Open Journal of Statistics, **2012, 2, 153-162**
http://dx.doi.org/10.4236/ojs.2012.22017 Published Online April 2012 (http://www.SciRP.org/journal/ojs)

A Universal Selection Method in Linear Regression Models

Eckhard Liebscher
Department of Computer Science and Communication Systems, Merseburg University of Applied Sciences, Merseburg, Germany
Email: eckhard.liebscher@hs-merseburg.de

Received January 27, 2012; revised February 29, 2012; accepted March 9, 2012

ABSTRACT

In this paper we consider a linear regression model with fixed design. A new rule for the selection of a relevant submodel is introduced on the basis of parameter tests. One particular feature of the rule is that subjective grading of the model complexity can be incorporated. We provide bounds for the mis-selection error. Simulations show that by using the proposed selection rule, the mis-selection error can be controlled uniformly.

Keywords: Linear Regression; Model Selection; Multiple Tests

1. Introduction

In this paper we consider a linear regression model with fixed design and deal with the problem of how to select a model from a family of models which fits the data well. The restriction to linear models is done for the sake of transparency. In applications the analyst is very often interested in simple models because these can be interpreted more easily. Thus a more precise formulation of our goal is to find the simplest model which fits the data reasonably well. We establish a principle for selecting this "best" model.

Over time the problem of model selection has been studied by a large number of authors. The papers [1,2] by Akaike and Mallows inspired statisticians to think about the comparisons of fitted models to a given dataset. Akaike, Mallows and later Schwarz (in [3]) developed information criteria which may be used for comparisons and in particular, may be applied to non-nested sets of models. The basic idea is the assessment of the trade-off between the improved fit of a larger model and the increased number of parameters. Akaike's approach is to penalise the maximised log-likelihood by twice the number of parameters in the model. The resulted quantity, the so called AIC, is maximised with respect to the parameters and the models. The disadvantage of this procedure is that it is not consistent; more precisely, the probability of overfitting the model tends to a positive value. Subsequently, a lot of other criteria have been developed. In a series of papers the consistency of procedures based on several information criteria (BIC, GIC, MDL, for example) are shown. The MDL-method was introduced by Rissanen in [4]. In the nineties of the last century a new class of model selection methods came into focus. The FDR procedure of Benjamini and Hochberg (see [5]) uses ideas from multiple testing and attempts to control the false discovery rate, which we will call the mis-selection rate in this paper. More recent papers of this direction are published by Bunea *et al.* [6], and by Benjamini and Gavrilov [7]. Surveys of the theory and existing results may be found in [8-11]. In a large number of papers the consistency and loss efficiency of the selection procedure is shown and the signal to noise ratio is calculated for the criterion under consideration. Among these papers we refer to [12-16], where consistency is proved in a rather general framework. A method for the submodel selection using graphs is studied in [17]. Leeb and Pötscher examine several aspects of the post-model-selection inference in [9,18,19]. The authors point out and illustrate the important distinction between asymptotic results and the small sample performance. Shao introduced in [20] a generalised information criterion, which includes many popular criteria or which is asymptotically equivalent to them. In this paper Shao proved convergence rates for the probability of mis-selection. In [21] a rather general approach using a penalised maximum likelihood criterion was considered for nested models.

Edwards and Havránek proposed in [22] a selection procedure aimed at finding sets of simplest models that are accepted by a test like a goodness-of-fit test. Unfortunately, it is not possible to use the typical statistical tests of linear models in Edwards and Havránek's procedure since the assumption (b) in the Section 2 of their paper is not fulfilled (cf. Section 4 of their paper).

In this paper we develop a new universal method for selecting a significant submodel from a linear regression model with fixed design, where the selection is done

from the whole set of all submodels. We point out the several new features of our approach:

1) A new selection procedure based on parameter tests is introduced. The procedure is not comparable with methods based on information criteria and it is different from Efroymson's algorithm of stepwise variable selection in [23].

2) We derive convergence rates for the probability of mis-selection which are better than those proved in papers about information criteria e.g. in [20].

3) Subjective grading of the model complexity can be incorporated.

Concerning 1) we consider tests on a set of parameters in contrast to FDR-methods, where several tests on only one parameter are applied. Moreover w.r.t. 2), many authors do not analyse the behaviour of mis-selection probabilities. The results on bounds or convergence rates of these probabilities are more informative than the consistency. The aspect 3) is of special interest from the point of view of model building. Typically model builder have some preference rules in mind when selecting the model. They prefer simple models with linear functions to models with more complex functions (exponential or logarithmic, for example). The crucial idea is to assign to each submodel a specific complexity number.

We do not assume that the errors are normally distributed. This ensures a wide-ranging applicability of the approach, but only asymptotic distributions of test statistics are available. From examples in Section 2, it can be seen that applications are possible in several directions, for instance to the one-factor-ANOVA model. The simulations show an advantage of the proposed method in that it controls the frequency of mis-selection uniformly. For models with a large number of regressors, the problem of establishing an effective selection algorithm is not discussed in this paper; we refer to the paper [24].

The paper is organised as follows: In Section 2 we introduce the regression model and several versions of submodels. The asymptotic behaviour of the basic statistic is also studied there. Section 3 is devoted to the model selection method. We provide convergence rates for the probability that the procedure selects the wrong model (mis-selection). We see that the behaviour is similar to that in the case of hypothesis testing. The results of simulations are discussed in Section 4. The reader finds the proofs in Section 5.

2. Models

Let us introduce the master model

$$Y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \quad \text{for} \quad i = 1,\cdots,n,$$

where $X = \left(x_{ij}\right)_{i=1,\cdots,n,\, j=1,\cdots,k} \in \mathbb{R}^{n\times k}$ is the design matrix, $\beta = \left(\beta_1,\cdots,\beta_k\right)^T \in \mathbb{R}^k$ is the parameter vector, and $\varepsilon_1,\cdots,\varepsilon_n$ are independent random variables. Assume that $\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\left|\varepsilon_i\right|^p < +\infty$ for some $p > 2$, and $\mathrm{Var}\left(\varepsilon_i\right) = \sigma^2$. $Y_1,\cdots,Y_n$ denote the values of the response variable. In short we can write

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

where $Y = \left(Y_1,\cdots,Y_n\right)^T$, $\varepsilon = \left(\varepsilon_1,\cdots,\varepsilon_n\right)^T$. The least square estimator for β is given by

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y.$$

This leads to the residual sum of squares

$$R_n = \left\|Y - X\hat{\beta}\right\|^2 = Y^T\left(I - X\left(X^T X\right)^{-1} X^T\right)Y,$$

where $\|.\|$ is the Euclidean vector norm.

The aim is to select model (1) or an appropriate submodel which fits the data well. Moreover, we search for a reasonably simple model. In the following we define the submodels of (1). The submodel with index $\nu \in \{1,\cdots,\bar{\nu}\}$ has the parameter vector $\gamma = \left(\gamma_1,\gamma_2,\cdots,\gamma_l\right)^T \in \mathbb{R}^l$, $l =: l(\nu)$, where the vector γ is related to β by $\beta = D_\nu\gamma$ with an appropriate matrix $D_\nu \in \mathbb{R}^{k\times l}$ having maximum rank $l < k$. In a large number of applications, the γ_i's coincide with different components of β. The submodel indices 1 and $\bar{\nu}$ correspond to the model function equal to zero (no parameters) and to the full model, respectively. Thus we can write the model equation for the submodel ν as

$$Y = X_\nu\gamma + \varepsilon, \tag{2}$$

where $X_\nu = XD_\nu$. The parameter space of submodel ν in (1) is given by $\Theta_\nu = \left\{D_\nu\gamma : \gamma \in \mathbb{R}^l\right\}$. Next we give several versions for the definition of submodels in different situations.

Example 1. We consider all submodels, where components of β are zero. More precisely, index ν is assigned to a submodel if $\gamma_1 = \beta_{i_1},\cdots,\gamma_l = \beta_{i_l}$ are the parameters of the submodel ($i_1 < i_2 < \cdots < i_l$), $\beta_j = 0$ for $j \notin J_\nu := \{i_1,\cdots,i_l\}$ and $\nu = 1 + \sum_{j=1}^{l} 2^{i_j - 1}$. Let $e_i = \left(0,\cdots,0,1_i,0,\cdots\right)^T \in \mathbb{R}^k$ be the *i*-th unit vector. Then $D_\nu = \left(e_{i_1}, e_{i_2},\cdots,e_{i_l}\right) \in \mathbb{R}^{k\times l}$, and $\Theta_\nu = \left\{\beta : \beta_j = 0 \text{ for all } j \notin J_\nu\right\}$. For example, for $k = 5$, the submodel with index $\nu = 14$ has the parameters $\gamma_1 = \beta_1$, $\gamma_2 = \beta_3$, $\gamma_3 = \beta_4$ and $\beta_2 = \beta_5 = 0$ holds. Moreover, we have

$$\boldsymbol{D_\nu} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \Theta_\nu = \left\{\beta \in \mathbb{R}^5 : \beta_2 = \beta_5 = 0\right\}$$

in this case. Here the digits "1" in the binary representation of $\nu-1$ give the indices of the parameters β_j available in the submodel ν. X_ν in (2) consists of the columns $i_1,\cdots,i_l$ of the design matrix X corresponding to the present parameters in submodel ν. □

Example 2. Let $k=3$. submodel 1: $\beta_1=0$, $\gamma=(\beta_2,\beta_3)^T$. Submodel 2: $\beta_1=1$, $\gamma=(\beta_2,\beta_3)^T$. Submodel 3: identity (1). □

Example 3. We consider the one-factor ANOVA model

$$Y_{ij}=\mu_i+\varepsilon_{ij} \quad \text{for} \quad i=1,\cdots,g, \quad j=1,\cdots,n_g,$$

where $Y=\left(Y_{11},Y_{12},\cdots,Y_{1n_1},Y_{21},\cdots,Y_{gn_g}\right)^T\in\mathbb{R}^n$, $n=\sum_{i=1}^g n_i$, $\beta=\left(\mu_1,\cdots,\mu_g\right)^T\in\mathbb{R}^k$ is the parameter vector. $\varepsilon_{11},\cdots,\varepsilon_{gn_g}$ are independent random variables. The submodels are characterised by the fact that several μ_i's are equal. Let $\bar{\nu}$ be the k-th Bell number. A Submodel with index $\nu\in\{1,\cdots,\bar{\nu}\}$ is determined by a partition $J_{\nu 1},\cdots,J_{\nu,l(\nu)}$ of $\{1,\cdots,g\}$ in the following way: $\Theta_\nu=\{\beta:\mu_j=\mu_k \text{ if } j,k\in J_{\nu i} \text{ for some } i\}$. The submodel with index ν has $l(\nu)$ parameters. Furthermore,

$$D_\nu=\left(d_{ij}\right)_{i=1,\cdots,k,\ j=1,\cdots,l(\nu)} \text{ holds},$$

$$\text{where } d_{ij}=\begin{cases}1 \text{ for } i\in J_{\nu j},\\ 0 \text{ otherwise.}\end{cases} \quad \square$$

Example 3 shows that the model selection problem occurs also in the context of ANOVA. In submodel (2) with index ν, the least square estimator $\hat{\gamma}_\nu$ and the residual sum of squares S_ν are given by

$$\begin{aligned}\hat{\gamma}_\nu&=\left(X_\nu^TX_\nu\right)^{-1}X_\nu^TY,\\ S_\nu&=\left\|Y-X_\nu\hat{\gamma}_\nu\right\|^2=Y^T\left(I-X_\nu\left(X_\nu^TX_\nu\right)^{-1}X_\nu^T\right)Y.\end{aligned}\tag{3}$$

What is an appropriate statistic for model selection? Let $M_n(\nu):=S_\nu-R_n$. Here we consider a quantity $\bar{M}_n(\nu)$, which is similar to F-statistics known from hypothesis testing in linear regression models with normal errors:

$$\begin{aligned}\bar{M}_n(\nu)&=\frac{M_n(\nu)}{\frac{1}{n-l(\nu)}S_\nu}\\ &=\frac{\left(\hat{\beta}-D_\nu\hat{\gamma}_\nu\right)^TX^TX\left(\hat{\beta}-D_\nu\hat{\gamma}_\nu\right)}{\frac{1}{n-l(\nu)}S_\nu}\end{aligned}\tag{4}$$

$$\text{for } \nu<\bar{\nu},\ \bar{M}_n(\bar{\nu})=0.$$

The main difference to classical F-statistics is that the estimator $\frac{1}{n-l(\nu)}S_\nu$ of the model variance in submodel ν appears in the denominator. The quantity $\frac{1}{n-l(\nu)}S_\nu$ is the proper estimator under the hypothesis of submodel ν. Classical F-statistics are used in Efroymson's algorithm of stepwise variable selection (see [23]).

In the remainder of this section we study the asymptotic behaviour of the statistic $\bar{M}_n(\nu)$ when β_0 is the true parameter of the model (1). For this reason, we first introduce some assumptions.

Assumption $\mathcal{A}$**.** Let $G_n=\frac{1}{n}X^TX$. Assume that

$$\operatorname{Rank}(G_n)=k,$$

$$\lim_{n\to\infty}G_n=\Gamma, \quad \Gamma \text{ regular.}$$

Moreover,

$$B_{np}:=\max_{j=1,\cdots,k}\sum_{i=1}^n\left|x_{ij}\right|^p=o\left(n^{p/2}\right). \quad \square$$

In a wide range of applications, the entries x_{ij} of the design matrix are uniformly bounded such that $B_{np}=O(n)=o\left(n^{p/2}\right)$ since $p>2$. The Assumption $\mathcal{A}$ may be weakened in some ways, but we use this assumption to reduce the technical effort. We introduce

$$\Gamma_\nu:=D_\nu^T\Gamma D_\nu\in\mathbb{R}^{l(\nu)\times l(\nu)} \text{ and}$$

$$K_\nu:=\beta_0^T\left(I-\Gamma D_\nu\Gamma_\nu^{-1}D_\nu^T\right)\Gamma\beta_0\in\mathbb{R}.$$

Proposition 2.1 clarifies the asymptotic behaviour of the statistic $\bar{M}_n(\nu)$.

Proposition 2.1. Assume that Assumption $\mathcal{A}$ is satisfied.

1) Assume that $\beta_0\in\Theta_\nu$ and $l(\nu)<k$. Then we have

$$\bar{M}_n(\nu)\xrightarrow{\mathcal{D}}\chi^2_{k-l(\nu)}.$$

2) Suppose that $\beta_0\notin\Theta_\nu$ and $l(\nu)<k$. Let $G_n=\Gamma+o\left(n^{-1/2}\right)$ be satisfied $n\to\infty$. Then we have

$$\bar{M}_n(\nu)=\left(\sigma^2+K_\nu\right)^{-1}K_\nu n+\sqrt{n}W_n+o_{\mathbb{P}}\left(\sqrt{n}\right),$$

where $W_n\xrightarrow{\mathcal{D}}\mathcal{N}\left(0,\sigma_W^2\right)$, $\sigma_W^2=4\left(\sigma^2+K_\nu\right)^{-2}\sigma^2K_\nu$.

Depending on whether the true parameter β_0 belongs to submodel ν or not, the statistic $\bar{M}_n(\nu)$ has a different asymptotic behaviour. In the first case, it has an asymptotic χ^2-distribution. In the second case it tends to infinity in probability with rate $\sqrt{n}$. Therefore, the statistic $\bar{M}_n(\nu)$ is suitable for model selection. In the next section a selection procedure is introduced based on $\bar{M}_n(\nu)$ serving as fundamental statistic.

3. The New Selection Rule

In this section we propose a selection rule which is based

on the statistic (4). We introduce a measure $d(\nu) \in \mathbb{Z}$ of the complexity for submodel ν with $0 \le d(\nu) \le d_{\max}$. With this quantity $d(\nu)$ it is possible to incorporate a subjective grading of the model complexity. The restriction to integers is made for simpler handling in the selection algorithm. The following examples should illustrate the applicability of the complexity measure.

Example 4. We consider the polynomial $p(x) = \beta_1 + \beta_2 x + \cdots + \beta_k x^{k-1}$. The regressor is observed at the measurement points $x_1, \cdots, x_n$. Hence $x_{ij} = x_i^{j-1}$ for $i = 1, \cdots, n$, $j = 1, \cdots, k$. Possible choices for $d = d(\nu)$ are:

1) d is the degree of the polynomial plus 1,

2) $d = l(\nu)$ is the number of parameters β_j available in the submodel, the other parameters β_j are zero,

3) $d = \frac{k(k-1)}{2} + l(\nu)$, where $l(\nu)$ is the number of parameters β_j available in the submodel. This choice has the advantage that a polynom of higher degree always gets a higher complexity number. □

Example 5. For a quasilinear model with regression function $f(x) = \beta_1 + \beta_2 x + \beta_3 \ln(x)$, we can define d as follows:

$$d = \begin{cases} 1 \text{ for submodel functions } f(x) = \beta_1, f(x) = \beta_2 x, \\ 2 \text{ for submodel function } f(x) = \beta_3 \ln(x) \\ 3 \text{ for submodel function } f(x) = \beta_1 + \beta_2 x \\ 4 \text{ for submodel functions } f(x) = \beta_1 + \beta_3 \ln(x), f(x) = \beta_2 x + \beta_3 \ln(x) \\ 5 \text{ for the full model} \end{cases}$$

This choice takes into account that the logarithm is a more complex function in comparison to constants or linear functions. □

Next we need restricted parameter sets defined by $\Theta_\nu^\circ = \Theta_\nu \setminus \bigcup_{i: i \ne \nu, d(i) \le d(\nu)} \Theta_i$. It is assumed that $\Theta_\nu^\circ \ne \varnothing$ for $\nu = 1, \cdots, \bar{\nu}$.

Example 1: If $d(\nu) = l(\nu)$ then $\Theta_\nu^\circ = \{\beta : \beta_j \ne 0 \text{ for all } j \in J_\nu, \beta_j = 0 \text{ for all } j \notin J_\nu\}$. □

Example 3: If $d(\nu) = l(\nu)$ then

$$\Theta_\nu^\circ = \{\beta : \mu_j \ne \mu_k \text{ for all } j \in J_{\nu l}, k \in J_{\nu l}, i \ne I, \mu_j = \mu_k \text{ if } j, k \in J_{iv} \text{ for some } i\}. \square$$

Given values $\alpha_n(0), \cdots, \alpha_n(d_{\max}) \in (0, \bar{\alpha})$, $\bar{\alpha} < 1$, we introduce special χ^2-quantiles:

$$\psi_n(d,l) = \chi^2_{k-l}(1 - \alpha_n(d))$$
$$\text{for } l = 0, \cdots, k-1, \ \psi_n(d,k) = 1.$$

Here $\psi_n(d,l)$ is just the quantile of order $1 - \alpha_n(d)$ of the asymptotic distribution of $\bar{M}_n(\nu)$ unter the null hypothesis $\beta_0 \in \Theta_\nu$, cf. part 1) of Proposition 2.1. The quantity $\alpha_n(d)$ will play the role of an asymptotic type-1 error probability later. A submodel is referred to as admissible if $\bar{M}_n(\nu) \le \psi_n(d(\nu), l(\nu))$ is satisfied, which in turn corresponds to the nonrejection of the hypothesis that the parameter belongs to the space Θ_ν of the submodel. The generalised information criterion introduced by Shao (see [20]) is given by $GIC_\nu = S_\nu + \lambda_n l(\nu) R_n / (n-k)$. We next show that there is a relationship between the both approaches. A submodel ν is admissible if

$$GIC_\nu < GIC_{\bar{\nu}},$$

where $\lambda_n = \psi_n(n-k) / \big((k - l(\nu))(n - k - \psi_n)\big)$. Moreover, note that our selection procedure is completely different from Shao's one. Whereas λ_n in information criteria is typically free of choice, the quantity ψ_n is well-defined and motivated. Let F_l be the distribution function of the χ^2_l-distribution. We introduce the following rule for the selection:

Select a model ν^* such that

$$d(\nu^*) = \min\{d(\nu) : 1 \le \nu \le \bar{\nu}, \bar{M}_n(\nu) \le \psi_n(d(\nu), l(\nu))\}$$

and

$$F_{k-l(\nu^*)}(\bar{M}_n(\nu^*)) = \min\{F_{k-l(\nu)}(\bar{M}_n(\nu)) : 1 \le \nu \le \bar{\nu}, d(\nu) = d(\nu^*)\}.$$

The central idea is to prefer any admissible model with lower complexity. If there is more than one admissible model with the same minimum complexity, then we take the model with maximum p-value of $\bar{M}_n(\nu)$.

The next step is to analyse the asymptotic behaviour of the probability that the wrong model is selected; *i.e.* the probability of mis-selection (PMS). Let $\beta_0 \in \Theta_\nu^\circ$, $\bar{d} = d(\nu)$, $\bar{l} = l(\nu)$. The following cases of mis-selection can occur:

(m1) $\bar{M}_n(\nu) > \psi_n(\bar{d}, \bar{l})$,

(m2) $\bar{M}_n(\nu) \le \psi_n(\bar{d}, \bar{l})$,

$F_{k-l(i)}(\bar{M}_n(i)) < F_{k-\bar{l}}(\bar{M}_n(\nu))$ for some $i : d(i) = \bar{d}$,

$\bar{M}_n(j) > \bar{M}_n(\nu)$ for all $j : d(j) < \bar{d}$,

(m3) $\bar{M}_n(\nu) \le \psi_n(\bar{d}, \bar{l})$, $\bar{M}_n(j) \le \psi_n(d(j), l(j))$

for some j with $d(j) < \bar{d}$.

The probability of mis-selection case (m2) may be decreased by reducing the number of submodels having the same complexity. The Theorem 3.1 below provides bounds for the selection error.

Theorem 3.1. Let $\beta_0 \in \Theta_\nu^\circ$. Assume that Assumption $\mathcal{A}$ is fulfilled, and

$$\lim_{n\to\infty} \frac{1}{n} \ln \alpha_n(d) = 0 \quad \text{for all} \quad d = 0, \cdots, d_{\max}.$$

1) If $G_n = \Gamma + o\left(n^{-1/2}\right)$ as $n \to \infty$, and $p \geq 3$, then

$$\mathbb{P}\{(m1)\} = \alpha_n(\bar{d})\left(1 + o(1)\right) + O\left(n^{-3/2} B_{n3}\right)$$

$$\text{with} \quad B_{n3} = \max_{j=1,\cdots,k} \sum_{i=1}^{n} \left|x_{ij}\right|^3.$$

2) If $\alpha_n(d) \geq \bar{C} n^{-a}$ for all $d = 0, \cdots, d_{\max}$ with some $a, \bar{C} > 0$, then

$$\mathbb{P}\{(m2)\} = O\left(B_{np} n^{-p}\right) \text{ and } \mathbb{P}\{(m3)\} = O\left(B_{np} n^{-p}\right).$$

The PMS of case (m1) behaves like a type-1-error in a statistical test. It approaches asymptotically $\alpha_n(\bar{d})$ under the assumptions of part 1). The additional term with rate $O\left(n^{-3/2} B_{n3}\right)$ comes from the application of the central limit theorem, and has rate $O\left(n^{-1/2}\right)$ in the case, where the x_{ij}'s are uniformly bounded. This theorem shows that the PMS of cases (m2) and (m3) tends to zero at rate $O\left(n^{1-p}\right)$ provided that the x_{ij}'s are uniformly bounded and $\alpha_n(d) \geq \bar{C} n^{-a}$ for all d and some $a > 0$. These rates of PMS are rather fast. They are better than in comparable cases in [20] (λ_n and ψ_n can be considered to have the same rate). One reason is that in this paper alternative techniques such as Fuk-Nagaev inequality are employed to obtain the convergence rates. The results of Theorem 3.1 recommend the selection rule above from the theoretical point of view. The behaviour in practice is discussed in the next section.

4. Simulations

Here we consider the polynomial model:

$$Y_{ni} = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \varepsilon_i \quad \text{for} \quad i = 1, \cdots, n$$

where $x_1, \cdots, x_n \in [0,1]$ are the observations of the regressor variable, and the ε_i's are i.i.d. random variables. For simplicity, we consider the case $x_i = \frac{i}{n}$. The complexity d is measured as given in Example 4(b). We compare the selection method of the previous section with procedures based on Schwarz's Bayesian information criterion (BIC, see [3]) and the Hannan-Quinn criterion (HQIC, see [25]). The **Tables 1-3** show the frequencies of mis-selection. The results are based on 10^6 replications of the model. We choose the following error probabilities: $\alpha_n(1) = 0.02$, $\alpha_n(2) = 0.022$, $\alpha_n(3) = 0.024$, $\alpha_n(4) = 0.026$ in the case $n = 100$, and $\alpha_n(i) = 0.01$ in the case $n = 400$.

Table 1. Frequencies for mis-selection (FM) in percent in the case n = 100, σ = 0.2, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

β_1	β_2	β_3	β_4	FM new method	FM BIC	FM HQIC
0	100	100	100	1.910	2.018	2.043
0.344	100	100	100	1.998	1.895	1.869
100	0	100	100	1.900	2.006	2.029
100	3	100	100	1.952	1.855	1.830
100	100	0	100	1.918	2.029	2.055
100	100	6.99	100	1.943	1.844	1.822
100	100	100	0	1.911	2.017	2.043
100	100	100	4.58	2.011	1.910	1.886
0	0	100	100	2.049	3.201	3.239
–0.3681	3.21	100	100	1.830	5.006	4.928
0	0	0	100	2.078	3.725	3.780
0.377	3.38	7.65	100	1.936	3.490	3.432
0	0	0	0	2.102	4.008	4.060
0.38872	3.39	7.8987	5.1754	1.825	5.269	5.178
–0.38872	3.39	7.8987	5.1754	1.830	6.309	6.198
0.38872	–3.39	7.8987	5.1754	1.893	5.297	5.213
0.38872	3.39	–7.8987	5.1754	1.873	8.039	7.900
0.38872	3.39	7.8987	–5.1754	1.897	6.452	6.347
–0.38872	–3.39	7.8987	5.1754	1.893	5.297	5.213
–0.38872	3.39	–7.8987	5.1754	1.864	14.207	13.95
–0.38872	3.39	7.8987	–5.1754	2.029	6.736	6.622

Table 2. FM in percent for different error distributions.

n	β_1	β_2	β_3	β_4	$\varepsilon_i \sim$	FM new meth.	FM BIC	FM HQIC
100	0	0	0	0	$\sigma \cdot t(3)$	1.735	3.895	3.951
	0.468	4.104	9.516	6.264	$\sigma \cdot t(3)$	2.043	2.966	2.943
400	0	0	0	0	$\mathcal{N}(0, \sigma^2)$	0.956	1.780	2.502
	0	0	0	0	$\sigma \cdot t(3)$	0.942	1.736	2.459
	–0.216	1.873	4.365	–2.863	$\mathcal{N}(0, \sigma^2)$	1.122	5.869	3.905
	–0.216	1.873	4.365	–2.863	$\sigma \cdot t(3)$	3.067	8.164	6.275

The selection rule of the previous section always gives FM-values near the given values of α_n. The methods based on BIC and HQIC partially show FM-values also near these α_n, but in some special cases the FM-values are much higher (for example, for $\beta_1 = -0.38872$, $\beta_2 = 3.39$, $\beta_3 = -7.8987$, $\beta_4 = 5.1754$ according to **Table 1**; $\beta_1 = -0.2569$, $\beta_2 = 2.227$, $\beta_3 = -5.197$, $\beta_4 = 3.405$ according to **Table 3**). By our method we

Table 3. FM in percent in the case n = 400, σ = 0.2, $\varepsilon_i \sim \sigma \cdot t(3)$.

β_1	β_2	β_3	β_4	FM new method	FM BIC	FM HQIC
0	0	0	0	0.942	1.736	2.459
0.2569	2.227	5.197	3.405	1.003	1.990	1.596
–0.2569	2.227	5.197	3.405	0.958	2.086	1.657
0.2569	–2.227	5.197	3.405	0.984	2.169	1.728
0.2569	2.227	–5.197	3.405	0.945	2.575	2.004
0.2569	2.227	5.197	–3.405	0.987	2.141	1.690
–0.2569	–2.227	5.197	3.405	1.011	2.005	1.606
–0.2569	2.227	–5.197	3.405	0.798	3.567	2.652
–0.2569	2.227	5.197	–3.405	1.015	2.299	1.823
0	100	100	100	1.200	1.064	1.427
0.217	100	100	100	0.983	1.059	0.890
100	0	100	100	1.004	0.873	1.217
100	1.89	100	100	0.969	1.046	0.868
100	100	0	100	0.984	0.844	1.202
100	100	4.38	100	0.976	1.059	0.876
100	100	100	0	0.948	0.817	1.168
100	100	100	2.87	0.992	1.070	0.887
100	0	0	100	0.973	1.0706	1.525
100	2.08	4.84	100	1.010	1.084	0.914
100	2.08	–4.84	100	0.868	2.2384	1.741

are able to control the FM-values by choosing an appropriate α_n.

5. Proofs

By C, we denote a positive generic constant which can vary from place to place and does not depend on other variates. Throughout this section, we assume that Assumption $\mathcal{A}$ is fulfilled. In the following we prove auxiliary statements which are used later in the proofs of the theorems.

Lemma 5.1.

$$\mathbb{P}\left\{\varepsilon^T XX^T\varepsilon \ge \eta\right\} \le C_1\left(B_{np}\eta^{-p/2} + e^{-C_0\eta/n}\right)$$

holds for all $\eta > 0$, where $C_0, C_1 > 0$ are constants not depending on η, n.

Proof: Obviously,

$$\mathbb{P}\left\{\varepsilon^T XX^T\varepsilon \ge \eta\right\} \le \sum_{j=1}^{k}\left\{\left(\sum_{i=1}^{n} x_{ij}\varepsilon_i\right)^2 \ge \eta k^{-1}\right\}$$

$$= \sum_{j=1}^{k}\mathbb{P}\left\{\left|\sum_{i=1}^{n} x_{ij}\varepsilon_i\right| \ge \sqrt{\eta}k^{-1/2}\right\},$$

and $\sum_{i=1}^{n} x_{ij}^2 \le n\mathrm{Trace}\left(G_n\right) = O\left(n\right)$.

Applying Fuk-Nagaev's inequality (see [26]), we obtain the assertion of the lemma:

$$\mathbb{P}\left\{\varepsilon^T XX^T\varepsilon \ge \eta\right\}$$

$$\le C\sum_{j=1}^{k}\left(\eta^{-p/2}\sum_{i=1}^{n}\left|x_{ij}\right|^p \mathbb{E}\left|\varepsilon_i\right|^p + \exp\left(-\frac{C\eta}{\sum_{i=1}^{n} x_{ij}^2\sigma^2}\right)\right)$$

$$\le C\left(B_{np}\eta^{-p/2} + \exp\left(-\frac{C\eta}{n}\right)\right). \quad \square$$

Lemma 5.2. Assume that $\beta_0 \in \Theta_\nu$ for some $\nu \in \{1, \cdots, \bar{\nu}\}$. Then

$$\mathbb{P}\left\{\left|\frac{1}{n-\bar{l}}S_\nu - \sigma^2\right| \ge \eta\right\} \le C_2\left(n^{-p/2}\eta^{-p/2} + e^{-C_3\eta^2 n}\right)$$

holds for $\eta n > 8\bar{l}\sigma^2$, $\eta \le 1$ and n large enough, where $\bar{l} := l(\nu)$ and $C_2, C_3 > 0$ are constants not depending on η, n. The same upper bound holds for $\mathbb{P}\left\{\left|\frac{1}{n-\bar{l}}\varepsilon^T\varepsilon - \sigma^2\right| \ge \eta\right\}$.

Proof: Observe that

$$S_\nu = \varepsilon^T\left(I - X_\nu\left(X_\nu^T X_\nu\right)^{-1}X_\nu^T\right)\varepsilon$$

by (3). Hence

$$\mathbb{P}\left\{\left|\frac{1}{n-\bar{l}}S_\nu - \sigma^2\right| \ge \eta\right\} \le \mathbb{P}\left\{\left|\frac{1}{n-\bar{l}}\varepsilon^T\varepsilon - \sigma^2\right| \ge \frac{\eta}{2}\right\}$$
$$+ \mathbb{P}\left\{\frac{1}{n-\bar{l}}\varepsilon^T X_\nu\left(X_\nu^T X_\nu\right)^{-1}X_\nu^T\varepsilon \ge \frac{\eta}{2}\right\}. \quad (5)$$

Further an application of Fuk-Nagaev's inequality from [26] leads to

$$\mathbb{P}\left\{\left|\frac{1}{n-\bar{l}}\varepsilon^T\varepsilon - \sigma^2\right| \ge \frac{\eta}{2}\right\} \le \mathbb{P}\left\{\left|\sum_{i=1}^{n}\left(\varepsilon_i^2 - \sigma^2\right)\right| \ge \frac{\eta n}{8}\right\}$$
$$\le C\left(\eta^{-p/2}n^{-p/2}\sum_{i=1}^{n}\mathbb{E}\left|\varepsilon_i\right|^p + \exp\left(-\frac{C\eta^2 n}{\mathbb{E}\varepsilon_i^4}\right)\right) \quad (6)$$
$$\le C\left(n^{-p/2}\eta^{-p/2} + e^{-C\eta^2 n}\right)$$

for $\eta n > 8\bar{l}\sigma^2$, $n \ge 2\bar{l}$. Since $\tilde{D}_\nu := D_\nu\left(\frac{1}{n}X_\nu^T X_\nu\right)^{-1}D_\nu^T \to D_\nu\Gamma_\nu D_\nu^T \in \mathbb{R}^{k,k}$ holds, and therefore, $\tilde{D}_\nu$ has a bounded norm, we deduce

$$\mathbb{P}\left\{\frac{1}{n-\bar{l}}\varepsilon^T X_\nu\left(X_\nu^T X_\nu\right)^{-1}X_\nu^T\varepsilon \ge \frac{\eta}{2}\right\}$$
$$\le \mathbb{P}\left\{\varepsilon^T XX^T\varepsilon \ge C\eta n^2\right\} \le C\left(B_{np}n^{-p}\eta^{-p/2} + e^{-C\eta n}\right). \quad (7)$$

by Lemma 5.1 for n large enough. A combination of Inequalities (5)-(7) yields the lemma. $\square$

Note that

$$S_\nu = \left(\varepsilon + X\beta_0\right)^T \left(I - X_\nu \left(X_\nu^T X_\nu\right)^{-1} X_\nu^T\right)\left(\varepsilon + X\beta_0\right) \quad (8)$$
$$= S_{\nu 1} + 2S_{\nu 2} + S_{\nu 3},$$

where

$$S_{\nu 1} = \varepsilon^T \left(I - X_\nu \left(X_\nu^T X_\nu\right)^{-1} X_\nu^T\right)\varepsilon,$$
$$S_{\nu 2} = \beta_0^T \left(I - G_n D_\nu G_{\nu n}^{-1} D_\nu^T\right) X^T \varepsilon,$$
$$S_{\nu 3} = n\beta_0^T \left(G_n - G_n D_\nu G_{\nu n}^{-1} D_\nu^T G_n\right)\beta_0, \quad G_{\nu n} := \frac{1}{n} X_\nu^T X_\nu.$$

Lemma 5.3. Suppose that $\beta_0 \notin \Theta_\nu$ for some $\nu \in \{1, \cdots, \bar{\nu}\}$. Let $\eta_0 > K_\nu + 2\sigma^2$, $\bar{l} = l(\nu)$. Then

1) $S_{\nu 3} = nK_\nu + o\left(\sqrt{n}\right)$ and

2)

$$\mathbb{P}\left\{\frac{1}{n-\bar{l}} S_\nu \geq \eta\right\} \leq C_4 \left(n^{-p/2}\left(\eta - \eta_0\right)^{-p/2} + e^{-C_5(\eta-\eta_2)^2 n}\right)$$

for $\eta > \eta_0$ and n large enough with constants $C_4, C_5 > 0$ not depending on n, η.

Proof: Part 1) is a consequence of $G_{\nu n} \to \Gamma_\nu$ and $G_n \to \Gamma$. 2) Using Lemmas 5.1 and 5.2, we deduce

$$\begin{aligned}
\mathbb{P}\left\{\frac{1}{n-\bar{l}} S_\nu \geq \eta\right\} &\leq \mathbb{P}\left\{\frac{1}{n-\bar{l}}\left(S_{\nu 1} + \frac{1}{2}S_{\nu 3}\right) \geq \frac{\eta}{2}\right\} + \mathbb{P}\left\{\frac{1}{n-\bar{l}}\left(S_{\nu 2} + \frac{1}{2}S_{\nu 3}\right) \geq \frac{\eta}{2}\right\} \\
&\leq \mathbb{P}\left\{\frac{1}{2}K_\nu + n^{-1/2}C + \frac{1}{n-\bar{l}}\varepsilon^T\varepsilon \geq \frac{\eta}{2}\right\} + \mathbb{P}\left\{\frac{1}{2}K_\nu + n^{-1/2}C + \frac{1}{n-\bar{l}}C\left\|X^T\varepsilon\right\| \geq \frac{\eta}{2}\right\} \\
&\leq \mathbb{P}\left\{\frac{1}{n-\bar{l}}\varepsilon^T\varepsilon \geq \frac{\eta - \eta_0 + 2\sigma^2}{2}\right\} + \mathbb{P}\left\{\varepsilon^T XX^T\varepsilon \geq C\left(\eta - \eta_0\right)^2 n^2\right\} \\
&\leq \mathbb{P}\left\{\left|\frac{1}{n-\bar{l}}\varepsilon^T\varepsilon - \sigma^2\right| \geq \frac{\eta - \eta_0}{2}\right\} + C\left(B_{np} n^{-p}\left(\eta - \eta_0\right)^{-p/2} + e^{-C(\eta-\eta_0)^2 n}\right) \leq C\left(n^{-p/2}\left(\eta - \eta_0\right)^{-p/2} + e^{-C(\eta-\eta_0)^2 n}\right)
\end{aligned}$$

for $n \geq 2\bar{l}$ large enough. This implies assertion 2) of the lemma. ⊡

An application of the central limit theorem and the Cramér-Wold device leads to the following lemma:

Lemma 5.4. Let $\xi_n := n^{-1/2} X^T \varepsilon$. $\tilde{x}_i$ denotes the *i*-th column of X^T. Then

$$\xi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{x}_i \varepsilon_i \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\Gamma\right).$$

In the second part of this section we provide the proofs of Proposition 2.1 and Theorem 3.1.

Proof of Proposition 2.1. 1) Let $\beta_0 \in \Theta_\nu$.

Then $D_\nu \gamma_0 = \beta_0$ holds with an appropriate vector $\gamma_0 \in \mathbb{R}^{l(\nu)}$. We have

$$\begin{aligned}
M_n(\nu) &= Y^T \left(X\left(X^T X\right)^{-1} X^T - X_\nu \left(X_\nu^T X_\nu\right)^{-1} X_\nu^T\right) Y^T \\
&= \varepsilon^T X \left(\left(X^T X\right)^{-1} - D_\nu \left(X_\nu^T X_\nu\right)^{-1} D_\nu^T\right) X^T \varepsilon \\
&= \xi_n^T \left(G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T\right)\xi_n.
\end{aligned}$$

Moreover, the identity

$$\lim_{n\to\infty} G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T = \Gamma^{-1} - D_\nu \Gamma_\nu^{-1} D_\nu^T \quad (9)$$

holds in view of Assumption $\mathcal{A}$. An application of Lemma 5.4 and the Cochran theorem leads to $\sigma^{-2} M_n(\nu) \xrightarrow{\mathcal{D}} \chi^2_{k-l(\nu)}$. Lemma 5.2 implies that

$$\frac{1}{n - l(\nu)} S_\nu \xrightarrow{\mathbb{P}} \sigma^2,$$

and therefore assertion 1) of Proposition 2 is proved.

2) Let $\beta_0 \notin \Theta_\nu$, and $W_n := 2\beta_0^T G_n \left(G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T\right)\xi_n$. By assumption, $G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T = \Gamma^{-1} - D_\nu \Gamma_\nu^{-1} D_\nu^T + o\left(n^{-1/2}\right)$ holds true.

We derive

$$\begin{aligned}
M_n(\nu) &= \left(\varepsilon^T X + \beta_0^T X^T X\right)\left(\left(X^T X\right)^{-1} - D_\nu \left(X_\nu^T X_\nu\right)^{-1} D_\nu^T\right)\left(X^T X\beta_0 + X^T\varepsilon\right) \\
&= \xi_n^T \left(G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T\right)\xi_n + \sqrt{n}W_n + n\beta_0^T \left(G_n - G_n D_\nu G_{\nu n}^{-1} D_\nu^T G_n\right)\beta_0 = nK_\nu + \sqrt{n}W_n + o_{\mathbb{P}}\left(\sqrt{n}\right).
\end{aligned}$$

From Lemma 5.4, (9) and $G_n \to \Gamma$, it follows that $W_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_0^2\right)$, where

$$\begin{aligned}
\sigma_0^2 &:= 4\sigma^2 \beta_0^T \Gamma\left(\Gamma^{-1} - D_\nu \Gamma_\nu^{-1} D_\nu^T\right)\Gamma\left(\Gamma^{-1} - D_\nu \Gamma_\nu^{-1} D_\nu^T\right)\Gamma\beta_0 \\
&= 4\sigma^2 K_\nu.
\end{aligned}$$

We obtain $\frac{1}{n - l(\nu)} S_{\nu 1} \xrightarrow{\mathbb{P}} \sigma^2$ using Lemma 5.2. Moreover, we deduce

$$S_{\nu 2} = \sqrt{n}\beta_0^T \left(I - G_n D_\nu G_{\nu n}^{-1} D_\nu^T\right)\xi_n = o_{\mathbb{P}}(n).$$

Hence by (8) and Lemma 5.3,

$\frac{1}{n-l(\nu)} S_\nu \xrightarrow{\mathbb{P}} \sigma^2 + K_\nu$, which completes the proof of assertion 2). ⊡

In the proof of Theorem 3.1, we need the following Lemma which is proved before.

Lemma 5.5. For $\beta_0 \notin \Theta_\nu$, $K_\nu > 0$ holds true.

Proof. Let $Q := I - \Gamma^{1/2} D_\nu \Gamma_\nu^{-1} D_\nu^T \Gamma^{1/2}$, and $y = \Gamma^{1/2}\beta_0$. Then $K_\nu = y^T Q y \geq 0$ since Q is symmetric and idempotent. Moreover,

$$\begin{aligned}\mathrm{Rank}(Q) &= \mathrm{Trace}(Q) = k - \mathrm{Rank}\left(D_\nu \Gamma_\nu^{-1} D_\nu^T\right) \\ &= k - l =: m.\end{aligned}$$

Therefore we have the following representation

$$Q = \sum_{i=1}^m h_i h_i^T = HDH^T, \quad D = \mathrm{diag}(1_1, \cdots, 1_m, 0_{m+1}, \cdots, 0_k),$$

and $h_1, \cdots, h_m$ are the first m columns of the orthogonal matrix $H \in \mathbb{R}^{k\times k}$. For $x \in \mathbb{R}^k$,

$$x^T Q x = \sum_{i=1}^m \left(x^T h_i\right)^2 = 0 \Leftrightarrow x \perp h_i$$

$$\text{for } i = 1, \cdots, m \Leftrightarrow x \in \mathcal{L}(h_{m+1}, \cdots, h_k).$$

We consider the linear independent vectors $z_1 = D_\nu e_1, \cdots, z_l = D_\nu e_l$, $e_j = (0, \cdots, 1_j, 0, \cdots)^T \in \mathbb{R}^l$ is the j-th unit vector, and obtain

$$z_j^T P z_j = e_j^T \left(D_\nu^T \Gamma D_\nu - D_\nu^T \Gamma D_\nu \Gamma_\nu^{-1} D_\nu^T \Gamma D_\nu\right) e_j = 0. \quad (10)$$

Since $\tilde{z}_1 = \Gamma^{1/2} z_1, \cdots, \tilde{z}_l = \Gamma^{1/2} z_l \in \mathcal{L}(h_{m+1}, \cdots, h_k)$ are linear independent, these vectors form a basis of $\mathcal{L}(h_{m+1}, \cdots, h_k)$. Assume that $K_\nu = 0$. Then there exist a $a \in \mathbb{R}^l$ such that

$$\Gamma^{1/2}\beta_0 = \sum_{i=1}^l a_i \tilde{z}_i = \Gamma^{1/2} D_\nu a, \text{ and hence } \beta_0 = D_\nu a \in \Theta_\nu$$

in contradiction to the assumption. This proves the lemma. ⊡

Proof of Theorem 3.1: One shows easily that $\frac{1}{n}\psi_n(d, l) \to 0$.

1) Let $\bar{\xi}_n = \sigma^{-1} G_n^{-1/2} \xi_n$ (ξ_n as in Lemma 5.4), and $\kappa > 0$ be a constant. Define

$\underline{\psi}_n := \left(\sigma^2 - \kappa n^{-1/4}\right)\psi_n(\bar{d}, \bar{l})$,

$\bar{\psi}_n := \left(\sigma^2 + \kappa n^{-1/4}\right)\psi_n(\bar{d}, \bar{l})$. Since $M_n(\nu) = \sigma^2 \bar{\xi}_n^T \bar{G}_n \bar{\xi}_n, \bar{G}_n = I - G_n^{1/2} D_\nu G_{\nu n}^{-1} D_\nu^T G_n^{1/2}$, we obtain by using Lemma 5.2

$$\begin{aligned}&\mathbb{P}\{(m1)\} = \mathbb{P}\left\{\bar{M}_n(\nu) > \psi_n(\bar{d}, \bar{l})\right\} \\ &\leq \mathbb{P}\left\{\bar{M}_n(\nu) > \psi_n(\bar{d}, \bar{l}), \sigma^2 - \kappa n^{-1/4} < \frac{1}{n-l(\nu)} S_\nu\right\} \\ &\quad + \mathbb{P}\left\{\left|\frac{1}{n-l(\nu)} S_\nu - \sigma^2\right| \geq \kappa n^{-1/4}\right\} \quad (11) \\ &\leq \mathbb{P}\left\{M_n(\nu) > \underline{\psi}_n\right\} + O\left(n^{-1/2}\right) \\ &\leq \mathbb{P}\left\{\bar{\xi}_n^T \bar{G}_n \bar{\xi}_n > \sigma^{-2}\underline{\psi}_n\right\} + O\left(n^{-1/2}\right)\end{aligned}$$

and analogously,

$$\mathbb{P}\left\{\bar{M}_n(\nu) > \psi_n(\bar{d}, \bar{l})\right\} \geq \mathbb{P}\left\{\bar{\xi}_n^T \bar{G}_n \bar{\xi}_n > \sigma^{-2}\bar{\psi}_n\right\} + O\left(n^{-1/2}\right). \quad (12)$$

Note that $\mathrm{cov}(\xi_n) = \sigma^2 G_n$, which implies $\mathrm{cov}(\bar{\xi}_n) = I$. Further by Assumption $\mathcal{A}$,

$$n^{-3/2} \sum_{i=1}^n \mathbb{E}\left\|G_n^{-1/2} \tilde{x}_i \varepsilon_i\right\|^3 = O\left(n^{-3/2} B_{n3}\right).$$

Since $\left\{z \in \mathbb{R}^k : z^T \bar{G} z \leq a\right\}$ is a convex set for all $a > 0$, we can apply Bhattacharya's theorem on a multivariate Berry-Esseen inequality (see [27])

$$\left|\mathbb{P}\left\{\bar{\xi}_n^T \bar{G}_n \bar{\xi}_n > \sigma^{-2}\bar{\psi}_n\right\} - \mathbb{P}\left\{Z^T \bar{G}_n Z^T > \sigma^{-2}\bar{\psi}_n\right\}\right| = O\left(n^{-1/2}\right),$$

where $Z \sim \mathcal{N}(0, I)$. The Cochran theorem implies that $Z^T \bar{G}_n Z^T \sim \chi^2_{k-\bar{l}}$. We denote the distribution function of the $\chi^2_{k-\bar{l}}$-distribution by $F_{k-\bar{l}}$. Hence

$$\begin{aligned}&\mathbb{P}\left\{Z^T \bar{G}_n Z^T > \sigma^{-2}\bar{\psi}_n\right\} = 1 - F_{k-\bar{l}}\left(\sigma^{-2}\bar{\psi}_n\right) \\ &= \left(1 - F_{k-\bar{l}}\left(\psi_n(\bar{d}, \bar{l})\right)\right)\left(1 + o(1)\right) = \alpha_n(\bar{d})\left(1 + o(1)\right), \\ &\text{and } \mathbb{P}\left\{Z^T \bar{G}_n Z^T > \sigma^{-2}\underline{\psi}_n\right\} = \alpha_n(\bar{d})\left(1 + o(1)\right).\end{aligned}$$

Combining these identities and (11), (12) we obtain assertion 1).

2) One can show that $\psi_n(\bar{d}, l(i)) = O(\ln(n))$. Let $\{\zeta_n\}$ be a sequence of real numbers with $\zeta_n \to \infty$, $\zeta_n = o\left(n\psi_n(\bar{d}, l(i))^{-1}\right)$. We deduce

$$\begin{aligned}\mathbb{P}\{(m2)\} &\leq \mathbb{P}\left\{\bar{M}_n(\nu) \leq \psi_n(\bar{d}, \bar{l}), F_{k-l(i)}\left(\bar{M}_n(i)\right) \leq F_{k-l(\nu)}\bar{M}_n(\nu) \text{ for some } i : d(i) = \bar{d}\right\} \\ &\leq \sum_{i\neq\nu : d(i)=\bar{d}} \mathbb{P}\left\{F_{k-l(i)}\left(\bar{M}_n(i)\right) \leq F_{k-\bar{l}}\left(\psi_n(\bar{d}, \bar{l})\right)\right\} = \sum_{i\neq\nu : d(i)=\bar{d}} \mathbb{P}\left\{F_{k-l(i)}\left(\bar{M}_n(i)\right) \leq 1 - \alpha_n(\bar{d})\right\} \\ &= \sum_{i\neq\nu : d(i)=\bar{d}} \mathbb{P}\left\{\bar{M}_n(i) \leq \chi^2\left(k - l(i), 1 - \alpha_n(\bar{d})\right)\right\} \\ &\leq \sum_{i\neq\nu : d(i)=\bar{d}} \left(\mathbb{P}\left\{M_n(i) \leq \psi_n(\bar{d}, l(i))\zeta_n, \frac{1}{n-l(i)} S_i \leq \zeta_n\right\} + \mathbb{P}\left\{\frac{1}{n-l(i)} S_i > \zeta_n\right\}\right).\end{aligned}$$

Define $\bar{K}_{ni} := \beta_0^T G_n \left(G_n^{-1} - D_i G_{in}^{-1} D_i^T \right) G_n \beta_0$. Let $i \neq \nu$ with $d(i) = \bar{d}$. Obviously, $\lim_{n\to\infty} \bar{K}_{ni} = K_i$ holds true. Since $\beta_0 \notin \Theta_i$, we have $K_i > 0$ by Lemma 5.5. Furthermore, by Lemma 5.1 we obtain

$$\begin{aligned}
&\mathbb{P}\left\{ M_n(i) \le \psi_n\left(\bar{d}, l(i)\right)\zeta_n \right\} = \mathbb{P}\left\{ n\bar{K}_{ni} + \xi_n^T \left(G_n^{-1} - D_i G_{in}^{-1} D_i^T \right)\xi_n + \sqrt{n} W_n \le \psi_n\left(\bar{d}, l(i)\right)\zeta_n \right\} \\
&\le \mathbb{P}\left\{ n\bar{K}_{ni} + \sqrt{n} W_n \le \psi_n\left(\bar{d}, l(i)\right)\zeta_n \right\} = \mathbb{P}\left\{ 2\beta_0^T G_n \left(G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T \right)\xi_n \le \psi_n\left(\bar{d}, l(i)\right)\zeta_n n^{-1/2} - \sqrt{n}\bar{K}_{ni} \right\} \\
&\le \mathbb{P}\left\{ \left| 2\beta_0^T G_n \left(G_n^{-1} - D_\nu G_{\nu n}^{-1} D_\nu^T \right)\xi_n \right| \ge \sqrt{n}\bar{K}_{ni} - \psi_n\left(\bar{d}, l(i)\right)\zeta_n n^{-1/2} \right\} \le \mathbb{P}\left\{ \|\xi_n\| \ge C\sqrt{n} \right\} \le \mathbb{P}\left\{ \varepsilon^T X X^T \varepsilon \ge Cn^2 \right\} = O\left(B_{np} n^{-p} \right)
\end{aligned}$$

for n large enough. On the other hand, we have

$$\mathbb{P}\left\{ \frac{1}{n - l(i)} S_i \ge \zeta_n \right\} = O\left(n^{-p/2} \zeta_n^{-p/2} + e^{-C\zeta_n^2 n} \right)$$

by Lemma 5.3. We choose $\zeta_n = n^{1-2/p}$. Then $\zeta_n = o\left(n\psi_n\left(\bar{d}, l(i)\right)^{-1} \right)$. This completes the proof of the bound for $\mathbb{P}\{(m2)\}$. Observe that

$$\mathbb{P}\{(m3)\} = \mathbb{P}\left\{ \bar{M}_n(i) \le \psi_n\left(d(i), l(i)\right) \text{ for some } i, d(i) < \bar{d} \right\} = \sum_{i \neq \nu : d(i) < \bar{d}} \mathbb{P}\left\{ \bar{M}_n(i) \le \psi_n\left(d(i), l(i)\right) \right\}.$$

The bound for $\mathbb{P}\{(m3)\}$ can now be established along the lines of the proof for $\mathbb{P}\{(m2)\}$. ⊡

REFERENCES

[1] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. 19, 1974, pp. 716-723. doi:10.1109/TAC.1974.1100705

[2] C. Mallows, "Some Comments on Cp," *Technometrics*, Vol. 15, No. 4, 1973, pp. 661-675. doi:10.2307/1267380

[3] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464. doi:10.1214/aos/1176344136

[4] J. Rissanen, "Modeling by Shortest Data Description," *Automatica*, Vol. 14, No. 5, 1978, pp. 465-471. doi:10.1016/0005-1098(78)90005-5

[5] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, Vol. 57, No. 1, 1995, pp. 289-300.

[6] F. Bunea, M. H. Wegkamp and A. Auguste, "Consistent Variable Selection in High Dimensional Regression via Multiple Testing," *Journal of Statistical Planning and Inference*, Vol. 136, No. 12, 2006, pp. 4349-4364. doi:10.1016/j.jspi.2005.03.011

[7] Y. Benjamini and Y. Gavrilov, "A Simple forward Selection Procedure Based on False Discovery Rate Control," *Annals of Applied Statistics*, Vol. 3, No. 1, 2009, pp. 179-198. doi:10.1214/08-AOAS194

[8] G. Claeskens and N. L. Hjort, "Model Selection and Model Averaging," Cambridge University Press, Cambridge, 2008.

[9] H. Leeb and B. M. Pötscher, "Model Selection," In: T. G. Andersen, *et al.*, Eds., *Handbook of Financial Time Series*, Springer, Berlin, 2009, pp. 889-925. doi:10.1007/978-3-540-71297-8_39

[10] A. D. R. McQuarrie and C.-L. Tsai, "Regression and Time Series Model Selection," World Scientific, Singapore City, 1998.

[11] A. J. Miller, "Subset Selection in Regression," 2nd Edition, Chapman & Hall, New York, 2002.

[12] B. Droge, "Asymptotic Properties of Model Selection Procedures in Linear Regression," *Statistics*, Vol. 40, No. 1, 2006, pp. 1-38. doi:10.1080/02331880500366050a

[13] R. Nishii, "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *Annals of Statistics*, Vol. 12, No. 2, 1984, pp. 758-765. doi:10.1214/aos/1176346522

[14] C. R. Rao and Y. Wu, "A Strongly Consistent Procedure for Model Selection in a Regression Problem," *Biometrika*, Vol. 76, No. 2, 1989, pp. 369-374. doi:10.1093/biomet/76.2.369

[15] J. Shao, "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, Vol. 7, 1997, pp. 221-264.

[16] C.-Y. Sin and H. White, "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, Vol. 71, No. 1-2, 1996, pp. 207-225. doi:10.1016/0304-4076(94)01701-8

[17] C. Gatu, P. I. Yanev and E. J. Kontoghiorghes, "A Graph Approach to Generate All Possible Regression Submodels," *Computational Statistics & Data Analysis*, Vol. 52, No. 2, 2007, pp. 799-815. doi:10.1016/j.csda.2007.02.018

[18] H. Leeb, "The Distribution of a Linear Predictor after Model Selection: Conditional Finite-Sample Distributions and Asymptotic Approximations," *Journal of Statistical Planning and Inference*, Vol. 134, No. 1, 2005, pp .64-89.

[19] H. Leeb and B. M. Pötscher, "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, Vol. 21, No. 1, 2005, pp. 21-59. doi:10.1017/S0266466605050036

[20] J. Shao, "Convergence Rates of the Generalized Information Criterion," *Journal of Nonparametric Statistics*, Vol. 9, No. 3, 1998, pp. 217-225. doi:10.1080/10485259808832743

[21] A. Chambaz, "Testing the Order of a Model," *Annals of Statistics*, Vol. 34, No. 3, 2006, pp. 1166-1203. doi:10.1214/009053606000000344

[22] D. E. Edwards and T. Havránek, "A Fast Model Selection Procedure for Large Families of Models," *Journal of the*

American Statistical Association, Vol. 82, No. 397, 1987, pp. 205-213. doi:10.2307/2289155

[23] M. A. Efroymson, "Multiple Regression Analysis," In: A. Ralston and H. S. Wilf, Eds., *Mathematical Methods for Digital Computers*, John Wiley, New York, 1960.

[24] M. Hofmann, C. Gatu and E. J. Kontoghiorghes, "Efficient Algorithms for Computing the Best-Subset Regression Models for Large Scale Problems," *Computational Statistics & Data Analysis*, Vol. 52, No. 1, 2007, pp. 16-29. doi:10.1016/j.csda.2007.03.017

[25] E. J. Hannan and B. G. Quinn, "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society*, *Series B*, Vol. 41, No. 2, 1979, pp.190-195.

[26] D. Kh. Fuk and S. N. Nagaev, "Probability Inequalities for Sums of Independent Random Variables," *Theory of Probability and Its Applications*, Vol. 16, 1971, pp. 643-660. doi:10.1137/1116071

[27] R. N. Bhattacharya, "On Errors of Normal Approximation," *Annals of Probability*, Vol. 3, No. 5, 1975, pp. 815-828. doi:10.1214/aop/1176996268

