# Selection consistency of Generalized Information Criterion for sparse logistic model

Jan Mielniczuk and Hubert Szymanowski

**Abstract** We consider selection rule for small-$n$-large-$P$ logistic regression which consists in choosing a subset of predictors minimizing Generalized Information Criterion over all subsets of variables of size not exceeding $k$. We establish consistency of such rule under weak conditions and thus generalize results of Chen and Chen (2012) to much broader regression scenario which also allows for a more general criterion function than considered there and $k$ depending on a sample size. The results are valid for number of predictors of exponential order of sample size.

## 1 Introduction

Let $X$ be $n \times (P+1)$ design matrix with rows $x'_{i,\cdot}$, columns $x_{\cdot,j}$ and $Y = (y_1, \ldots, y_n)'$ a response vector. All elements of $x_{\cdot,0}$ are equal to 1. We consider a standard logistic regression model in which response $y \in \{0,1\}$ is related to explanatory variable $x \in \mathbb{R}^{P+1}$ by the equation

$$\mathscr{P}(y=1|x) = \exp(x'\beta_0)/[1+\exp(x'\beta_0)], \tag{1}$$

where vector $\beta_0 = (\beta_{0,0}, \ldots, \beta_{0,P})'$ is a vector of parameters. The first coordinate $\beta_{0,0}$ corresponds to the column of ones in design matrix. Remaining coordinates pertain to $P$ explanatory variables. We assume that observations are either deterministic vectors in $\mathbb{R}^{P+1}$ or random variables distributed according to $\mathscr{P}_x$. Data consists of independent observations $(x'_{i,\cdot}, y_i), i = 1, \ldots, n$ and we assume that $x_{i,\cdot}$

Jan Mielniczuk
Institute of Computer Sciences Polish Academy of Sciences and Warsaw University of Technology
e-mail: miel@ipipan.waw.pl

Hubert Szymanowski
Institute of Computer Sciences Polish Academy of Sciences. e-mail: h.szymanowski@ipipan.waw.pl. Supported by POKL research fellowship.

are either deterministic or $x_{i,\cdot} \sim \mathscr{P}_x$ and conditional distribution of $y_i$ given $x_{i,\cdot} = x$ is specified by (1). In the paper we consider the problem of selecting unknown subset of relevant predictors with nonzero coefficients. Thus we want to estimate $s_0 = \{i \in \{1,2,...,P\} : \beta_{0,i} \neq 0\} \cup \{0\}$, where augmentation by 0 means that the fitted model always contains intercept. We assume that $s_0$ is fixed. From now on $\beta_0(s_0) = \beta_0$ will stand for the vector of true parameters in the model $s_0$ augmented by zeros to $(P+1)$-dimensional vector if necessary.
We consider the following Generalized Information Criterion GIC (cf [8])

$$GIC(s) = -2l(\hat{\beta}_s, Y|X_s) + a_n|s|, \tag{2}$$

where $s$ is a given submodel containing $|s|$ explanatory variables and an intercept, $l$ is defined in (4), $\hat{\beta}_s$ is a maximum likelihood estimator for model $s$ (augmented by zeros to $(P+1)$-dimensional vector if necessary) and $a_n$ is a chosen penalty. Observe that $a_n = \log(n)$ corresponds to BIC and $a_n = 2$ to AIC. Consideration of different penalties gained momentum after realization (cf [2]) that BIC penalty, although significantly larger than AIC, can also lead to choice of too many variables e.g. in case of linear model with many predictors. Solutions to this problem such as modified BIC (MBIC) ([1]) and Extended BIC (EBIC) ([3]) were proposed. EBIC criterion stems from putting a certain non-uniform prior on family of models and corresponds to $a_n$ in (2) equal $\log n + 2\gamma \log P$ for some $\gamma > 0$. We also mention extension to generalized linear models (GLMs) of three-stage procedure developed in [7]. For analysis of variable selection under sparsity in a general regression model we refer to [4]. Here, we consider the following selection method which looks for the minimum of GIC over a family of models with number of regressors bounded by a predetermined threshold. Namely, let $\mathscr{M} = \{s : |s| \leq k_n\}$ where $k_n$ is certain nondecreasing sequence of integers and

$$\hat{s}_0 = \arg\min_{s \in \mathscr{M}} GIC(s). \tag{3}$$

This selection method in the case of $k_n = k$ was introduced in [3] for the linear models and extended in [5] to the case of the generalized GLMs. Here, specializing GLM to the case of logistic regression we study behavior of general criterion function (2) under much weaker conditions on design and more general sequence $k_n$ allowing in particular that it diverges to infinity.
In order to heuristically justify $\hat{s}_0$ we need to know that $k_n \geq |s_0|$. As such knowledge is usually unavailable when $k_n$ is fixed, therefore it is natural to assume that $k_n$ is a sequence tending slowly to infinity. This is a principal motivation to extend results in [5] in this direction. We study consistency of $\hat{s}_0$ defined in (3) under fairly general assumptions on design and sequences $k_n$ and $a_n$. In particular we allow for random as well as deterministic predictors. As a byproduct we obtain a rate of consistency of maximum likelihood estimator which is uniform over supermodels of $s_0$.
The main technical improvement in comparison to [5] relies on application of exponential inequality for subgaussian random variables derived in [10]. This allows to circumvent Lemma 1 in [5] which seems unjustified under presented set of as-

sumptions (see line 7 on p. 586 of [5] in which condition that $(\sum_{i=1}^{n} a_{ni}^2)_n$ is bounded is tacitly used) and there is no obvious way to verify amended assumptions in the proof of their crucial Theorem 2. In particular, the condition on EBIC constant $\gamma$ in their result is still a conjecture, see Remark 3 for a result in this direction.

## 2 Main results

Under the logistic regression model (1) and letting $p(s) = 1/(1+e^{-s})$ the conditional log-likelihood function for the parameter $\beta \in \mathbb{R}^{P+1}$ is

$$l(\beta, Y|X) = \sum_{i=1}^{n} \{y_i \log[p(x_{i,\cdot}'\beta)] + (1-y_i)\log[1-p(x_{i,\cdot}'\beta)]\} \tag{4}$$

Maximum likelihood estimator (ML) of $\beta_0$ is denoted by $\hat{\beta}_0$. Note that the score function $S_n(\beta)$, derivative of $l(\beta, Y|X)$, equals $X'(Y-p(\beta))$, where $p(\beta) = (p(x_{1,\cdot}'\beta), \ldots, p(x_{n,\cdot}'\beta))'$. Negative Hessian $H_n(\beta)$ of loglikelihood, equals $X'\Pi(\beta)X$, where $\Pi(\beta) = \mathrm{diag}\{p(x_{1,\cdot}'\beta)(1-p(x_{1,\cdot}'\beta)), \ldots, p(x_{n,\cdot}'\beta)(1-p(x_{n,\cdot}'\beta)))\}$.

All the results of the section are proved for the case of random observations $x_{i,\cdot}$, however (see Remark 1) they remain true under slightly modified assumptions for the case of deterministic $x_{i,\cdot}$ which is the scenario considered in [5] for constant $k$. The proof for the random case requires more care. The conditions we impose on $P = P_n$, $k_n$ and penalty $a_n$ are $k_n^2 \log P_n = o(n)$ and $k_n \log P_n = o(a_n)$. They reduce for constant $k$ to $\log P_n = o(n)$ and $\log P_n = o(a_n)$. EBIC criterion corresponds to $a_n = \log n + 2\gamma \log P_n$ thus for $n \le P_n$ this is a boundary case of the condition $\log P_n = o(a_n)$. We indicate in Remark 3 that our results extend to the case of EBIC for large values of coefficient $\gamma$. Thus for constant $k$ our results are extensions of the results in [5] for EBIC for large penalty constants proved under less demanding conditions. In their paper the case of $P_n = O(\exp(n^\kappa))$ is considered and $\kappa < 1/3$ is assumed whereas our conditions stipulate only that $\kappa < 1$. Then the condition corresponding the first condition on $k_n$ is $k_n = o(n^{(1-\kappa)/2})$.

The lemma below is the main technical tool in proving GIC selection consistency. It follows from Zhang (2009) after noting that binary random variable satisfies $Ee^{t(\xi - E\xi)} \le e^{t^2/8}$ and taking $\sigma = 1/2$, $\varepsilon = \eta^{1/2} - \sigma$ in his Proposition 10.2.

**Lemma 1.** *Let $Y = (y_1, \ldots, y_n)'$ be a vector consisting of independent binary variables and $Z = Z(n \times n)$ be a fixed matrix. For any $\eta > 1/4$*

$$\mathscr{P}(||Z(Y-EY)||^2 \ge \mathrm{tr}(Z'Z)\eta) \le e^{-\eta/20}. \tag{5}$$

We apply the inequality to the case of logistic model when predictors are random. Let $Z = Z(X)$ be a random matrix. It is easily seen by conditioning that the following modification of the above inequality also holds. Namely

$$\mathscr{P}\big(||Z(Y-E(Y|X))||^2 \geq \mathrm{tr}(Z'Z)\eta\big) = E_X \mathscr{P}\big(||Z(Y-E(Y|X))||^2 \geq \mathrm{tr}(Z'Z)\eta|X\big) \leq e^{-\eta/20}.$$

In the following we will always assume that $|s_0| \leq k_n$ which is automatically satisfied for large $n$ if $k_n \to \infty$. We define two families of models: $A_0 = \{s : s_0 \subseteq s \wedge |s| \leq k_n\}$, i.e. family of true models consisting of at most $k_n$ predictors and $A_1 = \{s : s_0 \not\subseteq s \wedge 0 \in s \wedge |s| \leq k_n\}$. Let $\beta_0(s)$ for $s_0 \subseteq s$ denote $\beta_0(s_0)$ augmented by zeros for coordinates belonging to $s \setminus s_0$. The following conditions will be imposed.
$C1$: For every $\eta > 0$ there exist constants $0 < C_1, C_2 < +\infty$ such that for all $n$

$$\mathscr{P}\big\{C_1 \leq \min_{s\in A_1} \lambda_{min}(\frac{1}{n}H_n(\beta_0(s\cup s_0))) \leq \max_{s\in A_1} \lambda_{max}(\frac{1}{n}X'_{s\cup s_0}X_{s\cup s_0}) \leq C_2\big\} \geq 1-\eta.$$

$C2$: For every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $\eta > 0$ and every $n \geq n_0(\varepsilon,\delta,\eta)$, with $\leq_L$ denoting Loewner ordering

$$\mathscr{P}\big\{\forall_{s:|s|\leq k_n}\forall_{||\beta(s\cup s_0)-\beta_0(s\cup s_0)||\leq\delta}(1-\varepsilon)H_n(\beta_0(s\cup s_0)) \leq_L H_n(\beta(s\cup s_0)) \leq_L (1+\varepsilon)H_n(\beta_0(s\cup s_0))\big\} \geq 1-\eta.$$

*Remark 1.* Results of this section are valid for deterministic $X$ with slightly modified but simpler proofs with the following changes of assumptions. Condition $C1$ is replaced by:
$C1'$: There exist constants $0 < C_1, C_2 < +\infty$ such that for all $n$

$$C_1 \leq \min_{s\in A_1} \lambda_{min}(\frac{1}{n}H_n(\beta_0(s\cup s_0))) \leq \max_{s\in A_1} \lambda_{max}(\frac{1}{n}X'_{s\cup s_0}X_{s\cup s_0}) \leq C_2 \tag{6}$$

and $C2$ by:
$C2'$: For any $\varepsilon > 0$ there exists $\delta > 0$ such that for sufficiently large $n$

$$\forall_{s:|s|\leq k_n}\forall_{||\beta(s\cup s_0)-\beta_0(s\cup s_0)||\leq\delta}(1-\varepsilon)H_n(\beta_0(s\cup s_0)) \leq_L H_n(\beta(s\cup s_0)) \leq_L (1+\varepsilon)H_n(\beta_0(s\cup s_0)).$$

The first of this assumptions is a slight strengthening of assumption A4 in [5] whereas the second one is the same as their A5. Note that since $X'\Pi X \leq_L X'X$ condition $C1'$ implies that all subsets of columns of $X$ of size at most $k_n$ are linearly independent. It is shown in [9] that condition $C1$ (with Hessian $H_n$ replaced by moment matrix $X'X$) is satisfied for normal predictors under appropriate assumptions on their covariance matrices, $p$ and $k_n$. Condition $C'_2$ is analogous to condition (N) in [6] (cf (3.4), p. 348 there). Moreover, note that the assumption $\max_{s\in A_1} \mathrm{tr}(X'_{s\cup s_0}X_{s\cup s_0}) = O_P(k_n n)$ used in Theorem 1 below follows from $C1$. Further on it is replaced by $C1$.

**Theorem 1.** *Let $X = (x_{ij})$ $i = 1,...,n$; $j = 1,...,k_n$ be a random matrix such that $\max_{s\in A_1} tr(X'_{s\cup s_0}X_{s\cup s_0}) = O_P(k_n n)$. Then*

$$\max_{s\in A_1}||S_n(\beta_0(s\cup s_0))|| = O_P\left(k_n\sqrt{n\log P_n}\right). \tag{7}$$

*Proof.* Intersecting set $\{\max_{s\in A_1}||S_n(\beta_0(s\cup s_0))|| \geq M_1 k_n\sqrt{n\log P_n}\}$ with an event $\{\max_{s\in A_1}\mathrm{tr}(X'_{s\cup s_0}X_{s\cup s_0}) \leq M_2 k_n n\}$ and its complement we see that probability of the second set can be made arbitrarily small by a choice of $M_2$ whereas the first can be bounded using (1) by

$$\mathscr{P}\left(\max_{s\in A_1}||X'_{s\cup s_0}(Y-p(\beta(s\cup s_0)))|| \geq \frac{M_1}{\sqrt{M_2}}\sqrt{k_n\log P_n \mathrm{tr}(X'_{s\cup s_0}X_{s\cup s_0})}\right)$$
$$\leq P_n^{k_n}\exp(-\frac{M_1^2}{20M_2}k_n\log P_n).$$

For sufficiently large constant $M_1$ the last expression is arbitrarily small.

**Lemma 2.** *Let $k_n^2\log P_n = o(n)$. Then, under assumptions C1 and C2 we have*

$$\max_{s\in A_0}||\hat{\beta}_0(s)-\beta_0(s)|| = O_P\left(k_n\sqrt{\frac{\log P_n}{n}}\right). \tag{8}$$

*Proof.* Let $\beta_u(s) = \beta_0(s)+\gamma_n u$ for a vector $u$ such that $||u|| = 1$ and $\gamma_n = C_0 k_n\sqrt{\log P_n/n}$. For any $\delta > 0$ and sufficiently large $n$ in view of the condition on $k_n$ we have $||\beta_u(s)-\beta_0(s)|| \leq \delta$ and assumption $C2$ becomes applicable. We show that

$$\mathscr{P}\{\exists u : ||u|| = 1, \max_{s\in A_0, s\neq s_0}\{l_n(\beta_u(s))-l_n(\beta_0(s))\} > 0\} = o(1). \tag{9}$$

Let us fix $\varepsilon_0 > 0$ and let $\delta_0$ be the value of $\delta$ corresponding to $\varepsilon_0$ in assumption $C2$. Denote by $\mathscr{A}_n$ the following event

$$\{C_1 \leq \min_{s\in A_0}\lambda_{min}(\frac{1}{n}H_n(\beta_0(s))) \leq \max_{s\in A_0}\lambda_{max}(\frac{1}{n}X'_sX_s) \leq C_2\}$$
$$\cap\ \{\forall_{s:|s|\leq k_n}\forall_{||\beta(s\cup s_0)-\beta_0(s\cup s_0)||\leq\delta}(1-\varepsilon)H_n(\beta_0(s\cup s_0)) \leq_L H_n(\beta(s\cup s_0))$$
$$\leq_L (1+\varepsilon)H_n(\beta_0(s\cup s_0))\}.$$

Moreover, $\beta^*$ will stand for generic vector belonging to the line segment with endpoints $\beta_u(s)$ and $\beta_0(s)$ i.e. having the form $\lambda\beta_u(s)+(1-\lambda)\beta_0(s)$ for some $\lambda\in[0,1]$. It follows from assumptions $C1$ and $C2$ that $\mathscr{P}(\mathscr{A}_n)$ is arbitrarily close to 1 for large $n$ and sufficiently small $C_1$ and $C_2^{-1}$.
We have on $\mathscr{A}_n$ with some $\beta^*$

$$\mathscr{P}\{\exists u : ||u|| = 1, \max_{s\in A_0}\{l_n(\beta_u(s))-l_n(\beta_0(s))\} > 0\}$$
$$\leq P_n^{k_n}\max_{s\in A_0}\mathscr{P}\{\exists u : ||u|| = 1, u'S_n(\beta_0) > \frac{1}{2}\gamma_n u'H_n(\beta^*)u\}$$
$$\leq P_n^{k_n}\max_{s\in A_0}\mathscr{P}\{\exists u : ||u|| = 1, u'S_n(\beta_0) > \frac{1}{2}(1-\varepsilon_0)\gamma_n u'H_n(\beta_0)u\}$$

$$\leq P_n^{k_n} \max_{s\in A_0} \mathscr{P}\{||S_n(\beta_0)|| > \frac{C_1}{2}(1-\varepsilon_0)\gamma_n n\}$$

where the last inequality follows by taking $u = S_n(\beta_0)/||S_n(\beta_0)||$. Since $\mathscr{A}_n \subset \{\max_{s\in A_1\cup s_0} \mathrm{tr}(X_s'X_s) \leq C_2 k_n n\}$, we have from (1) that the last expression tends to zero when constant $C_1$ is sufficiently large. As $l_n(\beta(s))$ is a concave function for any $s$, it follows that, with probability tending to 1, estimator $\hat{\beta}_0(s)$ exists and belongs to $\gamma_n$-neighborhood of $\beta_0(s)$ uniformly for all $s \in A_0$.

*Remark 2.* Note that for $k_n = k$ and $P_n = O(\exp(n^\kappa))$ it follows from (8) that the uniform rate of convergence of $\hat{\beta}$ over supersets of $s_0$ is $k_n(\log P_n/n)^{1/2} = O(n^{\frac{\kappa-1}{2}})$, thus assuming $\kappa \in (0,1/3)$ as in [5] we obtain better rate of convergence than $O_P(n^{-1/3})$ determined in their Theorem 1.

**Theorem 2.** *Let assumptions C*1 *and C*2 *hold. Moreover, assume* $a_n = o(n)$ *and* $k_n^2 \log P_n = o(n)$. *Then*

$$\mathscr{P}(\min_{s\in A_1} GIC(s) \leq GIC(s_0)) \to 0. \tag{10}$$

*Proof.* Let $\varepsilon_0 > 0$ and $\delta_0$ corresponds to $\varepsilon_0$ in assumption $C2$. Moreover, $\tilde{s} = s \cup s_0$. Let us denote by $\breve{\beta}(\tilde{s})$ ML estimator $\hat{\beta}(s)$ augmented with zeros corresponding to the elements in $s_0 \setminus s$. Note that

$$||\breve{\beta}(\tilde{s}) - \beta_0(\tilde{s})|| \geq ||\beta_0(s_0 \setminus s)|| \geq \beta_{min} > 0 \tag{11}$$

for all $s \in A_1$ where $\beta_{min} = \min_{i\in s_0\setminus s} |\beta_{0,i}|$.
Let us fix $s \in A_1$ and denote $B = \{\beta : ||\beta(\tilde{s}) - \beta_0(\tilde{s})|| = r\}$, where
$r = \min\{\beta_{min}/2, \delta_0/2\}$. We have from Schwarz inequality and assumptions $C1$ and $C2$ on event $\mathscr{A}_n$ defined in the proof of Lemma 2 that for all $\beta \in B$ and some $\beta^*$ between $\beta(\tilde{s})$ and $\beta_0(\tilde{s})$ the difference $l_n(\beta(\tilde{s})) - l_n(\beta_0(\tilde{s}))$ equals

$$\begin{aligned}
&[\ \beta(\tilde{s}) - \beta_0(\tilde{s})]' S_n(\beta_0(\tilde{s})) - \frac{1}{2}[\beta(\tilde{s}) - \beta_0(\tilde{s})]' H_n(\beta^*)[\beta(\tilde{s}) - \beta_0(\tilde{s})] \\
&\leq ||\beta(\tilde{s}) - \beta_0(\tilde{s})|| \cdot ||S_n(\beta_0(\tilde{s}))|| - \frac{1}{2}(1-\varepsilon_0)[\beta(\tilde{s}) - \beta_0(\tilde{s})]' H_n(\beta_0(\tilde{s}))[\beta(\tilde{s}) - \beta_0(\tilde{s})] \\
&\leq ||\beta(\tilde{s}) - \beta_0(\tilde{s})|| \cdot ||S_n(\beta_0(\tilde{s}))|| - \frac{C_1}{2}(1-\varepsilon_0) \cdot n||\beta_0(\tilde{s}) - \beta(\tilde{s})||^2.
\end{aligned}$$

It follows from Theorem 1 and the definition of sphere $B$ that the last expression is bounded from above on a set of arbitrarily large positive measure by $-M_2 n$ for some positive constant $M_2$. By a concavity of function $l_n$ and a fact $\breve{\beta} \notin B$ on $\mathscr{A}_n$

$$\begin{aligned}
&l_n(\hat{\beta}(s)) - l_n(\hat{\beta}(s_0)) = l_n(\breve{\beta}(\tilde{s})) - l_n(\hat{\beta}(s_0)) \leq l_n(\breve{\beta}(\tilde{s})) - l_n(\beta_0(\tilde{s})) \\
&\leq l_n(\beta^*(\tilde{s})) - l_n(\beta_0(\tilde{s})) \leq -M_2 \cdot n,
\end{aligned}$$

where $\beta^*(\tilde{s})$ is any element of $B$. This and an assumption $a_n = o(n)$ yields that

$$\begin{aligned}
&\mathscr{P}(\min_{s\in A_1} GIC(s) \le GIC(s_0)) = \\
&= \mathscr{P}(\{\max_{s\in A_1}(l_n(\hat\beta(s)) - l_n(\hat\beta(s_0)) + a_n(|s_0| - |s|)) \ge 0\} \cap \mathscr{A}_n) + \mathscr{P}(\mathscr{A}_n^c) \\
&\le \mathscr{P}(-M_2 n + a_n|s_0| \ge 0) + \mathscr{P}(\mathscr{A}_n^c) \to 0.
\end{aligned}$$

The next result states that with probability tending to 1 GIC choses the smallest true model among all true models. Note that for $P_n = O(n^\kappa)$ the second condition on $k_n$ is implied by $k_n = O(n^{1/2-\varepsilon})$ for any $\varepsilon > 0$.

**Theorem 3.** *Under assumptions C*1 *and C*2 *for* $k_n \log P_n = o(a_n)$ *and* $k_n^2 \log P_n = o(n)$

$$\mathscr{P}(\min_{s\in A_0, s\neq s_0} GIC(s) \le GIC(s_0)) \to 0. \tag{12}$$

*Proof.* Let $s \in A_0$ and $s \neq s_0$. We have on event $\mathscr{A}_n$ and for some $\beta^*, \beta^{**}$ that

$$\begin{aligned}
&l_n(\hat\beta(s)) - l_n(\hat\beta(s_0)) \le l_n(\hat\beta(s)) - l_n(\beta_0(s)) \\
&= [\hat\beta(s) - \beta_0(s)]' S_n(\beta_0(s)) - \frac{1}{2}[\hat\beta(s) - \beta_0(s)]' H_n(\beta^{**})[\hat\beta(s) - \beta_0(s)].
\end{aligned}$$

Note that

$$S_n(\hat\beta(s)) - S_n(\beta_0(s)) = -H_n(\beta^*)[\hat\beta(s) - \beta_0(s)], \tag{13}$$

and in view of $C1$, $C2$ and Lemma 2 $H_n(\beta^*)$ is invertible. Thus

$$\hat\beta(s) - \beta_0(s) = H_n(\beta^*)^{-1} S_n(\beta_0(s)). \tag{14}$$

Therefore, the right side of inequality (13) can be rewritten as

$$[S_n(\beta_0(s))]' H_n(\beta^*)^{-1} S_n(\beta_0(s)) - \frac{1}{2}[S_n(\beta_0(s))]' H_n(\beta^*)^{-1} H_n(\beta^{**}) H_n(\beta^*)^{-1} S_n(\beta_0(s)).$$

Assumption $C2$ and a fact that $A \le_L B \Rightarrow A^{-1} \le_L A^{-1} B A^{-1}$ yields

$$l_n(\hat\beta(s)) - l_n(\hat\beta(s_0)) \le c[S_n(\beta_0(s))]' H_n(\beta_0(s))^{-1} S_n(\beta_0(s)) \tag{15}$$

for some constant $c$ independent of $s \in A_0$. Hence, on event $\mathscr{A}_n$ we have

$$\mathscr{P}(\max_{s\in A_0, s\neq s_0} (l_n(\hat\beta(s)) - l_n(\hat\beta(s_0)) - (|s| - |s_0|)a_n) \ge 0) \tag{16}$$

$$\le \max_{s\in A_0, s\neq s_0} P_n^{k_n} \mathscr{P}(c S_n(\beta_0(s))' H_n(\beta_0(s))^{-1} S_n(\beta_0(s)) \ge a_n(|s| - |s_0|)) \tag{17}$$

$$= \max_{s\in A_0, s\neq s_0} P_n^{k_n} \mathscr{P}(c||Z_s(Y - p(\beta(s)))||^2 \ge a_n(|s| - |s_0|)) \tag{18}$$

where $Z_s = (X_s' \Pi_s X_s)^{-\frac{1}{2}} X_s'$. It is seen that on $\mathscr{A}_n$ $tr(Z_s' Z_s) \le M_1|s|$. It follows now from Zhang's inequality that for fixed $s \in A_0$ on $\mathscr{A}_n$ with $M_2 = (20cM_1)^{-1}$

$$P_n^{k_n} \mathscr{P}(c||Z_s(Y - EY)||^2 \ge a_n(|s| - |s_0|)) \le P_n^{k_n} E\left[\exp\left(-\frac{a_n(|s| - |s_0|)}{20c \cdot \mathrm{tr}(Z_s' Z_s)}\right)\right]$$

$$\leq P_n^{k_n} \exp\left( -M_2 \frac{(|s|-|s_0|)a_n}{|s|} \right) \leq P_n^{k_n} \exp\left( -C \frac{(|s|-|s_0|)k_n \log P_n}{|s|} \right),$$

where $C$ may be chosen arbitrarily large and independent of $s$. Finally, we have as $\min_{s\in\mathscr{A}_0}(|s|-|s_0|)/|s| = 1/(|s_0|+1)$

$$\mathscr{P}(\min_{s\in\mathscr{A}_0} GIC(s) \leq GIC(s_0)) \tag{19}$$

$$\leq \mathscr{P}(\{\max_{s\in\mathscr{A}_0}(l_n(\hat{\beta}(s)) - l_n(\hat{\beta}(s_0)) + a_n(|s_0|-|s|)) \geq 0\} \cap \mathscr{A}_n) + \mathscr{P}(\mathscr{A}_n^c) \tag{20}$$

$$\leq P_n^{k_n} \exp\left( -C \frac{k_n \log P_n}{|s_0|+1} \right) + \mathscr{P}(\mathscr{A}_n^c) \to 0 \tag{21}$$

for sufficiently large constant $C$ independent of $s$.

*Remark 3.* Theorem 2 is applicable to EBIC as the sole condition on $a_n$ there is $a_n = o(n)$. Moreover, Theorem 3 remains true for EBIC in the case when $n = o(P_n)$ and constant $k_n = k$ if penalty coefficient $\gamma$ is large enough. Indeed, substitution $a_n = \log n + 2\gamma \log P_n$ in (19) leads to the inequality

$$P_n^k \mathscr{P}(c||Z_s(Y-EY)||^2 \geq a_n(|s|-|s_0|)) \leq P_n^k \exp\left( -M_2 \frac{2\gamma \log P_n}{|s_0|+1} \right). \tag{22}$$

Thus if $\gamma > 0.5k(|s_0|+1)M_2^{-1}$ (12) holds.

## References


1. Bogdan M, Doerge R and Ghosh J (2004) Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics, 167:989-999
2. Broman K and Speed T (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J Am Statist Assoc, 64:641-656
3. Chen J and Chen Z (2008) Extended Bayesian Information Criteria for model selection with large model spaces. Biometrika, 95:759-771
4. Comminges L and Dalalyan A (2012) Tight conditions for consistency of variable selection in the context of high dimensionality. Ann Statist. 40: 2667- 2696
5. Chen J and Chen Z (2012) Extended BIC for small-n-large-p sparse GLM. Statist Sinica, 22:555-574
6. Fahrmeir L and Kaufmann H (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann Statist, 1(13):342-368
7. Pokarowski P and Mielniczuk J (2014) Combined $\ell_1$ and greedy $\ell_0$ least squares for linear model selection. submitted
8. Sin C and White H (1996) Information criteria for selecting possibly misspecified parametric models. J Econometrics, 71:207-225
9. Wang H (2009) Forward regression for ultra-high dimensional variable screening. J Am Statist Assoc, 104:1512-1524
10. Zhang T (2009) Some sharp performance bounds for least squares regression with $L_1$ regularization. Ann Statist, 37:2109-2144