Multiple regression - information criteria for large data bases

Małgorzata Bogdan

University of Wroclaw

Bologna, 14-17/05/2019

Multiple regression model when $n > p$

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N(0, \sigma^2 I_{n\times n})$$

$$\hat{\beta}_{LS} = argmin_{\beta\in R^p}||Y - X\beta||^2 = (X'X)^{-1}X'Y$$

$$\hat{\beta}_{LS} \sim N(\beta, \sigma^2(X'X)^{-1})$$

$$\hat{\sigma}^2 = s^2 = \frac{||Y - X\hat{\beta}_{LS}||^2}{n-p} = \frac{RSS}{n-p}$$

Selection of important variables

T-tests,

$$T_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \ ,$$

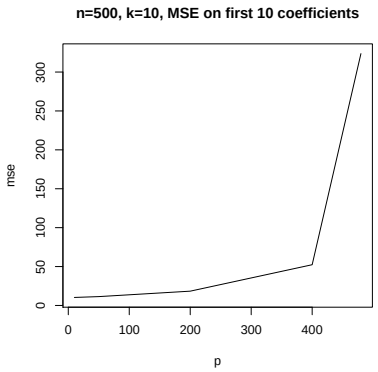
where $s(\hat{\beta}_i) = s^2(X'X)^{-1}[i,i]$

Problem - typically elements on the diagonal of $(X'X)^{-1}$ become large as p increases.

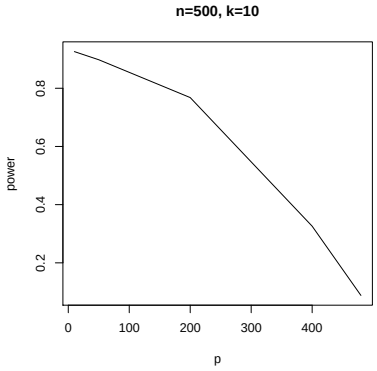
If elements of X are iid from $N(0, 1/\sqrt{n})$ then $X'X$ has a Wishart distribution and the elements on its diagonal have the expected value equal to 1.

But $(X'X)^{-1}$ has the inverse Wishart distribution and the expected values of the elements on the diagonal are equal to $\frac{n}{n-p-1}$ and become very large as p approaches n.

Inflation of MSE



Loss of Power



Model selection

Model selection in multiple regression - identification of important variables

Error in the training sample $RSS = ||Y - \hat{Y}||^2$ never increases when we add new variables into the model. Thus, minimization of RSS is not a good criterion for model selection.

Also, RSS is not a good measure of the prediction error.

Training and prediction error

Let's consider a new sample

$$Y^* = X\beta + \epsilon^* \ ,$$

where ϵ^* is independent on the noise term ϵ in the training sample
We use our training sample to build a good predictive model, i.e. the model which minimizes

$$PE = E||Y^* - \hat{Y}||^2$$

If $\mu = E(Y) = X\beta$, then $PE = E||\mu - \hat{\mu}||^2 + n\sigma^2 = E||\mu - \hat{Y}||^2 + n\sigma^2$

RSS measures the fit within the training sample, i.e. it adjusts to the specific realization of the noise term ϵ - this is overfitting. PE measures the fit with respect to the true expected value of Y, which indeed is an indication of predictive properties (i.e. how well we can predict new observations with different noise terms).

Prediction error in least squares regression

If $\hat{Y} = M_{n\times n} Y$ then

$$PE = E(RSS) + 2\sigma^2 Tr(M)$$

In least squares estimation

$$M = X(X'X)^{-1}X'$$

is the matrix of the orthogonal projection on the space spanned by columns of X and $Tr(M) = rank(X)$.

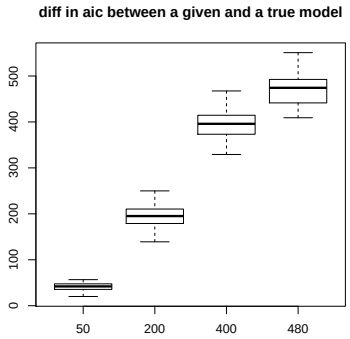
If $rank(X) = p$ then the unbiased estimator of the prediction error is equal to

$$\hat{PE} = RSS + 2\sigma^2 p \ .$$

Minimizing $\hat{PE}$ coincides with AIC criterion which suggests selecting the model for which $RSS + 2\sigma^2 p$ is minimal.

Properties of AIC (1)

In our example AIC identifies the true model among 5 models with different dimensions.



Can we use AIC to select variables in large data bases ?

Problem 1: Discrete optimization over 2^p of possible models - not doable in polynomial time.

In practice we often resort to heuristics which with large probability return models closed to being optimal.

Forward selection - we start from the empty model and add variables one by one. At each step we select the one which leads to the largest improvement of the criterion. We stop when the criterion is no longer improved.

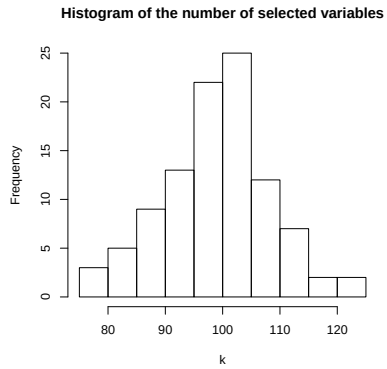
Backward elimination - we start from the full model and remove variables one by one until criterion is no longer improved.

Step-wise selection: alternating between forward selection and backward elimination

More complicated heuristics: genetic algorithms, simulated annealing etc.

Can we use AIC to select important variables in large data bases ?

bigstep - *R* library with many different search strategies, optimizing a variety of model selection criteria



Multiple testing explanation

Assume that $X'X = I$

Then AIC selects variables which satisfy

$$| < X_i, Y > | \geq \sqrt{2}\sigma \ .$$

When $\beta_i = 0$ then $< X_i, Y >=< X_i, \epsilon > \sim N(0, \sigma^2)$.

Thus probability of the type I error

$$P(X_i \text{ is selected} | \beta_i = 0) = 2(1 - \Phi(\sqrt{2})) = 0.16$$

When $p = 500$ and $k = 10$ we expect to see on average $490 \times 0.16 = 78$ false discoveries and the typical size of the selected model should be around k=88

In our simulations $k \approx 100$ due to additional disturbance by the sample correlations between columns of the design matrix and using the form of AIC with unknown σ

Would BIC help ?

BIC selects the model which minimizes

$$RSS + \sigma^2 k \log n$$

Thus BIC selects variables which satisfy

$$|< X_i, Y >| \geq \sqrt{\log n}\sigma \ .$$

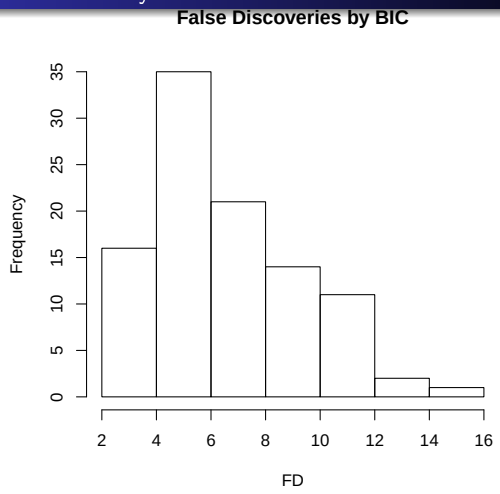
The probability of the type I error

$$P(X_i \text{ is selected} | \beta_i = 0) = 2(1 - \Phi(\sqrt{\log n}),$$

which for $n = 500$ is equal to 0.013

Thus we expect to see on average $p_0 * 0.013 = 490 * 0.013 \approx 6.5$ false discoveries

False Discoveries by BIC



Solution - multiple testing correction

In Risk Inflation Criterion (Foster and George 1994) the penalty depends on p

$$RSS + \sigma^2 2k \log p$$

Thus RIC selects variables which satisfy

$$|< X_i, Y >| \geq \sigma\sqrt{2 \log p} \ .$$

The probability of the type I error

$$P(X_i \text{ is selected} | \beta_i = 0) = 2(1 - \Phi(\sqrt{2 \log p})) \approx \frac{1}{\sqrt{\pi}} \frac{1}{p\sqrt{\log p}}.$$

Accuracy of approximation: for $p = 500$

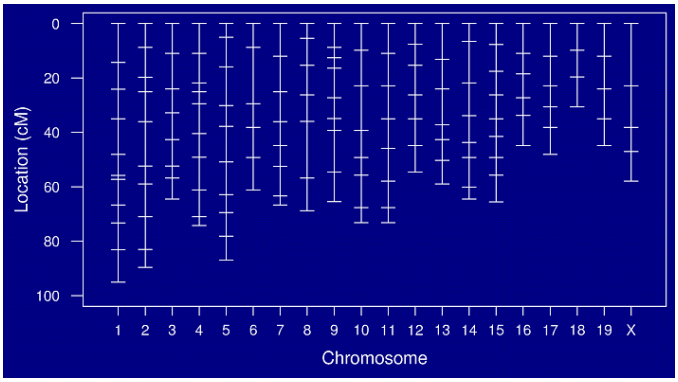
$$2(1 - \Phi(\sqrt{2 \log p}) = 0.000423, \quad \frac{1}{\sqrt{\pi}} \frac{1}{p\sqrt{\log p}} = 0.000453$$

Here the expected number of false discoveries is smaller than 1 and decreases with p

Modified BIC

- Motivation: QTL mapping and Genome Wide Association Studies
- Modified versions of BIC - Bayesian background and relationship to multiple testing.
- Simulation studies
- Asymptotic Optimality and Consistency

Locating Quantitative Trait Loci



Data for QTL mapping in backcross population and recombinant inbred lines

Y_i, $1 \leq i \leq n$ - trait values

Only two genotypes possible at a given locus

X_{ij}, $1 \leq i \leq n$, $1 \leq j \leq m$ - dummy variables encoding genotypes at m markers, $X_{ij} \in \{-1, 1\}$

Strong correlation between neigboring loci: backcross
d - distance in M , $\rho = e^(- 2d)$
$d = 0.1M$, $\rho = 0.82$
$d = 1M$, $\rho = 0.14$
Average chromosome length - 1.5 M, usually around 10-15 markers on each chromose
$m \approx 300$, $n > 200$

Data for GWAS

Three genotypes possible at a given locus
Usual coding

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

Weak and non-regular correlation between neigboring loci
Usually $n \approx k \times 100$ or $k \times 1000$, $m \approx k \times 100,000$

Multiple regression model

$$Y_i = \mu + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv} + \varepsilon_i, \qquad (1)$$

I - a subset of $N = \{1, \ldots, m\}$, U - a subset of $N \times N$,
$\epsilon_i \sim N(0, \sigma^2)$
Task : estimation of the number of influential genes and interaction effects

Bayesian Information Criterion (1)

M_i - i-th linear model

k_i - number of main effects, q_i - number of interactions

$k_i + q_i < n$

$\theta_i = (\beta_0, \beta_1, \ldots, \beta_{k_i}, \gamma_1, \ldots, \gamma_{q_i}, \sigma)$ - vector of model parameters

Bayesian Information Criterion (Schwarz, 1978)

maximize $BIC = \log L(Y|M_i, \hat{\theta}_i) - \frac{1}{2}(k_i + q_i) \log n$

If m is fixed, $n \to \infty$ and $X'X/n \to Q$, where Q is a positive definite matrix, then BIC is consistent - the probability of choosing the proper model converges to 1.

Surprise ? : - Broman and Speed (JRSS, 2002) report that BIC overestimates the number of regressors when applied to QTL mapping.

Explanation - Bayesian roots of BIC (1)

$f(\theta_i)$ - prior density of θ_i, $\pi(M_i)$ - prior probability of M_i

$m_i(Y) = \int L(Y|M_i, \theta_i) f(\theta_i) d\theta_i$ - integrated likelihood of the data given the model M_i

posterior probability of M_i : $P(M_i|Y) \propto m_i(Y)\pi(M_i)$

BIC neglects $\pi(M_i)$ and uses approximation

$$\log m_i(Y) \approx \log L(Y|M_i, \hat{\theta}_i) - 1/2(k_i + q_i + 2) \log n + R_i,$$

R_i is bounded in n.

Explanation - Bayesian roots of BIC (2)

neglecting $\pi(M_i) \equiv$ assigning the same probability to all models

$\equiv$ the prior on the number of effects is K is $B(m, \frac{1}{2})$

$E(K) = \frac{m}{2}$, $std(K) = \frac{\sqrt{m}}{2}$

distribution concentrated almost entirely on $[m/2 - 2\sqrt{m}, m/2 + 2\sqrt{m}]$

for $m = 400$ the prior distribution on K is almost entirely concentrated on $[160, 240]$

Modified version of BIC, mBIC

M. Bogdan, J.K. Ghosh,R.W. Doerge, *Genetics* (2004)

Solution - using an informative prior distribution on the number of main and interaction effects

Prior distribution on the number of main effects: $B(m, p_1)$

Prior distribution on the number of interactions: $B(N_e, p_2)$, where $N_e = m(m-1)/2$

$E(k) = mp_1 = c_1$, $E(q) = N_e p_2 = c_2$

mBIC: maximize

$$\log L(Y|\hat{\theta}) - \frac{1}{2}(k+q)\log(n) - k \log\left(\frac{m}{c_1} - 1\right) - q \log\left(\frac{N_e}{c_2} - 1\right)$$

Standard version of mBIC uses $c_1 = c_2 = 2.2$ to control the overall type I error at the level below 10%.

The overall type I error is approximately equally divided between main and interaction effects.

A similar $\log m$ penalty appears in RIC of Foster and George (1994)

Relationship to multiple testing - only main effects (1)

Orthogonal design: $X^T X = nI_{(m+1)\times(m+1)}$, (1)

BIC chooses those X_j's for which

$$\frac{n\hat{\beta}_j^2}{\sigma^2} > \log n$$

Under $\mathcal{H}_{0j}: \beta_j = 0, \quad Z_j = \frac{\sqrt{n}\hat{\beta}_j}{\sigma} \sim N(0,1)$

It holds that for large values of n

$$\alpha_n = 2P(Z_j > \sqrt{\log n}) \approx \sqrt{\frac{2}{\pi n \log n}}.$$

Relationship to multiple testing (2)

When n and m go to infinity and the number of true signals remains fixed, the expected number of "false discoveries" is of the rate $\frac{m}{\sqrt{n \log n}}$.

Corollary: BIC is not consistent when $\frac{m}{\sqrt{n \log n}} \to \infty$

.

Relationship to multiple testing (3)

Bonferroni correction for multiple testing : $\alpha_{n,m} = \frac{\alpha_n}{m}$

probability of detecting at least one "false positive": FWER $\leq \alpha_n$

$2(1 - \Phi(\sqrt{c_{Bon}})) = \frac{\alpha_n}{m}$

$$c_{Bon} = 2\log\left(\frac{m}{\alpha_n}\right)(1 + o_{n,m}) = (\log n + 2\log m)(1 + o_{n,m}) \ ,$$

where $o_{n,m}$ converges to zero when n or m tends to infinity.

$c_{mBIC} = \log n + 2\log\left(\frac{m}{c} - 1\right) \approx \log n + 2\log m - 2\log c$

Applications of mBIC for QTL mapping

1. Extending to intercross + a two-step version of mBIC : Baierl, Bogdan, Frommlet, Futschik *Genetics, 2006*

2. Robust versions based on M-estimates: Baierl, Futschik, Bogdan, Biecek *CSDA, 2007*

3. Rank version: Żak, Baierl, Bogdan, Futschik *Genetics, 2007*

4. Application for dense markers and interval mapping: Bogdan, Frommlet, Biecek, Cheng, Ghosh, Doerge, *Biometrics, 2008*

5. Application for the count data, based on the Zero-Inflated Generalized Poisson Regression: Earhardt, Bogdan, Czado *SAGMB, 2010*

Computer simulations(1)

Setting : $n = 200$, $m = 300$, entries of $X \sim N(0, \sigma = 0.5)$,

$k \sim Binomial(m, p)$, with $p = \frac{1}{30}$ $(mp = 10)$, $\beta_i \sim N(0, \sigma = 1.5)$,

$\varepsilon \sim N(0,1)$ and Tukey's gross error model:
$\varepsilon \sim Tukey(0.95, 100, 1) = 0.95 * N(0,1) + 0.05 * N(0,10)$.

Characteristics : Power, $FDR = \frac{FP}{AP}$, $MR = FP + FN$,
$l_2 = \sum_{j=1}^{m} (\beta_j - \hat{\beta}_j)^2$
mean value of the absolute prediction error based on 50 additional observations, d

Computer simulations, Bogdan et al. (QREI, 2008)

Tablica: **Results for 1000 replications.**

noise	N(0,1)			Tukey(0.95, 100, 1)		
citerion	BIC	mBIC	rBIC	BIC	mBIC	rBIC
FP	13.3	0.073	0.08	12.5	0.08	0.1
FN	1.84	2.97	3.45	3.95	6.11	4.29
Power	0.8155	0.7030	0.6586	0.6087	0.3923	0.5806
FDR	0.5889	0.0107	0.0116	0.6487	0.0210	0.0162
MR	15.1480	3.0410	3.5310	16.4440	6.1910	4.3910
l_2	2.3610	0.6025	0.8500	13.51	4.732	1.597
d	0.9460	0.8505	0.8687	1.714	1.503	1.298

$E|\varepsilon_1| \approx 0.8$, $E|\varepsilon_2| \approx 1.16$

Benjamini-Hochberg procedure, JRSS B, 1995

If $X^T X = nI_{(m+1)\times(m+1)}$ then $\hat{\beta}_j \sim N(\beta_j, \frac{\sigma^2}{n})$

H_{0j}: $\beta_j = 0$

p-values : $p_j = 2(1 - \Phi(|Z_j|))$, where $Z_j = \frac{\sqrt{n}\hat{\beta}_j}{\sigma}$

Benjamini and Hochberg procedure:
sorted p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$

$$k_F = argmax_j \left\{ p_{(j)} \leq \frac{j\alpha}{m} \right\} . \qquad (2)$$

BH rejects the hypothesis with p-values smaller or equal than $p_{(k_F)}$.

Asymptotic optimality under sparsity (1)

$\hat{\beta}_j \sim N(\beta_j, \sigma^2)$, $H_{0j} : \beta_j = 0$, $H_{Ai} : \beta_j \neq 0$

p - fraction of alternatives among all tests, sparsity: $p \to 0$ as $m \to \infty$

Abramovich, Benjamini, Donoho and Johnstone, Ann.Statist. 2006 - asymptotic minimax properties with respect to estimation loss $||\hat{\beta} - \beta||$, when $p_m \geq \frac{log^5 m}{m}$

Asymptotic optimality under sparsity (2)

Bogdan et al. Ann.Statist. 2011,
Frommlet and Bogdan, EJS 2013

Bayes risk, δ_0 - loss for type I error, δ_A - loss for type II error

$\hat{\beta} \sim N(\beta, \sigma^2/n)$, $n \geq C \log m$

$\beta_j \sim (1-p)\delta_0 + pF_A$, where F_A has a positive density at 0.

Bayes oracle $\to$ Bayes classifier

The rule is Asymptotically Bayes Optimal under Sparsity (ABOS) if $\lim \frac{R}{R_{opt}} \to 1$ (as $m \to \infty$)

Bonferroni correction at the FWER $\alpha \propto 1/\sqrt{n}$ is ABOS if $p \approx \frac{1}{m}$

BH at FDR $\alpha \propto 1/\sqrt{n}$ is ABOS if $p \to 0$ and $mp \to (0, \infty]$

mBIC2

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

In BH we look for $p_{(i)} < i\alpha_{n,m}$

this leads to $c_i^2 \approx (\log n + 2\log m - 2\log i)$

$$\sum_{i=1}^{k} \log i = \log(k!)$$

$$mBIC2 := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2k\log(m/4) + 2\log(k!)$$

Simulation results for GWAS (Frommlet, Ruhaltinger, Twarog and Bogdan, 2011, CSDA)

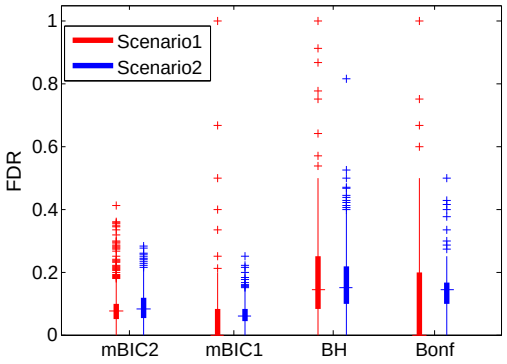
Population reference sample POPRES from dbGaP

- 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
  $Y = X_M\beta_M + \epsilon, \qquad \epsilon_i \sim (0, 1)$
- Simulation scenario:
  β_j equally distributed between 0.27 and 0.66

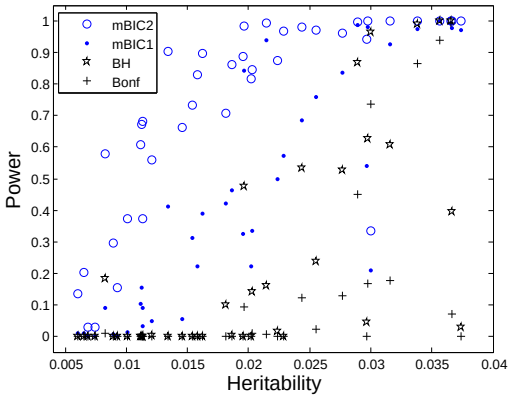
Search strategy

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 the model selection criterion
4. False positive - correlation with a causal SNP<0.9

FDR



Power



Extended BIC, EBIC

J. Chen, Z. Chen, *Biometrika* (2008)

Standard version - uniform prior on the number of main effects

$$EBIC := 2\log(L(Y|\hat{\theta})) - k\log(n) - 2\log\binom{m}{k} .$$

Caution - in EBIC $E(K) = \frac{m}{2}$.

Relationship between mBIC, mBIC2 and EBIC

If $\frac{\log k_{max}}{\log m} \to 0$ then $\frac{pen(EBIC(k))}{pen(mBIC(k))} \to 1$ uniformly for $k \in \{1, \ldots, k_{max}\}$

If $\frac{k_{max}}{m} \to 0$ then $\frac{pen(EBIC(k))}{pen(mBIC2(k))} \to 1$

mBIC2 is asymptotically equivalent to the Bayes rule based on the uniform prior on $\{0, \ldots, k_{max}\}$, where $\frac{k_{max}}{m} \to 0$.

Consistency (1)

Chen and Chen, 2008 - fixed true model dimension p_0, fixed maximal size of the model to search K

Identifiability condition: $\mu = EY$,
$H(s) = X(s)(X(s)^T X(s))^{-1} X(s)^T$, $\Delta_n(s) = ||(I - H(s))\mu||^2$,

$$\lim_{n\to\infty} \min \left(\frac{\Delta_n(s)}{\log n} : s \not\subset s_0, dim(s) \leq K \right) = \infty$$

Foygel and Drton, 2012 - random covariates,
There exists positive constants $a_1 < a_2$ such that for all $|J| \leq 2K$ the eigenvalues of $E[X_J X_J^T]$ are within $[a1, a2]$. The small true coefficients have bounded decay.

Consistency (2)

Chen and Luo, 2011, $p_0(n) \to \infty$, $K(n) \to \infty$,

$$\lim_{n\to\infty} \min \left\{ \frac{\Delta_n(s)}{p_0(n) \ln m_n} : s \not\subset s_0, dim(s) \leq K(n) \right\} = \infty \ ,$$

where $K_n = k p_0(n)$ for some fixed $k > 1$, $p_0(n) \ln m_n = o(n)$ and $\frac{\ln p_0}{\ln m_n} \to \delta \geq 0$.

Szulc, PMS, 2012 - showed consistency of mBIC and mBIC2 under slightly stronger assumptions

Open problem - asympotic optimality under non-orthogonal designs

Dense markers - Bogdan et al. (Biometrics, 2008)

Feingold, Brown and Siegmund, Genetics, 1993 - backcross

$$\begin{aligned} \alpha &= P_{H_0} \left(max_{j\in\{1,\ldots,p\}} LRT_j > c \right) \\ &\approx 1 - exp(-2[1 - \Phi(\sqrt{c})]) - 0.04 L \sqrt{c} \nu \left(\sqrt{0.04\delta} \right) \ , \end{aligned}$$

where

$$\nu(x) \approx e^{-0.583x} \ .$$

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

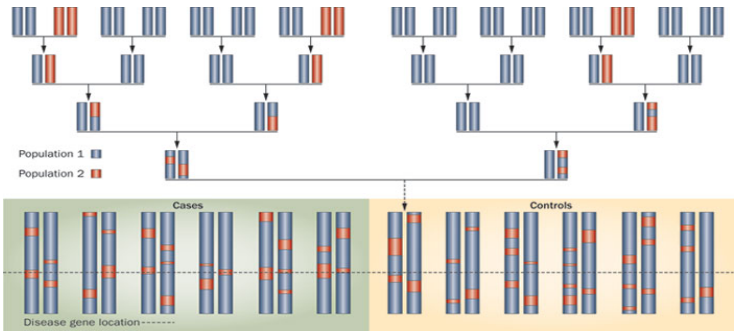
$$\alpha = P_{H_0} \left(\max_{i\in\{1,\ldots,p^{eff}\}} LRT_j > c \right) \approx 1 - \left[1 - 2 \left(1 - \Phi(\sqrt{(c)}) \right) \right]^{p^{eff}} \ .$$

The effective number of tests can be calculated as

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

Admixtures, Szulc, B,Frommlet, Tang (2017)

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 "causałloci 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 "causałloci, 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

If $\mu_b > \mu_B$, 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:

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

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

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:

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

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

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

$$Y_i = \beta_0 + \beta_Q Q_i + \sum_{j \in I} \beta_{Xj} X_{ij} + \sum_{j \in J} \beta_{Zj} Z_{ij} + \varepsilon_i, \qquad (3)$$

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).

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

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

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

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)

Tablica: 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:

$$\begin{aligned} \text{mBIC2:} n\log RSS &+ k_j(\log n + 2\log(m/4)) - 2\log(k_j!) \quad (4)\\ &+ \tilde{k}_j(\log n + 2\log(m^{eff}/4)) - 2\log(\tilde{k}_j!)\ , (5)\end{aligned}$$

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 $mBIC2$.

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, $m^{eff} = 4722$

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

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

Żecombination"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

Tablica: 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

Tablica: 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

ĆausalŚNPs 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

Tablica: 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)

Tablica: 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

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

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

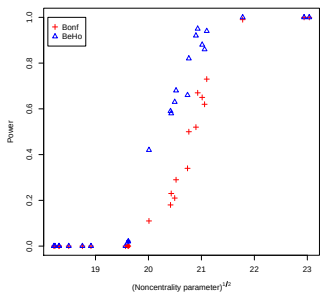
$Cov(Y - \beta_Q Q, X_1) = \beta_1 VarX_1 + \sum_{i=2}^{k} \beta_i Cov(X_1, X_i) + Cov(X_1, \epsilon)$

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

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

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

Power vs noncentrality parameter



Acknowledgments

1. Department of Statistics, Purdue University: Rebecca W. Doerge, Jayanta K. Ghosh, Riyan Cheng.
2. Institute of Statistics and Decision Support Systems, University of Vienna: Andreas Futschik, Florian Frommlet, Andreas Baierl, Felix Ruhaltinger.
3. Department of Statistics, University of Technology, Munich: Vinzenz Erhardt, Claudia Czado.
4. Indian Statistical Institute, Kolkata: Arijit Chakrabarti.
5. Institute of Mathematics and Computer Science, Wroclaw University of Technology: Przemyslaw Biecek, Malgorzata Zak-Szatkowska, Piotr Twarog, Piotr Szulc.
6. Erasmus University: Magdalena Murawska.
7. Department of Genetics, Stanford University: Hua Tang