Model selection strategies for identifying most relevant covariates in homoscedastic linear models

Aleksey Min[a,*], Hajo Holzmann[b], Claudia Czado[a]

[a] *Zentrum Mathematik, Technische Universität München, Germany*
[b] *Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany*

Abstract

A new method in two variations for the identification of most relevant covariates in linear models with homoscedastic errors is proposed. In contrast to many known selection criteria, the method is based on an interpretable scaled quantity. This quantity measures a maximal relative error one makes by selecting covariates from a given set of all available covariates. The proposed model selection procedures rely on asymptotic normality of test statistics, and therefore normality of the errors in the regression model is not required. In a simulation study the performance of the suggested methods along with the performance of the standard model selection criteria AIC, BIC, Lasso and relaxed Lasso is examined. The simulation study illustrates the favorable performance of the proposed method as compared to the above reference criteria, especially when regression effects possess influence of several orders in magnitude. The accuracy of the normal approximation to the test statistics is also investigated, it is already satisfactory for sample sizes 50 and 100. As an illustration the US college spending data from 1994 is analyzed.

Key words:
Asymptotic normality, linear regression, model selection, model validation, nested models, restricted least squares

1. Introduction

The choice of the relevant covariates in a linear regression model is an important and much studied problem. For this purpose, various methods have been suggested in the literature. One approach is via model selection criteria. Here one seeks to find the sub-model which minimizes a certain information criterion, for example Akaike's information criterion AIC (Akaike, 1974) or the Bayesian information criterion BIC (Schwarz, 1978). Typically, one of the various stepwise methods for subset selection in regression (c.f. Miller, 2002) is applied in order to find the sub-model which actually minimizes the information criterion in use. There is a wide variety of model selection criteria in the literature, apart from AIC and BIC we mention Mallows' (1973) C_p, the deviance information criterion DIC as discussed in Spiegelhalter et al. (2002) or the focused information criterion FIC of Claeskens and Hjort (2003). For further information see the monograph of Burnham and Anderson (2002). Another popular approach to model selection is the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996) and its extensions (see e.g. Meinshausen (2007) and Wang and Leng (2008)). The Lasso finds the ordinary least squares estimate with constrained sum of absolute regression coefficients. The l_1 constraint results in a model selection property of the Lasso since some of its estimates can be zero. In practice, the model selection problem is often also solved by sequentially testing the relevant linear restrictions determining the sub-models. For relationships between model testing using the $F-$test and certain information criteria see Teräsvirta and Mellin (1986).

In this paper we introduce a new type of test which is designed to validate a linear sub-model consisting of variables with strong effects on the response, and discuss its use for variable selection purposes. More

[*] Corresponding author: Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany
Email addresses: aleksmin@ma.tum.de (Aleksey Min), holzmann@mathematik.uni-marburg.de (Hajo Holzmann), cczado@ma.tum.de (Claudia Czado)

specifically, consider the homoscedastic linear regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the response vector, $X := [X_1, X_2] \in \mathbb{R}^{n\times(p+q)}$ is the known design matrix and $\boldsymbol{\beta} := (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')' \in \mathbb{R}^{p+q}$ denotes the unknown regression parameter vector of interest. For the moment the errors $\epsilon_1, \ldots, \epsilon_n$ constituting $\boldsymbol{\epsilon}$ in model (1) are assumed to be independent, identically normally distributed with $E(\epsilon_1) = 0$ and $Var(\epsilon_1) = \sigma^2$. However the distribution of the errors should not be necessarily a normal distribution and later we specify it more generally depending on aims we pursue. Suppose that we want to check the validity of the sub-model

$$\mathbf{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \tag{2}$$

where $X_1 \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\beta}_1 \in \mathbb{R}^p$. Classically one verifies model (2) by testing the point hypothesis

$$H_0 \ : \ \boldsymbol{\beta}_2 = \mathbf{0}$$

using the $F-$test. However, for many purposes it is not adequate to base a decision for or against the sub-model (2) on testing the hypothesis H_0, and some alternative methods have been developed. Toro-Vizcarrondo and Wallace (1968), see also Wallace (1972), observed that the sub-model may be superior to the complete model in terms of mean square error (MSE) even if the sub-model is incorrect. Therefore they suggested to test in which model the least squares estimator has smaller MSE. More precisely, let $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y}$ and $\hat{\boldsymbol{\beta}}_r = (X_1'X_1)^{-1}X_1'\mathbf{Y}$ denote the least squares (LS) estimator in the full model (1) and in the sub-model (2), respectively. Here $\hat{\boldsymbol{\beta}}_r$ is also considered as a $(p+q)-$dimensional vector by filling the last q entries by 0, i.e. $\hat{\boldsymbol{\beta}}_r = \left[\mathbf{Y}'X_1(X_1'X_1)^{-1}, \mathbf{0}'\right]'$. For an arbitrary estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ we let $MSE(\mathbf{b}) := E\left[(\mathbf{b}-\boldsymbol{\beta})(\mathbf{b}-\boldsymbol{\beta})'\right]$, and $MSE(\hat{\boldsymbol{\beta}}_r) \le MSE(\hat{\boldsymbol{\beta}})$ means that $MSE(\hat{\boldsymbol{\beta}}) - MSE(\hat{\boldsymbol{\beta}}_r)$ is positive semidefinite. Now they suggest to test the hypothesis

$$H_{MSE} \ : MSE(\hat{\boldsymbol{\beta}}_r) \le MSE(\hat{\boldsymbol{\beta}}) \qquad \text{versus} \qquad K_{MSE} \ : \ \text{Not } H_{MSE}.$$

Setting

$$\lambda := n\,\frac{d_n(\boldsymbol{\beta}_2)}{\sigma^2}, \quad d_n(\boldsymbol{\beta}_2) := \frac{1}{n}\boldsymbol{\beta}_2'X_2'Q_{X_1}X_2\boldsymbol{\beta}_2, \tag{3}$$

where $P_{X_1} := X_1(X_1'X_1)^{-1}X_1'$ is the projection matrix unto the column space spanned by X_1, $Q_{X_1} := I_n - P_{X_1}$ and I_n is the identity matrix of dimension n, Toro-Vizcarrondo and Wallace (1968) showed that H_{MSE} is equivalent to $\lambda \le 1$. Using the fact that under the assumption of normal errors, the F-statistic corresponding to H_{MSE} is non-central F-distributed (in the notation of Kotz and Johnsson (1970)) with non-centrality parameter λ, Toro–Vizcarrondo and Wallace (1968) constructed a uniformly most powerful test for H_{MSE} versus K_{MSE} based on the F-statistic. Hypotheses related to H_{MSE} were investigated by Wallace (1972) and by Yancey et al. (1973).

The hypothesis H_{MSE} still has some drawbacks. Instead of comparing models, it compares the performance of certain estimators. This is a somewhat arbitrary choice since there are other estimators (e.g. the ridge estimator, cf. Farbrother, 1975), which have smaller MSE than the LS estimator. Further, the hypothesis H_{MSE} compares the performance of pre-model selection estimators, while it should compare the performance of post model-selection estimators (cf. Leeb and Pötscher, 2003).

Moreover, even if the hypothesis H_{MSE} (or H_0) cannot be rejected with a large p-value, this does not imply that the hypothesis H_{MSE} (or H_0) is actually true, and therefore no evidence for sub-model (2) is provided. Hence, we suggest to test a hypothesis which focuses on validating the sub-model (2). A related approach to validating parametric functional forms of regression models (against nonparametric alternatives) was suggested by Dette and Munk (1998). For an extensive discussion on the methodological aspects of performing tests for model validation see Dette and Munk (2003).

In some variable selection problems one is interested not solely in identifying variables which have a nonzero regression coefficient but in identifying variables which have a strong effect on the response compared to the joint effect of the variables not selected. Therefore we need a distance measure between the restricted model (2) and the full model (1). Note that $d_n(\boldsymbol{\beta}_2)$ given in (3) is the squared normalized length (with factor n^{-1}) of the $n-$vector $X_2\boldsymbol{\beta}_2$, when projected onto the orthogonal complement of the space spanned by the columns of X_1 with projection matrix Q_{X_1}. Thus it provides a natural measure of distance between the restricted model (2) and the full model (1). We propose to validate sub-model (2) by testing the hypothesis that

$$H_{\Delta,n} \ : \ d_n(\boldsymbol{\beta}_2) > \Delta \quad \text{ against } K_{\Delta,n} \ : \ d_n(\boldsymbol{\beta}_2) \leq \Delta, \tag{4}$$

for some $\Delta > 0$. Thus, rejecting $H_{\Delta,n}$ actually is a decision in favor of sub-model (2), up to the prespecified precision Δ and significance level α. **Note that in contrast to the hypothesis H_{MSE}, the hypotheses $H_{\Delta,n}$ refers to the relative quality of the submodel (2) compared to model (1), and not to the quality of specific estimators such as LS estimators.** The test of $H_{\Delta,n}$ versus $K_{\Delta,n}$ will be based on asymptotic normality of an appropriate test statistic and therefore it does not require normality of the homoskedastic errors in the regression model specified in (1).

Section 2 deals with testing hypotheses related to $H_{\Delta,n}$. In Section 2.1 we introduce an asymptotic version H_Δ of $H_{\Delta,n}$, a test statistic for H_Δ and derive its asymptotic distribution. Section 2.2 is concerned with a nested testing situation, where we test a hypothesis $\tilde{H}_\Delta$, related to H_Δ, for model (2) against model (1) in the presence of another larger super-model containing model (1) as its sub-model. The results have their main application for deriving model selection method for identifying strong regression effects. In Section 3 we define quantities $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$, which can be interpreted as estimated maximal relative errors (with level α) that one makes when using the smaller sub-model. These are very convenient for model-selection purposes, and in Section 3 we discuss how they can be employed in a backward selection procedure.

Model selection criteria are classically divided into consistent and conservative criteria. In certain situations such as nested models containing the true model, consistent criteria like the BIC, choose the correct model, i.e. the minimal model that contains all covariates with $\beta_i \neq 0$, with a probability converging to 1. In these cases conservative criteria, like the AIC or Mallows' C_p, asymptotically choose models that are too large with a positive probability, but not models that are too small. Of course, criteria such as AIC have other advantages, e.g. when the linear model is misspecified, see Burnham and Anderson (2002) and Section 6 for further discussion.

In finite samples and in cases where there are many covariates with small but nonzero influence, both conservative and consistent criteria will typically include some but not all of these covariates in a somewhat arbitrary way. In contrast, model selection based on a threshold value of $\hat{D}_{\alpha,n}$ or $\tilde{D}_{\alpha,n}$ allows to discard covariates with a small influence in a controlled and interpretable way, namely as long as the relative error that arises remains below the chosen threshold t. Thus, model selection based on $\hat{D}_{\alpha,n}$ or $\tilde{D}_{\alpha,n}$ does not aim at finding all the covariates with nonzero influence. Rather, its goal is to find the relevant covariates in terms of maximal relative error.

The actual performance of $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$ for model selection purposes, as compared to the AIC, the BIC, the Lasso and the relaxed Lasso is investigated in Section 4 in an extensive simulation study. Here we also give illustrative examples for selection methods based on $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$, respectively and investigate the quality of the normal approximation of the test statistics introduced in Section 2. Further in Section 5, we illustrate the practical usefulness of our method by analyzing US college spending data from 1994. Finally, Section 6 closes the paper with conclusions and a discussion on future research. Technical assumptions and proofs are deferred to an appendix.

2. Asymptotic tests for identifying large regression effects

As an illustration, suppose for the moment that the errors are normally distributed, and let $SSE(\mathbf{b})$ denote the error sum of squares of an estimator $\mathbf{b}$ of $\boldsymbol{\beta}$. Then the statistic

$$T = \frac{SSE(\hat{\boldsymbol{\beta}}_r) - SSE(\hat{\boldsymbol{\beta}})}{q\hat{\sigma}^2}, \qquad \text{where } \hat{\sigma}^2 = \frac{1}{n-(p+q)}\left(\mathbf{Y} - X\boldsymbol{\beta}\right)'\left(\mathbf{Y} - X\boldsymbol{\beta}\right),$$

is F-distributed with degrees of freedom q and $(n-(p+q))$ and non-centrality parameter λ, given in (3). Since H_0 is equivalent to $\lambda = 0$ and H_{MSE} to $\lambda \leq 1$, it is then straightforward to construct tests for H_0 and H_{MSE} based on the F-distribution of T.

However, $H_{\Delta,n}$ is equivalent to $H_{\lambda,n} : \lambda > n\Delta/\sigma^2$. Since σ^2 is unknown we cannot construct even under normality an exact test of $H_{\Delta,n}$. However we can construct an asymptotic test for a limiting test hypothesis of $H_{\Delta,n}$, which does not require normal errors in the regression model (1). In Section 2.1, we consider this limiting version of the hypotheses $H_{\Delta,n}$, and construct an asymptotic test for this hypothesis in case the larger model is correct. Section 2.2 gives a corresponding test in case model (1) to which to (2) is compared is also incorrect, assuming that there is some larger valid super-model. In this situation, and in case of normally distributed errors, a test for H_{MSE} is discussed in Teräsvirta and Mellin (1986).

2.1. Testing when the larger model is correct

In this section we consider testing model (2) against the larger model (1) assuming that the larger model (1) is correct and the errors are not necessarily normally distributed. First we introduce an asymptotic version of $H_{\Delta,n}$. To do so, we consider the following condition under which $d_n(\boldsymbol{\beta}_2)$ converges as $n \to \infty$, say to $d(\boldsymbol{\beta}_2)$.

Assumption 1. The regressors X are non-random and we have $X'X/n \to G$ as $n \to \infty$, where $G \in \mathbb{R}^{(p+q)\times(p+q)}$ is a symmetric positive definite matrix.

Split G into blocks as follows

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}, \quad G_{11} \in \mathbb{R}^{p\times p}, \quad G_{22} \in \mathbb{R}^{q\times q}, \quad G_{12} = G_{21}' \in \mathbb{R}^{p\times q}.$$

Note that $d_n(\boldsymbol{\beta}_2)$ can be rewritten as

$$\begin{aligned} d_n(\boldsymbol{\beta}_2) &= \frac{1}{n}\boldsymbol{\beta}_2' X_2'(I_n - X_1(X_1'X_1)^{-1}X_1')X_2\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_2' \left[\frac{X_2'X_2}{n} - \frac{X_2'X_1}{n} \left(\frac{X_1'X_1}{n} \right)^{-1} \frac{X_1'X_2}{n} \right] \boldsymbol{\beta}_2 \end{aligned}$$

and from Assumption 1 it follows that

$$d_n(\boldsymbol{\beta}_2) \to d(\boldsymbol{\beta}_2) := \boldsymbol{\beta}_2'(G_{22} - G_{21}G_{11}^{-1}G_{12})\boldsymbol{\beta}_2 \quad \text{as} \quad n \to \infty. \tag{5}$$

Under Assumption 1, G is positive definite, which implies that $G_{22} - G_{21}G_{11}^{-1}G_{12}$ is also positive definite. Therefore we consider the following asymptotic version of test problem (4) given by

$$H_\Delta \;:\; d(\boldsymbol{\beta}_2) > \Delta \quad \text{against } K_\Delta \;:\; d(\boldsymbol{\beta}_2) \leq \Delta. \tag{6}$$

The test statistic R_n for (6) is now derived in the standard manner. To do this, we substitute the unknown $\boldsymbol{\beta}_2$ in $d(\boldsymbol{\beta}_2)$ by its consistent LS estimate $\hat{\boldsymbol{\beta}}_2$ from model (1), i.e. $R_n := d_n(\hat{\boldsymbol{\beta}}_2)$. It is not difficult to see that the test statistic R_n is also a normalized numerator of the $F-$statistic for $H_0 : \boldsymbol{\beta}_2 = 0$, i.e.

$$R_n = \frac{1}{n}\big(MSE(\hat{\boldsymbol{\beta}}_r) - MSE(\hat{\boldsymbol{\beta}})\big) = \frac{1}{n}\mathbf{Y}'\big(P_X - P_{X_1}\big)\mathbf{Y},$$

with $P_X := X(X'X)^{-1}X'$. In Theorem 1 we show that R_n is an asymptotic unbiased estimator of $d\left(\boldsymbol{\beta}_2\right)$ and derive the asymptotic distribution of R_n when $\boldsymbol{\beta}_2 \neq 0$.

Theorem 1.

(i) Suppose that in model (1) with independent zero mean homoscedastic errors **A***ssumption 1 is satisfied. Then $E(R_n) \to d(\boldsymbol{\beta}_2)$ as $n \to \infty$.*

(ii) Suppose that in model (1) with independent zero mean homoscedastic errors **A***ssumptions 3, 4 and 5 (cf. the appendix) are satisfied. If $d(\boldsymbol{\beta}_2) > 0$ we have that $\sqrt{n}\big(R_n - d(\boldsymbol{\beta}_2)\big) \xrightarrow{\mathcal{L}} N\big(0, 4\sigma^2 d(\boldsymbol{\beta}_2)\big)$ as $n \to \infty$.*

The proof of Theorem 1 is given in the appendix. Using Theorem 1, we construct an asymptotic test for $H_\Delta : d(\boldsymbol{\beta}_2) > \Delta$ versus $K_\Delta : d(\boldsymbol{\beta}_2) \leq \Delta$ as follows. Given $\Delta > 0$, reject H_Δ with level $\alpha > 0$ if

$$R_n \leq \Delta + 2\hat{\sigma}u_\alpha \sqrt{\Delta}/\sqrt{n}, \tag{7}$$

where u_α denotes the α-quantile of the standard normal distribution.

Remark 1.

(i) The results of Theorem 1 can be generalized to heteroscedastic linear models with scaled errors from a distribution having finite $(4+\delta)$*-th absolute moment as follows. Let*

$$\mathbf{Z} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}} = \tilde{X}_1\boldsymbol{\beta}_1 + \tilde{X}_2\boldsymbol{\beta}_2 + \tilde{\boldsymbol{\epsilon}}$$

be a heteroscedastic linear model, where $\tilde{\boldsymbol{\epsilon}} := (\sigma_1\epsilon_1, \ldots, \sigma_n\epsilon_n)'$ *with* $\epsilon_1, \ldots, \epsilon_n$ *being i.i.d. with mean zero and variance* 1*. Further let* $D := diag(\sigma_1^2, \ldots, \sigma_n^2)$ *and* $D^{-1/2} := diag(1/\sigma_1, \ldots, 1/\sigma_n)$*, where* $diag(a_1, \ldots, a_n)$ *denotes a diagonal matrix with* $a_1, \ldots, a_n$ *on the main diagonal. Since the linear model* $D^{-1/2}\mathbf{Z} = D^{-1/2}\tilde{X}_1\boldsymbol{\beta}_1 + D^{-1/2}\tilde{X}_2\boldsymbol{\beta}_2 + D^{-1/2}\tilde{\boldsymbol{\epsilon}}$ *has now homoscedastic errors, the statement of the corresponding theorem as well as its proof for the heteroscedastic errors are straightforward from Theorem 1 and its proof, respectively. Thus the corresponding statistics* R_n^{heter} *is given by*

$$R_n^{heter} := \mathbf{Z}'\left(D^{-1}\tilde{X}[\tilde{X}'D^{-1}\tilde{X}]^{-1}\tilde{X}'D^{-1} - D^{-1}\tilde{X}_1[\tilde{X}_1'D^{-1}\tilde{X}_1]^{-1}\tilde{X}_1'D^{-1}\right)\mathbf{Z}/n$$

and Assumption 1 is, for example, reformulated as $\tilde{X}'D^{-1}\tilde{X} \to G$*.*

(ii) Statement (i) of Theorem 1 remains valid if the design matrix X *contains random regressors independent of error vector* $\boldsymbol{\epsilon}$ *and* $E(X'X)/n \to G$*. This can be seen in decomposition* (A.4) *from the Appendix, where* $E(S_1|X)$ *and* $E(S_2|X)$ *would be independent of* X*. However statement (ii) on convergence of asymptotic distribution does not hold, since the leading term* S_2 *in* (A.4) *of asymptotics is a sum of dependent variables of unknown structure.*

2.2. *Testing when having a valid super-model*

Theorem 1 is only valid if the larger model (1) is correct. However, if we apply the test sequentially, then we will possibly also have the situation where the larger model is not true either, because we already excluded covariates with too small an influence, which are however non-zero. Therefore, we now study the situation where the larger model is also not correct. However, we assume that at least the linear regression model with all covariates is correct, thus, there is a valid super-model.

Changing the notation slightly, suppose that we already erroneously believe that at most the following sub-model

$$\mathbf{Y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{8}$$

of the true super-model

$$\mathbf{Y} = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad Z = [X_1, X_2, X_3], \quad \boldsymbol{\beta} = [\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3']', \tag{9}$$

contains all the relevant covariates, and that we want to check the validity of the smaller sub-model $Y = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. to Note that $X_1 \in \mathbb{R}^{n\times p}$, $X_2 \in \mathbb{R}^{n\times q_1}$, $X_3 \in \mathbb{R}^{n\times q_2}$, $\boldsymbol{\beta}_1 \in \mathbb{R}^p$, $\boldsymbol{\beta}_2 \in \mathbb{R}^{q_1}$, $\boldsymbol{\beta}_3 \in \mathbb{R}^{q_2}$. Further the errors ϵ_i's are independent identically distributed (i.i.d.) with $E(\epsilon_1 = 0)$, $Var(\epsilon_1) = \sigma^2$ and not necessarily normal. For convenience let $X := [X_1, X_2]$ and let $Q_{X,X_1} := P_X - P_{X_1}$. Now define

$$\tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) := \frac{1}{n}\left(\boldsymbol{\beta}_2'X_2' + \boldsymbol{\beta}_3'X_3'\right)Q_{X,X_1}\left(X_2\boldsymbol{\beta}_2 + X_3\boldsymbol{\beta}_3\right). \tag{10}$$

The quantity $\tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ is the normalized (by a factor of $1/n$) length of the $n-$vector $X_2\boldsymbol{\beta}_2 + X_3\boldsymbol{\beta}_3$ when orthogonally projected by Q_{X,X_1} onto the orthogonal complement of the space spanned by the column vectors of X_1 in the space spanned by those of X. As in Section 2.1 we impose here the following assumption on the design matrix Z.

Assumption 2. The regressors Z are non-random and we have $Z'Z/n \to G$ as $n \to \infty$, where $G \in \mathbb{R}^{(p+q_1+q_2)\times(p+q_1+q_2)}$ is a symmetric positive definite matrix.

Split G into blocks as follows

$$G = \begin{pmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{pmatrix},$$

where $G_{11} \in \mathbb{R}^{p\times p}$, $G_{22} \in \mathbb{R}^{q_1\times q_1}$, $G_{33} \in \mathbb{R}^{q_2\times q_2}$, $G_{12} = G'_{21} \in \mathbb{R}^{p\times q_1}$, $G_{13} = G'_{31} \in \mathbb{R}^{p\times q_2}$ and $G_{23} = G'_{32} \in \mathbb{R}^{q_1\times q_2}$. Under Assumption 2 it follows also for $n \to \infty$ that

$$X'X/n \to \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} =: G_{r_{1,2}}.$$

Lemma 1. *Let* **A***ssumption* (2) *hold for the design matrix Z of model* (9)*. Then*

$$\tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) \to \tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) := \boldsymbol{\beta}_2' A \boldsymbol{\beta}_2 + \boldsymbol{\beta}_3' B A^{-1} B' \boldsymbol{\beta}_3 + 2\boldsymbol{\beta}_3' B \boldsymbol{\beta}_2 \quad as \quad n \to \infty,$$

where

$$A := G_{22} - G_{21}G_{11}^{-1}G_{12}, \qquad B := G_{32} - G_{31}G_{11}^{-1}G_{12}. \tag{11}$$

The proof of Lemma 1 is given in the appendix.

For $\Delta > 0$ Lemma 1 allows to consider a testing problem

$$\tilde{H}_\Delta : \tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) \geq \Delta \qquad \text{against} \qquad \tilde{K}_\Delta : \tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) < \Delta, \tag{12}$$

which is an asymptotic version of

$$\tilde{H}_{\Delta,n} : \tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) \geq \Delta \qquad \text{against} \qquad \tilde{K}_{\Delta,n} : \tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) < \Delta.$$

The test statistic $\tilde{R}_n$ for (12) is derived by substituting $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ in $\tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ with their consistent LS estimates $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_3$ based on model (9), i.e. $\tilde{R}_n := \tilde{d}_n(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3)$. It should be noted that the test statistic $\tilde{R}_n$ can be rewritten as

$$\tilde{R}_n = \frac{1}{n}\left(MSE(\hat{\boldsymbol{\beta}}_{r_1}) - MSE(\hat{\boldsymbol{\beta}}_{r_{1,2}})\right) = \frac{1}{n}\mathbf{Y}'\left(P_X - P_{X_1}\right)\mathbf{Y},$$

where $\hat{\boldsymbol{\beta}}_{r_1}$ is the restricted LS estimator in model (2) and $\hat{\boldsymbol{\beta}}_{r_{1,2}}$ the restricted LS estimator in model (8). Here additional zeros have been added to obtain the same dimension as $\boldsymbol{\beta}$.

Theorem 2.

(i) Suppose that in model (9), **A***ssumption 2 is satisfied. Then $E(\tilde{R}_n) \to \tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ as $n \to \infty$.*

(ii) Suppose that in model (9), **A***ssumptions 3, 6 and 7 (cf. the appendix) are satisfied. If $\tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) > 0$ we have that $\sqrt{n}\left(\tilde{R}_n - \tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)\right) \xrightarrow{\mathcal{L}} N\left(0, 4\sigma^2\tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)\right)$ as $n \to \infty$.*

Theorem 2 allows to construct an asymptotic test for the testing problem (12). Indeed, given $\Delta > 0$, reject $\tilde{H}_\Delta$ versus $\tilde{K}_\Delta$ with level $\alpha > 0$ if

$$\tilde{R}_n \leq \Delta + 2\hat{\sigma}u_\alpha\sqrt{\Delta}/\sqrt{n}, \tag{13}$$

where $\hat{\sigma}$ is an estimate of σ based on the full super-model (9).

Remark 2.

(i) The proof of Theorem 2 is similar to the proof of Theorem 1 and is therefore omitted.

(ii) The results of Theorem 2 can be generalized to heterogeneous linear models in the same way as in Theorem 1 (compare to Remark 1 (i)).

(iii) Statement (i) of Theorem 2 remains valid if the design matrix Z contains random regressors independent of error vector $\boldsymbol{\epsilon}$ and $E(Z'Z)/n \to G$.

3. Model validation and model selection

Test decisions based on (7) or (13) will obviously strongly depend on the choice of Δ. For example, in (7), Δ is a threshold for $d(\boldsymbol{\beta}_2)$, the limit of the distance $d_n(\boldsymbol{\beta}_2)$, which as mentioned above measures the normalized (with factor n^{-1}) squared distance of the projected vector $X_2\boldsymbol{\beta}_2$. This has to be seen in relation with the total normalized squared length $\boldsymbol{\beta}'X'X\boldsymbol{\beta}/n$, and thus, a general recommendation for a numerical value of Δ (like 0.1) does not make sense. Therefore, in the following we suggest objective procedures for choosing Δ, and discuss how the resulting tests can be used for model selection purposes within a backward selection procedure.

3.1. *Model selection based on $\hat{D}_{\alpha,n}$*

First consider the hypothesis H_Δ in (6) for arbitrary Δ. Using (7) one can, for a given level α (e.g. $\alpha = 0.05$), determine a critical threshold $\hat{\Delta}_{\text{crit}}(\alpha, n)$ for which $H_{\hat{\Delta}_{\text{crit}}(\alpha,n)}$ can be rejected at level α, while H_Δ cannot be rejected for $\Delta < \hat{\Delta}_{\text{crit}}(\alpha, n)$, i.e. $\hat{\Delta}_{\text{crit}}(\alpha, n)$ is defined as

$$\hat{\Delta}_{\text{crit}}(\alpha, n) := \left(\left(R_n + \hat{\sigma}^2 u_\alpha^2/n\right)^{1/2} - \hat{\sigma} u_\alpha/\sqrt{n}\right)^2.$$

Then we suggest to normalize $\hat{\Delta}_{\text{crit}}(\alpha, n)$ by an estimate of the total normalized squared length and take square roots to obtain

$$\hat{D}_{\alpha,n} := \left(\frac{\hat{\Delta}_{\text{crit}}(\alpha, n)}{\hat{\boldsymbol{\beta}}' X' X \hat{\boldsymbol{\beta}}/n}\right)^{1/2},$$

where $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$ in (1). The quantity $\hat{D}_{\alpha,n}$ can be nicely interpreted as the estimated maximal relative error one makes (with level α) if one uses sub-model (2) instead of the full model (1). In fact, one has

$$\hat{D}_{\alpha,n} \to \left[d(\boldsymbol{\beta}_2)/(\boldsymbol{\beta}' G \boldsymbol{\beta})\right]^{1/2} := D$$

in probability as $n \to \infty$. Note that D depends on the unknown regression vector $\boldsymbol{\beta}$ which for brevity is suppressed in the following. Variable identification with strong effects on the response now proceeds in terms of $\hat{D}_{\alpha,n}$: If $\hat{D}_{\alpha,n}$ is less than some fixed value which we allow as maximal relative error (say 0.1), we identify the variables included in the smallest sub-model as variables with strong effects.

Let us describe how the above method can be used in a backward selection procedure. After fixing the level α, we compute $\hat{D}_{\alpha,n}$ for all sub-models of the full model (1) which exclude one covariate. Let the sub-model with minimal $\hat{D}_{\alpha,n}$, denoted by $\hat{D}^1_{\alpha,n}$, be M_1. Next we compute $\hat{D}_{\alpha,n}$, relative to the full model (1), for all sub-models of M_1 which exclude a further covariate. Let the sub-model among the sub-models of M_1 with minimal $\hat{D}_{\alpha,n}$, denoted by $\hat{D}^2_{\alpha,n}$, be M_2. Let us stress that in each step we compute $\hat{D}_{\alpha,n}$ relative to the full model (which we assume to be correct so that we have no misspecification), since we possibly already excluded covariates with small but still notable influence. In this way we obtain a decreasing sequence of sub-models $M_1 \supset M_2 \supset \ldots$ with increasing sequence $\hat{D}^1_{\alpha,n} \leq \hat{D}^2_{\alpha,n} \leq \ldots$ with corresponding relative errors w.r.t. the full model (1). One can now choose the model from the sequence M_i, e.g. as the model for which the relative error is just below some threshold t (e.g. 0.1). Note that we use $\hat{D}_{\alpha,n}$, which is always normalized by the same factor $\hat{\boldsymbol{\beta}} X' X \boldsymbol{\beta}/n$. Therefore, the order $M_1 \supset M_2 \supset \ldots$ in which the model is reduced is the same as for the F-test and the AIC and BIC, only the stopping is more transparent as it is based on the maximal relative error. We further discuss this in the simulation study in Section 4.

3.2. *Model selection based on $\tilde{D}_{\alpha,n}$*

For the testing problem $\tilde{H}_\Delta$ versus $\tilde{K}_\Delta$ in (12) and for a given level α, the threshold $\tilde{\Delta}_{\text{crit}}(\alpha, n)$ for which $\tilde{H}_{\Delta_{\text{crit}}(\alpha,n)}$ can be rejected at level α, while $\tilde{H}_\Delta$ cannot be rejected for $\Delta < \tilde{\Delta}_{\text{crit}}(\alpha, n)$, is determined by using (13) as follows:

$$\tilde{\Delta}_{\text{crit}}(\alpha, n) := \left(\left(\tilde{R}_n + \hat{\sigma}^2 u_\alpha^2/n\right)^{1/2} - \hat{\sigma} u_\alpha/\sqrt{n}\right)^2,$$

where $\hat{\sigma}$ is an estimator of σ in (9). We suggest to normalize $\tilde{\Delta}_{\text{crit}}(\alpha, n)$ by an estimate of the total normalized length in model (8)

$$\tilde{D}_{\alpha,n} := \left(\frac{\tilde{\Delta}_{\text{crit},\alpha,n}}{\hat{\boldsymbol{\beta}}'_{r_{1,2}} X' X \hat{\boldsymbol{\beta}}_{r_{1,2}}/n}\right)^{1/2},$$

where $\hat{\boldsymbol{\beta}}_{r_{1,2}}$ is the LSE in (8). The interpretation of $\tilde{D}_{\alpha,n}$ is also that of an estimated relative error, since $\tilde{d}(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ only involves the normalized length of $Z\boldsymbol{\beta}$ when orthogonally projected onto the orthogonal complement of the column space of X_1 in that of $X = [X_1, X_2]$. However, it takes into account the whole contribution of $X_2\boldsymbol{\beta}_2 + X_3\boldsymbol{\beta}_3$, and is therefore not exclusively relative to the intermediate model (8). Note that $\tilde{D}_{\alpha,n} \to \sqrt{d(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)/\boldsymbol{\beta}'_{r_{1,2}} G_{r_{1,2}} \boldsymbol{\beta}_{r_{1,2}}} =: \tilde{D}$ in probability as $n \to \infty$. Note that $\tilde{D}$ depends on $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, which we omit in the sequel.

Table 1: Simulation setups

Scenarios for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_6)'$		Sample size n	
1	$\boldsymbol{\beta} = (2, 2, 2, 2, 0.1, 0.1, 0.1)'$	a)	200
2	$\boldsymbol{\beta} = (10, 5, 5, 1, 1, 0.05, 0.05)'$	b)	400
3	$\boldsymbol{\beta} = (10, 15, 7, 3, 1.5, 0.7, 0.3)'$	c)	1000
		d)	5000
Error distribution			
i)	t-distribution with 5 degrees of freedom		
ii)	standard normal distribution		

The backward selection procedure based on $\tilde{D}_{\alpha,n}$ proceeds as follows. After fixing the level α and a threshold value t for $\tilde{D}_{\alpha,n}$, we compute $\tilde{D}_{\alpha,n}$ for all sub-models of the correct super-model M_0 which exclude one covariate. Let the sub-model of M_0 with minimal $\tilde{D}_{\alpha,n}$, denoted by $\tilde{D}^1_{\alpha,n}$, be $\tilde{M}_1$. Next we compute $D_{\alpha,n}$, relative to $\tilde{M}_1$, for all sub-models of $\tilde{M}_1$ which exclude a further covariate. Thus here we are possibly already in the situation where the intermediate model is incorrect, since we compare relative to the (possibly already incorrect) sub-model $\tilde{M}_1$. Let the sub-model among the sub-models of $\tilde{M}_1$ with minimal $\tilde{D}_{\alpha,n}$, denoted by $\tilde{D}^2_{\alpha,n}$, be $\tilde{M}_2$. In this way we obtain a decreasing sequence of sub-models $\tilde{M}_1 \supset \tilde{M}_2 \supset \dots$ with (not necessarily increasing) relative errors $\tilde{D}^1_{\alpha,n}, \tilde{D}^2_{\alpha,n}, \dots$. Now one chooses the first model from the sequence $\tilde{M}_i$ for which $\tilde{D}^i_{\alpha,n}$ is below the threshold t, while $\tilde{D}^{i+1}_{\alpha,n}$ is larger than the threshold t. Note that the denominator of $D^i_{\alpha,n}$ changes in each step, since the intermediate models (8) are changing. This implies that the order in which the models are reduced $\tilde{M}_1 \supset \tilde{M}_2 \supset \dots$ may be different from that for the AIC, the BIC and the method based on $\hat{D}_{\alpha,n}$.

4. Simulation study

In the simulation study we consider three different scenarios for the regression effects. First we choose strong equal regression effects (Scenario 1), one set of variables with very strong and one set with medium strength regression effects (Scenario 2) and finally regression effects which vary from very strong to weak effects (Scenario 3). The specific values of the regression coefficients are given in Table 1 where β_0 is an intercept.

For $i = 1, \dots, 6$ the covariate vector $\mathbf{x}_i$ is drawn independently from a uniform distribution on $[-1, 1]^n$. The $n-$dimensional unit vector $\mathbf{x}_0$ corresponds to the intercept β_0. For the corresponding design matrix $X = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6]$ Assumption 1 (Assumption 2) is satisfied with

$$X'X/n \to G = \operatorname{diag}(1, 1/3, \dots, 1/3) \qquad \left(Z'Z/n \to G = \operatorname{diag}(1, 1/3, \dots, 1/3)\right).$$

As error distribution we choose a $t-$distribution with 5 degrees of freedom (df) and the standard normal distribution. Note that the errors ϵ_i for $i = 1, \dots, n$ should be scaled by a factor SF in such a way that observations with signal to noise ratio (SNR) larger than 2 approximately form 40%–50% of the whole data. Here SNR is just a ratio of mean to standard deviation of an observation. Finally sample sizes $n = 200, 400, 1000$ and 5000 are investigated. Thus for each combination of $\boldsymbol{\beta}$, n and an error distribution given in Table 1 we apply our methods.

In Section 4.1 we illustrate in detail how our methods work in Scenario 1 and 2 for regression coefficients. Section 4.2 contains a simulation study in which we investigate the performance of model selection based on $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$ as compared to the standard model selection criteria AIC and BIC. We give results for two choices of the threshold t and discuss its choice. Finally in Section 4.3 we investigate the quality of the normal approximation in Theorems 1 and 2 for $t-$distributed errors. In all model selection methods, we use backward elimination and always keep the intercept, thus only choosing from the covariates.

4.1. Illustrating examples

Consider Scenario 1 where the vector of regression coefficients $\boldsymbol{\beta}$ is set as $\boldsymbol{\beta} = (2, 2, 2, 2, 0.1, 0.1, 0.1)'$. Now we chose $n = 200$ and $t-$distributed errors with 5 df scaled by the factor $SF = \sqrt{0.8}$. This results that

Table 2: Results of model selection procedure based on the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods for a data set of size $n = 200$ simulated with $\boldsymbol{\beta} = (2, 2, 2, 2, 0.1, 0.1, 0.1)'$ and $t-$distributed errors with 5 df scaled by $SF = \sqrt{0.8}$. The level α is equal to 0.05.

step i	sub-model	discarded cov.	$\hat{D}^i_{\alpha,n}$	D	$\tilde{D}^i_{\alpha,n}$	$\tilde{D}$
1	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$	$\mathbf{x}_6$	0.106	0.020	0.106	0.020
2	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5$	$\mathbf{x}_4$	0.107	0.028	0.107	0.020
3	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_5$	0.113	0.035	0.112	0.020
4	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_3$	$\mathbf{x}_2$	0.488	0.040	0.488	0.408
5	$\mathbf{x}_0, \mathbf{x}_3$	$\mathbf{x}_3$	0.643	0.578	0.503	0.447

Table 3: Results of model selection procedure based on the $\hat{D}_{\alpha,n}-$method for a data set simulated with $\boldsymbol{\beta} = (10, 5, 5, 1, 1, 0.05, 0.05)'$ and $t-$distributed errors with 5 df scaled by $SF = 4$. Note $\alpha = 0.05, \boldsymbol{\beta}_1 = (10, 5, 5)', \boldsymbol{\beta}_2 = (1, 1, 0.05, 0.05)'$ and $D \approx 0.08$.

Step i	Sub-Model	discarded covariate	$\hat{D}^i_{\alpha,n}$			
			$n =$			
			200	400	1000	5000
1	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_6$	$\mathbf{x}_5$	0.120		0.048	
	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5$	$\mathbf{x}_6$		0.076		0.022
2	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_5$		0.080		0.030
	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$	$\mathbf{x}_6$	0.120		0.057	
3	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_4$	$\mathbf{x}_3$				0.061
	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$	$\mathbf{x}_4$	0.125	0.105	0.070	
4	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_3$	0.147	0.127	0.094	
	$\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$	$\mathbf{x}_4$				0.081
5	$\mathbf{x}_0, \mathbf{x}_2$	$\mathbf{x}_1$	0.350	0.263		
	$\mathbf{x}_0, \mathbf{x}_1$	$\mathbf{x}_2$			0.293	0.272

the variance of scaled errors is equal to 1.33. Table 2 contains the results of our model selection procedures for one sample, together with the estimated $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$ as wells as with the true values D and $\tilde{D}$. Models which contain $\mathbf{x_1}, \mathbf{x_2}$ and $\mathbf{x_3}$ result in estimated relative error ($\hat{D}_{\alpha,n}$ or $\tilde{D}_{\alpha,n}$) around 0.1, however models which miss one of $\mathbf{x_1}, \mathbf{x_2}$ and $\mathbf{x_3}$ have estimated relative error larger than 0.4. Thus, there are only two reasonable choices for possible models. Either one is willing to except a relative error at about 0.2 and only keeps $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$, or one keeps all the covariates.

For the second illustration we consider Scenario 2 for regression coefficients, $t-$distributed errors with 5 df and sample sizes $n = 200, 400, 1000$ and 5000. Note that this scenario has several orders of magnitude for the regression coefficients β's. The scaling factor SF is set to 4 and it ensures for example that 48% observations have a SNR>2 for $n = 200$. Table 3 contains the results of the model selection procedure for one sample, together with the values of $\hat{D}_{\alpha,n}$. Here we are interested in identifying the large effects, namely covariates $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$ in the presence of moderate effects $\mathbf{x}_3$ and $\mathbf{x}_4$. Thus the regression vector $\boldsymbol{\beta}$ is split as $(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1 = (10, 5, 5)'$ and $\boldsymbol{\beta}_2 = (1, 1, 0.05, 0.05)'$. Further the true theoretical value of D for identifying $\boldsymbol{\beta}_1$ is equal to 0.08. If we would now use the threshold $t = 0.1$ then the important covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ should be chosen by the $\hat{D}_{\alpha,n}-$method. In each column of Table 3 the value of $\hat{D}^i_{\alpha,n}$ is bolded as soon as $D^i_{\alpha,n} > 0.1$ for a first time. This implies that the method chooses the previous model above. For $n = 200$ our model selection procedure with $t = 0.1$ does not choose any sub-model with 5 covariates and therefore the full model cannot be simplified. When now sample sizes increases to 400, then the sub-model with $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ and $\mathbf{x}_4$ is identified which contains the medium regression effects $\mathbf{x}_3$ and $\mathbf{x}_4$. For $n = 1000$ and $n = 5000$ the desired sub-model is detected with threshold $t = 0.1$, even though the empirical values of $\hat{D}_{\alpha,n}$ are slightly larger then the true value $D = 0.8$. This example shows that for small and moderate

Table 4: Number of times specific sub-models are chosen using $\hat{D}_{\alpha,n}$, $\tilde{D}_{\alpha,n}$, BIC and AIC among 1000 simulated data sets with $\boldsymbol{\beta} = (2,2,2,2,0.1,0.1,0.1)'$. The first column identifies models including an intercept through their sub-indices of covariates. We want to identify covariates $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$. Note $\boldsymbol{\beta}_1 = (2,2,2,2)'$, $\boldsymbol{\beta}_2 = (0.1,0.1,0.1)'$, $D \approx 0.04$ for the $\hat{D}_{\alpha,n}-$method and $\boldsymbol{\beta}_1 = (2,2,2,2)', \boldsymbol{\beta}_2 = 0.1, \boldsymbol{\beta}_3 = (0.1,0.1)'$, $\tilde{D} \approx 0.02$ for the $\tilde{D}_{\alpha,n}-$method.

Model	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.1$		$t = 0.15$		$t = 0.1$		$t = 0.15$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
t-errors with 5 degrees of freedom ($SF = \sqrt{0.8}$)												
"12345"	283	3	4	0	271	0	1	0	3	5	50	105
"12356"	271	11	4	1	261	6	2	1	2	16	60	103
"12346"	289	5	5	0	281	3	3	0	3	3	56	96
"1234"	44	118	54	1	41	59	23	0	64	92	165	155
"1235"	34	88	49	2	40	38	21	0	45	65	129	134
"1236"	46	76	42	0	49	29	15	0	50	61	132	145
"123"	33	699	842	996	57	865	935	999	833	758	408	262
Normal errors ($SF = \sqrt{4/3}$)												
"12345"	318	5	1	0	313	1	0	0	1	7	44	100
"12356"	277	4	3	0	272	1	1	0	3	6	55	123
"12346"	305	4	3	0	296	1	1	0	4	4	64	84
"1234"	26	106	48	0	28	38	18	0	48	70	163	159
"1235"	31	98	48	0	33	34	22	0	47	66	150	145
"1236"	23	100	47	0	23	41	15	0	40	73	113	133
"123"	20	683	850	1000	35	884	943	1000	857	774	411	256

sample sizes n such as 200 or 400 a correction for threshold value t is needed. We discuss this point in the next section in more detail.

4.2. *Model selection performance*

Here we report results of an extensive simulation study to compare the performance of our model selection criteria based on the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods with the AIC, the BIC, the Lasso and the relaxed Lasso. For the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods, in each simulation we choose the level $\alpha = 0.05$ and sample size 200 as wells as 400.

Table 4 displays the frequency of chosen sub-models for 1000 data sets simulated with $\boldsymbol{\beta} = (2,2,2,2,0.1,0.1,0.1)'$, a fixed design matrix X and the two error distributions from Table 1. The $t-$errors are scaled with $SF = \sqrt{0.8}$ while the normal errors are scaled with $SF = \sqrt{4/3}$. This implies that the both type of scaled errors have the same variance and 44% of observations have the SNR>2 for $n = 200$. Obviously, this regression model has three covariates $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ (model "123" in Table 4) corresponding to $\boldsymbol{\beta}_1 = (2,2,2,2)'$ which we want to identify. The true value of D ($\tilde{D}$) for identifying these is equal to 0.04 (0.02). However as we noticed in the previous section the threshold value t should be corrected for small and moderate sample sizes. Therefore we chose 0.1 or 0.15 as the threshold t for the $\hat{D}_{\alpha,n}-$method, which are obtained by rounding $3 \cdot 0.04$ and $4 \cdot 0.04$ to the nearest numbers 0.1, 0.15 or 0.2. From our experience the true relative error D ($\tilde{D}$) for choosing the threshold t for the $\hat{D}_{\alpha,n}-$method ($\tilde{D}_{\alpha,n}-$method) should be multiplied by the factor 2, 3 or even 4 (3,4 or even 5) for sample sizes 200 or 400 when about 50% observations have SNR larger than 2 and only 6 covariates are under consideration. For both methods we use the same threshold values in order to illustrate the difference between them.

For $t-$errors both methods choose in over 99% the sub-model with three important covariates $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ when $t = 0.15$ and $n = 400$. They clearly outperform BIC (AIC), which rather chooses the sub-model with $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ only in 76% (26%). If the sample size decreases from 400 to 200 then the $\hat{D}_{\alpha,n}-$method is comparable with BIC and outperforms AIC. The $\tilde{D}_{\alpha,n}-$ method performs here clearly better than BIC and AIC. If the threshold $t = 0.1$ then our methods are comparable with BIC only for $n = 400$ while for $n = 200$ their performance is very poor with respect to BIC and AIC. It should be noted that when a sample size increases then BIC and AIC start to detect the small and medium regression effects. In contrast, in this situation the precision of our methods increases and they start to choose a smaller model with all large regression effects. For normal errors we get similar results as for the $t-$error case.

We performed simulations for regression coefficients from Scenario 2 and the same design matrix X in a similar manner. The $t-$errors are now scaled with $SF = 4$ while the normal errors are scaled with $SF = 4\sqrt{5/3}$. This ensures that the scaled errors have the same variance and 48% of the observations have SNR>2 for $n = 200$. Here we would like to identify strong regression effects $\mathbf{x}_1$ and $\mathbf{x}_2$ (model "12" in Table 5) corresponding to $\boldsymbol{\beta}_1 = (10, 5, 5)'$ in the presence of the moderate effect regressors $\mathbf{x}_3$ and $\mathbf{x}_4$. The true value of D ($\tilde{D}$) for identifying them is equal to 0.08 (0.05). We set the threshold t for both methods to 0.15 and 0.2 which are argued by multiplying the true relative error $D = 0.08$ by 2 and 3 and then rounding them to the nearest numbers $0.1, 0.15$ or 0.2.

Table 5 displays how often sub-models have been selected for 1000 data sets simulated according to the above setup. We see that the correction of the true relative error by multiplying with 2 ($t = 0.15$) is not enough for $n = 200$. To see a better performance of our methods in comparison with BIC and AIC, the sample size should be increased to 400. The correction by the factor 3 shows that our methods clearly outperform BIC and AIC for the both sample sizes. A change of the error distribution does not change the results by much. If a larger proportion of observations has a SNR larger than 2 then a correction factor 2 becomes acceptable. Table 6 shows the same results as in Table 5 but with different scale factors SF's. The values for SF are chosen in such a way that 80% of observations have SNR>2. We see that for the both error distributions our methods identify the desired model with a correction factor 2 or 3 for the true threshold correctly, thus improving the precision of the methods. In contrast BIC and AIC start to choose strong and intermediate regression effects together. A change of the error distribution does not change the results by much.

Finally consider a scenario with no clearly separated orders of magnitudes for the β's. We choose $\boldsymbol{\beta} = (10, 15, 7, 3, 1.5, 0.7, 0.3)'$, i.e. Scenario 3 in Table 1. Thus regression effects decrease without having strong separation of effects. The $t-$errors are now scaled with $SF = 4$ while the normal errors are scaled with $SF = 4\sqrt{5/3}$. The scaled errors for both error distributions thus have the same variance and 47.5% of the observations have SNR>2 for $n = 200$. Now we want to identify covariates $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ (model "123" in Table 7) corresponding to $\boldsymbol{\beta}_1 = (\beta_0, \beta_1, \beta_2, \beta_3)'$. The true value of D ($\tilde{D}$) for identifying them is equal to 0.07 (0.06), which we use below to illustrate our methods at work. As above, we set the threshold as $t = 0.15$ and $t = 0.2$ which are motivated by multiplying 0.07 by 2 and 3. Table 7 contains the results of the model selection procedures for 1000 data sets simulated within Scenario 3. Here we see that a correction factor 2 ($t = 0.15$) works well for our both methods and they outperform BIC and AIC for sample sizes 200 and 400. Thus for $n = 200$ the $\hat{D}_{\alpha,n}-$method ($\hat{D}_{\alpha,n}-$method) chooses the sub-model with $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ in 64% (74%) while the BIC does so in 35%. If now the threshold 0.2 is used then the $\hat{D}_{\alpha,n}-$method identifies the desired model still better than the BIC (59% versus 35%) while $\tilde{D}_{\alpha,n}-$method fails to do this (25% versus 35%). Both our methods fail to identify the model with $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ if a slightly smaller threshold t than above is used. Thus for $t = 0.12$, $n = 200$ and $t-$errors the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods favor a larger model with $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ in 552 and 482 cases out of 1000, respectively. This indicates that when there are no clear strong effects then our methods become sensitive to the choice of the correction factor for the desired relative errors D and $\tilde{D}$ and this correction should be done carefully.

Following the referees' suggestion we also compared our model selection methods with the Lasso of Tibshirani (1996) and the relaxed Lasso of Meinshausen (2007). These two methods are implemented in the non-commercial statistical software `R` as packages `lasso2` (Lasso) and `relaxo` (relaxed Lasso). Thus we applied the Lasso and the relaxed Lasso to the same data from simulation Scenarios 1, 2 and 3 with $t-$errors corresponding to Tables 4, 5 and 7. Table 8 presents the model selection performance of both methods with penalty parameter chosen by cross-validation (s. Tibshirani (1996) and Meinshausen (2007)). It is well known that a model selection based on Lasso estimators is not a consistent variable selection procedure and chooses a larger set of covariates, however, containing important ones with probability 1 (s. Meinshausen (2007)). A comparison of the results for the Lasso from Table 8 with the corresponding results from Tables 4, 5 and 7 also exposes this fact, since the Lasso always prefers the full model with all six covariates. In contrast, in Scenario 1 the model selection based on the relaxed Lasso prefers the

Table 5: Number of times specific sub-models are chosen using $\hat{D}_{\alpha,n}$, $\tilde{D}_{\alpha,n}$, BIC and AIC among 1000 simulated data sets with $\boldsymbol{\beta} = (10, 5, 5, 1, 1, 0.05, 0.05)'$. The first column identifies models including an intercept through their sub-indices of covariates. We want to identify covariates $\mathbf{x}_0, \mathbf{x}_1$ and $\mathbf{x}_2$. Note $\boldsymbol{\beta}_1 = (10, 5, 5)'$, $\boldsymbol{\beta}_2 = (1, 1, 0.05, 0.05)'$, $D \approx 0.08$ for the $\hat{D}_{\alpha,n}-$method and $\boldsymbol{\beta}_1 = (10, 5, 5)', \boldsymbol{\beta}_2 = 1, \boldsymbol{\beta}_3 = (1, 0.05, 0.05)'$, $\tilde{D} \approx 0.05$ for the $\tilde{D}_{\alpha,n}-$method.

Model	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.15$		$t = 0.2$		$t = 0.15$		$t = 0.2$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
t-errors with 5 degrees of freedom ($SF = 4$)												
“12345”	10	1	0	0	7	0	0	0	3	5	40	99
“12356”	6	1	0	0	4	1	0	0	1	0	6	9
“12346”	12	0	0	0	11	0	0	0	3	4	60	102
“12456”	7	0	0	0	2	0	0	0	1	0	7	8
“1234”	130	3	3	0	70	0	0	0	58	179	230	435
“1235”	19	0	0	0	7	0	0	0	1	5	28	18
“1236”	26	0	1	1	10	0	0	0	6	2	32	22
“1245”	22	0	1	0	8	0	0	0	3	5	38	21
“1246”	29	0	2	0	13	0	0	0	5	3	41	11
“1256”	2	0	0	0	1	0	0	0	0	0	7	1
“123”	211	112	72	2	180	14	12	0	169	203	131	116
“124”	251	112	82	0	215	27	23	0	225	237	203	119
“125”	26	5	1	0	6	0	1	0	8	5	21	5
“126”	22	2	5	0	5	0	0	0	1	7	17	7
“12”	227	764	833	997	461	958	964	1000	516	345	139	27
Normal errors ($SF = \sqrt{80/3}$)												
“12345”	5	0	0	0	1	0	0	0	0	3	39	94
“12356”	5	0	0	0	1	0	0	0	0	0	8	7
“12346”	4	0	0	0	1	0	0	0	0	1	49	95
“12456”	3	0	0	0	1	0	0	0	0	0	8	7
“1234”	129	0	1	0	70	0	0	0	64	184	246	436
“1235”	19	0	0	0	12	0	0	0	8	4	37	21
“1236”	16	0	0	0	6	0	0	0	5	6	21	24
“1245”	25	0	0	0	10	0	0	0	5	4	37	17
“1246”	38	0	0	0	10	0	0	0	8	5	45	18
“1256”	4	0	0	0	2	0	0	0	0	1	4	3
“123”	228	92	70	1	190	12	11	0	176	187	152	101
“124”	273	98	78	0	234	10	18	0	233	217	202	130
“125”	27	4	6	0	10	0	0	0	10	3	15	5
“126”	26	5	3	0	14	0	0	0	9	13	20	6
“12”	198	801	842	999	438	978	970	1000	482	372	117	36

model with significant covariates $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$ in 610 (466) cases out of 1000 for $n = 200$ ($n = 400$). In the presence of strong as well as intermediate regression effects (Scenario 2), our methods and BIC outperform the relaxed Lasso. In Scenario 3 the relaxed Lasso is consistent with the BIC for $n = 200$ and as n increases to 400 a larger model with five covariates is favored. It should be noted that in the presence of regression effects of different magnitude, BIC, AIC, Lasso and relaxed Lasso start to include more and more covariates in a model as the sample size increases while our approaches, contrarily, choose a favored model in a small sample with higher probability when n gets larger.

Now we discuss some of our results using a false discovery rate (FDR) criterion for multiple testing problems introduced by Benjamini and Hochberg (1995). This criterion has been found especially useful when a large number of covariates are under consideration. Therefore its use became quite popular for

Table 6: Number of times specific sub-models are chosen using $\hat{D}_{\alpha,n}$, $\tilde{D}_{\alpha,n}$, BIC and AIC among 1000 simulated data sets with $\boldsymbol{\beta} = (10, 5, 5, 1, 1, 0.05, 0.05)'$. The first column identifies models including an intercept through their sub-indices of covariates. We want to identify covariates $\mathbf{x}_0, \mathbf{x}_1$ and $\mathbf{x}_2$. Note $\boldsymbol{\beta}_1 = (10, 5, 5)'$, $\boldsymbol{\beta}_2 = (1, 1, 0.05, 0.05)'$, $D \approx 0.08$ for the $\hat{D}_{\alpha,n}-$method and $\boldsymbol{\beta}_1 = (10, 5, 5)'$, $\boldsymbol{\beta}_2 = 1$, $\boldsymbol{\beta}_3 = (1, 0.05, 0.05)'$, $\tilde{D} \approx 0.05$ for the $\tilde{D}_{\alpha,n}-$method.

Model	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.15$		$t = 0.2$		$t = 0.15$		$t = 0.2$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
t-errors with 5 degrees of freedom ($SF = 2$)												
“12345”	0	0	0	0	0	0	0	0	14	25	129	160
“12356”	0	0	0	0	0	0	0	0	0	0	0	1
“12346”	0	0	0	0	0	0	0	0	11	10	147	148
“12456”	0	0	0	0	0	0	0	0	0	0	3	0
“1234”	0	0	0	0	0	0	0	0	621	919	637	690
“1235”	0	0	0	0	0	0	0	0	3	0	4	0
“1236”	0	0	0	0	0	0	0	0	2	2	5	0
“1245”	0	0	0	0	0	0	0	0	2	0	8	0
“1246”	0	0	0	0	0	0	0	0	4	0	4	0
“123”	13	1	0	0	0	0	0	0	116	11	19	0
“124”	18	0	0	0	2	0	0	0	197	31	43	1
“125”	0	0	0	0	0	0	0	0	1	0	0	0
“12”	969	999	1000	1000	998	1000	1000	1000	29	2	1	0
Normal errors ($SF = \sqrt{20/3}$)												
“12345”	0	0	0	0	0	0	0	0	19	14	139	154
“12356”	0	0	0	0	0	0	0	0	0	0	2	0
“12346”	0	0	0	0	0	0	0	0	11	25	124	153
“12456”	0	0	0	0	0	0	0	0	0	0	3	0
“1234”	0	0	0	0	0	0	0	0	618	921	663	691
“1235”	0	0	0	0	0	0	0	0	3	0	7	0
“1236”	0	0	0	0	0	0	0	0	6	0	5	0
“1245”	0	0	0	0	0	0	0	0	2	0	0	0
“1246”	0	0	0	0	0	0	0	0	5	1	7	1
“123”	8	0	0	0	1	0	0	0	125	16	18	0
“124”	11	0	0	0	1	0	0	0	188	22	32	1
“12”	981	1000	1000	1000	998	1000	1000	1000	23	1	0	0

microarray data (see e.g. Drigalenko and Elston (1997)), where an experimenter aims to detect a few genes relevant to a disease among many ten thousands or even hundred thousands of genes. In order to introduce the FDR criterion the related quantities such as the number of true positives (TP), the number of false negatives (FN), the number of false positives (FP) should be characterized. In a model selection procedure the relevant covariates to the response can be identified or not. Now the number of TP describes the number of correctly identified relevant covariates while the number of FN describes the number of relevant covariates which are not identified. Their sum results in the number p of the important covariates for the response (dimension of $\boldsymbol{\beta}_1$). Similarly, non-relevant covariates can be wrongly identified as relevant or not. The number FP is just the number of non-relevant regression effects which are wrongly identified as important. Now FDR is defined as follows

$$FDR := \frac{FP}{FP + TP}.$$

Thus FDR measures the rate of false discoveries among all discoveries. A low FDR with TP close to p

Table 7: Number of times specific sub-models are chosen using $\hat{D}_{\alpha,n}$, $\tilde{D}_{\alpha,n}$, BIC and AIC among 1000 simulated data sets with $\boldsymbol{\beta} = (10, 15, 7, 3, 1.5, 0.7, 0.3)'$. The first column identifies models including an intercept through their sub-indices of covariates. We want to identify covariates $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}_3$. Note $\boldsymbol{\beta}_1 = (10, 15, 7, 3)'$, $\boldsymbol{\beta}_2 = (1.5, 0.7, 0.3)'$, $D \approx 0.07$ for the $\hat{D}_{\alpha,n}-$method and $\boldsymbol{\beta}_1 = (10, 15, 7, 3)', \boldsymbol{\beta}_2 = 1.5, \boldsymbol{\beta}_3 = (0.7, 0.3)', \tilde{D} \approx 0.06$ for the $\tilde{D}_{\alpha,n}-$method

Model	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.15$		$t = 0.2$		$t = 0.15$		$t = 0.2$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
t-errors with 5 degrees of freedom ($SF = 4$)												
“12345”	4	0	0	0	3	0	0	0	57	156	297	496
“12356”	0	1	0	0	0	1	0	0	1	2	23	9
“12346”	3	0	0	0	1	0	0	0	16	26	127	146
“1234”	257	17	5	0	129	3	1	0	499	648	433	337
“1235”	37	2	1	0	8	0	0	0	50	33	41	6
“1236”	5	0	0	0	0	0	0	0	8	4	13	1
“1245”	3	0	0	0	0	0	0	0	3	0	0	0
“123”	642	929	568	157	743	693	256	13	349	131	65	5
“124”	41	14	35	2	40	4	4	0	11	0	1	0
“125”	0	0	1	0	1	0	0	0	1	0	0	0
“12”	8	37	390	841	75	299	739	987	5	0	0	0
Normal errors ($SF = \sqrt{80/3}$)												
“12345”	2	0	0	0	1	0	0	0	59	172	290	531
“12356”	1	0	0	0	0	0	0	0	4	8	18	20
“12346”	0	0	0	0	0	0	0	0	19	22	118	123
“1234”	270	14	1	0	134	1	0	0	505	633	434	311
“1235”	34	2	0	0	8	0	0	0	49	31	43	9
“1236”	5	0	0	0	0	0	0	0	7	9	16	2
“1245”	1	0	0	0	1	0	0	0	0	0	0	0
“1246”	2	0	0	0	0	0	0	0	2	0	0	0
“123”	643	932	593	152	753	697	255	11	345	125	80	4
“124”	28	7	39	0	30	0	3	0	8	0	1	0
“125”	1	0	1	0	0	0	0	0	0	0	0	0
“12”	13	45	366	848	73	302	742	989	2	0	0	0

indicates a good performance of the method. In Table 9 we give averaged values of TP, FN and FP for all investigated methods based on 1000 simulated data sets when errors are distributed according to the scaled $t-$distribution with 5 df. The FDR given in Table 9 is then computed using the average values of FP and TP. For Scenario 1 when $\boldsymbol{\beta} = (2, 2, 2, 2, 0.1, 0.1, 0.1)'$ the BIC performs better than AIC for $n = 200$ and 400. Our methods clearly outperform BIC for $n = 400$ and $t = 0.15$. For $n = 200$ a threshold of $t = 0.1$ is too small to achieve a better performance than BIC. For Scenario 2 when $\boldsymbol{\beta} = (10, 5, 5, 1, 1, 0.05, 0.05)'$ we obtain similar results, i.e. $t = 0.2$ and $n = 400$ outperform BIC. For Scenario 3 when $\boldsymbol{\beta} = (10, 15, 7, 3, 1.5, 0.7, 0.3)'$ we see a difference. Here a threshold $t = 0.15$ is outperforming BIC and AIC. However the average TP value is lower for our methods compared to AIC and BIC, i.e. too small models are identified by our methods. For normal errors we obtained similar results. Therefore we omit them for brevity. Overall we see that our tailored methods to identify relevant covariates outperform all purpose model selection criteria such as the BIC, the AIC, the Lasso and the relaxed Lasso.

4.3. Quality of the normal approximation

Finally, we investigate the quality of the normal approximations in Theorems 1 and 2. Since the true value of $d(\boldsymbol{\beta}_2)$ and $d(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ are usually not known we use $d_n(\boldsymbol{\beta}_2)$ and $d_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ instead of $d(\boldsymbol{\beta}_2)$ and

Table 8: Number of times specific sub-models are chosen using Lasso (L) and relaxed Lasso (RL) methods for scaled $t-$distributed errors with 5 df.

Model	Scenario 1				Scenario 2 ($SF = 4$)				Scenario 3			
	L		RL		L		RL		L		RL	
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
“123456”	365	513	0	0	253	338	0	0	516	704	0	0
“12345	175	125	58	97	185	219	94	127	259	205	285	430
“12356	128	131	42	97	32	16	13	10	8	3	5	17
“12346	158	128	52	86	216	209	100	150	137	65	102	115
“12456	0	0	0	0	36	35	16	19	1	0	0	0
“1234”	49	36	116	74	122	111	142	266	67	22	416	366
“1235”	54	28	61	80	24	6	13	9	1	1	17	10
“1236”	47	30	61	100	18	16	11	17	8	0	3	2
“1245”	0	0	0	0	35	17	22	13	0	0	0	0
“1246”	0	0	0	0	35	19	22	13	0	0	0	0
“1256”	0	0	0	0	8	0	0	2	0	0	0	0
“123”	24	9	610	466	10	7	95	84	3	0	168	60
“124”	0	0	0	0	16	5	165	127	0	0	1	0
“125”	0	0	0	0	6	0	9	3	0	0	0	0
“126”	0	0	0	0	2	1	5	3	0	0	0	0
“12”	0	0	0	0	2	1	293	157	0	0	3	0

$d(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ in Theorem 1 and Theorem 2, respectively. By virtue of Assumption 4 and Slutsky's theorem the above change does not affect the limiting normal distribution.

For Theorem 1, we test the complete model (1) against the model given in (2). We use $\boldsymbol{\beta} = (2, 2, 0.1, 0.1, 0.1, 2, 2)'$ and $t-$errors with 5 df scaled by the factor $SF = \sqrt{0.8}$. The design matrix X is constructed in a similar manner as in the previous sections. Model 2 is defined by excluding the covariate x_6 corresponding to the regression coefficient $\beta_6 = 2$. We simulate the statistic R_n 10000 times for $n = 50, 100$ and 200. For visualization in Figure 1 we use P-P plots. They show $\alpha \in [0, 1]$ on the $y-$axis and the empirical probability $\hat{\alpha}_n$ of the event $\{\sqrt{n}[R_n - d_n(\boldsymbol{\beta}_2)] \leq Q_{\alpha,n}\}$ on the $x-$axis. Here $Q_{\alpha,n}$ is the α-quantile of the asymptotic normal distribution of Theorem 1 with consistently estimated variance $4\hat{\sigma}^2 R_n$. Note that $\hat{\sigma}^2$ is the estimate of the error variance σ^2 in model (1).

Similarly for Theorem 2, we test a sub-model without the covariate x_6 against a sub-model where the covariate x_5 and x_6 are excluded, and simulate the statistic $\tilde{R}_n$ 10000 times for $n = 50, 100$ and 200 and scaled $t-$ distributed errors with 5 df as above. Thus $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ from Theorem 2 are both equal to 2. Figure 2 shows for each $\alpha \in [0, 1]$ on the $y-$axis the empirical probability $\hat{\alpha}_n$ of the event $\{\sqrt{n}[\tilde{R}_n - \tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)] \leq \tilde{Q}_{\alpha,n}\}$ on the $x-$axis. Here $\tilde{Q}_{\alpha,n}$ is the α-quantile of the asymptotic normal distribution of Theorem 2 with consistently estimated variance $4\hat{\sigma}^2\tilde{R}_n$. Note that $\hat{\sigma}^2$ is the estimate of the error variance σ^2 in model (9). ¿From the top row of Figures 1 and 2 we see that the asymptotic approximation is quite good already for rather small sample sizes. Note that for the test decisions (7) and (13), the approximations for small α's are relevant, which can be assessed using the bottom row.

5. College spending data

To illustrate our method in a practical application we analyze the college spending data from U.S. News and World Report 1994 College Guide. The complete data can be found in Dielman (1996) and its short description is given in Table 10. The variable of interest is educational spending per full-time equivalent (SPEND) given for 147 US colleges. A simple explorative data analysis shows that there is a presence of variance heterogeneity and a log transformation of the response SPEND is needed. Further, for numerical

Table 9: Averaged values of true positives (TP), false negatives (FN), false positives (FP) and the corresponding false discovery rate (FDR) based on the averaged values among 1000 simulated data sets for $t-$errors with 5 degrees of freedom.

$\boldsymbol{\beta} = (2, 2, 2, 2, 0.1, 0.1, 0.1)'$, $SF = \sqrt{0.8}$ and desired sub-model $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$												
	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.1$		$t = 0.15$		$t = 0.1$		$t = 0.15$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
TP	4	4	4	4	4	4	4	4	4	4	4	4
FN	0	0	0	0	0	0	0	0	0	0	0	0
FP	1.81	0.320	0.171	0.005	1.756	0.144	0.071	0.002	0.175	0.266	0.758	1.042
FDR	0.312	0.074	0.041	0.001	0.305	0.035	0.017	0.000	0.042	0.062	0.159	0.207
$\boldsymbol{\beta} = (10, 5, 5, 1, 1, 0.05, 0.05)'$, $SF = 4$ and desired sub-model $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$												
	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.15$		$t = 0.2$		$t = 0.15$		$t = 0.2$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
TP	3	3	3	3	3	3	3	3	3	3	3	3
FN	0	0	0	0	0	0	0	0	0	0	0	0
FP	1.071	0.243	0.174	0.004	0.696	0.044	0.696	0	0.573	0.867	1.463	1.917
FDR	0.263	0.075	0.055	0.001	0.188	0.014	0.188	0	0.160	0.224	0.328	0.390
$\boldsymbol{\beta} = (10, 15, 7, 3, 1.5, 0.7, 0.3)'$, $SF = 4$ and desired sub-model $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$												
	$\hat{D}_{\alpha,n}$				$\tilde{D}_{\alpha,n}$				BIC		AIC	
	$t = 0.15$		$t = 0.2$		$t = 0.15$		$t = 0.2$					
	$n =$		$n =$		$n =$		$n =$		$n =$		$n =$	
	200	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*	*200*	*400*
TP	3.948	3.949	3.574	3.157	3.884	3.697	3.257	3.013	3.980	4	3.999	4
FN	0.052	0.051	0.426	0.843	0.116	0.303	0.743	0.987	0.020	0	0.001	0
FP	0.360	0.035	0.042	0.002	0.186	0.009	0.005	0	0.723	1.053	1.382	1.646
FDR	0.084	0.009	0.012	0.001	0.046	0.002	0.002	0	0.154	0.208	0.257	0.292

Table 10: Variables of college spending data in USA from 1994

Notation	**Short description**
SAT	–average SAT score
TOP10	–freshmen in the top 10% of their high school class (in percentage)
ACCRATE	–acceptance rate (in percentage)
PHD	–faculty with PhD (in percentage)
RATIO	–student faculty ratio
SPEND	–educational spending per full-time equivalent student (in dollars)
GRADRATE	–graduation rate (in percentage)
ALUMNI	–alumni giving rate (in percentage)

stability, all variables including the response log(SPEND) are centered and normalized by their sample mean and sample standard deviation.

For our methods we set the desired relative errors D and $\tilde{D}$ equal to 0.1 and choose the nominal level $\alpha = 0.05$. Further we use the correction factor 3 for D and $\tilde{D}$ which works for $n = 200$ well as we have seen in the previous section. This results in the threshold value $t = 0.3$ for the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods. In the top part of Table 11, the results of a backward selection procedure for the $\hat{D}_{\alpha,n}-$method, the BIC and

Figure 1: P-P plots for $\sqrt{n}[R_n - d_n(\boldsymbol{\beta}_2)]$ based on 10000 replications when $\boldsymbol{\beta}_1 = (2, 2, 0.1, 0.1, 0.1, 2)'$, $\boldsymbol{\beta}_2 = 2$ and errors are scaled $t-$distributed with 5 df (top row $\alpha \in (0, 1)$, bottom row $\alpha \in (0, 0.1)$).

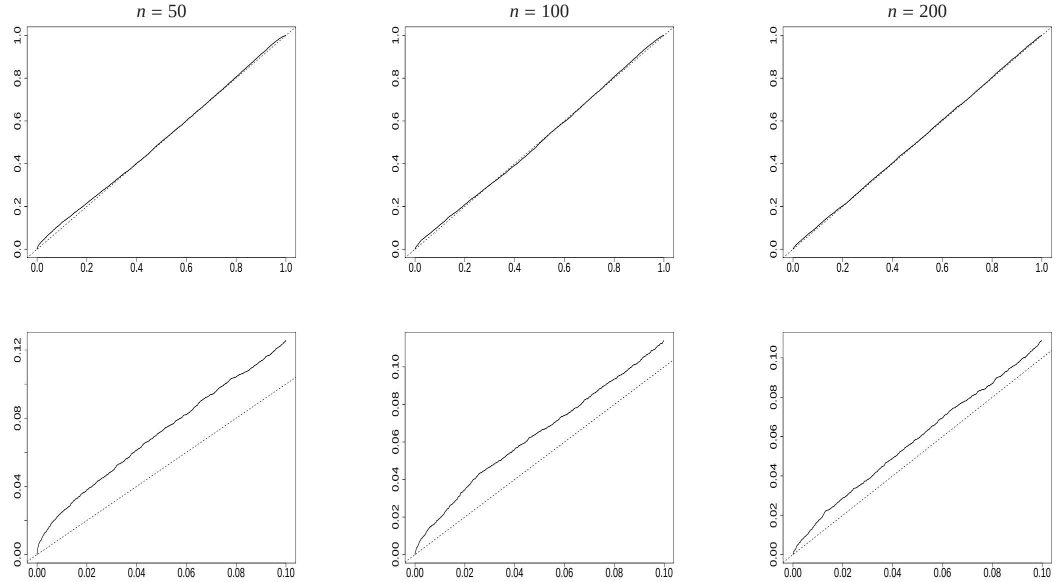
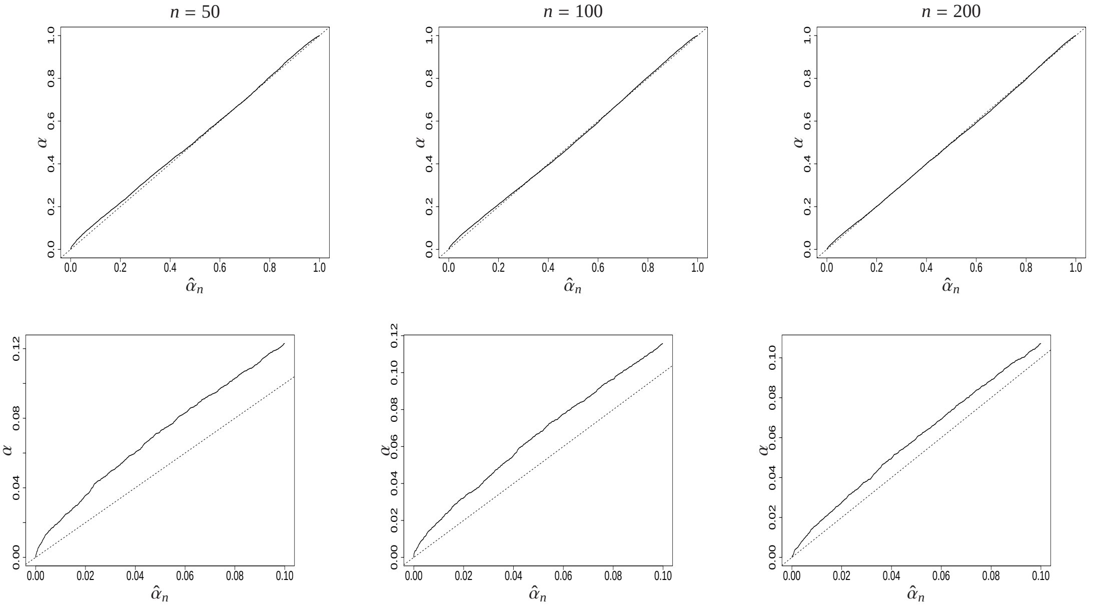


Figure 2: P-P plots for $\sqrt{n}[\tilde{R}_n - \tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)]$ based on 10000 replications when $\boldsymbol{\beta}_1 = (2, 2, 0.1, 0.1, 0.1)'$, $\boldsymbol{\beta}_2 = 2$, $\boldsymbol{\beta}_3 = 2$ errors are scaled $t-$distributed with 5 df (top row $\alpha \in (0, 1)$, bottom row $\alpha \in (0, 0.1)$).



the AIC, applied to the college spending data, are given.

As in the simulation study we always keep the intercept in the sub-models. One can see that the $\hat{D}_{\alpha,n}-$method, the BIC and the AIC have the same sequence of preferred sub-models when the number of the covariates decreases stepwise. However they choose completely different sub-models which are bolded in Table 11. The BIC chooses a sub-model consisting of the three covariates SAT, TOP10 and RATIO, and the AIC prefers a model with 4 covariates. Note that the values for the BIC for the sub-models with

Table 11: Results of a backward selection procedure for college spending data based on the $\hat{D}_{\alpha,n}-$, $\tilde{D}_{\alpha,n}-$methods, BIC and AIC.

Model selection based on $\hat{D}_{\alpha,n}$, BIC and AIC

step i	sub-model	discarded cov.	$\hat{D}^i_{\alpha,n}$	BIC	AIC
1	SAT, TOP10, ACCRATE, PHD, RATIO, ALUMNI	GRADRATE	0.164	253.7	229.8
2	SAT, TOP10, ACCRATE, PHD, RATIO	ALUMNI	0.165	248.8	227.9
3	SAT, TOP10, PHD,RATIO	ACCRATE	0.185	245.5	227.5
4	SAT, TOP10, RATIO	PHD	0.215	243.5	228.5
5	TOP10, RATIO	SAT	0.278	247.0	235.1
6	TOP10	RATIO	0.650	326.1	317.1

Model selection based on $\tilde{D}_{\alpha,n}$

step i	sub-model	discarded cov.	$\tilde{D}^i_{\alpha,n}$
1	SAT, TOP10, ACCRATE, PHD, RATIO, ALUMNI	GRADRATE	0.164
2	SAT, TOP10, ACCRATE, PHD, RATIO	ALUMNI	0.164
3	SAT, TOP10, PHD, RATIO	ACCRATE	0.389
4	SAT, TOP10, RATIO	PHD	0.394
5	SAT, TOP10	RATIO	0.484
6	SAT	TOP10	0.640

2 - 5 covariates, and the values of the AIC for 3-6 covariates are rather close together, thus making a clear decision in favor of any of these sub-models difficult. The $\hat{D}_{\alpha,n}-$method evidently chooses the sub-model with covariates TOP10 and RATIO.

The $\tilde{D}_{\alpha,n}-$method behaves somewhat differently as the bottom part of Table 11 displays. It prefers the sub-model with five covariates SAT, TOP10, ACCRATE, PHD and RATIO which are given in bold face. This sub-model is “almost” chosen by the AIC since the difference in AIC between 227.9 and 227.5 is negligible. Thus all four model selection procedures choose different sub-models consisting of 2-5 covariates which indicates a high model uncertainty in the college spending data. The $\tilde{D}_{\alpha,n}-$method is sensible to the uncertainty present and it results in the choice of the largest sub-model among all chosen sub-models. In the presence of model uncertainty the smaller sub-model with four covariates cannot be preferred to the larger sub-model with five covariates taking into account the existence of the super-model with all 6 covariates. In this situation the numerators of $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$ are approximately of the same order but the denominator of $\tilde{D}_{\alpha,n}$ is smaller than the denominator of $\hat{D}_{\alpha,n}$. Therefore we may observe large values of $\tilde{D}_{\alpha,n}$ compared to the corresponding ones of $\hat{D}_{\alpha,n}$ as for example Table 11 displays for steps $i = 3$ and 4. Note that the order in which the variables are discarded by the $\tilde{D}_{\alpha,n}-$method differs from that taken by the other methods, and $\tilde{D}_{\alpha,n}$ needs not be monotone. Finally we also applied Lasso and relaxed Lasso with optimal chosen penalty parameter to the college spending data. As expected the Lasso prefers the largest model among all methods, namely the full model reduced by GRADRATE. In contrast, the relaxed Lasso favors the model with SAT, TOP10 and RATIO chosen also by the BIC.

6. Conclusions and discussion

Model selection is one of the most important and difficult problems in statistics. Even in classical linear models with normal errors, where exact distributions of many statistics are known, there is no universal solution for selecting the relevant covariates. In this paper we solve this problem by means of relative errors $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$ for discarding non-relevant covariates from a starting full model with all covariates. In general, there are two reasonable approaches to measure these relative errors. The first approach consists in estimating a maximal relative error ($\hat{D}_{\alpha,n}$) of a sub-model with respect to the full model. In the second approach a maximal relative error ($\tilde{D}_{\alpha,n}$) of a sub-model is evaluated with respect to a larger model which is, in turn, nested under the full model. Obviously, the relative errors $\hat{D}_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$ allow for model selection strategies based on a threshold value t for them. These model selection methods, in contrast to many other selection criteria, have an interpretable distance (difference in relative errors) between compared sub-models, which makes them attractive.

The $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods rely on critical values of the tests given by (6) and (12). They do not require the normality of errors in a linear regression model and utilize the asymptotic normality of the test statistic for the above testing problems. The accuracy of the normal approximation in Theorems 1 and 2 is investigated in a simulation study. The corresponding PP-plots show that the asymptotic normal approximation can already be quite satisfactory for sample sizes $n = 50$ and 100.

A natural choice of the threshold t for the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods is a desired relative error that one allows when excluding non-relevant covariates. Our simulation study illustrates that for small and moderate small samples a correction of the desired relative errors D and $\tilde{D}$ is needed. For instance, in linear models with moderate covariate number and intercept for sample sizes $n = 200$ the desired relative errors D and $\tilde{D}$ should be multiplied by factor 3 or 4 when about 50% of observations have a SNR> 2. If a larger proportion of data have a SNR>2 then the correction factor can appropriately be decreased.

It is well known that classical model selection procedures such as the AIC and the BIC have certain optimality properties, i.e. in terms of a Bayesian a-posteriori rule (for the BIC) or in terms of efficiency (for the AIC). In contrast, our methods are designed to select a model with a certain maximal relative error. In this sense they intend to identify the most relevant covariates.

Theorems 1 and 2 remain valid if instead of LS estimators, distinct estimators such as ridge estimators are used to estimate the quantities $d_n(\boldsymbol{\beta}_2)$ and $\tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$. It would be of some interest to extend the approach of Toro-Vizcarrondo and Wallace (1968) in order to compare the performance of other estimators such as ridge estimates. Such generalization would allow for a construction of more general model selection methods than $D_{\alpha,n}$ and $\tilde{D}_{\alpha,n}$. This issue will be addressed in the future. Generalized linear models (GLM's) are an extension of classical linear models where the distribution of the response is a member of a general exponential family (see McCullagh and Nelder (1989)). In GLM's the AIC and the BIC are also widely used for model selection. It would be of substantial interest to have such interpretable model selection criteria for GLM's similar to the $\hat{D}_{\alpha,n}-$ and $\tilde{D}_{\alpha,n}-$methods. This is the subject of future research.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft under grants MU 1230/8-2, CZ 86/1-1 and HO 3260/2-1 is gratefully acknowledged. Hajo Holzman further gratefully acknowledges financial support of the Claussen-Simon Stiftung and of the Landesstiftung Baden-Württemberg, "Juniorprofessorenprogramm". The authors would like to thank the Associate Editor, two anonymous referees and Axel Munk for constructive comments.

Appendix A. Assumptions and proofs

Assumption 3. The errors $\epsilon_1, \dots \epsilon_n$ are i.i.d. with $E(\epsilon_1) = 0$, $Var(\epsilon_1) = \sigma^2$ and $E|\epsilon_1|^{4+\delta} < \infty$ for some $\delta > 0$.

Assumption 4. We have that

$$\sqrt{n}\big(n^{-1}X'X - G\big) \to 0 \quad \text{as} \quad n \to \infty. \tag{A.1}$$

Remark 3. Note that Assumption 4 is more restrictive than Assumption 1. In Assumption 4 we require the $o(n^{-1/2})$ rate of convergence in Assumption 1.

Assumption 5. The entries of the covariate matrix X_2 lie in a compact set $K \subset \mathbb{R}$ for all n.

Note that from Assumptions 4 and 5 it follows that

$$\sqrt{n}\Big[\big(\frac{1}{n}X'X\big)^{-1} - G^{-1}\Big] \to 0 \quad \text{as} \quad n \to \infty \tag{A.2}$$

since taking the inverse of a matrix is a Lipschitz continuous mapping on compact sets.

Proof of Theorem 1. (i) From Theil (1971, P. 146) it follows

$$P_X - P_{X_1} = Q_{X_1} X_2 (X_2' Q_{X_1} X_2)^{-1} X_2' Q_{X_1} =: Q_{X,X_1}, \tag{A.3}$$

where $Q_{X_1} = I_n - P_{X_1}$. The matrix Q_{X,X_1} is symmetric and idempotent and satisfies $Q_{X,X_1} X_1 = 0$. For convenience, here and in the sequel we drop the subindex of the matrix Q_{X,X_1}, i.e. we use Q instead of Q_{X,X_1}. Thus the statistic R_n can be decomposed as

$$\begin{aligned} R_n &= \frac{1}{n}\mathbf{Y}'(P_X - P_{X_1})\mathbf{Y} \\ &= \frac{1}{n}\boldsymbol{\epsilon}' Q \boldsymbol{\epsilon} + \frac{2}{n}\boldsymbol{\beta}_2' X_2' Q \boldsymbol{\epsilon} + \frac{1}{n}\boldsymbol{\beta}_2' X_2' Q X_2 \boldsymbol{\beta}_2 \\ &=: S_1 + S_2 + d_n(\boldsymbol{\beta}_2). \end{aligned} \tag{A.4}$$

Noting that $ES_1 = \sigma^2 tr(Q)/n = \sigma^2 q/n \to 0$ as $n \to \infty$ and $E(S_2) = 0$, the first statement of Theorem 1 now follows from (5).

(ii) First note that relationships (6) and (A.2) imply

$$\sqrt{n}\big(d_n(\boldsymbol{\beta}_2) - d(\boldsymbol{\beta}_2)\big) \to 0.$$

Since by assumption, $d(\boldsymbol{\beta}_2) > 0$, $d_n(\boldsymbol{\beta}_2)$ will be bounded away from 0 and we get

$$\sqrt{n}\,\frac{R_n - d(\boldsymbol{\beta}_2)}{2\sigma\sqrt{d_n(\boldsymbol{\beta}_2)}} = \sqrt{n}\,\frac{\frac{1}{n}\mathbf{Y}'(P_X - P_{X_1})\mathbf{Y} - d_n(\boldsymbol{\beta}_2)}{2\sigma\sqrt{d_n(\boldsymbol{\beta}_2)}} + o(1). \tag{A.5}$$

Consider now decomposition (A.4) for R_n. By virtue of Theorem 1.6 from Seber and Lee (2003), the variance of the first term S_1 in (A.4) is given by

$$Var\,(S_1) = \frac{1}{n^2}\left[(\mu_4 - 3\sigma^4)h'h + 2\sigma^4 tr(Q)\right],$$

where $\mu_4 := E(\epsilon_1^4)$ and h is the vector of diagonal elements of the matrix Q, for which $h'h \le q^2$. Thus

$$S_1 = O_P(|ES_1| + |S_1 - ES_1|) = O_P(n^{-1}).$$

Furthermore, $ES_2 = 0$ and

$$Var\,(S_2) = \frac{4}{n}\cdot\sigma^2 d_n(\boldsymbol{\beta}_2) \sim \frac{4}{n}\cdot\sigma^2 d(\boldsymbol{\beta}_2),$$

and therefore the term S_2 dominates the asymptotics in (A.4). It remains to show asymptotic normality of S_2. For this we check the Lyapounov condition

$$\frac{1}{n^{(4+\delta)/2}}\sum_{i=1}^{n} E\,|b_i\epsilon_i|^{4+\delta} = \frac{E|\epsilon_1|^{4+\delta}}{n^{2+\delta/2}}\sum_{i=1}^{n}|b_i|^3 \to 0 \quad \text{as} \quad n \to \infty,$$

where $\mathbf{b}' := 2\boldsymbol{\beta}_2' X_2' Q = (b_1, \ldots, b_n)$. It will be enough to show that the entries b_i divided by n are uniformly bounded. From Assumption 5 it follows that

$$\begin{aligned}
\max_{i=1,\ldots,n} \frac{|b_i|}{n} &= \frac{1}{n} \max_{i=1,\ldots,n} \left|[QX_2\boldsymbol{\beta}_2]_i\right| \\
&\leq \frac{1}{n} \max_{i=1,\ldots,n} \left\{ \sum_{k=1}^{n} |[Q]_{ik}| \cdot |[X_2\boldsymbol{\beta}_2]_k| \right\} \\
&\leq \frac{C}{n} \max_{i=1,\ldots,n} \left\{ \sum_{k=1}^{n} |[Q]_{ik}| \right\},
\end{aligned}$$

where $C > 0$ and $[\,\cdot\,]_{ik}$ denotes the (i,k)-th entry of the corresponding matrix. Since Q is symmetric and positive semi-definite, $|[Q]_{ik}| \leq (Q_{ii} + Q_{kk})/2$, and thus

$$\begin{aligned}
\max_{i=1,\ldots,n} \frac{|b_i|}{n} &\leq \frac{C}{n} \max_{i=1,\ldots,n} \left\{ \frac{1}{2} \sum_{k=1}^{n} ([Q]_{ii} + [Q]_{kk}) \right\} \\
&= \frac{C}{n} \max_{i=1,\ldots,n} \left\{ \frac{n}{2}[Q]_{ii} + \frac{1}{2} tr(Q) \right\} \\
&\leq \frac{C}{n} \frac{n+1}{2} tr(Q) \\
&\leq Cq \qquad \text{for} \quad n \geq 1
\end{aligned}$$

This finishes the proof of Theorem 1. □

Proof of Lemma 1. Using (10), (A.3) and noting $QX_1 = 0$, we have

$$\begin{aligned}
\tilde{d}_n(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3) &= \frac{1}{n}\boldsymbol{\beta}_2' X_2' Q_{X_1} X_2 \boldsymbol{\beta}_2 + \frac{2}{n}\boldsymbol{\beta}_3' X_3' Q_{X_1} X_2 \boldsymbol{\beta}_2 \\
&+ \frac{1}{n}\boldsymbol{\beta}_3' X_3' Q_{X_1} X_2 (X_2' Q_{X_1} X_2)^{-1} X_2' Q_{X_1} X_3 \boldsymbol{\beta}_3 \\
&=: T_1 + T_2 + T_3. \qquad \text{(A.6)}
\end{aligned}$$

Consider T_1 in (A.6). By virtue of Assumption 2 it follows that

$$\begin{aligned}
T_1 &:= \frac{1}{n}\boldsymbol{\beta}_2' X_2' Q_{X_1} X_2 \boldsymbol{\beta}_2 \\
&= \boldsymbol{\beta}_2' \left(\frac{X_2'X_2}{n} - \frac{X_2'X_1}{n} \left[\frac{X_1'X_1}{n} \right]^{-1} \frac{X_1'X_2}{n} \right) \boldsymbol{\beta}_2 \\
&\to \boldsymbol{\beta}_2' \left(G_{22} - G_{21} G_{11}^{-1} G_{12} \right) \boldsymbol{\beta}_2 \qquad \text{as} \quad n \to \infty \\
&= \boldsymbol{\beta}_2' A \boldsymbol{\beta}_2.
\end{aligned}$$

Similarly, one can show that

$$\begin{aligned}
T_2 &:= \frac{2}{n}\boldsymbol{\beta}_3' X_3' Q_{X_1} X_2 \boldsymbol{\beta}_2 \\
&\to 2\boldsymbol{\beta}_3' \left(G_{32} - G_{31} G_{11}^{-1} G_{12} \right) \boldsymbol{\beta}_2 \\
&= 2\boldsymbol{\beta}_3' B \boldsymbol{\beta}_2 \qquad \text{as} \quad n \to \infty
\end{aligned}$$

and

$$\begin{aligned}
T_3 &:= \frac{1}{n}\boldsymbol{\beta}_3' X_3' Q_{X_1} X_2 (X_2' Q_{X_1} X_2)^{-1} X_2' Q_{X_1} X_3 \boldsymbol{\beta}_3 \\
&\to \boldsymbol{\beta}_3' B A^{-1} B' \boldsymbol{\beta}_3 \qquad \text{as} \quad n \to \infty.
\end{aligned}$$

□

Assumption 6. We have that

$$\sqrt{n}\big(n^{-1}Z'Z - G\big) \to 0 \quad \text{as} \quad n \to \infty.$$

Remark 4. Note that Assumption 6 is more restrictive than Assumption 2. In Assumption 6 we require the $o(n^{-1/2})$ rate of convergence in Assumption 2.

Assumption 7. The entries of the covariate matrix $[X_2, X_3]$ lie in a compact set $K \subset \mathbb{R}$ for all n.

References

[1] Akaike, H., 1974. A new look at the statistical model identification. System identification and time-series analysis. IEEE Trans. Automatic Control 19, 716–723.

[2] Benjamini, Y. and Hochberg, Y., 1995. Controling the false discovery rate: a practical and powerful approach for multiple testing. J. R. Statist. Soc. B 57, 289–300.

[3] Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference. A practical information-theoretic approach. Springer-Verlag, New York.

[4] Claeskens, G., Hjort, N.L., 2003. The focused information criterion. J. Amer. Statist. Assoc. 98, 900–945.

[5] Dielman T.E., 1996. Applied regression analysis for business and economics. Duxbury Press, Belmont.

[6] Dette, H., Munk, A., 1998. Validation of linear regression models. Ann. Statist. 26, 778–800.

[7] Dette, H., Munk, A., 2003. Some methodological aspects of validation of models in nonparametric regression. Statist. Neerlandica 57, 207–244.

[8] Drigalenko, E.I., Elston, R.C., 1997. False discoveries in genome scanning. Genet. Epidemiol. 14, 779–784.

[9] Farebrother, R.W., 1975. The minimum mean square error linear estimator and ridge regression. Technometrics 17, 127–128.

[10] Johnson, N.L., Kotz, S., 1970. Distributions in statistics. Continuous univariate distributions **Volume** 2. Houghton Mifflin Co., Boston, Mass.

[11] Leeb, H, Pötscher, B., 2003. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. Econometric Theory 19, 100–142.

[12] McCullagh, P., Nelder. J.A., 1989. Generalized Linear Models. Chapman and Hall, London.

[13] Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.

[14] Meinshausen, N., 2007. Relaxed Lasso. Computational Statistics and Data Analysis 52, 374-393.

[15] Miller, A., 2002. Subset selection in regression. Chapman & Hall/CRC, Florida.

[16] Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

[17] Seber, G.A.F., Lee, A.J., 2003. Linear Regression Analysis. John Wiley & Sons, Hoboken, New Jersey.

[18] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. Royal Statist. Ser. B 64, 583–639.

[19] Teräsvirta, T., Mellin, I., 1986. Model selection criteria and model selection tests in regression models. Scand. J. Statist. 13, 159–171.

[20] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B. 58, 267-288.

[21] Theil, H., 1971. Principles of Econometrics. John Wiley & Sons, New York.

[22] Toro-Vizcarrondo, C., Wallace, T.D., 1968. A test of the mean square error criterion for restrictions in linear regression. J. Amer. Statist. Assoc. 63, 558–572.

[23] Wallace, T.D., 1972. Weaker criteria and tests for linear restrictions in regression. Econometrica 40, 689–698.

[24] Wang, H., Leng, C., 2008. A note on adaptive group LASSo. Computational Statistics and Data Analysis 52, 5277-5286.

[25] Yancey, T.A., Judge, G.G., Bock, M.E., 1973. Wallace's weak mean square error criterion for testing linear restrictions in regression: a tighter bound. Econometrica 41, 1203–1206.