Model Selection Procedures for Genome Wide Association Studies

Florian Frommlet

Department of Medical Statistics,
Medical University Vienna

Liege, May 2013

Contents

Genetic association studies

General purpose

Detect genomic regions which are associated with some trait
Traits might be e.g. quantitative or dichotomous

Genetic markers

Any observable characteristic with known location on the chromosome which varies between individuals

- Traditional: Genes that encode certain phenotypes
- Today: DNA sequence information
  e.g. SNPs, Copy number variation, etc.

Genetic association studies

General purpose

Detect genomic regions which are associated with some trait
Traits might be e.g. quantitative or dichotomous

Genetic markers

Any observable characteristic with known location on the chromosome which varies between individuals

- Traditional: Genes that encode certain phenotypes
- Today: DNA sequence information
  e.g. SNPs, Copy number variation, etc.

Genetic association studies

General purpose

Detect genomic regions which are associated with some trait
Traits might be e.g. quantitative or dichotomous

Genetic markers

Any observable characteristic with known location on the chromosome which varies between individuals

- Traditional: Genes that encode certain phenotypes
- Today: DNA sequence information
  e.g. SNPs, Copy number variation, etc.

Genetic association studies

General purpose

Detect genomic regions which are associated with some trait
Traits might be e.g. quantitative or dichotomous

Genetic markers

Any observable characteristic with known location on the chromosome which varies between individuals

- Traditional: Genes that encode certain phenotypes
- Today: DNA sequence information
  e.g. SNPs, Copy number variation, etc.

SNPs as genetic markers

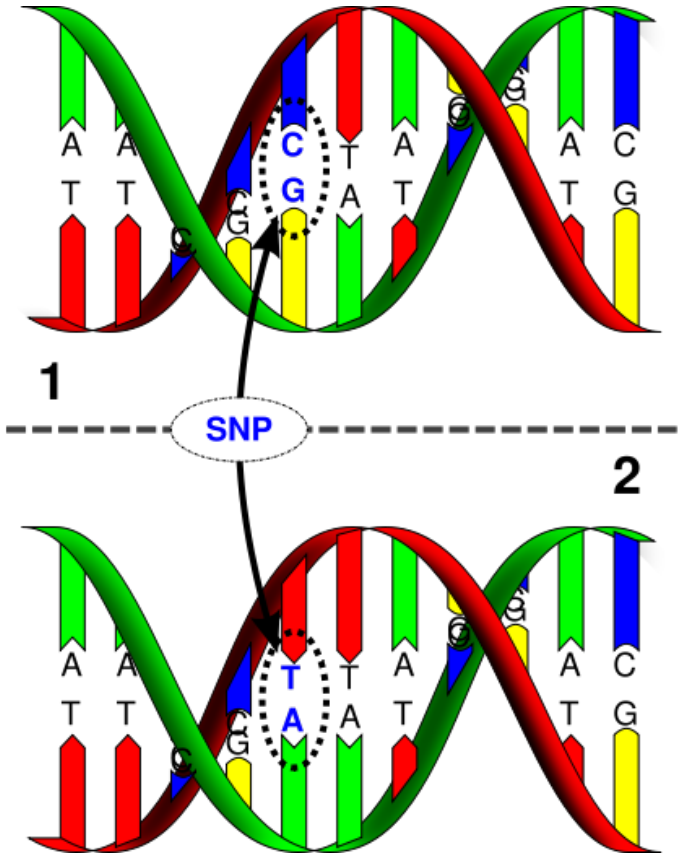
Single Nucleotide Polymorphism

SNP: Point mutation

Humans: Some 20 million SNPs known, figure increasing rapidly

SNP Arrays

Affymetrix 6: ca. 1 million SNPs
Latest Illumina: ca. 5 million SNPs



Wikipedia

SNPs as genetic markers

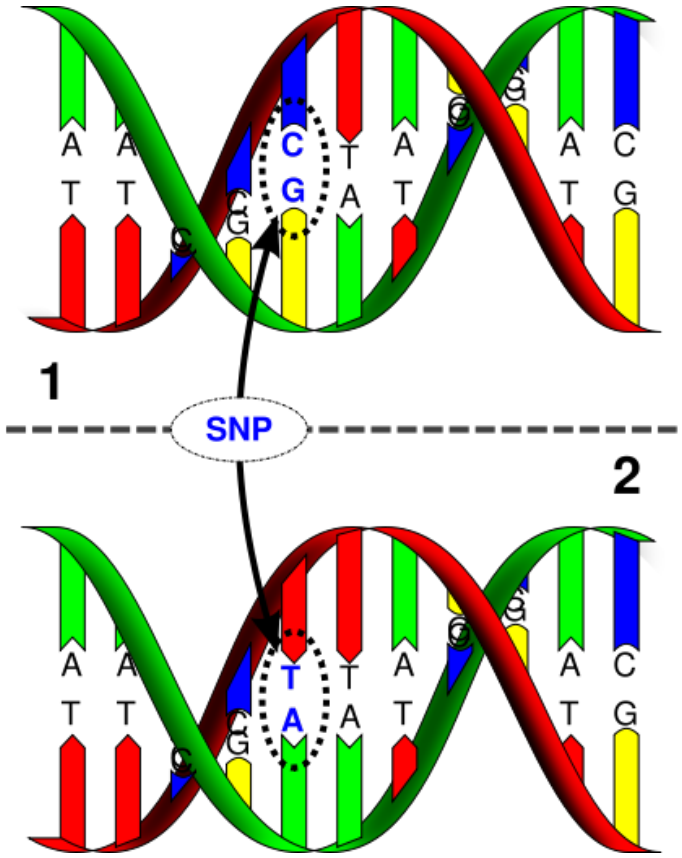
Single Nucleotide Polymorphism

SNP: Point mutation

Humans: Some 20 million SNPs known, figure increasing rapidly

SNP Arrays

Affymetrix 6: ca. 1 million SNPs
Latest Illumina: ca. 5 million SNPs



Wikipedia

Technology

SNP arrays

- Approximately 10 years around
- Similar to RNA Micro-Arrays
- Both variants of SNP on array (say A and a)
  $\Rightarrow$ Different intensities for genotypes AA, Aa, aa

First step

Image segmentation similar to RNA Micro-Arrays

Technology

SNP arrays

- Approximately 10 years around
- Similar to RNA Micro-Arrays
- Both variants of SNP on array (say A and a)
  $\Rightarrow$ Different intensities for genotypes AA, Aa, aa

First step

Image segmentation similar to RNA Micro-Arrays

Technology

SNP arrays

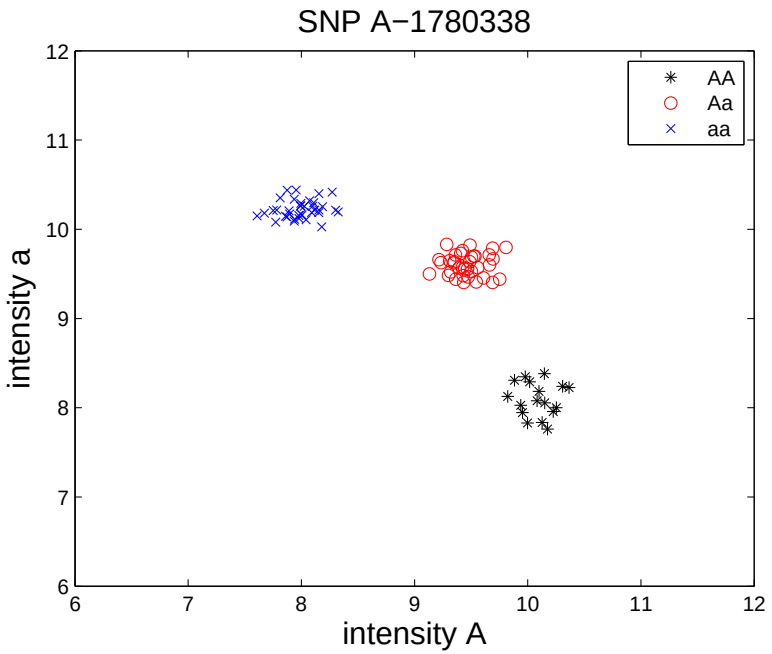
- Approximately 10 years around
- Similar to RNA Micro-Arrays
- Both variants of SNP on array (say A and a)
  $\Rightarrow$ Different intensities for genotypes AA, Aa, aa

First step

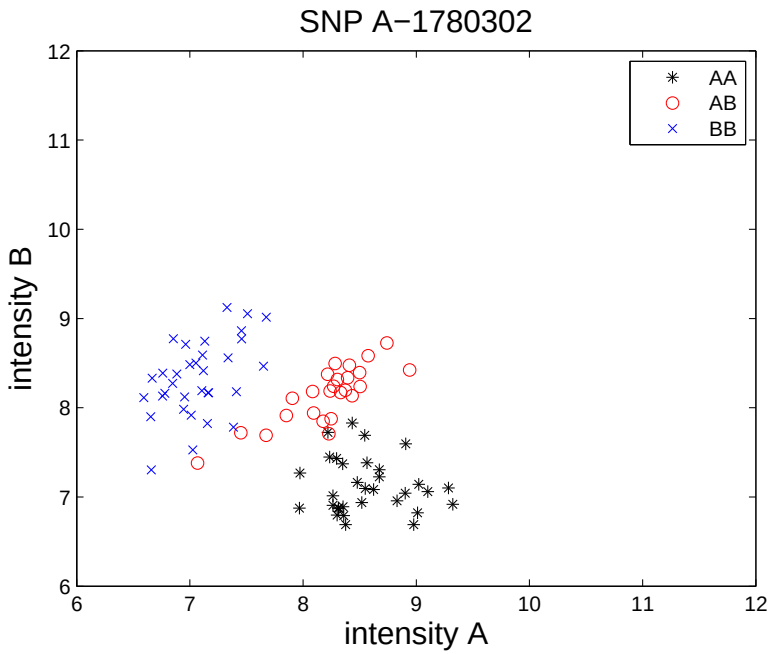
Image segmentation similar to RNA Micro-Arrays

Technology

Good for calling



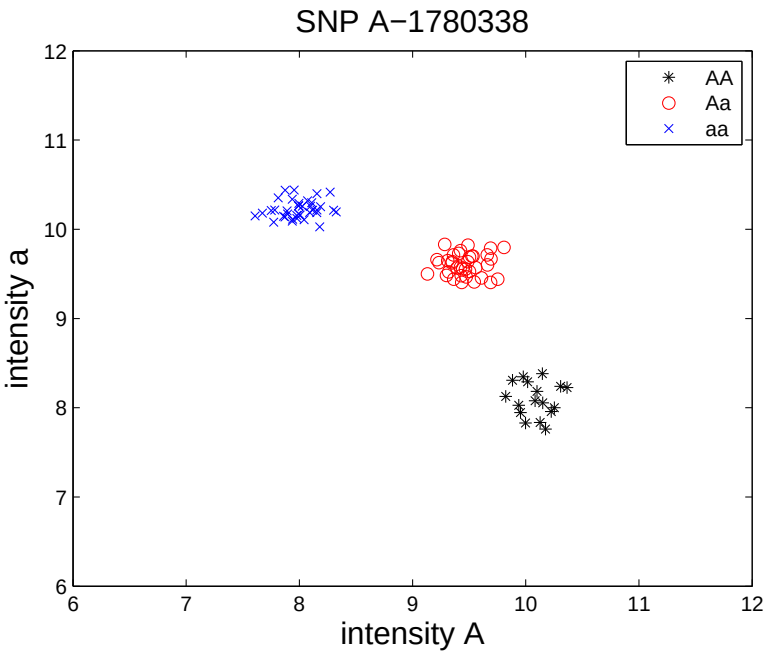
Bad for calling



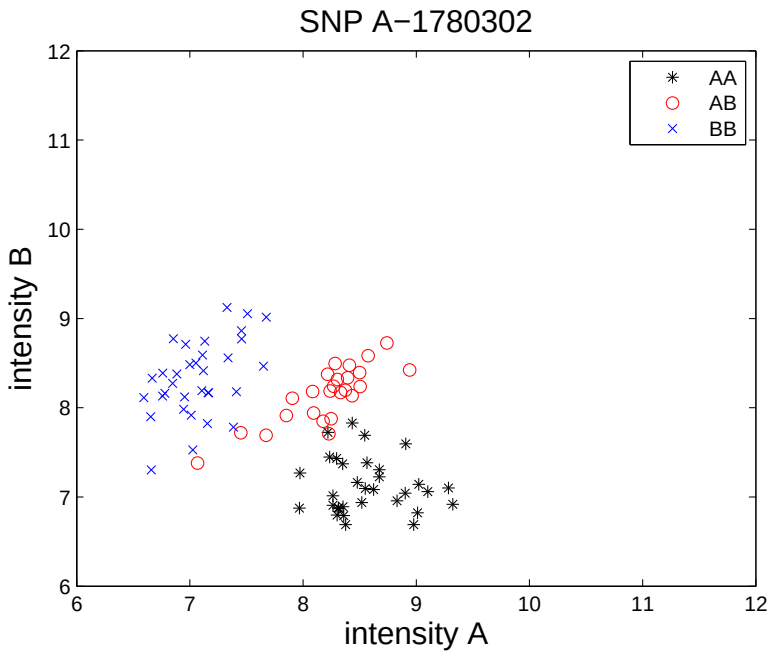
Quality measures are needed

Technology

Good for calling



Bad for calling



Quality measures are needed

Downstream Analysis

$Y \leftarrow X_1, \ldots, X_p$

- Y . . . quantitative (e.g. height) or categorical (e.g. disease status)
- $X_j \in \{-1, 0, 1\}$ for different genotypes
- n observations

Typical: $n > 10^3$, $p > 10^5$

Question:

Which X_j are associated with Y?

State of the art analysis: Single marker tests

- Test statistic for each SNP (ANOVA, χ^2, etc.)
- Multiple testing correction
  (Bonferroni, FDR control, permutation tests, . . .)

Downstream Analysis

$Y \leftarrow X_1, \ldots, X_p$

- Y . . . quantitative (e.g. height) or categorical (e.g. disease status)
- $X_j \in \{-1, 0, 1\}$ for different genotypes
- n observations

Typical: $n > 10^3$, $p > 10^5$

Question:

Which X_j are associated with Y?

State of the art analysis: Single marker tests

- Test statistic for each SNP (ANOVA, χ^2, etc.)
- Multiple testing correction
  (Bonferroni, FDR control, permutation tests, . . .)

Downstream Analysis

$Y \leftarrow X_1, \ldots, X_p$

- Y . . . quantitative (e.g. height) or categorical (e.g. disease status)
- $X_j \in \{-1, 0, 1\}$ for different genotypes
- n observations

Typical: $n > 10^3$, $p > 10^5$

Question:

Which X_j are associated with Y?

State of the art analysis: Single marker tests

- Test statistic for each SNP (ANOVA, χ^2, etc.)
- Multiple testing correction
  (Bonferroni, FDR control, permutation tests, . . .)

Alternative approach: Model selection

Model

Index vector $M = [j_1, \ldots, j_{k_M}]$

Quantitative Trait: Linear regression

$$\mathcal{M}: \; Y = X_M \beta_M + \epsilon, \quad X_M = [X_{j_1}, \ldots, X_{j_{k_M}}]$$

Columns of design matrix X_j:

- $(-1, 0, 1)$ additive effects
- $(1, 0, 1)$ dominance effects

Case control studies: Logistic regression

Two fundamental questions

1. How to evaluate what is a good model?
2. How to find a good model?

Alternative approach: Model selection

Model

Index vector $M = [j_1, \ldots, j_{k_M}]$

Quantitative Trait: Linear regression

$$\mathcal{M}: \quad Y = X_M \beta_M + \epsilon, \quad X_M = [X_{j_1}, \ldots, X_{j_{k_M}}]$$

Columns of design matrix X_j:

- $(-1, 0, 1)$ additive effects
- $(1, 0, 1)$ dominance effects

Case control studies: Logistic regression

Two fundamental questions

1. How to evaluate what is a good model?
2. How to find a good model?

Alternative approach: Model selection

Model

Index vector $M = [j_1, \ldots, j_{k_M}]$

Quantitative Trait: Linear regression

$$\mathcal{M}: \quad Y = X_M \beta_M + \epsilon, \quad X_M = [X_{j_1}, \ldots, X_{j_{k_M}}]$$

Columns of design matrix X_j:

- $(-1, 0, 1)$ additive effects
- $(1, 0, 1)$ dominance effects

Case control studies: Logistic regression

Two fundamental questions

1. How to evaluate what is a good model?
2. How to find a good model?

Alternative approach: Model selection

Model

Index vector $M = [j_1, \ldots, j_{k_M}]$

Quantitative Trait: Linear regression

$$\mathcal{M}: \quad Y = X_M \beta_M + \epsilon, \quad X_M = [X_{j_1}, \ldots, X_{j_{k_M}}]$$

Columns of design matrix X_j:

- $(-1, 0, 1)$ additive effects
- $(1, 0, 1)$ dominance effects

Case control studies: Logistic regression

Two fundamental questions

1. How to evaluate what is a good model?
2. How to find a good model?

Model selection criteria

Classical

Maximum likelihood L_M with penalties based on model size

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

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

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

More recent

LASSO: L_1- Penalty
Elastic Net: L_1 and L_2- Penalty
etc.

Model selection criteria

Classical

Maximum likelihood L_M with penalties based on model size

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

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

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

More recent

LASSO: L_1- Penalty
Elastic Net: L_1 and L_2- Penalty
etc.

Model selection criteria

Classical

Maximum likelihood L_M with penalties based on model size

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

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

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

More recent

LASSO: L_1- Penalty
Elastic Net: L_1 and L_2- Penalty
etc.

Model selection for $p > n$

Classical theory for AIC and BIC

Derived for constant p, while $n \to \infty$

Results for $p > n$ no longer correct
e.g. BIC no longer consistent

Problem

In case of sparsity and $p > n$ BIC chooses too large models

Model selection for $p > n$

Classical theory for AIC and BIC

Derived for constant p, while $n \to \infty$

Results for $p > n$ no longer correct
e.g. BIC no longer consistent

Problem

In case of sparsity and $p > n$ BIC chooses too large models

Schwarz BIC in case of sparsity

Source of problem

BIC derived in Bayesian context

$$P(M|Y) = \frac{P(Y|M)\pi(M)}{P(Y)}$$

BIC ignores model prior $\pi(M)$, i.e. equivalent with uniform prior for all models $\Rightarrow$ **informative prior for model size**

e.g. p models of size 1, $\binom{p}{p/2}$ models of size $p/2$

If one expects only few causal SNPs

$\Rightarrow$ BIC selects too large models

Solution

Use model prior $\pi(M)$ which takes into account p

Schwarz BIC in case of sparsity

Source of problem

BIC derived in Bayesian context

$$P(M|Y) = \frac{P(Y|M)\pi(M)}{P(Y)}$$

BIC ignores model prior $\pi(M)$, i.e. equivalent with uniform prior for all models $\Rightarrow$ **informative prior for model size**

e.g. p models of size 1, $\binom{p}{p/2}$ models of size $p/2$

If one expects only few causal SNPs

$\Rightarrow$ BIC selects too large models

Solution

Use model prior $\pi(M)$ which takes into account p

Schwarz BIC in case of sparsity

Source of problem

BIC derived in Bayesian context

$$P(M|Y) = \frac{P(Y|M)\pi(M)}{P(Y)}$$

BIC ignores model prior $\pi(M)$, i.e. equivalent with uniform prior for all models $\Rightarrow$ **informative prior for model size**

e.g. p models of size 1, $\binom{p}{p/2}$ models of size $p/2$

If one expects only few causal SNPs
$\Rightarrow$ BIC selects too large models

Solution

Use model prior $\pi(M)$ which takes into account p

Schwarz BIC in case of sparsity

Source of problem

BIC derived in Bayesian context

$$P(M|Y) = \frac{P(Y|M)\pi(M)}{P(Y)}$$

BIC ignores model prior $\pi(M)$, i.e. equivalent with uniform prior for all models $\Rightarrow$ **informative prior for model size**

e.g. p models of size 1, $\binom{p}{p/2}$ models of size $p/2$

If one expects only few causal SNPs
$\Rightarrow$ BIC selects too large models

Solution

Use model prior $\pi(M)$ which takes into account p

First modification of BIC

Original BIC [Schwarz (1978)]

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

mBIC [Bogdan et al. (2004)]

Model prior $\pi(M) = \omega^{k_M} \cdot (1-\omega)^{p-k_M}$ yields

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

Properties

- $\omega \ldots$ Prior probability of causal SNPs $\Rightarrow$ defines d
  Recommendation if no prior information: $d = -2 \log 4$
- Orthogonal design $\Rightarrow$ mBIC controls FWER (closely related to Bonferroni correction)

First modification of BIC

Original BIC [Schwarz (1978)]

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

mBIC [Bogdan et al. (2004)]

Model prior $\pi(M) = \omega^{k_M} \cdot (1-\omega)^{p-k_M}$ yields

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

Properties

- ω ... Prior probability of causal SNPs $\Rightarrow$ defines d
  Recommendation if no prior information: $d = -2 \log 4$
- Orthogonal design $\Rightarrow$ mBIC controls FWER (closely related to Bonferroni correction)

First modification of BIC

Original BIC [Schwarz (1978)]

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

mBIC [Bogdan et al. (2004)]

Model prior $\pi(M) = \omega^{k_M} \cdot (1-\omega)^{p-k_M}$ yields

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

Properties

- $\omega \ldots$ Prior probability of causal SNPs $\Rightarrow$ defines d
  Recommendation if no prior information: $d = -2 \log 4$
- Orthogonal design $\Rightarrow$ mBIC controls FWER (closely related to Bonferroni correction)

First modification of BIC

Original BIC [Schwarz (1978)]

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

mBIC [Bogdan et al. (2004)]

Model prior $\pi(M) = \omega^{k_M} \cdot (1-\omega)^{p-k_M}$ yields

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

Properties

- $\omega \ldots$ Prior probability of causal SNPs $\Rightarrow$ defines d
  Recommendation if no prior information: $d = -2 \log 4$
- Orthogonal design $\Rightarrow$ mBIC controls FWER (closely related to Bonferroni correction)

FDR-controlling modifications of BIC

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

mBIC2 [Frommlet et al. (2011)]

Model selection criterion which under orthogonality controls FDR

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

Properties

- Penalisation based on ideas of [Abramovich et al. (2006)]
- mBIC2 has certain optimality properties (as we will see)

FDR-controlling modifications of BIC

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

mBIC2 [Frommlet et al. (2011)]

Model selection criterion which under orthogonality controls FDR

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

Properties

- Penalisation based on ideas of [Abramovich et al. (2006)]
- mBIC2 has certain optimality properties (as we will see)

FDR-controlling modifications of BIC

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

mBIC2 [Frommlet et al. (2011)]

Model selection criterion which under orthogonality controls FDR

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

Properties

- Penalisation based on ideas of [Abramovich et al. (2006)]
- mBIC2 has certain optimality properties (as we will see)

FDR-controlling modifications of BIC

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

mBIC2 [Frommlet et al. (2011)]

Model selection criterion which under orthogonality controls FDR

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

Properties

- Penalisation based on ideas of [Abramovich et al. (2006)]
- mBIC2 has certain optimality properties (as we will see)

Ideas underlying mBIC2

Penalizing scheme by [Abramovich et al. (2006)]

$$\frac{RSS_M}{\sigma^2} + \sum_{i=1}^{k_M} q_N^2(\alpha i/2p) \tag{1}$$

with $q_N(\alpha)$ the $(1-\alpha)$ - quantile of standard normal

- Benjamini Hochberg (BH) corresponds to largest local minimum of (1)
- Corresponding step down procedure corresponds to smallest local minimum of (1)
- Approximation of penalty term using $\alpha = n^{-1/2}$ yields mBIC2

Ideas underlying mBIC2

Penalizing scheme by [Abramovich et al. (2006)]

$$\frac{RSS_M}{\sigma^2} + \sum_{i=1}^{k_M} q_N^2(\alpha i/2p) \tag{1}$$

with $q_N(\alpha)$ the $(1-\alpha)$ - quantile of standard normal

- Benjamini Hochberg (BH) corresponds to largest local minimum of (1)
- Corresponding step down procedure corresponds to smallest local minimum of (1)
- Approximation of penalty term using $\alpha = n^{-1/2}$ yields mBIC2

Ideas underlying mBIC2

Penalizing scheme by [Abramovich et al. (2006)]

$$\frac{RSS_M}{\sigma^2} + \sum_{i=1}^{k_M} q_N^2(\alpha i/2p) \tag{1}$$

with $q_N(\alpha)$ the $(1-\alpha)$ - quantile of standard normal

- Benjamini Hochberg (BH) corresponds to largest local minimum of (1)
- Corresponding step down procedure corresponds to smallest local minimum of (1)
- Approximation of penalty term using $\alpha = n^{-1/2}$ yields mBIC2

Optimality properties of mBIC2

Asymptotic Bayes optimality under sparsity (ABOS)

Topic of Bogdan et al. (2011), Frommlet et al. (2013) for multiple testing

Essential idea for regression

- Two groups model for regressors:

$$P(\beta_i \neq 0) = \eta, \quad \text{with} \quad \eta \text{ small}$$

  while n and p are large
- Compare misclassification rate of procedure with optimal Bayes rule

Optimality properties of mBIC2

Asymptotic Bayes optimality under sparsity (ABOS)

Topic of Bogdan et al. (2011), Frommlet et al. (2013) for multiple testing

Essential idea for regression

- Two groups model for regressors:

$$P(\beta_i \neq 0) = \eta, \quad \text{with} \quad \eta \text{ small}$$

  while n and p are large

- Compare misclassification rate of procedure with optimal Bayes rule

Optimality properties of mBIC2

Asymptotic Bayes optimality under sparsity (ABOS)

Topic of Bogdan et al. (2011), Frommlet et al. (2013) for multiple testing

Essential idea for regression

- Two groups model for regressors:

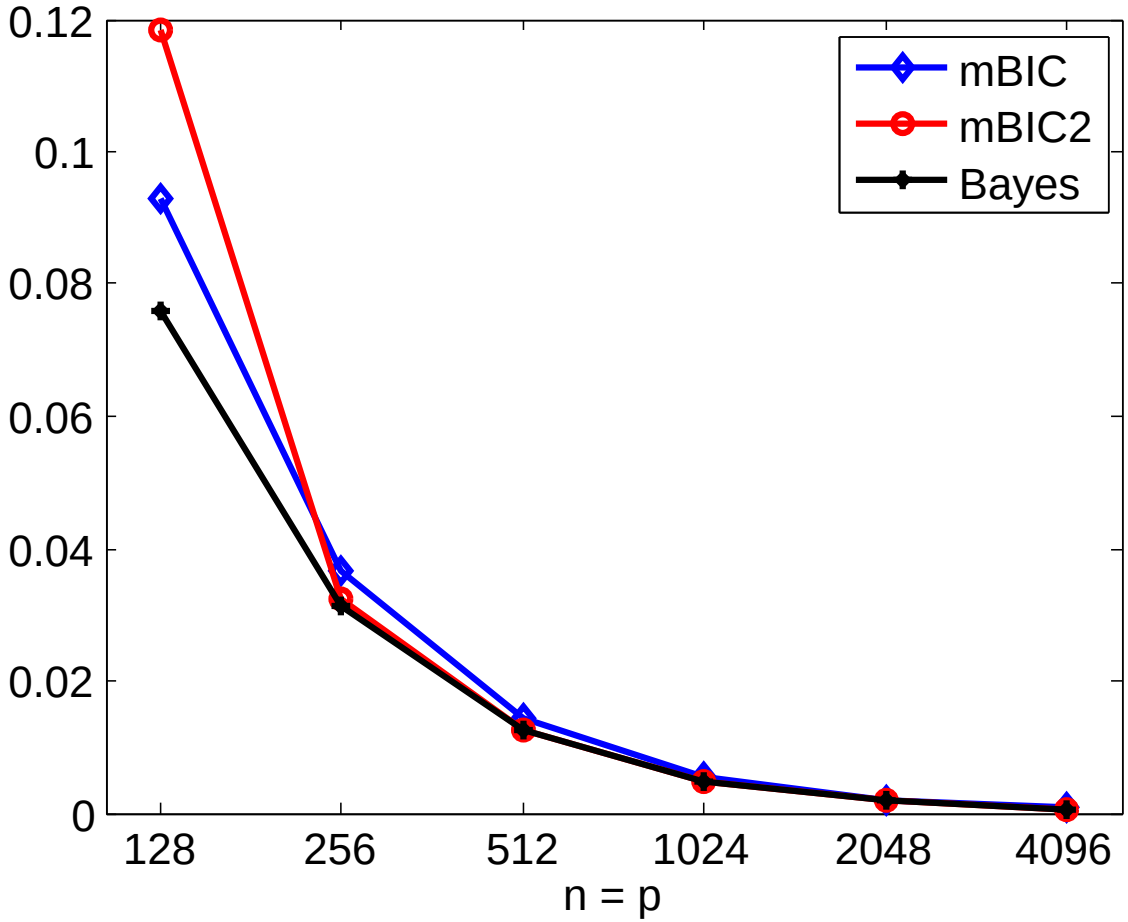
$$P(\beta_i \neq 0) = \eta, \quad \text{with} \quad \eta \text{ small}$$

  while n and p are large
- Compare misclassification rate of procedure with optimal Bayes rule

Simulation for unknown σ, $\eta \propto p^{-1}$

Orthogonal design, $p = n$

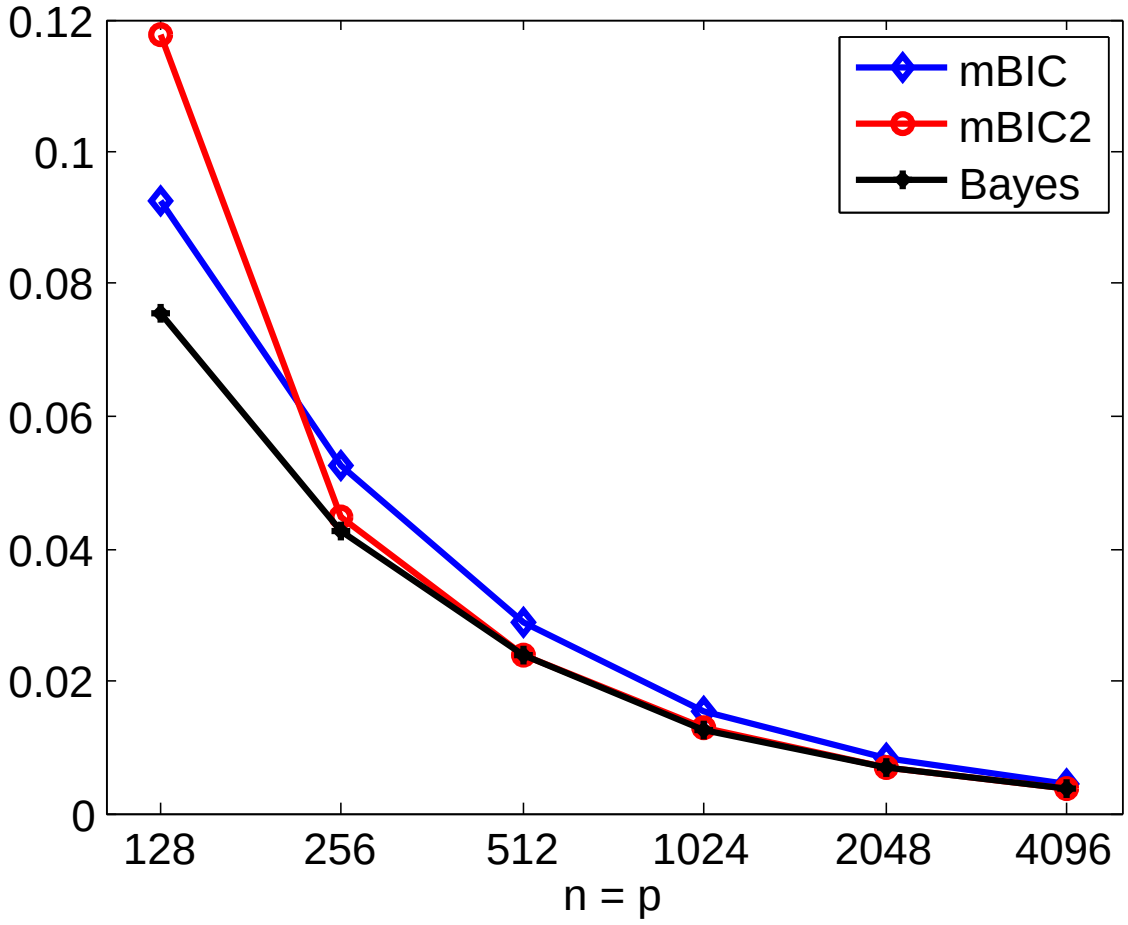
Misclassification rate as a function of p, $\eta(128) = 0.125$



Simulation for unknown σ, $\eta \propto p^{-1/2}$

Orthogonal design, $p = n$

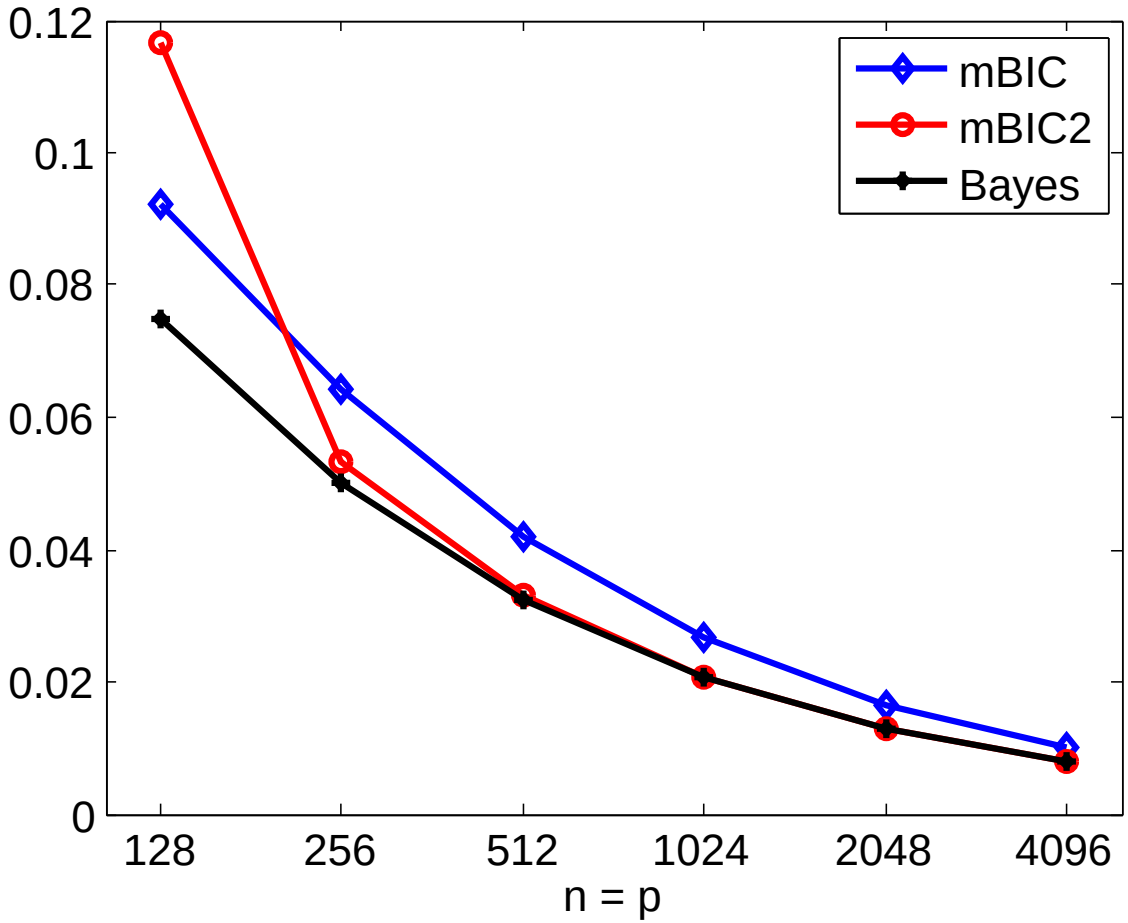
Misclassification rate as a function of p, $\eta(128) = 0.125$



Simulation for unknown σ, $\eta \propto p^{-1/4}$

Orthogonal design, $p = n$

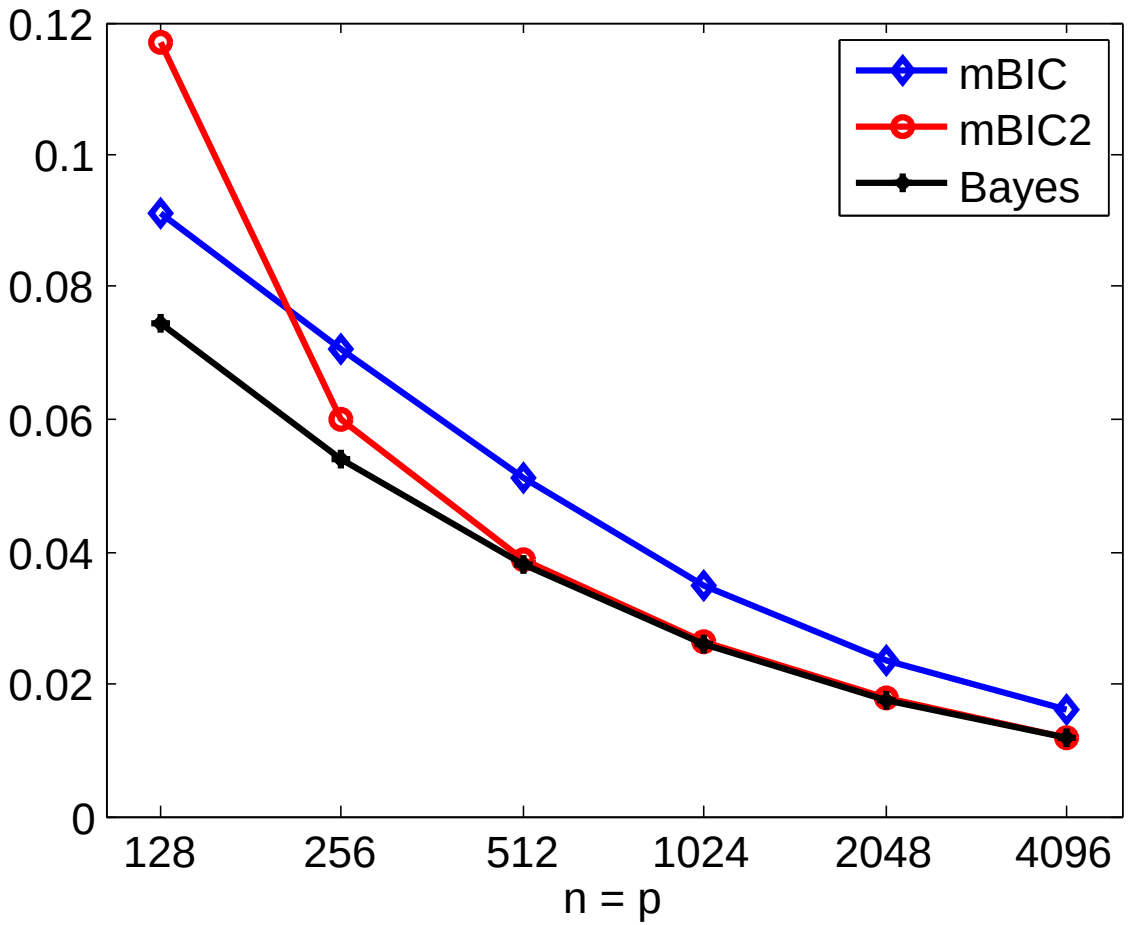
Misclassification rate as a function of p, $\eta(128) = 0.125$



Simulation for unknown σ, $\eta \propto p^{-1/8}$

Orthogonal design, $p = n$

Misclassification rate as a function of p, $\eta(128) = 0.125$



Closely related criterion: extended BIC

Chen and Chen (2008): Prior $\binom{p}{k_M}^{\kappa-1}$

$$EBIC = -2 \log L_M + k_M \log n + 2 \log \binom{p}{k_M}^{1-\kappa}$$

with $0 \leq \kappa \leq 1$.

- $\kappa = 1 \quad \Rightarrow \quad$ original BIC
- $\kappa = 0 \quad \Rightarrow \quad$ asymptotically equivalent with mBIC2

Chen and Chen (2008): Consistency results for EBIC

- Under certain assumptions on design matrix for non-orthogonal case
- Similar consistency results hold for mBIC2

Closely related criterion: extended BIC

Chen and Chen (2008): Prior $\binom{p}{k_M}^{\kappa-1}$

$$EBIC = -2 \log L_M + k_M \log n + 2 \log \binom{p}{k_M}^{1-\kappa}$$

with $0 \leq \kappa \leq 1$.

- $\kappa = 1 \quad \Rightarrow \quad$ original BIC
- $\kappa = 0 \quad \Rightarrow \quad$ asymptotically equivalent with mBIC2

Chen and Chen (2008): Consistency results for EBIC

- Under certain assumptions on design matrix for non-orthogonal case
- Similar consistency results hold for mBIC2

Closely related criterion: extended BIC

Chen and Chen (2008): Prior $\binom{p}{k_M}^{\kappa-1}$

$$EBIC = -2 \log L_M + k_M \log n + 2 \log \binom{p}{k_M}^{1-\kappa}$$

with $0 \leq \kappa \leq 1$.

- $\kappa = 1 \quad \Rightarrow \quad$ original BIC
- $\kappa = 0 \quad \Rightarrow \quad$ asymptotically equivalent with mBIC2

Chen and Chen (2008): Consistency results for EBIC

- Under certain assumptions on design matrix for non-orthogonal case
- Similar consistency results hold for mBIC2

Closely related criterion: extended BIC

Chen and Chen (2008): Prior $\binom{p}{k_M}^{\kappa-1}$

$$EBIC = -2 \log L_M + k_M \log n + 2 \log \binom{p}{k_M}^{1-\kappa}$$

with $0 \leq \kappa \leq 1$.

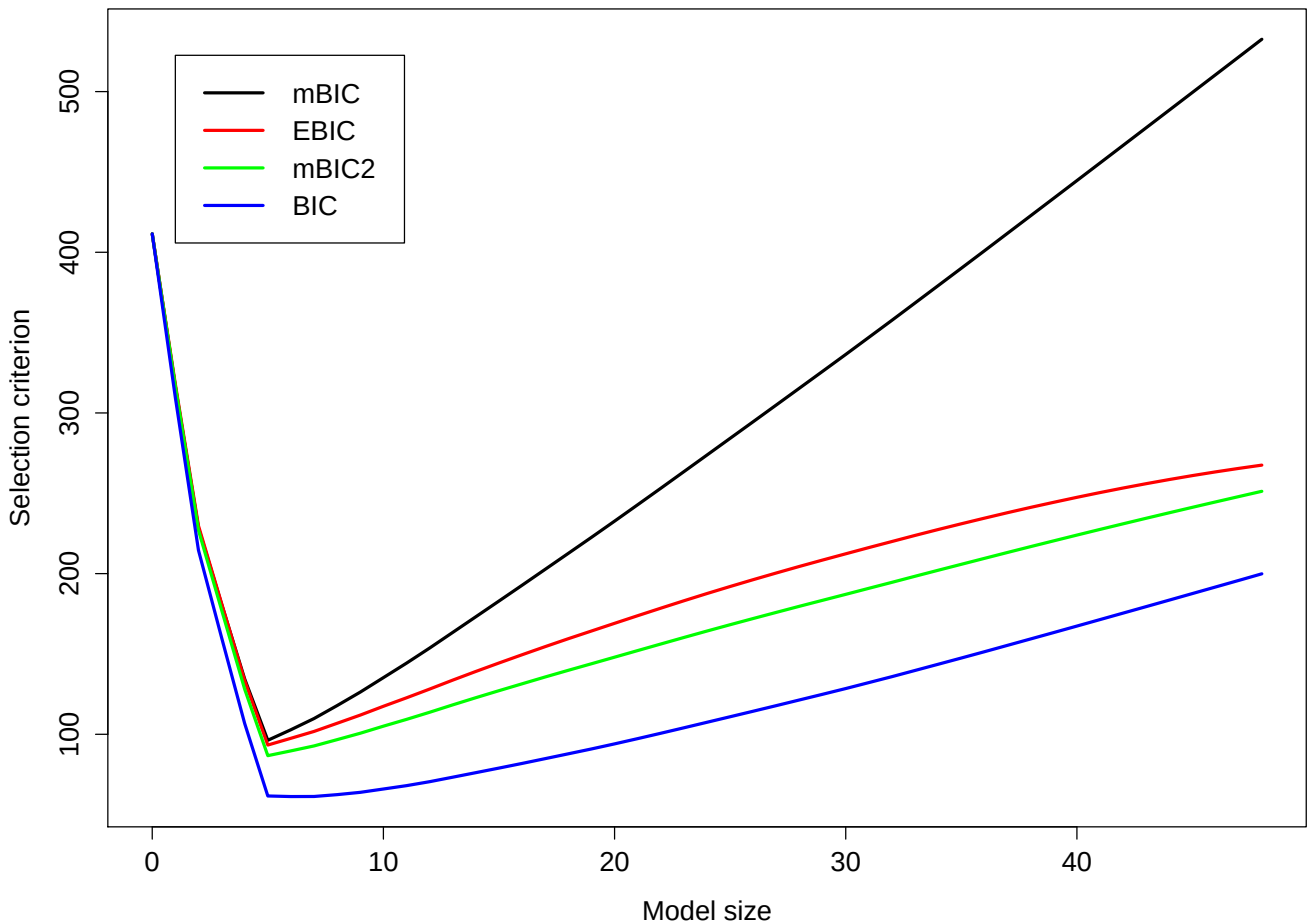
- $\kappa = 1 \quad \Rightarrow \quad$ original BIC
- $\kappa = 0 \quad \Rightarrow \quad$ asymptotically equivalent with mBIC2

Chen and Chen (2008): Consistency results for EBIC

- Under certain assumptions on design matrix for non-orthogonal case
- Similar consistency results hold for mBIC2

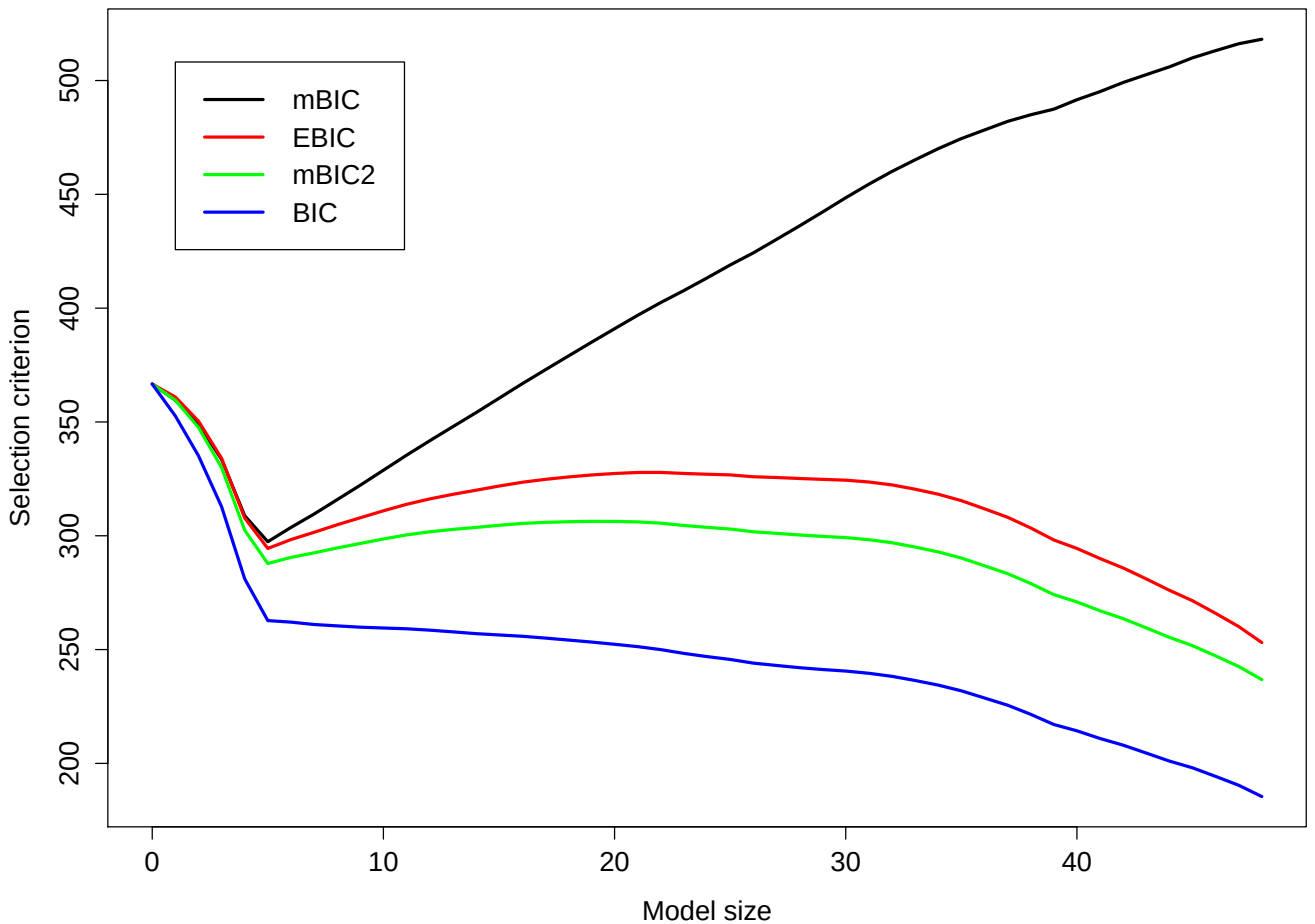
Comparison of criteria for known σ

Orthogonal design, $p = n = 64$



Comparison of criteria for unknown σ

Orthogonal design, $p = n = 64$



Simulation study for GWAS

Frommlet et al. (2011 b)

Real SNP Data: POPRES from dbGaP

- 309790 SNPs for 649 individuals (Caucasians)
- k = 40 **causal SNPs** chosen such that
  MAF between 0.3 and 0.5
  Correlation between -0.12 and 0.1
- Simulation of quantitative trait under additive model M

$$Y_i = \sum_{j=1}^{40} \beta_j X_{ij} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1)$$

  β_j equally spaced between 0.27 and 0.66

- 1000 simulation runs

Simulation study for GWAS

Frommlet et al. (2011 b)

Real SNP Data: POPRES from dbGaP

- 309790 SNPs for 649 individuals (Caucasians)
- k = 40 **causal SNPs** chosen such that
  MAF between 0.3 and 0.5
  Correlation between -0.12 and 0.1
- Simulation of quantitative trait under additive model M

$$Y_i = \sum_{j=1}^{40} \beta_j X_{ij} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1)$$

  β_j equally spaced between 0.27 and 0.66

- 1000 simulation runs

Simulation study for GWAS

Frommlet et al. (2011 b)

Real SNP Data: POPRES from dbGaP

- 309790 SNPs for 649 individuals (Caucasians)
- k = 40 **causal SNPs** chosen such that
  MAF between 0.3 and 0.5
  Correlation between -0.12 and 0.1
- Simulation of quantitative trait under additive model M

$$Y_i = \sum_{j=1}^{40} \beta_j X_{ij} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1)$$

β_j equally spaced between 0.27 and 0.66

- 1000 simulation runs

Simulation study for GWAS

Frommlet et al. (2011 b)

Real SNP Data: POPRES from dbGaP

- 309790 SNPs for 649 individuals (Caucasians)
- k = 40 **causal SNPs** chosen such that
  MAF between 0.3 and 0.5
  Correlation between -0.12 and 0.1
- Simulation of quantitative trait under additive model M

$$Y_i = \sum_{j=1}^{40} \beta_j X_{ij} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,1)$$

β_j equally spaced between 0.27 and 0.66

- 1000 simulation runs

Heritability

Total heritability:

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

Individual heritability:

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

Values in our simulation study

Total heritability: $H^2 \approx 0.81$.
Individual heritability: h_j^2 between 0.006 and 0.037

Heritability

Total heritability:

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

Individual heritability:

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

Values in our simulation study

Total heritability: $H^2 \approx 0.81$.
Individual heritability: h_j^2 between 0.006 and 0.037

Heritability

Total heritability:

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

Individual heritability:

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

Values in our simulation study

Total heritability: $H^2 \approx 0.81$.
Individual heritability: h_j^2 between 0.006 and 0.037

Definition of false positives and true positives

Problem

Causal SNPs are known

Frequently strongly correlated SNPs are selected:
Are these to be classified as true or false positives?

Common solution

Define threshold value C (E.g. $C = 0.7$ or $C = 0.9$).
If correlation between detected SNP and causal SNP larger than C
$\Rightarrow$ Classification as true positive

Definition of false positives and true positives

Problem

Causal SNPs are known

Frequently strongly correlated SNPs are selected:
Are these to be classified as true or false positives?

Common solution

Define threshold value C (E.g. $C = 0.7$ or $C = 0.9$).
If correlation between detected SNP and causal SNP larger than C

$\Rightarrow$ Classification as true positive

Comparison between model selection and single marker tests

Four methods:

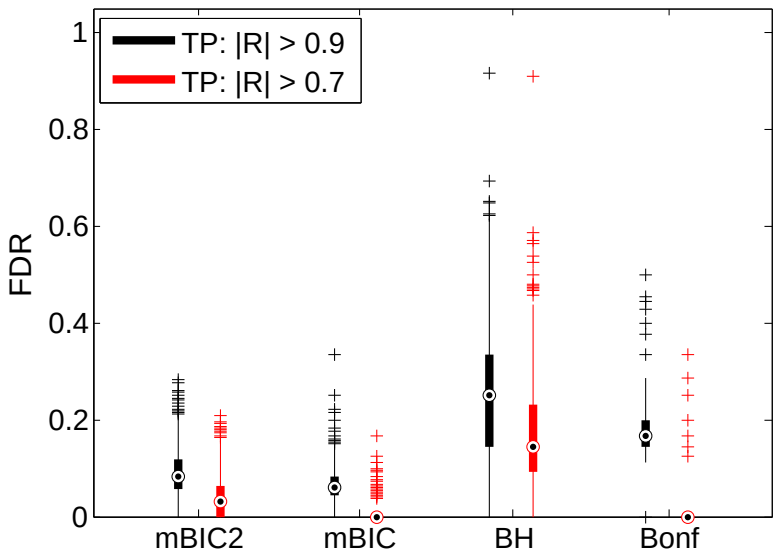
	Per marker	Model selection
Control of FWER	Bonferroni	mBIC
Control of FDR	Benjamini Hochberg	mBIC2

False Discovery Rate

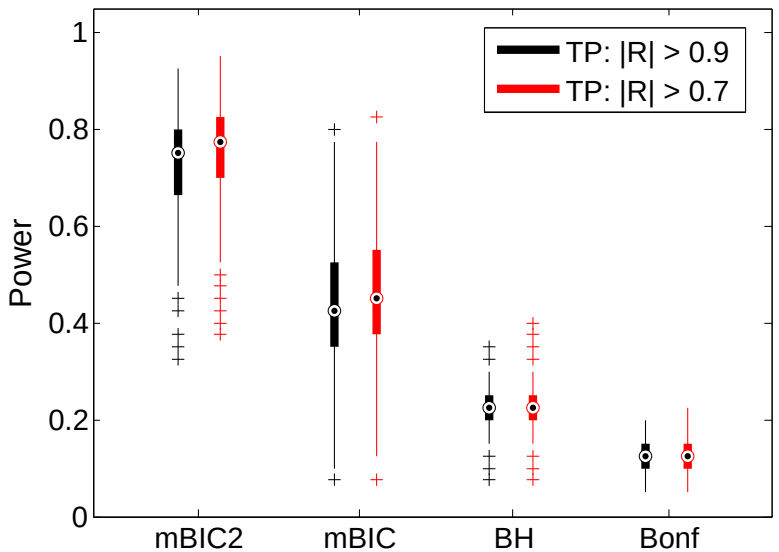
FDR and Power

Cutoffs define TP and FP for correlated marker

FDR



Power



Explanation for low power

Correct model: $Y_i = \beta_0 + \sum_{l \in M^*} \beta_l X_{il} + \epsilon_i$

Test statistic: $F_j = \frac{(n-2)MSS(X_j)}{RSS(X_j)}$

Non-centrality parameters:

$$\nu_{M,j} = \frac{\left(\sum_{l=1}^{k} \beta_l \text{Cov}\left(X_j, X_l\right)\right)^2}{\sigma^2 \text{Var}\left(X_j\right)}$$

$$\nu_{R,j} = \sum_{l \in M^* \setminus \{j\}} \sum_{r \in M^* \setminus \{j\}} \frac{\beta_l \beta_r}{\sigma^2} \left(\text{Cov}\left(X_l, X_r\right) - \frac{\text{Cov}\left(X_l, X_j\right)\text{Cov}\left(X_r, X_j\right)}{\text{Var}\left(X_j\right)}\right).$$

$\nu_{R,j}$ contains contribution of all other causal SNPs

Explanation for low power

Correct model: $Y_i = \beta_0 + \sum_{l \in M^*} \beta_l X_{il} + \epsilon_i$

Test statistic: $F_j = \frac{(n-2)MSS(X_j)}{RSS(X_j)}$

Non-centrality parameters:

$$\nu_{M,j} = \frac{\left(\sum_{l=1}^{k} \beta_l \text{Cov}\left(X_j, X_l\right)\right)^2}{\sigma^2 \text{Var}\left(X_j\right)}$$

$$\nu_{R,j} = \sum_{l \in M^* \setminus \{j\}} \sum_{r \in M^* \setminus \{j\}} \frac{\beta_l \beta_r}{\sigma^2} \left(\text{Cov}\left(X_l, X_r\right) - \frac{\text{Cov}\left(X_l, X_j\right)\text{Cov}\left(X_r, X_j\right)}{\text{Var}\left(X_j\right)} \right) .$$

$\nu_{R,j}$ contains contribution of all other causal SNPs

Explanation for low power

Correct model: $Y_i = \beta_0 + \sum_{l \in M^*} \beta_l X_{il} + \epsilon_i$

Test statistic: $F_j = \frac{(n-2)MSS(X_j)}{RSS(X_j)}$

Non-centrality parameters:

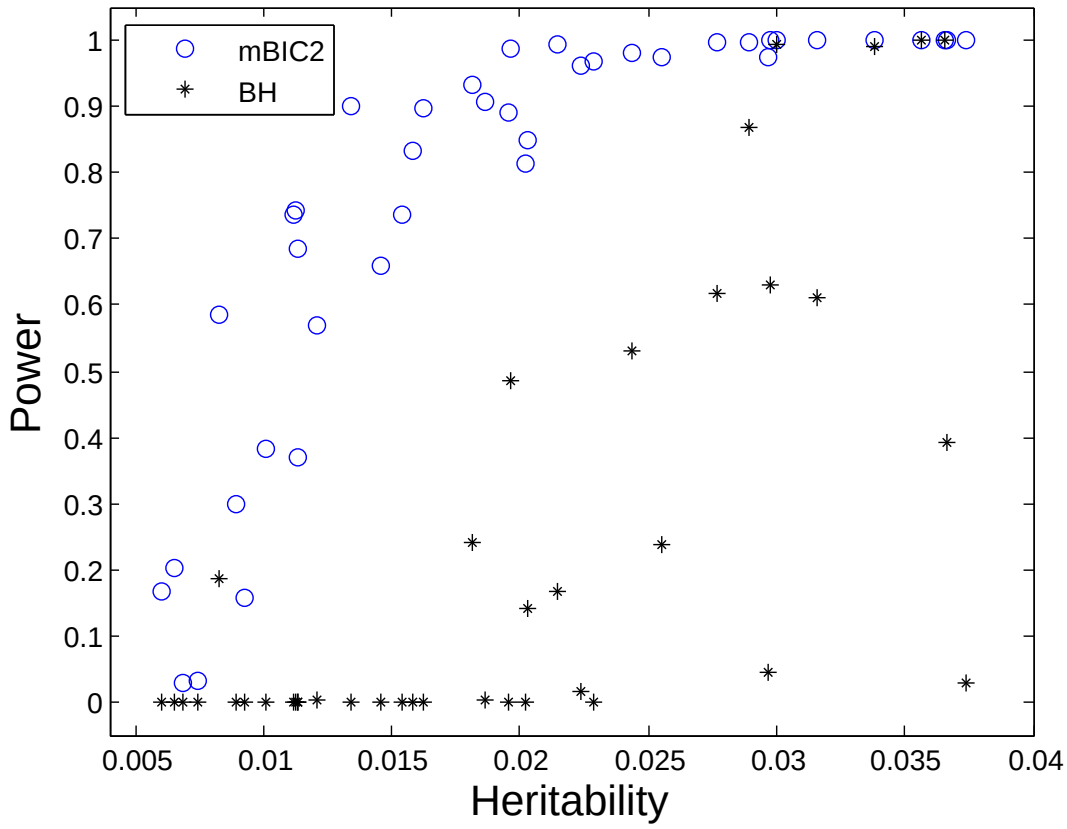
$$\nu_{M,j} = \frac{\left(\sum_{l=1}^{k} \beta_l \text{Cov}\left(X_j, X_l\right)\right)^2}{\sigma^2 \text{Var}\left(X_j\right)}$$

$$\nu_{R,j} = \sum_{l \in M^* \setminus \{j\}} \sum_{r \in M^* \setminus \{j\}} \frac{\beta_l \beta_r}{\sigma^2} \left(\text{Cov}\left(X_l, X_r\right) - \frac{\text{Cov}\left(X_l, X_j\right)\text{Cov}\left(X_r, X_j\right)}{\text{Var}\left(X_j\right)} \right).$$

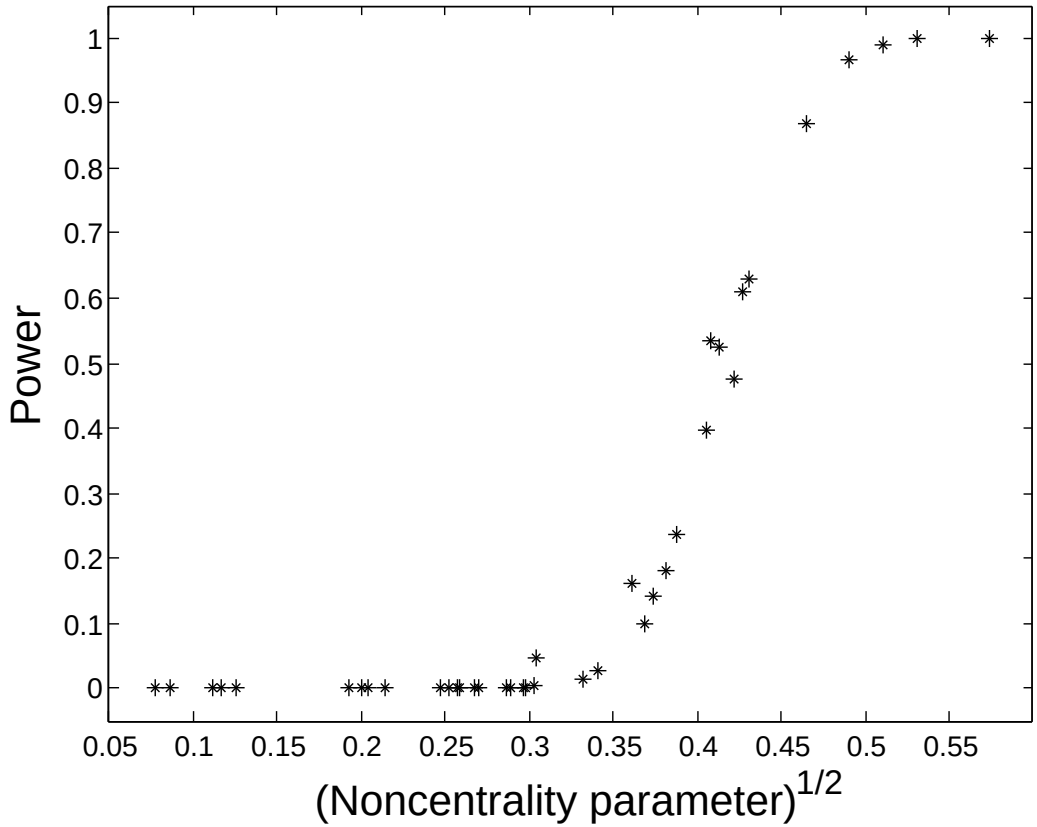
$\nu_{R,j}$ contains contribution of all other causal SNPs

Power for mBIC2 and BH

Bigger surprise:



Explanation for problems of BH



Non-centrality parameter:

$$\sqrt{\nu_{M,j}} = \left| \frac{\beta_j}{\sigma} \sqrt{\text{Var}\ (X_j)} + \frac{\sum_{l \neq j} \beta_l \text{Cov}\ (X_j, X_l)}{\sigma \sqrt{\text{Var}\ (x_j)}} \right|$$

Practical conclusions for GWAS analysis

In case of complex traits:

- Single marker tests (SMT) have low power
  $\Rightarrow$ One aspect in the discussion about "missing heritability"
- SMT have difficulties to rank the importance of causal SNPs
  $\Rightarrow$ Problem with replicability in GWAS
- For the same reason SMT systematically detect some false positives which are not correlated with any of the causal SNPs

Model selection approach helps to some extent

Practical conclusions for GWAS analysis

In case of complex traits:

- Single marker tests (SMT) have low power
  $\Rightarrow$ One aspect in the discussion about "missing heritability"
- SMT have difficulties to rank the importance of causal SNPs
  $\Rightarrow$ Problem with replicability in GWAS
- For the same reason SMT systematically detect some false positives which are not correlated with any of the causal SNPs

Model selection approach helps to some extent

Practical conclusions for GWAS analysis

In case of complex traits:

- Single marker tests (SMT) have low power
  $\Rightarrow$ One aspect in the discussion about "missing heritability"
- SMT have difficulties to rank the importance of causal SNPs
  $\Rightarrow$ Problem with replicability in GWAS
- For the same reason SMT systematically detect some false positives which are not correlated with any of the causal SNPs

Model selection approach helps to some extent

Practical conclusions for GWAS analysis

In case of complex traits:

- Single marker tests (SMT) have low power
  $\Rightarrow$ One aspect in the discussion about "missing heritability"
- SMT have difficulties to rank the importance of causal SNPs
  $\Rightarrow$ Problem with replicability in GWAS
- For the same reason SMT systematically detect some false positives which are not correlated with any of the causal SNPs

Model selection approach helps to some extent

Logistic Regression for Case Control

Usual model

Y_i is Bernoulli, $P(Y_i = 1) = p_i$, with

$$\log(p_i/(1-p_i)) = \beta_0 + \sum_{j \in M} \beta_j X_{ij}$$

We discuss three selected methods

- HLASSO: Hoggart, Balding (2008)
- GWASelect: He and Lin (2011)
- MOSGWA: Our own approach

Logistic Regression for Case Control

Usual model

Y_i is Bernoulli, $P(Y_i = 1) = p_i$, with

$$\log(p_i/(1-p_i)) = \beta_0 + \sum_{j \in M} \beta_j X_{ij}$$

We discuss three selected methods

- HLASSO: Hoggart, Balding (2008)
- GWASelect: He and Lin (2011)
- MOSGWA: Our own approach

HLASSO

Hoggart, . . . , Balding: Bioinformatics (2008)

More Bayesian approach using shrinkage priors on coefficients of logistic regression models

- Gaussian prior (implemented, but no selection)
- Double exponential prior (DE) $\Rightarrow$ equivalent to Lasso
- Normal exponential gamma (NEG) $\Rightarrow$ Hyper Lasso

Densities of DE and NEG

$$DE(\beta|\xi) = \int_{\sigma^2=0}^{\infty} N(\beta|0,\sigma^2)\, Ga(\sigma^2|1,\xi^2/2) d\sigma^2 = \frac{\xi}{2}\exp\left(-\xi|\beta|\right)$$

$$NEG(\beta|\lambda,\gamma) = \int_{\Psi=0}^{\infty}\int_{\sigma^2=0}^{\infty} N(\beta|0,\sigma^2)\, Ga(\sigma^2|1,\Psi)\, Ga(\Psi|\lambda,\gamma^2) d\sigma^2 d\Psi$$

HLASSO

Hoggart, . . . , Balding: Bioinformatics (2008)

More Bayesian approach using shrinkage priors on coefficients of logistic regression models

- Gaussian prior (implemented, but no selection)
- Double exponential prior (DE) $\Rightarrow$ equivalent to Lasso
- Normal exponential gamma (NEG) $\Rightarrow$ Hyper Lasso

Densities of DE and NEG

$$DE(\beta|\xi) = \int_{\sigma^2=0}^{\infty} N(\beta|0,\sigma^2)Ga(\sigma^2|1,\xi^2/2)d\sigma^2 = \frac{\xi}{2}\exp\left(-\xi|\beta|\right)$$

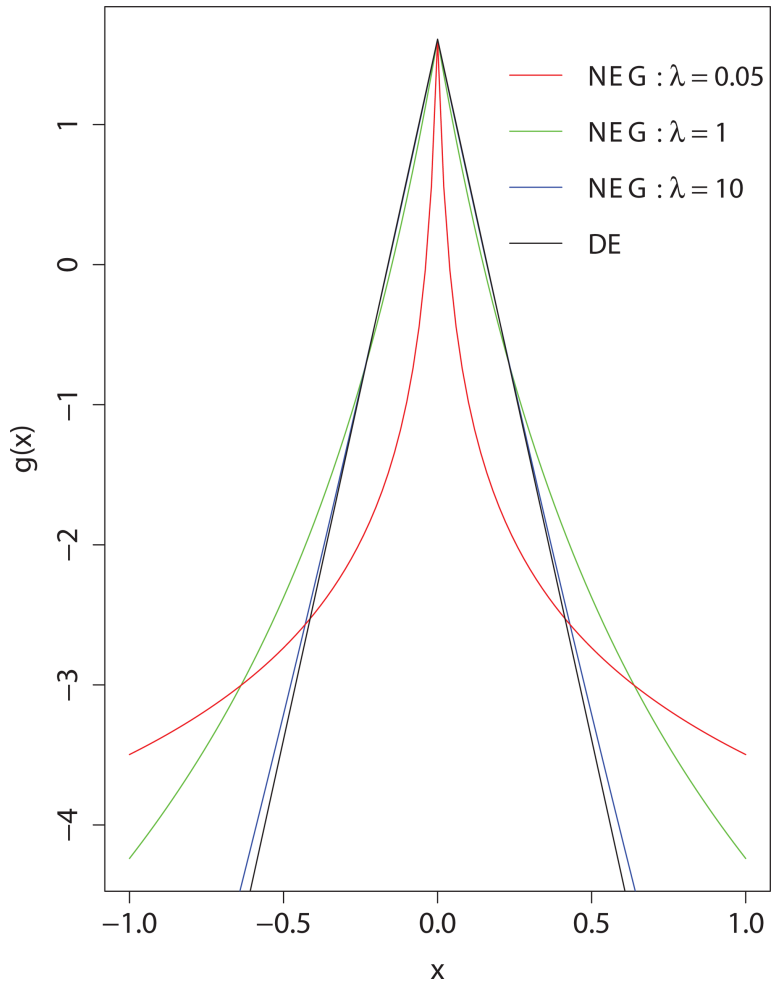
$$NEG(\beta|\lambda,\gamma) = \int_{\Psi=0}^{\infty}\int_{\sigma^2=0}^{\infty} N(\beta|0,\sigma^2)Ga(\sigma^2|1,\Psi)Ga(\Psi|\lambda,\gamma^2)d\sigma^2 d\Psi$$

NEG priors

Logarithm of prior densities fixed to have the same density at the origin (Taken from Hoggart et al., 2008)

$$NEG(\beta|\lambda,\gamma) \propto \exp\left(\frac{\beta^2}{4\gamma^2}\right) D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma}\right)$$

where D is parabolic cylinder function

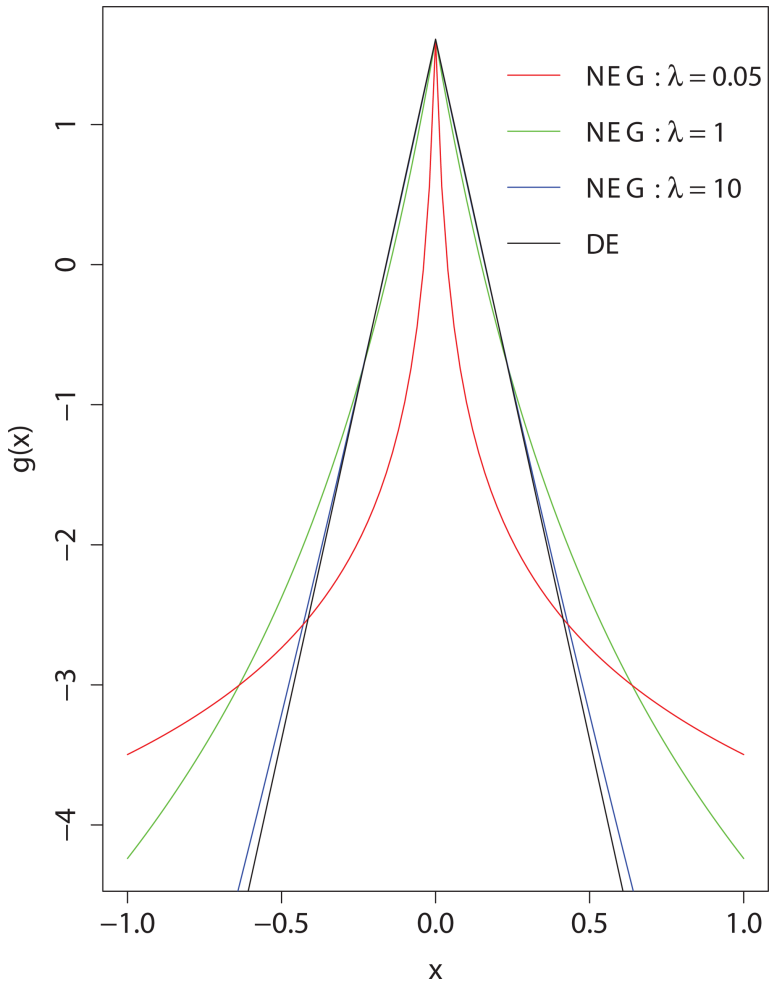


NEG priors

Logarithm of prior densities fixed to have the same density at the origin (Taken from Hoggart et al., 2008)

$$NEG(\beta|\lambda,\gamma) \propto \exp\left(\frac{\beta^2}{4\gamma^2}\right) D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma}\right)$$

where D is parabolic cylinder function



HLASSO

Optimisation algorithm

Not fully Bayesian, but searching only for posterior mode

$$\log p(\beta|X, Y) = \ell(\beta) - q(\beta) + const \ ,$$

with

$$\ell(\beta) := \log L(\beta|X, Y), \quad q(\beta) := -\log NEG(\beta|\lambda, \gamma)$$

HLASSO has rather efficient implementation to find optimum of $\log p$ CLG algorithm (cyclic coordinate descent) with clever bounds to speed up

HLASSO

Optimisation algorithm

Not fully Bayesian, but searching only for posterior mode

$$\log p(\beta|X, Y) = \ell(\beta) - q(\beta) + const \ ,$$

with

$$\ell(\beta) := \log L(\beta|X, Y), \quad q(\beta) := -\log NEG(\beta|\lambda, \gamma)$$

HLASSO has rather efficient implementation to find optimum of $\log p$ CLG algorithm (cyclic coordinate descent) with clever bounds to speed up

HLASSO

CLG - Basic idea

Run iteratively and repeatedly through all coefficients with component-wise Newton

$$\beta_j^{new} = \beta_j - \frac{\frac{\partial}{\partial \beta_j}\ell(\beta) - q'(\beta_j)}{\frac{\partial^2}{\partial \beta_j^2}\ell(\beta) - q''(\beta_j)}$$

If $\beta_j^{new} \cdot \beta_j < 0$ then set $\beta_j^{new} = 0$

Specifically if $\beta_j = 0$

Consider both limits $\beta_j = 0^+$ and $\beta_j = 0^-$
No change of sign is equivalent to

$$\left|\frac{\partial}{\partial \beta_j}\ell(\beta)\right|_{\beta_j=0} > q'(\beta_j = 0^+)$$

HLASSO

CLG - Basic idea

Run iteratively and repeatedly through all coefficients with component-wise Newton

$$\beta_j^{new} = \beta_j - \frac{\frac{\partial}{\partial \beta_j}\ell(\beta) - q'(\beta_j)}{\frac{\partial^2}{\partial \beta_j^2}\ell(\beta) - q''(\beta_j)}$$

If $\beta_j^{new} \cdot \beta_j < 0$ then set $\beta_j^{new} = 0$

Specifically if $\beta_j = 0$

Consider both limits $\beta_j = 0^+$ and $\beta_j = 0^-$
No change of sign is equivalent to

$$\left|\frac{\partial}{\partial \beta_j}\ell(\beta)\right|_{\beta_j=0} > q'(\beta_j = 0^+)$$

HLASSO

Parameter tuning to control FWER

Asymptotic normality under null gives $\hat{\beta}_j \sim \mathcal{N}\left(0, \frac{n_0+n_1}{n_0 n_1}\right)$

From last relationsship of previous slide one then gets relationship to determine type I error

$$q'(\beta_j = 0^+) = \sqrt{\frac{n_0 n_1}{n_0 + n_1}}\, \Phi^{-1}(1 - \alpha/2)$$

- Double Exponential: $q'(\beta_j = 0^+) = \xi$
- NEG: (Careful: Typo in paper)

$$q'(\beta_j = 0^+) = const \cdot \frac{2\lambda + 1}{\gamma}$$

Now actually two parameters to be fitted, Hoggart et al. use $\lambda = 0.05$ (after trying other parameters) and the γ follows

HLASSO

Parameter tuning to control FWER

Asymptotic normality under null gives $\hat{\beta}_j \sim \mathcal{N}\left(0, \frac{n_0+n_1}{n_0 n_1}\right)$

From last relationsship of previous slide one then gets relationship to determine type I error

$$q'(\beta_j = 0^+) = \sqrt{\frac{n_0 n_1}{n_0 + n_1}}\, \Phi^{-1}(1 - \alpha/2)$$

- Double Exponential: $q'(\beta_j = 0^+) = \xi$
- NEG: (Careful: Typo in paper)

$$q'(\beta_j = 0^+) = const \cdot \frac{2\lambda + 1}{\gamma}$$

Now actually two parameters to be fitted, Hoggart et al. use $\lambda = 0.05$ (after trying other parameters) and the γ follows

HLASSO

Parameter tuning to control FWER

Asymptotic normality under null gives $\hat{\beta}_j \sim \mathcal{N}\left(0, \frac{n_0+n_1}{n_0 n_1}\right)$

From last relationsship of previous slide one then gets relationship to determine type I error

$$q'(\beta_j = 0^+) = \sqrt{\frac{n_0 n_1}{n_0 + n_1}}\, \Phi^{-1}(1 - \alpha/2)$$

- Double Exponential: $q'(\beta_j = 0^+) = \xi$
- NEG: (Careful: Typo in paper)

$$q'(\beta_j = 0^+) = const \cdot \frac{2\lambda + 1}{\gamma}$$

Now actually two parameters to be fitted, Hoggart et al. use $\lambda = 0.05$ (after trying other parameters) and the γ follows

HLASSO

Parameter tuning to control FWER

HLASSO offers choice of parameter α which corresponds to uncorrected significance level

$\Rightarrow$ Choosing for example $\alpha = 0.05/p$ works

Difference between LASSO and HLASSO

- Lighter tails of DE distribution $\Rightarrow$ more shrinkage
  $\Rightarrow$ Correlated SNPs tend to enter model to explain full effect of causal SNP
- NEG prior has heavier tails $\Rightarrow$ less shrinkage

HLASSO

Parameter tuning to control FWER

HLASSO offers choice of parameter α which corresponds to uncorrected significance level

$\Rightarrow$ Choosing for example $\alpha = 0.05/p$ works

Difference between LASSO and HLASSO

- Lighter tails of DE distribution $\Rightarrow$ more shrinkage
  $\Rightarrow$ Correlated SNPs tend to enter model to explain full effect of causal SNP
- NEG prior has heavier tails $\Rightarrow$ less shrinkage

GWASelect

He and Lin: Bioinformatics (2011)

Rough outline of algorithm

1. Sure Independence Screening (SIS)
2. Lasso for logistic regression
3. 'Pruning' of correlated SNPs

These steps are iterated based on conditional score tests $\Rightarrow$ ISIS

Stability selection

Meinshausen, Bühlmann (2010)
Perform the whole procedure on 50 subsamples
(randomly select 50% cases and 50 % controls)
$\Rightarrow$ compute selection probabilities

GWASelect

He and Lin: Bioinformatics (2011)

Rough outline of algorithm

1. Sure Independence Screening (SIS)
2. Lasso for logistic regression
3. 'Pruning' of correlated SNPs

These steps are iterated based on conditional score tests $\Rightarrow$ ISIS

Stability selection

Meinshausen, Bühlmann (2010)
Perform the whole procedure on 50 subsamples
(randomly select 50% cases and 50 % controls)
$\Rightarrow$ compute selection probabilities

GWASelect

He and Lin: Bioinformatics (2011)

Rough outline of algorithm

1. Sure Independence Screening (SIS)
2. Lasso for logistic regression
3. 'Pruning' of correlated SNPs

These steps are iterated based on conditional score tests $\Rightarrow$ ISIS

Stability selection

Meinshausen, Bühlmann (2010)
Perform the whole procedure on 50 subsamples
(randomly select 50% cases and 50 % controls)
$\Rightarrow$ compute selection probabilities

GWASelect

First iteration

Preselection based on marginal tests (Cochran Armitage trend test),
SIS theory: Consider $0.9n/(4 \log n)$ SNPs with largest test statistic

Lasso

Model selection using Lasso on selected set of SNPs
(cyclic coordinate decent for optimization)
For 'dynamic' GWASelect tuning parameter λ determined by 5-fold cross validation

Pruning of Correlated SNPs

Remove SNPs from model which have pairwise correlation $|R| < 0.8$

GWASelect

First iteration

Preselection based on marginal tests (Cochran Armitage trend test),
SIS theory: Consider $0.9n/(4 \log n)$ SNPs with largest test statistic

Lasso

Model selection using Lasso on selected set of SNPs
(cyclic coordinate decent for optimization)
For 'dynamic' GWASelect tuning parameter λ determined by 5-fold cross validation

Pruning of Correlated SNPs

Remove SNPs from model which have pairwise correlation $|R| < 0.8$

GWASelect

First iteration

Preselection based on marginal tests (Cochran Armitage trend test),
SIS theory: Consider $0.9n/(4 \log n)$ SNPs with largest test statistic

Lasso

Model selection using Lasso on selected set of SNPs
(cyclic coordinate decent for optimization)
For 'dynamic' GWASelect tuning parameter λ determined by 5-fold cross validation

Pruning of Correlated SNPs

Remove SNPs from model which have pairwise correlation $|R| < 0.8$

GWASelect

Second and third iteration

After pruning t SNPs left in model, say $X_1, \ldots, X_t$
Interested to consider all influence of other SNPs conditional on SNPs already in the model, specifically for all $X_r, r > t$

$$\log(p_i/(1-p_i)) = \beta_0 + \sum_{j=1}^{t} \beta_j X_{ij} + \gamma X_r$$

we would like to know if $\gamma = 0$
LRT too timeconsuming, but **Scoretest** very fast alternative
$\Rightarrow$ SIS step based on Score test statistics
Keep $0.05n/(4 \log n)$ best SNPs
Then again Lasso and pruning of correlated SNPs

GWASelect

Second and third iteration

After pruning t SNPs left in model, say $X_1, \ldots, X_t$
Interested to consider all influence of other SNPs conditional on SNPs already in the model, specifically for all $X_r, r > t$

$$\log(p_i/(1-p_i)) = \beta_0 + \sum_{j=1}^{t} \beta_j X_{ij} + \gamma X_r$$

we would like to know if $\gamma = 0$
LRT too timeconsuming, but **Scoretest** very fast alternative

$\Rightarrow$ SIS step based on Score test statistics
Keep $0.05n/(4 \log n)$ best SNPs
Then again Lasso and pruning of correlated SNPs

GWASelect

Second and third iteration

After pruning t SNPs left in model, say $X_1, \ldots, X_t$
Interested to consider all influence of other SNPs conditional on SNPs already in the model, specifically for all $X_r, r > t$

$$\log(p_i/(1-p_i)) = \beta_0 + \sum_{j=1}^{t} \beta_j X_{ij} + \gamma X_r$$

we would like to know if $\gamma = 0$
LRT too timeconsuming, but **Scoretest** very fast alternative
$\Rightarrow$ SIS step based on Score test statistics
Keep $0.05n/(4 \log n)$ best SNPs
Then again Lasso and pruning of correlated SNPs

GWASelect

Software

- Stand alone program, difference between GWASelect and d-GWASelect
- In d-GWASelect only parameter to choose is threshold from stability selection
  Recommended values: between 0.1 and 0.2

MOSGWA for Case Control

Based on criterion mBIC2

$$mBIC2 = -2 \log L_M + k_M[\log(np^2/4)] - 2 \log k_m!$$

Specific issue with logistic regression

$p > n \quad \Rightarrow \quad$ problem of **complete separation** occurs even for large n

$\Rightarrow \quad$ we use Firth correction

$$L^*(\beta|\ Y, X) := L(\beta) \cdot |I(\beta)|^{1/2}$$

$|I(\beta)|^{1/2}$ ... Jeffreys prior , $\quad I(\beta) = -E\left(\frac{\partial^2 \beta}{\partial \beta_r \partial \beta_s} \log L(\beta)\right)^2$

Neither HLASSO nor GWASelect have this problem, as they implicitly penalize too large β_j

MOSGWA for Case Control

Based on criterion mBIC2

$$mBIC2 = -2 \log L_M + k_M[\log(np^2/4)] - 2 \log k_m!$$

Specific issue with logistic regression

$p > n \quad \Rightarrow \quad$ problem of **complete separation** occurs even for large n

$\Rightarrow \quad$ we use Firth correction

$$L^*(\beta|\ Y, X) := L(\beta) \cdot |I(\beta)|^{1/2}$$

$|I(\beta)|^{1/2}$ ... Jeffreys prior , $\qquad I(\beta) = -E\left(\frac{\partial^2 \beta}{\partial \beta_r \partial \beta_s} \log L(\beta)\right)^2$

Neither HLASSO nor GWASelect have this problem, as they implicitly penalize too large β_j

MOSGWA for Case Control

Based on criterion mBIC2

$$mBIC2 = -2 \log L_M + k_M[\log(np^2/4)] - 2 \log k_m!$$

Specific issue with logistic regression

$p > n \quad \Rightarrow \quad$ problem of **complete separation** occurs even for large n

$\Rightarrow \quad$ we use Firth correction

$$L^*(\beta|\ Y, X) := L(\beta) \cdot |I(\beta)|^{1/2}$$

$|I(\beta)|^{1/2}$ ... Jeffreys prior , $\qquad I(\beta) = -E\left(\frac{\partial^2 \beta}{\partial \beta_r \partial \beta_s} \log L(\beta)\right)^2$

Neither HLASSO nor GWASelect have this problem, as they implicitly penalize too large β_j

MOSGWA, Model Search

Major issue for practical application $\Rightarrow$ 2^p potential models

Strategies

Computation of ML much more time consuming than for linear regression $\Rightarrow$ Even more important than for linear regression to keep models small

- Preselection of markers using marginal tests (Compare SIS)
- Heuristic greedy search procedures (as described in the next slide)
- Genetic Algorithms (as described later)

MOSGWA, Model Search

Major issue for practical application $\Rightarrow$ 2^p potential models

Strategies

Computation of ML much more time consuming than for linear regression $\Rightarrow$ Even more important than for linear regression to keep models small

- Preselection of markers using marginal tests (Compare SIS)
- Heuristic greedy search procedures (as described in the next slide)
- Genetic Algorithms (as described later)

MOSGWA, Model Search

Major issue for practical application $\Rightarrow$ 2^p potential models

Strategies

Computation of ML much more time consuming than for linear regression $\Rightarrow$ Even more important than for linear regression to keep models small

- Preselection of markers using marginal tests (Compare SIS)
- Heuristic greedy search procedures (as described in the next slide)
- Genetic Algorithms (as described later)

MOSGWA, Model Search

Major issue for practical application $\Rightarrow$ 2^p potential models

Strategies

Computation of ML much more time consuming than for linear regression $\Rightarrow$ Even more important than for linear regression to keep models small

- Preselection of markers using marginal tests (Compare SIS)
- Heuristic greedy search procedures (as described in the next slide)
- Genetic Algorithms (as described later)

MOSGWA, Model Search

Major issue for practical application $\Rightarrow$ 2^p potential models

Strategies

Computation of ML much more time consuming than for linear regression $\Rightarrow$ Even more important than for linear regression to keep models small

- Preselection of markers using marginal tests (Compare SIS)
- Heuristic greedy search procedures (as described in the next slide)
- Genetic Algorithms (as described later)

Model Search strategy

Step 1

- Preselection and sorting based on marginal tests (CAT)
- Fast stepwise search (to be specified)

Step 2

- Preselection and sorting based on Score tests conditional on model obtained in Step 1
  Idea: For logistic regression Score test much faster than LRT
- Fast stepwise search (to be specified)

Described strategy good, but tends to get stuck in too small models

$\Rightarrow$ Start with search using milder criterion

Model Search strategy

Step 1

- Preselection and sorting based on marginal tests (CAT)
- Fast stepwise search (to be specified)

Step 2

- Preselection and sorting based on Score tests conditional on model obtained in Step 1
  Idea: For logistic regression Score test much faster than LRT
- Fast stepwise search (to be specified)

Described strategy good, but tends to get stuck in too small models

$\Rightarrow$ Start with search using milder criterion

Model Search strategy

Step 1

- Preselection and sorting based on marginal tests (CAT)
- Fast stepwise search (to be specified)

Step 2

- Preselection and sorting based on Score tests conditional on model obtained in Step 1
  Idea: For logistic regression Score test much faster than LRT
- Fast stepwise search (to be specified)

Described strategy good, but tends to get stuck in too small models

$\Rightarrow$ Start with search using milder criterion

Model Search strategy

Fast stepwise search

Starting point: Regressors sorted by test statistic
Iterate the following three steps till no further improvement

- **Directed forward search:** Find the first regressor in sorted list which decreases mBIC2 and add to model
- **Exchange step:** See if substituting any regressor in model with candidate SNP foe exchange decreases mBIC2
  Candidates: neighboring SNPs
  or in first step SNPs preselected with CAT
- **Backward step:** Routine backward elimination

Fast enough to deal with full GWAS data sets

Model Search strategy

Fast stepwise search

Starting point: Regressors sorted by test statistic
Iterate the following three steps till no further improvement

- **Directed forward search:** Find the first regressor in sorted list which decreases mBIC2 and add to model
- **Exchange step:** See if substituting any regressor in model with candidate SNP foe exchange decreases mBIC2
  Candidates: neighboring SNPs
  or in first step SNPs preselected with CAT
- **Backward step:** Routine backward elimination

Fast enough to deal with full GWAS data sets

Simulation study

Simulation study

Comparison of MOSGWA, GWASelect and HLasso
Again SNP data from POPRES (dbGaP), now more than 4000 individuals

First simulation under global null

For four different sets of SNPs
Chr1, Chr1 + Chr2, Chr1 - Chr4, Chr1 - Chr6

Second simulation

24 causal SNPs (uncorrelated, MAF > 0.3)
Simulate 200 instances under logistic regression model

Effect β_j sizes between 0.2 and 0.26
Half of causal SNPs removed before search

Simulation study

Simulation study

Comparison of MOSGWA, GWASelect and HLasso
Again SNP data from POPRES (dbGaP), now more than 4000 individuals

First simulation under global null

For four different sets of SNPs
Chr1, Chr1 + Chr2, Chr1 - Chr4, Chr1 - Chr6

Second simulation

24 causal SNPs (uncorrelated, MAF > 0.3)
Simulate 200 instances under logistic regression model

Effect β_j sizes between 0.2 and 0.26
Half of causal SNPs removed before search

Simulation study

Simulation study

Comparison of MOSGWA, GWASelect and HLasso
Again SNP data from POPRES (dbGaP), now more than 4000 individuals

First simulation under global null

For four different sets of SNPs
Chr1, Chr1 + Chr2, Chr1 - Chr4, Chr1 - Chr6

Second simulation

24 causal SNPs (uncorrelated, MAF > 0.3)
Simulate 200 instances under logistic regression model

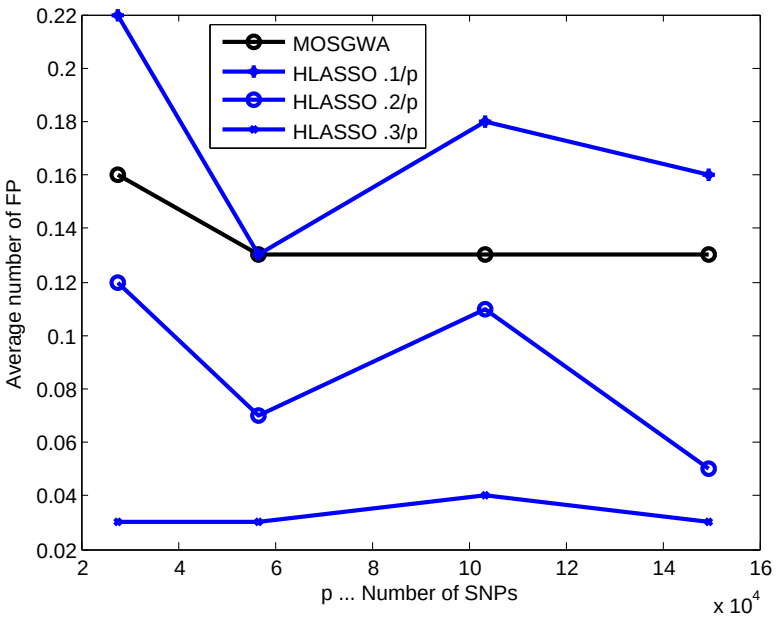
Effect β_j sizes between 0.2 and 0.26
Half of causal SNPs removed before search

Simulation under global null

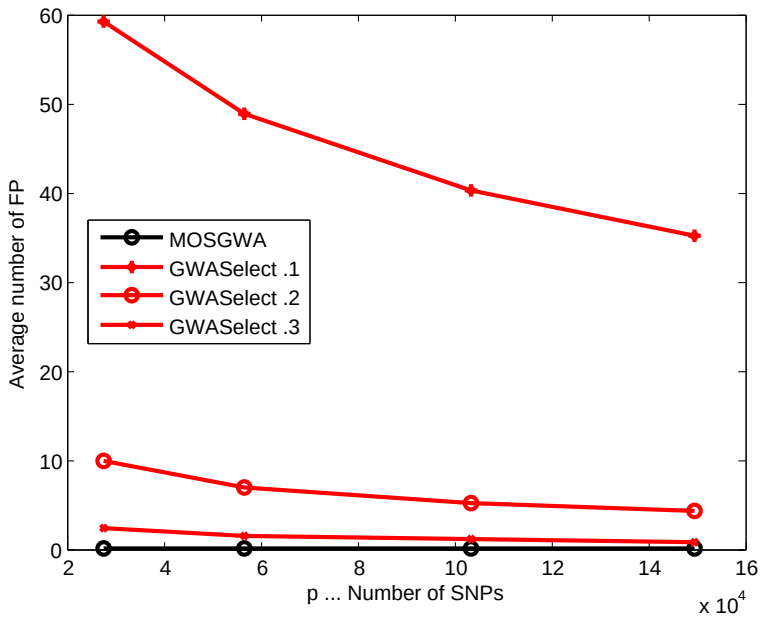
Average number of False Positives

HLASSO with parameters $0.1/p, 0.2/p, 0.3/p$
GWASelect with parameters $0.1, 0.2, 0.3$

MOSGWA and HLASSO



MOSGWA and GWASelect

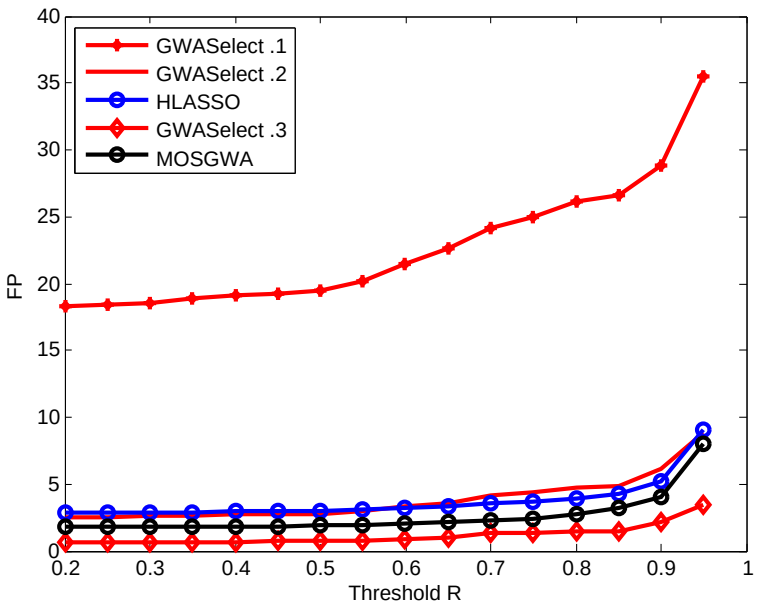


Simulation under Model ($k^* = 24$)

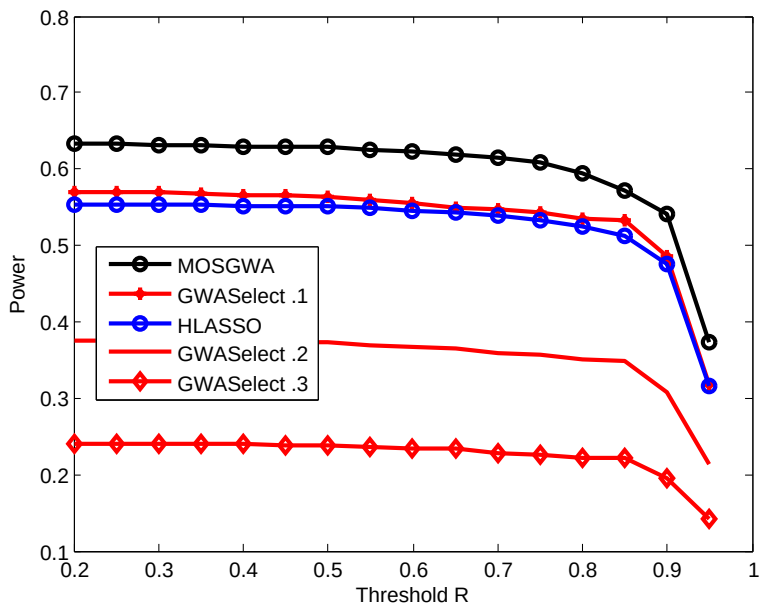
False positives and Power

Cutoffs define TP and FP for correlated marker

FP



Power



MOSGWA

Model **S**election for **G**enome **W**ide **A**ssociation

Package

- Written in C++ by Bodenstorfer, Dolejsi, Ruhaltinger
- **Aim**: Professional software for genetic researchers
- Currently
  - SNP array data (PLINK format and HDF5)
  - Linear and Logistic Regression
  - Allows for inclusion of covariates
- First version on-line since this week!
  https://sourceforge.net/projects/mosgwa/

MOSGWA

Model **S**election for **G**enome **W**ide **A**ssociation

Package

- Written in C++ by Bodenstorfer, Dolejsi, Ruhaltinger
- **Aim**: Professional software for genetic researchers
- Currently
  - SNP array data (PLINK format and HDF5)
  - Linear and Logistic Regression
  - Allows for inclusion of covariates
- First version on-line since this week!
  https://sourceforge.net/projects/mosgwa/

MOSGWA

Model **S**election for **G**enome **W**ide **A**ssociation

Package

- Written in C++ by Bodenstorfer, Dolejsi, Ruhaltinger
- **Aim**: Professional software for genetic researchers
- Currently
  - SNP array data (PLINK format and HDF5)
  - Linear and Logistic Regression
  - Allows for inclusion of covariates
- First version on-line since this week!
  https://sourceforge.net/projects/mosgwa/

MOSGWA

Planned Extensions

- Better search strategies Specifically genetic algorithm for search
- Mixed models (relatively easy for QT)
- Sequencing data, including methods for rare SNPs
- Admixture mapping
- Logic regression for interactions, etc.

MOSGWA

Planned Extensions

- Better search strategies Specifically genetic algorithm for search
- Mixed models (relatively easy for QT)
- Sequencing data, including methods for rare SNPs
- Admixture mapping
- Logic regression for interactions, etc.

MOSGWA

Planned Extensions

- Better search strategies Specifically genetic algorithm for search
- Mixed models (relatively easy for QT)
- Sequencing data, including methods for rare SNPs
- Admixture mapping
- Logic regression for interactions, etc.

MOSGWA

Planned Extensions

- Better search strategies Specifically genetic algorithm for search
- Mixed models (relatively easy for QT)
- Sequencing data, including methods for rare SNPs
- Admixture mapping
- Logic regression for interactions, etc.

MOSGWA

Planned Extensions

- Better search strategies Specifically genetic algorithm for search
- Mixed models (relatively easy for QT)
- Sequencing data, including methods for rare SNPs
- Admixture mapping
- Logic regression for interactions, etc.

Memetic algorithm

Basic idea of genetic algorithm

- Work with population of models
- mBIC2 as measure of fitness of a model
- Use evolutionary dynamic to increase fitness of population
  Selection, recombination, mutation

Memetic algorithm

Also hybrid genetic algorithm
Incorporates local optimization to increase performance

Memetic algorithm

Basic idea of genetic algorithm

- Work with population of models
- mBIC2 as measure of fitness of a model
- Use evolutionary dynamic to increase fitness of population
  Selection, recombination, mutation

Memetic algorithm

Also hybrid genetic algorithm
Incorporates local optimization to increase performance

Memetic algorithm

Basic idea of genetic algorithm

- Work with population of models
- mBIC2 as measure of fitness of a model
- Use evolutionary dynamic to increase fitness of population
  Selection, recombination, mutation

Memetic algorithm

Also hybrid genetic algorithm
Incorporates local optimization to increase performance

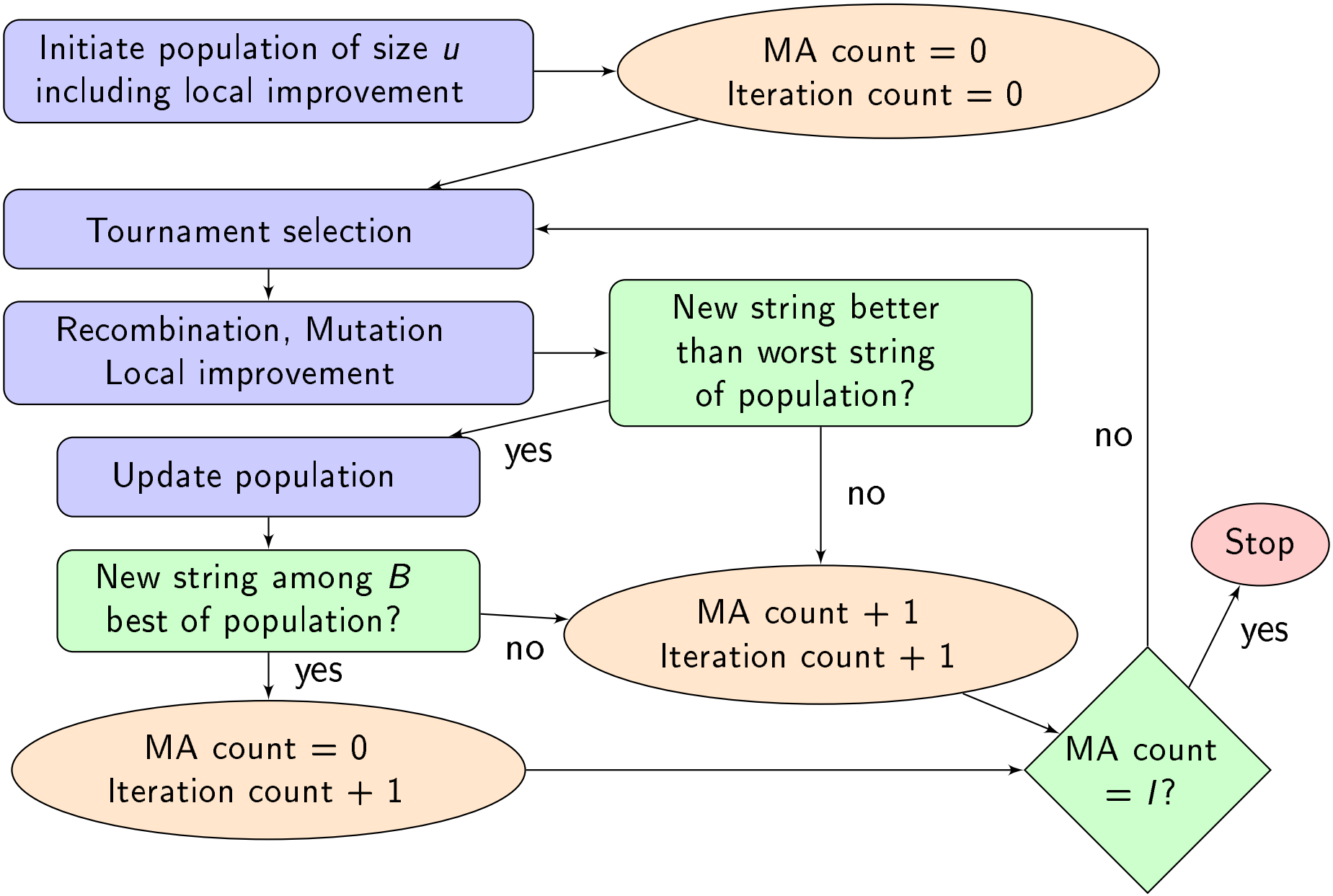
Memetic algorithm

Basic idea of genetic algorithm

- Work with population of models
- mBIC2 as measure of fitness of a model
- Use evolutionary dynamic to increase fitness of population
  Selection, recombination, mutation

Memetic algorithm

Also hybrid genetic algorithm
Incorporates local optimization to increase performance


Initiate population of size u including local improvement
MA count = 0
Iteration count = 0
Tournament selection
Recombination, Mutation
Local improvement
New string better than worst string of population?
yes
no
Update population
New string among B best of population?
no
yes
MA count + 1
Iteration count + 1
MA count = 0
Iteration count + 1
MA count = l?
no
yes
Stop

Memetic algorithms

MA particularly designed for our application

- Frommlet et al. (2012): Simulations for QTL mapping:
  $\Rightarrow$ GA finds frequently better solution than stepwise search
- Implementation for GWAS finished, not yet integrated in MOSGWA

Estimation of marker posteriors

MA gives many good solutions $\Rightarrow$ model posteriors

$$\sum_{M\in\mathcal{M}} P(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} P(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} \exp\left(-\text{mBIC}(M)/2\right) .$$

Then marker posteriors

$$P(j_r|Y) \approx \frac{\sum_{M\in\mathcal{M}_r^{Pool}} \exp\left(-\text{mBIC}(M)/2\right)}{\sum_{M'\in\mathcal{M}^{Pool}} \exp\left(-\text{mBIC}(M')/2\right)}$$

Memetic algorithms

MA particularly designed for our application

- Frommlet et al. (2012): Simulations for QTL mapping:
  $\Rightarrow$ GA finds frequently better solution than stepwise search
- Implementation for GWAS finished, not yet integrated in MOSGWA

Estimation of marker posteriors

MA gives many good solutions $\Rightarrow$ model posteriors

$$\sum_{M\in\mathcal{M}} P(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} P(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} \exp\left(-\text{mBIC}(M)/2\right) .$$

Then marker posteriors

$$P(j_r|Y) \approx \frac{\sum_{M\in\mathcal{M}_r^{Pool}} \exp\left(-\text{mBIC}(M)/2\right)}{\sum_{M'\in\mathcal{M}^{Pool}} \exp\left(-\text{mBIC}(M')/2\right)}$$

Memetic algorithms

MA particularly designed for our application

- Frommlet et al. (2012): Simulations for QTL mapping:
  $\Rightarrow$ GA finds frequently better solution than stepwise search
- Implementation for GWAS finished, not yet integrated in MOSGWA

Estimation of marker posteriors

MA gives many good solutions $\Rightarrow$ model posteriors

$$\sum_{M\in\mathcal{M}} \mathrm{P}(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} \mathrm{P}(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} \exp\left(-\mathrm{mBIC}(M)/2\right) .$$

Then marker posteriors

$$\mathrm{P}(j_r|Y) \approx \frac{\sum_{M\in\mathcal{M}_r^{Pool}} \exp\left(-\mathrm{mBIC}(M)/2\right)}{\sum_{M'\in\mathcal{M}^{Pool}} \exp\left(-\mathrm{mBIC}(M')/2\right)}$$

Memetic algorithms

MA particularly designed for our application

- Frommlet et al. (2012): Simulations for QTL mapping:
  $\Rightarrow$ GA finds frequently better solution than stepwise search
- Implementation for GWAS finished, not yet integrated in MOSGWA

Estimation of marker posteriors

MA gives many good solutions $\Rightarrow$ model posteriors

$$\sum_{M\in\mathcal{M}} \mathrm{P}(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} \mathrm{P}(Y|M)\cdot\pi(M) \approx \sum_{M\in Pool} \exp\left(-\mathrm{mBIC}(M)/2\right) .$$

Then marker posteriors

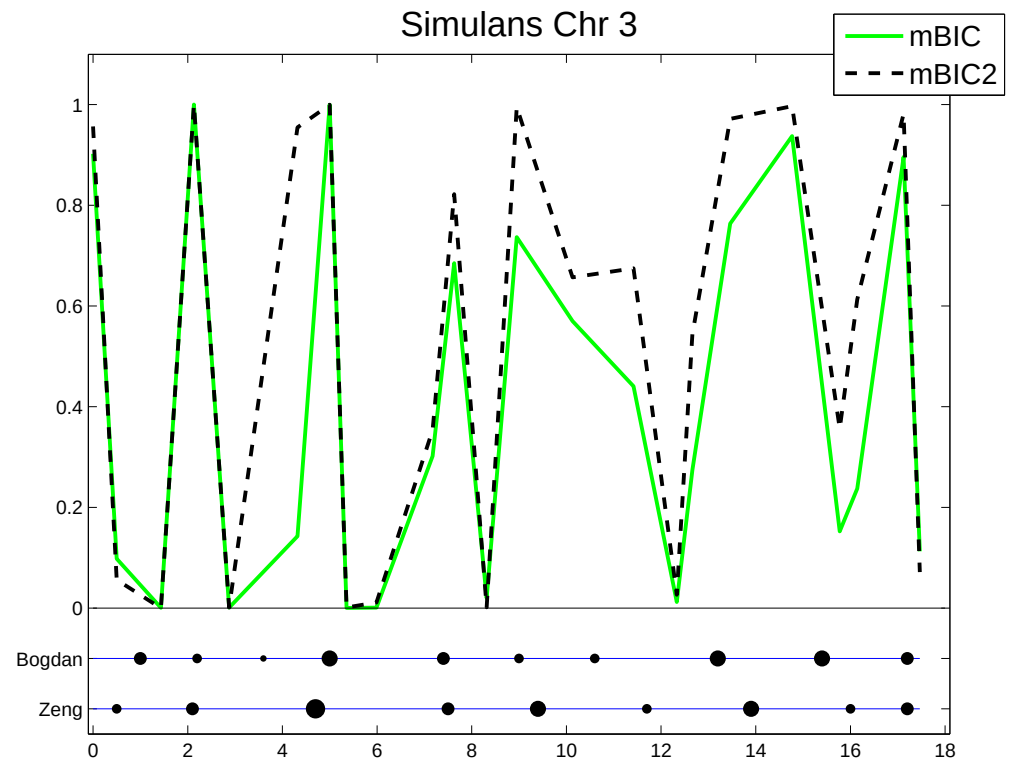
$$\mathrm{P}(j_r|Y) \approx \frac{\sum_{M\in\mathcal{M}_r^{Pool}} \exp\left(-\mathrm{mBIC}(M)/2\right)}{\sum_{M'\in\mathcal{M}^{Pool}} \exp\left(-\mathrm{mBIC}(M')/2\right)}$$

Marker posteriors from MA

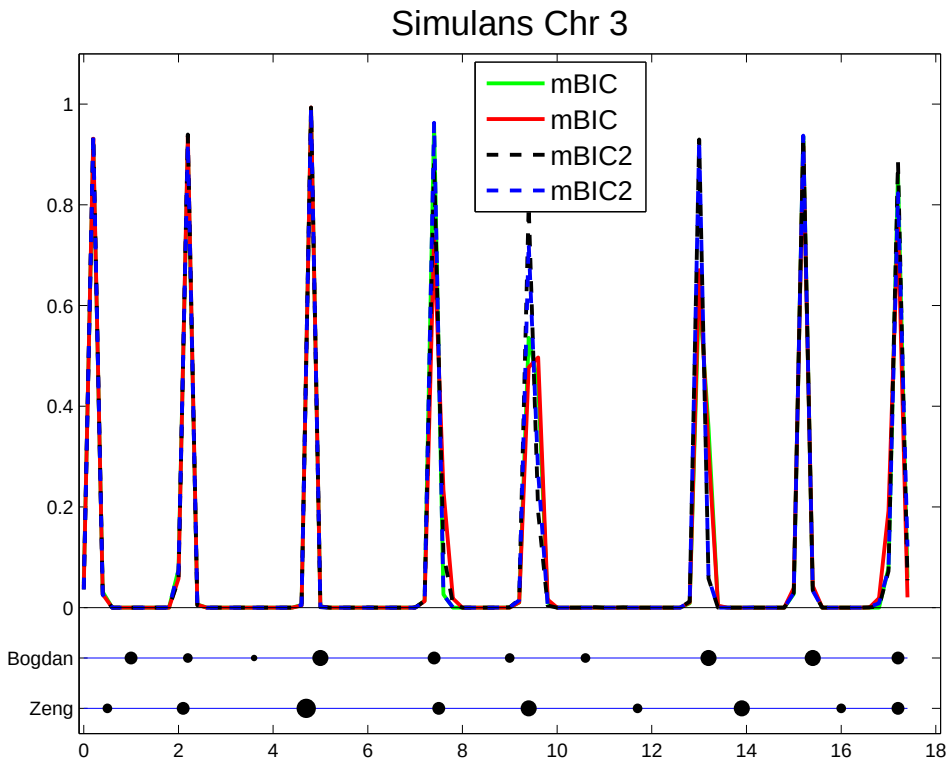
Example from real data analysis

Morphological differences in Drosophila Simulans (Zeng et al. (2000))

Search over markers



With imputation (IM)



Literature

Schwarz (1978) Estimating the Dimension of a Model. *Ann. Statist.* **6**(2), 461 - 464.

Benjamini, Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57**, 289 - 300.

Bogdan, Ghosh, Doerge (2004) Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitive trait loci. *Genetics*, **167**, 989 - 999.

Abramovich, Benjamini, Donoho, Johnstone (2006) Adapting to unknown sparsity by controlling the false discovery rate *Ann. Statist.***34**, 584 - 653.

Chen, Chen (2008) Extended Bayesian Information criteria for model selection with large model spaces. *Biometrika* **95**, 759 - 771.

Literatur

Baierl, Bogdan, Frommlet, Futschik, (2006). On locating multiple interacting quantitative trait loci in intercross designs. *Genetics* **173**, 1693 - 1703.

Bogdan, Frommlet, Biecek, Cheng, Ghosh, Doerge, (2008) Extending the Modified Bayesian Information Criterion (mBIC) to Dense Markers and Multiple Interval Mapping. *Biometrics* **64**, 1162 - 1169.

Bogdan, Chakrabati, Frommlet, Ghosh (2011) Asymptotic Bayes Optimality under sparsity of some multiple testing procedures. *Ann. Statist.* **39**, 1551 - 1579

Frommlet, Bogdan (2013) Some optimality properties of FDR controlling rules under sparsity, *Electron. J. Statist.* **7**, 1328-1368.

Frommlet, Ruhaltinger, Twarog, Bogdan (2011) Modified versions of Bayesian Information Criterion for genome-wide association studies. *CSDA*, **56**, 1038 - 1051

Frommlet, Ljubic, Arnardottir, Bogdan (2012) QTL Mapping Using a Memetic Algorithm with modifications of BIC as fitness function, *Stat. Appl. in Genet. and Molec. Biol.*, **11**(4), Article 2