What Do We Choose When We Err? Model Selection and Testing for Misspecified Logistic Regression Revisited

Jan Mielniczuk and Paweł Teisseyre

Abstract The problem of fitting logistic regression to binary model allowing for missppecification of the response function is reconsidered. We introduce two-stage procedure which consists first in ordering predictors with respect to deviances of the models with the predictor in question omitted and then choosing the minimizer of Generalized Information Criterion in the resulting nested family of models. This allows for large number of potential predictors to be considered in contrast to an exhaustive method. We prove that the procedure consistently chooses model t^* which is the closest in the averaged Kullback-Leibler sense to the true binary model t. We then consider interplay between t and t^* and prove that for monotone response function when there is genuine dependence of response on predictors, t^* is necessarily nonempty. This implies consistency of a deviance test of significance under misspecification. For a class of distributions of predictors, including normal family, Rudd's result asserts that $t^* = t$. Numerical experiments reveal that for normally distributed predictors probability of correct selection and power of deviance test depend monotonically on Rudd's proportionality constant η.

Keywords Incorrect model specification · Variable selection · Logistic regression

J. Mielniczuk (✉)
Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warsaw, Poland
e-mail: miel@ipipan.waw.pl

J. Mielniczuk · P. Teisseyre
Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5,
01-248 Warsaw, Poland
e-mail: teisseyrep@ipipan.waw.pl
url: http://www.ipipan.waw.pl

S. Matwin and J. Mielniczuk (eds.), *Challenges in Computational Statistics and Data Mining*, Studies in Computational Intelligence 605,
DOI 10.1007/978-3-319-18781-5_15

1 Introduction

We consider a general binary regression model in which responses $y \in \{0, 1\}$ are related to explanatory variables $\mathbf{x} = (1, x_1, \ldots, x_p)' \in R^{p+1}$ by the equation

$$P(y = 1|\mathbf{x}) = q(\mathbf{x}'\boldsymbol{\beta}), \tag{1}$$

where vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is an unknown vector of parameters and $q : R \to (0, 1)$ is a certain unknown response function. To the data pertaining to (1) we fit the logistic regression model i.e. we postulate that the posterior probability that $y = 1$ given $\mathbf{x}$ is of the form

$$p(\mathbf{x}'\boldsymbol{\gamma}) = \exp(\mathbf{x}'\boldsymbol{\gamma})/[1 + \exp(\mathbf{x}'\boldsymbol{\gamma})], \tag{2}$$

where $\boldsymbol{\gamma} \in R^{p+1}$ is a parameter. Our main interest here is the situation when the logistic model is misspecified i.e. $p \neq q$. Let $t = \{0\} \cup \{1 \leq k \leq p : \beta_k \neq 0\}$ be the true model i.e. consisting of indices of nonzero coefficients corresponding to true predictors and of the intercept denoted by 0. Our task may be either to identify model t when incorrectly specified model (2) is fitted or, less ambitiously, to verify whether t contains indices corresponding to predictors i.e. whether response depends on predictors at all. The situation of incorrect model specification is of importance because of obvious reasons as in real applications usually we have no prior knowledge about data generation process and, moreover, goodness-of-fit checks may yield inconclusive results. Thus investigating to what extent selection and testing procedures are resistant to response function misspecification is of interest. This is especially relevant with large number of possible features and sparsity when selecting true predictors is a challenge in itself and is further exacerbated by possible model misspecification. Moreover, some data generation mechanisms lead directly to misspecified logistic model. As an example we mention [6] who consider the case of logistic model when each response is mislabeled with a certain fixed probability.

In the paper we consider selection procedures specially designed for large p scenario which use Generalized Information Criterion (GIC). This criterion encompasses, for specific choices of parameters, such widely used criteria as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC is known to overestimate the dimension of the true model (see e.g. [4]) whereas BIC in the case of correctly specified linear model with fixed p is consistent [7]. There are many modifications of AIC and BIC which among others are motivated by the phenomenon that for large p depending on the sample size BIC also choses too large number of variables. We mention in particular modified BIC [3, 23], Extended BIC (EBIC) which consists in adding a term proportional to log p to BIC [8, 9] and Risk Inflation Criterion [15]. Qian and Field [20] consider GIC and proved its consistency under correct specification. In this line of research [9] propose minimization of EBIC over all possible subsets variables of sizes not larger than k when k is some sufficiently large number. However, this approach becomes computationally prohibitive for even moderate k. Other important approach is based on l_1-penalized loglikelihood and its extensions and modifications such as Elastic Net (see [24]) and SCAD [14]. It is known that l_1-penalization leads to cancelation of some coefficients and thus can be considered as model selection method. For discussion of other approaches we refer to [5, 10, 17] and references there.

The aims of the paper are twofold. We first introduce two-step modification of a procedure based on GIC, the minimizer of which over the family of all possible models is used as a selector of relevant variables. In the case when number of possible predictors is large such an approach is practically unfeasible due to high computational cost of calculating GIC for all possible subsets. This is a reason, likely the only one, why these methods are not frequently used and sequential greedy methods are applied in practice. However, greedy methods lack theoretical underpinning and it is known that they may miss true predictors. We thus propose a specific two-stage greedy method which consists in first ranking the predictors according to residual deviances of the models containing all variables but the considered one. Then in the second stage GIC is minimized over the nested family of models pertaining to increasing sets of the most important variables. We prove that such procedure picks with probability tending to 1 the logistic model t^* which minimizes averaged Kullback-Leibler distance from the binary model (1). This is to the best of our knowledge the first formal result on the consistency of greedy selection procedure for logistic regression even in the case when $p = q$. As a by-product we obtain the known result concerning behaviour of GIC optimized over the family of all models due to [22]. As in their paper the very general framework is considered for which stringent assumptions are needed we note that it is possible to prove the result under much weaker conditions (cf. their Proposition 4.2 (i), (ii) and Theorem 2 below). In view of the result the nature of the interplay between t^* and t becomes relevant. However, it seems that the problem, despite its importance, has failed to attract much attention. Addressing this question, admittedly partially, is the second aim of the paper. We discuss Rudd's (1983) result in this context which states that for certain distributions of predictors $\boldsymbol{\beta}^* = \eta\boldsymbol{\beta}$ for some $\eta \in R$, where $\boldsymbol{\beta}^*$ which minimizes averaged Kullback-Leibler distance from the binary model to logistic regressions. This obviously implies that $t^* = t$ if $\eta \neq 0$. As our main result in this direction we prove in Theorem 4 if t contains genuine regressors so does t^* provided that q is monotone and not constant. This implies in particular that in such a case significance test for regressors constructed under logistic model is consistent under misspecification. We also discuss the relevance of proved results in practice by investigating probability of correct model selection for two-stage procedure and power of test of significance for moderate sample sizes. In particular, we empirically verify that, surprisingly, misspecification of the model may lead to larger probabilities of correct selection and positive selection rate than for correct specification and stress the importance of the proportionality constant η in this context. Namely, it turns out that this phenomenon occurs mostly in the cases when $\eta > 1$. Moreover, we established that probability of correct selection and power of deviance test depend monotonically on η.

Generalization to the case when p is large in comparison to n is left for further study. As the fitting of the full model in the first stage of the procedure excludes its application when $p > n$ an initial screening of variables which is commonly done in applications (see e.g. [9]) would be necessary.

The paper is structured as follows. Section 2 contains preliminaries, in Sect. 3 we introduce and prove consistency of two-step greedy GIC procedure. Interplay between t and t^* is discussed in Sect. 4 together with its consequence for consistency of deviance test under misspecification. In Sect. 5 we describe our numerical experiments and Appendix contains proofs of auxiliary lemmas.

2 Preliminaries

Observe that the first coordinate of $\boldsymbol{\beta}$ in (1) corresponds to the intercept and remaining coefficients to genuine predictors which are assumed to be random variables. We assume that $\boldsymbol{\beta}$ is uniquely defined. The data consists of n observations $(y_i, \mathbf{x}_i)$ which are generated independently from distribution $P_{\mathbf{x},y}$ such that conditional distribution $P_{y|\mathbf{x}}$ is given by Eq. (1) and distribution of attribute vector $\mathbf{x}$ is $(p+1)$-dimensional with first coordinate equal to 1. We consider the case when $\mathbf{x}$ is random since in this situation behaviour of $\boldsymbol{\beta}^*$ of maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ for incorrect model specification can be more easily described (cf. definition (6) below, see however [13] for analogous development for deterministic predictors).

As a first remark note that as distribution $P_{\mathbf{x},y}$ which satisfies (1) with parameters q and $\boldsymbol{\beta}$ satisfies also (1) for parameters $\tilde{q}$ and $c\boldsymbol{\beta} + \alpha$ where $c > 0$ and $\tilde{q}(s) = q((s-\alpha)/c)$. It follows that when q is unknown only *the direction* of the vector $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)'$ may be possibly recovered.

Let $\mathbf{X}$ be $n \times (p+1)$ design matrix with rows $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{Y} = (y_1, \ldots, y_n)'$ be a response vector. Under the logistic regression model, the conditional log-likelihood function for the parameter $\boldsymbol{\gamma} \in R^{p+1}$ is

$$\begin{aligned} l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X}) &= \sum_{i=1}^{n} \{y_i \log[p(\mathbf{x}_i'\boldsymbol{\gamma})] + (1-y_i)\log[1-p(\mathbf{x}_i'\boldsymbol{\gamma})]\} \\ &= \sum_{i=1}^{n} \{y_i \mathbf{x}_i'\boldsymbol{\gamma} - \log[1+\exp(\mathbf{x}_i'\boldsymbol{\gamma})]\}. \end{aligned}$$

Note that we can alternatively view $l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})$ defined above as an empirical risk corresponding to the logistic loss. Define also the score function for the parameter $\boldsymbol{\gamma} \in R^{p+1}$

$$s_n(\boldsymbol{\gamma}) = \frac{\partial l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} [y_i - p(\mathbf{x}_i'\boldsymbol{\gamma})]\mathbf{x}_i = \mathbf{X}'(\mathbf{Y} - \mathbf{p}(\boldsymbol{\gamma})), \qquad (3)$$

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

$$J_n(\boldsymbol{\gamma}) = -\frac{\partial l^2(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} = \sum_{i=1}^{n}\{p(\mathbf{x}_i'\boldsymbol{\gamma})[1 - p(\mathbf{x}_i'\boldsymbol{\gamma})]\}\mathbf{x}_i\mathbf{x}_i' = \mathbf{X}'\Pi(\boldsymbol{\gamma})\mathbf{X}, \quad (4)$$

where $\Pi(\boldsymbol{\gamma}) = \text{diag}\{p(\mathbf{x}_1'\boldsymbol{\gamma})(1-p(\mathbf{x}_1'\boldsymbol{\gamma})), \ldots, p(\mathbf{x}_n'\boldsymbol{\gamma})(1-p(\mathbf{x}_n'\boldsymbol{\gamma}))\}$. Under assumption $\mathbf{E}(x_k^2) < \infty$, for $k = 1, \ldots, p$ it follows from the Law of Large Numbers that

$$n^{-1}J_n(\boldsymbol{\gamma}) \xrightarrow{P} \mathbf{E}_{\mathbf{x}}\{\mathbf{x}\mathbf{x}'p(\mathbf{x}'\boldsymbol{\gamma})[1 - p(\mathbf{x}'\boldsymbol{\gamma})]\} =: J(\boldsymbol{\gamma}). \quad (5)$$

Observe that in the case of incorrect model specification $\text{cov}[s_n(\boldsymbol{\gamma})|\mathbf{x}_1, \ldots, \mathbf{x}_n] = \sum_{i=1}^{n}\{q(\mathbf{x}_i'\boldsymbol{\gamma})[1 - q(\mathbf{x}_i'\boldsymbol{\gamma})]\}\mathbf{x}_i\mathbf{x}_i'$ is not equal to negative Hessian $J_n(\boldsymbol{\gamma})$ as in the case of correct model specification when $p(\cdot) = q(\cdot)$.

The maximum likelihood estimator (ML) $\hat{\boldsymbol{\beta}}$ of parameter $\boldsymbol{\beta}$ is defined to be

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\gamma}\in R^{p+1}} l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X}).$$

Moreover define

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\gamma}\in R^{p+1}} E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})]\},$$

where

$$\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})] = q(\mathbf{x}'\boldsymbol{\beta})\log\frac{q(\mathbf{x}'\boldsymbol{\beta})}{p(\mathbf{x}'\boldsymbol{\gamma})} + [1 - q(\mathbf{x}'\boldsymbol{\beta})]\log\frac{1 - q(\mathbf{x}'\boldsymbol{\beta})}{1 - p(\mathbf{x}'\boldsymbol{\gamma})}$$

is the Kulback-Leibler distance from the true Bernoulli distribution with the parameter $q(\mathbf{x}'\boldsymbol{\beta})$ to the postulated one with the parameter $p(\mathbf{x}'\boldsymbol{\gamma})$. Thus $\boldsymbol{\beta}^*$ is the parameter corresponding to the logistic model closest to binary model with respect to Kullback-Leibler divergence. It follows from [16] that

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}^* \quad (6)$$

Using the fact that $\partial p(\mathbf{x}'\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = p(\mathbf{x}'\boldsymbol{\gamma})[1 - p(\mathbf{x}'\boldsymbol{\gamma})]\mathbf{x}$ it is easy to see that

$$\mathbf{E}\left[\frac{\partial\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}}\right] = \mathbf{E}[-q(\mathbf{x}'\boldsymbol{\beta})\mathbf{x} + p(\mathbf{x}'\boldsymbol{\gamma})\mathbf{x}]$$

and

$$\mathbf{E}\left[\frac{\partial^2\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\gamma})]}{\partial\boldsymbol{\gamma}\boldsymbol{\gamma}'}\right] = \mathbf{E}\{p(\mathbf{x}'\boldsymbol{\gamma})[1 - p(\mathbf{x}'\boldsymbol{\gamma})]\mathbf{x}\mathbf{x}'\}$$

is positive-semidefinite. Thus from the first of the above equations we have

$$\mathbf{E}[q(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}] = \mathbf{E}[p(\mathbf{x}'\boldsymbol{\beta}^*)\mathbf{x}] = \mathbf{E}(y\mathbf{x}). \tag{7}$$

Note that as the first coordinate of $\mathbf{x}$ is equal one which corresponds to intercept, the pertaining equation is

$$\mathbf{E}[q(\mathbf{x}'\boldsymbol{\beta})] = \mathbf{E}[p(\mathbf{x}'\boldsymbol{\beta}^*)] = \mathbf{E}(y). \tag{8}$$

Using (3) and (7) we obtain

$$\begin{aligned} \mathrm{cov}\{\mathbf{E}[s_n(\boldsymbol{\beta}^*)|\mathbf{x}_1, \dots \mathbf{x}_n]\} &= n\mathbf{E}\{\mathbf{x}\mathbf{x}'[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]^2\} \\ &\quad - n\mathbf{E}\{\mathbf{x}[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]\}\{E\{\mathbf{x}[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]\}\}' \\ &= n\mathbf{E}\{\mathbf{x}\mathbf{x}'[q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]^2\}. \end{aligned}$$

We also have

$$E\{\mathrm{cov}[s_n(\boldsymbol{\beta}^*)|\mathbf{x}_1, \dots, \mathbf{x}_n]\} = n\mathbf{E}\{\mathbf{x}\mathbf{x}'q(\mathbf{x}'\boldsymbol{\beta})[1 - q(\mathbf{x}'\boldsymbol{\beta})]\}.$$

Let $K_n(\boldsymbol{\gamma}) = \mathrm{cov}[s_n(\boldsymbol{\gamma})]$ be covariance matrix of score function $s_n(\boldsymbol{\gamma})$. From above facts we have

$$\begin{aligned} &n^{-1}K_n(\boldsymbol{\beta}^*) \\ &\quad = \mathbf{E}\left\{\mathbf{x}\mathbf{x}'\{q(\mathbf{x}'\boldsymbol{\beta})[1 - q(\mathbf{x}'\boldsymbol{\beta})] + [q(\mathbf{x}'\boldsymbol{\beta}) - p(\mathbf{x}'\boldsymbol{\beta}^*)]^2\}\right\} =: K(\boldsymbol{\beta}^*). \end{aligned} \tag{9}$$

The form of $K_n(\boldsymbol{\beta}^*)$ will be used in the proof of Lemma 2. From (6) it is also easy to see that

$$\boldsymbol{\beta}^* = \arg \min_{\boldsymbol{\gamma} \in R^{p+1}} E\{-l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})\}.$$

It follows from [19] that $\boldsymbol{\beta}^*$ exists provided $0 < q(\boldsymbol{\beta}'x) < 1$ almost everywhere with respect to $P_{\mathbf{x}}$ and is unique provided $E||\mathbf{x}|| < \infty$. In the following we will always assume that $\boldsymbol{\beta}^*$ exists and is unique. In the case of correct specification, when $p(\cdot) = q(\cdot)$ we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}$. In general $\boldsymbol{\beta}^*$ may be different from $\boldsymbol{\beta}$. The most immediate example is when $q(s) = p(-s)$ which corresponds to logistic model with switched classes. In this case $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$. Li and Duan [19], p. 1019 give an example when supports of β and β^* are disjoint for a loss different than logistic. Let $t^* = \{0\} \cup \{1 \le k \le p : \beta_k^* \ne 0\}$. In Sect. 4 we discuss the relationships between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ as well as between t and t^* in more detail. In Sect. 3 we give conditions under which set t^* is identified consistently. Under certain assumptions we can also have $t^* = t$ and thus identification of set t is possible.

Let us discuss the notation used in this paper. Let $m \subseteq f := \{0, 1, \dots, p\}$ be any subset of variable indices and $|m|$ be its cardinality. Each subset m is associated with a model with explanatory variables corresponding to this subset. In the following f stands for the full model containing all available variables and by $null$ we denote model containing only intercept (indexed by 0). We denote by $\hat{\boldsymbol{\beta}}_m$ a maximum likelihood estimator calculated for model m and by $\boldsymbol{\beta}^*_m$ the minimizer of averaged Kullback-Leibler divergence when only predictors belonging to m are considered. Thus $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*_f$. Moreover, $\boldsymbol{\beta}^*(m)$ stands for $\boldsymbol{\beta}^*$ restricted to m. Depending on the context these vectors will be considered as $|m|$-dimensional or as their $(p+1)$-dimensional versions augmented by zeros. We need the following fact stating that when $m \supseteq t^*$ then $\boldsymbol{\beta}^*_m$ is obtained by restricting $\boldsymbol{\beta}^*$ to m.

Lemma 1 *Let $m \supseteq t^*$ and assume $\boldsymbol{\beta}^*$ is unique. Then $\boldsymbol{\beta}^*_m = \boldsymbol{\beta}^*(m)$.*

Proof The following inequalities hold

$$\begin{aligned} E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'_m\boldsymbol{\beta}^*_m)]\} &\geq E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'\boldsymbol{\beta}^*)]\} \\ &= E\{\Delta_{\mathbf{x}}[q(\mathbf{x}'\boldsymbol{\beta}), p(\mathbf{x}'_m\boldsymbol{\beta}^*(m))]\}. \end{aligned}$$

From the definition of projection the above inequality is actually equality and from the uniqueness the assertion follows.

3 Consistency of Two-Step Greedy GIC Procedure

We consider the following model selection criterion

$$GIC(m) = -2l(\hat{\boldsymbol{\beta}}_m, \mathbf{Y}|\mathbf{X}) + a_n|m|,$$

where m is a given submodel containing $|m|$ variables, $\hat{\boldsymbol{\beta}}_m$ is a maximum likelihood estimator calculated for model m (augmented by zeros to p-dimensional vector) and a_n is penalty. Observe that $a_n = \log(n)$ corresponds to Bayesian Information Criterion and $a_n = 2$ corresponds to Akaike Information Criterion. GIC was considered e.g. by [22]. We would like to select a model which minimizes GIC over a family

$$\mathcal{M} := \{\{0\} \cup s : s \subseteq \{1, \ldots, p\}\},$$

i.e. the family of all submodels of f containing intercept. Denote the corresponding selector by $\hat{t}^*$. As $\mathcal{M}$ consists of 2^p models and determination of $\hat{t}^*$ requires calculation of GIC for all of them this becomes computationally unfeasible for large p. In order to restrict the space of models over which the optimal value of criterion function is sought we propose the following two-stage procedure.

Step 1. The covariates $\{1, \ldots, p\}$ are ordered with respect to the residual deviances

$$D_{f\setminus\{i_1\}f} \geq D_{f\setminus\{i_2\}f} \geq \cdots \geq D_{f\setminus\{i_p\}f}.$$

Step 2. The considered model selection criterion GIC is minimized over a family

$$\mathcal{M}_{\text{nested}} := \{\{0\}, \{0\} \cup \{i_1\}, \{0\} \cup \{i_1, i_2\}, \ldots, \{0\} \cup \{i_1, i_2, \ldots, i_p\}\}.$$

We define $\hat{t}^*_{gr}$ as the minimizer of GIC over $\mathcal{M}_{\text{nested}}$. The intuition behind the first step of the procedure is that by omitting the true regressors from the model their corresponding residual deviances are increased significantly more than when spurious ones are omitted. Thus the first step may be considered as screening of the family $\mathcal{M}$ and reducing it to $\mathcal{M}_{\text{nested}}$ by whittling away elements likely to be redundant.

The following assumption will be imposed on $P_{\mathbf{x}}$ and penalization constants a_n

(A1) $J(\boldsymbol{\beta}^*)$ is positive definite matrix.
(A2) $E(x_k^2) < \infty$, for $k = 1, \ldots, p$.
(A3) $a_n \to \infty$ and a_n/n is nonincreasing and tends to 0 as $n \to \infty$.

The main result of this section is the consistency of the greedy procedure defined above.

Theorem 1 *Under assumptions (A1)–(A3) greedy selector $\hat{t}^*_{gr}$ is consistent i.e. $P(\hat{t}^*_{gr} = t^*) \to 1$ when $n \to \infty$.*

The following two results which are of independent interest constitute the proof of Theorem 1. The first result asserts consistency of $\hat{t}^*$. This is conclusion of Proposition 4.2 (i) and (iii) in [22]. However, as the framework in the last paper is very general, it is possible to prove the assertions there under much milder assumptions without assuming e.g. that loglikelihood satisfies weak law of large numbers uniformly in β and similar assumption on J_n. Theorem 3 states that after performing the first step of the procedure relevant regressors will precede the spurious ones with probability tending to 1. Consistency of GIC in the almost sure sense was proved by [20] for deterministic regressors under some extra conditions.

Theorem 2 *Assume (A1)–(A3). Then $\hat{t}^*$ is consistent i.e.*

$$P(\hat{t}^* = t^*) = P[\min_{m \in \mathcal{M}, m \neq t^*} GIC(m) > GIC(t^*)] \to 1.$$

Consider two models j and k and denote by

$$D^n_{jk} = 2[l(\hat{\boldsymbol{\beta}}_k, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_j, \mathbf{Y}|\mathbf{X})] \tag{10}$$

deviance of the model k from the model j.

Theorem 3 *Assume conditions (A1)–(A2). Then for all $i \in t^* \setminus \{0\}$ and $j \notin t^* \setminus \{0\}$ we have*

$$P[D_{f\setminus\{i\}f} > D_{f\setminus\{j\}f}] \to 1, \text{ as } n \to \infty.$$

Proof (Theorem 1) As the number of predictors is finite and does not depend on n the assertion in Theorem 3 implies that with probability tending to one model t^* will be included in $\mathcal{M}_{\text{nested}}$. This in view of Theorem 2 yields the proof of Theorem 1.

The following lemmas will be used to prove Theorem 2. Define sequence

$$d_n^2 = \min\{[\max_{1\le i\le n} ||\mathbf{x}_i||^2]^{-1}, [\min_{k\in t^*, 1\le k\le p} (1/2)\beta_k^*]^2\}. \tag{11}$$

Remark 1 It follows from Lemma 6 that under assumptions (A2) and (A3) if $t^* \setminus 0 \neq \emptyset$ we have $nd_n^2/a_n \xrightarrow{P} \infty$.

Two lemmas below are pivotal in proving Theorem 2. The proofs are in the appendix.

Lemma 2 *Let $c \supseteq m \supseteq t^*$. Assume (A1)–(A2). Then $D_{mc} = O_P(1)$.*

Lemma 3 *Let $w \not\supseteq t^*$ and $c \supseteq t^*$. Assume (A1)–(A2). Then $P(D_{wc} > \alpha_1 nd_n^2) \to 1$ as $n \to \infty$, for some $\alpha_1 > 0$.*

Proof (Theorem 3) It follows from Lemma 3 that for $i \in t$ we have $P[D^n_{f\setminus\{i\}f} > \alpha_1 nd_n^2] \to 1$, for $\alpha_1 > 0$ and by Remark 1 $nd_n^2 \xrightarrow{P} \infty$. By Lemma 2 we have that $D_{f\setminus\{j\}f} = O_P(1)$ for $j \in t^*$, which end the proof.

Proof (Theorem 2) Consider first the case $t^* = \{0\} \cup m$, $m \neq \emptyset$. We have to show that for all models $m \in \mathcal{M}$ such that $m \neq t^*$

$$P[-2l(\hat{\boldsymbol{\beta}}_{t^*}, \mathbf{Y}|\mathbf{X}) + |t^*|a_n < -2l(\hat{\boldsymbol{\beta}}_m, \mathbf{Y}|\mathbf{X}) + |m|a_n] \to 1,$$

as $n \to \infty$ which is equivalent to $P[D_{mt^*} > a_n(|t^*| - |m|)] \to 1$. In the case of $m \not\supseteq t^*$ this follows directly from Lemma 3 and $nd_n^2/a_n \xrightarrow{P} \infty$. Consider the case of $m \supset t^*$. By Lemma 2 $D_{mt^*} = O_P(1)$. This ends the first part of the proof in view of $a_n(|t^*| - |m|) \to -\infty$. For $t^* = \{0\}$ we only consider the case $m \supset t^*$ and the assertion $P[D_{mt^*} > a_n(1 - |m|)] \to 1$ follows again from Lemma 2.

4 Interplay Between *t* and *t**

In view of the results of the previous section t^* can be consistently selected by two-step GIC procedure. As we want to choose t not t^*, the problem what is the connection between these two sets naturally arises. First we study the problem whether it is possible that t^* is $\{0\}$ whereas t does contain genuine regressors. Fortunately, the answer under some mild conditions on the distribution $P_{\mathbf{x},y}$, including monotonicity of response function q, is negative. We proceed by reexpressing the fact that $t^* = \{0\}$ in terms of conditional expectations and then showing that the obtained condition for monotone q can be satisfied only in the case when y and $\mathbf{x}$ are independent.

Let $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)$, $\tilde{\boldsymbol{\beta}}^* = (\beta_1^*, \ldots, \beta_p^*)$ and $\tilde{\mathbf{x}} = (x_1, \ldots, x_p)$. The first proposition (proved in the appendix) gives the simple equivalent condition for $t^* = \{0\}$.

Proposition 1 $E(\mathbf{x}|y=1) = E(\mathbf{x}|y=0)$ *if and only* $t^* = \{0\}$.

Let $f(\tilde{\mathbf{x}}|y=1)$ and $f(\tilde{\mathbf{x}}|y=0)$ be the density functions of $\tilde{\mathbf{x}}$ in classes $y=1$ and $y=0$, respectively and denote by $F(\tilde{\mathbf{x}}|y=1)$ and $F(\tilde{\mathbf{x}}|y=0)$ the corresponding probability distribution functions. Note that the above proposition in particular implies that in the logistic model for which expectations of $\mathbf{x}$ in both classes are equal we necessarily have $\tilde{\boldsymbol{\beta}} = 0$. The second proposition asserts that this is true for a general binary model under mild conditions. Thus in view of the last proposition under these conditions $t^* = \{0\}$ is equivalent to $t = \{0\}$.

Proposition 2 *Assume that q is monotone and densities $f(\tilde{\mathbf{x}}|y=1)$, $f(\tilde{\mathbf{x}}|y=0)$ exist. Then $E(\tilde{\mathbf{x}}|y=1) = E(\tilde{\mathbf{x}}|y=0)$ implies $f(\tilde{\mathbf{x}}|y=1) = f(\tilde{\mathbf{x}}|y=0)$ a.e., i.e. y and $\tilde{\mathbf{x}}$ are independent.*

Proof Define $h(\tilde{\mathbf{x}})$ as the density ratio of $f(\tilde{\mathbf{x}}|y=1)$ and $f(\tilde{\mathbf{x}}|y=0)$. Observe that as

$$h(\tilde{\mathbf{x}}) = \frac{f(\tilde{\mathbf{x}}|y=1)}{f(\tilde{\mathbf{x}}|y=0)} = \frac{P(y=0)}{P(y=1)} \frac{q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})}{1 - q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})} \tag{12}$$

we have that $h(\tilde{\mathbf{x}}) = w(\tilde{\mathbf{x}}'\tilde{\beta})$ and w is monotone.

Consider first the case $p = 1$. It follows from the monotone likelihood ratio property (see [18], Lemma 2, Sect. 3) that since $h(\tilde{\mathbf{x}})$ is monotone then conditional distributions $F(\tilde{\mathbf{x}}|y=1)$ and $F(\tilde{\mathbf{x}}|y=0)$ are ordered and as their expectations are equal this implies $F(\tilde{\mathbf{x}}|y=1) = F(\tilde{\mathbf{x}}|y=0)$ and thus the conclusion for $p = 1$.

For $p > 1$ assume without loss of generality that $\beta_1 \neq 0$ and consider the transformation $\mathbf{z} = (z_1, \ldots, z_p) = (\tilde{\boldsymbol{\beta}}'\tilde{\mathbf{x}}, x_2, \ldots, x_p)'$. Denote by $\tilde{f}(\mathbf{z}|y=1)$ and $\tilde{f}(\mathbf{z}|y=0)$ densities of $\mathbf{z}$ in both classes. It is easy to see that we have

$$\tilde{f}(\mathbf{z}|y=1) = \beta_1^{-1} f\left((z_1 - \beta_2 z_2 - \cdots - \beta_p z_p)/\beta_1, z_2, \ldots, z_p \middle| y=1\right),$$

$$\tilde{f}(\mathbf{z}|y=0) = \beta_1^{-1} f\left((z_1 - \beta_2 z_2 - \cdots - \beta_p z_p)/\beta_1, z_2, \ldots, z_p \middle| y=0\right)$$

and

$$\frac{\tilde{f}(\mathbf{z}|y=1)}{\tilde{f}(\mathbf{z}|y=0)} = w\left(\tilde{\boldsymbol{\beta}}'((z_1 - \beta_2 z_2, \ldots, \beta_p z_p)/\beta_1, z_2, \ldots, z_p)\right) = w(z_1). \tag{13}$$

It follows from (13) that marginal densities $\tilde{f}_1(z_1|y=1)$, $\tilde{f}_1(z_1|y=0)$ satisfy $\tilde{f}_1(z_1|y=1)/\tilde{f}_1(z_1|y=0) = w(z_1)$ and the first part of the proof yields $\tilde{f}_1(z_1|y=1) = \tilde{f}_1(z_1|y=0)$.

Thus we have for fixed z_1

$$\frac{\tilde{f}(\mathbf{z}|y=1)}{\tilde{f}(\mathbf{z}|y=0)} = \frac{\tilde{f}(z_2,\ldots,z_p|z_1,y=1)\tilde{f}_1(z_1|y=1)}{\tilde{f}(z_2,\ldots,z_p|z_1,y=0)\tilde{f}_1(z_1|y=0)}$$
$$= \frac{\tilde{f}(z_2,\ldots,z_p|z_1,y=1)}{\tilde{f}(z_2,\ldots,z_p|z_1,y=0)} = w(z_1),$$

which implies that for any z_1 we have $\tilde{f}(z_2,\ldots,z_p|z_1,\mathrm{y}=1) = \tilde{f}(z_2,\ldots,z_p|z_1, y=0)$ and thus $\tilde{f}(\mathbf{z}|y=1) = \tilde{f}(\mathbf{z}|y=0)$ and consequently $f(\tilde{\mathbf{x}}|y=1) = f(\tilde{\mathbf{x}}|y=0)$ which ends the proof.

Observe now that in view of (12) if $f(\tilde{\mathbf{x}}|y=1) = f(\tilde{\mathbf{x}}|y=0)$ then $q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})$ is constant and thus $\tilde{\boldsymbol{\beta}} = 0$ if $1, x_1, \ldots, x_p$ are linearly independent with probability 1 i.e. $\mathbf{x}'\mathbf{b} = b_0$ a.e. implies that $\mathbf{b} = 0$ (or equivalently that $\Sigma_{\mathbf{x}} > 0$). Thus we obtain

Theorem 4 *If q is monotone and not constant and $1, x_1, \ldots, x_p$ are linearly independent with probability 1 then $t^* = \{0\}$ is equivalent to $t = \{0\}$ or, $\tilde{\boldsymbol{\beta}}^* \neq 0$ is equivalent to $\tilde{\boldsymbol{\beta}} \neq \mathbf{0}$.*

Now we address the question when $t = t^*$. The following theorem has been proved in [21], see also [19] for a simple proof based on generalized Jensen inequality.

Theorem 5 *Assume that $\boldsymbol{\beta}^*$ is uniquely defined and there exist $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in R^p$ such that*

(R) $E(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}'\boldsymbol{\beta} = z) = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 z.$

Then $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$, for some $\eta \in R$.

It is well known that Rudd's condition (R) is satisfied for eliptically contoured distributions. In particular multivariate normal distribution satisfies this property (see e.g. [19], Remark 2.2). The case when $\eta \neq 0$ plays an important role as it follows from the assertion of Theorem 5 that then $t^* = t$. Note that in many statistical problems we want to consistently estimate the direction of vector $\boldsymbol{\beta}$ and not its length. This is true for many classification methods when we look for direction such that projection on this direction will give maximal separation of classes. Theorem 4 implies that under its conditions η in the assertion of Theorem 5 is not equal zero. Thus we can state

Corollary 1 *Assume (A1)–(A3), (R) and conditions of Theorem 4. Then*

$$P(\hat{t}^*_{gr} = t) \to 1$$

i.e. two-stage greedy GIC is consistent for t.

Proof Under (R) it follows from Theorem 5 that $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$ and as q is monotone and not constant it follows from Theorem 4 that $\eta \neq 0$ and thus $t = t^*$. This implies the assertion in view of Theorem 2.

In the next section by means of numerical experiments we will indicate that magnitude of η plays an important role for probability of correct selection. In particular we will present examples showing that when regressors are jointly normal and thus Ruud's condition is satisfied, probability of correct selection of t by two-step greedy GIC can be significantly larger under misspecification than under correct specification.

The analogous result to Corollary 1 follows for $\hat{t}^*$ when GIC is minimized over the whole family of 2^p models.

The important consequence of Theorem 4 is that power of significance test will increase to 1 when there is dependence of y on $\mathbf{x}$ even when logistic model is misspecified and critical region is constructed for such model. Namely, consider significance test for $H_0 : \tilde{\boldsymbol{\beta}} = 0$ with critical region

$$\mathcal{C}_{1-\alpha} = \{D_{null,\hat{t}^*_{gr}} > \chi^2_{|\hat{t}^*_{gr}|-1,1-\alpha}\} \tag{14}$$

where $\chi^2_{k,1-\alpha}$ is quantile of order $1-\alpha$ of chi-squared distribution with k degrees of freedom. Observe that if $p = q$ it follows from Theorem 2 and [12] that under null hypothesis $P(C_{1-\alpha}|H_0) \to \alpha$ what explains the exact form of the threshold of the rejection region when the logistic model is fitted. We have

Corollary 2 *Assume that conditions of Theorem 4 are satisfied and* $\tilde{\boldsymbol{\beta}} \neq 0$. *Consider test of* $H_0 : \tilde{\boldsymbol{\beta}} = \mathbf{0}$ *against* $H_1 : \tilde{\boldsymbol{\beta}} \neq \mathbf{0}$ *with critical region* $\mathcal{C}_{1-\alpha}$ *defined in (14). Then the test is consistent i.e.* $P(D_{null,\hat{t}^*_{gr}} \in C_{1-\alpha}|H_1) \to 1$.

Observe that if $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}$ then in view of Remark 1 $nd_n^2 \to \infty$. Then the main results and Lemma 3 imply that when $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}$ $P[D_{null,\hat{t}^*_{gr}} > \chi^2_{|\hat{t}^*_{gr}|-1,1-\alpha}] \to 1$ for any $\alpha > 0$ and the test is consistent. But in view of Theorem 4 $\tilde{\boldsymbol{\beta}}^* \neq \mathbf{0}$ is implied by $\tilde{\boldsymbol{\beta}} \neq 0$.

5 Numerical Experiments

In this section we study how the incorrect model specification affects the model selection and testing procedures, in particular how it influences probability of correct model selection, positive selection rate, false discovery rate and power of a test of significance. In the case when attributes are normally distributed we investigate how these measures depend on proportionality constant η appearing in Rudd's theorem.

Recall that t denotes the minimal true model. Convention that $\boldsymbol{\beta}_t$ is subvector of $\boldsymbol{\beta}$ corresponding to t is used throughout. We consider the following list of models.

(M1) $t = \{10\}$, $\beta_t = 0.2$,
(M2) $t = \{2, 4, 5\}$, $\boldsymbol{\beta}_t = (1, 1, 1)'$,
(M3) $t = \{1, 2\}$, $\boldsymbol{\beta}_t = (0.5, 0.7)'$,
(M4) $t = \{1, 2\}$, $\boldsymbol{\beta}_t = (0.3, 0.5)'$,
(M5) $t = \{1, \ldots, 8\}$, $\boldsymbol{\beta}_t = (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)'$.

Models (M3)–(M5) above are considered in [9]. The number of all potential attributes is initially set to be $p = 15$ so the proportion of relevant variables varies from 6.66 % (for model M1) to 53.33 % (for model M5). Recall that $q(\cdot)$ denotes a true response function, i.e. for a given $\mathbf{x}$, y is generated from Bernoulli distribution with success probability $q(\mathbf{x}'\boldsymbol{\beta})$. The logistic model defined in (2) is fitted. Let $F_{N(0,1)}(\cdot)$ denote distribution function of standard normal random variable and $F_{Cauchy(u,v)}(\cdot)$ distribution function of Cauchy distribution with location u and scale v. In the case of incorrect model specification, the following response functions are considered:

$$q_1(s) = F_{N(0,1)}(s) \quad \text{(Probit model)},$$

$$q_2(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (0.1, 0.8) \\ 0.1 & \text{for } F_{N(0,1)}(s) \leq 0.1 \\ 0.8 & \text{for } F_{N(0,1)}(s) \geq 0.8, \end{cases}$$

$$q_3(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (0.2, 0.7) \\ 0.2 & \text{for } F_{N(0,1)}(s) \leq 0.2 \\ 0.7 & \text{for } F_{N(0,1)}(s) \geq 0.7, \end{cases}$$

$$q_4(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } |s| > 1 \\ 0.5 + 0.5\cos[4\pi F_{N(0,1)}(s)]F_{N(0,1)}(s) & \text{for } |s| \leq 1, \end{cases}$$

$$q_5(s) = F_{Cauchy(0,1)}(s),$$

$$q_6(s) = F_{Cauchy(0,2)}(s),$$

Studied response functions are shown in Fig. 1. Dashed line there corresponds to fitted logistic response function $p(\cdot)$.

We consider two distributions of attributes, in both cases attributes are assumed to be independent. In the first scenario x_j have $N(0, 1)$ distribution and in the second x_j are generated from Gaussian mixture $0.95N(0, 1) + 0.05N(5, 1)$. Thus in the first case condition (R) of Theorem 5 is satisfied. This implies $\tilde{\boldsymbol{\beta}}^* = \eta\tilde{\boldsymbol{\beta}}$, for some $\eta \in R$. One of our main goals is to investigate how the value of η affects the performance of model selection and testing procedures.

Recall that although Rudd's condition is a property of distribution of predictors and $\boldsymbol{\beta}$ it follows from definition of $\boldsymbol{\beta}^*$ that η depends on the model as well as on misspecified response $q(\cdot)$. Table 1 shows values of estimated proportionality constant η, denoted by $\hat{\eta}$. To calculate $\hat{\eta}$, for each variable $k \in t$, the value $\hat{\beta}_k/\beta_k$, where $\hat{\boldsymbol{\beta}}$ is based on $n = 10^6$ observations is computed and then the values are averaged over all attributes. The first column corresponds to $\eta = 1$ and it allows to gauge the variability of $\hat{\eta}$. Note also that the smallest value of $\hat{\eta}$ equal 0.52 and the second largest (equal 1.74) are obtained for the model M2 and responses q_6 and q_1, respectively. It follows that in the first case estimated $\boldsymbol{\beta}$ is on average two times smaller than the true one and around 1.7 times larger in the second case. Observe also that when $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are approximately proportional, for $q(s)$ such that $q(s) > p(s)$ for $s > 0$ we can expect that $\hat{\boldsymbol{\beta}} > \boldsymbol{\beta}$ as we try to match $q(\mathbf{x}_i'\boldsymbol{\beta})$ with $p(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$.

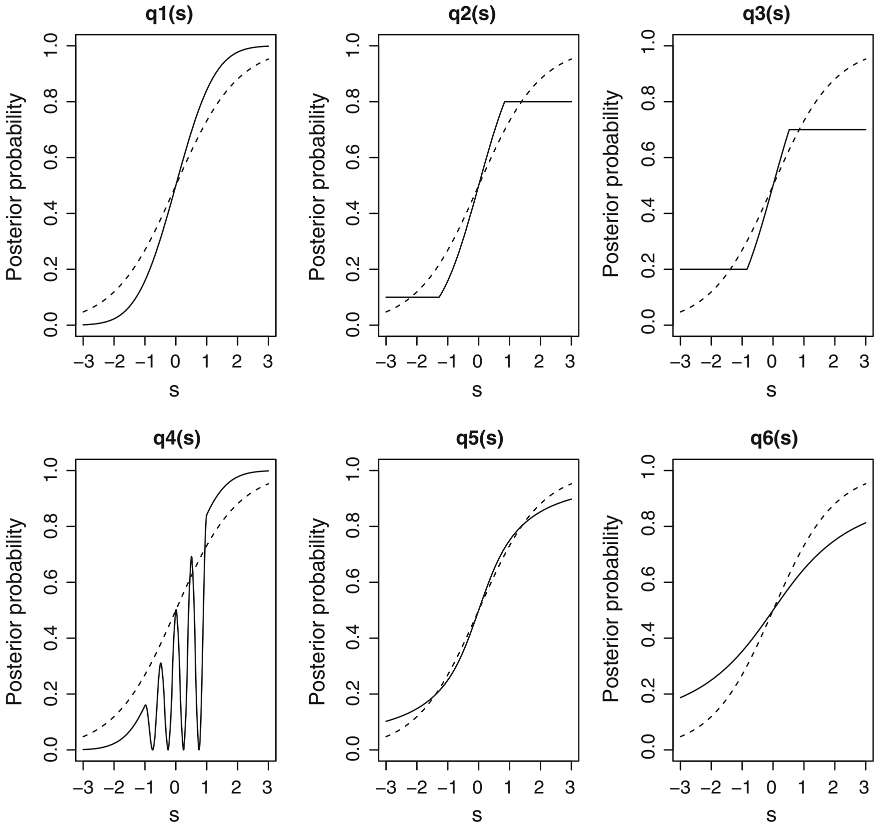

Fig. 1 Responses functions. *Dashed line* corresponds to fitted logit model $p(\cdot)$

Table 1 Values of $\hat{\eta}$ for considered models

Model	$p(\cdot)$	$q_1(\cdot)$	$q_2(\cdot)$	$q_3(\cdot)$	$q_4(\cdot)$	$q_5(\cdot)$	$q_6(\cdot)$
M1	0.988	1.642	1.591	1.591	0.788	1.241	0.651
M2	1.005	1.741	0.863	0.537	1.735	0.874	0.522
M3	0.993	1.681	1.352	0.968	1.524	1.045	0.580
M4	1.005	1.644	1.510	1.236	1.293	1.140	0.610
M5	1.013	1.779	0.897	0.552	1.724	0.879	0.532

This results in $\hat{\eta} > 1$. Thus as expected for q_1, $\hat{\eta}$ is greater than 1, whereas for q_6 it is smaller than 1.

It is noted in [2] (Sect. 4.2) that the probit function can be approximated by the scaled logit function as $q_1(s) \approx p(a \cdot s)$, where the scaling constant $a = \sqrt{8/\pi} \approx 1.6$ is chosen so that the derivatives of the two curves are equal for $s = 0$. Observe that constant a is very close to $\hat{\eta}$ calculated for q_1 (see Table 1).

In order to select the final model we use the two-step greedy procedure with Bayesian Information Criterion (BIC) described in Sect. 3. All fitted models include intercept.

Let $\hat{t}^*$ denote the model selected by a given selection criterion. As the measures of performance we use the following indices:

- probability of correct model selection (CS): $P(\hat{t}^* = t)$,
- positive selection rate (PSR): $\mathbf{E}(|\hat{t}^* \cap t|/|t|)$,
- false discovery rate (FDR): $\mathbf{E}(|\hat{t}^* \setminus t|/|\hat{t}^*|)$,
- power of significance test (POWER): $P(D_{null,\hat{t}^*} \in C_{1-\alpha}|H_1)$, where $\mathcal{C}_{1-\alpha}$ is critical region and H_1 corresponds to models M1–M5. Level $\alpha = 0.05$ was adopted throughout.

Empirical versions of the above measures are calculated and the results are averaged over 200 simulations. In the case of difficult models containing several predictors with small contributions CS can be close to zero and thus PSR and FDR are much more revealing measures of effectiveness. Observe that PSR is an average fraction of correctly chosen variables with respect to all significant ones whereas FDR measures a fraction of false positives (selected variables which are not significant) with respect to all chosen variables. Thus PSR $= 1$ means that all significant variables are included in the chosen model whereas FDR $= 0$ corresponds to the case when no spurious covariates are present in the final model. Instead of using critical region based on asymptotic distribution defined in (14) for which the significance level usually significantly exceeded assumed one, Monte Carlo critical value is calculated. For a given n and p 10000 datasets from null model are generated, for each one $\hat{t}^*$ and $D_{null,\hat{t}^*}$ is computed and this yields distribution of $D_{null,\hat{t}^*}$. The critical value is defined as empirical quantile of order $(1-\alpha)$ for $D_{null,\hat{t}^*}$.

Table 2 shows the results for $n = 200$. The highlighted values are maximal value in row (minimal values in case of FDR) and the last column pertains to maximal standard deviation in row. Observe that the type of response function influences greatly all considered measures of performance. Values of POWER are mostly larger than CS as detection of at least one significant variable usually leads to rejection of the null hypothesis. The most significant differences are observed for model M5 for which it is difficult to identify all significant variables as some coefficients are close to zero but it is much easier to reject the null model. However, when there is only one significant variable in the model, the opposite may be true as it happens for model M1. Note also that CS, PSR and POWER are usually large for large $\hat{\eta}$. To make this point more clear Fig. 2 shows the dependence of CS, PSR, POWER on $\hat{\eta}$. Model M1 is not considered for this graph as it contains only one significant predictor. In the case of CS, PSR and POWER monotone dependence is evident. However FDR is unaffected by the value of η which is understandable in view of its definition.

Table 3 shows the results for $n = 200$ when attributes x_j are generated from Gaussian mixture $0.95N(0, 1) + 0.05N(5, 1)$. Observe that the greatest impact of the change of $\mathbf{x}$ on CS occurs for truncated probit responses q_2 and q_3 for which in the case of M2–M5 CS drops dramatically. The change affects also PSR but to a lesser extent.

Table 2 CS, PSR, FDR and POWER for $x_j \sim N(0, 1)$ with $n = 200$, $p = 15$

Model		$p(\cdot)$	$q_1(\cdot)$	$q_2(\cdot)$	$q_3(\cdot)$	$q_4(\cdot)$	$q_5(\cdot)$	$q_6(\cdot)$	max sd
M1	CS	0.100	**0.410**	**0.410**	0.400	0.070	0.190	0.060	0.035
	PSR	0.170	**0.530**	**0.530**	0.520	0.110	0.300	0.080	0.036
	FDR	0.218	0.198	0.198	0.198	**0.142**	0.234	0.243	0.030
	POWER	0.080	**0.200**	**0.200**	**0.200**	0.110	0.120	0.040	0.028
M2	CS	0.820	0.760	0.850	0.550	0.770	**0.870**	0.590	0.035
	PSR	**1.000**	**1.000**	**1.000**	0.860	**1.000**	**1.000**	0.867	0.016
	FDR	0.050	0.072	0.040	0.051	0.061	**0.038**	0.064	0.011
	POWER	**1.000**	**1.000**	**1.000**	0.970	**1.000**	**1.000**	0.970	0.012
M3	CS	0.680	**0.790**	0.760	0.670	0.680	0.660	0.250	0.034
	PSR	0.920	**0.995**	0.975	0.910	0.985	0.940	0.590	0.023
	FDR	0.068	0.073	0.082	**0.060**	0.103	0.095	0.087	0.013
	POWER	0.980	**1.000**	**1.000**	0.950	**1.000**	0.990	0.550	0.035
M4	CS	0.300	**0.700**	0.680	0.440	0.380	0.380	0.050	0.035
	PSR	0.650	**0.940**	0.920	0.795	0.740	0.765	0.310	0.023
	FDR	0.130	0.078	**0.073**	0.113	0.140	0.103	0.153	0.021
	POWER	0.700	**1.000**	0.990	0.890	0.870	0.830	0.290	0.033
M5	CS	0.000	0.090	0.010	0.000	**0.110**	0.000	0.000	0.022
	PSR	0.647	**0.821**	0.601	0.391	0.815	0.595	0.372	0.012
	FDR	0.033	0.031	0.034	0.047	**0.024**	0.038	0.068	0.010
	POWER	**1.000**	**1.000**	**1.000**	0.950	**1.000**	**1.000**	0.930	0.018

To investigate this effect further we consider the probit function truncated at levels c and $1 - c$

$$q_7(s) = \begin{cases} F_{N(0,1)}(s) & \text{for } F_{N(0,1)}(s) \in (c, 1-c) \\ 0.2 & \text{for } F_{N(0,1)}(s) \leq c \\ 0.7 & \text{for } F_{N(0,1)}(s) \geq 1 - c, \end{cases}$$

which is a generalization of q_2 and q_3. Figure 7 shows how parameter c influences CS, PSR and FDR when the response is generated from q_7 and attributes are generated from Gaussian mixture $0.95N(0, 1) + 0.05N(5, 1)$.

To illustrate the result concerning the consistency of greedy two-step model selection procedure stated in Corollary 1 we made an experiment in which dependency on n is investigated. Figures 3 and 4 show considered measures of performance with respect to n for models M4 and M5. Somehow unexpectedly in some situations the results for incorrect model specification are better than for the correct specification, e.g. for model (M4) CS is larger for q_1, q_2 and q_4 than for $q(\cdot) = p(\cdot)$ (cf. Fig. 3). The results for q_6 are usually significantly worse than for p, which is related to the fact that $\hat{\eta}$ for this response is small (see again Table 1). Observe also that the type of response function clearly affects the PSRs whereas FDRs are similar in all cases.

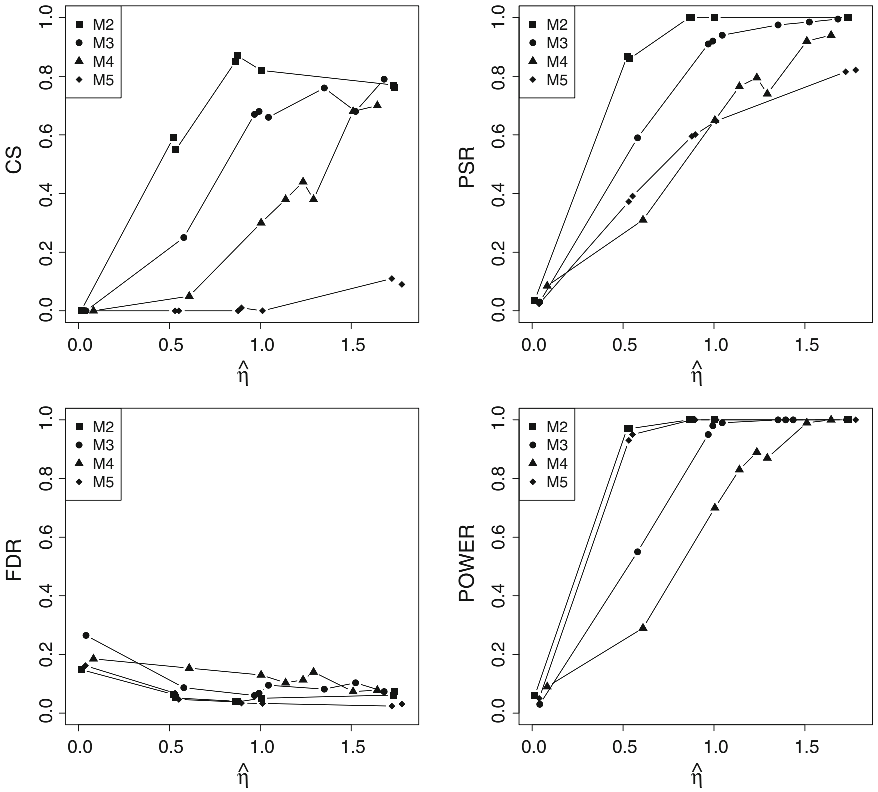

Fig. 2 CS, PSR, FDR, POWER versus $\hat{\eta}$ for $n = 200$, $p = 15$. Each point corresponds to different response function

Figure 5 shows how the power of the test of significance for the selected model and for the full model depends on the value of coefficient corresponding to the relevant variable in model M1. We see that for both correct and incorrect specification the power for selected model is slightly larger than for the full model for sufficiently large value of coefficient β_{10}. The difference is seen for smaller values of $\boldsymbol{\beta}$ in case of misspecification.

Finally we analysed how the number of potential attributes p influences the performance measures. The results shown in Fig. 6 for model M1 and $n = 500$ indicate that FDR increases significantly when spurious variables are added to the model. At the same time CS decreases when p increases, however, PSR is largely unaffected.

In conclusion we have established that when predictors are normal quality of model selection and power of the deviance test depend on the magnitude of Rudd's constant η. When $\eta > 1$ one can expect better results than for correct specification. Moreover, values of CS, PSR and POWER depend monotonically on η.

Table 3 CS, PSR, FDR and POWER for $x_j \sim 0.95N(0, 1) + 0.05N(5, 1)$ with $n = 200$, $p = 15$

Model		$p(\cdot)$	$q_1(\cdot)$	$q_2(\cdot)$	$q_3(\cdot)$	$q_4(\cdot)$	$q_5(\cdot)$	$q_6(\cdot)$	max sd
M1	CS	0.140	**0.540**	0.490	0.370	0.270	0.220	0.060	0.036
	PSR	0.220	**0.700**	0.670	0.490	0.330	0.320	0.110	0.036
	FDR	0.403	0.263	0.270	**0.233**	0.452	0.344	0.245	0.034
	POWER	0.220	**0.460**	0.450	0.240	0.340	0.260	0.090	0.035
M2	CS	**0.790**	0.730	0.180	0.050	0.780	0.720	0.350	0.034
	PSR	0.993	**1.000**	0.943	0.573	**1.000**	0.977	0.777	0.021
	FDR	**0.052**	0.070	0.278	0.227	0.056	0.084	0.094	0.016
	POWER	**1.000**	**1.000**	0.990	0.740	**1.000**	**1.000**	0.980	0.031
M3	CS	0.600	**0.740**	0.140	0.090	0.700	0.440	0.140	0.035
	PSR	0.925	**1.000**	0.915	0.725	0.990	0.855	0.600	0.021
	FDR	0.103	**0.095**	0.338	0.283	0.106	0.169	0.163	0.019
	POWER	**1.000**	**1.000**	0.990	0.840	**1.000**	**1.000**	0.790	0.029
M4	CS	0.330	**0.670**	0.120	0.040	0.410	0.210	0.010	0.035
	PSR	0.690	**0.920**	0.700	0.620	0.800	0.685	0.385	0.020
	FDR	0.148	**0.077**	0.235	0.230	0.127	0.147	0.248	0.027
	POWER	0.950	**1.000**	0.930	0.760	**1.000**	0.890	0.460	0.035
M5	CS	0.010	**0.140**	0.000	0.000	0.070	0.000	0.000	0.025
	PSR	0.641	**0.834**	0.338	0.194	0.792	0.573	0.324	0.011
	FDR	**0.013**	0.020	0.188	0.185	0.017	0.034	0.054	0.015
	POWER	**1.000**	**1.000**	0.970	0.720	**1.000**	**1.000**	0.960	0.032

In addition to tests on simulated data we performed an experiment on real data. We used Indian Liver Patient Dataset publicly available at UCI Machine Learning Repository [1]. This data set contains 10 predictors: age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. The binary response indicates whether the patient has a liver disease or not. Our aim was to use real explanatory variables describing the patients to generate an artificial response from different response functions. This can mimic the situation in which the liver disease cases follow some unknown distribution depending on explanatory variables listed above. We applied the following procedure. Predictors chosen by stepwise backward selection using BIC were considered. Estimators pertaining to 3 chosen variables (1st-age, 4th-direct Bilirubin and 6th-albumin) are treated as new true parameters corresponding to significant variables whereas the remaining variables are treated as not significant ones. Having the new parameter $\boldsymbol{\beta}$ and vectors of explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the data we generate new $y_1, \ldots, y_n$ using considered response functions $p, q_1, \ldots, q_6$.

Table 4 shows fraction of simulations in which the given variable was selected to the final model when the two-step procedure was applied. Note that this measure is less restrictive than CS used in previous experiments. Observe that the choice of response function affects the probabilities, e.g. direct Bilirubin is chosen in 80 %

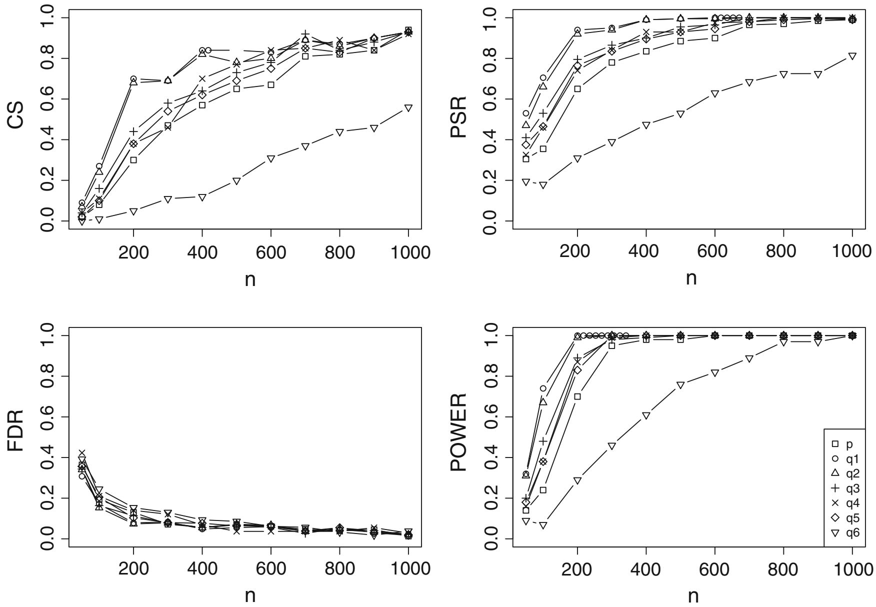

Fig. 3 CS, PSR, FDR, POWER versus *n* for model (M4), $p = 15$. Note change of the scale for FDR

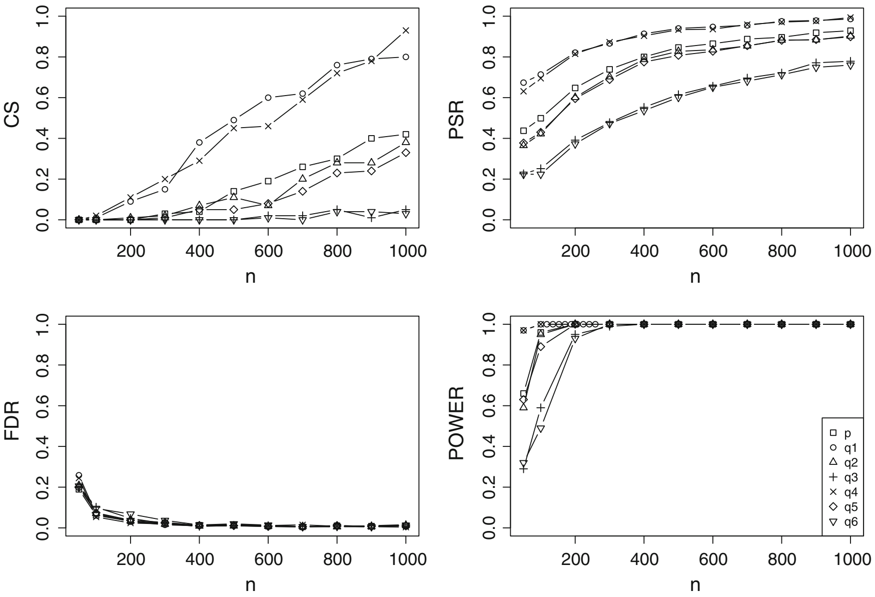

Fig. 4 CS, PSR, FDR, POWER versus *n* for model (M5), $p = 15$. Note change of the scale for FDR

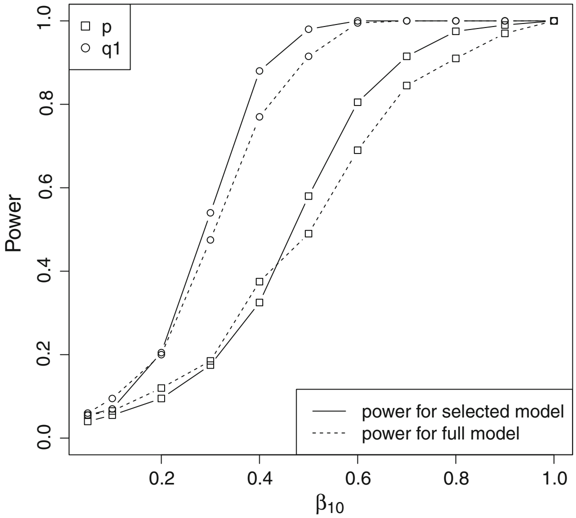

Fig. 5 Power versus β_{10} for selected model and full model, with $n = 200$, $p = 15$

simulations for correct specification and only in 12 % simulations for q_3. The significant variables are most often chosen to the final model for p and q_1. It is seen that direct Bilirubin is less likely to be selected in the case of most of the considered response functions (Fig. 7).

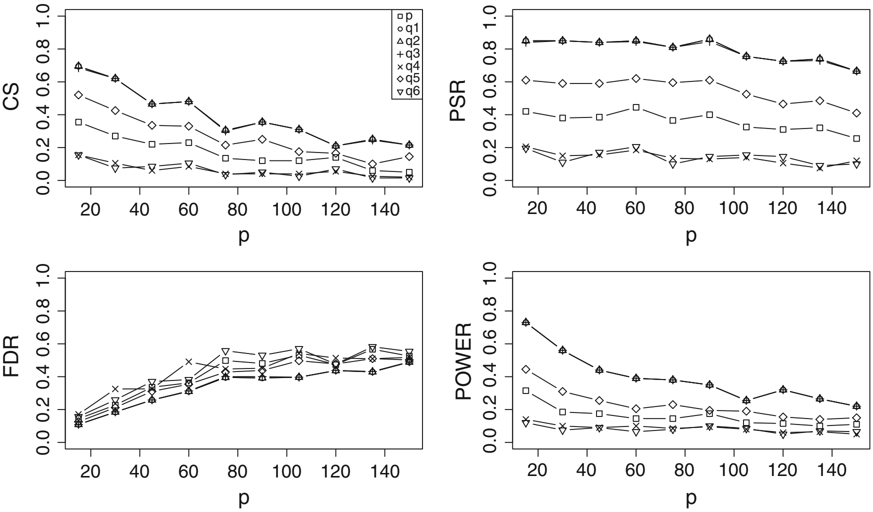

Fig. 6 CS, PSR, FDR, POWER versus p for model M1 with $n = 500$

Table 4 Probabilities of selecting variables to the final model for Indian liver patient dataset

Relevant variable	$\boldsymbol{\beta}$	p	q_1	q_2	q_3	q_4	q_5	q_6
1	−0.02	0.95	1.00	1.00	0.88	0.87	0.95	0.62
0	0.00	0.12	0.13	0.20	0.11	0.09	0.09	0.11
0	0.00	0.23	0.23	0.16	0.07	0.18	0.19	0.27
1	−0.67	0.80	0.77	0.36	0.12	0.30	0.60	0.63
0	0.00	0.11	0.15	0.26	0.08	0.17	0.10	0.17
1	−0.02	1.00	1.00	0.44	0.10	0.95	0.84	0.72
0	0.00	0.17	0.17	0.27	0.05	0.09	0.22	0.19
0	0.00	0.23	0.16	0.13	0.01	0.08	0.15	0.16
0	0.00	0.28	0.15	0.06	0.02	0.08	0.18	0.17
0	0.00	0.22	0.14	0.06	0.04	0.10	0.16	0.12

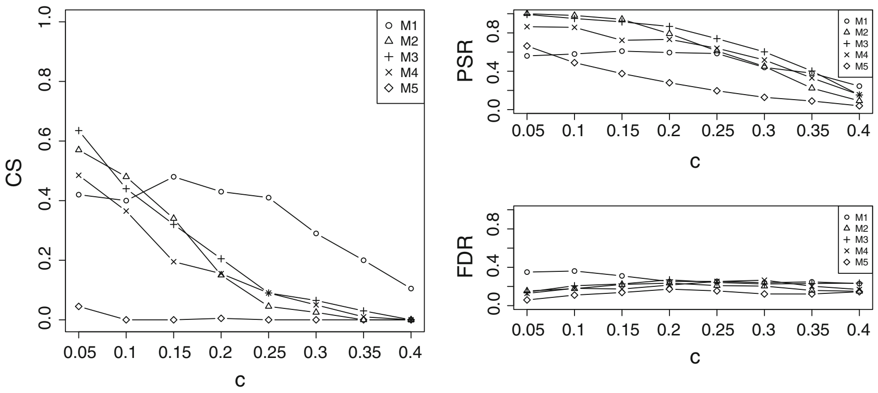

Fig. 7 CS, PSR, FDR versus c for q_7, $x_j \sim 0.95N(0, 1) + 0.05N(5, 1)$, $n = 200$ and $p = 15$

Appendix A: Auxiliary Lemmas

This section contains some auxiliary facts used in the proofs. The following theorem states the asymptotic normality of maximum likelihood estimator.

Theorem 6 *Assume (A1) and (A2). Then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, J^{-1}(\boldsymbol{\beta}^*)K(\boldsymbol{\beta}^*)J^{-1}(\boldsymbol{\beta}^*))$$

where J and K are defined in (5) and (9), respectively.

The above Theorem is stated in [11] (Theorem 3.1) and in [16] ((2.10) and Sect. $5B$).

Lemma 4 *Assume that* $\max_{1\le i\le n}|\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta})|\le C$ *for some* $C>0$ *and some* $\boldsymbol{\gamma}\in R^{p+1}$. *Then for any* $\mathbf{c}\in R^{p+1}$

$$\exp(-3C)\mathbf{c}'J_n(\boldsymbol{\beta})\mathbf{c}\le\mathbf{c}'J_n(\boldsymbol{\gamma})\mathbf{c}\le\exp(3C)\mathbf{c}'J_n(\boldsymbol{\beta})\mathbf{c},\quad a.e.$$

Proof It suffices to show that for $i=1,\ldots,n$

$$\exp(-3C)p(\mathbf{x}_i'\boldsymbol{\beta})[1-p(\mathbf{x}_i'\boldsymbol{\beta})]\le p(\mathbf{x}_i'\boldsymbol{\gamma})[1-p(\mathbf{x}_i'\boldsymbol{\gamma})]\le\exp(3C)p(\mathbf{x}_i'\boldsymbol{\beta})[1-p(\mathbf{x}_i'\boldsymbol{\beta})].$$

Observe that for $\boldsymbol{\gamma}$ such that $\max_{i\le n}|\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta})|\le C$ there is

$$\frac{p(\mathbf{x}_i'\boldsymbol{\gamma})[1-p(\mathbf{x}_i'\boldsymbol{\gamma})]}{p(\mathbf{x}_i'\boldsymbol{\beta})[1-p(\mathbf{x}_i'\boldsymbol{\beta})]}=e^{\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta})}\left[\frac{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\gamma}}}\right]^2\ge e^{-C}\left[\frac{e^{-\mathbf{x}_i'\boldsymbol{\beta}}+1}{e^{-\mathbf{x}_i'\boldsymbol{\beta}}+e^{C}}\right]^2\ge e^{-3C}. \tag{15}$$

By replacing $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (15) we obtain the upper bound for $\mathbf{c}'J_n(\boldsymbol{\gamma})\mathbf{c}$.

Lemma 5 *Assume (A1) and (A2). Then* $l(\hat{\boldsymbol{\beta}},\mathbf{Y}|\mathbf{X})-l(\boldsymbol{\beta}^*,\mathbf{Y}|\mathbf{X})=O_P(1)$.

Proof Using Taylor expansion we have for some $\bar{\boldsymbol{\beta}}$ belonging to the line segment joining $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$

$$l(\hat{\boldsymbol{\beta}},\mathbf{Y}|\mathbf{X})-l(\boldsymbol{\beta}^*,\mathbf{Y}|\mathbf{X})=\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'[J_n(\bar{\boldsymbol{\beta}})/n]\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/2, \tag{16}$$

Define set $A_n=\{\boldsymbol{\gamma}:||\boldsymbol{\gamma}-\boldsymbol{\beta}^*||\le s_n\}$, where s_n is an arbitrary sequence such that $ns_n^2\to 0$. Using Schwarz and Markov inequalities we have for any $C>0$

$$\begin{aligned}P[\max_{i\le i\le n}|\mathbf{x}_i'(\boldsymbol{\gamma}-\boldsymbol{\beta}^*)|>C]&\le P[\max_{1\le i\le n}||\mathbf{x}_i||s_n>C]\\ &\le n\max_{i\le i\le n}P[||\mathbf{x}_i||>Cs_n^{-1}]\le C^{-2}ns_n^2\mathbf{E}(||\mathbf{x}||^2)\to 0.\end{aligned}$$

Thus using Lemma 4 the quadratic form in (16) is bounded with probability tending to 1 from above by

$$\exp(3C)\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)'[J_n(\boldsymbol{\beta}^*)/n]\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)/2,$$

which is $O_P(1)$ as $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)=O_P(1)$ in view of Theorem 6 and $n^{-1}J_n(\boldsymbol{\beta}^*)\xrightarrow{P}J(\boldsymbol{\beta}^*)$.

A.1 Proof of Lemma 2

As $\boldsymbol{\beta}_m^* = \boldsymbol{\beta}_c^*$ we have for $c \supseteq m \supseteq t^*$

$$l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_m, \mathbf{Y}|\mathbf{X}) = [l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\beta}_c^*, \mathbf{Y}|\mathbf{X})] + [l(\boldsymbol{\beta}_m^*, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_m|\mathbf{X}, \mathbf{Y})],$$

which is $O_P(1)$ in view of Remark 1 and Lemma 5.

A.2 Proof of Lemma 3

The difference $l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_w, \mathbf{Y}|\mathbf{X})$ can be written as

$$[l(\hat{\boldsymbol{\beta}}_c, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X})] + [l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X}) - l(\hat{\boldsymbol{\beta}}_w|\mathbf{X}, \mathbf{Y})]. \tag{17}$$

It follows from Lemma 5 and Remark 1 that the first term in (17) is $O_P(1)$. We will show that the probability that the second term in (17) is greater or equal $\alpha_1 n d_n^2$, for some $\alpha_1 > 0$ tends to 1. Define set $A_n = \{\boldsymbol{\gamma} : ||\boldsymbol{\gamma} - \boldsymbol{\beta}^*|| \le d_n\}$. Using the Schwarz inequality we have

$$\sup_{\boldsymbol{\gamma}\in A_n} \max_{i\le n} |\mathbf{x}_i'(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)| < \max_{1\le i\le n} ||\mathbf{x}_i||d_n \le 1, \tag{18}$$

with probability one. Define $H_n(\boldsymbol{\gamma}) = l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X})$. Note that $H(\boldsymbol{\gamma})$ is convex and $H(\boldsymbol{\beta}^*) = 0$. For any incorrect model w, in view of definition (11) of d_n, we have $\hat{\boldsymbol{\beta}}_w \notin A_n$ for sufficiently large n. Thus it suffices to show that $P(\inf_{\boldsymbol{\gamma}\in\partial A_n} H_n(\boldsymbol{\gamma}) > \alpha_1 n d_n^2) \to 1$, as $n \to \infty$, for some $\alpha_1 > 0$. Using Taylor expansion for some $\bar{\boldsymbol{\gamma}}$ belonging to the line segment joining $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}^*$

$$l(\boldsymbol{\gamma}, \mathbf{Y}|\mathbf{X}) - l(\boldsymbol{\beta}^*, \mathbf{Y}|\mathbf{X}) = (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' s_n(\boldsymbol{\beta}^*) - (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)/2$$

and the last convergence is implied by

$$P[\sup_{\boldsymbol{\gamma}\in\partial A_n} (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' s_n(\boldsymbol{\beta}^*) > \inf_{\boldsymbol{\gamma}\in\partial A_n} (\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)/2 - \alpha_1 n d_n^2] \to 0. \tag{19}$$

It follows from Lemma 4 and (18) that for $\boldsymbol{\gamma} \in A_n$

$$(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma} - \boldsymbol{\beta}^*) \ge e^{-3}(\boldsymbol{\gamma} - \boldsymbol{\beta}^*)' J_n(\boldsymbol{\beta}^*)(\boldsymbol{\gamma} - \boldsymbol{\beta}^*). \tag{20}$$

Let $\tau = \exp(-3)/2$. Using (20), the probability in (19) can be bounded from above by

$$P[\sup_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'s_n(\boldsymbol{\beta}) > \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta})) - \alpha_1 n d_n^2] \\ + P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma}-\boldsymbol{\beta})/2 < \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))]. \tag{21}$$

Let $\lambda_1^- = \lambda_{\min}(J(\boldsymbol{\beta}))/2$. Assuming $\alpha_1 < \lambda_1^-\tau$, the first probability in (21) can be bounded by

$$P[d_n||s_n(\boldsymbol{\beta})|| > \tau n d_n^2\lambda_1^- - \alpha_1 n d_n^2] + P[\lambda_{\min}(J_n(\boldsymbol{\beta})) < \lambda_1^- n] \\ \leq P[||s_n(\boldsymbol{\beta})|| > (\tau\lambda_1^- - \alpha_1)n^{1/2}a_n^{1/2}] \\ + P[nd_n < n^{1/2}a_n^{1/2}] + P[\lambda_{\min}(J_n(\boldsymbol{\beta})) < \lambda_1^- n]. \tag{22}$$

Consider the first probability in (22). Note that $s_n(\boldsymbol{\beta}^*)$ is a random vector with zero mean and the covariance matrix $K_n(\boldsymbol{\beta}^*)$. Using Markov's inequality, the fact that $\text{cov}[s_n(\boldsymbol{\beta}^*)] = nK(\boldsymbol{\beta}^*)$ and taking $\alpha_1 < \lambda^-\tau$ it can be bounded from above by

$$\frac{tr\{\text{cov}[s_n(\boldsymbol{\beta}^*)]\}}{(\tau\lambda^- - \alpha_1)^2 n^2 d_n^2} = \frac{tr[K_n(\boldsymbol{\beta}^*)]}{(\tau\lambda^- - \alpha_1)^2 n^2 d_n^2} \leq \frac{n\kappa p}{(\tau\lambda^- - \alpha_1)^2 n^2 d_n^2} \\ \leq \frac{\kappa p}{(\tau\lambda^- - \alpha_1)^2 a_n} \to 0, \tag{23}$$

where the last convergence follows from $a_n \to \infty$.

The convergence to zero of the second probability in (22) follows from $nd_n^2/a_n \xrightarrow{P} \infty$. As eigenvalues of a matrix are continuous functions of its entries, we have $\lambda_{\min}(n^{-1}J_n(\boldsymbol{\beta}^*)) \xrightarrow{P} \lambda_{\min}(J(\boldsymbol{\beta}^*))$. Thus the convergence to zero of the third probability in (22) follows from the fact that in view of (A1) matrix $J(\boldsymbol{\beta}^*)$ is positive definite. The second term in (21) is bounded from above by

$$P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'J_n(\bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma}-\boldsymbol{\beta})/2 < \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))] \\ \leq P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'[J_n(\bar{\boldsymbol{\gamma}}) - 2\tau J_n(\boldsymbol{\beta})](\boldsymbol{\gamma}-\boldsymbol{\beta})/2 \\ + 2\tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))/2 < \tau d_n^2\lambda_{\min}(J_n(\boldsymbol{\beta}))] \\ \leq P[\inf_{\boldsymbol{\gamma}\in\partial A_n}(\boldsymbol{\gamma}-\boldsymbol{\beta})'[J_n(\bar{\boldsymbol{\gamma}}) - 2\tau J_n(\boldsymbol{\beta})](\boldsymbol{\gamma}-\boldsymbol{\beta})/2 < 0] \to 0,$$

where the last convergence follows from Lemma 4 and (18).

Lemma 6 *Assume (A2) and (A3). Then we have* $\max_{i\leq n}||\mathbf{x}_i||^2 a_n/n \xrightarrow{P} 0$.

Proof Using Markov inequality, (A2) and (A3) we have that $||\mathbf{x}_n||^2 a_n/n \xrightarrow{P} 0$. We show that this implies the conclusion. Denote $g_n := \max_{1\leq i\leq n}||\mathbf{x}_i||^2 a_n/n$ and $h_n := ||\mathbf{x}_n||^2 a_n/n$. Define sequence n_k such that $n_1 = 1$ and $n_{k+1} = \min\{n > n_k : \max_{i\leq n}||\mathbf{x}_i||^2 > \max_{i\leq n_k}||\mathbf{x}_i||^2\}$ (if such n_{k+1} does not exist put $n_{k+1} = n_k$). Without loss of generality we assume that for $A = \{n_k \to \infty\}$ we have $P(A) = 1$

as on A^c the conclusion is trivially satisfied. Observe that $g_{n_k} = h_{n_k}$ and $h_{n_k} \xrightarrow{P} 0$ as a subsequence of $h_n \xrightarrow{P} 0$ and thus also $g_{n_k} \xrightarrow{P} 0$. This implies that for any $\epsilon > 0$ there exists $n_0 \in \mathbf{N}$ such that for $n_k > n_0$ we have $P[|g_{n_k}| \leq \epsilon] \geq 1 - \epsilon$. As for $n \in (n_k, n_{k+1})$ $g_n \leq g_{n_k}$ since a_n/n is nonincreasing we have that if $n \geq n_0$ $P[|g_n| \leq \epsilon] \geq 1 - \epsilon$ i.e. $g_n \xrightarrow{P} 0$.

A.3 Proof of Proposition 1

Assume first that $\tilde{\boldsymbol{\beta}}^* = 0$ and note that this implies $p(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}}^*) = p(\beta_0) = C \in (0, 1)$. From (8) we have

$$P(y = 1) = \mathbf{E}(y) = \mathbf{E}[\mathbf{E}(y|\tilde{\mathbf{x}})] = \mathbf{E}[q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})] = \mathbf{E}[p(\beta_0^* + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}}^*)] = C. \quad (24)$$

Using (24) and (7) we get

$$\begin{aligned}\mathbf{E}(\tilde{\mathbf{x}}y) &= \mathbf{E}\{\mathbf{E}[\tilde{\mathbf{x}}y|\tilde{\mathbf{x}}]\} = \mathbf{E}\{\tilde{\mathbf{x}}\mathbf{E}[y|\tilde{\mathbf{x}}]\} = \mathbf{E}[\tilde{\mathbf{x}}q(\beta_0 + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}})] \\ &= \mathbf{E}[\tilde{\mathbf{x}}p(\beta_0^* + \tilde{\mathbf{x}}'\tilde{\boldsymbol{\beta}}^*)] = \mathbf{E}(\tilde{\mathbf{x}})C.\end{aligned} \quad (25)$$

From (24) we also have

$$\mathbf{E}(\tilde{\mathbf{x}}y) = \mathbf{E}\tilde{\mathbf{x}}I\{y = 1\} = \mathbf{E}(\tilde{\mathbf{x}}|y = 1)P(y = 1) = \mathbf{E}(\tilde{\mathbf{x}}|y = 1)C.$$

Comparing the last equation and right-side term in (25) we obtain $\mathbf{E}(\tilde{\mathbf{x}}|y = 1) = E\tilde{\mathbf{x}} = \mathbf{E}(\tilde{\mathbf{x}}|y = 0)$. Assume now $\mathbf{E}(\tilde{\mathbf{x}}|y = 1) = \mathbf{E}(\tilde{\mathbf{x}}|y = 0)$ which implies as before that that $\mathbf{E}(\tilde{\mathbf{x}}|y = 1) = \mathbf{E}(\tilde{\mathbf{x}})$. Thus

$$\mathbf{E}(\tilde{\mathbf{x}}y) = \mathbf{E}(\tilde{\mathbf{x}}|y = 1)\mathbf{E}(y) = \mathbf{E}(\tilde{\mathbf{x}})\mathbf{E}(y). \quad (26)$$

Since $(\beta_0^*, \tilde{\boldsymbol{\beta}}^*)$ is unique it suffices to show that (7) and (8) are satisfied for $\tilde{\boldsymbol{\beta}}^* = 0$ and β_0^* such that $Ep(\beta_0^*) = P(Y = 1)$. This easily follows from (26).

References

1. Bache K, Lichman M (2013) UCI machine learning repository. University of California, Irvine
2. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
3. Bogdan M, Doerge R, Ghosh J (2004) Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. Genetics 167:989–999
4. Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analitycal extensions. Psychometrika 52:345–370
5. Burnham K, Anderson D (2002) Model selection and multimodel inference. A practical information-theoretic approach. Springer, New York

6. Carroll R, Pederson S (1993) On robustness in the logistic regression model. J R Stat Soc B 55:693–706
7. Casella G, Giron J, Martinez M, Moreno E (2009) Consistency of Bayes procedures for variable selection. Ann Stat 37:1207–1228
8. Chen J, Chen Z (2008) Extended Bayesian Information Criteria for model selection with large model spaces. Biometrika 95:759–771
9. Chen J, Chen Z (2012) Extended BIC for small-n-large-p sparse glm. Statistica Sinica 22: 555–574
10. Claeskens G, Hjort N (2008) Model selection and model averaging. Cambridge University Press, Cambridge
11. Czado C, Santner T (1992) The effect of link misspecification on binary regression inference. J Stat Plann Infer 33:213–231
12. Fahrmeir L (1987) Asymptotic testing theory for generalized linear models. Statistics 1:65–76
13. Fahrmeir L (1990) Maximum likelihood estimation in misspecified generalized linear models. Statistics 4:487–502
14. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Stat Assoc 96:1348–1360
15. Foster D, George E (1994) The risk inflation criterion for multiple regression. Ann Stat 22: 1947–1975
16. Hjort N, Pollard D (1993) Asymptotics for minimisers of convex processes. Unpublished manuscript
17. Konishi S, Kitagawa G (2008) Information criteria and statistical modeling. Springer, New York
18. Lehmann E (1959) Testing statistical hypotheses. Wiley, New York
19. Li K, Duan N (1991) Slicing regression: a link-free regression method. Ann Stat 19(2):505–530
20. Qian G, Field C (2002) Law of iterated logarithm and consistent model selection criterion in logistic regression. Stat Probab Lett 56:101–112
21. Ruud P (1983) Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. Econometrica 51(1):225–228
22. Sin C, White H (1996) Information criteria for selecting possibly misspecified parametric models. J Econometrics 71:207–225
23. Zak-Szatkowska M, Bogdan M (2011) Modified versions of Baysian Information Criterion for sparse generalized linear models. Comput Stat Data Anal 5:2908–2924
24. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67(2):301–320