Model selection in logistic regression using p-values and greedy search

Jan Mielniczuk[1,2] and Paweł Teisseyre[1] ⋆

[1] Institute of Computer Science, Polish Academy of Sciences, Ordona 21, 01–237 Warsaw, Poland
miel@ipipan.waw.pl teisseyrep@ipipan.waw.pl
[2] Warsaw University of Technology, Faculty of Mathematics and Information Science, Politechniki Sq. 1, 00–601 Warsaw, Poland

Abstract. We study new logistic model selection criteria based on p-values. The rules are proved to be consistent provided suitable assumptions on design matrix and scaling constants are satisfied and the search is performed over the family of all submodels. As a byproduct, consistency of Bayesian Information Criterion (BIC) for logistic regression models proved by Qian and Field in [11] is obtained under milder assumptions. Moreover, we investigate practical performance of the introduced criteria in conjunction with greedy search methods such as initial ordering, forward and backward search and genetic algorithm which restrict the range of family of models over which an optimal value of the respective criterion is sought. Scaled minimal p-value criterion with initial ordering turns out to be a promising alternative to BIC.

Keywords: logistic regression, model selection, greedy search methods, p-values

1 Introduction

Model selection and properties of ensuing postmodel selection estimators is one of the central subjects in theoretical statistics and its applications. In particular, variable selection in regression models with dichotomous response e.g. a logistic models is widely used (cf. e.g. [6]). In the paper we focus on the first from the two main related problems of statistical modelling which are explanation (i.e. finding an adequate model) and prediction. The present paper provides some insights into behaviour of logistic model selection criteria based on p-values. In this approach introduced for parametric families of densities by Pokarowski and Mielniczuk ([9]) competing models are viewed as alternative hypotheses with null hypothesis being the minimal model and choosing the model for which appropriately scaled p-value of LRT test statistic is the smallest one. In the paper we investigate basic property of such rules concerning identification of the true model namely their consistency which means that probability of choosing a minimal

⋆ corresponding author

true model tends to 1. Moreover, we focus on the situation when the number of potential regressors is large and only search of an optimal model over a restricted family of submodels feasible. That is, we investigate in numerical experiments the performance of the considered criteria coupled with greedy search methods such as initial ordering, forward and backward search and genetic algorithm. We provide some evidence that a scaled minimal p-value criterion introduced in Section 2 in conjunction with forward search compares favourably with Bayesian Information Criterion.
The paper is organized as follows. In Section 2 we outline some of the basics for logistic regression models and introduce the considered criteria, in Section 3 the required conditions, some auxiliary results and main results are presented. Section 4 provides conclusions from numerical experiments. The outline of proofs are deferred to the appendix.

2 Logistic regression model and model selection criteria

2.1 Logistic regression model

Let $y_1,\ldots,y_n$ be a sequence of independent random variables such that y_i has Bernoulli distribution $P(y_i=1)=1-P(y_i=0)=\pi_i$. Let $\mathbf{x}_1',\ldots,\mathbf{x}_n'$ be a sequence of associated covariates, $\mathbf{x}_i=(x_{i,1},\ldots,x_{i,M})'$ with $\mathbf{x}'$ denote a transpose of $\mathbf{x}$. Suppose that the expectations of response variables are related to explanatory variables by the logistic model

$$\pi_i(\boldsymbol{\beta})=\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}. \tag{1}$$

Vector $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_M)'$ is an unknown vector of parameters. Denote $\mathbf{Y}_n=(y_1,\ldots,y_n)'$ as the response vector and $\mathbf{X}_n=(\mathbf{x}_1,\ldots,\mathbf{x}_n)'$ as the design matrix. The conditional log-likelihood function for the parameter $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta},\mathbf{Y}_n|\mathbf{X}_n)=\sum_{i=1}^{n}\{y_i\log[\pi_i(\boldsymbol{\beta})]+(1-y_i)\log[1-\pi_i(\boldsymbol{\beta})]\}. \tag{2}$$

The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ is defined to be $\hat{\boldsymbol{\beta}}=\arg\max_{\boldsymbol{\beta}\in\mathbf{R}^M} l(\boldsymbol{\beta},\mathbf{Y}_n|\mathbf{X}_n)$. Let $\mathit{\Pi}(\boldsymbol{\beta})=\mathrm{diag}\{\pi_1(\boldsymbol{\beta})(1-\pi_1(\boldsymbol{\beta})),\ldots,\pi_n(\boldsymbol{\beta})(1-\pi_n(\boldsymbol{\beta}))\}$. A useful quantity is the Fisher information matrix for the parameter $\boldsymbol{\beta}$ which is defined as

$$\mathbf{I}_n(\boldsymbol{\beta})=-\mathbf{E}\frac{\partial^2 l(\boldsymbol{\beta},\mathbf{Y}_n|\mathbf{X}_n)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}=-\frac{\partial^2 l(\boldsymbol{\beta},\mathbf{Y}_n|\mathbf{X}_n)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}=\mathbf{X}_n'\mathit{\Pi}(\boldsymbol{\beta})\mathbf{X}_n.$$

Define also the score function $s_n(\boldsymbol{\beta})=\frac{\partial l(\boldsymbol{\beta},\mathbf{Y}_n|\mathbf{X}_n)}{\partial\boldsymbol{\beta}}$.

Suppose now that some covariates do not contribute to the prediction of expectation of $\mathbf{Y}$ in a sense that the corresponding coefficients are zero. It is assumed that the true model is a submodel of (1). As any submodel of (1) containing p_j variables $(x_{i,j_1},\ldots,x_{i,j_{p_j}})'$ is described by set of indexes $j=\{j_1,\ldots,j_{p_j}\}$

it will be referred to as model j. The minimal true model will be denoted by t. So p_t is the number of nonzero coefficients in equation (1). The empty model for which $P(y_i = 1) = 1 - P(y_i = 0) = \frac{1}{2}$ will be denoted briefly by 0 and the full model (1) by $f = \{1, \ldots, M\}$. Vector $\boldsymbol{\beta}_j$ of parameters for model j is augmented to $M \times 1$ vector in such a way that $\beta_k = 0$, for $k \notin j$. Let $\hat{\boldsymbol{\beta}}_j$ be a Maximum Likelihood Estimator (MLE) of $\boldsymbol{\beta}$ calculated for the model j also augmented by zeros to $M \times 1$ vector. We denote $\hat{\boldsymbol{\beta}}_f$, MLE in the full model, briefly by $\hat{\boldsymbol{\beta}}$. Let $\mathcal{M}$ be a family of all subsets of a set f.

2.2 Model Selection Criteria

The main objective is to identify the minimal true model t using data $(\mathbf{X}_n, \mathbf{Y}_n)$. Consider two models j and k such that the j model is nested within the model k. Denote by D^n_{jk} likelihood ratio test (LRT) statistic, based on conditional likelihoods given $\mathbf{X}_n$, for testing H_0 : model j is adequate against hypothesis H_1 : model k is adequate whereas j is not, equal to

$$D^n_{jk} = 2[l(\hat{\boldsymbol{\beta}}_k, \mathbf{Y}_n|\mathbf{X}_n) - l(\hat{\boldsymbol{\beta}}_j, \mathbf{Y}_n|\mathbf{X}_n)]. \tag{3}$$

Let F and G be univariate cumulative distribution functions and T be a test statistic which has distribution function G not necessarily equal to F. Let $p(t|F) = 1 - F(t)$. By p-value of a test statistic T given the reference distribution F (which will correspond to the approximate null distribution) we will mean $p(T|F)$. We will consider p-values of statistic D^n_{jk} given chi square distribution with $p_k - p_j$ degrees of freedom in view of Fahrmeir (1987) ([4]) who established asymptotic distribution of D^n_{jk} for generalized linear model (see the statement of it in Theorem 1 for the logistic regression). In order to make notation simpler, $p(D^n_{jk}|\chi^2_{p_k - p_j})$ will be denoted as $p(D^n_{jk}|p_k - p_j)$. From now on deviance D^n_{jk} of the model k from the model j will be formally defined by (3) even if j is not nested within k.

We define the model selection criteria based on p-values of D^n_{jk} (cf [9]).

Minimal P-Value Criterion (mPVC)

$$M^n_m = \mathrm{argmin}_{j \in \mathcal{M}} e^{p_j a_n} p(D^n_{0j}|p_j),$$

where $p(D^n_{00}|0) = e^{a_n}/\sqrt{n}$. Observe that when $a_n = 0$ then from among the pairs $\{(H_0, H_j)\}$ we choose a pair for which we are most inclined to reject H_0 and we select the model corresponding to the most convincing alternative hypothesis. If $a_n > 0$ the scaling factor $e^{p_j a_n}$ is interpreted as additional penalization for the complexity of a model. We will assume throughout that $a_n = O(\log(n))$.

Maximal P-Value Criterion (MPVC)

$$M^n_M = \mathrm{argmax}_{j \in \mathcal{M}} e^{-p_j a_n} p(D^n_{jf}|M - p_j),$$

where $p(D^n_{ff}|0) = 1$, $a_n \to \infty$ and $a_n = O(\log(n))$. The motivation is similar as in the case of mPVC, namely we choose a model which we are least inclined to reject when compared to the full model f. We stress that the additional

assumption $a_n \to \infty$ needed for consistency of MPVC is not required to prove consistency of mPVC.
Bayesian Information Criterion (BIC) is defined as
Bayesian Information Criterion (BIC)

$$BIC^n = \text{argmin}_{j \in \mathcal{M}}[-2l(\hat{\boldsymbol{\beta}}_j, \mathbf{Y}_n|\mathbf{X}_n) + p_j \log(n)].$$

2.3 Model Selection Criteria Based on a Restricted Search

Selection rules given above require calculations for all members of $\mathcal{M}$ what for large number of possible regressors carries considerable and often enormous computational cost. In order to mend this drawback we consider the following methods whose aim is to restrict the family of models over which the optimal value of the criterion is sought. The restricted search can be applied for any of the criteria considered above. Assume temporarily that the minimum of the criterion is sought.

1. **Initial ordering (I0).** The covariates $\{j_1, j_2, \ldots, j_M\}$ are ordered with respect to the decreasing values of LRT statistics
$$D^n_{(f-\{j_1\})f} \geq D^n_{(f-\{j_2\})f} \geq \cdots \geq D^n_{(f-\{j_M\})f}.$$
Let $\mathcal{M}_{IO} = \{\{0\}, \{j_1\}, \{j_1, j_2\}, \ldots, \{j_1, j_2, \ldots, j_M\}\}$. The selection criteria with initial ordering $M^n_{m,IO}$, $M^n_{M,IO}$, BIC^n_{IO} are defined analogously as M^n_m, M^n_M, BIC^n. The difference is that the optimization is now performed over set $\mathcal{M}_{IO}$. Note that now only $M+1$ instead of 2^M possible models are fitted.
2. **Forward selection (FS).**The procedure begins with the null model and at each stage adds the attribute that yields the greatest decrease in the given criterion function. The final model is obtained when none of the remaining variables leads to the decrease of the criterion.
3. **Backward elimination (BE).** A nested sequence of models of decreasing dimensionality beginning with the full model is constructed. At each step a variable is omitted that yields the greatest decrease in the criterion function. FS and BE are widely used techniques for model selection (see e.g. [7])
4. **Genetic algorithm (GA).** We used an algorithm proposed in [12] with the settings considered there. Each model, also called an individual, is described by a binary vector $\mathbf{z} = (z_1, \ldots, z_M)'$, where j^{th} gene $z_j = 1$ indicates that j^{th} variable is included in the model. Each generation consists of 40 individuals (models) . The initial population is randomly generated in such a way that $z_j = 1$ with probability 0.9. Instead of using fitness proportionate selection as in [12] we applied truncation selection (see e.g. in [8]) which performed better. Namely the two individuals with the smallest values of the given criterion function are selected as parents. To create the offspring two integer points are randomly selected from the interval $[0, M-1]$, and ordered so that $v_2 \geq v_1$. The offspring gets the first v_1 genes from the first parent, the next $v_2 - v_1$ genes from the second parent and the last $M - v_2$ genes again from the first parent. This procedure is repeated 40 times to match the size of

the previous generation.The individuals of each generation are also mutated before model estimation. Each gene of each individual is flipped, from zero to one or vice versa, with probability 0.01. The procedure outlined above is repeated until convergence is achieved.

3 Consistency properties of introduced criteria

We first state some properties of LRT statistic D^n_{jk}. They are necessary to prove the consistency of selection rules M^n_m and M^n_M introduced in the previous section. We discuss now some technical conditions imposed on the logistic model. We assume throughout that $\mathbf{X}'_n\mathbf{X}_n$ has full rank. This condition will ensure that the information matrix $\mathbf{I}_n(\boldsymbol{\beta})$ is positive definite for all $\boldsymbol{\beta} \in \mathbf{R}^M$ as $\Pi(\boldsymbol{\beta})$ is positive definite. Let $\lambda_{\min}$ ($\lambda_{\max}$) denote the smallest (the largest) eigenvalue of a symmetric matrix. Let $\mathbf{A}^{1/2}$ be a left square root of positive definite matrix $\mathbf{A}$, i.e. $\mathbf{A}^{1/2}(\mathbf{A}^{1/2})' = \mathbf{A}$. The right square root is defined as $\mathbf{A}^{T/2} = (\mathbf{A}^{1/2})'$. As a left square root one can take $\mathbf{Q}\Lambda^{1/2}\mathbf{Q}'$, where $\mathbf{Q}\Lambda\mathbf{Q}'$ is a spectral decomposition of $\mathbf{A}$ or the lower triangular matrix from Cholesky decomposition. $\mathbf{A}^{-1/2}$ will denote the inverse of $\mathbf{A}^{1/2}$. $W_n = O_P(1)$ means that the sequence of random variables is bounded in probability and $\xrightarrow{d}$ ($\xrightarrow{P}$) denotes convergence in distribution (in probability). The following conditions will be needed.

(A1) $\gamma n \leq \lambda_{\min}(\mathbf{I}_n(\boldsymbol{\beta}_t)) \leq \lambda_{\max}(\mathbf{I}_n(\boldsymbol{\beta}_t)) \leq \kappa n$ holds for some positive constants γ and κ.

(A2) $\max_{1\leq i\leq n} ||\mathbf{x}_i||^2 \log(n)/n \to 0$, as $n \to \infty$.

As $\log(n)/n$ is decreasing, condition (A2) is equivalent to a condition $||\mathbf{x}_n||^2 \log(n)/n \to 0$. Define the sequence $N_n(\delta)$, $\delta > 0$, of neighborhoods of $\boldsymbol{\beta}_t$ as

$$N_n(\delta) = \{\boldsymbol{\beta} : ||\mathbf{I}_n(\boldsymbol{\beta}_t)^{T/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_t)|| \leq \delta\}, \quad n = 1, 2, \ldots.$$

The following auxiliary condition is assumed for proving Theorem 1.

(F) For all $\delta > 0$,

$$\max_{\boldsymbol{\beta}\in N_n(\delta)} ||\mathbf{I}_n(\boldsymbol{\beta}_t)^{-1/2}\mathbf{I}_n(\boldsymbol{\beta})\mathbf{I}_n(\boldsymbol{\beta}_t)^{-T/2} - \mathbf{I}|| \to 0,$$

as $n \to \infty$.

The following Theorem (cf [4]) states the asymptotic result of LRT statistic D^n_{jk}.

Theorem 1 *Assume $\lambda_{\min}(\mathbf{I}_n(\boldsymbol{\beta}_t)) \to \infty$ as $n \to \infty$ and (F). Then $D^n_{jk} \xrightarrow{d} \chi^2_{p_k - p_j}$ as $n \to \infty$ provided that model j is true.*

Remark 1 *Assume $\max_{1\leq i\leq n} ||\mathbf{x}_i||^2/n \to 0$, as $n \to \infty$ and (A1). Then condition (F) holds. Namely letting $\delta_n^2 = \max_{1\leq i\leq n} \mathbf{x}'_i\mathbf{I}_n^{-1}(\boldsymbol{\beta}_t)\mathbf{x}_i$ and in view of (A1)-(A2)*

$$\delta_n^2 \leq \max_{1\leq i\leq n} ||\mathbf{x}_i||^2 \lambda_{\max}(\mathbf{I}_n^{-1}(\boldsymbol{\beta}_t)) \leq \frac{\max_{1\leq i\leq n} ||\mathbf{x}_i||^2}{n\gamma} \to 0.$$

It follows now from Corollary 2 in [5] that the convergence $\delta_n^2 \to 0$ implies (F).

In particular it follows from the above Remark that conditions (A1) and (A2) imply (F). Recall that $\boldsymbol{\beta}_t = (\beta_{t,1}, \ldots, \beta_{t,p_t})'$ is a vector of parameters for model t. Let $d_n^2 = \min\{[\max_{1\le i\le n} ||x_i||^2]^{-1}, [\min_k 1/2\beta_{t,k}]^2\}$ and observe that $d_n^2 n/\log(n) \to \infty$ as $n \to \infty$. Below we state two propositions. The main idea here is to prove that under mild conditions on the design matrix certain properties of averaged deviance hold, which are weaker than its law of large numbers. However, these properties are sufficient for consistency of BIC and p-valued criteria which we prove in Theorem 2. Consider now two models w and c where the first model is a wrong one (i.e. it does not include at least one explanatory variable with corresponding coefficient not equal zero) and the second model is a correct model (although it is not necessarily the simplest one).

Proposition 1 *Under (A1), (A2) $P(D_{wc}^n \ge \alpha_1 n d_n^2) \to 1$, as $n \to \infty$, for some $\alpha_1 > 0$.*

Proposition 2 *Assume (F) and that for some $\varepsilon > 0$ and for some $\alpha > 0$ $\max_{1\le i\le n} ||\mathbf{x}_i||n^{-\varepsilon} \le \alpha$, as $n \to \infty$. Then $n^{-(1+\varepsilon)}D_{0c}^n = O_P(1)$ as $n \to \infty$.*

Note that the larger ε results in a weaker conclusion of the Proposition 2. In view of Remark 1, (A1) and (A2) imply assumptions of the 2 for $\varepsilon = 1/2$. Apart from the asymptotic results of LRT statistic the following approximation of $p(x|p_j)$ for $x \to \infty$ will be used. For $x > 0$ and $p \in \mathbf{N}$ define

$$C(x,p) = e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^{\frac{p}{2}-1}\left[\Gamma\left(\frac{p}{2}\right)\right]^{-1}$$

and

$$B(x,p) = C(x,p)\left[\frac{x}{x-(p-2)}\right].$$

Lemma 1 *If $Z \sim \chi_p^2$ then*

(i) for $p = 1$ and $x > 0$, $B(x,1) \le P(Z > x) \le C(x,1)$;
(ii) for $p > 1$ and $x > 0$, $C(x,p) \le P(Z > x)$, if $p > 1$ and $x > p-2$, $P(Z > x) \le B(x,p)$;
(iii) for $x \to \infty$ $P(Z > x) = C(x,p)[1 + O(x^{-1})]$.

The above Lemma is proved in [9].

Now we state consistency property of Bayesian Information Criterion and the introduced selectors M_m^n and M_M^n.

Theorem 2 *Under (A1), (A2) BIC , M_m^n and M_M^n are consistent i.e. $P(\hat{t} = t) \to 1$, as $n \to \infty$ when $\hat{t}$ denotes any one of these selectors.*

The strong consistency of Bayesian Information Criterion is proved in [11] where assumption (A1) is also imposed. Condition (C.2) in [11] after taking into account (A1) can be restated as $\max_{1\le i\le n} ||\mathbf{x}_i||^2 \log\log(n)/n \to 0$ i.e. it is slightly weaker than our condition (A2). However, we avoid assuming any extra conditions, in particular condition (C.5) in [11], a certain technical condition which seems hard to verify. From the main result there follows consistency of the greedy counterparts of the method.

Corollary 1 *Under (A1), (A2)* BIC_{IO}^n*,* $M_{m,IO}^n$ *and* $M_{M,IO}^n$ *are consistent.*

In order to explain the main lines of reasoning we prove Theorem 1 in the case of BIC and Corollary 1 here. More technically involved proofs of the remaining part of Theorem 1 as well as proofs of auxiliary results are relegated to the Appendix.

Proof of Theorem 2 (BIC case) Consider the case $j \supset t$ i.e. model t is a proper subset of a model j. We have to show that

$$P[-2l(\hat{\boldsymbol{\beta}}_t, \mathbf{Y}_n|\mathbf{X}_n) + p_t \log(n) < -2l(\hat{\boldsymbol{\beta}}_j, \mathbf{Y}_n|\mathbf{X}_n) + p_j \log(n)] \to 1,$$

as $n \to \infty$ which is equivalent to $P[D_{jt}^n > \log(n)(p_t - p_j)] \to 1$ as $n \to \infty$. The last convergence follows from the fact that $D_{jt}^n = O_P(1)$ which is implied by Theorem 1. The convergence for $j \not\supseteq t$ follows directly from Proposition 1 and assumption $nd_n^2/\log(n) \to \infty$.

Proof of Corollary 1 Let j_c be an index corresponding to the variable in t and j_w an index corresponding to the variable which is not in t. Note that

$$P[D_{(f-\{j_c\})f}^n \geq D_{(f-\{j_w\})f}^n] \to 1$$

as $n \to \infty$ which follows from the fact that by Proposition 1 $D_{(f-\{j_c\})f}^n \to \infty$ in probability and by Theorem 1 $D_{(f-\{j_w\})f}^n = O_P(1)$. This implies the convergence $P(t \in M_{IO}) \to 1$ which in conjunction with Theorem 2 yields the consistency of a respective two-step rule with an initial ordering.

4 Numerical Experiments

In this section the finite-sample performance of the discussed variable selection procedures is investigated. We considered Bayesian Information Criterion (BIC) and two scaled p-value criteria with scalings which performed well in the simulations, namely minimal p-value criterion with $a_n = \log(n)/2$ (mPVC2)and maximal p-value criterion with the same a_n (MPVC2). Every of the three pertaining criteria was considered in conjunction with any of four search methods resulting in twelve final methods. Our objective is to study the impact of both a criterion function and a search method on the probability of the minimal true model identification in the case when the number of possible variables M is large compared to the number p_t of the true ones. Let $\hat{t}$ be a model selected by the considered rule. As the measures of performance, besides $P(\hat{t} = t)$, we also consider positive selection rate (PSR) defined as $\mathbf{E}(p_{t\cap\hat{t}}/p_t)$ and and the false discovery rate (FDR) $\mathbf{E}(p_{\hat{t}\setminus t}/p_{\hat{t}})$. The last two measures are more appropriate when the probabilities of correct model selection are low (cf. model S3 below). The simulation experiments were carried out for $n = 100$ and repeated $N = 200$ times.

The following logistic regression models have been considered:

(S1) $t = 1$, $\beta_1 = 1$,

(S2) $t = (1, 2)$, $\boldsymbol{\beta} = (1, -1)'$,
(S3) $t = (1, 2, 3, 4)$, $\boldsymbol{\beta} = (0.75, 0.75, 1, 1.25)'$,
(S4) $t = (1, 2, 3)$, $\boldsymbol{\beta} = (1, 1, 1)'$.

The covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ were generated independently from the standard normal M-dimensional distribution and the binary outcome is drawn as Bernoulli r.v. with probability defined in (1). Results of our simulation study show that for the given criterion employed search method can affect considerably the probability of the true model selection. Moreover, the analogous results with variable M (cf. Figures 1 and 2 below) indicate that the differences between search methods become larger with increasing M. It is also interesting to note that the search method for which given criterion works the best depends on the criterion used, e.g. for BIC it is usually initial ordering method, whereas for methods based on p-values it is forward search. When for the given criterion and the model the best search method is chosen we see that MPVC2 criterion (with forward search) works better than BIC with any of the considered search methods in the case of model S1 and the same is true for mPVC2 criterion (with forward search) for models S2 and S4. The latter also behaves comparably to MPVC2 in the case of S1. We have shown in Figures 1 and 2 probabilities of correct identification as a function of horizon M for BIC and mPVC2 for models S1 and S4, whereas in Tables 1 and 2 indices for all the methods and $M = 30$ are given for models S2 and S3.

Note that in the case of the model S3, when $P(\hat{t} = t)$ is small overall and for a fixed search method, both FDR and PSR are larger for BIC than for p-based methods indicating that BIC has a tendency for choosing too large subset of variables, whereas p-value based methods choose a proper subset of true variables but rarely include superfluous ones. This is also true for other considered models. The first observation is concordant with [1] and [2]. We also noted (results not shown) that generating dependent predictors with covariance matrix $\Sigma = (\rho_{ij} = \rho^{|i-j|})$ may result in a change of optimal search method for a given criterion. For $\rho = 0.8$ IO is replaced by FS as the the best search method for BIC (actually, BIC with FS became the best method overall). Note also that the genetic algorithm worked uniformly worst for any of the criteria and model considered.

We also investigated in more detail usefulness of initial ordering as the search method. Figure 3 shows probabilities of correct ordering (i.e. true variables preceding superfluous ones) together with $P(t = \hat{t})$ as the function of M in model S1 with $\beta_1 = 1$ and $\beta_1 = 0.5$ for BIC and mPVC2. For mPVC2 they do not differ significantly indicating that the crucial problem is choice of a restricted family of models over which criterion is optimized. Discrepancy between $P(BIC^n = t)$ and probability of correct ordering is mainly due to choice of too large model.

Summarizing, mPVC2 method turns out to be a worthy competitor for BIC when used with an appropriate search method in the case when the number of potential predictors in logistic model is large. Performance of combined selection rule seems worth investigating. This is also confirmed by a real data example we considered. Namely, we investigated performance of BIC and mPVC2 with

IO and FS for `urine` data set ([10], $n = 77$) by the means of parametric bootstrap. Two variables (calcium and mmho having the smallest p-values in the full model) were chosen as predictors and logistic regression model was fitted with $\mathbf{Y}$ being occurrence of crystals in urine. The value of $\hat{\boldsymbol{\beta}}$ equals $(0.5725, -0.1186)'$. A parametric bootstrap (see e.g. [3]) was employed to check how the considered selection criteria perform for this data set. The true model was the fitted logistic model with the original two regressors, $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ from which 200 samples and additional superfluous explanatory variables were created in pairs by drawing from the two-dimensional normal distribution with independent components, which mean and variance vector matched that of the original predictors. We considered $k = 3, 7, \ldots, 19$ additional pairs what amounted to horizons $M = 8, 16, \ldots, 40$ when the true variables were accounted for. Figure 4 shows summary of the results. For both search methods mPVC2 performs considerably better, however in the case of IO it behaves much more stably when M increases.

	IO	BE	FS	GA
BIC	0.40	0.30	0.32	0.18
	(0.03)	(0.03)	(0.03)	(0.03)
mPVC2	0.62	0.68	0.70	0.46
	(0.03)	(0.03)	(0.03)	(0.03)
MPVC2	0.42	0.43	0.51	0.28
	(0.04)	(0.04)	(0.04)	(0.03)

(a) Fractions of correct model selection and their SEs.

		IO	BE	FS	GA
BIC	PSR	0.945	0.965	0.963	0.958
	FDR	0.261	0.332	0.285	0.407
mPVC2	PSR	0.875	0.917	0.905	0.917
	FDR	0.095	0.086	0.061	0.187
MPVC2	PSR	0.657	0.675	0.760	0.695
	FDR	0.164	0.152	0.019	0.271

(b) Positive selection rates and false discovery rates.

Table 1: Simulation results for model S2 with $M = 30$.

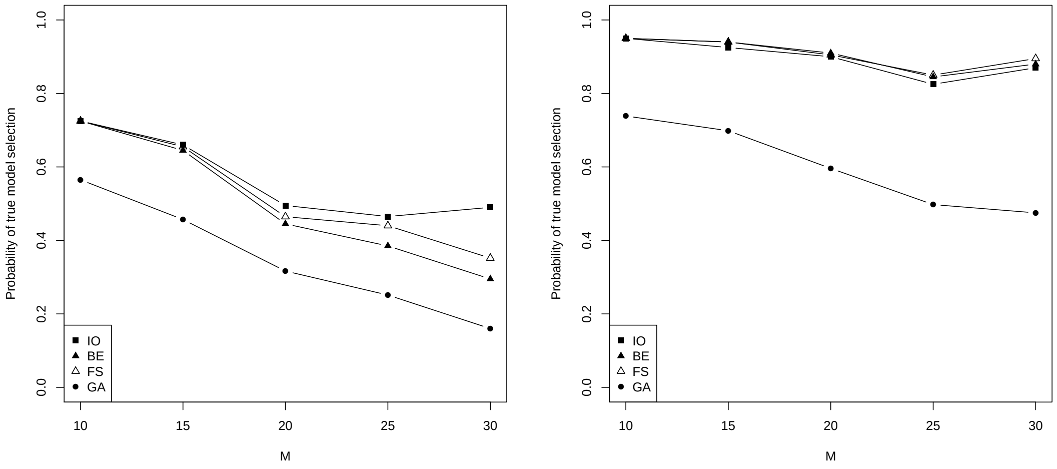
	IO	BE	FS	GA
BIC	0.12	0.14	0.19	0.12
	(0.02)	(0.02)	(0.03)	(0.02)
mPVC2	0.08	0.12	0.12	0.10
	(0.02)	(0.02)	(0.03)	(0.02)
MPVC2	0.04	0.02	0.07	0.04
	(0.01)	(0.01)	(0.02)	(0.01)

(a) Fractions of correct model selection and their SEs.

		IO	BE	FS	GA
BIC	PSR	0.814	0.858	0.845	0.843
	FDR	0.272	0.341	0.214	0.319
mPVC2	PSR	0.631	0.714	0.650	0.698
	FDR	0.102	0.199	0.065	0.204
MPVC2	PSR	0.497	0.545	0.540	0.544
	FDR	0.104	0.186	0.034	0.161

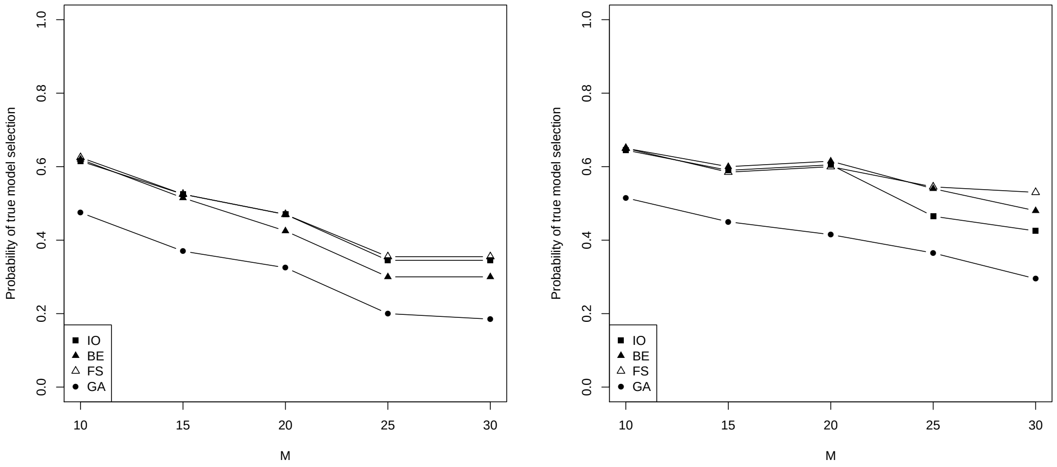
(b) Positive selection rates and false discovery rates.

Table 2: Simulation results for model S3 with $M = 30$.



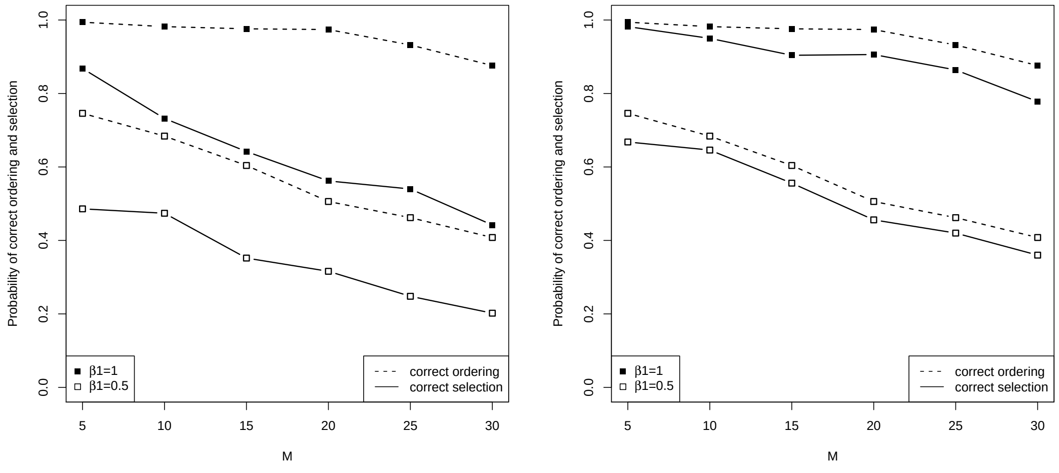
(*a*) (*b*)

Fig. 1: Estimated probabilities of correct model selection with respect to M for BIC (figure (a)) and mPVC2 (figure (b)) for model S1



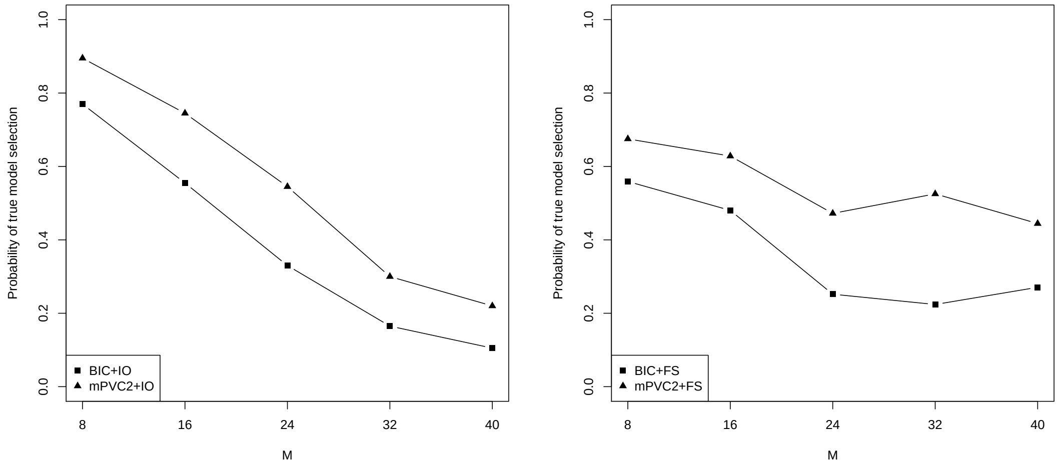
(*a*) (*b*)

Fig. 2: Estimated probabilities of correct model selection with respect to M for BIC (figure (a)) and mPVC2 (figure (b)) for model S4.



(*a*) (*b*)

Fig. 3: Estimated probabilities of correct model selection and correct ordering in IO method with respect to M for BIC (figure (a)) and mPVC2 (figure (b)) for model S1 with $\beta_1 = 1$ and $\beta_1 = 0.5$.



(*a*) (*b*)

Fig. 4: Estimated probabilities of correct model selection with respect to M for IO (figure (a)) and FS (figure (b)) for `urine` dataset.

Appendix

Proof of Theorem 2. (Consistency of M_m^n). Assume first that $t \neq 0$ and consider the case $j \not\supseteq t$. As the family of logistic models is finite it is sufficient to show that as $n \to \infty$

$$P[e^{p_t a_n} p(D_{0t}^n|p_t) \geq e^{p_j a_n} p(D_{0j}^n|p_j)] \to 0, \tag{4}$$

The probability in (4) does not decrease when D_{0j}^n is replaced by $\max(D_{0j}^n, 2)$ Thus for $p_j > 1$ in view of Lemma 1 (ii) the above probability is bounded from above by

$$P\{e^{p_t a_n} p(D_{0t}^n|p_t) \geq e^{p_j a_n} e^{-\max(D_{0j}^n,2)/2} \max(D_{0j}^n/2, 1)^{\frac{p_j}{2}-1} \mathit{\Gamma}^{-1}(\frac{p_j}{2})\}.$$

As $\max(D_{0j}^n/2, 1)^{\frac{p_j}{2}-1} \geq 1$ in view of Lemma 1 (iii) it suffices to show that

$$P\{e^{(D_{0t}^n - \max(D_{0j}^n,2))/2} \leq e^{(p_t-p_j)a_n} \mathit{\Gamma}(\frac{p_j}{2}) \mathit{\Gamma}^{-1}(\frac{p_t}{2})(D_{0t}^n/2)^{\frac{p_t}{2}-1}[1+O(1/D_{0t}^n)]\} \to 0,$$

as $n \to \infty$. The above convergence follows easily form Propositions 1 and 2. For $p_j = 1$ we apply part (i) of Lemma 1 together with

$$e^{-x/2} \left(\frac{x}{2}\right)^{-1/2} \left(\frac{x}{x+1}\right) \geq \frac{2}{3} e^{-x/2-\log(x/2)/2} \geq \frac{2}{3} e^{-x}.$$

for $x = \max(D_{0j}^n, 2) \geq 2$. For $p_j = 0$ the proof is similar. Consider now the case $j \supset t$. We have to show (4). Using Lemma 1 (iii) we obtain for $p_t \geq 1$

$$\begin{aligned}
&P\Big[\frac{1}{2} D_{jt}^n \leq \Big(\frac{p_t}{2} - 1\Big) \log\left(\frac{D_{0t}^n}{2}\right) - \Big(\frac{p_j}{2} - 1\Big) \log\left(\frac{D_{0j}^n}{2}\right) + \log \mathit{\Gamma}^{-1}\left(\frac{p_t}{2}\right) - \log \mathit{\Gamma}^{-1}\left(\frac{p_j}{2}\right) \\
&+ \log[1 + O(1/D_{0j}^n)] - \log[1 + O(1/D_{0t}^n)] + a_n(p_t - p_j)\Big] = \\
&P\Big[\frac{1}{2} D_{jt}^n \leq \Big(\frac{p_t}{2} - 1\Big) \frac{D_{jt}^n}{2} (D_{jt}^{n*})^{-1} - \Big(\frac{p_j}{2} - \frac{p_t}{2}\Big) \log\left(\frac{D_{0j}^n}{2}\right) + \log \mathit{\Gamma}^{-1}\left(\frac{p_t}{2}\right) - \log \mathit{\Gamma}^{-1}\left(\frac{p_j}{2}\right) \\
&+ \log[1 + O(1/D_{0j}^n)] - \log[1 + O(1/D_{0t}^n)] + a_n(p_t - p_j)\Big] \to 0,
\end{aligned}$$

where D_{jt}^{n*} belongs to the segment joining $D_{0j}^n/2$ and $D_{0t}^n/2$. The above convergence follows from $p_j > p_t$ and Theorem 1 and Proposition 1 which imply that $D_{jt}^n = O_P(1)$ and D_{0t}^n, $D_{0j}^n \to \infty$. For $p_t = 0$ we have to show that $P[e^{a_n}/\sqrt{n} > e^{p_j a_n} p(D_{0j}^n|p_j)] \to 0$, which follows from the fact that $p(D_{0j}^n|p_j) \xrightarrow{d} \mathcal{U}([0, 1])$ as $n \to \infty$ and $p_j \geq 1$. □

Consistency of M_M^n. Assume first that $t \neq f$ and consider the case $j \not\supseteq t$. We have to show that

$$P[e^{-p_t a_n} p(D_{tf}^n|M - p_t) \leq e^{-p_j a_n} p(D_{jf}^n|M - p_j)] \to 0, \tag{5}$$

as $n \to \infty$. It follows from Theorem 1 that $p(D_{tf}^n|M - p_t) \xrightarrow{d} \mathcal{U}([0, 1])$ and from Proposition 1 that $D_{jf}^n \to \infty$. Thus using Lemma 1 (iii) it sufficies to show

$$P\Big[\frac{1}{2} D_{jf}^n \leq \left(\frac{M - p_j}{2} - 1\right) \log\left(\frac{D_{jf}^n}{2}\right) + \log \mathit{\Gamma}^{-1}\left(\frac{M - p_j}{2}\right) + \log[1 + O(1/D_{jf}^n)] + a_n(p_t - p_j)\Big] \to 0,$$

as $n \to \infty$. The above convergence follows easily from Proposition 1.
Consider the case $j \supset t$. We have to show (5). For $j \neq f$ the desired convergence follows from Theorem 1 which implies that $p(D^n_{tf}|M - p_t)$, $p(D^n_{jf}|M - p_j) \xrightarrow{d} \mathcal{U}([0,1])$, as $n \to \infty$. For $j = f$ this is implied by $P(e^{-p_f a_n} < (e^{-p_t a_n} p(D_{tf}|M - p_t) \to 1$ which in its turn follows from $p_t < M$ and $a_n \to \infty$. For the case $t = f$ the proof is similar and uses the assumption that $a_n = O(\log n)$. □

Proof of Proposition 1 First we will show that under (A1) and (A2)

$$\mathbf{I}_n(\boldsymbol{\beta}) \geq \tau \mathbf{I}_n(\boldsymbol{\beta}_t) \tag{6}$$

for some positive constant τ and for $\boldsymbol{\beta} \in A_n = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_t|| \leq d_n\}$. Recall that $d_n^2 = \min\{[\max_{1\leq i\leq n} ||x_i||^2]^{-1}, [\min_k 1/2\beta_{t,k}]^2\}$. Using Cauchy-Schwarz inequality we have

$$\sup_{\boldsymbol{\beta}\in A_n} |\mathbf{x}_n'(\boldsymbol{\beta} - \boldsymbol{\beta}_t)| \leq \sup_{\boldsymbol{\beta}\in A_n} ||\mathbf{x}_n|| \cdot ||\boldsymbol{\beta} - \boldsymbol{\beta}_t|| \leq 1 \tag{7}$$

In order to prove (6) it suffices to show that there exists a positive constant $\tau > 0$ such that $\pi_i(\boldsymbol{\beta})(1 - \pi_i(\boldsymbol{\beta})) > \tau \pi_i(\boldsymbol{\beta}_t)(1 - \pi_i(\boldsymbol{\beta}_t))$, for all $i = 1, \ldots, n$ and $\boldsymbol{\beta} \in A_n$. This follows easily from (7) as it implies that it is enough to show that

$$\inf_i \inf_{\boldsymbol{\beta}\in A_n} \frac{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}_t}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}} > \inf_i \inf_{\boldsymbol{\beta}\in A_n} \max\left(\frac{1}{e^{-\mathbf{x}_i'\boldsymbol{\beta}_t} + e^{\mathbf{x}_i'(\boldsymbol{\beta}-\boldsymbol{\beta}_t)}}, \frac{e^{-\mathbf{x}_i'\boldsymbol{\beta}_t}}{e^{-\mathbf{x}_i'\boldsymbol{\beta}_t} + e^{\mathbf{x}_i'(\boldsymbol{\beta}-\boldsymbol{\beta}_t)}}\right) > 0,$$

which is easy to verify by the second application of (7).
The difference $l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}_n|\mathbf{X}_n) - l(\hat{\boldsymbol{\beta}}_w, \mathbf{Y}_n|\mathbf{X}_n)$ can be written as

$$[l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}_n|\mathbf{X}_n) - l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n)] + [l(\boldsymbol{\beta}_t, \mathbf{Y}_n|\mathbf{X}_n) - l(\hat{\boldsymbol{\beta}}_w|\mathbf{X}_n, \mathbf{Y}_n)]. \tag{8}$$

It can be shown using one term Taylor expansion, proof of Theorem 1 in [5] and condition (F) that the first term in (8) is $O_P(1)$. We omit the details. We will show that the probability that the second term is greater or equal $\alpha_1 n d_n^2$, for some $\alpha_1 > 0$ tends to 1. Define $H_n(\boldsymbol{\beta}) = l(\boldsymbol{\beta}_t, \mathbf{Y}_n|\mathbf{X}_n) - l(\boldsymbol{\beta}, \mathbf{Y}_n|\mathbf{X}_n)$. Note that $H(\boldsymbol{\beta})$ is convex and $H(\boldsymbol{\beta}_t) = 0$. For any incorrect model w we have $\hat{\boldsymbol{\beta}}_w \notin A_n$. Thus it suffices to show that $P(\inf_{\boldsymbol{\beta}\in\partial A_n} H_n(\boldsymbol{\beta}) < \alpha_1 n d_n^2) \to 0$, as $n \to \infty$, for some $\alpha_1 > 0$. Consider the following Taylor expansion

$$l(\boldsymbol{\beta}, \mathbf{Y}_n|\mathbf{X}_n) - l(\boldsymbol{\beta}_t, \mathbf{Y}_n|\mathbf{X}_n) = (\boldsymbol{\beta} - \boldsymbol{\beta}_t)' s_n(\boldsymbol{\beta}_t) - (\boldsymbol{\beta} - \boldsymbol{\beta}_t)' \mathbf{I}_n(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}_t)/2,$$

where $\tilde{\boldsymbol{\beta}}$ belongs to the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_t$. Note that $s_n(\boldsymbol{\beta}_t)$ is a random vector with zero mean and the covariance matrix $\mathbf{I}_n(\boldsymbol{\beta}_t)$.

Using the equality above, assumption (A1), (6) and Markov's inequality we have, taking $\alpha_1 < \gamma\tau$

$$P[\inf_{\boldsymbol{\beta}\in\partial A_n} H_n(\boldsymbol{\beta}) < \alpha_1 n d_n^2] \leq P[\sup_{\boldsymbol{\beta}\in\partial A_n} (\boldsymbol{\beta} - \boldsymbol{\beta}_t)' s_n(\boldsymbol{\beta}_t) \geq (\gamma\tau - \alpha_1) n d_n^2) =$$

$$P[||s_n(\boldsymbol{\beta}_t)||d_n \geq (\gamma\tau - \alpha_1) n d_n^2) \leq \frac{tr(I_n(\boldsymbol{\beta}_t)) d_n^2}{(\gamma\tau n d_n^2 - \alpha_1 n d_n^2)^2} \leq \frac{M\kappa n d_n^2}{(\gamma\tau n d_n^2 - \alpha_1 n d_n^2)^2} \to 0,$$

as $n \to \infty$. □

Proof of Proposition 2 Call $\max_{1\le i\le n} ||\mathbf{x}_i|| n^{-\varepsilon} \le \alpha$ assumption (A). We have the following decomposition

$$D_{0c}^n = l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}_n|\mathbf{X}_n) - l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n) + \\ l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n) - \mathbf{E}l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n) + \mathbf{E}l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n) - n\log(1/2). \qquad (9)$$

It was proved in the proof of Proposition 1 that under assumption (F) $l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}_n|\mathbf{X}_n) - l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n) = O_P(1)$ and thus $n^{-(1+\varepsilon)}[l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}_n|\mathbf{X}_n) - l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n)] \xrightarrow{P} 0$. We will show that

$$n^{-(1+\varepsilon)}[l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n) - \mathbf{E}l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n)] \xrightarrow{P} 0.$$

This follows from the Law of Large Numbers using Schwarz inequality since

$$Var[n^{-(1+\varepsilon)}l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n)] = n^{-2(1+\varepsilon)}Var(\sum_{i=1}^n y_i \mathbf{x}_i'\boldsymbol{\beta}_c) \le ||\boldsymbol{\beta}_c||^2 n^{-2(1+\varepsilon)} \sum_{i=1}^n ||\mathbf{x}_i||^2 \to 0$$

as $n \to \infty$ by assumption (A). In view of (9) it suffices to show that $|n^{-(1+\varepsilon)}\mathbf{E}l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n)| \le \alpha_2$, for $\alpha_2 > 0$. The following inequality holds

$$|n^{-(1+\varepsilon)}\mathbf{E}l(\boldsymbol{\beta}_c, \mathbf{Y}_n|\mathbf{X}_n)| \le n^{-(1+\varepsilon)}(\sum_{i=1}^n |\mathbf{x}_i'\boldsymbol{\beta}_c| + \sum_{i=1}^n \log(1+e^{\mathbf{x}_i'\boldsymbol{\beta}_c})). \qquad (10)$$

The first term in (10) is bounded in view of the Schwarz inequality and assumption (A). The following inequality holds

$$\log(1+x) \le 2\log(x)\mathbf{1}\{x>2\} + x\mathbf{1}\{x\le 2\} \le 2\log(x)\mathbf{1}\{x>2\} + 2. \qquad (11)$$

Using (11) and the Schwarz inequality the second term in (10) is bounded from above by $n^{-(1+\varepsilon)}[2\sum_{i=1}^n |\mathbf{x}_i'\boldsymbol{\beta}_c| + 2n]$ which is bounded by assumption (A). □

References

1. Broman K. W and Speed T. P.: A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J Roy Stat Soc B 64, 641–656, 731–775 (2002)
2. Chen J. and Chen Z.: Extended Bayesian criteria for model selection with large model spaces. Biometrika 95(3), 759–771 (1995)
3. Davison, A. and Hinkley, D.: Bootstrap Methods and Their Applications. Cambridge University Press (1997)
4. Fahrmeir, L.: Asymptotic testing theory for generalized linear models. Statistics 1, 65–76 (1987)
5. Fahrmeir, L. and Kaufmann, H.: Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics 1(13), 342–368 (1985)
6. Harrell, F.E.: Regression Modelling Strategies: with Applications to Linear Models, Logistic Regression and Survival Analysis, Springer, New York (2001)

7. Hastie, T. J. and Pregibon, D.: Generalized Linear Models. Wadsworth and Brooks/Cole (1992)
8. Mühlenbein, H. and Schlierkamp-Voosen, D.:Predictive Models for the Breeder Genetic Algorithm, I: Continuous Parameter Optimization. Evolutionary Computation 1(1), 25–49 (1993)
9. Pokarowski, P. and Mielniczuk, J.:P-values of likelihood ratio statistic for consistent model selection and testing. In preparation (2011)
10. SAS datasets, `http://ftp.sas.com/samples/A56902`
11. Qian, G. and Field, C.: Law of iterated logarithm and consistent model selection criterion in logistic regression. Statistics and Probability Letters 56, 101–112 (2002)
12. Tolvi, J.: Genetic algorithms for outlier detection and variable selection in linear regression models. Soft Comput. 8(8), 527-533 (2004)