Selection Consistency of GIC for Small-n-Large-p Sparse Logistic Regression Model

Hubert Szymanowski[1] and Jan Mielniczuk[1,2]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

[2] Warsaw University of Technology, Faculty of Mathematics and Information Science,
ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract. We consider selection procedure for high-dimensional logistic regression problem which consists in choosing a subset of predictors which minimizes Generalized Information Criterion (GIC) over all subsets of variables of size not exceeding preset value k_n which may depend on a sample size. Nonasymptotic bound on probability of erroneous selection is proved which yields a range for GIC penalty parameter for which the procedure is consistent under mild assumptions and thus generalizes results of [1] and [2]. Various modifications of the procedure are analyzed using numerical examples.

1 Introduction

Let n be the number of observations and P_n be the number of variables which may depend on n. We consider a regression problem with matrix of experiment X of dimension $n \times (P_n + 1)$ and a binary response vector Y. Rows $x'_{i,\cdot}$ of X are thus transposed observations and its columns $x_{\cdot,j}$ contain predictors' values. The first column of X consisting of ones corresponds to the intercept. We assume that observations pertain to the standard logistic regression model with probability of success given observation $x_{i,\cdot}$ described by formula

$$\mathcal{P}(Y_i = 1 | x_{i,\cdot}) = \frac{1}{1 + \exp(-\beta_0 x_{i,\cdot})}$$

where $\beta_0 = (\beta_{0,0}, \beta_{0,1}, ..., \beta_{0,P_n})$ is (P_n+1)-dimensional vector of true coefficients. We denote by s_0 the minimal true model $\{j : \beta_{0,j} \neq 0\}$. The conditions we impose later imply that s_0 is identifiable. We assume that the minimal true model always contains intercept i.e $0 \in s_0$.

We consider the problem of constructing selectors of s_0 incorporating Generalized Information Criterion (cf [3]) with objective function

$$GIC(s) = -2l_n(\hat{\beta}(s), Y | X(s)) + a_n |s|,$$

where s is a given submodel containing $|s|$ explanatory variables and an intercept, $\hat{\beta}(s)$ is a maximum likelihood estimator calculated for model s (augmented by zeros to (P_n+1)-dimensional vector if necessary), $X(s)$ is a matrix of experiment restricted to columns from s and a_n is a chosen penalty. Specific values of penalty term lead to popular selection criteria such as Bayesian Information Criterion (BIC) with $a_n = \log(n)$ or Akaike Information Criterion (AIC) with $a_n = 2$. It was established (cf [4]) that both of this criteria tend to choose too many predictors in the case when number of potential predictors is large. Therefore, in the last few years number of Information Criteria with penalty larger than $\log(n)$ have been proposed. In [5] generalization of BIC called Extended BIC (EBIC) with penalty $a_n = \log n + 2\gamma \log P_n$ for some $\gamma \geq 0$ is considered. It stems from putting a certain non-uniform prior on family of models. Note that EBIC penalty depends on the number of potential predictors and is of order $\log P_n$ when P_n is of a higher order than n.

Selection procedure based on GIC involves looking for a subset $\hat{s}_0$ which minimizes GIC objective function over predefined family of models $\mathcal{M}$. We consider $\mathcal{M} = \{s : |s| \leq k_n\}$ with threshold k_n which may depend on n. This selection method in the case of $k_n = k$ was introduced in [5] for the linear models and extended in [1] to the case of the generalized linear models (GLMs). In [2] properties of this selection method restricted to the standard logistic regression model were studied under two assumptions: Sparse Riesz Condition (SRC) for both Hessian matrix of loglikelihood function and experimental matrix and assumption of uniform continuity of the Hessian. Here we generalize and improve these results. We prove a nonasymptotic bound on the probability of erroneous selection from which selection consistency follows under certain relations between the minimal eigenvalue of a moment matrix, norms of observations and GIC penalty.

The paper is organized as follows. In Section 2 we introduce preliminaries and in Section 3 we state and prove the main results. In Section 4 numerical experiments are discussed.

2 Preliminaries

We partition all models including intercept of size not exceeding k_n into two disjoint families

$$\mathcal{A}_0 = \{s : |s| \leq k_n \wedge s \supseteq s_0\}$$

and its complement

$$\mathcal{A}_1 = \{s : |s| \leq k_n \wedge s \not\supseteq s_0 \wedge 0 \in s\}.$$

Let $p(t) = 1/(1 + \exp(-t))$ and $\sigma^2(t) = p(t)(1 - p(t))$. For the standard logistic regression with logit link function, conditional likelihood for a given model $s \in$

$\mathcal{A}_0 \cup \mathcal{A}_1$ and parameter $\beta \in \mathbb{R}^{|s|+1}$ is

$$\begin{aligned} l(\beta, Y|X(s)) &= \sum_{i=1}^{n} \{y_i \log[p(x'_{i,\cdot}(s)\beta)] + (1-y_i)\log[1-p(x'_{i,\cdot}(s)\beta)]\} \\ &= \sum_{i=1}^{n} \{y_i x'_{i,\cdot}(s)\beta - \log[1+\exp(x'_{i,\cdot}(s)\beta)]\}, \end{aligned}$$

where $X(s)$ stands for the design matrix X restricted to the columns from s and $x'_{i,\cdot}(s)$ is the i-th row of this matrix.
We denote by $\beta(s)$ $|s|$-dimensional vector augmented by zeros to higher-dimensional vector when necessary. The maximum likelihood estimator (ML) $\hat{\beta}(s)$ of parameter $\beta_0(s)$ is defined as

$$\hat{\beta}(s) = \arg \max_{\beta \in \mathbb{R}^{|s|+1}} l(\beta, Y|X(s)).$$

Define also the score function

$$S_n(\beta) = \frac{\partial l(\beta, Y|X)}{\partial \beta} = \sum_{i=1}^{n} [y_i - p(x'_{i,\cdot}\beta)]x_{i,\cdot} = X'(Y - \mathbf{p}(\beta)), \tag{1}$$

where $\mathbf{p}(\beta) = (p(x'_1\beta), \ldots, p(x'_n\beta))'$. The negative Hessian matrix will be denoted by

$$H_n(\beta) = -\frac{\partial^2 l(\beta, Y|X)}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \sigma^2(x'_{i,\cdot}\beta)x_{i,\cdot}x'_{i,\cdot} = X'\Pi(\beta)X, \tag{2}$$

where $\Pi(\beta) = \text{diag}\{\sigma^2(x'_{1,\cdot}\beta), \ldots, \sigma^2(x'_{n,\cdot}\beta)\}$.
Define

$$\tilde{\lambda}_{min} = \min_{s \in \mathcal{A}_1} \lambda_{min}(X'(s \cup s_0)X(s \cup s_0)),$$

$$N = \max_{i=1,2,\ldots,n} ||x_{i,\cdot}(s_0)||,$$

$$\tilde{N} = \max_{s \in \mathcal{A}_1} \max_{i=1,2,\ldots,n} ||x_{i,\cdot}(s \cup s_0)||.$$

Results of the paper are proved under certain assumptions involving relations between the three quantities above and penalty a_n.

3 Main results

We assume throughout that $P_n > 2$ for every n.

Lemma 1 *Let $Y = (y_1, \ldots, y_n)'$ be a vector consisting of independent binary variables having not necessarily the same distribution and*

$$A(s) = X(s \cup s_0)[X'(s \cup s_0)X(s \cup s_0)]^{-1}X'(s \cup s_0)$$

for $s \in \mathcal{A}_0 \cup \mathcal{A}_1$. Then for every $\varepsilon > 0$ and $c(\varepsilon) = 0.5(4\varepsilon - \sqrt{9+8\varepsilon} + 3) > 0$ the following inequalities hold for all n

$$\mathcal{P}(\max_{s\in\mathcal{A}_1} ||A(s)(Y-EY)||^2 > (\frac{5}{4}+\varepsilon)(k_n+|s_0|)\log P_n) \leq P_n^{-c(\varepsilon)(k_n+|s_0|)}. \quad (3)$$

$$\mathcal{P}(\max_{s\in\mathcal{A}_0} ||A(s)(Y-EY)||^2 > (\frac{5}{4}+\varepsilon)(k_n)\log P_n) \leq P_n^{-c(\varepsilon)k_n}. \quad (4)$$

$$\mathcal{P}(\max_{s\in\mathcal{A}_0, s\neq s_0} [||A(s)(Y-EY)||^2 - (\frac{5}{4}+\varepsilon)(|s|-|s_0|)\log P_n] > 0) \leq \exp(P_n^{-c(\varepsilon)}) - 1. \quad (5)$$

Proof

First we prove inequality (3). Since $A(s)$ is an idempotent matrix for any s, we have $\mathrm{tr}A^2(s) = \mathrm{tr}A(s) = |s \cup s_0|$ and $\lambda_{max}(A(s)) = 1$. It follows from Theorem 2.1 in [6] that

$$P(||A(s)(Y-EY)||^2 > \frac{1}{4}(\mathrm{tr}(A(s)) + 2\sqrt{\mathrm{tr}(A^2(s))t} + 2\lambda_{max}(A(s))t)) < e^{-t}.$$

Let $t = (1+c(\varepsilon))(k_n+|s_0|)\log P_n$. Note that $\sqrt{1+c(\varepsilon)} = \sqrt{2\varepsilon+9/4} - 1/2$. We have

$$\begin{aligned}
&\mathcal{P}(\max_{s\in\mathcal{A}_1} ||A(s)(Y-EY)||^2 > (\frac{5}{4}+\varepsilon)(k_n+|s_0|)\log P_n) \\
&= \mathcal{P}(\max_{s\in\mathcal{A}_1} ||A(s)(Y-EY)||^2 > \frac{1}{4}(1+2\sqrt{1+c(\varepsilon)}+2(1+c(\varepsilon)))(k_n+|s_0|)\log P_n) \\
&\leq \sum_{j=1}^{k_n} \binom{P_n}{j} \max_{s:|s|=j} P(||A(s)(Y-EY)||^2 \\
&\quad > \frac{1}{4}(k_n+|s_0| + 2(k_n+|s_0|)\sqrt{(1+c(\varepsilon))\log P_n} + 2(1+c(\varepsilon))(k_n+|s_0|)\log P_n)) \\
&\leq \exp(-(1+c(\varepsilon))(k_n+|s_0|)\log P_n) \sum_{j=1}^{k_n} \binom{P_n}{j} \\
&\leq \exp(-(1+c(\varepsilon))(k_n+|s_0|)\log P_n) \sum_{j=1}^{k_n+|s_0|} \frac{P_n^j}{j!} \\
&\leq \exp(-(1+c(\varepsilon))(k_n+|s_0|)\log P_n) \frac{P_n^{k_n+|s_0|}}{(k_n+|s_0|-1)!} \leq P_n^{-c(\varepsilon)(k_n+|s_0|)}.
\end{aligned}$$

For the last two inequalities we use an $\binom{n}{k} \leq n^k/k!$ and the fact that sequence $n^k/k!$ is non decreasing for fixed n and $k = 1, 2, ..., n$.

Proof of inequality (4) is similar. In order to prove (5) change t in the reasoning above to $t = |s| \log P_n$ and note that

$$\sum_{j=|s_0|}^{k_n} \binom{P_n}{j} \big(\exp(-(1+c(\varepsilon)) \log P_n)\big)^j \leq (1 + \frac{1}{P_n^{1+c(\varepsilon)}})^{P_n} - 1 \leq \exp(P_n^{-c(\varepsilon)}) - 1.$$

□

Remark 1 *If the number of variables P_n is constant, Lemma 1 does not give a suitable bounds on considered probability. In such a case we use a slightly modified version of the Lemma. For $P_n = P$ we set $k_n = P$. Let $f = \{0, 1..., P\}$ be a full model and 2^f be a set of all possible models. In the considered setting for $n > P$ and any constant $M \geq P$ we have*

$$\mathcal{P}(\max_{s \in 2^f, 0 \in s} ||A(s)(Y - EY)||^2 > \frac{5}{4}M) = P(||A(f)(Y - EY)||^2 > \frac{5}{4}M)$$
$$\leq \mathcal{P}(||A(f)(Y - EY)||^2 > \frac{1}{4}(P + 2\sqrt{PM} + 2M)) < e^{-M}$$

and

$$\mathcal{P}(\max_{s \in \mathcal{A}_0, s \neq s_0} [||A(s)(Y - EY)||^2 - \frac{5}{4}(|s| - |s_0|)M] \geq 0)$$
$$\leq \mathcal{P}(||A(F)(Y - EY)||^2] \geq \frac{5}{4}M) \leq e^{-M}.$$

In order to ensure that the minimal true model is selected with a large probability we need to find conditions under which the behaviour of $l_n(\hat{\beta}(s)) - l_n(\hat{\beta}_0(s))$ can be controlled uniformly over $s \in \mathcal{A}_1$ and $s \in \mathcal{A}_0 \setminus \{s_0\}$. This will be done using the following notion. For a given $s \in \mathcal{A}_0 \cup \mathcal{A}_1$ define

$$B(s, r) = \{\beta : ||X(s \cup s_0)(\beta(s) - \beta_0(s))||^2 \leq r^2\}. \quad (6)$$

Lemma 2 and Theorem 2 state conditions under which $\hat{\beta}(s) \in B(s, r)$ for $s \in \mathcal{A}_0$ whereas for $s \in \mathcal{A}_1$ we have $\hat{\beta}(s) \notin B(s, r)$. Define

$$\mathcal{B} = \{\forall s \in \mathcal{A}_0 \;\; \hat{\beta}(s) \in B(s, \frac{\sqrt{\tilde{\lambda}_{min}}}{\tilde{N}})\}. \quad (7)$$

Lemma 2 *For all n such that inequality*

$$\frac{\sqrt{\tilde{\lambda}_{min}}}{\tilde{N}} \exp(-N||\beta_0||) \geq e\sqrt{(80 + 64\varepsilon)k_n \log P_n} \quad (8)$$

holds, we have

$$P(\mathcal{B}) \geq 1 - \exp(-c(\varepsilon)k_n \log P_n).$$

Proof

It is easily seen that $\beta \in B(s,r)$ can be represented as $\beta_0(s) + \gamma(X'(s)X(s))^{-\frac{1}{2}}u$ where $\gamma \in [0,r]$ and u is a vector with $||u|| = 1$. For any index i, $s \in \mathcal{A}_0$ and $\beta \in B(s,r)$ we have

$$\begin{aligned}
\sigma^2(x'_{i,\cdot}\beta) &= \sigma^2(x'_{i,\cdot}\beta_0 + \gamma x'_{i,\cdot}(X(s)'X(s))^{-\frac{1}{2}}u) \\
&\geq \sigma^2(||x'_{i,\cdot}(s_0)|| \cdot ||\beta_0|| + r\sqrt{x'_{i,\cdot}(X(s)'X(s))^{-1}x_{i,\cdot}}) \\
&\geq \sigma^2(||x'_{i,\cdot}(s_0)|| \cdot ||\beta_0|| + r||x'_{i,\cdot}(s)||/\sqrt{\lambda_{min}(X(s)'X(s))}) \\
&\geq \sigma^2(N||\beta_0|| + r\tilde{N}/\sqrt{\tilde{\lambda}_{min}}).
\end{aligned}$$

Let $\beta_u = \beta_0 + r(X(s)'X(s))^{-\frac{1}{2}}u$ for some u such that $||u|| = 1$. Note that β_u is a boundary point of $B(s,r)$. Using concavity of $l_n(\cdot)$ we have

$$P(\exists s \in \mathcal{A}_0 \;\; \hat{\beta}(s) \notin B(s, \frac{\sqrt{\tilde{\lambda}_{min}}}{\tilde{N}})) \leq P(\exists u : ||u|| = 1, \max_{s \in \mathcal{A}_0} l_n(\beta_u) \geq l_n(\beta_0))$$

and the bound above is in its turn not larger than

$$\begin{aligned}
&P(\exists u : ||u|| = 1, \max_{s \in \mathcal{A}_0} \big[u'(X(s)'X(s))^{-\frac{1}{2}}X(s)'(Y - EY) \\
&-\frac{1}{2}ru'(X(s)'X(s))^{-\frac{1}{2}}H(\beta^*)(X(s)'X(s))^{-\frac{1}{2}}u\big] \geq 0) \\
&\leq P(\max_{s \in \mathcal{A}_0} ||A(s)(Y - EY)|| \geq \frac{1}{2}r\sigma^2(N||\beta_0|| + r\tilde{N}/\sqrt{\tilde{\lambda}_{min}})) \\
&\leq P(\max_{s \in \mathcal{A}_0} ||A(s)(Y - EY)|| \geq \frac{1}{8}r\exp(-N||\beta_0|| - r\tilde{N}/\sqrt{\tilde{\lambda}_{min}}))
\end{aligned}$$

for some β^* belonging to the line segment between β_u and β_0. We used the fact that the scalar product $u'v$ for a given vector v and $||u|| = 1$ is maximized by $u = v/||v||$. The last inequality follows from the fact that $\sigma^2(t) = e^{-|t|}/(1 + e^{-|t|})^2 \geq 0.25e^{-|t|}$. For $r = \sqrt{\tilde{\lambda}_{min}}/\tilde{N}$ by Lemma 1 the right hand side is bounded from above by $\exp(-c(\varepsilon)k_n \log P_n)$ if inequality (8) is satisfied. □

Theorem 1 *For all n such that inequality (8) and*

$$a_n \geq (5 + 4\varepsilon)e \log P_n e^{N||\beta_0||} \tag{9}$$

hold simultaneously, we have

$$P(\min_{s \in A_0, s \neq s_0} GIC(s) \leq GIC(s_0)) \leq P_n^{-c(\varepsilon)k_n} + \exp(P_n^{-c(\varepsilon)}) - 1$$

Proof
Let $s \in \mathcal{A}_0$. For some β^* being a vector belonging to the line segment with endpoints $\hat{\beta}(s)$ and $\beta_0(s)$ we have

$$\begin{aligned} & l_n(\hat{\beta}(s)) - l_n(\hat{\beta}_0(s)) \leq l_n(\hat{\beta}(s)) - l_n(\beta_0(s)) \\ = & [\hat{\beta}(s) - \beta_0(s)]' S_n(\beta_0(s)) - \frac{1}{2}[\hat{\beta}(s) - \beta_0(s)]' H_n(\beta^*)[\hat{\beta}(s) - \beta_0(s)]. \end{aligned}$$

From convexity $\beta^* \in B(s,r)$ and on event $\mathcal{B}$ defined by (7) we have in view of the proof of Lemma 2 that $\sigma^2(x'_{i,\cdot}\beta^*) \geq \sigma^2(N||\beta_0|| + 1)$ for any i. On the event $\mathcal{B}$ we also have

$$\hat{\beta}(s) - \beta_0 = \gamma_s (X'(s)X(s))^{-\frac{1}{2}} u'_s$$

for some $\gamma_s \in [0,r]$ and vector u_s with $||u_s|| = 1$. This implies that on $\mathcal{B}$

$$\begin{aligned} l_n(\hat{\beta}(s)) - l_n(\hat{\beta}_0(s)) & \leq \gamma_s ||(X'(s)X(s))^{-\frac{1}{2}} X(s)'(Y - EY)|| - \frac{1}{2}\gamma_s^2 \sigma^2(N||\beta_0|| + 1) \\ & \leq \frac{||(X'(s)X(s))^{-\frac{1}{2}} X(s)'(Y - EY)||^2}{2\sigma^2(N||\beta_0|| + 1)} = \frac{||A(s)(Y - EY)||^2}{2\sigma^2(N||\beta_0|| + 1)} \end{aligned}$$

Therefore,

$$\begin{aligned} & P(\min_{s \in \mathcal{A}_0, s \neq s_0} GIC(s) \leq GIC(s_0)) \\ = & \mathcal{P}(\max_{s \in \mathcal{A}_0, s \neq s_0} (l_n(\hat{\beta}(s)) - l_n(\hat{\beta}(s_0)) - \frac{(|s| - |s_0|)a_n}{2}) \geq 0) \\ \leq & \mathcal{P}(\max_{s \in \mathcal{A}_0, s \neq s_0} [||A(s)(Y - EY)||^2 - (|s| - |s_0|)a_n \sigma^2(N||\beta_0|| + 1))] \geq 0) \\ \leq & \mathcal{P}(\max_{s \in \mathcal{A}_0, s \neq s_0} [||A(s)(Y - EY)||^2 - \frac{(|s| - |s_0|)a_n \exp(-N||\beta_0||)}{4e}] \geq 0). \end{aligned}$$

By Lemma 1 the last expression is bounded from above by $\exp(P_n^{-c(\varepsilon)}) - 1$ if condition (9) is satisfied. Since it follows from Lemma 2 that $P(\mathcal{B}) \leq \exp(-c(\varepsilon)k_n \log P_n)$ the last step is to use inequality $P(\mathcal{C}) \leq P(\mathcal{C} \cap B) + P(B')$ for $\mathcal{C} = \{\min_{s \in \mathcal{A}_0, s \neq s_0} GIC(s) \leq GIC(s_0)\}$.

□

Theorem 2 *Let $\beta_{min} = \min_{i \in s_0} |\beta_{0,i}|$. Fix $\eta \in (0,1)$. For all n such the following inequalities hold*

$$\tilde{N}\beta_{min} > 1 \tag{10}$$

$$\frac{\eta}{4e} \frac{\tilde{\lambda}_{min}}{\tilde{N}^2} e^{-N||\beta_0||} \geq a_n \tag{11}$$

and

$$(1 - \eta) \frac{\sqrt{\tilde{\lambda}_{min}}}{\tilde{N}} e^{-N||\beta_0||} \geq e\sqrt{(80 + 64\varepsilon)(k_n + |s_0|) \log P_n} \tag{12}$$

we have

$$P(\min_{s\in\mathcal{A}_1} GIC(s) \le GIC(s_0)) \le P_n^{-c(\varepsilon)(k_n+|s_0|)}$$

Proof

Consider set $B(s,r)$ defined in (6) for $r = \sqrt{\tilde{\lambda}_{min}}/\tilde{N}$ and $s \in \mathcal{A}_1$. Note that,

$$||X(s\cup s_0)(\beta(s)-\beta_0(s))|| \ge \sqrt{\lambda_{min}(X'(s\cup s_0)X(s\cup s_0))}||\beta(s)-\beta_0(s)|| \ge \sqrt{\tilde{\lambda}_{min}}\beta_{min}.$$

Thus if $r < \beta_{min}\sqrt{\tilde{\lambda}_{min}}$ then $\hat{\beta}(s) \notin \mathcal{A}_1$ and the last inequality is satisfied in view of (10). Using concavity of $l_n(\cdot)$ again we have

$$\begin{aligned}
&P(\min_{s\in\mathcal{A}_1} GIC(s) \le GIC(s_0)) \\
&\le P(\max_{s\in\mathcal{A}_1}\ l_n(\hat{\beta}(s)) - l_n(\beta_0) \ge -\frac{a_n}{2}) \\
&\le P(\exists u : ||u|| = 1 \max_{s\in\mathcal{A}_1}\ l_n(\beta_u) - l_n(\beta_0) \ge -\frac{a_n}{2}) \\
&\le P(\exists u : ||u|| = 1, \max_{s\in\mathcal{A}_1}(u'(X(s\cup s_0)'X(s\cup s_0))^{-\frac{1}{2}}X(s\cup s_0)'(Y-EY) \\
&\quad -\frac{1}{2}ru'(X(s\cup s_0)'X(s\cup s_0))^{-\frac{1}{2}}H(\beta^*)(X(s\cup s_0)'X(s\cup s_0))^{-\frac{1}{2}}u) \ge -\frac{a_n}{2r}) \\
&\le P(\max_{s\in\mathcal{A}_1}||A(s)(Y-EY)|| \ge \frac{1}{2}r\sigma^2(N||\beta_0|| + \frac{\tilde{N}}{\sqrt{\tilde{\lambda}_{min}}}r) - \frac{a_n}{2r}) \\
&\le P(\max_{s\in\mathcal{A}_1}||A(s)(Y-EY)|| \ge \frac{1}{8e}r\exp(-N||\beta_0||) - \frac{a_n}{2r}) \\
&\le P(\max_{s\in\mathcal{A}_1}||A(s)(Y-EY)|| \ge \frac{1-\eta}{8e}r\exp(-N||\beta_0||))
\end{aligned}$$

where (11) is used for the last inequality. By Lemma 1 the last probability is bounded from above by $\exp(-c(\varepsilon)(k_n + |s_0|\log P_n)$ if

$$\frac{1-\eta}{8e}r\exp(-N||\beta_0||) \ge \sqrt{(\frac{5}{4}+\varepsilon)(k_n+|s_0|)\log P_n}$$

which is equivalent to (12).

□

Corollary 1 *It follows from Theorems 1 and 2 that for all n such that inequalities (9) (10), (11),(12) hold for some $\eta \in (0,1)$ and $\varepsilon > 0$, we have*

$$P(\min_{s\in\mathcal{A}_0\cup\mathcal{A}_1, s\neq s_0} GIC(s) \le GIC(s_0)) \le 2P_n^{-c(\varepsilon)(k_n+|s_0|)} + \exp(P_n^{-c(\varepsilon)}) - 1.$$

Remark 2 *Write $w_n << z_n$ if $w_n = o(z_n)$ for $n \to \infty$. It follows from Theorems 1 and 2 that if*

$$1 << \tilde{N} \tag{13}$$

$$k_n \log P_n << \frac{\tilde{\lambda}_{min}}{\tilde{N}^2} e^{-2N||\beta_0||} \tag{14}$$

and

$$e^{N||\beta_0||} \log P_n << a_n << \frac{\tilde{\lambda}_{min}}{\tilde{N}^2} e^{-N||\beta_0||} \tag{15}$$

we have

$$P(\min_{s \in \mathcal{A}_0 \cup \mathcal{A}_1, s \neq s_0} GIC(s) \leq GIC(s_0)) \to 0$$

when n tends to infinity.

Remark 3 *If number of variables P is constant we use in Lemma 2 and Theorems 1 and 2 inequalities from Remark 1. This leads to the following conditions for consistency of GIC. If condition (13) is satisfied and*

$$e^{N||\beta_0||} << a_n << \frac{\tilde{\lambda}_{min}}{\tilde{N}^2} e^{-N||\beta_0||} \tag{16}$$

we have

$$P(\min_{s \in 2^f, s \neq s_0} GIC(s) \leq GIC(s_0)) \to 0$$

where f and 2^f are defined in Remark 1. Note that in considered case we have $\tilde{\lambda}_{min} = \lambda_{min}(X'(f)X(f))$ and $\tilde{N} = \max_{i=1,\ldots,n} ||x_{i,\cdot}||$.
If condition (13) does not hold, then in the proof of Theorem 2 we take $r = A\sqrt{\tilde{\lambda}_{min}}$ with $A < \beta_{min}$. This leads to the following conditions on consistency of GIC. If

$$1 << a_n << \tilde{\lambda}_{min} \tag{17}$$

we have

$$P(\min_{s \in 2^f, 0 \in s, s \neq s_0} GIC(s) \leq GIC(s_0)) \to 0.$$

4 Discussion of the assumptions

In this section we examine behavior of $\tilde{\lambda}_{min}$, $\tilde{N}$ and N when design matrix has some specific structure. We find the following lemma useful. It is a version of Preposition 1 in [7] for unnormalized predictors.

Lemma 3 *Let $\rho_{ij} = x'_{\cdot,i} x_{\cdot,j}$. The following inequality holds*

$$\tilde{\lambda}_{min} \geq \min_j \rho_{jj} - \max_{|s|=k_n} \inf_{\alpha > 1} \left[\sum_{i \in s \cup s_0} \left(\sum_{j \in s \cup s_0 \setminus \{i\}} |\rho_{ij}|^{\alpha/(\alpha-1)} \right)^{\alpha-1} \right]^{1/\alpha}. \tag{18}$$

Since all the matrices $X'(s \cup s_0)X(s \cup s_0)$ are positively defined, the lemma is nontrivial if the right hand side of (18) is positive.

Proof

Fix s such that $|s| \leq k_n$ and denote by $j_1, j_2, ..., j_{|s \cup s_0|}$ elements of $s \cup s_0$. Let $b(s \cup s_0) = b = (b_{j_1}, ..., b_{j_{|s \cup s_0|}})'$ be an eigenvector of $X'(s \cup s_0)X(s \cup s_0)$ corresponding to its minimal eigenvalue $\lambda_{min}(s \cup s_0) = \lambda_{min}$. From the definition of eigenvector for any $j \in s \cup s_0$ we have

$$\sum_{i \in s \cup s_0} \rho_{ji} b_i = \lambda_{min} b_j.$$

Therefore by Hölder's inequality

$$\min_{j=1,\ldots,p} |\lambda_{min} - \rho_{jj}|^{\alpha} \sum_{j \in s \cup s_0} |b_j|^{\alpha} \leq \sum_{j \in s \cup s_0} |(\lambda_{min} - \rho_{jj}) b_j|^{\alpha}$$

$$= \sum_{j \in s \cup s_0} \Bigg| \sum_{i \in s \cup s_0 \setminus \{j\}} \rho_{ji} b_i \Bigg|^{\alpha} \leq \sum_{j \in s \cup s_0} \Bigg(\sum_{i \in s \cup s_0 \setminus \{i\}} |\rho_{ij}|^{\alpha/(\alpha-1)} \Bigg)^{\alpha-1} \sum_{i \in s \cup s_0} |b_i|^{\alpha}.$$

Let $\delta = \max_{|s|=k_n} \inf_{\alpha>1} \left[\sum_{i \in s \cup s_0} \left(\sum_{j \in s \cup s_0 \setminus \{i\}} |\rho_{ij}|^{\alpha/(\alpha-1)} \right)^{\alpha-1} \right]^{1/\alpha}$. After dividing both sides by $\sum_{i \in s \cup s_0} |b_i|^{\alpha}$ we obtain $\min_j |\lambda_{min} - \rho_{jj}| \leq \delta$ which implies that $\lambda_{min} \geq \min_j \rho_{jj} - \delta$. Since the right hand side does not depend on choice of s the lemma is proved.

□

Let

$$\rho_n = \max_{i \neq j} |\rho_{ij}|, \quad \tau_n = \min_j ||x_{\cdot,j}||, \quad M_n = \max_{i=1,\ldots,n; j=1,\ldots,P_n} |x_{ij}|.$$

Then (18) and Schwarz inequality implies

$$\frac{\sqrt{\tilde{\lambda}_{min}}}{\tilde{N}} \exp(-N||\beta_0||) > \frac{\sqrt{\tau_n^2 - (k_n + |s_0|)\rho_n}}{\sqrt{k_n + |s_0|} M_n} \exp(-\sqrt{s_0} M_n ||\beta_0||).$$

The lower bound is positive if $\rho_n < \tau_n^2 / k_n$ and the inequality (8) holds if

$$\tau_n^2 - (k_n + |s_0|)\rho_n > e^2 (80 + 64\varepsilon)(k_n + |s_0|)^2 \log P_n M_n^2 \exp(2\sqrt{s_0} M_n ||\beta_0||).$$

The assumption frequently used in the literature is Sparse Riesz Condition (SRC) see e.g. [7]. We say that the design matrix X satisfies left-sided SRC with rank k_n and a spectrum bound $0 < C_1 < +\infty$ if

$$\forall s : |s| \leq k_n \quad \forall v \in \mathbb{R}^{|s|} \quad C_1 \leq \frac{||X(s)v||^2}{n||v||^2}$$

which is equivalent to

$$\min_{s:|s| \leq k_n} \lambda_{min}(X'(s)X(s)) \geq C_1 n$$

Corollary 2 *Assume that matrix X satisfies left-sided SRC with rank $k_n + |s_0|$ and constant C_1. For all n such that*

$$e^2(80 + 64\varepsilon)M_n^2(k_n + |s_0|)^2 \log P_n \exp(2\sqrt{s_0}M_n||\beta_0||) < C_1(1-\eta)^2 n \tag{19}$$

$$1 < C_1 k_n \beta_{min}^2 \tag{20}$$

$$(5 + 4\varepsilon)\exp(M_n\sqrt{s_0}||\beta_0||)\log P_n < a_n < \frac{\eta C_1 n}{4e(k_n + |s_0|)M_n^2}\exp(-M_n\sqrt{s_0}||\beta_0||) \tag{21}$$

hold for some $\eta \in (0,1)$, we have

$$P(\min_{s\in\mathcal{A}_0\cup\mathcal{A}_1, s\neq s_0} GIC(s) \le GIC(s_0)) \le 2P_n^{-c(\varepsilon)(k_n+|s_0|)} + P_n^{-c(\varepsilon)}.$$

Note that if $M_n \le M$, inequality (19) reduces to

$$n > A(k_n + |s_0|)^2 \log P_n \quad \text{with} \quad A = \frac{e^2(80 + 64\varepsilon)M^2 \exp(\sqrt{s_0}M||\beta_0||)}{C_1(1-\eta)^2}$$

and inequality (21) reduces to

$$B_1 \log P_n < a_n < B_2 \frac{n}{k_n + |s_0|}$$

$$\text{with} \quad B_1 = (5 + 4\varepsilon)\exp(M\sqrt{s_0}||\beta_0||) \quad \text{and} \quad B_2 = \frac{\eta C_1}{4eM^2}\exp(-M\sqrt{s_0}||\beta_0||).$$

Proof

We show that conditions (19)-(21) imply assumptions of Corollary 1. Left-sided SRC with rank $k_n + |s_0|$ implies that for fixed s with $|s| \le k_n$, we have

$$C_1 n \le \lambda_{min}(X'(s\cup s_0)X(s\cup s_0)) \le \frac{1}{|s\cup s_0|}\mathrm{tr}(X'(s\cup s_0)X(s\cup s_0))$$
$$= \frac{1}{|s\cup s_0|}\sum_{i=1}^{n}||x_{i,\cdot}(s\cup s_0)||^2 \le \frac{n}{|s\cup s_0|}\max_{i=1,\dots,n}||x_{1,\cdot}(s\cup s_0)||^2.$$

Thus, when left-sided SRC is satisfied and all absolute values of design entries x_{ij} are bounded from above by M_n the inequality (12) holds for some $\eta \in (0,1)$ if

$$n > AM_n^2(k_n + |s_0|)^2 \log P_n \exp(2\sqrt{s_0}M_n||\beta_0||) \quad \text{with} \quad A = \frac{e^2(80 + 64\varepsilon)}{C_1(1-\eta)^2}.$$

Analogously, inequality (19) implies (12). Moreover, the string of the inequalities above implies that

$$\sqrt{k_n + |s_0|}M_n \ge \tilde{N} \ge C_1\sqrt{k_n + |s_0|} \quad \text{and} \quad \sqrt{|s_0|}M_n \ge N \ge C_1\sqrt{|s_0|}$$

which shows that (9) and (12) follow from (21) and (20) implies (10). □

5 Simulation study

In this section we compare different methods of variable selection for high-dimensional logistic regression. We give detailed description of the experiment, its results and conclusions.

We perform simulations for 4 artificially generated data sets described in Table 1. Number of observations is equal to 100 for the first data set and it is increased by 80 for every subsequent data set. Number of variables is a function of n given by $\lfloor \exp((n-20)^{0.37}) \rfloor$ and number of relevant variables vary from 3 for the first data set to 6 for the last one. Vector of true coefficients β_0 is equal to $(-3.5, 1.5, -2)$ for the first data set and is augmented alternately by -2 or 2 for each new relevant variable. The same setting is considered in [8].

For each data set we consider three different dependence structures between variables, namely observations are generated independently from multivariate normal distribution with zero mean and covariance matrix Σ with $\Sigma(i,j) = \rho^{|i-j|}$ for $\rho = -0.5, 0, 0.5$.

Due to computational burden we cannot directly optimize GIC for all subsets of variables containing no more than k_n variables even if k_n is relatively small. Hence, we perform two-stage procedure to find minimal true model. In the first stage we screen moderate number of valuable variables and in the second one we optimize GIC on some subfamily of models consisting of this chosen variables only. There are many statistical procedures such as LASSO, SCAD, Dantzig Selector or Random Forests, which results in ordering variables according to some measure of importance and so can be used as screening methods. In the experiment we order variables according to LASSO for GLM. The most important variable is the one for which corresponding coefficient became nonzero for the largest value of penalty parameter in the LASSO objective function.

We compare three searching procedures: hierarchical (denoted by *hier*), exhaustive (*exh*) and semi-exhaustive (*semexh*). Hierarchical procedure involves minimization of GIC objective function on the nested family of 40 variables chosen in the first step. Since we take into account an empty model- intercept only model- the number of fitted models is 41. In exhaustive procedure we minimize GIC objective function on the family of all submodels of 10 variables chosen by LASSO. The number of fitted models is 1024. Semi-exhaustive method is a version of step forward algorithm with different stop condition. First we fit 40 models, one for each variable chosen by LASSO. Then we choose the best one, so the one witch minimize GIC. Next step is to fit 39 models with two variables- the one chosen previously and each remaining one. We chose the best pair of variables and proceed. The last fitted model is a full model. Including an empty model, we fit $\binom{41}{2} + 1 = 821$ models.

In the first part of the experiment we examine quality of LASSO for GLM as a screening method in considered scenarios. Figure 1 shows estimated probability that after initial screening relevant variables are separated from spurious ones for given values of ρ. This is equivalent to saying that minimal true model s_0 belongs to the nested family of 40 most important variables. We see that values differ significantly. In the easiest case, for $\rho = -0.5$ estimated probability is nearly equal to 1 whereas in the most difficult case, for $\rho = 0.5$ it occurs to be nearly 0.

In the second part we compare searching procedures by taking into account probability of selecting s_0 and selection error. We use EBIC penalty with $\gamma = 1$ which was chosen as the best value in preliminary simulations. The estimated probability of selecting s_0 is shown in Figure 2. In the easiest case for $\rho = -0.5$ hierarchical method works significantly better than remaining two. However, for independent predictors when ordering after first step is of lower quality, exhaustive and semi exhaustive methods are superior to hierarchical one. The tendency is even stronger for positively correlated variables. When LASSO fails in ordering variables semi exhaustive method appears to be the best. In this case probability of selecting s_0 by hierarchical method is close to 0.

We measure error of each searching procedure by mean sum of false positives (FP) and false negatives (FN). Let s_j a set of features chosen in the j-th run. The measure is given by

$$FP + FN = \frac{\sum_{j=1}^{N} |(s_j \cup s_0) \setminus (s_j \cap s_0)|}{N}.$$

Figure (3) shows the result. Conclusions are in line with those from Figure (2). The case of independent variables is the only one when with growth of n error systematically decreases. For negative ρ we see again dominance of hierarchical method with error varying from 0.6 to 0.8. For positive ρ all methods work worse, with error close to the number of relevant variables. Although in this case the best method is semi exhaustive one.

Model	$\|s_0\|$	n	$p = \lfloor \exp((n-20)^{0.37}) \rfloor$	$\beta_0(s_0)$
1	3	100	158	(-3,1.5,-2)
2	4	180	692	(-3,1.5,-2,2)
3	5	260	1993	(-3,1.5,-2,2,-2)
4	6	340	4680	(-3,1.5,-2,2,-2,2)

Table 1: Model specifications.

Acknowledgements

The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL “Information technologies: Research and their interdisciplinary applications”, Agreement UDA-POKL.04.01.01-00-051/10-00.

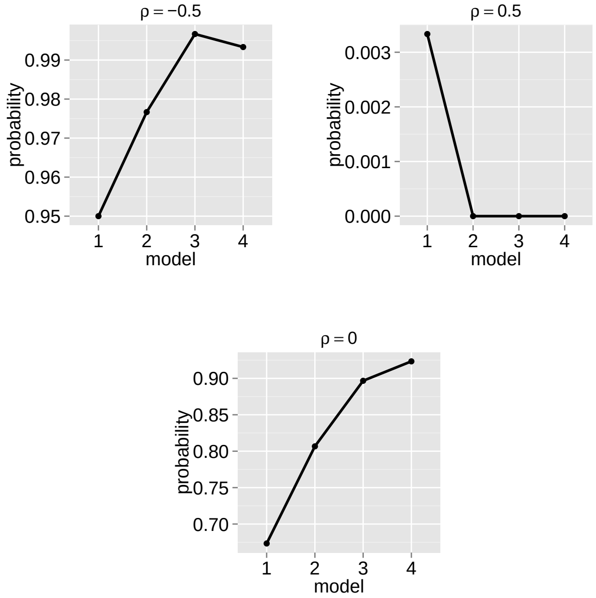

Fig. 1: Estimated probability that all relevant variables proceed the spurious ones after screening for $\rho = -0.5, 0, 0.5$.

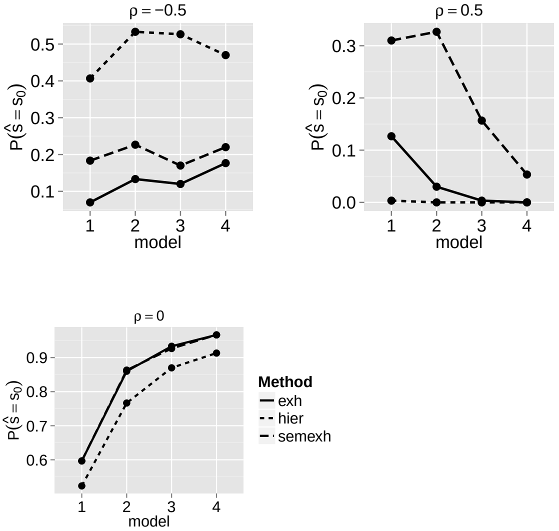

Fig. 2: Estimated probability of selecting true model.

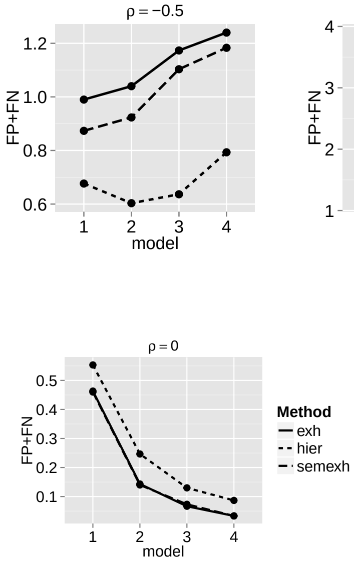

Fig. 3: Mean sum of false positives and false negatives.

References

1. Chen, J., Chen, Z.: Extended BIC for small-n-large-p sparse GLM. Statistica Sinica **22** (2012) 555–574
2. Mielniczuk, J., Szymanowski, H.: Selection consistency of generalized information criterion for sparse logistic model. In: Stochastic Models, Statistics and Their Applications. Volume 122 of Springer Proceedings in Mathematics & Statistics. (2015) 111–119
3. Sin, C., White, H.: Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics **71** (1996) 207–225
4. Broman, K., Speed, T.: A model selection approach for the identification of quantitative trait loci in experimental crosses. Journal of the American Statistical Association **64** (2002) 641–656
5. Chen, J., Chen, Z.: Extended Bayesian Information Criteria for model selection with large model spaces. Biometrika **95** (2008) 759–771
6. Hsu, D., Kakade, S.M., T., Z.: A tail inequality for quadratic forms of subgaussian random vectors. Electronic Communications in Probability **17**(52) (2012) 1–6
7. Zhang, C.H., Huang, J.: The sparsity and bias of the lasso selection in high-dimensional linear regression. Annals of Statistics **36**(4) (2008) 1567–1594
8. Fan, Y., Tang, C.Y.: Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society: Series B **75** (2013) 531–552