*Communications in Statistics—Theory and Methods*, 41: 4484–4502, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2011.573161

Taylor & Francis
Taylor & Francis Group

# Selection of Regression and Autoregression Models with Initial Ordering of Variables

JAN MIELNICZUK AND PAWEŁ TEISSEYRE

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

*We consider an information criterion for model selection in random design linear regression and autoregression case which allows for a general penalty and a general averaging factor for sum of squared residuals replacing reciprocal of a sample size. This leads to a consistent selection of a set of non zero coefficients. The search over all subsets may be replaced by search over nested family when predictors are pre-ordered with respect to their significance in the largest model. We show that such procedure detects the significant variables in both regression setups even when the number of models increases with a sample size.*

**Keywords** Akaike and Schwarz rule; Consistency; Generalized information criterion; Penalty; Prediction error; Regression and autoregression.

**Mathematics Subject Classification** 62J05; 62H12.

## 1. Introduction

Consider the linear regression model $\mathcal{M}_{1,\ldots,M_n}$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations which variability we would like to explain, $\mathbf{X}$ is a random $n \times M_n$ design matrix consisting of vectors of $M_n$ attributes (regressors) collected from $n$ objects and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an unknown vector of errors, assumed to have $N(0, \sigma^2\mathbf{I})$ distribution. Here, and throughout, $\mathbf{a}'$ denotes transposition of a column vector $\mathbf{a}$. Vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M_n})'$ is an unknown vector of parameters. When all potential regressors are included in the model $\mathcal{M}_{1,\ldots,M_n}$, we face the possibility that part of them is redundant, i.e., respective coefficients in (1) are 0. In order to detect the minimal adequate model corresponding only to non zero parameters it is common to consider the list of all submodels $\mathcal{M}_{i_1,\ldots,i_k}$ of $\mathcal{M}_{1,\ldots,M_n}$ with $1 \le i_k \le M_n$ and $k \le M_n$ such that only coefficients $\beta_{i_1}, \ldots, \beta_{i_k}$ are not

Received July 8, 2010; Accepted March 8, 2011
Address correspondence to Paweł Teisseyre, Institute of Computer Science, Polish Academy of Sciences, 21 J.K., Ordona 01-237 Warsaw, Poland; E-mail: teisseyrep@ipipan.waw.pl

set equal to 0 in $\mathcal{M}_{i_1,\ldots,i_k}$. Then a one of the considered models is chosen, usually the minimizer of a certain criterion which reflects goodness of fit and, at the same time, complexity of the model under consideration. A common approach to construct such a criterion is to base it on some estimator of a predictive loglikelihood, i.e., loglikelihood calculated at future observations. More precisely, the object of interest is -2× loglikelihood, i.e., $-2\log \mathcal{L}_{\widehat{\boldsymbol{\beta}}(\mathcal{M}_{i_1,\ldots,i_k})}(Y_1^0,\ldots,Y_n^0)$ , where $Y_1^0,\ldots,Y_n^0$ are future values of dependent variable observed at the same design points as $Y_1,\ldots,Y_n$ and $\widehat{\boldsymbol{\beta}}(\mathcal{M}_{i_1,\ldots,i_k})$ is maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ calculated for the considered model $\mathcal{M}_{i_1,\ldots,i_k}$. The aim is then to find a model for which some estimator of averaged predictive likelihood is minimized. Akaike (1970) shown that when $\sigma^2$ is assumed known and the linear regression model is correctly specified, asymptotically unbiased estimate of the predictive loglikelihood for linear model (1) is $-2\log \mathcal{L}_{\widehat{\boldsymbol{\beta}}(\mathcal{M}_{i_1,\ldots,i_k})}(Y_1,\ldots,Y_n)+2k$, which up to an additive constant equals to $RSS(\mathcal{M}_{i_1,i_2,\ldots,i_k})/\sigma^2+2k$. Here, $RSS(\mathcal{M}_{i_1,i_2,\ldots,i_k})$ denotes the residual sum of squares, i.e., sum of squared residuals from the fit of the model $\mathcal{M}_{i_1,i_2,\ldots,i_k}$. This leads to frequently used Akaike Information Criterion (AIC)

$$AIC(\mathcal{M}_{i_1,\ldots,i_k}) = RSS(\mathcal{M}_{i_1,i_2,\ldots,i_k}) + 2k\widehat{\sigma}^2, \tag{2}$$

where $\widehat{\sigma}^2$ is an estimator of the error variance based on the full model $\mathcal{M}_{1,\ldots,M_n}$. Criterion (2) is equivalent to $C_p$ criterion introduced by Mallows (1973). An analogous reasoning for the case of unknown variance leads to a criterion

$$\widetilde{AIC}(\mathcal{M}_{i_1,\ldots,i_k}) = n\log\left[\frac{RSS(\mathcal{M}_{i_1,i_2,\ldots,i_k})}{n}\right]+2k. \tag{3}$$

The above procedures are subject to various modifications and have led to different proposals, which usually introduce various forms of a penalty. It is frequently of the form $ka_n\widehat{\sigma}^2$ and $ka_n$ for (2) and (3), respectively, i.e., a penalty is a function of a sample size. In particular, $a_n=\log n$ corresponds to Bayesian Information Criterion introduced by Schwarz (1978) (see Haughton, 1988 for formal justification of its properties) and $a_n=\log\log n$ to the Hannan and Quinn (1979) proposal. Choosing the minimal adequate model from the list of $2^{M_n}-1$ models is hindered by a considerable, and, for large $M_n$, enormous computational cost. Moreover, and equally importantly, for large $M_n$ comparable to the sample size $n$ such procedures may be inconsistent in the sense that they pick too large subset of attributes. Indeed, for an orthogonal matrix $\mathbf{X}$ it is easily seen that minimizing AIC corresponds to multiple testing problem of sequence of hypotheses $H_{i0}:\beta_i=0$ for which many false rejections happen for large $M_n$. In order to mend this drawback, Zheng and Loh (1997) proposed a two-step procedure, the first step of which corresponds to fitting the full model and ordering the regressors in decreasing order of importance according to the absolute values of $t$ statistics. Denote the resulting permutation of indices by $\{j_1,j_2,\ldots,j_{M_n}\}$. Then the second step consists of finding the minimizer of generalized AIC for *nested* list of models $\mathcal{M}_{j_1},\mathcal{M}_{j_1,j_2},\ldots,\mathcal{M}_{j_1,\ldots,M_n}$ having length $M_n$. It is easy to observe that the ordering according to the values of $F$ test statistics for the full model and a model with a successive variable omitted yields the same permutation as $F=t^2$. Zheng and Loh (1997) proved consistency of such rule when $M_n/n\to 0$ and at the same time $a_n/M_n\to\infty$ and have shown that the proposal behaves promisingly in limited simulation experiments for regression as

well autoregression models. We will call such procedure a selection rule with initial ordering of variables. Despite its obvious advantages, Zheng and Loh's proposal has not attracted much attention, an exception is Bunea et al. (2007) who combined pre-ordering of variables with Hochberg-Benjamini approach.

The aim of this article is twofold. First, we establish consistency of a generalization of $\widetilde{AIC}$ in (3), which is called $\widetilde{GIC}$, for regression with random explanatory variables and autoregressive models under similar conditions to those considered by Zheng and Loh. This seems to be worthwhile as generalizations of $\widetilde{AIC}$ are commonly used for model selection. Generalization of $\widetilde{AIC}$ consists not only in considering a more general form of penalty but also in replacing $RSS(\mathscr{M}_{i_1,\ldots,i_k})/n$ by a general form of variance estimator $RSS(\mathscr{M}_{i_1,\ldots,i_k})/(n-c_{n,k})$, where $0 \leq c_{n,k} \leq M_n$. We prove in Theorem 2.1 that when $\widetilde{GIC}$ is minimized over all possible non empty subsets of $\{1, 2, \ldots, M_n\}$, the resulting selection rule consistently identifies the set of true regressors. Moreover, minimization over nested family of models is sufficient for consistent estimation of a maximal index of non zero coefficient. In both results, the length of the list $M_n$ is allowed to grow with the sample size. This is important as the upper bound on the number of significant predictors is rarely known in practice. The proved results imply consistency of two-step procedure with initial ordering of variables for the considered criterion. The analogous results for non random case are also briefly discussed. Second, we investigate by simulations two problems: how likely is to choose a correct model by one of these methods and, at the same time, what is prediction error of such proposals. Specifically, we consider a post-model selection estimator of $\mathbf{X}\boldsymbol{\beta}$ equal $\mathbf{X}\widehat{\boldsymbol{\beta}}(\widehat{\mathscr{M}})$, where $\widehat{\boldsymbol{\beta}}(\widehat{\mathscr{M}})$ is ML estimator of $\boldsymbol{\beta}$ in the model $\widehat{\mathscr{M}}$ chosen by the considered criterion. The problem that good consistency properties do not necessarily mean that post-model selection estimators are good predictors is known; see in Shao (1997). This is, however, rarely investigated for specific regression problems. As the simulation experiments show, this modification of variance estimator turns out to be important, especially in regression context, where variance estimator with $c_{n,k} = k$ works much better than ML estimator.

This article is organized as follows. In Sec. 2, we discuss technical preliminaries, state the main results, and discuss them. Sec. 3 presents results of the simulation experiments. The results are proved in the Appendix.

## 2. Preliminaries and Results

It is important to discuss, in detail, the assumed framework as the theoretical properties of selection methods and post-model selection estimators depend on it in a crucial way. We consider models with random regressors (explanatory variables), i.e., we assume that the rows $\mathbf{X}_1', \ldots, \mathbf{X}_n'$ of a matrix $\mathbf{X}(n \times M_n)$ are iid, $\mathbf{X}_i = \mathbf{X}_i^{(n)} = (X_{i,1}^{(n)}, \ldots X_{i,M_n}^{(n)})'$. Thus, $\{\mathbf{X}_1^{(n)'}, \ldots, \mathbf{X}_n^{(n)'}\}$ constitute rows in an array of iid sequences of $M_n$-dimensional random variables. We impose the condition that $M_n$ is non decreasing and that the law of the first $M_n$ coordinates of $\mathbf{X}_1^{(n+1)}$ coincides with that of $\mathbf{X}_1^{(n)}$, i.e., the distribution of attributes considered for a certain sample size remains the same for larger sample sizes. We assume throughout that $E[\mathbf{X}_i^{(n)}\mathbf{X}_i^{(n)'}]$ is finite, i.e., the second moments of all coordinates exist. Errors $\varepsilon_1, \ldots, \varepsilon_n$ are assumed independent and normal with mean 0 and a common unknown variance $\sigma^2$. Throughout we also impose the assumption that data is generated according to random design linear regression model with fixed $\boldsymbol{\beta}$ and non random number $j_0$

of regressors, such that it is a submodel of a model $\mathcal{M}_{1,2,\ldots,M_n}$ for sufficiently large $n$. This means that for such $n$, $M_n \geq j_{\max}$, where $j_{\max}$ is the maximal index of non zero coefficient (cf. assumption (A1.3) below). Observe that $j_0 = j_{\max}$ when all true coefficients proceed spurious ones. Let

$$\Gamma = \{i : 1 \leq i \leq M_n, \beta_i \neq 0\} \tag{4}$$

be the index set of true parameters. Throughout this article we will assume that $\mathbf{X}'\mathbf{X}$ is invertible with probability approaching 1. We will impose the following assumption.

(A0) For each $n$ matrix $\mathbf{E}[\mathbf{X}_1^{(n)}\mathbf{X}_1^{(n)'}]$ is invertible.

Depending on the context we will be using some of the following additional conditions on $a_n$, $M_n$, and matrix $\mathbf{X}$.

(A1.1) $a_n/n \to 0$ as $n \to \infty$.
(A1.2) $a_n/M_n \to \infty$ as $n \to \infty$.
(A1.3) $\liminf_{n\to\infty} M_n \geq j_{\max}$.
(A1.4) $M_n/n \to 0$ as $n \to \infty$.
(A1.5) The minimum eigenvalue $\kappa_n$ of $\mathbf{E}[\mathbf{X}_1^{(n)}\mathbf{X}_1^{(n)'}]$ is bounded away from zero, i.e. $\kappa_n > \kappa > 0$ for some $\kappa > 0$ and $n \in \mathbf{N}$.
(A1.6) For some $\eta > 0$, $n^{-1}M_n^{1+\eta} \to 0$ and

$$\sup_n \sup_{\|\mathbf{d}\|=1} \mathbf{E}|\mathbf{d}'\mathbf{Z}^{(n)}|^{4\lceil 2/\eta \rceil} < \infty,$$

where $\mathbf{Z}^{(n)} = \mathbf{E}[(\mathbf{X}_1^{(n)}\mathbf{X}_1^{(n)'})^{-1/2}\mathbf{X}_1^{(n)}]$ is the standardized vector $\mathbf{X}_1^{(n)}$ and $\lceil 2/\eta \rceil$ is the smallest integer greater than or equal to $2/\eta$.

The assumptions (A1.2) and (A1.4) imply that the length $M_n$ of the list of models has to be small not only when compared to sample size but also with respect to penalty. In particular, BIC criterion satisfies the assumptions only when $M_n = o(\log n)$. The assumptions (A1.5) and the second part of (A1.6), used in Zheng and Loh (1997), imply, in particular, that with probability tending to one $(\mathbf{X}'\mathbf{X})^{-1}$ exists and therefore $\widehat{\boldsymbol{\beta}}$ is unique. Similar conditions were used by Mammen (1993) to study the asymptotic behavior of bootstrap estimators of contrasts in linear models of increasing dimension.

### 2.1. *All Subsets Search*

Our objective here is consistent estimation of set $\Gamma$. The following generalized information criterion is proposed. We choose the minimal model $\mathcal{M}_{\widehat{\Gamma}}$ such that

$$\widehat{\Gamma} = \arg\min_{i_1,\ldots,i_j} \widetilde{GIC}(\mathcal{M}_{i_1,\ldots,i_j}), \tag{5}$$

where

$$\widetilde{GIC}(\mathcal{M}_{i_1,\ldots,i_j}) = n \log\left[\frac{RSS(\mathcal{M}_{i_1,\ldots,i_j})}{n - c_{n,j}}\right] + ja_n, \tag{6}$$

$1 \le j \le M_n$ and $0 \le c_{n,j} \le M_n$. Thus, $\widetilde{GIC}$ is minimised over $2^{M_n}$-1 non empty subsets of $\{1, 2, \ldots, M_n\}$.

**Theorem 2.1.** *Under conditions* (A0) *and* (A1.1)–(A1.3) *and* (A1.5)–(A1.6), $\widehat{\Gamma}$ *is consistent estimator of* $\Gamma$, *i.e.*, $P(\widehat{\Gamma} = \Gamma) \to 1$ *as* $n \to \infty$.

**Remark 2.1.** For the case that the coordinates $X_{i,1}^{(n)}, \ldots, X_{i,M_n}^{(n)}$ of $\mathbf{X}_i^{(n)}$ are i.i.d. conditions (A1.5) and (A1.6) may be replaced by the conditions (A1.4) and $\mathbf{E}X_{1,1}^4 < \infty$.

### 2.2. *Nested Family Search*

Recall that $j_{\max} = \max\{j : \beta_j \in \Gamma\}$ is defined as the largest index of non zero coefficients in the true model. Consider now the situation when $\widetilde{GIC}$ is minimized not over all subsets of $\{1, 2, \ldots, M_n\}$ but only over the nested list of models $\mathcal{M}_{1,\ldots,j}$, $j = 1, \ldots, M_n$. Then it turns out that, under weaker conditions than in Theorem 2.1, $j_{\max}$ can be consistently identified. Namely, let

$$\hat{j}_{\max} = \arg \min_{1 \le j \le M_n} \widetilde{GIC}(\mathcal{M}_{1,\ldots,j}). \tag{7}$$

In this case, assumptions (A1.5) and (A1.6) will be replaced by the following assumption. Let $\mathbf{X}_{i,j_{\max}} = (X_{i,1}, \ldots, X_{i,j_{\max}})'$. We assume the following assumption.

(A1.5') $\mathbf{E}(\mathbf{X}_{1,j_{\max}}\mathbf{X}'_{1,j_{\max}})$ is positive definite matrix.

Let $\mathbf{D}_{j_{\max}}$ be a $M_n \times j_{\max}$ matrix of zeros and ones such that $\mathbf{X}\mathbf{D}_{j_{\max}}$ consists of only the first $j_{\max}$ columns of $\mathbf{X}$. Observe that in view of law of large numbers $n^{-1}(\mathbf{X}\mathbf{D}_{j_{\max}})'(\mathbf{X}\mathbf{D}_{j_{\max}}) \xrightarrow{P} \mathbf{E}(\mathbf{X}_{1,j_{\max}}\mathbf{X}'_{1,j_{\max}})$ as as $n \to \infty$, which is positive definite in view of (A1.5'). Then the following result holds.

**Theorem 2.2.** *Under conditions* (A0) *and* (A1.1), (A1.2), (A1.3), (A1.4), (A1.5') $\hat{j}_{\max} \xrightarrow{P} j_{\max}$ *as* $n \to \infty$.

### 2.3. *Two-step Procedure*

Theorem 2.2 indicates that, provided that the coordinates corresponding to non zero coefficients are placed ahead of the spurious ones, a search of hierarchical list of models suffices to consistently determine $\Gamma$. This method requires less computation as we consider only $M_n$ different models instead of searching through all $2^{M_n} - 1$ subsets. Thus, it is desirable to consider methods of ordering which, with high probability, yield the correct order in this sense. The two-step procedure introduced in Zheng and Loh (1997) relies on ordering according to the absolute values of $t$-statistics in the full model. To introduce this method define the following quantities. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{M_n})'$ be the least squares estimator of $\beta$ based on the full model $\mathcal{M}_{1,\ldots,M_n}$ and let

$$T_i = \frac{\widehat{\beta}_i}{\widehat{\sigma}\left(\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{i,i}}\right)}, \quad i = 1, \ldots, M_n \tag{8}$$

be the corresponding t-statistic. In the above definition, $\widehat{\sigma}^2 = (n - M_n)^{-1} RSS(\mathscr{M}_{1,\ldots,M_n})$ is the usual variance estimator based on the full model and $\mathbf{A}_{i,i}$ denotes the $i$th diagonal element of matrix $\mathbf{A}$. In view of (A1.6), matrix $(\mathbf{X}'\mathbf{X})^{-1}$ exists with probability tending to one and therefore vector $\widehat{\beta}$ is unique. The analogu of Zheng and Loh's procedure for $\widetilde{GIC}$ consists of the following.

**Step 1.** Order the absolute values of t-statistics $|T_{i_1}| \geq |T_{i_2}| \geq \cdots \geq |T_{i_{M_n}}|$.
**Step 2.** Apply criterion (7) to the ordered variables $i_1, i_2, \ldots, i_{M_n}$ and choose model

$$\widehat{\Gamma} = \{i_1, i_2, \ldots, i_{\hat{j}_0}\}, \tag{9}$$

where $\hat{j}_0 = \min\{j_0^* : j_0^* = \arg\min_{1 \leq j \leq M_n} \widetilde{GIC}(\mathscr{M}_{i_1,\ldots,i_j})\}$.

To prove consistency of the ordering in Step 1, conditions (A1.5) and (A1.6) considered in Zheng and Loh (1997) are assumed. Result of Zheng and Loh (1997) together with Theorem 2.2 now yields

**Theorem 2.3.** *Under conditions* (*A*0) *and* (*A*1.1)–(*A*1.3) *and* (*A*1.5)–(*A*1.6),

$$\lim_{n\to\infty} P[\min_{i\in\Gamma} |T_i| > \max_{i\notin\Gamma} |T_i|] = 1.$$

*Moreover, criterion* (9) *is consistent estimator of* $\Gamma$.

Observe that in view of Theorem 2.2 *any* pre-ordering which puts all true regressors ahead of all spurious ones with probability tending to 1 can be used in place of Zheng and Loh's ordering.

### 2.4. *The Case of Deterministic Covariates*

In this section we will briefly discuss the case when the design matrix $\mathbf{X}$ is non random. In this case, we will impose some conditions on $\mathbf{X}$ which in the previous section follow from the random structure of regressors. We assume that, for each $n$ ($M_n \times M_n$), matrix $\mathbf{X}'\mathbf{X}$ is invertible. The notions introduced in Secs. 2.2 and 2.3 are used in the sequel. In the case of all subset search we will replace conditions (A1.5) and (A1.6) by the following assumption:

(A2.1) The minimum eigenvalue $\tilde{\kappa}_n$ of $n^{-1}\mathbf{X}'\mathbf{X}$ is bounded away from zero, i.e., $\tilde{\kappa}_n > \tilde{\kappa} > 0$ for some $\tilde{\kappa} > 0$ and $n \in \mathbf{N}$.

(A2.1) is used to prove that $P(\min_{i\in\Gamma} \widehat{\sigma}^2 T_i^2 < a_n) \to 1$ for any $a_n$ such that $a_n = o(n)$ (cf. end of the proof of Theorem 2.1). It follows that the analogous result to Theorem 2.1 holds.

**Theorem 2.4.** *Under conditions* (*A*1.1)-(*A*1.4) *and* (*A*2.1), $\widehat{\Gamma}$ *is consistent estimator of* $\Gamma$ *i.e.* $P(\widehat{\Gamma} = \Gamma) \to 1$ *as* $n \to \infty$.

The proof of Theorem 2.4 is analogous to the proof of Theorem 2.1, the main difference being now that the property following from (A2.1) listed above is used. We omit the details. It is also easily seen that Theorem 2.3 holds for deterministic regressors under conditions of Theorem 2.4. In the case of nested family search,

let $Q(\mathcal{M}_{1,\ldots,j})$ be a projection on the column space spanned by the regressors corresponding to coefficients in a given model $\mathcal{M}_{1,\ldots,j}$. In this case, we will replace condition (A1.5') by the following commonly used assumption on $\mathbf{X}$ (see Shao, 1993 and Zhang, 1992).

(A2.2) $\min_{j<j_{\max}}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-Q(\mathcal{M}_{1,\ldots,j}))(\mathbf{X}\boldsymbol{\beta}) > \delta n$, for some $\delta > 0$.

Then the following result holds

**Theorem 2.5.** *Under conditions* (A1.1), (A1.2), (A1.3), (A1.4), (A2.2) $\widehat{j}_{\max} \xrightarrow{P} j_{\max}$ *as* $n \to \infty$.

The assumption (A2.2) can be weakened to the condition

$$\min_{j<j_{\max}}(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-Q(\mathcal{M}_{1,\ldots,j}))(\mathbf{X}\boldsymbol{\beta}) > \delta_n \log(n),$$

where $\delta_n \to \infty$ and $[\delta_n \log(n)]^{-1} a_n \to 0$, as $n \to \infty$ (cf. Zheng and Loh, 1995). The proof which follows the lines of the proof of Theorem 2.2 is omitted.

### 2.5. *Autoregressive Models*

In this section we apply the criterion defined in Sec. 2.2 to the model selection of autoregressive processes. The process $(Y_i)_{-\infty}^{\infty}$ is assumed to be autoregressive of order $M_n$ ($AR(M_n)$), i.e., it satisfies

$$Y_i = \beta_1 Y_{i-1} + \cdots + \beta_{M_n} Y_{i-M_n} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{10}$$

where $\varepsilon_i$ are i.i.d. normal with mean zero and variance $\sigma^2$. We consider the case when process $Y_i$ is stationary, which holds when the autoregressive polynomial $\phi(z) = \beta_1 z^{M_n} + \beta_2 z^{M_n-1} + \cdots + \beta_{M_n}$ does not have zeros on the unit circle. Model (10) can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $i$-th row of the matrix $\mathbf{X}$ is given by $\mathbf{X}_i' = (Y_{i-1}, \ldots, Y_{i-M_n})$. All the previous notation applies. We assume that the true model is $AR(j_{\max})$, that is $\beta_{j_{\max}} \neq 0$, $\beta_j = 0$, $j > j_{\max}$ and $j_{\max} \in \mathbf{N}$ is independent of $n$. Our goal now is to estimate consistently the maximal index of non zero coefficient. We apply criterion (7) defined in Section 2.2. We make the following additional assumption on $M_n$.

(A1.4') ${M_n}^2/n \to 0$ as $n \to \infty$.

**Theorem 2.6.** *Under conditions* (A0) *and* (A1.1), (A1.2), (A1.3), (A1.4'), (A1.5') $\widehat{j}_{\max} \xrightarrow{P} j_{\max}$ *as* $n \to \infty$.

## 3. Numerical Experiments

The aim of this section is to study the finite-sample performance of the considered variable selection procedure. We consider two forms of the Generalized Information Criterion:

$$GIC(\mathcal{M}_{i_1,\ldots,i_j}) = RSS(\mathcal{M}_{i_1,\ldots,i_j}) + j a_n \widehat{\sigma}^2$$

and $\widetilde{GIC}$ defined in (6). The first criterion was investigated in Zheng and Loh (1997).

### 3.1. *Linear Regression*

The theoretical background for $\widetilde{GIC}$ is discussed in Sec. 2. We compared the following forms of penalty: $a_n = 2$, $a_n = \log(n)$ and $a_n = \log\log(n)$, which for $c_{n,j} = 0$ correspond to Akaike, Schwarz, and Hannan-Quinn proposals, respectively. We also considered $GIC(\mathcal{M}_{i_1,\ldots,i_j})$ and $\widetilde{GIC}(\mathcal{M}_{i_1,\ldots,i_j})$ with penalty $a_n = n^{\alpha}$ proposed by Zheng and Loh. We used their values of $\alpha$ which were chosen as a function of $M_n$ and independently of the model. We considered two estimators of the variance: the unbiased variance estimator for which $c_{n,j} = j$ and the ML estimator with $c_{n,j} = 0$. The simulation experiments were carried out with sample sizes $n = 300$ and $n = 1,000$ repeated $N = 200$ times. The following regression models have been considered:

(m1) $\Gamma = \{1\}$ with $\beta_1 = -0.3$ and $M_n = 5$;
(m2) $\Gamma = \{10\}$ with $\beta_{10} = 0.2$ and $M_n = 30$;
(m3) $\Gamma = \{1, 2, 5, 6\}$ with $\boldsymbol{\beta} = (0.9, -0.8, -0.4, 0.2)'$ and $M_n = 20$.

The second model was also considered in Zheng and Loh (1997). The rows of the design matrix $\mathbf{X}$ were generated independently from normal distribution with mean and variance equal to 0 and 2, respectively. The distribution of $\varepsilon_i$ was standard normal. The investigated selection method was two-step procedure, described in the previous section. In the first step the covariates were ordered by values of their $t$-statistics and then the considered selection rules were applied to the nested list of models. Table 3 presents estimated probabilities of correct ordering of variables, e.g., the probabilities that the coordinates corresponding to non zero coefficients are placed ahead the spurious ones. It is seen that for $n \geq 100$ a correct ordering is recovered practically always. We assess the effectiveness of the selection rule in terms of the probability of true model selection $P(\widehat{\Gamma} = \Gamma)$, where $\widehat{\Gamma}$ is a model selected by the considered rule and mean squared error $\mathbf{E}(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathcal{M}_{\widehat{\Gamma}})\|^2)$, where $\widehat{\boldsymbol{\beta}}(\mathcal{M}_{\widehat{\Gamma}})$ is the post-model selection estimator of $\boldsymbol{\beta}$. In the experiments, estimates of these measures calculated as the empirical means of respective quantities were considered. The influence of the size of the list $M_n$ on the effectiveness of selection rules was also investigated. Tables 1 and 2 indicate that pre-ordering coupled with

**Table 1**
Estimated probability of correct regression model selection and its standard error for $n = 300$ based on $N = 200$ trials

| Model | Method | $M_n$ | $a_n = 2$ | $a_n = \log(n)$ | $a_n = \log\log(n)$ | $a_n = n^{\alpha}$ |
|---|---|---|---|---|---|---|
| (m1) | $GIC$ | 5 | 0.51 (0.03) | 0.93 (0.01) | 0.78 (0.02) | 0.93 (0.01) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.5 (0.03) | 0.94 (0.01) | 0.78 (0.02) | 0.93 (0.01) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.71 (0.03) | 0.95 (0.01) | 0.84 (0.02) | 0.94 (0.01) |
| (m2) | $GIC$ | 30 | 0.02 (0.01) | 0.63 (0.03) | 0.20 (0.02) | 0.99 (0.01) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.01 (0.01) | 0.63 (0.03) | 0.19 (0.02) | 0.99 (0.01) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.10 (0.02) | 0.74 (0.03) | 0.37 (0.03) | 0.99 (0.01) |
| (m3) | $GIC$ | 20 | 0.07 (0.02) | 0.75 (0.03) | 0.32 (0.03) | 0.53 (0.03) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.06 (0.01) | 0.74 (0.03) | 0.27 (0.03) | 0.52 (0.03) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.18 (0.02) | 0.83 (0.02) | 0.56 (0.03) | 0.52 (0.03) |

**Table 2**
Sample mean of prediction error and standard deviation in the case of regression model for $n = 300$ based on $N = 200$ trials

| Model | Method | $M_n$ | $a_n = 2$ | $a_n = \log(n)$ | $a_n = \log\log(n)$ | $a_n = n^{\alpha}$ |
|---|---|---|---|---|---|---|
| (m1) | $GIC$ | 5 | 3.46 (0.24) | 1.63 (0.17) | 2.43 (0.21) | 1.63 (0.17) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 3.48 (0.24) | 1.63 (0.17) | 2.43 (0.22) | 1.63 (0.17) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 2.71 (0.23) | 1.56 (0.17) | 2.09 (0.20) | 1.60 (0.17) |
| (m2) | $GIC$ | 30 | 17.14 (0.58) | 4.67 (0.39) | 10.11 (0.52) | 1.32 (0.23) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 17.93 (0.60) | 4.70 (0.39) | 10.61 (0.54) | 1.32 (0.23) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 12.45 (0.54) | 3.64 (0.36) | 7.35 (0.47) | 1.32 (0.23) |
| (m3) | $GIC$ | 20 | 13.01 (0.45) | 6.00 (0.35) | 9.37 (0.42) | 14.33 (0.80) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 13.33 (0.46) | 6.06 (0.36) | 9.60 (0.41) | 14.42 (0.80) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 10.45 (0.42) | 5.61 (0.38) | 7.55 (0.39) | 14.42 (0.80) |

the penalties $a_n = \log(n)$ and $a_n = n^{\alpha}$ is the most effective in terms of both accuracy measures. In concordance with Zheng and Loh (1997), the value of parameter $\alpha$ was set to 0.3 for $M_n = 5$ and 0.7 for $M_n = 30$ and 60. For other values of $M_n$ values of $\alpha$ were linearly approximated. This choice is appropriate for some models (see model m2 in Table 1) but seems to work poorly for others (see model m3 in Table 1 and Fig. 1). The modified Schwarz rule with $c_{n,j} = j$ performs consistently better than for $c_{n,j} = 0$ which shows that the introduction of the correction factor is useful. Akaike criterion $\widetilde{GIC}$ with $a_n = 2$ and $c_{n,j} = 0$ performs much worse, especially in sparse model (m2) in which only one regressor among 30 potential regressors is significant. The same is true in even more pronounced form for Hannan-Quinn criterion. Even in the case of Akaike and Hannan-Quinn criteria the change from $c_{n,j} = 0$ to $c_{n,j} = j$ yields a significant improvement of performance. This modification always improved the performance in all cases but Zheng and Loh's proposal. The results also indicate that model m2 with the only one significant variable placed at position 10 is the most difficult for selection for the models considered. Figures 1 and 2 show that performance of the Schwarz rule is heavily influenced by the choice of the horizon $M_n$, however, the selection pertaining to $c_{n,j} = j$ is the least affected. The penalty $a_n = n^{\alpha}$ proposed by Zheng and Loh seems to be too heavy in this case, especially for large $M_n$. Note that in this case probability of correct model selection is close to 0 for $M_n \geq 20$. It seems that preliminary

**Table 3**
Estimated probability of correct ordering and its standard error in the case of regression model based on $N = 200$ trials

| Model | $M_n$ | $n = 50$ | $n = 100$ | $n = 300$ |
|---|---|---|---|---|
| (m1) | 5 | 0.86 (0.02) | 0.99 (0.01) | 1 (0) |
| (m2) | 30 | 0.93 (0.01) | 0.99 (0.01) | 1 (0) |
| (m3) | 20 | 0.95 (0.01) | 1 (0) | 1 (0) |

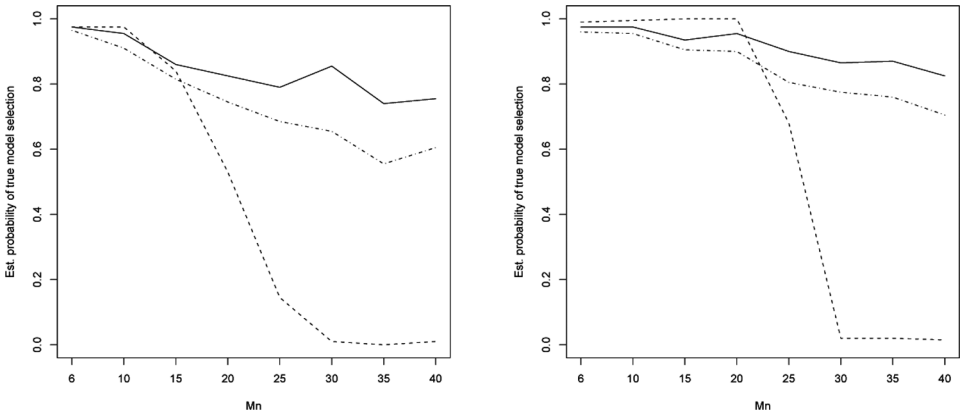


**Figure 1.** Estimated probabilities of correct regression model selection with respect to $M_n$, for $GIC$ with $c_{n,j} = 0$, $a_n = n^{\alpha}$ (dashed line), $\widetilde{GIC}$ with $c_{n,j} = 0$, $a_n = \log(n)$ (dot-dashed line), and $\widetilde{GIC}$ with $c_{n,j} = j$, $a_n = \log(n)$ (solid line). The true model is m3 and sample sizes are: $n = 300$ (a) and $n = 1,000$ (b).

estimation of horizon would further enhance the behaviour of this selection rule. This will be a subject of further study.

### 3.2. *Autoregression*

We consider analogous models m1–m3 as in the regression case; the past value of autoregressive sequence now plays a role of regressors. The investigated selection method was also a two-step procedure. We tested two cases in which $c_{n,j} = 0$ and $c_{n,j} = j$, where $j$ corresponds to the number of nonzero coefficients. The results

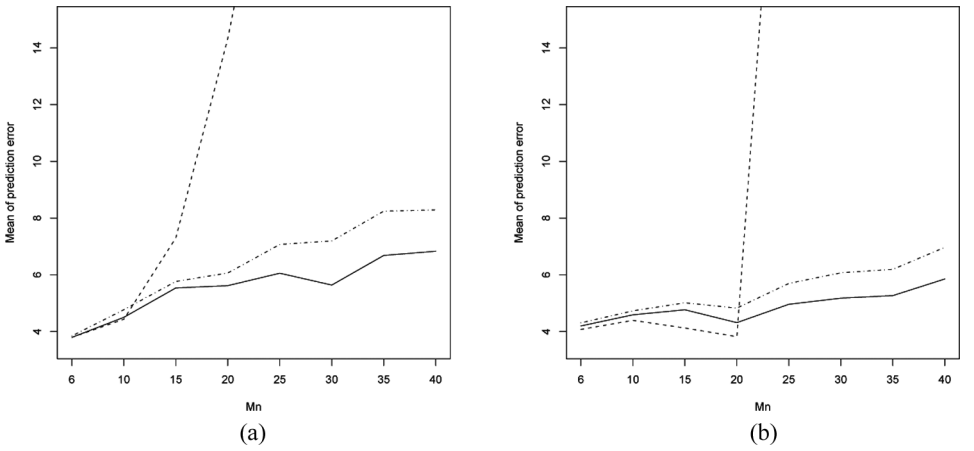


**Figure 2.** Means of prediction errors with respect to $M_n$, for $GIC$ with $c_{n,j} = 0$, $a_n = n^{\alpha}$ (dashed line), $\widetilde{GIC}$ with $c_{n,j} = 0$, $a_n = \log(n)$ (dot-dashed line), and $\widetilde{GIC}$ with $c_{n,j} = j$, $a_n = \log(n)$ (solid line). The true model is m3 and sample sizes are: $n = 300$ (a) and $n = 1,000$ (b).

**Table 4**
Estimated probability of correct autoregression model selection and its standard error for $n = 300$ based on $N = 200$ trials

| Model | Method | $M_n$ | $a_n = 2$ | $a_n = \log(n)$ | $a_n = \log\log(n)$ | $a_n = n^{\alpha}$ |
|---|---|---|---|---|---|---|
| (m1) | $GIC$ | 5 | 0.46 (0.03) | 0.95 (0.01) | 0.76 (0.03) | 0.95 (0.01) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.46 (0.03) | 0.95 (0.01) | 0.77 (0.02) | 0.95 (0.01) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.7 (0.03) | 0.97 (0.01) | 0.87 (0.02) | 0.97 (0.01) |
| (m2) | $GIC$ | 30 | 0.01 (0.00) | 0.48 (0.03) | 0.16 (0.02) | 0.69 (0.03) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.01 (0.00) | 0.48 (0.03) | 0.15 (0.02) | 0.69 (0.03) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.10 (0.02) | 0.57 (0.03) | 0.31 (0.03) | 0.69 (0.03) |
| (m3) | $GIC$ | 20 | 0.18 (0.02) | 0.43 (0.03) | 0.34 (0.03) | 0.36 (0.03) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.18 (0.02) | 0.43 (0.03) | 0.32 (0.03) | 0.36 (0.03) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.28 (0.03) | 0.45 (0.03) | 0.38 (0.03) | 0.34 (0.03) |

given in Tables 4 and 5 and Figs. 3 and 4 are similar to that for regression in that the ordering of methods with respect to both considered measures remains the same. Probabilities of correct model selection are smaller and prediction errors larger on average than for corresponding regression problem.

### 3.3. *Real Data Example*

We consider `bodyfat` data set (Johnson, 1996) consisting of records of the percentage of fat in the body (dependent variable) together with 13 independent variables for $n = 252$ individuals. Three independent variables were selected having the smallest p-values when the full linear model was fitted. They were abdomen, hip, and wrist circumference and when used as predictors resulted in the fitted model with a coefficient of determination $R^2 = 0.96$, a vector of estimated coefficients $\widehat{\boldsymbol{\beta}} =$

**Table 5**
Sample mean of prediction error and standard deviation in the case of autoregression for $n = 300$ based on $N = 200$ trials

| Model | Method | $M_n$ | $a_n = 2$ | $a_n = \log(n)$ | $a_n = \log\log(n)$ | $a_n = n^{\alpha}$ |
|---|---|---|---|---|---|---|
| (m1) | $GIC$ | 5 | 3.50 (0.22) | 1.45 (0.14) | 2.35 (0.19) | 1.45 (0.14) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 3.51 (0.22) | 1.45 (0.14) | 2.32 (0.19) | 1.45 (0.14) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 2.59 (0.20) | 1.33 (0.13) | 1.87 (0.17) | 1.33 (0.13) |
| (m2) | $GIC$ | 30 | 16.70 (0.57) | 7.01 (0.49) | 10.81 (0.53) | 5.74 (0.54) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 17.38 (0.57) | 7.09 (0.50) | 11.26 (0.54) | 5.74 (0.54) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 12.66 (0.57) | 6.10 (0.48) | 8.49 (0.49) | 5.74 (0.54) |
| (m3) | $GIC$ | 20 | 13.47 (0.58) | 11.17 (0.69) | 10.90 (0.60) | 19.97 (1.05) |
| | $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 13.60 (0.57) | 11.19 (0.69) | 11.39 (0.62) | 20.42 (1.08) |
| | $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 11.60 (0.59) | 11.58 (0.72) | 10.98 (0.64) | 21.39 (1.14) |

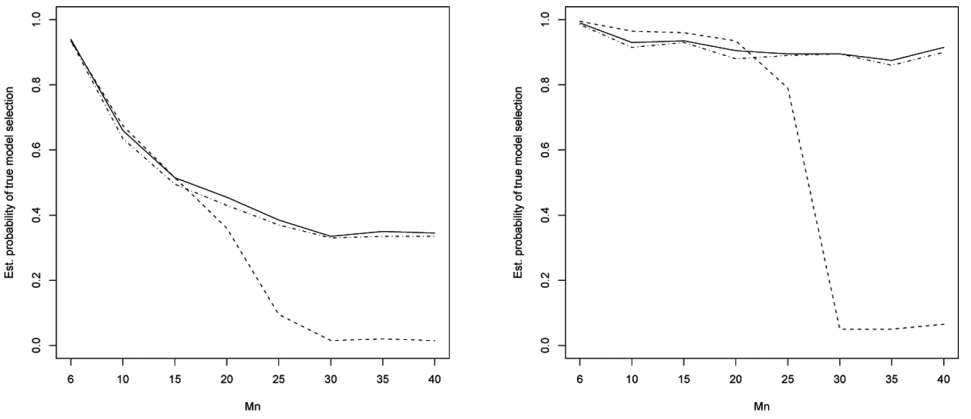


**Figure 3.** Estimated probabilities of correct autoregression model selection with respect to $M_n$, for $GIC$ with $c_{n,j} = 0$, $a_n = n^{\alpha}$ (dashed line), $\widetilde{GIC}$ with $c_{n,j} = 0$, $a_n = \log(n)$ (dot-dashed line), and $\widehat{GIC}$ with $c_{n,j} = j$, $a_n = \log(n)$ (solid line). The true model is m3 and sample sizes are: $n = 300$ (a) and $n = 1000$ (b).

$(0.92, -0.32, -1.86)'$ and a variance of residuals $\widehat{\sigma}^2 = 4.45$. A parametric bootstrap (see, e.g., Davison and Hinkley, 1997) was employed to check how the considered selection criteria perform for this data set. Namely, the true model was the fitted linear model with the original three regressors, $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and the normal errors with the variance equal to $\widehat{\sigma}^2$. Additional superfluous explanatory variables were created in triples by drawing from the three-dimensional normal distribution with independent components, which mean and variance vector matched that of the original predictors. We considered $k = 1, 2, \ldots, 15$ additional triples what amounted

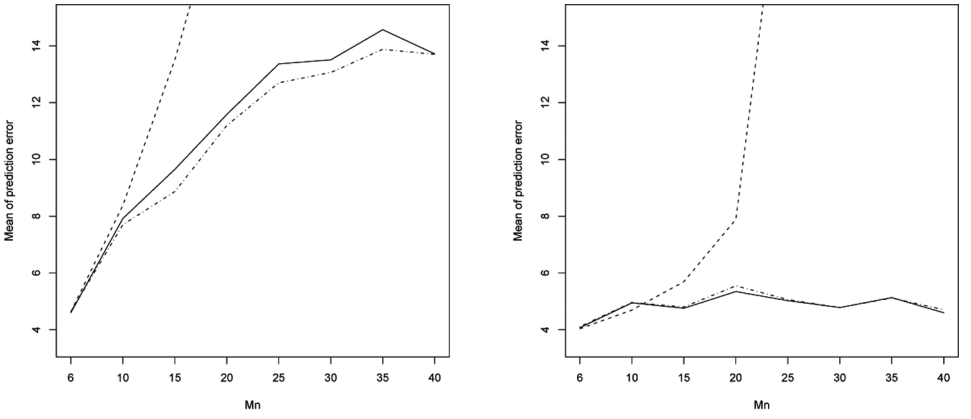


**Figure 4.** Means of prediction errors with respect to $M_n$, for $GIC$ with $c_{n,j} = 0$, $a_n = n^{\alpha}$ (dashed line), $\widetilde{GIC}$ with $c_{n,j} = 0$, $a_n = \log(n)$ (dot-dashed line), and $\widehat{GIC}$ with $c_{n,j} = j$, $a_n = \log(n)$ (solid line). The true model is m3 and sample sizes are: $n = 300$ (a) and $n = 1000$ (b).

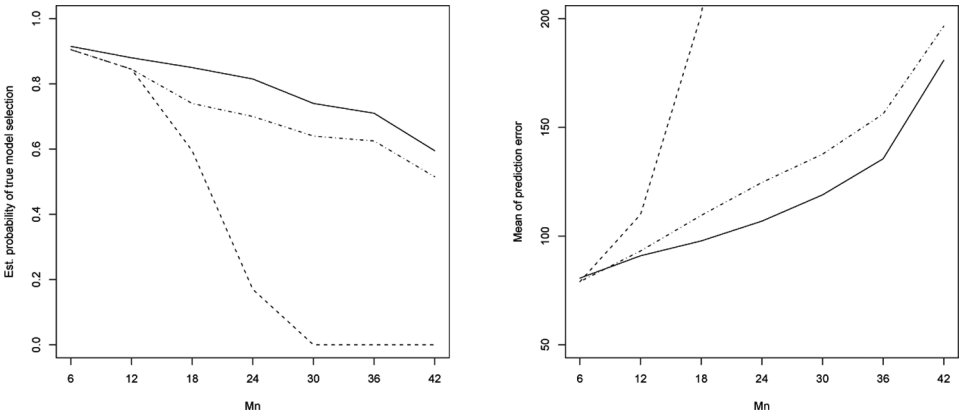


**Figure 5.** Estimated probabilities of correct model selection (a) and means of prediction errors (b) with respect to $M_n$, for $GIC$ with $c_{n,j} = 0$, $a_n = n^{\alpha}$ (dashed line), $\widetilde{GIC}$ with $c_{n,j} = 0$, $a_n = \log(n)$ (dot-dashed line), and $\widetilde{GIC}$ with $c_{n,j} = j$, $a_n = \log(n)$ (solid line).

**Table 6**
Estimated probability of correct model selection and its standard error

| Method | $M_n$ | $a_n = 2$ | $a_n = \log(n)$ | $a_n = \log\log(n)$ | $a_n = n^{\alpha}$ |
|---|---|---|---|---|---|
| $GIC$ | 18 | 0.12 (0.02) | 0.75 (0.03) | 0.41 (0.02) | 0.6 (0.03) |
| $\widetilde{GIC}$ $(c_{n,j} = 0)$ | | 0.11 (0.02) | 0.74 (0.03) | 0.42 (0.03) | 0.57 (0.03) |
| $\widetilde{GIC}$ $(c_{n,j} = j)$ | | 0.32 (0.03) | 0.85 (0.02) | 0.61 (0.03) | 0.54 (0.03) |

to horizons $M_n = 6, 9, \ldots, 48$ when the true variables were accounted for. Thus, $M_n/n$ ranged from 0.02 to $-0.2$. Two hundred parametric bootstrap samples consisting of 252 observations each were created to mimic the original sample and the considered selection criteria were employed to choose subset of potential $M_n$ variables. Figure 5 and Table 6 present the results in the same format as before, outcomes in the table are estimated probabilities of correct selection and they are shown for the chosen horizon $M_n = 18$. The results are similar to that of simulation experiments indicating that $\widetilde{GIC}$ with $c_{n,j} = j$ and $a_n = \log(n)$ performs the best in this case, and the second best is $\widetilde{GIC}$ with $c_{n,j} = 0$ and $a_n = \log(n)$.

## Appendix

For brevity, we assume in the proofs that $\sigma^2 = 1$. Let $J_j = \{i_1, \ldots, i_j\}$ denote the set of indexes of covariates in the given model $\mathcal{M}_{i_1,\ldots,i_j}$, where $1 \le j \le M_n$. Recall that $\Gamma$ is a set of indexes of non zero coefficients in the true model and $j_0$ denotes the cardinality of $\Gamma$. $RSS(\mathcal{M}_{i_1,\ldots,i_j})$ and $\widetilde{GIC}(\mathcal{M}_{i_1,\ldots,i_j})$ will be denoted by $RSS(J_j)$ and $\widetilde{GIC}(J_j)$, respectively.

### *A.1 Proof of Theorem* 2.1

The proof proceeds in two steps. Consider first the case $J_j \supset \Gamma$, where $\supset$ denotes strict inclusion. Recall that assumption (A1.3) implies that for sufficiently large $n$ $\Gamma \subseteq \{1, 2, \ldots, M_n\}$. We will prove that $P[\widetilde{GIC}(\Gamma) > \min_{J_j \supset \Gamma} \widetilde{GIC}(J_j)] \to 0$ as $n \to \infty$. For $j > j_0$ we have

$$\widetilde{GIC}(J_j) - \widetilde{GIC}(\Gamma) \geq n \log \frac{RSS(J_{M_n})}{n} - n \log \frac{RSS(\Gamma)}{n - c_{n,j_o}} + a_n.$$

Let $\bar{n} = (n - c_{n,j_0})/n$ and $\bar{b}_n = n[\bar{n} \exp(a_n/n) - 1]$, which can be written as

$$\bar{b}_n = a_n \left\{ -\frac{c_{n,j_0}}{a_n} + \bar{n} \left[ 1 + \frac{a_n}{n} + o\left(\frac{a_n}{n}\right) \right] \right\}.$$

The assumptions (A1.1), (A1.2), and $c_{n,j_0} \leq M_n$ yield $\bar{b}_n / M_n \to \infty$. Thus,

$$\begin{aligned}
&P[\widetilde{GIC}(\Gamma) > \min_{J_j \supset \Gamma} \widetilde{GIC}(J_j)] \\
&\leq P\left[ n \log \frac{RSS(J_{M_n})}{n} - n \log \frac{RSS(\Gamma)}{n - c_{n,j_o}} + a_n < 0 \right] = P\left[ \frac{RSS(\Gamma)}{RSS(J_{M_n})} > \bar{n} \exp\left(\frac{a_n}{n}\right) \right] \\
&= P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(J_{M_n}) - \mathbf{Q}(\Gamma)]\boldsymbol{\varepsilon} > \bar{b}_n n^{-1} \boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{Q}(J_{M_n})]\boldsymbol{\varepsilon}\} \\
&\leq P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(J_{M_n}) - \mathbf{Q}(\Gamma)]\boldsymbol{\varepsilon} > \bar{b}_n n^{-1}(n - M_n - d_n)\} \\
&\quad + P\{\boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{Q}(J_{M_n})]\boldsymbol{\varepsilon} \leq n - M_n - d_n\},
\end{aligned}$$

where $\mathbf{Q}(J_j)$ denotes projection on the column space spanned by the regressors corresponding to coefficients in a given model $J_j$ and $d_n = (n - M_n)^{(1+\delta)/2}$, for some $\delta \in (0, 1)$. Assumption (A1.6) implies that $\mathbf{X}'\mathbf{X}$ has rank $M_n$ with probability tending to 1 and we can assume without loss of generality that $\mathbf{X}'\mathbf{X}$ is invertible (see the proof of Theorem 2.2 in Zheng and Loh, 1997). Then it follows that $\boldsymbol{\varepsilon}'[\mathbf{Q}(J_{M_n}) - \mathbf{Q}(\Gamma)]\boldsymbol{\varepsilon} \sim \chi^2_{M_n - j_0}$ and $\boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{Q}(J_{M_n})]\boldsymbol{\varepsilon} \sim \chi^2_{n - M_n}$ (since $\sigma^2 = 1$). By an inequality for cumulative distribution function of a chi-square distribution,

$$P(\chi^2_k \leq k - \delta_0) \leq \exp\{-(4k)^{-1}\delta_0^2\},$$

for $\delta_0 > 0$ (see Shibata, 1981). Thus, we have

$$P\{\boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{Q}(J_{M_n})]\boldsymbol{\varepsilon} \leq n - M_n - d_n\} \leq \exp\left[ -\frac{d_n^2}{4(n - M_n)} \right] \to 0,$$

as $n \to \infty$, since $M_n / n \to 0$. Moreover, by Markov inequality as $\boldsymbol{\varepsilon}'[\mathbf{Q}(J_{M_n}) - \mathbf{Q}(\Gamma)]\boldsymbol{\varepsilon}$ is non negative

$$P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(J_{M_n}) - \mathbf{Q}(J_{j_0})]\boldsymbol{\varepsilon} > \bar{b}_n n^{-1}(n - M_n - d_n)\} \leq \frac{(M_n - j_0)n}{\bar{b}_n(n - M_n - d_n)} \to 0,$$

as $n \to \infty$, since $\bar{b}_n / M_n \to \infty$ and $(n - M_n - d_n)/n \to 1$. This completes the first part of the proof.

It remains to show that $P[\widetilde{GIC}(\Gamma) > \min_{J_j \not\supseteq \Gamma} \widetilde{GIC}(J_j)] \to 0$ as $n \to \infty$. Let $A_n = n \log \frac{RSS(J_{M_n})}{n} + a_n(j_0 + 1)$. Note that $A_n \leq \min_{J_j \supset \Gamma} \widetilde{GIC}(J_j)$. Since

$$P\left[\widetilde{GIC}(\Gamma) > \min_{J_j \not\supseteq \Gamma} \widetilde{GIC}(J_j)\right] \leq P[\widetilde{GIC}(\Gamma) > A_n] + P\left[A_n > \min_{J_j \not\supseteq \Gamma} \widetilde{GIC}(J_j)\right]$$

it suffices to show that $P[A_n > \min_{J_j \not\supseteq \Gamma} \widetilde{GIC}(J_j)] \to 0$ as convergence $P[\widetilde{GIC}(\Gamma) > A_n] \to 0$ follows from the first part of the proof. Let $i^* \in \Gamma$ be the index such that $RSS(J_{M_n} - \{i^*\}) = \min_i RSS(J_{M_n} - \{i\})$. For $J_j \not\supseteq \Gamma$ we have

$$\widetilde{GIC}(J_j) \geq n \log \frac{RSS(J_{M_n} - \{i^*\})}{n} + a_n.$$

Thus,

$$P\left[A_n > \min_{J_j \not\supseteq \Gamma} \widetilde{GIC}(J_j)\right] \leq P\left[\log \frac{RSS(J_{M_n} - \{i^*\})}{RSS(J_{M_n})} < \frac{a_n}{n}\right].$$

Noting that

$$\frac{RSS(J_{M_n} - \{i^*\})}{RSS(J_{M_n})} = \frac{T_{i^*}^2}{n - M_n} + 1,$$

where $T_i^2$ is a t-statistic corresponding to $\beta_{i^*}$, we obtain that

$$P\left[\log \frac{RSS(J_{M_n} - \{i^*\})}{RSS(J_{M_n})} < \frac{a_n}{n}\right] \leq P\left[T_{i^*}^2 < (n - M_n)(\exp(a_n/n) - 1)\right]$$
$$\leq P\left(\min_{i \in \Gamma} T_i^2 < (n - M_n)(\exp(a_n/n) - 1)\right).$$

Since $\exp(a_n/n) - 1 = a_n/n + o(a_n/n)$ it sufficies to show that $P[\min_{i \in \Gamma} T_i^2 < a_n] \to 0$. This follows from the proof of Theorem 2.2 in Zheng and Loh (1997) who proved that under conditions of Theorem 2.1 $P[\min_{i \in \Gamma} \widehat{\sigma}^2 T_i^2 < n^{1/(1+\eta)}] \to 0$, $\eta > 0$. Now the required convergence follows from assumptions (A1.1), (A1.2), and the fact that $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$.

*Proof of Remark* 2.1. We will use the following easily verifiable lemma.

**Lemma A.1.** *Let* $\mathbf{K} = (K_1, \ldots, K_n)'$ *be the random vector of identically distributed random variables such that* $\mathbf{E}(K) = 0$ *and the covariance matrix* $\Sigma_{\mathbf{K}} = \mathbf{I}$. *Assume* $\mathbf{E}(K_1^4) < \infty$. *Let* $P$ *be a symmetric, idempotent matrix. Then,*

$$Var(\mathbf{K}' P \mathbf{K}) = [\mathbf{E}(K_1^4) - 3] \sum_{i=1}^{n} P_{i,i}^2 + 2tr(P).$$

In order to prove the remark we assume, without loss of generality, that columns of $\mathbf{X}$ have mean zero. The first part of the proof, for $J_j \supset \Gamma$ proceeds the same

as in the proof of Theorem 2.1. It sufficies to consider the case $J_j \not\supseteq \Gamma$. Reasoning analogously to the proof of Theorem 2.1 we have that, in order to show $P[\widetilde{GIC}(\Gamma) > \min_{J_j \not\supseteq \Gamma} \widetilde{GIC}(J_j)] \to 0$, it suffices to prove

$$P\left[\log \frac{RSS(J_{M_n} - \{i^*\})}{RSS(J_{M_n})} < \frac{a_n}{n}\right] \to 0,$$

as $n \to \infty$. As $RSS(J_{M_n}) = \boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{Q}(J_{M_n}))\boldsymbol{\varepsilon}$ given $\mathbf{X}$ has chi-square distribution with $n - M_n$ degrees of freedom, it is easy to see that $n^{-1}RSS(J_{M_n}) \xrightarrow{P} \sigma^2$. Since $a_n/n \to 0$ it sufficies to show that $n^{-1}RSS(J_{M_n} - \{i^*\}) \to \sigma^2 + \lambda$, $\lambda > 0$. Let $c_i(X)$ denote the $i$th column vector of the matrix $\mathbf{X}$, for $i = 1, \ldots, M_n$. Define

$$\Lambda_{n,M_n} = n^{-1}(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(J_{M_n} - \{i^*\})](\mathbf{X}\boldsymbol{\beta}) = n^{-1}[c_{i^*}(\mathbf{X})]'[\mathbf{I} - \mathbf{Q}(J_{M_n} - \{i^*\})][c_{i^*}(\mathbf{X})].$$

Note that $\mathbf{E}(\Lambda_{n,M_n}) \to 1$, as $n \to \infty$. Under assumptions that the coordinates $X_{i,1}, \ldots, X_{i,M_n}$ are i.i.d. and $\mathbf{E}X_{1,1}^4 < \infty$ it follows from Lemma A.1 that $Var(\Lambda_{n,M_n}) \to 0$ and thus $\Lambda_{n,M_n} \xrightarrow{P} \lambda = 1$. We have the folllowing decomposition:

$$\begin{aligned} n^{-1}RSS(J_{M_n} - \{i^*\}) = {} & n^{-1}\boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{Q}(J_{M_n} - \{i^*\})]\boldsymbol{\varepsilon} \\ & + n^{-1}2(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(J_{M_n} - \{i^*\}))]\boldsymbol{\varepsilon} + \Lambda_{n,M_n}. \end{aligned}$$

The first summand converges in probability to $\sigma^2$. Provided that $\mathbf{X}'\mathbf{X}$ is invertible, $n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{Q}(J_{M_n} - \{i^*\}))\boldsymbol{\varepsilon}$ given $\mathbf{X}$ has $N(0, v_n)$ distribution, where $v_n = n^{-1}\Lambda_{n,j} \xrightarrow{P} 0$. Thus, $n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \mathbf{Q}(J_{M_n} - \{i^*\}))\boldsymbol{\varepsilon} \xrightarrow{P} 0$. This completes the proof of the remark.

*Proof of Theorem* 2.2. The first part of the proof, for $J_j \supset J_{j_{\max}}$, proceeds the same as the first part of the proof of Theorem 2.1. It suffices to consider the case $j < j_{\max}$. Define $\Lambda_{n,j} = n^{-1}(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(J_j)](\mathbf{X}\boldsymbol{\beta}) > 0$. Let $\mathbf{D}_j$ be a $M_n \times j$ matrix of zeros and ones such that $\mathbf{X}\mathbf{D}_j$ consists of the first $j$ columns of $\mathbf{X}$. By assumption (A1.5') and the fact that $\mathbf{X}\boldsymbol{\beta} = (\mathbf{X}\mathbf{D}_{j_{\max}})\bar{\boldsymbol{\beta}}$ where $\bar{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_{j_{\max}})'$, we have $\Lambda_{n,j} \xrightarrow{P} \lambda > 0$ as $n \to \infty$. The assertion follows from the fact that for $j < j_{\max}$

$$n^{-1}(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(J_j)](\mathbf{X}\boldsymbol{\beta}) = n^{-1}\bar{\boldsymbol{\beta}}'\mathbf{A}\bar{\boldsymbol{\beta}}, \tag{11}$$

where

$$\begin{aligned} \mathbf{A} = {} & [(\mathbf{X}\mathbf{D}_{j_{\max}})'(\mathbf{X}\mathbf{D}_{j_{\max}})] \\ & - [(\mathbf{X}\mathbf{D}_{j_{\max}})'(\mathbf{X}\mathbf{D}_{j_{\max}})]\bar{\mathbf{D}}_j[\bar{\mathbf{D}}_j'(\mathbf{X}\mathbf{D}_{j_{\max}})'(\mathbf{X}\mathbf{D}_{j_{\max}})\bar{\mathbf{D}}_j]^{-1}\bar{\mathbf{D}}_j'[(\mathbf{X}\mathbf{D}_{j_{\max}})'(\mathbf{X}\mathbf{D}_{j_{\max}})] \end{aligned}$$

and $\bar{\mathbf{D}}_j$ is a $j_{\max} \times j$ matrix such that $\mathbf{X}\mathbf{D}_j = (\mathbf{X}\mathbf{D}_{j_{\max}})\bar{\mathbf{D}}_j$. Matrix $\mathbf{W} = \mathbf{E}(\mathbf{X}_{1,j_{\max}}\mathbf{X}_{1,j_{\max}}')$, defined in (A1.5') is a positive definite matrix. Thus it can be decomposed as $\mathbf{W} = \mathbf{W}^{1/2}\mathbf{W}^{1/2}$ where $\mathbf{W}^{1/2} = \mathbf{U}\Xi^{1/2}\mathbf{U}'$, $\mathbf{U}$ is an orthogonal matrix

and $\Xi$ is a diagonal matrix with positive diagonal. The right-hand side of (11) converges in probability to

$$\begin{aligned}\lambda &= \bar{\boldsymbol{\beta}}'[\mathbf{W} - \mathbf{W}\bar{\mathbf{D}}_j(\bar{\mathbf{D}}_j'\mathbf{W}\bar{\mathbf{D}}_j)^{-1}\bar{\mathbf{D}}_j'\mathbf{W}]\bar{\boldsymbol{\beta}} \\ &= (\mathbf{W}^{1/2}\bar{\boldsymbol{\beta}})'[\mathbf{I} - \mathbf{W}^{1/2}\bar{\mathbf{D}}_j(\bar{\mathbf{D}}_j'\mathbf{W}\bar{\mathbf{D}}_j)^{-1}\bar{\mathbf{D}}_j'(\mathbf{W}^{1/2})']\mathbf{W}^{1/2}\bar{\boldsymbol{\beta}} > 0\end{aligned}$$

since the columns of $\mathbf{W}^{1/2}$ are linearly independent. Let $p = j_{\max} - j$ and $\tilde{n} = (n - c_{n,j})/(n - c_{n,j_{\max}})$. For $j \le j_{\max}$ we have

$$\begin{aligned}&P\{[\widetilde{GIC}(J_j) - \widetilde{GIC}(J_{j_{\max}})] < 0\} \\ &\quad \le P\left[n\log\frac{RSS(J_j)}{n - c_{n,j}} - n\log\frac{RSS(J_{j_{\max}})}{n - c_{n,j_{\max}}} < a_n p\right] = P\left[\frac{RSS(J_j)}{RSS(J_{j_{\max}})} < \tilde{n}\exp\left(\frac{a_n p}{n}\right)\right].\end{aligned}$$

As $RSS(J_{j_{\max}}) = \boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{Q}(J_{j_{\max}}))\boldsymbol{\varepsilon}$ given $\mathbf{X}$ has chi-square distribution with $M_n - j_{\max}$ degrees of freedom, it is easy to see that $n^{-1}RSS(J_{j_{\max}}) \xrightarrow{P} \sigma^2$. Note that $\tilde{n}\exp\left(\frac{a_n p}{n}\right) \to 1$ as $n \to \infty$. In order to prove

$$P\left[\frac{RSS(J_j)}{RSS(J_{j_{\max}})} < \tilde{n}\exp\left(\frac{a_n p}{n}\right)\right] \to 0$$

it sufficies to show that $n^{-1}RSS(J_j) \xrightarrow{P} \sigma^2 + \lambda$, $\lambda > 0$. This is shown analogously to the proof of the Remark 2.1. This completes the proof since

$$P\left[\widetilde{GIC}(J_{j_{\max}}) > \min_{j < j_{\max}} \widetilde{GIC}(J_j)\right] \le \sum_{j=0}^{j_{\max}-1} P\{[\widetilde{GIC}(J_j) - \widetilde{GIC}(J_{j_{\max}})] < 0\}.$$

*Proof of Theorem* 2.6. The following fact will be useful in the proof of Theorem 2.6.

**Lemma A.2.** *Under conditions of Theorem* 2.6 $\boldsymbol{\varepsilon}'\mathbf{Q}(J_{M_n})\boldsymbol{\varepsilon} = O_P(M_n)$.

The proof of the above lemma follows from the proof of Theorem 2.3 in Zheng and Loh (1997). We note, in particular, that Lemma A.2 implies consistency of $\widehat{\sigma}^2$ provided that $M_n/n \to 0$.

In order to prove Theorem 2.6 we first show that $P[\widetilde{GIC}(J_{j_{\max}}) > \min_{J_j \supset J_{j_{\max}}} \widetilde{GIC}(J_j)] \to 0$, as $n \to \infty$. Reasoning analogously to the proof of Theorem 2.1, we have for $j > j_{\max}$

$$\begin{aligned}&P\left[\widetilde{GIC}(J_{j_{\max}}) > \min_{J_j \supset J_{j_{\max}}} \widetilde{GIC}(J_j)\right] \\ &\quad \le P\{\boldsymbol{\varepsilon}'[\mathbf{Q}(J_{M_n}) - \mathbf{Q}(J_{j_{\max}})]\boldsymbol{\varepsilon} > \bar{b}_n n^{-1}(n - M_n - d_n)\} \\ &\qquad + P\{\boldsymbol{\varepsilon}'[\mathbf{I} - \mathbf{Q}(J_{M_n})]\boldsymbol{\varepsilon} \le n - M_n - d_n\},\end{aligned}$$

Lemma A.2 and the fact that $\boldsymbol{\varepsilon}'\mathbf{Q}(J_{j_{\max}})\boldsymbol{\varepsilon} \ge 0$, $\bar{b}_n/M_n \to \infty$ and $(n - M_n - d_n)/n \to 1$ imply that the first probability tends to zero. The second probability can be bounded

from above by

$$P(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \le n - M_n - d_n + c_n) + P(\boldsymbol{\varepsilon}'\mathbf{Q}(J_{M_n})\boldsymbol{\varepsilon} \ge c_n),$$

where $d_n = (n - M_n)^{(1+\delta)/2}$, for some $\delta \in (0, 1)$ and $c_n = d_n/2$. As in the proof of Theorem 2.1, using an inequality for cumulative distribution function of a chi-square distribution, we have

$$P\{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \le n - M_n - d_n + c_n\} \le \exp\left[-\frac{(M_n + d_n - c_n)^2}{4n}\right] \to 0,$$

as $n \to \infty$. From Lemma A.2 and the fact that $M_n/c_n \to 0$, we obtain $P(\boldsymbol{\varepsilon}'\mathbf{Q}(J_{M_n})\boldsymbol{\varepsilon} \ge c_n) \to 0$ which completes the first part of the proof.

It remains to show that $P[\widetilde{GIC}(J_{j_{\max}}) > \min_{j<j_{\max}} \widetilde{GIC}(J_j)] \to 0$. Let $p = j_{\max} - j$. For $j < j_{\max}$ we have, as in the proof of Theorem 2.1 with $\tilde{\sigma}^2(J_j) = RSS(J_j)/n$,

$$P\left[\widetilde{GIC}(J_{j_{\max}}) > \min_{j<j_{\max}} \widetilde{GIC}(J_j)\right] \le \sum_{j=0}^{j_{\max}-1} P\{[\widetilde{GIC}(J_j) - \widetilde{GIC}(J_{j_{\max}})] < 0\}$$

$$\le \sum_{j=0}^{j_{\max}-1} P\left[\log\frac{\tilde{\sigma}^2(J_j)}{\tilde{\sigma}^2(J_{j_{\max}})} \le \frac{a_n p}{n} - \log\left(\frac{n - c_{n,j_{\max}}}{n - c_{n,j}}\right)\right].$$

In order to prove that the above probability tends to zero we will show that

$$\liminf_n \log\frac{\tilde{\sigma}^2(J_j)}{\tilde{\sigma}^2(J_{j_{\max}})} > 0 \quad a.s. \tag{12}$$

For $j < j_{\max}$ we have

$$n[\tilde{\sigma}^2(J_j) - \tilde{\sigma}^2(J_{j_{\max}})] = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{Q}(J_j)\mathbf{X}\boldsymbol{\beta}\|^2 + 2(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(J_j)]\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}'\mathbf{Q}(J_{j_{\max}})\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\mathbf{Q}(J_j)\boldsymbol{\varepsilon}.$$

It follows from Lai and Wei (1982), Theorems 2.4 and 2.3 that under assumption $\sup_i \mathbf{E}(|\varepsilon_i^{\alpha}|) < \infty$, for some $\alpha > 2$ we have $2(\mathbf{X}\boldsymbol{\beta})'[\mathbf{I} - \mathbf{Q}(J_j)]\boldsymbol{\varepsilon} = o(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{Q}(J_j)\mathbf{X}\boldsymbol{\beta}\|^2)$ and $\boldsymbol{\varepsilon}'\mathbf{Q}(J_j)\boldsymbol{\varepsilon} = o(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{Q}(J_j)\mathbf{X}\boldsymbol{\beta}\|^2)$. Since $\boldsymbol{\varepsilon}'\mathbf{Q}(J_{j_{\max}})\boldsymbol{\varepsilon} > 0$, we have $n[\tilde{\sigma}^2(J_j) - \tilde{\sigma}^2(J_{j_{\max}})] \ge \|\mathbf{X}\boldsymbol{\beta} - \mathbf{Q}(J_j)\mathbf{X}\boldsymbol{\beta}\|^2 + o(\|\mathbf{X}\boldsymbol{\beta} - \mathbf{Q}(J_j)\mathbf{X}\boldsymbol{\beta}\|^2)$. From the proof of Theorem 3.1 in Pötscher (1989) we get $\liminf_n n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{Q}(J_j)\mathbf{X}\boldsymbol{\beta}\|^2 > 0$ and thus $\liminf_n[\tilde{\sigma}^2(J_j) - \tilde{\sigma}^2(J_{j_{\max}})] > 0$ which, together with $\tilde{\sigma}^2(J_{j_{\max}}) \xrightarrow{P} \sigma^2$, implies (12).

## References


Akaike, H. (1970). Statistical predictor identification. *Ann. Instit. Statist. Math.* 22:203–217.

Davison, A., Hinkley, D. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

Hannan, W. J., Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. B* 41:190–195.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* 16(1):342–355.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *J. Statist. Educat.* 4(1).

Lai, T. L., Wei, C. Z. (1982). Asymptotic properties of projections with applications to stochastic regression problems. *J. Multivariate Anal.* 12:346–370.

Mallows, C. L. (1973). Some comments on Cp. *Technometrics* 15:661–675.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* 21:255–285.

Pötscher, B. M. (1989). Model selection under nonstationarity: autoregressive models and stochastic linear regression models. *Ann. Statist.* 17(3):1257–1274.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6:461–464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* 88: 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* 7:221–264.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* 63:117–126.

Zhang, P. (1992). On the distributional properties of model selection criteria. *J. Amer. Statist. Assoc.* 21:291–313.

Zheng, X., Loh, W. -Y. (1995). Consistent variable selection in linear models. *J. Amer. Statist. Assoc.* 90:151–156.

Zheng, X., Loh, W. -Y. (1997). A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica* 7:311–325.