A Modified KOO Method for Selection of Variables in Large-dimensional Regression and Estimation of its Selection Probability

Tetsuro Sakurai*, Takayuki Yamada** and Yasunori Fujikoshi***

**School of General and Management Studies, Suwa University of Science,*
5000-1 Toyohira, Chino, Nagano 391-0292, Japan,

***Faculty of Data Science, Kyoto Women's University,*
35 Kitahiyoshi-cho, Imakumano, Higashiyama-ku, Kyoto 605-8501, Japan,

****Department of Mathematics, Graduate School of Science,*
Hiroshima University,
1-3-1 Kagamiyama, Higashi Hiroshima, Hiroshima 739-8626, Japan

Abstract

This paper is concerned with the selection of variables in large-dimensional regression using Knock-one-out (KOO) method. One of studies of KOO methods, Bai et al. (2025) treats the method based on the Lawley-Hotelling statistic. Their theoretical result guarantees the strong consistency, but the threshold in the method involves unknown quantity. Instead of it, they suggested to use the threshold calculated by bootstrap. In this paper we modify Bai et al. (2025)'s KOO method by using a threshold which does not contain unknown quantity. We show that the limiting probability of selecting the true model is equal to 1 when the dimension and the sample are large. Furthermore, a method to estimate its selection probability is provided in this paper. Tendencies of our method are explored numerically through a Monte Carlo simulation.

AMS 2000 subject classification: primary 62H12; secondary 62H30
Key Words and Phrases: Consistency, Large-dimensional regressio, Modified KOO methods, Selection probability

1. Introduction

This paper is concerned with the variable selection problems in multivariate regression model with large-dimensional regression under non-normality. In general, model selection approaches based on model selection criteria such as AIC, BIC and Cp involves computational problems as the number of regression variables increases. As an approach to overcome this problem, we consider KOO approach due to Nishii et al. (1988) and Zhao et al. (1986). Suppose that there are n observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ on p response variables $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and n observations $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n$ on k explanatory variables $\boldsymbol{x} = (x_1, \ldots, x_k)'$. Here, $\boldsymbol{y}_i$ and $\widetilde{\boldsymbol{x}}_i$ are observations of $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively, for the ith subject. Let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and

$$\mathbf{X} = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) = (\boldsymbol{x}_j, j \in \boldsymbol{\omega}).$$

The multivariate regression model is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \boldsymbol{\mathcal{E}}, \tag{1.1}$$

where $\boldsymbol{\Theta}$ is a $k \times p$ regression coefficient matrix, and $\boldsymbol{\mathcal{E}} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)'$ is the error matrix. It is assumed that the $\boldsymbol{\epsilon}_i$'s are independently and identically distributed as a p-variate distribution with a mean $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}$. In order to explore a simpler linear structure, we consider to select the explanatory variables. In general, the selection of $\boldsymbol{x}_i$'s may be decided by whether the ith row $\boldsymbol{\theta}_i$ of $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$ is the null vector or not. We consider such a selection problem, assuming that $\boldsymbol{\Sigma}$ is unknown positive definite.

For notational simplicity, let us identify the set $\{x_1, \ldots, x_k\}$ with $\boldsymbol{\omega} = \{1, \ldots, k\}$. Let $\boldsymbol{j}$ be a subset of $\boldsymbol{\omega}$, and $\mathbf{X}_{\boldsymbol{j}} = (\boldsymbol{x}_j, j \in \boldsymbol{j})$. Let $k_{\boldsymbol{j}}$ be the cardinality of $\boldsymbol{j}$. Denote the model based on $\mathbf{X}_{\boldsymbol{j}} = (\boldsymbol{x}_j, j \in \boldsymbol{j})$ by

$$M_{\boldsymbol{j}} : \quad \mathbf{Y} = \mathbf{X}_{\boldsymbol{j}}\boldsymbol{\Theta}_{\boldsymbol{j}} + \boldsymbol{\mathcal{E}}. \tag{1.2}$$

Bai et al. (2025) introduced variable selection methods based on the following KOO statistics:

$$\mathcal{K}_j = \mathrm{tr}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\backslash j}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}\right) - p, \quad j = 1, \ldots, k, \tag{1.3}$$

where for any $\boldsymbol{j}$,

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

and $\boldsymbol{\omega}\backslash j$ is the set obtained by removing an element j from the set $\boldsymbol{\omega}$. Note that $\mathcal{K}_j$ is the Lawley-Hotelling statistic (Fujikoshi et al. (2010)) for testing $\boldsymbol{\theta}_j = \mathbf{0}$. They proposed a consistent method, but the method involves unknown quantities. They also proposed a method of approximating the KOO statistic by Bootstrap approximation.

In this paper, we improve the method proposed by Bai et al. (2025) by avoiding the use of the bootstrap approach. Under the same assumptions as Bai et al. (2025), we also show that the limiting probability of selecting the true model is equal to 1. Furthermore, we also provide a method for estimating the selection probability of explanatory variables in our proposed method.

The remainder of this paper is organized as follows: In Section 2, we propose a modified KOO method of Bai et al. (2025) which is illustrated in diagrams. In Section 3, we show that our KOO has a consistency property. In Section 4, our method is explored numerically through a Monte Carlo simulation. In Section 5, we describe an estimation method for the selection probability of explanatory variables using our proposed KOO method. In Section 6, the accuracy of the proposed estimation method is evaluated through a Monte Carlo simulation. Finally, Section 7 concludes the study with a summary and future research directions.

2. A modified KOO method

Using the KOO statistic $\mathcal{K}_j$ in (1.3), Bai et al. (2025) proposed the following KOO method for selection of explanatory variables:

$$\widehat{\boldsymbol{j}}_{Z1} = \left\{ j \in \boldsymbol{\omega} \mid \mathcal{K}_j > \frac{c_n(1+\vartheta)}{1-\alpha_n-c_n} \right\}. \tag{2.1}$$

Here ϑ is a constant satisfying $\vartheta \in (0, \min_{j\in\boldsymbol{j}_*}\{\lim \eta_j\})$, and $c_n = p/n$, $\alpha_n = k/n$. Further,

$$\eta_j = p^{-1}\xi_j^2, \quad \xi_j^2 = \boldsymbol{x}_j'\mathbf{Q}_j\boldsymbol{x}_j\boldsymbol{\theta}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_j. \tag{2.2}$$

The selection method (2.1) means that "$j \in \widehat{\boldsymbol{j}}_{Z1}$" $\Leftrightarrow$ "we select x_j". Note that $\widehat{\boldsymbol{j}}_{Z1}$ is the estimator of the true model, and it is known (Bai et al. (2025)) that the KOO method (2.1) is consistent under the following conditions (C1) $\sim$ (C4):

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

(C2) : The true model M_* is included in the full model M, $M_* \subset M$, and the cordinality $|M_*|$ is allowed to diverge as $k \to \infty$.

(C3) : The entries ϵ_{ij} of $\boldsymbol{\mathcal{E}}$ are independent and identically distributed with zero means, unite variances, and finite fourth moments, i.e., $\tau = \mathrm{E}(\epsilon_{ij}^4) - 3 \in (-\infty, \infty)$.

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

However, the KOO method (2.1) involves unknown quantities, and so the authors suggest to approximate the distribution of $\mathcal{K}_j$ by a Bootstrap method, and propose the following selection method:

$$\widehat{\boldsymbol{j}}_{Z2} = \{ j \in \boldsymbol{\omega} \mid \mathcal{K}_j > \mathcal{K}_\mu \}, \tag{2.3}$$

where $\mathcal{K}_\mu$ is the critical value with at significance level μ, which is estimated by an Algorithm (for the details, see Bai et al (2025)).

In this paper we suggest to use the following modified KOO method:

$$\widehat{\boldsymbol{j}}_F = \left\{ j \in \boldsymbol{\omega} \mid \mathcal{K}_j - \frac{c_n}{1-\alpha_n-c_n} > a\sigma_n \right\}, \tag{2.4}$$

where a is a constant sasifying O(1), and

$$\sigma_n^2 = \frac{2(1-\alpha_n)c_n^2}{(1-\alpha_n-c_n)^3}. \tag{2.5}$$

The quantity σ_n^2 is obtained from σ_{nj}^2 by putting $\eta_j = 0$, where

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

Further, it is known (Bai et al. (2025)) that under (C1) $\sim$ (C7)

$$\sqrt{p}\left(\mathcal{K}_j - \frac{c_n(1+\eta_j)}{1-c_n-\alpha_n}\right)/\sigma_{nj} \to \mathrm{N}(0,1).$$

Here (C5) $\sim$ (C7) are given as follows:

(C5) : $\mathrm{E}(e_{ij}^3) = 0$.

(C6) : As $\min\{p, n, k\} \to \infty$,

$$\| \boldsymbol{a}_j \|_\infty = \mathrm{o}(1), \quad \boldsymbol{x}_j'\mathbf{Q}_j\boldsymbol{x}_j \| \boldsymbol{\theta}_j'\boldsymbol{\Sigma}^{-1/2} \|_\infty^2 = \mathrm{o}(p),$$

where for $\boldsymbol{b} = (b_1, \ldots, b_p)'$, $\| \boldsymbol{b} \|_\infty = \max_{j=1,\ldots,p} |b_j|$.

(C7) : As $\min\{p, n, k\} \to \infty$, η_j tends to a constsnt.

Our method $\widehat{\boldsymbol{j}}_F$ will be more executable than $\widehat{\boldsymbol{j}}_{Z1}$ and $\widehat{\boldsymbol{j}}_{Z2}$. In the next section we note that $\widehat{\boldsymbol{j}}_F$ has a consistency property.

We explain our method, by using the simulation setting as in Bai et al. (2025). Simulation replications are 10^3. The errors are assumed that $e_{ij} \sim \mathrm{N}(0,1)$ and independent, Further, $\boldsymbol{\Sigma} = \mathbf{I}_p$, $n = 200$, $p = 40$, $k = 40$, $k_* = 5$. The components of $\mathbf{X}$ were generated from $U(1,5)$. The true parameter matrices are setted from: $\mathbf{1}_5\boldsymbol{\theta}_*'$, where $\boldsymbol{\theta}_*' = ((0.5)^0, \ldots, (0.5)^{p-1})$. Under these assumptions we consider the distribution of $\mathcal{K}_j$; $j = 1, 2, \ldots, k$.

Figure 1 presents a box plot of the behavior of $\mathcal{K}_j$ on the vertical axis, while the horizontal axis represents the variable indices. The true model is given by $\boldsymbol{j}_* = \{1,2,3,4,5\}$. From this figure, it can be observed that $\mathcal{K}_j$ tends to deviate from zero when j is included in the true variables, whereas $\mathcal{K}_j$ stays close to zero when j is not included in the true variables. Then, the red and blue lines in the figure are given as follows:

$$\begin{aligned} \text{Red line:} j \notin \boldsymbol{j}_* \quad & \frac{c_n}{1-\alpha_n-c_n}, \\ \text{Blue line:} j \in \boldsymbol{j}_* \quad & \frac{c_n}{1-\alpha_n-c_n}(1+\vartheta), \quad \vartheta \in (0, \min_{j\in\boldsymbol{j}_*}\{\eta_j\}). \end{aligned}$$

Our method draws a green line between the red and blue lines. Here, for example, the green line in Figure 1 corresponds to the case of $a=1$. Therefore, the green line is mostly over the variation of $\mathcal{K}_j$ for $j \notin \boldsymbol{j}_*$ and mostly under variation of $j \in \boldsymbol{j}_*$.

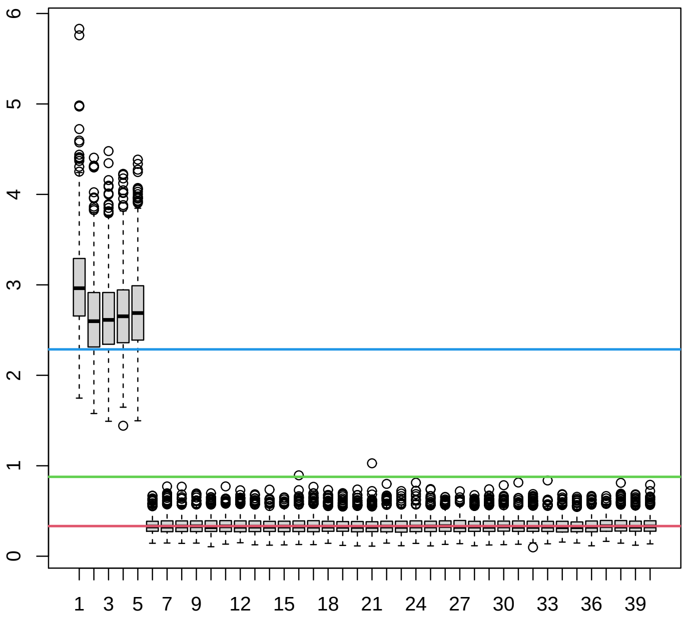

Figure 1. The values of $\mathcal{K}_j$.

3. Strong consistency

Related to the modified KOO method (2.4), let

$$\widetilde{T}_j = \mathcal{K}_j - \frac{c_n}{1-\alpha_n-c_n} - a\sigma_n. \tag{3.1}$$

Then our modified KOO method is expressed as

$$\widehat{\boldsymbol{j}}_{\mathrm{F}} = \{j \in \boldsymbol{\omega} \mid \widetilde{T}_j > 0\}. \tag{3.2}$$

It is known (Bai et al. (2014)) that the KOO method $\widehat{\boldsymbol{j}}_{Z1}$ has a consistency under conditions (C1) $\sim$ (C4). This was deduced from the strong limits of the KOO statistics $\mathcal{K}_j$ which is as follows: uniformly in $j \in \boldsymbol{\omega}$,

$$\mathcal{K}_j = \begin{cases} \dfrac{c_n}{1-\alpha_n-c_n} + o_{a.s.}(1), & \text{if } j \notin \boldsymbol{j}_*, \\ (1+\eta_j)\left(\dfrac{c_n}{1-\alpha_n-c_n} + o_{a.s.}(1)\right), & \text{if } j \in \boldsymbol{j}_*. \end{cases}$$

It holds that

$$\widetilde{T}_j > 0 \iff \mathcal{K}_j > \frac{c_n}{1-\alpha_n-c_n}\left(1 + a\sqrt{\frac{2(1-\alpha_n)}{(1-\alpha_n-c_n)}}\right).$$

Then we find that $\widehat{\boldsymbol{j}}_{\mathrm{F}} \subset \widehat{\boldsymbol{j}}_{Z1}$ if we set $\tau \in (0, \min_{j\in\boldsymbol{j}_*}\{\lim \eta_j\})$ as

$$\tau = a\sqrt{\frac{2(1-\alpha_n)}{(1-\alpha_n-c_n)}}.$$

This is summarized as the following theorem.

Theorem 3.1. *Assume that conditions* (C1) $\sim$ (C4) *hold and* $\lim \eta_j > 0$ *for all* $j \in \boldsymbol{j}_*$. *Then,* $\lim_{n,p\to\infty} \widehat{\boldsymbol{j}}_F \overset{a.s.}{\to} \boldsymbol{j}_*$ *if*

$$a\sqrt{\frac{2(1-\alpha_n)}{(1-\alpha_n-c_n)}} \in \left(0, \min_{j\in\boldsymbol{j}_*}\{\lim \eta_j\}\right).$$

4. Numerical experiments: consistency

In this section we give simulation results on the estimation method $\widehat{\boldsymbol{j}}_F$ in (2.4). We considered the two cases $a = 1, 1/\sqrt{2}$. Our simulation setting is based on Figure 1 in Bai et al. (2025). However we did not attempt the cases of $n = 1000, 2000$.

- The number of trials:10^3.
- $n = 100, 500$, $c_n = p/n = 0.2, 0.4$, $\alpha_n = k/n = 0.2, 0.4$, $k_* = 5$.
- The components of $\mathbf{X}$ were constructed from a sample of $\mathrm{U}(1, 5)$.
- The regression coefficient matrix is given as follows: $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_* \quad \mathbf{O})$, $\boldsymbol{\Theta}_* = \mathbf{1}_5 \boldsymbol{\theta}_*$. Here $\mathbf{1}_5$ is a five-dimensional vector of ones and $\boldsymbol{\theta}_* = ((0.5)^0, \ldots, (0.5)^{p-1})$.
- As the distribution of e_{ij}, we considered; (i) Standard normal distribution, (ii) Uniform distribution; $U(0, 1)$, (iii) Biomial distribution; $Bin(1, \rho)$, $\rho = (6 - \sqrt{6})/12$, (iv) Chi-square distribution χ^2 with 12 degrees of freedom, (v) t-distribution with 10 degrees of freedom, (vi) Poisson distribution with parameter 1; $Pos(1)$, (vii) Exponential distribution with parameter 1; $Exp(1)$, (viii) Chi-square distribution χ^2 with 2 degrees of freesdom. Here, all the distributions are normalized as the means 0 and variances 1.
- $a = 1, 1/\sqrt{2}$.

The selection rates associated with our method are given in Tables 1 to 2.

From Tables 1 and 2, we can identify the following tendncies.

(1) The estimator is consistent since the probabilities of selecting the true models are near 1, except for the case $a = 1, n = 500, p = 200, k = 200$.

(2) The probabilities of selecting the true models increase as n increases, for all distributions, and for any given p and k. However, these probabilities decrease as $p + k$ increases.

Table 1: Selection rates of the true models under (i) $\sim$ (iv).

					$a = 1$			$a = 1/\sqrt{2}$	
Dist.	n	p	k	Under	True	Over	Under	True	Over
Normal	100	20	20	0.00	0.97	0.03	0.00	0.87	0.13
Normal	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
Normal	100	40	20	0.16	0.83	0.00	0.03	0.94	0.04
Normal	500	200	100	0.00	1.00	0.00	0.00	1.00	0.00
Normal	100	20	40	0.01	0.91	0.08	0.00	0.66	0.34
Normal	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
Normal	100	40	40	0.89	0.10	0.01	0.67	0.25	0.08
Normal	500	200	200	0.67	0.33	0.00	0.08	0.92	0.00
Uniform	100	20	20	0.00	0.99	0.02	0.00	0.89	0.11
Uniform	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
Uniform	100	40	20	0.17	0.84	0.00	0.02	0.94	0.04
Uniform	500	200	100	0.00	1.00	0.00	0.00	1.00	0.00
Uniform	100	20	40	0.00	0.91	0.09	0.00	0.69	0.31
Uniform	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
Uniform	100	40	40	0.81	0.18	0.01	0.50	0.41	0.09
Uniform	500	200	200	0.87	0.14	0.00	0.22	0.78	0.00
Binomial	100	20	20	0.00	0.98	0.03	0.00	0.86	0.14
Binomial	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
Binomial	100	40	20	0.24	0.76	0.00	0.06	0.91	0.04
Binomial	500	200	100	0.00	1.00	0.00	0.00	1.00	0.00
Binomial	100	20	40	0.00	0.91	0.09	0.00	0.67	0.33
Binomial	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
Binomial	100	40	40	0.77	0.22	0.01	0.47	0.45	0.08
Binomial	500	200	200	0.80	0.20	0.00	0.16	0.84	0.00
χ^2_{12}	100	20	20	0.00	0.97	0.03	0.00	0.85	0.15
χ^2_{12}	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
χ^2_{12}	100	40	20	0.23	0.77	0.00	0.05	0.92	0.03
χ^2_{12}	500	200	100	0.00	1.00	0.00	0.00	1.00	0.00
χ^2_{12}	100	20	40	0.11	0.82	0.08	0.03	0.64	0.33
χ^2_{12}	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
χ^2_{12}	100	40	40	0.73	0.25	0.02	0.43	0.46	0.11
χ^2_{12}	500	200	200	0.79	0.21	0.00	0.17	0.84	0.00

Table 2: Selection rates of the true models under (v) $\sim$ (viii).

				$a = 1$			$a = 1/\sqrt{2}$		
Dist.	n	p	k	Under	True	Over	Under	True	Over
t_{10}	100	20	20	0.00	0.97	0.03	0.00	0.87	0.13
t_{10}	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
t_{10}	100	40	20	0.15	0.84	0.01	0.02	0.96	0.02
t_{10}	500	200	100	0.00	1.00	0.00	0.00	1.00	0.00
t_{10}	100	20	40	0.01	0.92	0.08	0.00	0.71	0.29
t_{10}	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
t_{10}	100	40	40	0.86	0.14	0.01	0.61	0.32	0.07
t_{10}	500	200	200	0.73	0.27	0.00	0.14	0.86	0.00
Poisson	100	20	20	0.00	0.97	0.03	0.00	0.86	0.14
Poisson	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
Poisson	100	40	20	0.17	0.83	0.00	0.02	0.94	0.03
Poisson	500	200	100	0.00	1.00	0.00	0.00	1.00	0.00
Poisson	100	20	40	0.01	0.91	0.08	0.00	0.67	0.33
Poisson	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
Poisson	100	40	40	0.67	0.31	0.02	0.34	0.54	0.12
Poisson	500	200	200	0.85	0.15	0.00	0.26	0.75	0.00
Exp.	100	20	20	0.00	0.98	0.02	0.00	0.88	0.12
Exp.	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
Exp.	100	40	20	0.24	0.75	0.00	0.06	0.91	0.04
Exp.	500	200	100	0.01	0.99	0.00	0.00	1.00	0.00
Exp.	100	20	40	0.01	0.90	0.09	0.00	0.71	0.28
Exp.	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
Exp.	100	40	40	0.75	0.23	0.01	0.48	0.44	0.09
Exp.	500	200	200	0.62	0.38	0.00	0.10	0.90	0.00
χ^2_2	100	20	20	0.00	0.98	0.02	0.00	0.88	0.12
χ^2_2	500	100	100	0.00	1.00	0.00	0.00	1.00	0.00
χ^2_2	100	40	20	0.17	0.83	0.00	0.03	0.95	0.03
χ^2_2	500	200	100	0.01	0.99	0.00	0.00	1.00	0.00
χ^2_2	100	20	40	0.02	0.91	0.07	0.00	0.69	0.31
χ^2_2	500	100	200	0.00	1.00	0.00	0.00	1.00	0.00
χ^2_2	100	40	40	0.74	0.24	0.02	0.46	0.45	0.10
χ^2_2	500	200	200	0.71	0.29	0.00	0.15	0.85	0.00

(3) Therefore, the probabilities of selecting the true models are relatively low in cases with $a = 1$, especially when $n = 100, p = 200, k = 200$.

(4) However, observing the tendencies of selecting the true models in cases of $n = 100$ and $n = 500$, we see that these probabilities tend toward 1.

(5) For the cases $n = 500$ and $a = 1$, the consistency is relatively weak, especially when p and k become large.

(6) The case $a = 1/\sqrt{2}$ selects significant variables more easily than the case $a = 1$. This is because the threshold becomes smaller when $a = 1/\sqrt{2}$.

(7) Thus, in the case $a = 1/\sqrt{2}$, the probabilities of selecting the true models are higher compared to the case $a = 1$.

5. Estimation of selection probabilities

In this section, we propose a method to estimate the selection probabilities of explanatory variables using our proposed KOO method. From Bai et al. (2025), it is shown under (C1) $\sim$ (C7) that

$$\sqrt{p}\left(\mathcal{K}_j - \frac{c_n(1+\eta_j)}{1-c_n-\alpha_n}\right)/\sigma_{nj} \to \mathrm{N}(0,1).$$

Therefore, the selection probabilities of explanatory variables using our proposed KOO method is expressed as follows:

$$\begin{aligned} p_j &= P\left(\mathcal{K}_j - \frac{c_n}{1-\alpha_n-c_n} > a\sigma_n\right) \\ &\approx 1 - \Phi\left(\sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\eta_j\right)\Big/\sigma_{nj}\right) \end{aligned}$$

Here, $\Phi(x)$ is the cumulative-distribution function of the standard normal distribution. Then, we consider the following estimator $\hat{p}_j$ for p_j.

$$\hat{p}_j = 1 - \Phi\left(\sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\hat{\eta}_j\right)\Big/\hat{\sigma}_{nj}\right).$$

Here,

$$\hat{\eta}_j = \frac{1}{p}\boldsymbol{x}_j'\mathbf{Q}_j\boldsymbol{x}_j\left(\frac{n-k-p-1}{n}\hat{\boldsymbol{\theta}}_j'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}_j - pv_{ii}\right),$$
$$\hat{\boldsymbol{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\hat{\boldsymbol{\theta}}_j), \quad \hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}, \quad (\mathbf{X}'\mathbf{X})^{-1} = (v_{ij}),$$

where,

$$\boldsymbol{x}_j'\mathbf{Q}_j\boldsymbol{x}_j = \frac{1}{v_{jj}}, \quad \mathcal{K}_j = \frac{p}{n}\left(\frac{1}{p}\boldsymbol{x}_j'\mathbf{Q}_j\boldsymbol{x}_j\hat{\boldsymbol{\theta}}_j'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\theta}}_j\right). \tag{5.1}$$

Thus, we obtain

$$\mathcal{K}_j = \frac{p}{n}\frac{n}{n-k-p-1}(\hat{\eta}_j + 1) = \frac{c_n}{1-c_n-\alpha_n}(\hat{\eta}_j+1) + O(n^{-1}).$$

From this result and Bai et al. (2025), it is shown under (C1) ∼ (C7) that

$$\sqrt{p}\frac{c_n}{1-c_n-\alpha_n}\left(\hat{\eta}_j - \eta_j\right)/\sigma_{nj} \to \mathrm{N}(0,1).$$

Furthermore, it follows that,

$$\hat{\eta}_j \xrightarrow{p} \eta_j, \quad \hat{\sigma}_{nj} = 2c_n^2\frac{(1-\alpha_n)(1+2\hat{\eta}_j)+c_n\hat{\eta}_j^2}{(1-\alpha_n-c_n)^3} \xrightarrow{p} \sigma_{nj}.$$

Under (C1) ∼ (C7), the estimator $\hat{p}_j$ of the selection probability p_j has the following properties. Noting that $\Phi(x)$ is a monotonically increasing function with an inverse function $\Phi^{-1}(x)$, we obtain

$$\begin{aligned}
&P(\hat{p}_j < p_j)\\
&= P\Bigg(1 - \Phi\left(\sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\hat{\eta}_j\right)\Big/\hat{\sigma}_{nj}\right)\\
&\qquad\qquad < 1 - \Phi\left(\sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\eta_j\right)\Big/\sigma_{nj}\right)\Bigg) + o(1)\\
&= P\Bigg(\sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\hat{\eta}_j\right)\Big/\hat{\sigma}_{nj}\\
&\qquad\qquad > \sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\eta_j\right)\Big/\sigma_{nj}\Bigg) + o(1)\\
&= P\left(\sqrt{p}\frac{c_n}{1-\alpha_n-c_n}\left(\hat{\eta}_j - \eta_j\right)\Big/\sigma_{nj} > 0\right) + o(1) = \frac{1}{2} + o(1)
\end{aligned}$$

This is summarized as the following theorem.

Theorem 5.1. *Assume that conditions (C1)–(C7) hold. Then, the estimator $\hat{p}_j$ for p_j is asymptotically median-unbiased.*

6. Numerical experiments: selection probabilities

In this section we give simulation results on estimate the selection probabilities of explanatory variables using our proposed KOO method. Here, the selection probability p_j and its estimator $\hat{p}_j$ are given as follows.

$$\begin{aligned} p_j &= P\left(\mathcal{K}_j - \frac{c_n}{1-\alpha_n-c_n} > a\sigma_n\right) \\ \hat{p}_j &= 1 - \Phi\left(\sqrt{p}\left(a\sigma_n - \frac{c_n}{1-\alpha_n-c_n}\hat{\eta}_j\right) \Big/ \hat{\sigma}_{nj}\right) \end{aligned}$$

We considered the following cases:

- The number of trials:10^3.
- $n = 100, 500$, $c_n = p/n = 0.2, 0.4$, $\alpha_n = k/n = 0.2, 0.4$, $k_* = 5$.
- The components of $\mathbf{X}$ were constructed from a sample of $\mathrm{U}(1,5)$.
- The regression coefficient matrix is given as follows: $\mathbf{\Theta} = (\mathbf{\Theta}_* \quad \mathbf{O})$, $\mathbf{\Theta}_* = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_5) = (\theta_{ij})$. For $n = 100$, the elements θ_{ij} are independently generated from the following distributions. $\theta_{i1} \sim U(0.20, 0.25)$, $\theta_{i2} \sim U(0.25, 0.30)$, $\theta_{i3} \sim U(0.30, 0.35)$, $\theta_{i4} \sim U(0.35, 0.40)$, $\theta_{i5} \sim U(0.40, 0.45)$. For $n = 500$, the elements θ_{ij} are independently generated from the following distributions. $\theta_{i1} \sim U(0.130, 0.135)$, $\theta_{i2} \sim U(0.135, 0.140)$, $\theta_{i3} \sim U(0.140, 0.145)$, $\theta_{i4} \sim U(0.145, 0.150)$, $\theta_{i5} \sim U(0.150, 0.155)$.
- As the distribution of e_{ij}, we considered; (i) Standard normal distribution, (ii) Uniform distribution; $U(0,1)$, (iii) Biomial distribution;

$Bin(1, \rho)$, $\rho = (6 - \sqrt{6})/12$, (iv) Chi-square distribution χ^2 with 12 degrees of freedom, (v) t-distribution with 10 degrees of freedom, (vi) Poisson distribution with parameter 1; $Pos(1)$, (vii) Exponential distribution with parameter 1; $Exp(1)$, (viii) Chi-square distribution χ^2 with 2 degrees of freesdom. Here, all the distributions are normalized as the means 0 and variances 1.

- $a = 1$.

The results of the selection probabilities and their estimators are given in Tables 3 to 6.

In Tables 3 to 6, p_j denotes the selection probability, and "med" denotes the median of its estimator $\hat{p}_j$. The column $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_5$ indicates the selection probabilities of parameters included in the true model, $\min_{i=6,\ldots,k}$ denotes the minimum selection probability among variables not included in the true model, and $\max_{i=6,\ldots,k}$ denotes the maximum selection probability among those not included in the true model.

From Tables 3 to 6, we can identify the following tendncies.

(1) The median of $\hat{p}_j$ is closer to p_j for $n = 500$ than for $n = 100$.

(2) As p or k increases, the median of $\hat{p}_j$ becomes increasingly distant from p_j.

(3) The selection probabilities p_j for variables not included in the true model are essentially zero, and the minimum and maximum median values of their estimators $\hat{p}_j$ are also zero.

7. Concluding Remarks

Bai et al. (2025) proposed two KOO methods for selection of variables in large-dimensional regression. One is based on KOO statistics which are test

Table 3: Selection probabilities under (i) and (ii).

Dist.	n	p	k		$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\min_{i=6,\ldots,k}$	$\max_{i=6,\ldots,k}$
Normal	100	20	20	p_j	0.24	0.70	0.89	0.96	1.00	0.00	0.01
				med	0.22	0.66	0.86	0.92	0.97	0.00	0.00
	100	20	40	p_j	0.12	0.27	0.58	0.98	0.91	0.00	0.01
				med	0.06	0.23	0.55	0.93	0.87	0.00	0.00
	100	40	20	p_j	0.12	0.38	0.60	0.92	0.99	0.00	0.00
				med	0.07	0.34	0.57	0.87	0.96	0.00	0.00
	100	40	40	p_j	0.08	0.10	0.28	0.52	0.90	0.00	0.00
				med	0.02	0.01	0.18	0.45	0.81	0.00	0.00
	500	100	100	p_j	0.96	0.99	1.00	1.00	1.00	0.00	0.00
				med	0.94	0.97	0.99	0.99	1.00	0.00	0.00
	500	100	200	p_j	0.50	0.45	0.81	0.52	0.90	0.00	0.00
				med	0.48	0.43	0.78	0.50	0.87	0.00	0.00
	500	200	100	p_j	0.71	0.86	0.97	0.99	0.99	0.00	0.00
				med	0.68	0.83	0.96	0.97	0.98	0.00	0.00
	500	200	200	p_j	0.05	0.08	0.08	0.27	0.55	0.00	0.00
				med	0.02	0.04	0.04	0.24	0.51	0.00	0.00
Uniform	100	20	20	p_j	0.35	0.75	0.86	0.98	1.00	0.00	0.00
				med	0.33	0.72	0.82	0.95	0.97	0.00	0.00
	100	20	40	p_j	0.19	0.31	0.65	0.88	1.00	0.00	0.01
				med	0.13	0.27	0.62	0.83	0.96	0.00	0.00
	100	40	20	p_j	0.18	0.64	0.82	0.96	1.00	0.00	0.00
				med	0.13	0.60	0.78	0.92	0.97	0.00	0.00
	100	40	40	p_j	0.08	0.23	0.20	0.67	0.58	0.00	0.00
				med	0.02	0.17	0.11	0.61	0.52	0.00	0.00
	500	100	100	p_j	0.96	0.98	1.00	1.00	1.00	0.00	0.00
				med	0.94	0.96	0.99	0.99	1.00	0.00	0.00
	500	100	200	p_j	0.30	0.63	0.61	0.93	0.88	0.00	0.00
				med	0.29	0.61	0.60	0.91	0.86	0.00	0.00
	500	200	100	p_j	0.88	0.97	0.95	1.00	1.00	0.00	0.00
				med	0.88	0.94	0.92	0.99	0.99	0.00	0.00
	500	200	200	p_j	0.13	0.08	0.23	0.30	0.65	0.00	0.00
				med	0.10	0.04	0.19	0.25	0.64	0.00	0.00

Table 4: Selection probabilities under (iii) and (iv).

Dist.	n	p	k		$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\min_{i=6,\ldots,k}$	$\max_{i=6,\ldots,k}$
Binomial	100	20	20	p_j	0.35	0.75	0.86	0.98	1.00	0.00	0.00
				med	0.33	0.72	0.82	0.95	0.97	0.00	0.00
	100	20	40	p_j	0.19	0.31	0.65	0.88	1.00	0.00	0.01
				med	0.13	0.27	0.62	0.83	0.96	0.00	0.00
	100	40	20	p_j	0.18	0.64	0.82	0.96	1.00	0.00	0.00
				med	0.13	0.60	0.78	0.92	0.97	0.00	0.00
	100	40	40	p_j	0.08	0.23	0.20	0.67	0.58	0.00	0.00
				med	0.02	0.17	0.11	0.61	0.52	0.00	0.00
	500	100	100	p_j	0.96	0.98	1.00	1.00	1.00	0.00	0.00
				med	0.94	0.96	0.99	0.99	1.00	0.00	0.00
	500	100	200	p_j	0.30	0.63	0.61	0.93	0.88	0.00	0.00
				med	0.29	0.61	0.60	0.91	0.86	0.00	0.00
	500	200	100	p_j	0.88	0.97	0.95	1.00	1.00	0.00	0.00
				med	0.88	0.94	0.92	0.99	0.99	0.00	0.00
	500	200	200	p_j	0.13	0.08	0.23	0.30	0.65	0.00	0.00
				med	0.10	0.04	0.19	0.25	0.64	0.00	0.00
χ^2_{12}	100	20	20	p_j	0.31	0.63	0.88	0.98	1.00	0.00	0.00
				med	0.30	0.62	0.83	0.94	0.98	0.00	0.00
	100	20	40	p_j	0.22	0.34	0.67	0.86	0.96	0.00	0.01
				med	0.19	0.27	0.62	0.80	0.90	0.00	0.00
	100	40	20	p_j	0.13	0.64	0.86	0.96	1.00	0.00	0.00
				med	0.08	0.59	0.80	0.91	0.97	0.00	0.00
	100	40	40	p_j	0.07	0.17	0.33	0.56	0.77	0.00	0.00
				med	0.01	0.08	0.25	0.50	0.69	0.00	0.00
	500	100	100	p_j	0.99	0.99	1.00	1.00	1.00	0.00	0.00
				med	0.97	0.98	0.99	0.99	1.00	0.00	0.00
	500	100	200	p_j	0.49	0.75	0.66	0.91	0.89	0.00	0.00
				med	0.47	0.74	0.65	0.86	0.86	0.00	0.00
	500	200	100	p_j	0.89	0.92	0.98	0.99	0.98	0.00	0.00
				med	0.86	0.89	0.96	0.99	0.96	0.00	0.00
	500	200	200	p_j	0.04	0.17	0.22	0.42	0.31	0.00	0.00
				med	0.02	0.15	0.19	0.39	0.30	0.00	0.00

Table 5: Selection probabilities under (v) and (vi).

Dist.	n	p	k		$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\min\limits_{i=6,\ldots,k}$	$\max\limits_{i=6,\ldots,k}$
t_{10}	100	20	20	p_j	0.39	0.49	0.92	0.98	0.99	0.00	0.01
				med	0.36	0.47	0.87	0.94	0.97	0.00	0.00
	100	20	40	p_j	0.14	0.38	0.63	0.75	0.99	0.00	0.01
				med	0.08	0.36	0.60	0.70	0.96	0.00	0.00
	100	40	20	p_j	0.17	0.41	0.74	0.99	1.00	0.00	0.00
				med	0.12	0.37	0.68	0.96	0.98	0.00	0.00
	100	40	40	p_j	0.09	0.14	0.28	0.37	0.86	0.00	0.00
				med	0.03	0.07	0.20	0.29	0.78	0.00	0.00
	500	100	100	p_j	0.93	0.98	0.99	1.00	1.00	0.00	0.00
				med	0.90	0.96	0.98	0.99	1.00	0.00	0.00
	500	100	200	p_j	0.29	0.69	0.78	0.90	0.77	0.00	0.00
				med	0.25	0.68	0.76	0.87	0.75	0.00	0.00
	500	200	100	p_j	0.80	0.84	0.98	1.00	0.99	0.00	0.00
				med	0.76	0.82	0.97	0.99	0.98	0.00	0.00
	500	200	200	p_j	0.05	0.10	0.22	0.33	0.62	0.00	0.00
				med	0.02	0.07	0.19	0.30	0.59	0.00	0.00
Poisson	100	20	20	p_j	0.25	0.65	0.90	0.98	1.00	0.00	0.00
				med	0.20	0.63	0.85	0.93	0.98	0.00	0.00
	100	20	40	p_j	0.19	0.32	0.54	0.79	0.97	0.00	0.01
				med	0.12	0.28	0.52	0.75	0.91	0.00	0.00
	100	40	20	p_j	0.17	0.39	0.86	0.99	1.00	0.00	0.00
				med	0.11	0.35	0.79	0.94	0.96	0.00	0.00
	100	40	40	p_j	0.08	0.21	0.48	0.73	0.75	0.00	0.00
				med	0.01	0.10	0.43	0.65	0.68	0.00	0.00
	500	100	100	p_j	0.99	0.94	1.00	0.99	1.00	0.00	0.00
				med	0.98	0.91	0.99	0.99	1.00	0.00	0.00
	500	100	200	p_j	0.43	0.71	0.63	0.76	0.83	0.00	0.00
				med	0.43	0.68	0.61	0.74	0.81	0.00	0.00
	500	200	100	p_j	0.82	0.92	0.98	0.99	1.00	0.00	0.00
				med	0.81	0.88	0.97	0.98	0.99	0.00	0.00
	500	200	200	p_j	0.04	0.10	0.20	0.36	0.27	0.00	0.00
				med	0.01	0.07	0.18	0.36	0.22	0.00	0.00

Table 6: Selection probabilities under (vii) and (viii).

Dist.	n	p	k		$\boldsymbol{\theta}_1$	$\boldsymbol{\theta}_2$	$\boldsymbol{\theta}_3$	$\boldsymbol{\theta}_4$	$\boldsymbol{\theta}_5$	$\min\limits_{i=6,\ldots,k}$	$\max\limits_{i=6,\ldots,k}$
Exp.	100	20	20	p_j	0.51	0.71	0.88	1.00	1.00	0.00	0.00
				med	0.49	0.67	0.84	0.98	0.99	0.00	0.00
	100	20	40	p_j	0.30	0.38	0.50	0.92	0.97	0.00	0.01
				med	0.26	0.33	0.47	0.86	0.92	0.00	0.00
	100	40	20	p_j	0.20	0.42	0.91	0.98	1.00	0.00	0.00
				med	0.14	0.38	0.83	0.93	0.98	0.00	0.00
	100	40	40	p_j	0.05	0.20	0.25	0.72	0.70	0.00	0.00
				med	0.01	0.12	0.18	0.64	0.62	0.00	0.00
	500	100	100	p_j	0.99	0.98	0.99	1.00	1.00	0.00	0.00
				med	0.98	0.96	0.99	1.00	1.00	0.00	0.00
	500	100	200	p_j	0.51	0.59	0.63	0.80	0.75	0.00	0.00
				med	0.50	0.57	0.60	0.78	0.73	0.00	0.00
	500	200	100	p_j	0.87	0.86	0.98	0.95	1.00	0.00	0.00
				med	0.85	0.82	0.97	0.92	0.99	0.00	0.00
	500	200	200	p_j	0.05	0.12	0.16	0.29	0.34	0.00	0.00
				med	0.03	0.11	0.12	0.26	0.33	0.00	0.00
χ^2_2	100	20	20	p_j	0.22	0.65	0.86	0.98	1.00	0.00	0.00
				med	0.18	0.63	0.80	0.94	1.00	0.00	0.00
	100	20	40	p_j	0.19	0.39	0.70	0.75	0.97	0.00	0.01
				med	0.15	0.37	0.66	0.70	0.92	0.00	0.00
	100	40	20	p_j	0.12	0.45	0.80	0.98	0.98	0.00	0.00
				med	0.05	0.42	0.76	0.94	0.93	0.00	0.00
	100	40	40	p_j	0.08	0.11	0.36	0.54	0.72	0.00	0.00
				med	0.01	0.04	0.31	0.48	0.64	0.00	0.00
	500	100	100	p_j	0.99	0.99	0.99	1.00	1.00	0.00	0.00
				med	0.98	0.98	0.98	1.00	0.99	0.00	0.00
	500	100	200	p_j	0.47	0.67	0.78	0.80	0.90	0.00	0.00
				med	0.46	0.66	0.76	0.77	0.89	0.00	0.00
	500	200	100	p_j	0.70	0.86	0.95	0.99	0.99	0.00	0.00
				med	0.68	0.83	0.93	0.98	0.98	0.00	0.00
	500	200	200	p_j	0.18	0.15	0.37	0.44	0.50	0.00	0.00
				med	0.14	0.11	0.36	0.40	0.47	0.00	0.00

statistics for each of the regression coefficients. However, the method involves unknown parameters. The other method based on Bootstrap distribution of KOO statistic. In this paper we propose a modified KOO method for the variable selection problem. The method is easily computed. The consistency property is shown, and its property has been confirmed through a Monte Carlo simulation. In addition, we provide a method to estimate the selection probabilities of explanatory variables in our proposed KOO method, and its accuracy has been confirmed through a Monte Carlo simulation. It is left to compare the present method to the methods under normality due to Fujikoshi (2022) and Oda and Yanagihara (2020, 2021).

Acknowledgements

The second author's research was partially supported by JSPS KAKENHI Grant Number 21K03371.

References

[1] Bai, Z., Choi, K. P., Fujikoshi, Y. and Hu, J. (2018). Asymptotics of AIC, BIC, and Cp model selection rulues in high-dimensional regression. *Bernoulli*, **28**, 2375-2403.

[2] Bai, Z., Choi, K. P., Fujikoshi, Y. and Hu, J. (2025). KOO approach for scalable variable selection problem in large-dimensional regression. To appear in Statistica Sinica.

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

[4] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* Wiley, Hobeken, N.J.

[5] Nishii, R. , Bai, Z. D. and Krishnaia, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.*, **18**, 451–462.

[6] Oda, R., and Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron J. Statist.*, **14**, 1386–1412.

[7] Oda, R., and Yanagihara, H. (2021). A consistent likelihood-based variable selection method in normal multivariate linear regression. *Intelligent Decision Technologies*, I. Czarnowski et al. (eds.), **238**, 391-401,

[8] Zhao, L. C. , Krishnaiah, P. R. and Bai, Z. D. (1986). On determination of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.