A Model Selection Approach for Genome Wide Association Studies

Florian Frommlet, Piotr Twarog, Malgorzata Bogdan

Department of Statistics and Decision Support Systems,
University of Vienna, Austria

Paris, August 2010

Genome Wide Association Studies

Data structure: $Y \leftarrow X_1, \dots, X_p$

Up to one million SNPs $X_1, \dots, X_p$
Trait Y quantitative or categorical (case control)

Question:

Which X_i are actually associated with trait?

Virtually all GWAS published so far: Single marker analysis

Model selection approach

Model specified by index vector $M = [i_1, \dots, i_{k_M}]$

$$\mathcal{M}: \; Y = X_M \beta_M + \epsilon, \quad X_M = [X_{i_1}, \dots, X_{i_{k_M}}]$$

Genome Wide Association Studies

Data structure: $Y \leftarrow X_1, \ldots, X_p$

Up to one million SNPs $X_1, \ldots, X_p$
Trait Y quantitative or categorical (case control)

Question:

Which X_i are actually associated with trait?

Virtually all GWAS published so far: Single marker analysis

Model selection approach

Model specified by index vector $M = [i_1, \ldots, i_{k_M}]$

$$\mathcal{M}: \; Y = X_M \beta_M + \epsilon, \quad X_M = [X_{i_1}, \ldots, X_{i_{k_M}}]$$

Genome Wide Association Studies

Data structure: $Y \leftarrow X_1, \ldots, X_p$

Up to one million SNPs $X_1, \ldots, X_p$
Trait Y quantitative or categorical (case control)

Question:

Which X_i are actually associated with trait?

Virtually all GWAS published so far: Single marker analysis

Model selection approach

Model specified by index vector $M = [i_1, \ldots, i_{k_M}]$

$$\mathcal{M} : \; Y = X_M \beta_M + \epsilon, \quad X_M = [X_{i_1}, \ldots, X_{i_{k_M}}]$$

Classical model selection criteria

Selection criteria based on likelihood L_M

Penalization of model size

$$-2 \log L_M + \text{Penalty} \cdot k_M$$

Examples: AIC, BIC, RIC, Mallows C, etc.

AIC . . . Penalty $= 2$, BIC . . . Penalty $= \log n$

L_1- penalization: LASSO
etc.

Classical model selection criteria

Selection criteria based on likelihood L_M

Penalization of model size

$$-2 \log L_M + \text{Penalty} \cdot k_M$$

Examples: AIC, BIC, RIC, Mallows C, etc.

AIC $\ldots$ Penalty $= 2$, BIC $\ldots$ Penalty $= \log n$

L_1- penalization: LASSO

etc.

Classical model selection criteria

Selection criteria based on likelihood L_M

Penalization of model size

$$-2\log L_M + \text{Penalty}\cdot k_M$$

Examples: AIC, BIC, RIC, Mallows C, etc.

AIC $\dots$ Penalty $= 2$, BIC $\dots$ Penalty $= \log n$

L_1- penalization: LASSO
etc.

Situation when $p > n$

Classical theory for AIC and BIC

Developed for p constant and $n \to \infty$

Results no longer valid when $p > n$
e.g. BIC no longer consistent

Sparsity

Theory possible when number of true signals $k \ll p$

Reasonable assumption, only few SNPs expected to be associated with trait

Surprise

Under sparsity and $p > n$ BIC is choosing too large models

Situation when $p > n$

Classical theory for AIC and BIC

Developed for p constant and $n \to \infty$

Results no longer valid when $p > n$
e.g. BIC no longer consistent

Sparsity

Theory possible when number of true signals $k \ll p$

Reasonable assumption, only few SNPs expected to be associated with trait

Surprise

Under sparsity and $p > n$ BIC is choosing too large models

Situation when $p > n$

Classical theory for AIC and BIC

Developed for p constant and $n \to \infty$

Results no longer valid when $p > n$
e.g. BIC no longer consistent

Sparsity

Theory possible when number of true signals $k \ll p$

Reasonable assumption, only few SNPs expected to be associated with trait

Surprise

Under sparsity and $p > n$ BIC is choosing too large models

Modifications of BIC

$$BIC = -2 \log L_M + k_M \log n$$

For situation $p > n$ under sparsity [Bogdan et al. (2004)]

$$mBIC = -2 \log L_M + k_M \log(np^2 + d)$$

In a particular sense controlling FWE (related to Bonferroni)

FDR - controlling model selection criterion

$$mBIC2 = -2 \log L_M + k_M \log(np^2 + d) - 2 \log k_m!$$

Adaptivity to level of sparsity [Abramovich et al. (2006)]

Modifications of BIC

$$BIC = -2 \log L_M + k_M \log n$$

For situation $p > n$ under sparsity [Bogdan et al. (2004)]

$$mBIC = -2 \log L_M + k_M \log(np^2 + d)$$

In a particular sense controlling FWE (related to Bonferroni)

FDR - controlling model selection criterion

$$mBIC2 = -2 \log L_M + k_M \log(np^2 + d) - 2 \log k_m!$$

Adaptivity to level of sparsity [Abramovich et al. (2006)]

Modifications of BIC

$$BIC = -2 \log L_M + k_M \log n$$

For situation $p > n$ under sparsity [Bogdan et al. (2004)]

$$mBIC = -2 \log L_M + k_M \log(np^2 + d)$$

In a particular sense controlling FWE (related to Bonferroni)

FDR - controlling model selection criterion

$$mBIC2 = -2 \log L_M + k_M \log(np^2 + d) - 2 \log k_m!$$

Adaptivity to level of sparsity [Abramovich et al. (2006)]

Theoretical papers

ABOS: Asymptotic Bayes optimality under sparsity

Multiple Testing, normal mixtures

M. Bogdan, A. Chakrabarti, F. Frommlet, J.K. Ghosh.
Bayes oracle and asymptotic optimality of multiple testing procedures under sparsity. Arxiv 1002.3501

General priors, model selection

Florian Frommlet, Malgorzata Bogdan, Arijit Chakrabarti
Asymptotic Bayes optimality under sparsity of selection rules for general priors. Arxiv 1005.4753

Theoretical papers

ABOS: Asymptotic Bayes optimality under sparsity

Multiple Testing, normal mixtures

M. Bogdan, A. Chakrabarti, F. Frommlet, J.K. Ghosh.
Bayes oracle and asymptotic optimality of multiple testing procedures under sparsity. Arxiv 1002.3501

General priors, model selection

Florian Frommlet, Malgorzata Bogdan, Arijit Chakrabarti
Asymptotic Bayes optimality under sparsity of selection rules for general priors. Arxiv 1005.4753

Simulation scenario

Population reference sample POPRES from dbGaP

- 309790 SNPs for 649 individuals of European ancestry
- k = 40 SNPs selected to be causal
  MAF between 0.3 and 0.5,
  pairwise correlation between -0.12 and 0.1
- Simulation of 1000 replicates from additive model M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim \mathcal{N}(0,1)$

Two scenarios

1. effect size for all SNPs constant at $\beta_j = 0.5$
2. β_j equally distributed between 0.27 and 0.66

Simulation scenario

Population reference sample POPRES from dbGaP

- 309790 SNPs for 649 individuals of European ancestry
- k = 40 SNPs selected to be causal
  MAF between 0.3 and 0.5,
  pairwise correlation between -0.12 and 0.1
- Simulation of 1000 replicates from additive model M
  $Y = X_M\beta_M + \epsilon, \qquad \epsilon_i \sim \mathcal{N}(0,1)$

Two scenarios

1. effect size for all SNPs constant at $\beta_j = 0.5$
2. β_j equally distributed between 0.27 and 0.66

Simulation scenario

Population reference sample POPRES from dbGaP

- 309790 SNPs for 649 individuals of European ancestry
- k = 40 SNPs selected to be causal
  MAF between 0.3 and 0.5,
  pairwise correlation between -0.12 and 0.1
- Simulation of 1000 replicates from additive model M
  $Y = X_M \beta_M + \epsilon, \qquad \epsilon_i \sim \mathcal{N}(0,1)$

Two scenarios

1. effect size for all SNPs constant at $\beta_j = 0.5$
2. β_j equally distributed between 0.27 and 0.66

Simulation scenario

Population reference sample POPRES from dbGaP

- 309790 SNPs for 649 individuals of European ancestry
- k = 40 SNPs selected to be causal
  MAF between 0.3 and 0.5,
  pairwise correlation between -0.12 and 0.1
- Simulation of 1000 replicates from additive model M
  $Y = X_M\beta_M + \epsilon, \qquad \epsilon_i \sim \mathcal{N}(0,1)$

Two scenarios

1. effect size for all SNPs constant at $\beta_j = 0.5$
2. β_j equally distributed between 0.27 and 0.66

Heritability

Overall heritability is defined as

$$H^2 = \frac{\text{Var}\left(X_M \beta_M\right)}{1 + \text{Var}\left(X_M \beta_M\right)}$$

Heritability of an individual effect defined as

$$h_j^2 = \frac{\beta_j^2 \text{Var}\left(X_j\right)}{1 + \text{Var}\left(X_M \beta_M\right)} \,,$$

Scenario 1

Overall heritability: $H^2 \approx 0.82$.
Individual effect: $h_j^2 \sim 0.022$.

Scenario 2

Overall heritability: $H^2 \approx 0.81$.
Individual effect: h_j^2 ranging from 0.006 till 0.037

Heritability

Overall heritability is defined as

$$H^2 = \frac{\text{Var}\,(X_M\beta_M)}{1 + \text{Var}\,(X_M\beta_M)}$$

Heritability of an individual effect defined as

$$h_j^2 = \frac{\beta_j^2 \text{Var}\,(X_j)}{1 + \text{Var}\,(X_M\beta_M)}\ ,$$

Scenario 1

Overall heritability: $H^2 \approx 0.82$.
Individual effect: $h_j^2 \sim 0.022$.

Scenario 2

Overall heritability: $H^2 \approx 0.81$.
Individual effect: h_j^2 ranging from 0.006 till 0.037

Heritability

Overall heritability is defined as

$$H^2 = \frac{\text{Var}\,(X_M\beta_M)}{1 + \text{Var}\,(X_M\beta_M)}$$

Heritability of an individual effect defined as

$$h_j^2 = \frac{\beta_j^2 \text{Var}\,(X_j)}{1 + \text{Var}\,(X_M\beta_M)}\ ,$$

Scenario 1

Overall heritability: $H^2 \approx 0.82$.
Individual effect: $h_j^2 \sim 0.022$.

Scenario 2

Overall heritability: $H^2 \approx 0.81$.
Individual effect: h_j^2 ranging from 0.006 till 0.037

Heritability

Overall heritability is defined as

$$H^2 = \frac{\text{Var}\left(X_M\beta_M\right)}{1 + \text{Var}\left(X_M\beta_M\right)}$$

Heritability of an individual effect defined as

$$h_j^2 = \frac{\beta_j^2 \text{Var}\left(X_j\right)}{1 + \text{Var}\left(X_M\beta_M\right)} \ ,$$

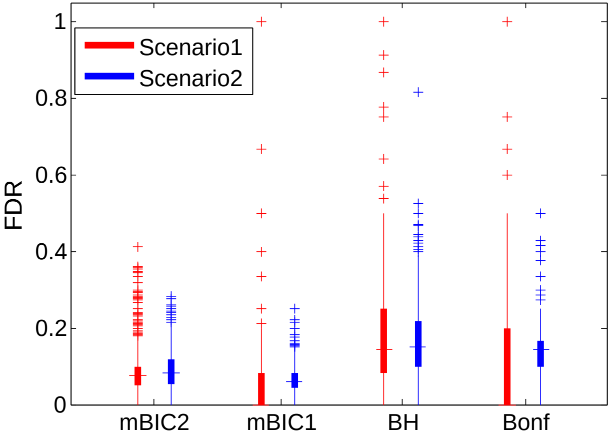
Scenario 1

Overall heritability: $H^2 \approx 0.82$.
Individual effect: $h_j^2 \sim 0.022$.

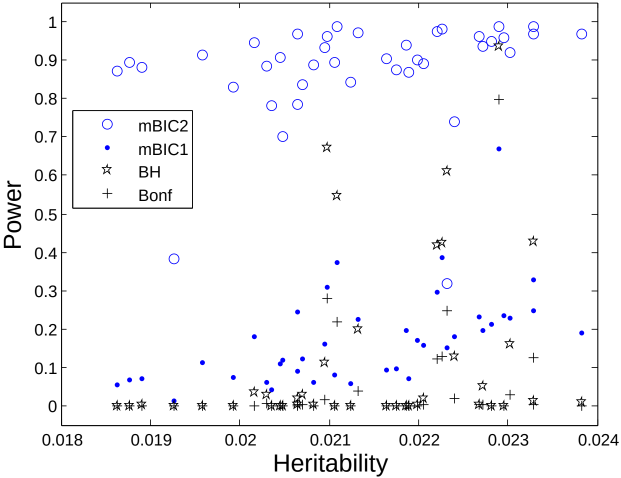
Scenario 2

Overall heritability: $H^2 \approx 0.81$.
Individual effect: h_j^2 ranging from 0.006 till 0.037

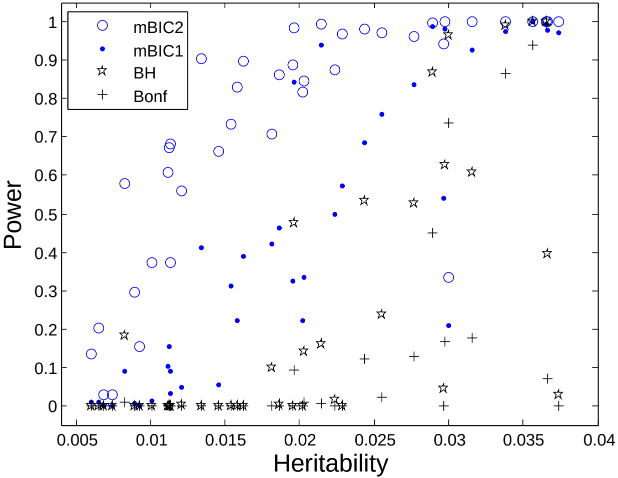
FDR for both Scenarios



Power for Scenario 1



Power for Scenario 2



Important conclusions

Power

Model selection has larger power than multiple testing procedures.
In general both mBIC2 and mBIC are performing much better than multiple testing procedures

Heritability

Power of model selection procedures quite erratic in terms of individual heritability
This observation extremely important!
Order of p-values not necessarily corresponds with order of importance of a SNP for the trait

Important conclusions

Power

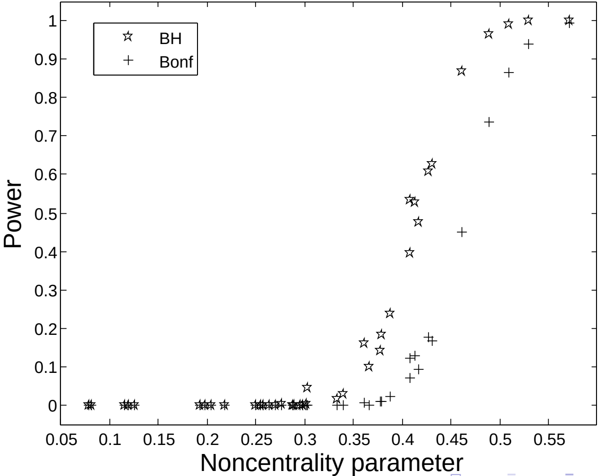
Model selection has larger power than multiple testing procedures.
In general both mBIC2 and mBIC are performing much better than multiple testing procedures

Heritability

Power of model selection procedures quite erratic in terms of individual heritability
This observation extremely important!
Order of p-values not necessarily corresponds with order of importance of a SNP for the trait

Power for Scenario 2

Ordered by noncentrality parameter $\frac{\left(\sum_{l=1}^{k} \beta_l Cov(x_j, x_l)\right)^2}{\sigma^2 Var(x_j)}$



15 most frequent false positives

mBIC2			BH		
SNP	freq	corr	SNP	freq	corr
'243410'	668	0.8958	'243410'	708	0.8958
'182913'	203	0.7728	'188154'	182	0.2628
'119266'	105	0.8416	'119266'	78	0.8416
'125713'	85	0.8311	'125713'	74	0.8311
'4613'	82	0.7683	'255836'	71	0.8351
'271397'	80	0.8162	'221042'	70	0.1116
'145745'	63	0.7230	'291932'	64	0.6255
'291932'	54	0.6255	'181596'	55	0.0970
'150321'	50	0.7659	'27741'	40	0.1137
'301398'	46	0.7669	'267989'	38	0.1008
'255836'	38	0.8351	'264343'	36	0.1007
'106264'	33	0.7277	'27668'	29	0.5742
'11081'	26	0.7187	'227937'	26	0.8372
'227937'	25	0.8372	'11020'	22	0.0896
'243472'	22	0.8954	'283397'	21	0.0875

15 most frequent false positives

mBIC2			BH		
SNP	freq	corr	SNP	freq	corr
'243410'	668	0.8958	'243410'	708	0.8958
'182913'	203	0.7728	'188154'	182	0.2628
'119266'	105	0.8416	'119266'	78	0.8416
'125713'	85	0.8311	'125713'	74	0.8311
'4613'	82	0.7683	'255836'	71	0.8351
'271397'	80	0.8162	'221042'	70	0.1116
'145745'	63	0.7230	'291932'	64	0.6255
'291932'	54	0.6255	'181596'	55	0.0970
'150321'	50	0.7659	'27741'	40	0.1137
'301398'	46	0.7669	'267989'	38	0.1008
'255836'	38	0.8351	'264343'	36	0.1007
'106264'	33	0.7277	'27668'	29	0.5742
'11081'	26	0.7187	'227937'	26	0.8372
'227937'	25	0.8372	'11020'	22	0.0896
'243472'	22	0.8954	'283397'	21	0.0875

15 most frequent false positives

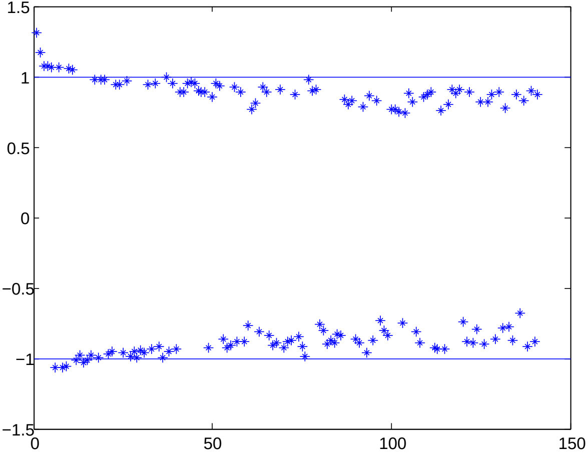
mBIC2			BH		
SNP	freq	corr	SNP	freq	corr
'243410'	668	0.8958	'243410'	708	0.8958
'182913'	203	0.7728	'188154'	182	0.2628
'119266'	105	0.8416	'119266'	78	0.8416
'125713'	85	0.8311	'125713'	74	0.8311
'4613'	82	0.7683	'255836'	71	0.8351
'271397'	80	0.8162	'221042'	70	0.1116
'145745'	63	0.7230	'291932'	64	0.6255
'291932'	54	0.6255	'181596'	55	0.0970
'150321'	50	0.7659	'27741'	40	0.1137
'301398'	46	0.7669	'267989'	38	0.1008
'255836'	38	0.8351	'264343'	36	0.1007
'106264'	33	0.7277	'27668'	29	0.5742
'11081'	26	0.7187	'227937'	26	0.8372
'227937'	25	0.8372	'11020'	22	0.0896
'243472'	22	0.8954	'283397'	21	0.0875

15 most frequent false positives

mBIC2			BH		
SNP	freq	corr	SNP	freq	corr
'243410'	668	0.8958	'243410'	708	0.8958
'182913'	203	0.7728	'188154'	182	0.2628
'119266'	105	0.8416	'119266'	78	0.8416
'125713'	85	0.8311	'125713'	74	0.8311
'4613'	82	0.7683	'255836'	71	0.8351
'271397'	80	0.8162	'221042'	70	0.1116
'145745'	63	0.7230	'291932'	64	0.6255
'291932'	54	0.6255	'181596'	55	0.0970
'150321'	50	0.7659	'27741'	40	0.1137
'301398'	46	0.7669	'267989'	38	0.1008
'255836'	38	0.8351	'264343'	36	0.1007
'106264'	33	0.7277	'27668'	29	0.5742
'11081'	26	0.7187	'227937'	26	0.8372
'227937'	25	0.8372	'11020'	22	0.0896
'243472'	22	0.8954	'283397'	21	0.0875

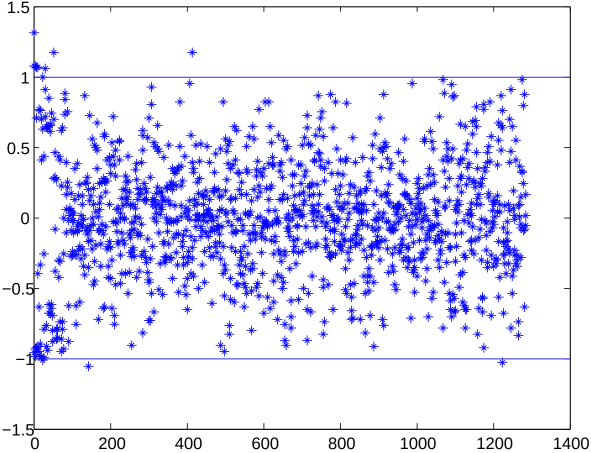
Sum of correlations of FP under BH

Ordered by number of simulations in which SNP occurs as FP



Sum of correlations of FP under mBIC2

Ordered by number of simulations in which SNP occurs as FP



Conclusion

- Problems with multiple testing approach to GWAS when many causal SNPs are influencing traits
  small random correlations of genotypes determine which SNPs are selected
- Possible explanation for ”Missing heritability” in GWAS
- Model selection approach can help
  - much larger power to detect causal SNPs
  - ”False positives” are rather likely to be correlated with causal SNP

Conclusion

- Problems with multiple testing approach to GWAS when many causal SNPs are influencing traits
  small random correlations of genotypes determine which SNPs are selected
- Possible explanation for ”Missing heritability” in GWAS
- Model selection approach can help
  - much larger power to detect causal SNPs
  - ”False positives” are rather likely to be correlated with causal SNP

Conclusion

- Problems with multiple testing approach to GWAS when many causal SNPs are influencing traits
  small random correlations of genotypes determine which SNPs are selected
- Possible explanation for "Missing heritability" in GWAS
- Model selection approach can help
  - much larger power to detect causal SNPs
  - "False positives" are rather likely to be correlated with causal SNP

Conclusion

- Problems with multiple testing approach to GWAS when many causal SNPs are influencing traits
  small random correlations of genotypes determine which SNPs are selected
- Possible explanation for "Missing heritability" in GWAS
- Model selection approach can help
  - much larger power to detect causal SNPs
  - "False positives" are rather likely to be correlated with causal SNP