

FEATURE SELECTION WITH HIGH-DIMENSIONAL DATA: CRITERIA AND PROCEDURES

Zehua Chen

Department of Statistics & Applied Probability
National University of Singapore

Conference in Honour of Grace Wahba, June 4-6, 2014

OUTLINE

- 1 INTRODUCTION
- 2 SELECTION CRITERIA
- 3 SELECTION PROCEDURES
- 4 THE ORACLE PROPERTY
- 5 SIMULATION STUDIES
- 6 REFERENCES

HIGH-DIMENSIONAL DATA AND FEATURE SELECTION

- High-dimensional data arises from many important fields such as genetic research, financial studies, web information analysis, etc.
- High-dimensionality causes difficulties in feature selection: effect of relevant features could be masked by irrelevant features; hard to distinguish between relevant and irrelevant variables; computationally challenging.
- Feature selection involves two components: selection criteria and selection procedures.
- Traditional criteria and procedures in general no longer work for high-dimensional data.
- In this talk, we focus on the selection criteria and procedures for high-dimensional feature selection.

TRADITIONAL CRITERIA

- Traditional criteria include Akaike's information criterion (AIC), Bayes information criterion (BIC), Mallow's C_p , cross validation (CV) and generalized cross validation (GCV).
- Traditional criteria are in general too liberal for high-dimensional feature selection.
- Theoretically, traditional criteria are not selection consistent in the case of high-dimensional data.
- In this talk, we focus on an extension of BIC for high-dimensional feature selection.

THE BAYESIAN FRAMEWORK AND BIC

- Prior on model s : $p(s)$.
- Prior on the parameters of model s : $\pi(\beta(s))$.
- Probability density of data given s and $\beta(s)$: $f(\mathbf{Y}|\beta(s))$.
- Marginal density of data given s :

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

- Posterior probability of s :

$$p(s|\mathbf{Y}) = \frac{m(\mathbf{Y}|s)p(s)}{\sum_s m(\mathbf{Y}|s)p(s)}.$$

- BIC is essentially $-2 \ln p(s|\mathbf{Y})$.
- In the derivation of BIC: (i) $p(s)$ is taken as a constant; (ii) $m(\mathbf{Y}|s)$ is approximated by Laplace approximation.

DRAWBACK OF CONSTANT PRIOR

- Partition the model space as

$$\mathcal{S} = \cup_j \mathcal{S}_j.$$

With the constant prior, the prior probability on $\mathcal{S}_j$ is proportional to $\tau(\mathcal{S}_j)$, the size of $\mathcal{S}_j$.

- If $\mathcal{S}_j$ is the set of models consisting of exactly j variables, then

$$p(\mathcal{S}_j) = c \binom{p}{j}.$$

In particular,

$$p(\mathcal{S}_1) = cp,$$

$$p(\mathcal{S}_2) = cp(p-1)/2 = P(\mathcal{S}_1)(p-1)/2, \dots$$

- The constant prior prefers models with more variables.

THE EXTENDED BIC (EBIC)

- Chen and Chen (2008) considered the prior

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

which leads to EBIC.

- General definition: Let $\mathcal{S}_j$ be the class of models with the same natures indexed by j . For $s \in \mathcal{S}_j$, the EBIC for model s is defined as

$$\text{EBIC}_\gamma(s) = -2 \ln L_n(\hat{\beta}(s)) + |s| \ln n + 2\gamma \ln \tau(\mathcal{S}_j), \quad \gamma = 1 - \xi \geq 0,$$

where $|\cdot|$ is the cardinality of a set, $\hat{\beta}(s)$ is the MLE of the parameters under model s .

- For additive main-effect models, $\mathcal{S}_j$ is classified as the set of models consisting of exactly j variables. Then, for $s \in \mathcal{S}_j$,

$$\text{EBIC}_\gamma(s) = -2 \ln L_n(\hat{\beta}(s)) + j \ln n + 2\gamma \ln \binom{p}{j}.$$

EBIC (CONT.)

- For interactive models, let $j = (j_M, j_I)$ and S_j be classified as the set of models consisting of exactly j_M main-effect features and j_I interaction features, then, for $s \in S_j$,

$$\begin{aligned} & \text{EBIC}_{\gamma_M \gamma_I}(s) \\ &= -2 \ln L_n(\hat{\beta}(s)) + (j_M + j_I) \ln n + 2\gamma_M \ln \binom{p}{j_M} + 2\gamma_I \ln \binom{\frac{p(p-1)}{2}}{j_I}, \end{aligned}$$

with a modification on the prior in the general definition.

- For q -variate regression models, let $j = (j_1, \dots, j_q)$, and S_j be the models such that there are exactly j_k covariates for the k th component of the response variable, then for $s \in S_j$,

$$\begin{aligned} & \text{EBIC}_{\gamma}(s) \\ &= -2 \ln L_n(\hat{\beta}(s)) + \sum_{k=1}^q j_k \ln n + 2\gamma \sum_{k=1}^q \ln \binom{p}{j_k}. \end{aligned}$$

EBIC (CONT.)

- For Gaussian graphical models, let S_j be classified as the set of models consisting of exactly j edges, then, for $s \in S_j$,

$$\text{EBIC}_\gamma(s) = -2 \ln L_n(\hat{\beta}(s)) + j \ln n + 2\gamma \ln \binom{\frac{p(p-1)}{2}}{j}.$$

- Properties of EBIC are studied for
 - linear models by Chen and Chen (2008), Luo and Chen (2013a);
 - generalized linear models by Chen and Chen (2012), Luo and Chen (2013b);
 - q -variate regression models by Luo and Chen (2014a);
 - Gaussian graphical models by Foygel and Drton (2010);
 - survival models by Luo, Xu and Chen (2014);
 - interactive models by He and Chen (2014).

THE SELECTION CONSISTENCY OF EBIC

- The selection consistency of EBIC refers to the following property:

$$P\left\{\min_{s: |s| \leq c|s_0|} \text{EBIC}_\gamma(s) > \text{EBIC}_\gamma(s_0)\right\} \rightarrow 1,$$

for any fixed positive number $c > 1$, where s_0 denotes the true model.

- Under certain reasonable conditions, the selection consistency holds for
 - Gaussian graphical models, if $\gamma > 1 - \frac{\ln n}{4 \ln p}$;
 - interactive models, if $\gamma_M > 1 - \frac{\ln n}{2 \ln p}$ and $\gamma_I > 1 - \frac{\ln n}{4 \ln p}$;
 - other models, if $\gamma > 1 - \frac{\ln n}{2 \ln p}$.
- When $p > \sqrt{n}$, the consistency range $\gamma > 0$, which reveals that, in this situation, the original BIC (with *gamma* = 0) is not selection consistent.

EXISTING PROCEDURES

- The existing feature selection procedures can be roughly classified into sequential and non-sequential ones.
- Some sequential procedures: forward regression, OMP, LARS, etc.
- Some non-sequential procedures: penalized likelihood approaches (Lasso, SCAD, Bridge, Adaptive lasso, Elastic net, MCP), Dentzig selector, etc..
- Sequential procedures are computationally more appealing.
- We concentrate on a sequential procedure in this talk. The remainder of the talk is based on Luo and Chen (2014b) and He and Chen (2014).

SEQUENTIAL PENALIZED LIKELIHOOD (SPL)

APPROACH: THE IDEA

- Notation:
 - $\mathbf{y}$: response vector;
 - X : design matrix;
 - S : column index set of X ;
 - $X(s), s \in S$: submatrix of X with column indices in s .
- The idea of sequential penalized approach is to select features sequentially by minimizing partially penalized likelihoods

$$-2 \ln L(\mathbf{y}, X\beta) + \lambda^* \sum_{j \notin s^*} |\beta_j|, \quad (1)$$

where s^* is the index set of already selected features, and λ^* is set at the largest value that allows at least one of the β_j 's, $j \notin s^*$, to be estimated nonzero.

SPL APPROACH: THE ALGORITHM

- For the purpose of feature selection, we don't really need to carry out the minimization of (1). The only thing we need is the active set in the minimization. The active set is related to the partial profile score defined below.

- Let $\ell(X\beta) = \ln L(\mathbf{y}, X\beta)$ and

$$\tilde{\ell}(\beta(s^c)) = \max_{\beta(s)} \ell(X(s)\beta(s) + X(s^c)\beta(s^c)).$$

For $j \in s^c$, the partial profile score is defined as

$$\psi(\mathbf{x}_j|s) = \left. \frac{\partial \tilde{\ell}(\beta_{s^c})}{\partial \beta_j} \right|_{\beta_{s^c}=\mathbf{0}}.$$

- The active features $\mathbf{x}_j$'s in the minimization of (1) satisfy

$$|\psi(\mathbf{x}_j|s^*)| = \max_{l \notin s^*} |\psi(\mathbf{x}_l|s^*)|.$$

- In the case of linear models,

$$\psi(\mathbf{x}_j|s^*) = |\mathbf{x}_j^\tau \tilde{\mathbf{y}}|,$$

where $\tilde{\mathbf{y}} = [I - H(s^*)]\mathbf{y}$.

SPL APPROACH: THE ALGORITHM (CONT.)

The computation algorithm for the SPL approach is as follows:

- Initial step: set $s_{*0} = \emptyset$, then
 - Compute $\psi(\mathbf{x}_j | s_{*0})$ for $j \in S$;
 - Identify $s_{\text{TEMP}} = \{j : |\psi(\mathbf{x}_j | s_{*0})| = \max_{I \in S} |\psi(\mathbf{x}_I | s_{*0})|\}$.
 - Let $s_{*1} = s_{\text{TEMP}}$ and compute $\text{EBIC}(s_{*1})$.
- General step $k (\geq 1)$:
 - Compute $\psi(\mathbf{x}_j | s_{*k})$ for $j \in s_{*k}^c$;
 - Identify $s_{\text{TEMP}} = \{j : |\psi(\mathbf{x}_j | s_{*k})| = \max_{I \in s_{*k}^c} |\psi(\mathbf{x}_I | s_{*k})|\}$.
 - Let $s_{*k+1} = s_{*k} \cup s_{\text{TEMP}}$, and compute $\text{EBIC}(s_{*k+1})$.
 - If $\text{EBIC}(s_{*k+1}) > \text{EBIC}(s_{*k})$, stop; otherwise, continue.

SPL APPROACH: INTERACTIVE MODELS

Let $S = \{1, \dots, p\}$, $\Psi = \{(j, k) : j < k \leq p\}$. Let $\mathbf{x}_j$'s denote main-effect features and $\mathbf{z}_{jk}$'s denote interaction features. The algorithm at each step is modified as

- Compute $\psi(\mathbf{x}_j | s_{*k})$ for $j \in S \setminus s_{*k}$ and identify

$$s_{\text{TEMP}}^M = \{j : |\psi(\mathbf{x}_j | s_{*k})| = \max_{l \in S \setminus s_{*k}} |\psi(\mathbf{x}_l | s_{*k})|\}.$$

- Compute $\psi(\mathbf{z}_{jk} | s_{*0})$ for $(j, k) \in \Psi \setminus s_{*k}$ and identify

$$s_{\text{TEMP}}^I = \{(j, k) : |\psi(\mathbf{z}_{jk} | s_{*0})| = \max_{(j, k) \in \Psi \setminus s_{*k}} |\psi(\mathbf{z}_{jk} | s_{*0})|\}.$$

- If $\text{EBIC}(s_{*k} \cup s_{\text{TEMP}}^M) < \text{EBIC}(s_{*k} \cup s_{\text{TEMP}}^I)$, let $s_{*k+1} = s_{*k} \cup s_{\text{TEMP}}^M$, otherwise let $s_{*k+1} = s_{*k} \cup s_{\text{TEMP}}^I$.
- If $\text{EBIC}(s_{*k+1}) > \text{EBIC}(s_{*k})$, stop; otherwise, continue.

SELECTION CONSISTENCY

Let $s_{*1}, s_{*2}, \dots, s_{*k}, \dots$ be the sequence of models selected by the SPL procedure without stopping. We have the following general theorem:

THEOREM

Under certain mild conditions, there exists a $k = k^*$ such that

$$Pr(s_{*k^*} = s_0) \rightarrow 1, \text{ as } n \rightarrow \infty,$$

where s_0 is the exact set of relevant features.

SELECTION CONSISTENCY (CONT.)

The following theorem gives the selection consistency of the SPL procedure for main-effect models with EBIC as the stopping rule (the same result holds for interactive models):

THEOREM

Let $s_{*1} \subset s_{*2} \subset \dots \subset s_{*k} \subset \dots$ be the sets generated by the sequential penalized procedure. Then, under the conditions of the previous theorems,

(I) *Uniformly, for k such that $|s_{*k}| < p_0$,*

$$P(EBIC_{\gamma}(s_{*k+1}) < EBIC_{\gamma}(s_{*k})) \rightarrow 1, \text{ when } \gamma > 0.$$

(II) *$P(\min_{p_0 < |s_{*k}| \leq cp_0} EBIC_{\gamma}(s_{*k}) > EBIC_{\gamma}(s_0)) \rightarrow 1$, when $\gamma > 1 - \frac{\ln n}{2 \ln p}$, where $c > 1$ is an arbitrarily fixed constant.*

ASYMPTOTIC NORMALITY OF PARAMETER ESTIMATORS

The following theorem is for the case of linear models. Similar result holds for generalized linear models.

Let $\mathbf{a} = (a_1, a_2, \dots)$ be an infinite sequence of constants. For any index set s , let $\mathbf{a}(s)$ denote the vector with components $a_j, j \in s$.

THEOREM

Let $\mathbf{z}_i^\tau$ be the i th row vector of $\mathbf{X}(s_0)$, $i = 1, \dots, n$. Assume that

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} \mathbf{z}_i^\tau [\mathbf{X}(s_0)^\tau \mathbf{X}(s_0)]^{-1} \mathbf{z}_i \rightarrow 0. \quad (2)$$

Then, for any fixed sequence $\mathbf{a}$,

$$\frac{\mathbf{a}(s^*)^\tau [\hat{\beta}(s^*) - \beta(s^*)]}{\sqrt{\mathbf{a}(s^*)^\tau [\mathbf{X}(s^*)^\tau \mathbf{X}(s^*)]^{-1} \mathbf{a}(s^*)}} \rightarrow_d N(0, \sigma^2),$$

where $\sigma^2 = \text{Var}(Y_i)$.

COVARIATE COVARIANCE STRUCTURES

For structure A1 — A5, $(n, p_{0n}, p_n) = (n, [4n^{0.16}], [5 \exp(n^{0.3})])$.

- **A1:** All the p_n features are statistically independent.
- **A2:** The Σ satisfies $\Sigma_{ij} = \rho^{|i-j|}$ for all $i, j = 1, 2, \dots, p_n$.
- **A3:** Let Z_j, W_j i.i.d. $N(0, I)$.

$$X_j = \frac{Z_j + W_j}{\sqrt{2}}, \text{ for } j \in s_{0n}; \quad X_j = \frac{Z_j + \sum_{k \in s_{0n}} Z_k}{\sqrt{1 + p_{0n}}} \text{ for } j \notin s_{0n}.$$

- **A4:** For $j \in s_{0n}$, X_j 's have constant pairwise correlations. For $j \notin s_{0n}$,

$$X_j = \epsilon_j + \frac{\sum_{k \in s_{0n}} X_k}{p_{0n}},$$

where $\epsilon_j \sim N(0, 0.08 * \mathbf{I}_n)$.

COVARIATE COVARIANCE STRUCTURES (CONT.)

- **A5:** Same as A4 except that, for $j \in s_{0n}$, X_j 's have correlation matrix $\Sigma = (\rho^{|i-j|})$ and $s_{0n} = \{1, 2, \dots, p_{0n}\}$.
- **B:** $(n, p_n, p_{0n}) = (100, 1000, 10)$ and $\sigma = 1$.
The relevant features are generated as i.i.d. standard normal variables. The coefficients of the relevant features are $(3, 3.75, 4.5, 5.25, 6, 6.75, 7.5, 8.25, 9, 9.75)$. The irrelevant features are generated as

$$X_j = 0.25Z_j + \sqrt{0.75} \sum_{k \in s_0} X_k, j \notin s_0,$$

where Z_j 's are i.i.d. standard normal and independent from the relevant features.

SIMULATION RESULTS: A1 AND A2

Setting	Methods	MSize	PDR	FDR	PMSE
A1	ALasso	49.0	1.00	.791	10.937
	SCAD	13.6	1.00	.283	8.638
	SIS+SCAD	8.7	.793	.181	12.355
	FSR	9.4	1.00	.035	8.688
	SPL	9.4	1.00	.035	8.683
A2	ALasso	40.6	.941	.735	15.297
	SCAD	23.7	.931	.612	14.159
	SIS+SCAD	8.1	.661	.255	14.715
	FSR	7.9	.846	.035	13.541
	SPL	7.8	.796	.073	14.462

SIMULATION RESULTS: A3 AND A4

Setting	Methods	MSize	PDR	FDR	PMSE
A3	ALasso	25.5	.956	.507	4.205
	SCAD	9.1	.972	.031	3.963
	SIS+SCAD	8.9	.864	.128	4.498
	FSR	9.2	.708	.311	4.688
	SPL	9.2	.873	.148	4.272
A4	ALasso	13.3	1.00	.215	2.186
	SCAD	9.0	1.00	.000	2.320
	SIS+SCAD	8.7	.449	.535	3.327
	FSR	9.3	.993	.037	2.207
	SPL	9.2	1.00	.023	2.183

SIMULATION RESULTS: A5 AND B

Setting	Methods	MSize	PDR	FDR	PMSE
A5	ALasso	15.7	.986	.276	5.303
	SCAD	9.0	.999	.000	5.199
	SIS+SCAD	7.8	.681	.206	7.975
	FSR	9.4	.943	.091	5.545
	SPL	9.3	1.00	.024	5.241
B	ALasso	69.6	.852	.877	7.893
	SCAD	8.8	.583	.308	28.737
	SIS+SCAD	4.3	.000	1.00	58.334
	FSR	18.2	.785	.561	8.684
	SPL	9.8	.754	.262	19.470

SUMMARY OF FINDINGS:

- Under A1 and A2, SLasso and FSR are better than the others. SPL and FSR are comparable while FSR is slightly better.
- Under A3 - A5, SCAD is better than all the others. SPL is close to SCAD and better than the others.
- Under B, SPL is better than all the others.
- SPL is robust: it always has a very low FDR and is always the best or close to the best. On the contrast, SCAD and FSR are erratic over the settings. They are the best in certain settings but perform much worse in other settings.

REFERENCES

- J. Chen and Z. Chen. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Space. *Biometrika*, **95**: 759-771.
- J. Chen and Z. Chen (2012). Extended BIC for small- n -large- P sparse GLM. *Statistica Sinica* **22** 555-574.
- Y. He and Z. Chen (2014). The EBIC and a sequential procedure for feature selection from big data with interactive linear models. AISM, accepted.
- S. Luo and Z. Chen (2013a). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. JSPI. 143, 494-504.
- S. Luo and Z. Chen (2013b). Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters. *Statistics and Its Interface*, **6**, 275-284.

- S. Luo and Z. Chen (2014a) Edge detection in sparse Gaussian graphical models. Computational Statistics and Data Analysis, **70** 138-152.
- S. Luo and Z. Chen (2014b) Sequential Lasso cum EBIC for feature selection with ultra-high dimensional feature space. JASA, DOI: 10.1080/01621459.2013.877275. to appear.
- S. Luo, J. Xu and Z. Chen (2014) Extended Bayesian information criterion in the Cox model with high-dimensional feature space. AISM, to appear.