Model selection approach for genome wide association studies

Malgorzata Bogdan, Florian Frommlet, Piotr Szulc, Hua Tang

Wroclaw University of Technology
Medical University of Vienna
Stanford University

Piza, 7.12.2014

Outline

- Genomewide Association Studies
- Admixtures: Genotype and Ancestry Information
- Modified versions of BIC for admixtures
- Simulation Study

Genome-Wide Association Studies

MAIN PURPOSE: finding the mutations in DNA sequence, that influence the trait of interest.

Y - quantitative trait

Examples: blood pressure, cholesterol level, gene expression level

Data structure

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

$G_{n \times m}$ - matrix of SNP genotypes (Affymetrix, Illumina, Roche)

Caution - SNPs are just markers. We can not assume that the causal mutation is represented on the SNP microarray.

Short range and irregular dependency structure

Regions of a low linkage disequilibrium - weak correlation

High density of markers required

Usually $n \approx k \times 100$ or $k \times 1000$, $m \approx k \times 10,000$ or $m \approx k \times 100,000$

Single marker tests

Multiple testing : separate tests at every SNP
One way ANOVA or Simple Regression - usual coding

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

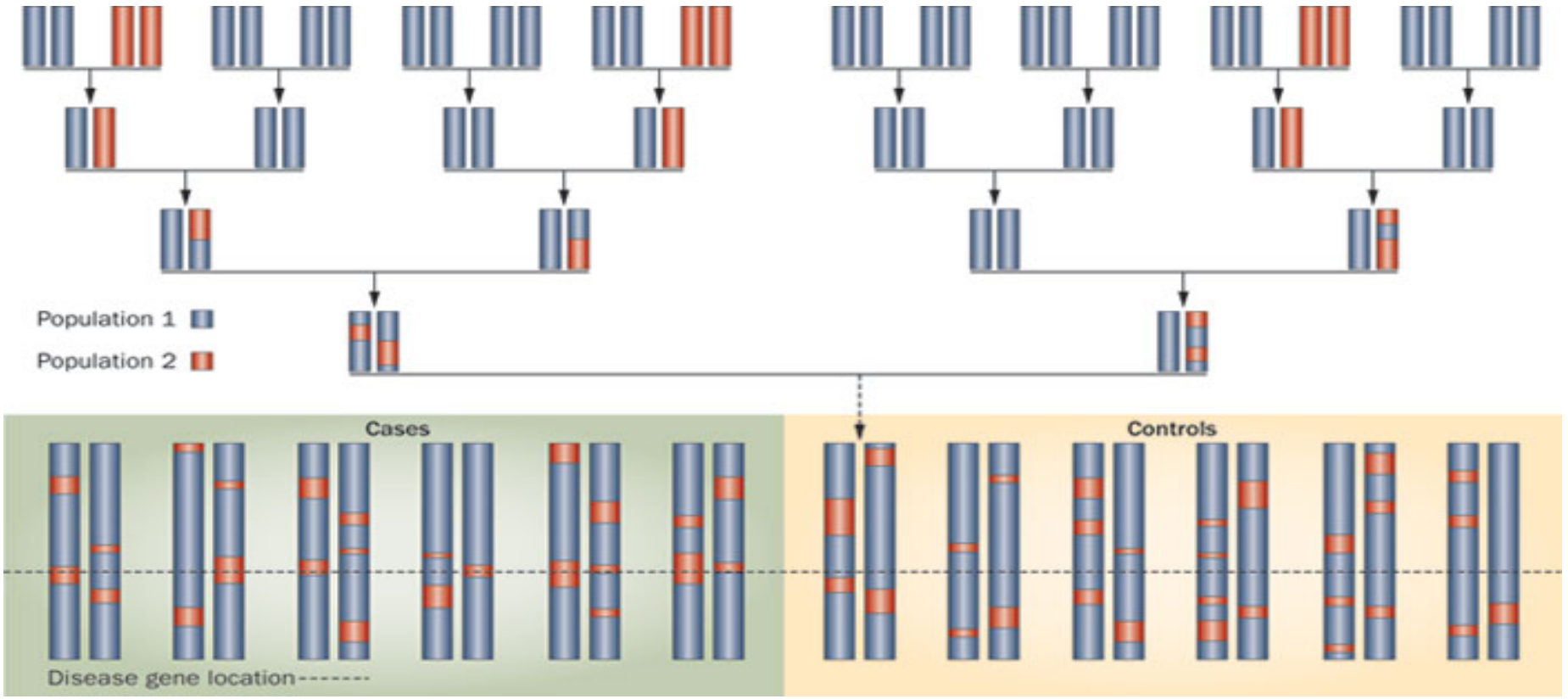
Simple regression model: $Y_i = \beta_0 + \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$

$$T_j = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

If $\beta_j = 0$ then $T_j \sim t_{n-2}$
Multiple testing procedures: e.g. Bonferroni and Benjamini-Hochberg

Population admixtures

Picture from Rosset, Tzur, Behar, Wasser and Karl Skorecki, Nature Reviews Nephrology 7, 313-326 (June 2011)



Ancestry state

Locus-specific ancestry can be accurately estimated based on the genotype data from standard genotyping platforms and distribution of haplotypes in ancestral population (see e.g. methods based on Hidden Markov models in Tang et al. (2006, Am. J. Hum. Gen.) or Price et al. (2009, PLOS Genet.)).

Strong correlation structure - reduced correction for multiple testing

Coding :

$$Z_{ij} = \begin{cases} 0 & \text{if} \quad A_{ij} = bb \\ 1 & \text{if} \quad A_{ij} = bB \\ 2 & \text{if} \quad Z_{ij} = BB \end{cases}$$

Admixture mapping - looking for association between the ancestry and the trait

When is ancestry information useful ? (1)

Assumption - the trait is determined by the genotype at “causal" loci X_j, $j \in \{1, \ldots, k\}$.

Notation: $p_{jb}(a)$ - frequency of a allele at jth locus in the population b

If $p_{jb}(a) = 0$ and $p_{jB}(a) = 1$ then $Z_j = X_j$

If $p_{jb}(a) = p_{jB}(a)$ then $\rho(Z_j, X_j) = 0$

Corollary : Admixture mapping can detect only those “causal" loci, for which the allelic distribution differs between admixing population.

When is ancestry information useful ? (2)

q_j - average jth locus specific ancestry in the considered population

$Cov(X_j, Z_j) = 2q_j(1 - q_j)(p_{jB} - p_{jb})$

If $q_j = 0.5$ then

$$\rho(X_j, Z_j) = \frac{p_{jB} - p_{jb}}{\sqrt{(p_{jB} + p_{jb})(2 - (p_{jB} + p_{jb}))}}.$$

If the maximal correlation between X_j and the genotypes of neigboring markers is comparable or smaller than $\rho(X_j, Z_j)$ then the admixture mapping will typically have a larger power than the association mapping.

Admixture mapping can help to detect genes in the regions of a low linkage disequilibrium and such that their allelic frequencies differ between parental populations.

False Associations

μ_b - expected value of the trait in the population b

Assumptions: $\mu_b > \mu_B$, e.g. due to the polygenic effects, $p_{jb}(a) > p_{jB}(a)$

$$\rho(Y, X_j) > 0$$

Spourious association between X and Y

Solution - conditioning on Q - genomewide ancestry for i-th individual

Statistical models for single marker tests:

$$Y_i = \beta_0 + \beta_Q Q_i + \beta_j X_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$Y_i = \beta_0 + \beta_Q Q_i + \beta_j Z_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Tang, Siegmund, Johnson, Romieu, London: (2010, Genet. Epidemiol.) - Combine ancestry and genotype information in a new two degrees of freedom "TDT" test.
In the context of regression one could consider a joint test for:

$$H_0 : \beta_{Xj} = \beta_{Zj} = 0$$

$$Y_i = \beta_0 + \beta_Q Q_i + \beta_{Xj} X_{ij} + \beta_{Zj} Z_{ij} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \ .$$

In many cases one of these variables would be sufficient to detect a gene. Two degrees of freedom - unnecessary inflation of critical values - loss of power.

mBIC2

$$Y_i = \beta_0 + \beta_Q Q_i + \sum_{j \in I} \beta_{Xj} X_{ij} + \sum_{j \in J} \beta_{Zj} Z_{ij} + \varepsilon_i, \qquad (1)$$

I, J - subsets of $N = \{1, \ldots, m\}$, $\epsilon_i \sim N(0, \sigma^2)$

Żak-Szatkowska, Bogdan (CSDA, 2011), Frommlet et al. (CSDA, 2012), for similar criteria see also Foster and George (Biometrika 2004) and Abramovich et al. (Ann. Statist. 2006)

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

Derived by the analogy to BH

Ancestry dummy variables - adjustment for correlation, Bogdan et al. (Biometrics, 2008)

Hybrid isolation model: $\rho = Corr(Z_j, Z_{j+1}|Q = q) = exp(-t\Delta)$, where t is the time from the admixing event and Δ is the distance between loci (in Morgans).

$$Y_i = \mu + \beta_0 Q_i + \beta_j Z_{ij} \quad .$$

Feingold, Brown and Siegmund, Genetics, 1993 - Modelling the distribution of the t-test statistics by the Gaussian process

$$P_{H_0}\left(max_j LRT_j > c\right) \approx 1 - exp(-2[1-\Phi(\sqrt{c})]) - 0.02mt\Delta\sqrt{c}\nu\left(\sqrt{0.02t\Delta c}\right) \quad ,$$

where

$$\nu(t) \approx \frac{(2/t)(\Phi(t/2) - 0.5)}{(t/2)\Phi(t/2) + \phi(t/2)}.$$

Effective number of tests (1)

Alternatively, FWER resulting from performing m^{eff} independent test is

$$\alpha = P_{H_0}\left(\max_{i\in\{1,\ldots,m^{eff}\}} LRT_j > c\right) \approx 1 - \left[1 - 2\left(1 - \Phi(\sqrt{(c)})\right)\right]^{m^{eff}} \quad .$$

The effective number of tests can be calculated as

$$m^{eff} = \log(1-\alpha)/\log\left(2\Phi(\sqrt{c}) - 1\right) \quad .$$

$\overline{\log \rho}$ - the average of the logarithms of the correlations between ancestry dummy variables at neigboring markers

$t\Delta := -\overline{\log \rho}$

m_{eff} may be also calculated based on the simulations/permutations

Effective number of tests (2)

Table : Effective number of tests for 22 chromosomes.

Chr	L_{tot}	$\bar{L}$	m	m_{eff}
1	278.09	0.0075	37173	397
2	263.45	0.0066	39958	376
3	224.62	0.0067	33385	314
4	213.19	0.0073	29290	295
5	203.98	0.0067	30587	281
6	193.02	0.0060	32204	266

Model selection for admixtures:

$$\begin{aligned} \text{mBIC2}: n \log RSS &+ k_{\mathbf{j}}(\log n + 2\log(m/4)) - 2\log(k_{\mathbf{j}}!) && (2) \\ &+ \tilde{k}_{\mathbf{j}}(\log n + 2\log(m^{eff}/4)) - 2\log(\tilde{k}_{\mathbf{j}}!)\ , && (3) \end{aligned}$$

Search strategy

1. Aggregated forward selection based on BIC
2. Stepwise selection starting with the model constructed in 1.
3. Threshold for stepwise selection is determined by *mBIC*2.

Simulation Study (1)

Hybrid isolation admixture model. Basic populations - African Americans, Europeans

482 298 SNPs from Illumina 650K microarray (X chromosome is excluded), 1000 individuals, $m^{eff} = 4722$

$Q \sim Beta(7, 3)$, $E(Q) = 0.7$

$T \sim 15 * Beta(2, 4) + 5$, $E(T) = 10$

"Recombination" points are generated according to $d \sim Exp(\lambda = T)$ distribution. At recombination points ancestry is randomly generated as a Bernoulli variable, P(A)=Q. Block genotypes are randomly sampled from the HapMap data for the given population.

Scenario 1

Table : SNPs selected for Scenario 1

	SNP's name	AF	MAF	LD
1	ch01_27796	0.000	0.455	0.994
2	ch03_10846	0.000	0.418	0.990
3	ch05_07371	0.000	0.414	0.991
4	ch10_00444	0.000	0.488	0.990
5	ch02_39189	0.000	0.432	0.943
6	ch17_04306	0.000	0.495	0.942
7	ch19_06378	0.000	0.466	0.991
8	ch22_00033	0.000	0.485	0.947
9	ch01_32763	0.803	0.430	0.872
10	ch04_05127	0.765	0.461	0.993
11	ch06_25838	0.743	0.428	0.895
12	ch11_12611	0.719	0.491	0.807
13	ch12_03421	0.808	0.419	0.977
14	ch14_06999	0.821	0.414	0.996
15	ch15_03859	0.785	0.401	0.932
16	ch16_04525	0.720	0.426	0.868
17	ch01_19810	0.715	0.497	0.363
18	ch08_15190	0.583	0.400	0.377
19	ch02_22034	0.634	0.456	0.379
20	ch10_08265	0.646	0.492	0.377
21	ch11_20057	0.718	0.447	0.358
22	ch18_01031	0.650	0.431	0.382
23	ch19_01377	0.656	0.499	0.376
24	ch03_02703	0.654	0.497	0.460

Scenario 2

Table : SNPs selected for Scenario 2

SNP's no.	SNP's name	AF	MAF	LD
1	ch01_00531	0.674	0.483	0.347
2	ch01_19810	0.715	0.497	0.364
3	ch04_22846	0.745	0.500	0.505
4	ch08_12075	0.812	0.407	0.624
5	ch02_16712	0.755	0.409	0.650
6	ch11_20899	0.779	0.428	0.682
7	ch03_26157	0.769	0.425	0.691
8	ch05_16192	0.741	0.433	0.899
9	ch15_03859	0.785	0.401	0.931
10	ch07_05936	0.824	0.404	0.954
11	ch12_03421	0.808	0.419	0.977
12	ch14_06999	0.821	0.415	0.996
13	ch13_05394	0.458	0.410	0.396
14	ch20_12128	0.450	0.401	0.429
15	ch19_00410	0.467	0.411	0.499
16	ch21_02904	0.453	0.419	0.599
17	ch18_01592	0.447	0.421	0.698
18	ch16_06363	0.446	0.451	0.904
19	ch22_03194	0.458	0.486	0.912
20	ch17_11568	0.458	0.459	0.996

Simulation Study (3)

Statistical model:

$$Y_i = 0.5 \sum_{j=1}^{k} X_j + \epsilon_j \ ,$$

where $\epsilon_j \sim N(0, 1)$.

LD - maximal correlation with 50 neigboring SNPs on each side

AF - difference in allelic frequencies between ancestral populations

"Causal" SNPs are removed from the data set used to locate them.

Simulation study (3)

100 simulation runs

Average power - percentage of detected causal genes

Average empirical FDR - proposition of false discoveries among all discoveries

What is the true/false positive ?

We used the 0.5 correlation cutoff for [X,causal X] or [Z, causal Z].

Multiple testing procedures - concept of scan statistics (Siegmund, Biometrika 2010). Detected SNP + its 0.5 correlation neigborhood are classified as a one (true or false) discovery.

FWER

Table : Familywise Error Rate, 1000 simulations (no differences between mBIC and mBIC2).

Matrix X	Matrix X+Z
0.016	0.037

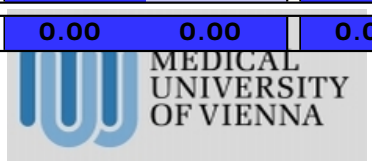
Results

BMIX - Shriner et al (PLOS Comput. Biol., 2011)

Table : Summary results: TP, FP and FDR

	Bonf		B-H		BMIX	mBIC2		
	X	Z	X	Z	X+Z	X	Z	X+Z
	Scenario1							
TP	8.04	4.68	11.95	8.26	6.65	15.41	9.43	20.81
FP	0.21	0.23	2.31	1.01	0.29	2.18	0.51	0.69
FDR	0.03	0.16	0.05	0.11	0.04	0.12	0.05	0.03
	Scenario2							
TP	5.56	6.30	7.32	9.90	9.74	9.82	8.54	15.14
FP	0.52	0.44	2.72	1.83	0.69	1.98	0.68	0.63
FDR	0.08	0.07	0.27	0.16	0.07	0.17	0.07	0.04

	Bonf		BH		mBIC2		
	X	Z	X	Z	X	Z	X+Z
1	0.99	0.00	1.00	0.00	1.00	0.00	1.00 (Z: 0.00)
2	0.73	0.00	0.94	0.00	0.99	0.00	1.00 (Z: 0.00)
3	1.00	0.00	1.00	0.00	1.00	0.00	1.00 (Z: 0.00)
4	0.50	0.00	0.82	0.00	1.00	0.00	0.97 (Z: 0.00)
5	1.00	0.00	1.00	0.00	1.00	0.00	1.00 (Z: 0.00)
6	0.34	0.00	0.66	0.00	1.00	0.00	0.99 (Z: 0.00)
7	0.65	0.00	0.88	0.00	1.00	0.00	1.00 (Z: 0.00)
8	0.29	0.00	0.68	0.00	1.00	0.00	1.00 (Z: 0.00)
9	0.18	0.52	0.59	0.85	0.72	0.92	0.92 (Z: 0.63)
10	0.67	0.56	0.95	0.85	1.00	0.66	0.99 (Z: 0.03)
11	0.21	0.20	0.63	0.54	1.00	0.62	0.99 (Z: 0.21)
12	0.00	0.00	0.02	0.10	0.87	0.09	0.76 (Z: 0.23)
13	0.62	0.79	0.86	0.95	1.00	0.88	1.00 (Z: 0.14)
14	0.11	0.30	0.42	0.68	0.96	0.91	0.92 (Z: 0.15)
15	0.23	0.10	0.58	0.48	0.87	0.73	0.94 (Z: 0.21)
16	0.52	0.85	0.92	0.98	1.00	0.99	1.00 (Z: 0.03)
17	0.00	0.29	0.00	0.55	0.00	0.59	0.89 (Z: 0.89)
18	0.00	0.00	0.00	0.04	0.00	0.07	0.17 (Z: 0.17)
19	0.00	0.00	0.00	0.03	0.00	0.34	0.54 (Z: 0.54)
20	0.00	0.56	0.00	0.89	0.00	0.69	0.85 (Z: 0.85)
21	0.00	0.21	0.00	0.51	0.00	0.55	0.95 (Z: 0.95)
22	0.00	0.23	0.00	0.61	0.00	0.83	0.85 (Z: 0.85)
23	0.00	0.37	0.00	0.75	0.00	0.66	0.71 (Z: 0.71)
24	0.00	0.00	0.00	0.00	0.00	0.02	0.24 (Z: 0.24)



	Bonf		BH		mBIC2		
	X	Z	X	Z	X	Z	X+Z
1	0.00	0.53	0.00	0.85	0.00	0.75	0.95 (Z: 0.95)
2	0.00	0.60	0.00	0.87	0.00	0.78	0.89 (Z: 0.89)
3	0.00	0.05	0.00	0.23	0.00	0.45	0.88 (Z: 0.88)
4	0.06	0.96	0.15	1.00	0.40	0.95	0.98 (Z: 0.98)
5	0.02	0.80	0.07	0.97	0.63	0.89	0.95 (Z: 0.91)
6	0.00	0.15	0.03	0.55	0.07	0.44	0.48 (Z: 0.34)
7	0.00	0.30	0.08	0.73	0.23	0.64	0.86 (Z: 0.72)
8	0.08	0.08	0.27	0.24	0.81	0.21	0.78 (Z: 0.06)
9	0.58	0.16	0.79	0.34	0.98	0.16	0.99 (Z: 0.00)
10	0.53	0.62	0.8	0.92	0.97	0.44	0.98 (Z: 0.29)
11	0.79	0.84	0.95	0.99	1.00	0.96	0.99 (Z: 0.09)
12	1.00	1.00	1.00	1.00	1.00	1.00	0.99 (Z: 0.02)
13	0.00	0.00	0.00	0.00	0.00	0.00	0.01 (Z: 0.01)
14	0.00	0.01	0.00	0.09	0.00	0.12	0.32 (Z: 0.32)
15	0.00	0.01	0.00	0.04	0.00	0.06	0.02 (Z: 0.02)
16	0.03	0.05	0.15	0.25	0.42	0.11	0.62 (Z: 0.16)
17	0.00	0.25	0.01	0.71	0.34	0.23	0.49 (Z: 0.12)
18	0.78	0.06	0.93	0.45	1.00	0.36	0.96 (Z: 0.00)
19	0.85	0.00	0.98	0.01	1.00	0.00	1.00 (Z: 0.00)
20	0.54	0.00	0.85	0.00	0.96	0.00	1.00 (Z: 0.00)

Multiple regression vs Single marker tests

$$\hat{\beta} \approx \frac{Cov(Y-\beta_Q Q, X)}{VarX}$$

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

$$Cov(Y - \beta_Q Q, X_1) = \beta_1 VarX_1 + \sum_{i=2}^{k} \beta_i Cov(X_1, X_i) + Cov(X_1, \epsilon)$$

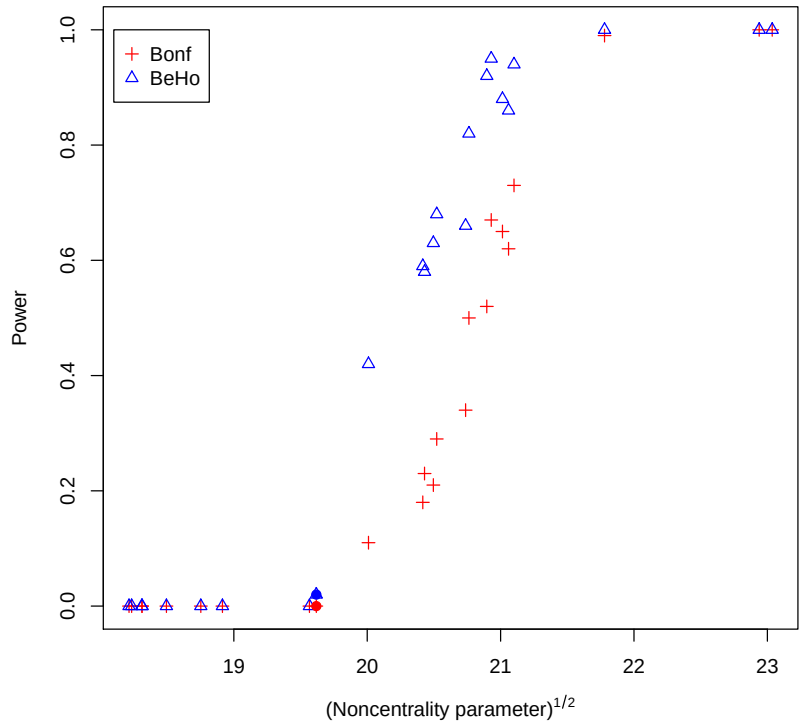
Assume that for $i > 1$, $Cov(X_1, X_i) \sim N(0, \sigma_c^2)$

$$E \sum_{i=2}^{k} \beta_i Cov(X_1, X_i) = 0$$

$$Var(\sum_{i=2}^{k} \beta_i Cov(X_1, X_i)) \approx \sum_{i=2}^{k} \beta_i^2 \sigma_c^2$$

Politechnika
Wrocławska

Power vs noncentrality parameter



Politechnika
Wrocławska

Acknowledgment

This research received funding from

- European Union's 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552
- Fulbright Program of the US Department of State