Adaptive EBIC and its application to genome-wide association studies

Zehua Chen
National University of Singapore

jointly with

Jiahua Chen
University of British Columbia

OUTLINE

- Tournament screening cum EBIC procedure for GWAS
- EBIC and its consistency
- Adaptive EBIC procedure
- Numerical Studies

1. Tournament screening cum EBIC procedure for GWAS

• GWAS and small-n-large-p problem

In genome-wide association studies, there are three characteristics:

1. The number of features (SNPs, genes, etc.) to consider is huge.
2. The sample size is relatively small.
3. The causal features are sparse.

GWAS is a typical feature selection problem with small-n-large-p and sparse causal features.

- **Drawbacks of single-locus approaches**

  The control of family-wise type I error is irrelevant.

  The family of hypotheses in traditional multiple tests is equivalent to a SINGLE overall hypothesis. It makes sense to control the family-wise type I error.

  In GWAS, the hypotheses in multiple tests are independent, the control of family-wise type I error is un-necessary.

  The Bonferroni adjusted critical value is problematic. It turns the advantage of large-p into a disadvantage — the larger the p, the harder a causal feature can be detected.

  It is prone to have more false discoveries because of the high spurious correlation in a small-n-large-p sample.

- **TS cum lower dimensional EBIC procedure**

  Dimension reduction is a natural way to proceed with small-n-large-p problems.

  Three components of the TS cum lower dimensional EBIC procedure:

  1. The tournament screening procedure which screens out obvious non-significant features.
  2. A lower dimensional approach to search candidate models.
  3. The EBIC as the criterion for model selection.

- **Three desired properties:**

  1. The sure-screening property, i.e., causal features must be retained by the screening procedure with probability 1 asymptotically.

     The tournament screening has the sure-screening property, see Chen and Chen (2009a).

  2. The low-dimensional search must pass the model consisting of the exact causal features.

     This can be satisfied if a penalized likelihood method has the oracle property and the method is used to form nested models.

  3. The model selection criterion must be consistent.

     The consistency property is satisfied by EBIC, see Chen and Chen (2008, 2009b).

2. The EBIC and its consistency

- **Traditional model selection criteria fail in small-n-large-p problems**

  The criteria AIC, BIC, CV, etc. tend to select too many superfluous features due to the phenomena of high spurious correlation when the number of features is huge.

  – AIC, CV select models by minimizing prediction error, they do not discriminate causal and non-causal features, they don't care how many features are selected.

  – BIC favors models with more features when dimensionality of feature space is huge.

- **Extended Bayes information criteria (EBIC)**

  The Bayesian paradigm

  The prior on models: $p(s)$.

  The prior on parameters: $\pi\{\boldsymbol{\beta}(s)\}$.

  The marginal density of the data $\boldsymbol{Y}$ given the priors is:

$$m(\boldsymbol{Y}|s) = \int f\{\boldsymbol{Y}; \boldsymbol{\beta}(s)\}\pi\{\boldsymbol{\beta}(s)\}d\boldsymbol{\beta}(s),$$

  The posterior probability of a model s:

$$p(s|\boldsymbol{Y}) = \frac{m(\boldsymbol{Y}|s)p(s)}{\sum_{s\in\mathcal{S}} p(s)m(\boldsymbol{Y}|s)}.$$

  The Bayesian paradigm is to choose the model with the largest posterior probability. It entails the minimization of

$$-2\ln L(\hat{\boldsymbol{\beta}}(s)) + \nu(s)\ln n - 2\ln p(s).$$

BIC and its drawbak:

$$BIC(s) = -2 \ln L(\hat{\boldsymbol{\beta}}(s)) + \nu(s) \ln n.$$

It takes $p(s)$ as a constant free of s.

This prior favors models with more features as illustrated below:

Partition the model space as

$$\mathcal{S} = \cup_{j=0}^{p} \mathcal{S}_j,$$

$\mathcal{S}_j$: the set of models which contain exactly j features.

Note that

$$\text{Pior}(S_j) \propto \binom{p}{j}.$$

$$\text{Thus} \quad \text{Pior}(S_2) = \frac{(p-1)}{2} \text{Pior}(S_1), \quad \cdots$$

Definition of EBIC:

For $s \in \mathcal{S}_j$,

$$\begin{aligned} \text{EBIC}_\gamma(s) \;=\; & -2 \ln L_n(\hat{\boldsymbol{\beta}}(s)) + \nu(s) \ln n \\ & +2 \ln[\tau(\mathcal{S}_j)]^\gamma, \;\; \gamma \geq 0. \end{aligned}$$

$\tau(\mathcal{S}_j)$: number of models in $\mathcal{S}_j$,

EBIC is equivalent to take

$$p(s) \propto \tau(\mathcal{S}_j)^{-\gamma}, \quad \text{for} \quad s \in \mathcal{S}_j.$$

- **The consistency theorem**:

  Under proper conditions, for $p = O(n^{\kappa})$, and $\gamma > 1 - \frac{1}{2\kappa}$,

$$P\{\min\{\mathrm{EBIC}_{\gamma}(s) : \nu(s) = j\} > \mathrm{EBIC}_{\gamma}(s_0)\} \to 1$$

  for $j = 1, 2, \ldots, K(> \nu(s_0))$, as $n \to \infty$, where s_0 is the model consisting of exactly the causal features, s is any other model and $\nu(s)$ denotes the number of features in model s.

- **Condition for Gaussian model:**

$$\lim_{n\to\infty} \min\{(\log n)^{-1}\Delta_n(s) : s \neq s_0, \nu(s) \leq K_0\} = \infty,$$

  where

$$\Delta_n(s) = \|[I - X_n(s)\{X_n^{\mathrm{t}}(s)X_n(s)\}^{-1}X_n^{\mathrm{t}}(s)]\mu_n\|^2.$$

- **Condition for GLM with canonical link**:

  Let

$$H_n(\boldsymbol{\beta}) = -\frac{\partial^2 l_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau}.$$

  where l_n is the log likelihood function.

  – There exist $c_1 > 0, c_2 > 0$ such that for all sufficiently large n,

$$c_1 \leq \lambda_{\min}(n^{-1} H_n(\boldsymbol{\beta}_0(s))) \leq \lambda_{\max}(n^{-1} H_n(\boldsymbol{\beta}_0(s))) \leq c_2,$$

    for all s such that $\nu(s) \leq K$.

  – For any $\epsilon > 0$, there exists $\delta > 0$ such that, when n is sufficiently large,

$$(1-\epsilon) H_n(\boldsymbol{\beta}_0(s)) \leq H_n(\boldsymbol{\beta}(s)) \leq (1+\epsilon) H_n(\boldsymbol{\beta}_0(s))$$

    for all s and $\boldsymbol{\beta}(s)$ such that $\nu(s) \leq K$ and $\|\boldsymbol{\beta}(s) - \boldsymbol{\beta}_0(s)\| \leq \delta$.

  – Denote by x_{ij} the jth component of $\boldsymbol{x}_i$.

$$\max_{1 \leq j \leq P} \max_{1 \leq i \leq n} \left\{ \frac{x_{ij}^2}{\sum_{i=1}^n x_{ij}^2 \sigma_i^2} \right\} = o((\log n)^{-1}).$$

3. Adaptive EBIC procedure

- **Issues on the choice of γ**

  Two measures are of primary concern for feature selection:

  FDR — false discovery rate,
  PDR — positive discovery rate.

  The implication of the consistency of EBIC: if the screening procedure has a sure screening property and the penalized likelihood has a oracle property, then FDR and PDR of the feature selection approach will converge to 0 and 1 respectively as sample size goes to infinity, while γ is in the convergence range.

  For finite sample size, different values of gamma result in different FDR and PDR.

  A choice of γ is in order.

• A procedure for assessing FDR

1. For a given value of γ, apply the TS cum EBIC procedure to select features and estimate effects and their variances.

2. Generate a number m^* according to a Poisson random variable with mean being the number of features selected in step 1.

3. Sample m^* pseudo-causal features from the original feature set.

4. Assign evenly the estimated effects to the m^* pseudo causal features as the expected effects. Generate the pseudo-effects according to normal distributions with means and variances as the estimated effects and variances.

5. Generate the pseudo-responses using the assumed model and the generated pseudo-causal features and pseudo-effects.

6. Apply the selection procedure with the same γ value to the generated data. Compare the resultant features with the pseudo-causal features to compute FDR.

7. Repeat steps 2-6 for a large number of times. Take the average of the simulated FDR as the estimate of the true FDR.

- **Final feature selection based on estimated FDR**.

  Usually there are only a few different models which can be selected by EBIC with different γ values, each model corresponding to a range of γ values.

  The upper bound of each such range can be taken as a measure of the relative significance of the features contained in the corresponding model.

  The FDR with the upper bound for each model is evaluated by the procedure in the previous slide.

  Eventually, the final model is selected by a consideration of the FDR and the number of selected features.

4. Numerical studies

- **Simulation studies**

  Simulation settings: feature variables are generated in batches of 50, each batch is generated as a multivariate normal vector with mean zero, varinace 1 and correlation $\rho^{|i-j|}$. The response variable is generated as a normal variable by a linear predictor and a error term. The effects of the linear predictor are generated for each replication as $\beta = (-1)^u * (a + \text{abs}(\text{rnorm}(\text{pm})))$ where u is a Bernoulli vector with probability of success 0.4, $a = 5 * \log(n)/\sqrt{n}$, pm is the number of causal effects. The standard deviation of the error term is determined such that a certain heritability is achieved. Simulation size is 500.

Simulation 1: Number of causal features = 8, total number of features = 1000, sample size = 200.

		ρ			
		0		0.75	
σ	γ	FDR	$\widehat{\text{FDR}}$	FDR	$\widehat{\text{FDR}}$
5.5	0.50	0.134	0.159	0.215	0.245
	0.75	0.074	0.080	0.173	0.178
	1.00	0.040	0.041	0.146	0.132
8.0	0.50	0.176	0.212	0.274	0.272
	0.75	0.069	0.097	0.206	0.168
	1.00	0.024	0.088	0.151	0.144

Simulation 2: Number of causal features = 18, total number of features = 1500, sample size = 400.

		ρ			
		0		0.75	
σ	γ	FDR	$\widehat{\text{FDR}}$	FDR	$\widehat{\text{FDR}}$
1.0	0.50	0.022	0.029	0.091	0.103
	0.75	0.008	0.009	0.090	0.098
	1.00	0.004	0.004	0.089	0.096
2.5	0.50	0.024	0.032	0.131	0.160
	0.75	0.009	0.012	0.127	0.148
	1.00	0.005	0.006	0.124	0.142

- **An real example**

  Data description:

  The quantitative trait: A measure on the mRNA expression level of the EBNA-3A gene in the lymphoblastoid cell lines (LCLs) transformed from B lymphocytes extracted from blood samples of individuals.

  Sample: 233 individuals from 16 pedigrees.

  Candidate features: Genotypes at 2155 SNPs spread over 23 chromosomes.

  A data-clean procedure retains 1414 SNPs.

The results:

γ	No. of SNP	FDR
0.00 - 0.03	20	—
0.04 - 0.26	18	—
0.27 - 0.44	9	—
0.45 - 0.50	8	0.4386
0.51 - 1.22	2	0.0735
1.23 - 1.49	1	0.0079
1.50 -	0	—

Indices of selected SNP:

$$42,\ 291,\ 349,\ 685,\ 790,\ 835^{**},\ 923,\ 1231^{*}$$

** appears in all three models.

* appears in two models.

Others only appear in one model.

References

Chen and Chen (2008),
Biometrika.

Chen and Chen (2009a),
Science in China, Series A.

Chen and Chen (2009b),
submitted, under revision.

Chen and Chen (2009c),
manuscript