
ARI
INT
ACONT
0.0
0.2
0.4
0.6
0.8
1.0
SSC
LRSC
MLCC++
aSSC
sPCA
COV

bigstep (P.Szulc) - kryteria informacyjne dla modelu regresji w "grubych" bazach danych

Pakiety w R

bigstep (P.Szulc) - kryteria informacyjne dla modelu regresji w "grubych" bazach danych

SLOPE (E. Paterson), *geneSLOPE* (P.Sobczyk), *grpSLOPE* (A. Gossmann) - optymalizacja wypukła dla modelu regresji w "grubych" bazach danych

Pakiety w R

bigstep (P.Szulc) - kryteria informacyjne dla modelu regresji w "grubych" bazach danych

SLOPE (E. Paterson), *geneSLOPE* (P.Sobczyk), *grpSLOPE* (A. Gossmann) - optymalizacja wypukła dla modelu regresji w "grubych" bazach danych

PESEL (P. Sobczyk) - estymacja liczby składowych głównych

Pakiety w R

bigstep (P.Szulc) - kryteria informacyjne dla modelu regresji w "grubych" bazach danych

SLOPE (E. Paterson), *geneSLOPE* (P.Sobczyk), *grpSLOPE* (A. Gossmann) - optymalizacja wypukła dla modelu regresji w "grubych" bazach danych

PESEL (P. Sobczyk) - estymacja liczby składowych głównych

varclust (P. Sobczyk) - klastrowanie w przestrzeniach nisko-wymiarowych, wersja robocza dostępna na *github*

Pakiety w R

bigstep (P.Szulc) - kryteria informacyjne dla modelu regresji w "grubych" bazach danych

SLOPE (E. Paterson), *geneSLOPE* (P.Sobczyk), *grpSLOPE* (A. Gossmann) - optymalizacja wypukła dla modelu regresji w "grubych" bazach danych

PESEL (P. Sobczyk) - estymacja liczby składowych głównych

varclust (P. Sobczyk) - klastrowanie w przestrzeniach nisko-wymiarowych, wersja robocza dostępna na *github*

Komentarze i uwagi mile widziane

Pakiety w R

bigstep (P.Szulc) - kryteria informacyjne dla modelu regresji w "grubych" bazach danych

SLOPE (E. Paterson), *geneSLOPE* (P.Sobczyk), *grpSLOPE* (A. Gossmann) - optymalizacja wypukła dla modelu regresji w "grubych" bazach danych

PESEL (P. Sobczyk) - estymacja liczby składowych głównych

varclust (P. Sobczyk) - klastrowanie w przestrzeniach nisko-wymiarowych, wersja robocza dostępna na *github*

Komentarze i uwagi mile widziane

Do zrobienia i przebadania - logistyczne SLOPE i "Ordered Dantzig Selector" - mamy wersje matlabowe