Statistica Sinica Preprint No: SS-2023-0153

Title	KOO Approach for Scalable Variable Selection Problem in Large-dimensional Regression
Manuscript ID	SS-2023-0153
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0153
Complete List of Authors	Zhidong Bai, Kwok Pui Choi, Yasunori Fujikoshi and Jiang Hu
Corresponding Authors	Jiang Hu
E-mails	huj156@nenu.edu.cn

Notice: Accepted version subject to English editing.

Statistica Sinica

KOO APPROACH FOR SCALABLE VARIABLE SELECTION PROBLEM IN LARGE-DIMENSIONAL REGRESSION

Zhidong Bai[12], Kwok Pui Choi[3], Yasunori Fujikoshi[4] and Jiang Hu[1]*

[1] *KLASMOE and School of Mathematics & Statistics, Northeast Normal University, China,*

[2] *School of Mathematics and Statistics, Xi'an Jiaotong University, China*

[3] *Department of Statistics and Data Science, National University of Singapore, Singapore.*

[4] *Department of Mathematics, Graduate School of Science, Hiroshima University, Japan*

Abstract: An important issue in many multivariate regression problems is to eliminate candidate predictors with null predictor vectors. In large-dimensional (LD) setting where the numbers of responses and predictors are large, model selection encounters the scalability challenge. Knock-one-out (KOO) statistics hold promise to meet this challenge. In this paper, the almost sure limits and the central limit theorem of the KOO statistics are derived under the LD setting and mild distributional assumptions (finite fourth moments) of the errors by random matrix theory. These theoretical results guarantee the strong consistency of a subset selection rule based on the KOO statistics with a general threshold. For enhancing the robustness of the selection rule, we also propose a bootstrap threshold for the KOO approach. Simulation results support our

* Corresponding author.

conclusions and demonstrate the selection probabilities by the KOO approach with the bootstrap threshold outperform the methods using Akaike information threshold, Bayesian information threshold and Mallow's C_p threshold. We compare the proposed KOO approach with those based on information threshold to a chemometrics dataset and a yeast cell-cycle dataset, which suggests our proposed method identifies useful models.

Key words and phrases: High-dimensional Regression, AIC, BIC, Information criteria, Multi-response regression, KOO, Variable selection, RMT.

1. Introduction

In multivariate statistical analysis, linear regression is a basic and commonly used type of approach. The overall idea of regression is to examine which variables in particular are significant predictors of the outcome variables, and in what way do they indicated by the magnitude and sign of the outcome variables. Specifically,

$$\mathbf{Y} = \mathbf{X}\mathbf{\Theta} + \mathbf{E}\mathbf{\Sigma}^{1/2}, \tag{1.1}$$

where the $n \times p$ response matrix $\mathbf{Y} = (y_{ij}) = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$, the $n \times k$ predictor matrix $\mathbf{X} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)' = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, the $k \times p$ regression coefficient matrix $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$, the $n \times p$ random errors matrix $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_p) = (e_{ij})$, $(\mathbf{\Sigma}^{1/2})^2 = \mathbf{\Sigma}$, and the $p \times p$ covariance matrix $\mathbf{\Sigma}$ is of full rank. A main goal in multivariate linear regression (MLR) is to estimate the regression coefficients $\mathbf{\Theta}$. The estimates should be such that the estimated regression plane explains the variation in the values of the responses with great accuracy.

Model (1.1) (referred to hereinafter as the full model), however, is not always satisfactory because some of the predictors may be uncorrelated with the responses. We take a simple example to illustrate this fact. Let $\mathbf{j}$ be a subset of $[k] = \{1, 2, \ldots, k\}$, $\mathbf{X}_{\mathbf{j}} = (\mathbf{x}_j, j \in \mathbf{j})$ and $\mathbf{\Theta}_{\mathbf{j}} = (\boldsymbol{\theta}_j, j \in \mathbf{j})'$. Denote model $\mathbf{j}$ by

$$M_{\mathbf{j}}: \quad \mathbf{Y} = \mathbf{X}_{\mathbf{j}}\mathbf{\Theta}_{\mathbf{j}} + \mathbf{E}\mathbf{\Sigma}^{1/2}. \tag{1.2}$$

The classical linear least-squares solution is to estimate the matrix of regression coefficients $\widehat{\mathbf{\Theta}}$ of the full model (1.1) by

$$\widehat{\mathbf{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

which minimizes the sum of the squares of errors, i.e.,

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \operatorname{tr}(\mathbf{Y} - \mathbf{X}\mathbf{\Theta})(\mathbf{Y} - \mathbf{X}\mathbf{\Theta})'.$$

If there exists a predictor vector $\boldsymbol{\theta}_j = \mathbf{0}$, then the least-squares estimator of the regression coefficients of model $M_{[k]\backslash j}$ is

$$\widehat{\mathbf{\Theta}}_{[k]\backslash j} = (\mathbf{X}'_{[k]\backslash j}\mathbf{X}_{[k]\backslash j})^{-1}\mathbf{X}'_{[k]\backslash j}\mathbf{Y},$$

It is known that in this case the mean squared error (MSE) of the predictions from $\widehat{\mathbf{\Theta}}_{[k]\backslash j}$ is smaller than that from $\widehat{\mathbf{\Theta}}$ under some mild conditions. Moreover, even though the elements of $\boldsymbol{\theta}_j$ are not equal to zero but small enough, the MSE of the predictions from $\widehat{\mathbf{\Theta}}_{[k]\backslash j}$ is also smaller than that from $\widehat{\mathbf{\Theta}}$ (e.g., Fujikoshi et al. (2010)). Therefore, removing these "non-significant" predictors from the full model improves the model. How to determine the significance of each predictor for the response and to select the true model from the full model are important problems in multiple regression model. Here, the true model is the data-generating model and is denoted by

$$M_{\mathbf{j}_*}: \quad \mathbf{Y} = \mathbf{X}_{\mathbf{j}_*}\mathbf{\Theta}_{\mathbf{j}_*} + \mathbf{E}\mathbf{\Sigma}^{1/2}, \tag{1.3}$$

where for all $j \in [k]\backslash\mathbf{j}_*$, $\boldsymbol{\theta}_j = \mathbf{0}$.

To measure the significance of the predictors for the response, one can make use of the regression coefficients, the partial correlation or the multiple correlation coefficient between each predictor and the responses. However, these direct measures are unstable under high-dimensional regression because they all highly depend on the values of each predictor. Instead, we consider removing one predictor vector from the full model and measuring how much "information" we lose. Hence, we refer to this kind of statistics KOO (knock-one-out or kick-one-out) statistics in the technical report (Bai et al., 2018). This KOO idea can be traced back to Nishii et al. (1988), who investigated the discriminant analysis and canonical correlation analysis under fixed dimensions. In this paper, we study the KOO statistics in high-dimensional responses and predictors.The KOO method was motivated to address the issue of computational complexity in traditional AIC and BIC methods. Moreover, we find that the KOO method exhibits excellent stability, particularly in high-dimensional response settings.

There has been a lot of recent interest in variable selection problems for high-dimensional linear regression models because of the increasingly frequent and important in diverse fields of economics, finance and machine learning. For univariate (or single) response case (i.e., $p = 1$), a variety of methods have been developed. This includes the penalty-based methods such as the least angle and shrinkage selection operator (LASSO, Tibshirani (1996)), the adaptive LASSO (Zou, 2006), the smoothly clipped absolute deviation (SCAD Fan and Li (2001)), the minimax convex penalty (MCP, Zhang (2010)); the screening-based methods such as the sure independence screening (SIS, Fan and Lv (2008)), the covariate assisted screening estimates (CASE, Ke et al. (2014)); the testing based methods such as the multiple testing approach by the false discovery rate (FDR) (Liu and Luo, 2014; Xia et al., 2018) and many other related methods. We refer to some recent review papers (Shao, 1997; Fan and Lv, 2010; Huang et al., 2012; Anzanello and Fogliatto, 2014; Heinze et al., 2018; Desboulets, 2018; Lee et al., 2019; Cai et al., 2023) for more details. However, there is comparatively less literature available for multiple responses (i.e. $p > 1$). Xia (2017) proposed a row-wise multiple testing procedure when p is fixed; Kong et al. (2017) suggested a screening method via the distance correlations of the responses and each covariate for high-dimensional multi-response interaction models. For $p \to \infty$, following Bai et al. (2014), Bai et al. (2022) investigated the asymptotic properties of the classical AIC, BIC and C_p criteria; and Sakurai and Fujikoshi (2020); Oda and Yanagihara (2020) established the consistencies of the KOO methods with AIC, BIC and C_p thresholds under normality errors.

Main contributions of this paper are: (1) We obtain the asymptotic distributions of the KOO statistics $\mathcal{K}_j$ for any $j = 1, \ldots, k$ under some mild moment conditions and 3L asymptotic framework: large-response ($p \to \infty$), large-model ($k \to \infty$) and large-sample ($n \to \infty$). These theoretical results are applicable to many other model selection rules, such as growth curve model, multiple discriminant analysis, principal component analysis, canonical correlation analysis, and graphical model (e.g., Fujikoshi and Sakurai (2019); Oda et al. (2020); Fujikoshi et al. (2023)). (2) A scalable model selection method based on the KOO statistics is proposed. In practice, we use a multiplier bootstrap procedure to estimate the asymptotic thresholds. Simulation studies and real data analyses suggest the proposed model selection method performs favorably against the existing KOO methods with AIC, BIC and C_p thresholds.

The remainder of this paper is organized as follows. In Section 2, we state the main results of this paper, which include the almost sure limit and central limit theorem (CLT) of the KOO statistics. In Section 3, we propose a subset selection rule for the high-dimensional linear regression model based on the KOO statistics, and the strong consistency of this rule is presented as well. For enhancing the robustness of the selection rule, we also propose a bootstrap threshold for the KOO approach in Section 3. In Sections 4 and 5, we conduct some simulation studies and real data analysis, respectively. Proofs of the main theorems under normality are given in Section 6 since they are less technical and of independent interests. Proofs for general error distributions using random matrix theory are provided in the supplementary material for interested readers.

2. KOO statistics

2.1 Notation and preliminary

We begin this section with some basic notation and definitions. In this paper, matrices and vectors are denoted by boldface uppercase and lowercase letters, respectively. Let $\mathbf{I}_n$ denote the identity matrix of order n,

$$\widehat{\mathbf{\Sigma}}_{\mathbf{j}} = \frac{1}{n}\mathbf{Y}'\mathbf{Q}_{\mathbf{j}}\mathbf{Y}, \quad \mathbf{Q}_{\mathbf{j}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{j}}, \quad \mathbf{P}_{\mathbf{j}} = \mathbf{X}_{\mathbf{j}}(\mathbf{X}_{\mathbf{j}}'\mathbf{X}_{\mathbf{j}})^{-1}\mathbf{X}_{\mathbf{j}}', \tag{2.4}$$

$|\mathbf{j}|$ the cardinality of subset $\mathbf{j}$, and $|\widehat{\mathbf{\Sigma}}_{\mathbf{j}}|$ the determinant $\widehat{\mathbf{\Sigma}}_{\mathbf{j}}$. Note that $\mathbf{P}_{\mathbf{j}}$ is an orthogonal projection of rank $|\mathbf{j}|$ onto the subspace spanned by $\mathbf{X}_{\mathbf{j}}$, and $\mathbf{Q}_{\mathbf{j}}$ is the orthogonal projection of rank $n - |\mathbf{j}|$ onto the orthogonal complement subspace spanned by $\mathbf{X}_{\mathbf{j}}$. For brevity, we suppress the subscript $[k]$ for full model, and denote the true model subscript by $*$ and the subscript of model $[k]\backslash j$ by j (e.g., $\mathbf{Q} := \mathbf{Q}_{[k]}$, $\mathbf{Q}_{\mathbf{j}_*} := \mathbf{Q}_*$ and $\mathbf{Q}_j := \mathbf{Q}_{[k]\backslash j}$). The identity matrix, all-zero matrix, all-one vector and all-zero vector, whose orders are often clear from the context and thus will not be indicated, are denoted by $\mathbf{I}$, $\mathbf{O}$, $\mathbf{1}$, and $\mathbf{0}$, respectively. We call j (or variable $\mathbf{x}_j$) true if $j \in \mathbf{j}_*$, and j (or variable $\mathbf{x}_j$) is spurious if $j \notin \mathbf{j}_*$. We call a model $\mathbf{j}$ is over-specified if $\mathbf{j} \supset \mathbf{j}_*$ and under-specified if $\mathbf{j}^c \cap \mathbf{j}_*$ is not empty.

For a matrix $\mathbf{A}$, its spectral norm and maximum norm are denoted by $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_\infty$, respectively. The largest and smallest eigenvalues of $\mathbf{A}$

are denoted by $\lambda_{\max}^{\mathbf{A}}$ and $\lambda_{\min}^{\mathbf{A}}$, respectively. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, $\mathbf{A} \circ \mathbf{B}$ stands for the Hadamard product of $\mathbf{A}$ and $\mathbf{B}$. We denote the probability by $\mathbb{P}$, the expectation by $\mathbb{E}$, and the trace by tr. Define $c_n := p/n$ and $\alpha_n := k/n$. Throughout this paper, we use $o(1)$ (respectively, $o_p(1)$, $o_{a.s.}(1)$) to denote (respectively, in probability, almost surely) scalar negligible entries. And the notations $O(1)$, $O_p(1)$ and $O_{a.s.}(1)$ are used in a similar way.

We now introduce the KOO statistics

$$\mathcal{K}_j = \operatorname{tr}(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}_j) - p.$$

It is known that for testing $\boldsymbol{\theta}_j = \mathbf{0}$ under normality, the Lawley-Hotelling trace statistic can be expressed as $(n-k)(\mathcal{K}_j + p)$. Next we will investigate the statistical properties of $\mathcal{K}_j$ under the 3L asymptotic framework: large-model (k), large-sample (n) and large-dimensional response (p). Before presenting our main theoretical results, we briefly analyze the statistic $\mathcal{K}_j$. Let

$$\mathbf{a}_j = \mathbf{Q}_j\mathbf{x}_j / \left\|\mathbf{Q}_j\mathbf{x}_j\right\|.$$

By Sylvester's determinant theorem, we have that

$$n\widehat{\mathbf{\Sigma}}_j = n\widehat{\mathbf{\Sigma}} + \mathbf{Y}'\mathbf{a}_j\mathbf{a}_j'\mathbf{Y} \tag{2.5}$$

which implies

$$\mathcal{K}_j = n^{-1}\mathbf{a}_j'\mathbf{Y}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{Y}'\mathbf{a}_j. \tag{2.6}$$

If we plug the model (1.3) into the jth KOO statistic, we have

$$\mathcal{K}_j = (\mathbf{a}_j'\mathbf{X}_*\mathbf{\Theta}_*\mathbf{\Sigma}^{-1/2} + \mathbf{a}_j'\mathbf{E})(\mathbf{E}'\mathbf{Q}\mathbf{E})^{-1}(\mathbf{E}'\mathbf{a}_j + \mathbf{\Sigma}^{-1/2}\mathbf{\Theta}_*'\mathbf{X}_*'\mathbf{a}_j).$$

When j is spurious (i.e., $j \notin \mathbf{j}_*$), $\mathbf{a}_j$ and $\mathbf{X}_*$ are orthogonal. Thus, in this case,

$$\mathcal{K}_j = \mathbf{a}_j'\mathbf{E}(\mathbf{E}'\mathbf{Q}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_j.$$

On the other hand, when j is true (i.e., $j \in \mathbf{j}_*$), then

$$\mathcal{K}_j \asymp \mathbf{a}_j'\mathbf{E}(\mathbf{E}'\mathbf{Q}\mathbf{E})^{-1}\mathbf{E}'\mathbf{a}_j + \mathbf{x}_j'\mathbf{Q}_j\mathbf{x}_j\boldsymbol{\theta}_j'\mathbf{\Sigma}^{-1/2}(\mathbf{E}'\mathbf{Q}\mathbf{E})^{-1}\mathbf{\Sigma}^{-1/2}\boldsymbol{\theta}_j.$$

We emphasize that, for spurious j, the KOO statistics $\mathcal{K}_j$ are independent of the population covariance matrix $\mathbf{\Sigma}$. This property is highly desirable as it eliminates the involvement of unknown parameters. Furthermore, the term $\mathbf{x}_j'\mathbf{Q}_j\mathbf{x}_j\boldsymbol{\theta}_j'\mathbf{\Sigma}^{-1/2}(\mathbf{E}'\mathbf{Q}\mathbf{E})^{-1}\mathbf{\Sigma}^{-1/2}\boldsymbol{\theta}_j > 0$ becomes a key indicator to distinguish between spurious and true variables, with its value serving as a crucial factor in the determination process. The detailed discussion is stated in the next subsection.

2.2 Asymptotical properties of the KOO statistics

In this subsection, we state the asymptotics of the KOO statistics and illustrate how the KOO statistics of true variables behave differently from that of the KOO statistics of spurious variables under some mild conditions. Before stating these results, we collect the needed conditions below.

(C1) As $\min\{k, p, n\} \to \infty$, $c_n \to c \in (0, 1)$ and $\alpha_n \to \alpha \in [0, 1)$ satisfying $\alpha + c < 1$.

(C2) The true model $\mathbf{j}_* \subset [k]$, and $|\mathbf{j}_*|$ is allowed to diverge as $k \to \infty$.

(C3) The entries e_{ij} of $\mathbf{E}$ are independent and identically distributed (i.i.d.) with zero means, unit variances, and finite fourth moments, i.e., $\tau = \mathbb{E}e_{ij}^4 - 3 \in (-\infty, \infty)$.

(C4) Matrix $\mathbf{X}'\mathbf{X}$ is positive definite for all $n > k + p$.

Our main results of this paper are stated below. The proofs, under normality of errors, will be given in Section 6; and the general proofs without assuming normality of errors will be given in the Appendix.

Let

$$\delta_j := \delta_{nj} = p^{-1}\mathbf{x}_j'\mathbf{Q}_j\mathbf{x}_j\boldsymbol{\theta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_j. \tag{2.7}$$

The following theorem identifies the strong limits of the KOO statistics $\mathcal{K}_j$

for all $j \in [k]$.

Theorem 1. *Under conditions* (C1) – (C4)*, we have uniformly in* $j \in [k]$*,*

$$\mathcal{K}_j = \begin{cases} \frac{c_n}{1-c_n-\alpha_n} + o_{a.s.}(1), & \textit{if } j \notin \mathbf{j}_*, \\ (1+\delta_j)\left[\frac{c_n}{1-c_n-\alpha_n} + o_{a.s.}(1)\right], & \textit{if } j \in \mathbf{j}_*. \end{cases}$$

As $\boldsymbol{\theta}_j$ and $\boldsymbol{\Sigma}$ are typically unknown in practice, the limits of $\mathcal{K}_j$'s for $j \in \mathbf{j}_*$ are unknown. However, the fluctuations of the $\mathcal{K}_j$'s for spurious variables are pretty simple, which is described in the following theorem.

Theorem 2. *Under conditions* (C1) – (C4)*, for any fixed integer* $q > 0$ *and* $\{j_1, \ldots, j_q\} \subset [k] \backslash \mathbf{j}_*$*, the random vector*

$$\sqrt{p}\mathbf{G}_q^{-1/2}\left[(\mathcal{K}_{j_1}, \ldots, \mathcal{K}_{j_q})' - \frac{c_n}{1-c_n-\alpha_n}\mathbf{1}_q\right]$$

converges weakly to the standard q*-dimensional Gaussian random vector, where*

$$\mathbf{G}_q = \frac{c_n^2}{(1-\alpha_n-c_n)^2}\left[\frac{2(1-\alpha_n)}{(1-\alpha_n-c_n)}(\boldsymbol{\mathcal{A}}_q'\boldsymbol{\mathcal{A}}_q)^2 + \tau(\boldsymbol{\mathcal{A}}_q \circ \boldsymbol{\mathcal{A}}_q)'(\boldsymbol{\mathcal{A}}_q \circ \boldsymbol{\mathcal{A}}_q)\right],$$

and $\boldsymbol{\mathcal{A}}_q = (\mathbf{a}_{j_1}, \ldots, \mathbf{a}_{j_q})$ *is an* $n \times q$ *non-random matrix.*

Theorem 2 is of independent interest: As $\mathcal{K}_j$'s are the basic statistics for testing the hypothesis that $\boldsymbol{\theta}_j = \mathbf{0}$, this theorem can be used to obtain the CLTs of these statistics under the null hypothesis. Moreover, if $\tau = 0$

(e.g., $\{e_{ij}\}$ come from a standard normal distribution), then the second term in $\mathbf{G}_q$ vanishes; or if $\max_{j\in[k]\backslash\mathbf{j}_*}\|\mathbf{a}_j\|_\infty = o(1)$, then the second term in the covariance matrix $\mathbf{G}_q$ tends to 0 as $n \to \infty$.

When $\tau \neq 0$, we propose an estimator of τ,

$$\hat{\tau} = \left\{p^{-1}\mathrm{tr}[(\mathbf{Y}'\mathbf{Q}\mathbf{Y} - (n-k)\mathbf{I}) \circ (\mathbf{Y}'\mathbf{Q}\mathbf{Y} - (n-k)\mathbf{I})] - 2(n-k)\right\} / \mathrm{tr}(\mathbf{Q}\circ\mathbf{Q}),$$

which is shown to be unbiased and weakly consistent in Theorem 3 below.

Theorem 3. *Under the conditions* (C1) – (C4)*, $\hat{\tau}$ is an unbiased and weakly consistent estimator of τ.*

Combining Theorems 2 and 3, the rejection region of the KOO statistics for testing whether some variables are spurious can be constructed. However, in order to know the power, we also need to know the fluctuations for the statistics of the true variables. The following theorem states that under some additional assumptions, the KOO statistic of the true variable is comparable to that of the spurious variables.

Theorem 4. *In addition to the conditions* (C1) – (C4)*, for $j \in \mathbf{j}_*$, we assume that*

(C5): $\mathbb{E}e_{11}^3 = 0$.

(C6): *As* $\min\{p, n, k\} \to \infty$, $\|\mathbf{a}_j\|_\infty = o(1)$, $\mathbf{x}_j'\mathbf{Q}_j\mathbf{x}_j\|\boldsymbol{\theta}_j{}'\boldsymbol{\Sigma}^{-1/2}\|_\infty^2 = o(p)$.

(C7): *As* $\min\{p,n,k\} \to \infty$, δ_j *tends to a constant.*

Then,

$$\sqrt{p}\left(\mathcal{K}_j - \frac{c_n(1+\delta_j)}{1-c_n-\alpha_n}\right)/\sigma_{nj} \xrightarrow{D} N(0,1),$$

where $\sigma_{nj}^2 = 2c_n^2[(1-\alpha_n)(1+2\delta_j)+c_n\delta_j^2]/(1-\alpha_n-c_n)^3$.

2.3 Some remarks on the theorems

Remark 1. The condition, $c > 0$, in (C1) is due to technical reasons: our main tools are from random matrix theory (RMT) and RMT generally assumes the limit p/n exists and is positive. Note further that we make no explicit use of the unknown limits α and c in all the theorems below. Rather, we used α_n and c_n, which are always positive, in our results.

Remark 2. If the model size k is greater than the sample size n but the true model size k_* is fixed, one can first apply screening methods (such as the sure independence screening method based on the distance correlation (Li et al., 2012), and interaction pursuit via distance correlation (Kong et al., 2017)) to ensure condition (C1) holds. For further details on the screening methods, see (Fan and Lv, 2008, 2010).

Remark 3. If the entries e_{ij} of $\mathbf{E}$ are independent with common kurtosis, but not necessarily identically distributed, our results in this paper continue to hold provided an additional Lindeberg-type condition:

$$\frac{1}{\eta^4 n^2}\sum_{i,j}\mathbb{E}\left[|e_{ij}|^4 \mathbb{1}\left\{|e_{ij}| \geq \eta\sqrt{n}\right\}\right] = o(1),$$

for any $\eta > 0$. Here, $\mathbb{1}\{\cdot\}$ stands for the indicator function. The proofs are analogous but slightly more tedious, and we do not pursue this extension in this paper.

Remark 4. From Theorems 2 and 4, we can theoretically investigate the asymptotic power of whether a variable is spurious. However, for testing whether a variable is true, the asymptotic distribution of the true KOO statistic (i.e., Theorem 4) cannot be applied directly since δ_j is unknown when j is a true variable. Variable selection problem will be discussed in the next section in detail.

3. Selection criteria based on the KOO statistics

Theorem 1 highlights the crucial role of δ_j in differentiating the true variables from the spurious ones. For spurious variables, $\mathcal{K}_j$'s should be close to the point $c_n/(1-c_n-\alpha_n)$ when n, p, k are large. Since δ_j is always positive for $j \in \mathbf{j}_*$, the true variables would be separated from $c_n/(1-c_n-\alpha_n)$ and thus can be identified by the largest $\mathcal{K}_j$'s. Moreover, we can deduce a strongly consistent estimator for the true variables from this theorem. Let

$$\hat{\mathbf{j}}_\vartheta = \left\{ j \in [k] | \mathcal{K}_j > \frac{c_n(1+\vartheta)}{1-\alpha_n - c_n} \right\}, \quad \vartheta > 0.$$

Then, we have the following corollary of Theorem 1.

Corollary 1. *Assume that conditions* (C1) – (C4) *hold and* $\lim \delta_j > 0$ *for all* $j \in \mathbf{j}_*$. *Then, for any fixed value* $\vartheta \in (0, \min_{j\in\mathbf{j}_*}\{\lim \delta_j\})$,

$$\lim_{n,p\to\infty} \hat{\mathbf{j}}_\vartheta \overset{a.s.}{\to} \mathbf{j}_*.$$

Remark 5. This corollary implies the strong consistency for the KOO methods with AIC, BIC and C_p thresholds if δ_j satisfies the conditions.

In practice, however, choosing a suitable ϑ is important but very challenging because (1) the largest spurious KOO statistic may converge to its limit slowly; (2) the spurious KOO statistics are correlated; and (3) the limits of the true KOO statistics are unknown. Hence, we propose a high-dimensional multiplier bootstrap procedure to approximate the distribution of the largest spurious KOO statistic $\mathcal{K}_j$, from which a selection criterion for the linear regression model (1.1) under the 3L framework is formulated.

Denote the estimator of the true model be

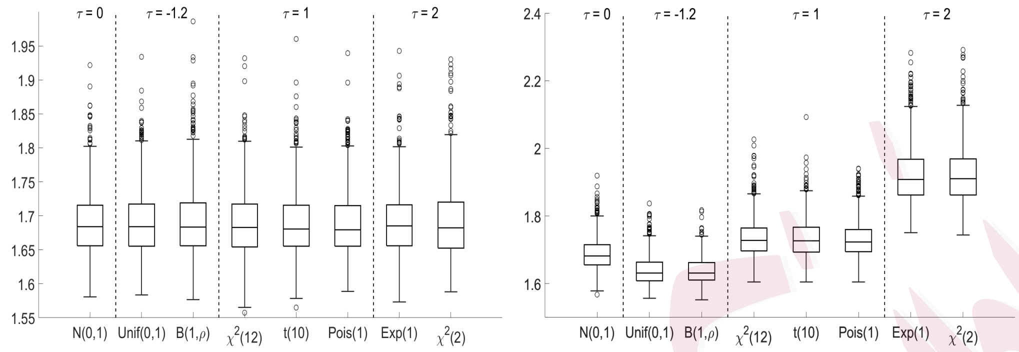
$$\hat{\mathbf{j}}_* = \{ j \in [k] : \mathcal{K}_j > K_\nu \},$$

Algorithm 1: Estimation of K_ν

```
Input: $\nu$, $\mathbf{Y}$, $\mathbf{X}$ and estimator $\hat{\tau}$ based on $\{\mathbf{Y}, \mathbf{X}\}$
Output: Estimator $\hat{K}_\nu$
1 Compute $\boldsymbol{\mathcal{A}}_k = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$.
2 Generate a random matrix $\tilde{\mathbf{E}}$ with $n \times p$ i.i.d. zero mean, unit
   variance and $\hat{\tau}$ excess kurtosis elements.
3 Compute $\mathbf{K} = \boldsymbol{\mathcal{A}}_k' \tilde{\mathbf{E}} (\tilde{\mathbf{E}}' \mathbf{Q} \tilde{\mathbf{E}})^{-1} \tilde{\mathbf{E}}' \boldsymbol{\mathcal{A}}_k$.
4 Compute the largest value of the diagonal elements of $\mathbf{K}$ and
   denote it by $\tilde{\mathcal{K}}^{(1)}$.
5 Repeat $N$ times of the above procedures 2–4, and obtain
   $\{\tilde{\mathcal{K}}^{(1)}, \ldots, \tilde{\mathcal{K}}^{(N)}\}$.
6 Compute the $100(1-\nu)$th quantile of $\{\tilde{\mathcal{K}}^{(1)}, \ldots, \tilde{\mathcal{K}}^{(N)}\}$ and denote
   it by $\hat{K}_\nu$.
```

where K_ν is the critical value with at significance level ν, which is estimated by Algorithm 1.

From Theorem 2, the critical value K_ν may depend on $\|\mathbf{a}_i\|_\infty$ or the excess kurtosis but not on the exact distribution of the errors. The boxplots of the spurious KOO statistics $\mathcal{K}_j$'s for different distributions presented in Fig. 1 support this claim. In this simulation, we set $\boldsymbol{\Theta} = \mathbf{O}$, $\boldsymbol{\Sigma} = \mathbf{I}$ and generate two predictor matrices: the first one is a 2000×600 matrix with i.i.d. entries from $U(1,5)$; and the second one is a 2000×600 diagonal matrix. As the values of the diagonal elements do not affect the result, the diagonal entries were chosen to be 1 in our simulation. We examine six different distributions of the errors: standard normal distribution $N(0,1)$, standardized uniform distribution $U(0,1)$, standardized Bernoulli distribution $B(1,\rho)$ with parameter $\rho = (6-\sqrt{6})/12$, standardized chi-square distribution with 12 degrees of freedom $\chi^2(12)$, standardized t-distribution with 10 degrees of freedom t_{10}, standardized Poisson distribution with parameter 1 $Pois(1)$, standardized exponential distribution with rate parameter 1 $Exp(1)$ and standardized chi-square distribution with 2 degrees of freedom $\chi^2(2)$. Note that $\|\mathbf{a}_i\|_\infty \to 0$ for the random predictor matrix, $\|\mathbf{a}_i\|_\infty = 1$ for the rectangular diagonal predictor matrix, the excess kurtosis of $N(0,1)$ is 0, the excess kurtoses of $Exp(1)$ and $\chi^2(2)$ are 2, the excess kurtoses of $\chi^2(12)$, t_{10} and $Pois(1)$ are 1, and the excess kurtoses of $U(0,1)$ and $B(1,(6-\sqrt{6})/12)$ are $-6/5$. Hence, in practice for convenience, we can use standardized χ^2 distribution with $12/\hat{\tau}$ degrees of freedom if $\hat{\tau} > 0$ and standardized Bernoulli distribution $B(1,\rho)$ with parameter ρ satisfying $\rho(1-\rho) = 1/(6-\hat{\tau})$ if $\hat{\tau} < 0$. Of course, if $\max_i \|\mathbf{a}_i\|_\infty \to 0$, we can use the standard normal distribution directly.



(a) Random predictor matrix

(b) Rectangular diagonal predictor matrix

Figure 1: Boxplots of the spurious KOO statistics $\{\mathcal{K}^{(j)},\ \ j = 1, \ldots, 1000\}$ with six different normalized distributions and two predictor matrices. The y-axis represents the values of $\mathcal{K}^{(j)}$'s.

4. Simulation studies

In this section, we numerically examine the properties of the proposed KOO method in a 3L framework with different settings. For comparison, we also report the results of KOO methods with AIC, BIC and C_p thresholds as proposed by Nishii et al. (1988) and implemented by Fujikoshi and Sakurai (2019); Oda et al. (2020); Nakagawa et al. (2021); Fujikoshi (2022). Specifically, the KOO methods with AIC, BIC and C_p thresholds, respectively, choose the model

$$\hat{\mathbf{j}}_*^A = \{j \in [k] : \log(1+\mathcal{K}_j) > 2c_n\},$$
$$\hat{\mathbf{j}}_*^B = \{j \in [k] : \log(1+\mathcal{K}_j) > \log(n)c_n\},$$
$$\hat{\mathbf{j}}_*^C = \{j \in [k] : (1-\alpha_n)\mathcal{K}_j > 2c_n\}.$$

For simplicity, we abbreviate the KOO method with our bootstrapping threshold to KBT, the KOO method with AIC threshold as KAIC. Similar abbreviations KBIC and KCp are used. We consider the following two settings.

Setting I: Fix $k_* = 5$, $p/n = \{0.2, 0.4\}$ and $k/n = \{0.2, 0.4\}$ with $n = 100, 500, 1000, 2000$. The results for $n = 2000$ are given in the Appendix. Set $\boldsymbol{\Sigma} = \mathbf{I}$, $\mathbf{X} = (x_{ij})_{n\times k}$, $\boldsymbol{\Theta}_{\mathbf{j}_*} = \mathbf{1}_5\boldsymbol{\theta}_*$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_{\mathbf{j}_*}, \mathbf{0})$, where $\{x_{ij}\}$ are i.i.d. generated from the continuous uniform distributions $U(1,5)$, $\mathbf{1}_5$ is a five-dimensional vector of ones and $\boldsymbol{\theta}_* = ((-0.5)^0, \ldots, (-0.5)^{p-1})$.

Setting II: Same as Setting I, except $\mathbf{X} = (\mathbf{I}_k, \mathbf{O}_{k\times(n-k)})'$ and $\boldsymbol{\Theta}_{\mathbf{j}_*} = \sqrt{n}\mathbf{1}_5\boldsymbol{\theta}_*$.

For Setting I, we consider three cases for the distribution of $\mathbf{E}$:

(i) Standard normal distribution, $e_{ij} \sim N(0,1)$;

(ii) Standardized t distribution with three degrees of freedom, i.e., $e_{ij} \sim t_3/\sqrt{5/3}$;

(iii) Standardized chi-square distribution with two degrees of freedom, i.e., $e_{ij} \sim (\chi^2(3) - 3)/\sqrt{6}$.

Since $\|\mathbf{a}_j\|_\infty \to 0$ in Setting I, we use $\tilde{\mathbf{E}}$ with the standard normal distribution to estimate $\hat{K}_\nu$. We emphasize that the excess kurtosis of distribution t_3 is infinite.

For Setting II, we consider three cases for the distribution of $\mathbf{E}$:

(iv) Standardized exponential distribution with rate parameter 1, i.e., $e_{ij} \sim Exp(1) - 1$;

(v) Standardized Poisson distribution with parameter 1, i.e., $e_{ij} \sim Pois(1) - 1$;

(vi) Standardized uniformly distribution, i.e., $e_{ij} \sim U(-\sqrt{3}, \sqrt{3})$.

Since $\|\mathbf{a}_j\|_\infty = 1$ in Setting II, we use $\tilde{\mathbf{E}}$ with standardized χ^2 distribution and standardized Bernoulli distribution, respectively, to estimate $\hat{K}_\nu$ with some suitably chosen parameter values.

In all the simulation studies, we choose two critical points in the KOO methods:.

$$\hat{\mathbf{j}}_*^{(0)} = \{j \in [k] : \mathcal{K}_j > \hat{K}_0\} \quad \text{and} \quad \hat{\mathbf{j}}_*^{(5)} = \{j \in [k] : \mathcal{K}_j > \hat{K}_{0.05}\},$$

where $\hat{K}_0$ and $\hat{K}_{0.05}$ are the largest and the 95th percentile of 1,000 bootstrap values, respectively.

We first explain our choices of the settings and the distributions. Since the KOO criteria depend on the values $\delta_j = p^{-1}\mathbf{x}_j'\mathbf{Q}_j\mathbf{x}_j\theta_j'\mathbf{\Sigma}^{-1}\theta_j$, it suffices to set $\mathbf{\Sigma} = \mathbf{I}$ and vary $\mathbf{\Theta}_*$ and $\mathbf{X}$ in conducting our simulation studies. Settings I and II both ensure δ_j are bounded above. For the case $\delta_j \to \infty$, the KOO statistics for the true variables and spurious variables are well separated, and all the compared selection methods will not show significant differences. The selection of distributions comprises five continuous distributions and one discrete distribution. The distribution described in (ii) only has finite second moment. This selection was made to investigate the implications of not satisfying the condition of finite fourth moment. To measure in greater detail the performance of these selection rules, the numbers of times, in 1000 repetitions, a selection rule under-specifies the true model, exactly identifies it and over-specifies it were tabulated. When the selection rule over-specifies the true model, we also report the average number of spurious variables selected in the last row of each sub-table. Due to space consideration, we present selected results, but typical, of Setting I (i) and Setting II (iv) in Tables 1 and 2, respectively. Full set of results, including those for $n = 2000$, can be found in the Appendix and the code can be accessed from https://github.com/huj156/KOO.git.

Based on our simulation results, the following observations are made: (1) The proposed KBT are the most robust among the compared methods, especially when the sample size n is large. (2) If the sample size n is small, we recommend choosing a bigger ν in order to avoid missing the true variables. After all, admitting a small number of spurious variables is a better tradeoff than missing some true variables. (3) Choosing a bigger ν may select more spurious variables, but unlike the KAIC and KCp, the number of spurious variables selected is still under control. (4) The simulation results are very similar across different distributions of errors, which suggests these selection rules are rather robust against the distributions of errors. (5) When $\max_i \|\mathbf{a}_i\|_\infty \to 0$, our proposed methods also work well even the finite fourth moment condition does not hold, suggesting that our theorems continue to hold even under weaker conditions. Our guess is that finite second moment of the underlying error distributions is enough. (6) The performances of KAIC, KBIC and KCp are not acceptable under our settings: KAIC and KCp frequently over-specify the true models quite substantially, and KBIC frequently under-specifies the true models. Under some special cases, KBIC has good selection times, however, KBT in general outperforms KBIC.

$\alpha = 0.2$, $c = 0.4$															
	$n = 100$					$n = 500$					$n = 1000$				
	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$
U-S	0	938	0	640	19	0	1000	0	0	0	0	1000	0	0	0
T-S	35	62	0	360	940	2	0	0	1000	953	23	0	0	998	957
O-S	965	0	1000	0	41	998	0	1000	0	47	977	0	1000	2	43
A-S	3.69	–	7.06	–	1.05	6.86	–	46.30	–	1.04	3.92	–	95.28	1	1.05

$\alpha = 0.4$, $c = 0.2$															
	$n = 100$					$n = 500$					$n = 1000$				
	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$
U-S	0	42	0	828	41	0	129	0	0	0	0	729	0	0	0
T-S	0	923	3	172	919	0	871	0	998	965	0	271	41	1000	954
O-S	1000	35	997	0	40	1000	0	1000	2	35	1000	0	959	0	46
A-S	16.50	1.09	6.67	–	1.12	100.87	–	8	1	1	213.52	–	3.28	–	1.02

Table 1: Selection times of the KOO methods with AIC, BIC, C_p thresholds and bootstrap methods under Settings (I) and (i) based on 1,000 replications. Here U-S, T-S, O-S and A-S stand for number of times a selection method under-specified the true model, number of times a selection method identified the true model exactly, number of times a selection method over-specified the true model, and the average number of spurious variables a selection method identified when it over-specified the model, respectively.

	$\alpha = 0.2$, $c = 0.4$														
	$n = 100$					$n = 500$					$n = 1000$				
	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$
U-S	0	925	0	991	622	0	1000	0	2	0	0	1000	0	0	0
T-S	2	74	0	9	361	0	0	0	997	934	0	0	0	1000	938
O-S	998	1	1000	0	17	1000	0	1000	1	66	1000	0	1000	0	62
A-S	4.74	1	7	–	1.06	16.40	–	45.88	1	1	18.74	–	95.36	–	1.03

	$\alpha = 0.4$, $c = 0.2$														
	$n = 100$					$n = 500$					$n = 1000$				
	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$	$\hat{\mathbf{j}}_*^A$	$\hat{\mathbf{j}}_*^B$	$\hat{\mathbf{j}}_*^C$	$\hat{\mathbf{j}}_*^{(0)}$	$\hat{\mathbf{j}}_*^{(5)}$
U-S	0	4	0	999	597	0	7	0	7	0	0	27	0	0	0
T-S	0	348	0	1	386	0	993	0	993	961	0	973	0	1000	939
O-S	1000	648	1000	0	17	1000	0	1000	0	39	1000	0	1000	0	61
A-S	15.31	1.67	9.23	–	1	94.64	–	28.61	–	1	198.92	–	31.12	–	1.03

Table 2: Selection times of the KOO methods with AIC, BIC, C_p thresholds and bootstrap methods under Settings (II) and (iv) based on 1,000 replications. Here U-S, T-S, O-S and A-S stand for number of times a selection method under-specified the true model, number of times a selection method identified the true model exactly, number of times a selection method over-specified the true model, and the average number of spurious variables a selection method identified when it over-specified the model, respectively.

5. Real data analysis

We apply the proposed methods to two real examples. Due to the limitation of article length, the analysis of the other real example is postponed in the supplementary material. The first example is a multivariate yeast cell-cycle dataset from Spellman et al. (1998), which can be found in the R package “spls”. This data set contains 542 cell-cycle-related genes (i.e., $n = 542$). Each gene contains 106 binding levels of transcription factors (i.e., $k = 106$) and 18 time points covering two cell cycles (i.e., $p = 18$). The binding levels of the transcription factors play a role in determining which genes are expressed and help delineate the process behind eukaryotic cell cycles. Further explanations of the dataset can be found in (Wang et al., 2007; Chun and Keleş, 2010; Chen and Huang, 2012; Kong et al., 2017). Our results are presented in Figure 2. The transcription factors {SWI5, STE12, ACE2, NDD1}, corresponding to the four largest $\mathcal{K}_j$-values, have been confirmed to be related to the cell cycle regulation by experiment Wang et al. (2007). And the other two transcription factors {RME1, HIR2} could possibly be related to the cell cycle regulation. KBIC, however, will have missed identifying the TFs {STE12, ACE2, NDD1} in the yeast cell-cycle. On the other hand, KAIC and KCp will have identified more TFs, many of which may not be related to the yeast cell-cycle.

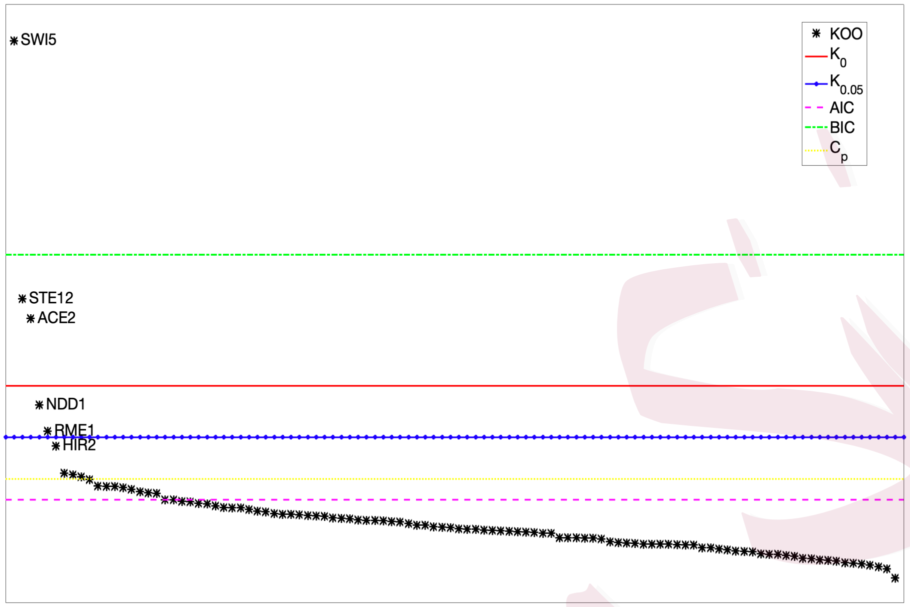


Figure 2: Scatterplots for the yeast cell-cycle dataset.

6. Proofs of Theorems 1, 2 and 4 under normality

If the errors follow the standard normal distribution, the KOO statistic can be written as the quotient of two independent chi-squared random variables. As a result, the proofs of Theorems 1, 2 and 4 are easier to present. The proofs may also be of independent interest. Hence, we prove these results under normality in this section. The proofs of these theorems for general error distributions via random matrix theory are postponed in the Appendix for interested readers. Note that there is no need to prove Theorem 3 when the errors follow the standard normal distribution.

6. PROOFS OF THEOREMS 1, 2 AND 4 UNDER NORMALITY

Recall the KOO statistic

$$\mathcal{K}_j = \mathbf{v}_j' \mathbf{W}^{-1} \mathbf{v}_j,$$

where

$$\mathbf{v}_j = \boldsymbol{\Sigma}^{-1/2} \mathbf{Y}' \mathbf{a}_j \quad \text{and} \quad \mathbf{W} = \mathbf{E}' \mathbf{Q} \mathbf{E}.$$

When $\mathbf{E}$ has the standard normal distribution, it follows that

$$\mathbf{W} \sim W_p(n-k, \mathbf{I}_p), \qquad \mathbf{v}_j \sim N_p(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Theta}_* \mathbf{X}_*' \mathbf{a}_j, \mathbf{I}_p), \quad j = 1, \ldots, k,$$

and $\mathbf{W}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ are independent. Note that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are not necessarily independent. If $j \notin \mathbf{j}_*$, $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Theta}_* \mathbf{X}_*' \mathbf{a}_j = \mathbf{0}$, on the other hand if $j \in \mathbf{j}_*$, $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Theta}_* \mathbf{X}_*' \mathbf{a}_j \neq \mathbf{0}$. Moreover, under the assumption of normality, $\tau = 0$ in assumption (C3). Next, we state a preliminary lemma.

Lemma 1. *Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_q)$ be a $p \times q$ random matrix with $q \leq p$, and let $\mathbf{W}$ be a $p \times p$ random matrix which is distributed as Wishart distribution $W_p(m, \mathbf{I}_p)$. Assume that $\mathbf{V}$ and $\mathbf{W}$ are independent. Let $\mathbf{H}$ be a $p \times p$ random orthogonal matrix such that the first q columns are $\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1/2}$, that is, $\mathbf{H} = (\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1/2}, \cdot)$. Let*

$$\mathbf{Z} = \mathbf{H}'\mathbf{W}\mathbf{H} = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22} \end{pmatrix}, \quad \mathbf{Z}_{11\cdot 2} = \mathbf{Z}_{11} - \mathbf{Z}_{12}\mathbf{Z}_{22}^{-1}\mathbf{Z}_{21}, \tag{6.8}$$

and $\mathbf{Z}_{21}$ *is a* $(p-q)\times q$ *matrix. Then,*

$$\mathbf{Z} \sim W_p(m, \mathbf{I}_p), \tag{6.9}$$

$$\mathbf{V}'\mathbf{W}^{-1}\mathbf{V} = (\mathbf{V}'\mathbf{V})^{1/2}\mathbf{Z}_{11\cdot 2}^{-1}(\mathbf{V}'\mathbf{V})^{1/2}, \tag{6.10}$$

$$\mathbf{Z}_{11\cdot 2} \sim W_q(m-(p-q), \mathbf{I}_q). \tag{6.11}$$

When $q = 1$*,*

$$\mathbf{v}_1'\mathbf{W}^{-1}\mathbf{v}_1 = \frac{\mathbf{v}_1'\mathbf{v}_1}{\mathbf{Z}_{11\cdot 2}},$$

where $\mathbf{Z}_{11\cdot 2} \sim \chi^2(m-(p-1))$ *is a chi-square variate with* $m-(p-1)$ *degrees of freedom, and* $\mathbf{v}_1'\mathbf{v}_1$ *and* $\mathbf{Z}_{11\cdot 2}$ *are independent. Further, if* $\mathbf{v}_1 \sim N_p(\boldsymbol{\mu}, \mathbf{I}_p)$*,*

$$\mathbf{v}_1'\mathbf{W}^{-1}\mathbf{v}_1 = \frac{\mathbf{v}_1'\mathbf{v}_1}{\mathbf{Z}_{11\cdot 2}} \sim \frac{\chi^2(p;\boldsymbol{\mu}'\boldsymbol{\mu})}{\chi^2(m-(p-1))}.$$

Here, $\chi^2(p;\boldsymbol{\mu}'\boldsymbol{\mu})$ *denotes a noncentral chi-square variate with* p *degrees of freedom and noncetrality parameter* $\boldsymbol{\mu}'\boldsymbol{\mu}$*, and* $\chi^2(m-(p-1))$ *denotes a chi-square variate with* $m-(p-1)$ *degrees of freedom, and they are independent.*

Proof. The result (6.9) is straightforward by considering the conditional distribution of $\mathbf{Z}$ given $\mathbf{H}$, and noting that the obtained result does not depend on $\mathbf{H}$. Next we consider the result (6.10). Noting that $\mathbf{H}$ is orthogonal, we have

$$\begin{aligned}\mathbf{V}'\mathbf{W}^{-1}\mathbf{V} &= \mathbf{V}'\mathbf{H}(\mathbf{H}'\mathbf{W}\mathbf{H})^{-1}\mathbf{H}'\mathbf{V}\\ &= \left[(\mathbf{V}'\mathbf{V})^{1/2},\ \mathbf{O}\right]'\mathbf{Z}^{-1}\left[(\mathbf{V}'\mathbf{V})^{1/2},\ \mathbf{O}\right],\end{aligned}$$

which implies (6.10). (6.11) is a well known result on Wishart distribution (e.g., Theorem 2.2.3 in (Fujikoshi et al., 2010)).

Next we consider the case of $q = 1$. Note that the $(1, 1)$ element of $\mathbf{Z}^{-1}$ is $\mathbf{Z}_{11\cdot 2}^{-1}$. Then

$$\begin{aligned}\mathbf{v}_1'\mathbf{W}^{-1}\mathbf{v}_1 =&\mathbf{v}_1\mathbf{H}'(\mathbf{H}'\mathbf{W}\mathbf{H})^{-1}\mathbf{H}'\mathbf{v}_1\\ =&((\mathbf{v}_1'\mathbf{v}_1)^{1/2}, 0, \dots, 0)\mathbf{Z}^{-1}((\mathbf{v}_1'\mathbf{v}_1)^{1/2}, 0, \dots, 0)'\\ =&\mathbf{v}_1'\mathbf{v}_1\mathbf{Z}_{11\cdot 2}^{-1}.\end{aligned}$$

The required result follows from the fact that $\mathbf{v}_1'\mathbf{v}_1 \sim \chi^2(p; \chi^2(p; \boldsymbol{\mu}'\boldsymbol{\mu}))$ and $\mathbf{Z}_{11\cdot 2} \sim \chi^2(m - (p - 1))$. Then we complete the proof of this lemma. □

6.1 Proof of Theorem 1

By Lemma 1, we can express $\mathcal{K}_j$ as a ratio of two independent chi-square variates as

$$\mathcal{K}_j = \begin{cases} \frac{\chi^2(p)}{\chi^2(\tilde{m})} & \text{if } \ j \notin \mathbf{j}_* \\ \frac{\chi^2(p;p\delta_j)}{\chi^2(\tilde{m})} & \text{if } \ j \in \mathbf{j}_* \end{cases},$$

where $\delta_j = p^{-1}\mathbf{x}_j'\mathbf{Q}_j\mathbf{x}_j\theta_j'\boldsymbol{\Sigma}^{-1}\theta_j$ and $\tilde{m} = n - k - p + 1$. For $j \notin \mathbf{j}_*$, let

$$Z_1 = \frac{\chi^2(p) - p}{\sqrt{2p}} \quad \text{and} \quad Z_2 = \frac{\chi^2(\widetilde{m}) - \widetilde{m}}{\sqrt{2\widetilde{m}}}.$$

Then, it is clear that

$$\begin{aligned}\mathcal{K}_j &= \frac{p+\sqrt{2p}Z_1}{\widetilde{m}+\sqrt{2\widetilde{m}}Z_2} = \frac{p/n+\sqrt{2p}Z_1/n}{\widetilde{m}/n+\sqrt{2\widetilde{m}}Z_2/n} \\ &= \frac{c_n}{1-c_n-\alpha_n} + o_{a.s.}(1).\end{aligned}$$

For $j \in \mathbf{j}_*$, let

$$\widetilde{Z}_1 = \frac{\chi^2(p; p\delta_j) - p(1+\delta_j)}{\sqrt{2p(1+2\delta_j)}}.$$

Note that $\widetilde{Z}_1 \overset{D}{\to} N(0,1)$ as $p \to \infty$ or $p\delta_j \to \infty$. Thus we can find that

$$\begin{aligned}\mathcal{K}_j &= \left\{p(1+\delta_j) + \sqrt{2p(1+2\delta_j)}\widetilde{Z}_1\right\}\left\{\widetilde{m}+\sqrt{2\widetilde{m}}Z_2\right\}^{-1} \\ &= \frac{p}{\widetilde{m}}\left\{1+\delta_j+\sqrt{2(1+2\delta_j)p^{-1}}\widetilde{Z}_1\right\}\left\{1-\sqrt{2\widetilde{m}^{-1}}Z_2\right\}^{-1},\end{aligned}$$

which implies

$$\mathcal{K}_j - \frac{c_n(1+\delta_j)}{1-c_n-\alpha_n} = o_{a.s.}(1+\delta_j).$$

Then we complete the proof of Theorem 1.

6.2 Proof of Theorem 2

For simplicity, we consider the case $q=2$, and assume that $\{1,2\} \subset [k]\backslash\mathbf{j}_*$. To prove Theorem 2, it is sufficient to show that for any non-null vector $\mathbf{h}=(h_1,h_2)'$, $\sqrt{p}[h_1\mathcal{K}_1 + h_2\mathcal{K}_q - \frac{c_n}{1-c_n-\alpha_n}(h_1+h_2)]$ converges weakly to a normal distribution with mean zero and variance $\frac{2c^2(1-\alpha)}{(1-\alpha-c)^3}\mathbf{h}'(\boldsymbol{\mathcal{A}}_2'\boldsymbol{\mathcal{A}}_2)^2\mathbf{h}$.

6. PROOFS OF THEOREMS 1, 2 AND 4 UNDER NORMALITY

Under the normality assumption, we can express $\mathcal{K}_1$ and $\mathcal{K}_2$ as follows:

$$\mathcal{K}_1 = \mathbf{v}_1'\mathbf{W}^{-1}\mathbf{v}_1, \quad \mathcal{K}_2 = \mathbf{v}_2'\mathbf{W}^{-1}\mathbf{v}_2. \tag{6.12}$$

Here, $\mathbf{v}_i \sim N_p(\mathbf{0}, \mathbf{I}_p), i = 1, 2$, $\mathbf{W} \sim W_p(m, \mathbf{I}_p)$, $m = n - k$, $\{\mathbf{v}_1, \mathbf{v}_2\}$ and $\mathbf{W}$ are independent, but $\mathbf{v}_1$ and $\mathbf{v}_2$ are not independent. Let $\mathcal{K}_0 = \mathbf{v}_1'\mathbf{W}^{-1}\mathbf{v}_2$ and

$$\boldsymbol{\mathcal{K}} = \begin{pmatrix} \mathcal{K}_1 & \mathcal{K}_0 \\ \mathcal{K}_0' & \mathcal{K}_2 \end{pmatrix}.$$

Note that

$$h_1\mathcal{K}_1 + h_2\mathcal{K}_2 = \mathrm{tr} D_h \boldsymbol{\mathcal{K}},$$

where $D_h = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}$.

Let $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2)$. Using Lemma 1, we can write $\boldsymbol{\mathcal{K}}$ as

$$\begin{aligned} \boldsymbol{\mathcal{K}} &= \mathbf{V}'\mathbf{W}^{-1}\mathbf{V} \\ &= \left(\frac{1}{p}\mathbf{V}'\mathbf{V}\right)^{1/2} \left(\frac{1}{\widetilde{m}}\mathbf{Z}_{11\cdot 2}\right)^{-1} \left(\frac{1}{p}\mathbf{V}'\mathbf{V}\right)^{1/2} \frac{p}{\widetilde{m}}, \end{aligned} \tag{6.13}$$

where $\mathbf{Z}_{11\cdot 2} \sim W_2(\tilde{m}, \mathbf{I}_p)$ is defined in (6.8) and $\widetilde{m} = m - (p-2)$. Note that

$$\mathbf{V}'\mathbf{V} \sim W_2(p, \boldsymbol{\Lambda}), \quad \boldsymbol{\Lambda} = \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix},$$

where $\lambda = \mathbf{a}_1'\mathbf{a}_2$. Let

$$\mathbf{F} = \begin{pmatrix} f_1 & f_3 \\ f_3 & f_2 \end{pmatrix} = \sqrt{p}\left(\frac{1}{p}\mathbf{V}'\mathbf{V} - \mathbf{\Lambda}\right), \tag{6.14}$$

$$\mathbf{G} = \begin{pmatrix} g_1 & g_3 \\ g_3 & g_2 \end{pmatrix} = \sqrt{\widetilde{m}}\left(\frac{1}{\widetilde{m}}\mathbf{Z}_{11\cdot 2} - \mathbf{I}_2\right). \tag{6.15}$$

It follows from the asymptotic distribution of a Wishart matrix (e.g., Theorem 2.5.1 in (Fujikoshi et al., 2010)) that the limiting distribution of $(f_1, f_2, f_3)'$ (respectively, $(g_1, g_2, g_3)'$) is a 3-variate normal distribution with mean zero and covariance matrix

$$\begin{pmatrix} 2 & 2\lambda^2 & 2\lambda \\ 2\lambda^2 & 2 & 2\lambda \\ 2\lambda & 2\lambda & 1+\lambda^2 \end{pmatrix}, \left(\text{respectively, } \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right).$$

Consequently, it is straightforward to show that

$$\mathrm{tr} D_h \mathbf{F} \overset{D}{\to} N(0, 2\mathrm{tr}(D_h\mathbf{\Lambda})^2) \tag{6.16}$$

and

$$\mathrm{tr}\mathbf{\Lambda}^{1/2} D_h \mathbf{\Lambda}^{1/2}\mathbf{G} \overset{D}{\to} N(0, 2\mathrm{tr}(D_h\mathbf{\Lambda})^2). \tag{6.17}$$

Then, by substituting

$$\frac{1}{p}\mathbf{V}'\mathbf{V} = \Lambda + \frac{1}{\sqrt{p}}\mathbf{F}$$

and

$$\left(\frac{1}{\widetilde{m}}\mathbf{Z}_{11\cdot 2}\right)^{-1} = \left(\mathbf{I}_2 + \frac{1}{\sqrt{\widetilde{m}}}\mathbf{G}\right)^{-1} = \mathbf{I}_2 - \frac{1}{\sqrt{\widetilde{m}}}\mathbf{G} + \frac{1}{\widetilde{m}}\left(\mathbf{I}_2 + \frac{1}{\sqrt{\widetilde{m}}}\mathbf{G}\right)^{-1}\mathbf{G}^2$$

into (6.13), we can expand $\boldsymbol{\mathcal{K}}$ as

$$\boldsymbol{\mathcal{K}} = \left\{\Lambda + \frac{1}{\sqrt{p}}\mathbf{F} - \frac{1}{\sqrt{\widetilde{m}}}(\Lambda + \frac{1}{\sqrt{p}}\mathbf{F})^{1/2}\mathbf{G}(\Lambda + \frac{1}{\sqrt{p}}\mathbf{F})^{1/2} + \emptyset\right\}\frac{c_n}{1 - c_n - \alpha_n}, \tag{6.18}$$

where $\emptyset$ denotes the terms of order $O_p(n^{-1})$. Using (6.18), we have

$$\begin{aligned}&\sqrt{p}\left\{h_1\mathcal{K}_1 + h_2\mathcal{K}_2 - \frac{c_n}{1 - c_n - \alpha_n}(h_1 + h_2)\right\}\\ &= \frac{c_n}{1 - c_n - \alpha_n}\left\{\mathrm{tr}D_h\mathbf{F} - \left(\frac{c_n}{1 - c_n - \alpha_n}\right)^{1/2}\mathrm{tr}\boldsymbol{\Lambda}^{1/2}D_h\boldsymbol{\Lambda}^{1/2}\mathbf{G}\right\} + O_p(n^{-1/2}).\end{aligned} \tag{6.19}$$

By (6.16) and (6.17), we can see that the limiting distribution of (6.19) is normal with mean zero and variance

$$\frac{2c^2}{(1 - c - \alpha)^2}\left(1 + \frac{c}{1 - c - \alpha}\right)\mathrm{tr}(D_h\boldsymbol{\Lambda})^2 = \frac{2c^2(1 - \alpha)}{(1 - \alpha - c)^3}\mathbf{h}'(\boldsymbol{\mathcal{A}}_2'\boldsymbol{\mathcal{A}}_2)^2\mathbf{h}.$$

This completes the proof of Theorem 2 .

6.3 Proof of Theorem 4

In the proof of Theorem 1, recall that for $j \in \mathbf{j}_*$, $\mathcal{K}_j$ can be expressed as a ratio of two independent chi-square variates:

$$\mathcal{K}_j = \frac{\chi^2(p; p\delta_j)}{\chi^2(\tilde{m})},$$

where $\chi^2(p;p\delta_j)$ denotes a noncentral chi-square variate with p degrees of freedom and noncentrality parameter $p\delta_j$, and $\chi^2(\tilde{m})$ denotes a chi-square variate with $\tilde{m}=n-k-p+1$ degrees of freedom, and they are independent. Let

$$\widetilde{Z}_1=\frac{\chi^2(p;p\delta_j)-p-p\delta_j}{\sqrt{2(p+2p\delta_j)}},\quad \widetilde{Z}_2=\frac{\chi^2(\widetilde{m})-\widetilde{m}}{\sqrt{2\widetilde{m}}}.$$

Then, it is checked that $\widetilde{Z}_1$ and $\widetilde{Z}_2$ converge to the standard normal distribution. Note that

$$\begin{aligned}\mathcal{K}_j&=\left\{(p+p\delta_j)+\sqrt{2(p+2p\delta_j)}\widetilde{Z}_1\right\}\left\{\widetilde{m}+\sqrt{2\widetilde{m}}\widetilde{Z}_2\right\}^{-1}\\&=\frac{p}{\widetilde{m}}\left\{1+\delta_j+\sqrt{2p^{-1}(1+2\delta_j)}\widetilde{Z}_1\right\}\left\{1+\sqrt{2\widetilde{m}^{-1}}\widetilde{Z}_2\right\}^{-1}.\end{aligned}$$

This implies that

$$\begin{aligned}&\sqrt{p}\left\{\mathcal{K}_j-\frac{p}{\widetilde{m}}(1+\delta_j)\right\}\\&\quad=\frac{p}{\widetilde{m}}\left\{\sqrt{2(1+2\delta_j)}\widetilde{Z}_1-(1+\delta_j)\left(\frac{2p}{\widetilde{m}}\right)^{1/2}\widetilde{Z}_2\right\}+O_p(n^{-1/2}).\end{aligned}$$

Theorem 4 follows from noting that $\widetilde{Z}_1$ and $\widetilde{Z}_2$ independently converge to the standard normal distribution.

Supplementary Materials

Supplementary material includes additional simulation studies, additional real data analysis and proofs of the main theorems for general error distributions using random matrix theory.

Acknowledgements

The authors thank the AE and the reviewers for their careful reading of the manuscript and their helpful comments that improve this paper. The authors are also grateful to Professor Xuming He for his helpful suggestions and help in polishing the English. Zhidong Bai was partially supported by National Natural Science Foundation of China (Grant Nos. 12171198, 12271536), and Team Project of Jilin Provincial Department of Science and Technology No. 20210101147JC. Jiang Hu was partially supported by National Natural Science Foundation of China (Grant Nos. 12292980, 12292982, 12171078), National Key R & D Program of China No. 2020YFA0714102 and Fundamental Research Funds for the Central Universities No. 2412023YQ003.

References

Anzanello, M. J. and F. S. Fogliatto (2014). A review of recent variable selection methods in industrial and chemometrics applications. *European J. of Industrial Engineering 8*(5), 619.

Bai, Z., K. P. Choi, Y. Fujikoshi, and J. Hu (2022). Asymptotics of AIC, BIC and Cp model selection rules in high-dimensional regression. *Bernoulli 28*(4), 2375–2403.

Bai, Z., Z. Fang, and Y.-C. Liang (2014). *Spectral Theory of Large Dimensional Random Matrices and Its Applications to Wireless Communications and Finance Statistics: Random Matrix Theory and Its Applications*. Singapore: World Scientific Publishing Co. Pte Ltd.

Bai, Z., Y. Fujikoshi, and J. Hu (2018). Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression. https://arxiv.org/abs/1810.12609v3.

Cai, T. T., Z. Guo, and Y. Xia (2023). Statistical inference and large-scale multiple testing for high-dimensional regression models. *TEST 32*(4), 1135–1171.

Chen, L. and J. Z. Huang (2012). Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection. *Journal of the American Statistical Association 107*(500), 1533–1545.

Chun, H. and S. Keleş (2010). Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 72*(1), 3–25.

Desboulets, L. D. D. (2018). A Review on Variable Selection in Regression Analysis. *Econometrics 6*(4), 45.

Fan, J. and R. Li (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Associ-*

ation 96(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*, 101–148.

Fujikoshi, Y. (2022). High-dimensional consistencies of KOO methods in multivariate regression model and discriminant analysis. *Journal of Multivariate Analysis 188*, 104860.

Fujikoshi, Y. and T. Sakurai (2019). Consistency of test-based method for selection of variables in high-dimensional two-group discriminant analysis. *Japanese Journal of Statistics and Data Science 2*(1), 155–171.

Fujikoshi, Y., T. Sakurai, and T. Yamada (2023). High-dimensional consistencies of KOO methods for selecting graphical models. pp. Hiroshima Statistical Research Group, TR–22–6.

Fujikoshi, Y., V. V. Ulyanov, and R. Shimizu (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* Wiley Series in Probability and Statistics. Wiley: Hoboken, N.J.

Heinze, G., C. Wallisch, and D. Dunkler (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal 60*(3), 431–449.

Huang, J., P. Breheny, and S. Ma (2012). A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science 27*(4), 481–499.

Ke, Z. T., J. Jin, and J. Fan (2014). Covariate Assisted Screening and Estimation. *The Annals of Statistics 42*(6), 2202–2242.

Kong, Y., D. Li, Y. Fan, and J. Lv (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics 45*(2), 897–922.

Lee, E. R., J. Cho, and K. Yu (2019). A systematic review on model selection in high-dimensional regression. *Journal of the Korean Statistical Society 48*(1), 1–12.

Li, R., W. Zhong, and L. Zhu (2012). Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Liu, W. and S. Luo (2014). Hypothesis Testing for High-dimensional Regression Models. pp. Technical report.

Nakagawa, T., H. Watanabe, and M. Hyodo (2021). Kick-one-out-based variable selection method for Euclidean distance-based classifier in high-dimensional settings. *Journal of Multivariate Analysis 184*, 104756.

Nishii, R., Z. Bai, and P. R. Krishnaiah (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Mathematical Journal 18*(3), 451–462.

Oda, R., Y. Suzuki, H. Yanagihara, and Y. Fujikoshi (2020). A consistent variable selection method in high-dimensional canonical discriminant analysis. *Journal of Multivariate Analysis 175*, 104561.

Oda, R. and H. Yanagihara (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electronic Journal of Statistics 14*(1), 1386–1412.

Sakurai, T. and Y. Fujikoshi (2020). Exploring Consistencies of Information Criterion and Test-Based Criterion for High-Dimensional Multivariate Regression Models Under Three Covariance Structures. In T. Holgersson and M. Singull (Eds.), *Recent Developments in Multivariate and Random Matrix Analysis: Festschrift in Honour of Dietrich von Rosen*, pp. 313–334. Cham: Springer International Publishing.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica 7*(2), 221–242.

Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher (1998). Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell 9*(12), 3273–3297.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wang, L., G. Chen, and H. Li (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics 23*(12), 1486–1494.

Xia, Y. (2017). Testing and support recovery of multiple high-dimensional covariance matrices with false discovery rate control. *TEST 26*(4), 782–801.

Xia, Y., T. Cai, and T. T. Cai (2018). Two-Sample Tests for High-Dimensional Linear Regression with an Application to Detecting Interactions. *Statistica Sinica 28*(1), 63–92.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association 101*(476), 1418–1429.