

AN ADAPTIVE SCREEN-SELECTION PROCEDURE FOR FEATURE SELECTION IN SMALL- n -LARGE- p PROBLEMS

Science Lunchtime Talk
Department of Statistics & Applied Probability
National University of Singapore

OUTLINE

- 1 SMALL- n -LARGE- p AND ITS CHALLENGES
- 2 SCREENING-SELECTION APPROACH
- 3 ADAPTIVE SCREENING-SELECTION
- 4 NUMERICAL STUDIES
- 5 CONCLUSION
- 6 REFERENCES

SMALL- n -LARGE- p AND ITS CHALLENGES

Small- n -large- p is common in conventional statistics

- **Example 1:** Genome-wide association study (GWAS) In a typical GWAS, hundreds of thousands SNPs are available for investigation. But the sample size is relatively small.
- **Example 2:** Microarray data of gene expression levels. A microarray chip contains expression levels of thousands of genes. But the number of chips are small.
- **Example 3:** Stock market data. Thousands of shares are available. But for the design of investment portfolios, only short term data is useful, say, in months, which results in small sample sizes.

SMALL- n -LARGE- p AND ITS CHALLENGES

The characteristics of small- n -large- p problem

- The number of features to consider is huge — **large** p .
- The sample size is small — **small** n .
- The number of causal or relevant features is only a few — **sparsity**.

The causal features are like needles hidden in a haystack.

SMALL- n -LARGE- p AND ITS CHALLENGES

The challenges

- The spurious correlations are high even if all the features are stochastically independent when the dimensionality of the feature space is huge.
- The required computation amount is prohibitive.
- The off-shelf statistical methods are either inapplicable or inefficient.

SCREENING-SELECTION APPROACH

Dimensionality deduction is a natural way to proceed. The screen-selection approach provides a solution.

The components and desired properties of a screening-selection approach

- A screening procedure to reduce the feature set to a small one. The procedure must retain causal features all most surely, the property is referred to as the sure-screening property.
- A lower dimensional procedure to search for candidate models. A ideal such search procedure must result in a set of candidate models which contains the exact model consisting of all and only the causal features.
- A model selection criterion. The criterion must be able to identify the exact model from other models. This property is referred to as selection consistency.

SCREENING-SELECTION APPROACH

Screening procedures

- **Sure independence screening (SIS) (Fan and Lv)**
 - Fit a univariate model for each feature and compute the p -value of the model.
 - Order the p -values in ascending order.
 - Retain the features with orders below a cut-off point.

SCREENING-SELECTION APPROACH

- **Tournament screening (TS) (Chen and Chen)**

- The screening procedure mimics the competitions in a tournament.
- At each round, all the features subjected to selection are randomly divided into groups. A given number of features in each group are selected by a penalized likelihood mechanism.

E.g., Lasso:

$$l_p(\beta) = -2 \log L(\mathbf{y}, X' \beta) + \lambda \sum_{j=1}^p |\beta_j|.$$

- All selected features in the previous round are mixed and randomly divided into groups again. A new round of selection is carried out.
- The procedure continues until a desired low dimension of the feature space is reached.

SCREENING-SELECTION APPROACH

• Pros and cons of SIS and TS

- Both SIS and TS satisfy the sure screening property.
- SIS completely inherits the spurious correlation in the original feature set, which entails difficulty to identify the exact model in the low dimensional process.
- TS can reduce the spurious correlation to a great extent.

SCREENING-SELECTION APPROACH

Low-dimensional model search procedures

- Traditional stepwise procedure — greedy and computation intensive. It does not guarantee the exact model is among the models assessed.
- Penalized likelihood approach: A penalized likelihood is used to rank the features by starting with a large penalizing parameter and then gradually relaxing the parameter. Thus a sequence of nested models can be formed. With proper version of the penalized likelihood, the exact model is one of the models in the sequence.

SCREENING-SELECTION APPROACH

Model selection criteria

The traditional criteria AIC, BIC, CV, etc. tend to select too many superfluous features due to high spurious correlation when the number of features is huge. Selection consistency is not satisfied by these criteria.

- 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.
- For feature selection in small- n -large- p problems, a family of extended BIC (EBIC) is developed by Chen and Chen (2008).

SCREENING-SELECTION APPROACH

The Bayesian paradigm

- The prior on models: $p(s)$.
- The prior on parameters: $\pi\{\beta(s)\}$.
- The marginal density of data give a model:

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

- The posterior probability of a model s :

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

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

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

SCREENING-SELECTION APPROACH

BIC and its drawback

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

$$\text{Pior}(\mathcal{S}_j) \propto \binom{p}{j}.$$

$$\text{Thus} \quad \text{Pior}(\mathcal{S}_2) = \frac{(p-1)}{2} \text{Pior}(\mathcal{S}_1), \quad \dots$$

SCREENING-SELECTION APPROACH

Definition of EBIC

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

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

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

- EBIC is equivalent to take

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

SCREENING-SELECTION APPROACH

Consistency of EBIC

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

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

for $j = 1, 2, \dots, K(> \nu(s_0))$, as $n \rightarrow \infty$, 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:**

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

where

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

SCREENING-SELECTION APPROACH

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

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

where l_n is the log likelihood function.

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

$$c_1 \leq \lambda_{\min}(n^{-1}H_n(\beta_0(s))) \leq \lambda_{\max}(n^{-1}H_n(\beta_0(s))) \leq c_2,$$

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

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

$$(1 - \epsilon)H_n(\beta_0(s)) \leq H_n(\beta(s)) \leq (1 + \epsilon)H_n(\beta_0(s))$$

for all s and $\beta(s)$ such that $\nu(s) \leq K$ and $\|\beta(s) - \beta_0(s)\| \leq \delta$.

- Denote by x_{ij} the j th component of \mathbf{x}_i .

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

SCREENING-SELECTION APPROACH

TS cum EBIC screening-selection in a nutshell

- The tournament screening procedure is used for the screening stage. TS has the sure screening property and is able to reduce spurious correlation.
- Penalized likelihood ranking is used for model searching. The method guarantees the inclusion of the exact model in the candidate models to be assessed.
- EBIC is used for final model selection. It is selection consistent.
- The FDR and PDR of the procedure converge to 0 and 1 respectively, as sample size gets large.
 - FDR — proportion of false discoveries in the set of selected features.
 - PDR — proportion of selected causal features among all causal features.

ADAPTIVE SCREENING-SELECTION

A procedure for the estimation of FDR with finite sample size

- For a given value of γ , apply the screening-selection procedure. Obtain the estimates of the effects of the selected features and their variances.
- Generate a number m^* according to a Poisson random variable with mean being the number of features selected in step 1.
- Sample m^* pseudo-causal features from the original feature set.
- 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.

ADAPTIVE SCREENING-SELECTION

A procedure for the estimation of FDR (cont.)

- Generate the pseudo-responses using the assumed model and the generated pseudo-causal features and pseudo-effects.
- Apply the screening-selection procedure with the same γ value to the generated data. Compare the resultant features with the pseudo-causal features to compute FDR.
- Repeat steps 2-6 for a large number of times. Take the average of the simulated FDR as the estimate of the true FDR.

ADAPTIVE SCREENING-SELECTION

The adaptive procedure

- Apply the screening-selection procedure with γ values equally spaced in an interval $[0, G]$. Identify the subintervals which result in different models.
- For the upper bound of each sub-interval, estimate the FDR by applying the procedure described in the previous slide.
- Use the estimated FDR together with the number of selected features of each model to make final selection and guide further studies. Features in the model with small FDR can be taken as confirmed causal ones. Features in models with larger but still reasonable FDR can be taken for further studies.

NUMERICAL STUDIES

Simulation settings:

Feature variables are generated in batches of 50, each batch is generated as a multivariate normal vector with mean zero, variance 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.

NUMERICAL STUDIES

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

NUMERICAL STUDIES

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

NUMERICAL STUDIES

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.

NUMERICAL STUDIES

- **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 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.

CONCLUSION

- The screening-selection approach is a natural way for feature selection with large number of features.
- The family of EBIC is developed especially for feature selection with high-dimensional feature space.
- The family is consistent in a range of γ values under assumptions of GLM with some mild conditions.
- The addaptive screening-selection procedure with EBIC is efficient.
- Numerical studies vindicated the validity of the addaptive procedure.

REFERENCES

- J. Chen and Z. Chen. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Space. *Biometrika*, **95**: 759-771.
- Z. Chen, J. Chen. (2009). Tournament Screening cum EBIC for Feature Selection with High Dimensional Feature Spaces. *Science in China, Series A*, **52**(6): 1327-1341.
- Chen, Z. and Chen, J. Adaptive EBIC for feature selection with high-dimensional feature spaces, working paper.
- Chen, J. and Chen, Z. Extended BIC for small- n -large- P sparse GLM, Manuscript.