# A fast and consistent variable selection method for conditional independence under large-dimensional Gaussian case

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

*Faculty of Data Science, Kyoto Women's University,
35 Kitahiyoshi-cho, Imakumano, Higashiyama-ku, Kyoto 605-8501, Japan
**Center of General Education, Tokyo University of Science, Suwa,
5000-1 Toyohira, Chino, Nagano 391-0292, 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 problem to select non-zero partial correlations under normality-assumed population. It is cumbersome to calculate variable selection criteria for all subsets of pairs of variables when the number of variables is large even though less than sample size. To tackle this problem, we propose a fast and consistent variable selection method based on Baysian information criterion (BIC). The consistency of the method is provided in a high-dimensional asymptotic framework such that the sample size and the number of variables both tend toward infinity under a certain rule. Through numerical simulations, it is shown that the proposed method has a high probability of selecting the true subset of pairs of non-zero partial correlation.

*AMS 2000 Subject Classification*: primary 62H22; secondary 60F10

*Key Words and Phrases*: Variable selection problem, General information criteria, High-dimensional consistency, New model selection criteria, Partial correlations.

*Abbreviated title*: Fast and Consistent Variable Selection of Conditional Independence Under Normality

# 1. Introduction

Let $\boldsymbol{X} = (X_1, \ldots, X_p)'$ be a $p$-dimensional random vector following a multivariate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with unknown mean $\boldsymbol{\mu}$ and unknown nonsingular covariance matrix $\boldsymbol{\Sigma}$. We are interested in identifying or estimating the set of nonzero partial correlations. This problem is called the covariance selection problem (Dempster (1972)) or the Gaussian concentration graph selection problem (see, e.g., Cox and Wermuth, 1996). Here, the partial correlation of $X_i$ and $X_j$ is defined as the usual correlation after removing the effects of the other variables.

We often express the $j_1, j_2$ components of $\boldsymbol{X}$ by $(X_{j_1}, X_{j_2})$ and the $j_1, j_2$ components of $\boldsymbol{\Sigma}$ by $\sigma_{j_1 j_2}$. Suppose that $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})}$ denotes the partial correlation between $X_{j_1}$ and $X_{j_2}$ after removing the effects of all the other variables, denoted by $(-\boldsymbol{j})$, where $\boldsymbol{j} = (j_1, j_2)$. Let $\boldsymbol{\omega}$ be the full set or model. We are interested in finding the true model of which the pairs satisfy $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0$. The true model is described as

$$J_* = \{(j_1, j_2) \mid \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0, j_1, j_2 \in \{1, 2, \ldots, p\}, j_1 < j_2\}.$$

Note that $J_*$ is a subset of $\boldsymbol{\omega} = \{(j_1, j_2) \mid j_1, j_2 = 1, 2, \ldots, p, j_1 < j_2\}$, but is unknown.

Now, we have $k = 2^{p(p+1)/2}$ candidate models by considering whether $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0$ or $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = 0$ for each $(j_1, j_2) \in \boldsymbol{\omega}$. These candidate models are denoted by $M_J$ or simply $J$, which is a subset of $\boldsymbol{\omega}$.

Often, we use AIC or BIC to find the true model, where these criteria are as follows:

$$\mathrm{GIC}_{J,d} = -2 \log L(\widehat{\mathbf{\Sigma}}_J) + dg_J, \tag{1.1}$$

where $L(\widehat{\mathbf{\Sigma}}_J)$ is the maximum likelihood, $dg_J$ is the penalty term, and $g_J$ is the number of unknown parameters. The $d$ values for AIC and BIC are 2 and $\log n$, respectively, and $g_J$ is equal to $p$ plus the number of nonzero partial correlations.

For selecting non-zero partial correlations problem, the best subset deduced from a variable selection criterion such as AIC and BIC typically defined as the subset of pairs of variables which minimize the criterion among all candidate subsets. Identifying this best subset generally involves searching through all candidate subsets. To search all candidate subsets, it needs to calculate the variable selection criteria for $2^{p(p-1)/2} - 1$ subsets. Thus, searching through all candidate subsets will not be feasible when $p$ is large, since the possible number of candidate models become large, and so we need another search method. A practicable selection method was proposed by Nishii et al. (1988) and Zhao et al. (1986) in searching problem of the true subset of explanatory variables in some multivariate models. This method is focussing on the difference of the variable selection criteria for the full set and the one with removing a variable in the full set. If the variable selection criterion for the subset where a variable is removed from full set is larger than the one for full set, then the removed variable is regarded as an element of the best subset. This is called knock-one-out (KOO) method. Fujikoshi (2022) reviewed KOO methods for multivariate regression and discriminant analysis. Applying KOO method with the problem of the selecting non-zero partial correlations, the best subset can be obtained as the pairs of variables such that the value of variable selection criterion for $\boldsymbol{\omega}$ with removing the pair is larger than the one for $\boldsymbol{\omega}$.

Consistency is an important property of the variable selection method. This is defined as the probability that the selected model coincides with the true model converges to 1 as the sample size goes to infinity. Often, it is discussed that BIC has consistency but AIC does not. Nishii et al. (1988) noted that BIC has consistency under the KOO method. Such a consistent variable selection method is going to be likely to select a true variable. Since one cannot know the true subset, one should use the consistent variable selection method with high probability of selecting true variable. It may be noted that there are many cases that the number of variables $p$ is less than the sample size $n$, but both are close (as an example, see Appendix B). In such a case, the probability of selecting true variable by consistent variable selection method may be low.

It may be noted that the theory of variable selection methods would be improved by using a high-dimensional asymptotic framework such that the sample size and the number of variables both tend toward infinity while the ratio of them converges a positive constant less than 1 (see, e.g., Fujikoshi et al., 2010). For example, Yanagihara et al. (2015) improved AIC for multivariate linear models and propose an improvement for searching through all candidate subsets. Besides, Yanagihara et al. (2015) noted that BIC does not have a consistency under a high-dimensional asymptotic framework. Consistent KOO methods are studied by Bai et al. (2018), Sakurai and Fujikoshi (2020) and Oda and Yanagihara (2020, 2021) for multivariate regression problem; Fujikoshi and Sakurai (2019), Oda et al. (2020) for discriminant analysis. It is not studied for selecting non-zero partial correlation problem.

In this paper, we consider the consistency of the KOO method based on the BIC for selecting non-zero partial correlation problem, which is defined as (1.1) with $d$ being $\log n$, under a high-dimensional asymptotic framework:

A1: As $n, p \to \infty$ through $(n, p) \in \mathfrak{I}_1 = \{(n, p) \in \mathbb{N}^2 \ : \ n > p^{4p/(n-p+1)}, n - p > 11\}$, $p/n \to 0$ and $(\log p)/\log n \to \ell \in [0, 1/4)$.

Note that each of the followings are sufficient conditions for A1.

- $p = O(n^{\delta})$, $\delta \in (0, 1/4)$.
- $p = O(\log n)$.

Assumption A1 means that $n$ always tends toward infinity, but $p$ is allowed to diverge under restriction that $(\log p)/\log n \to \ell \in [0, 1/4)$. This means that BIC with KOO method might have high probability for selecting true set for the case that $n$ is much larger than $p$.

We also propose a new consistent KOO method under a high-dimensional asymptotic framework:

A2: As $n, p \to \infty$ through $\mathfrak{I}_2 = \{(n, p) \in \mathbb{N}^2 \ : \ n - p > 11\}$, either one of the following convergences holds.

- $p/n \to 0$ and $(\log p)/\log n \to \ell \in [0, 1)$.
- $p/n \to c \in (0, 1)$.

Since A2 $\Longleftarrow$ A1, A2 is milder than A1. From assumption A2, the true model is selected with high probability for the case in which $p$ is large compared to $n$, but $p < n$.

In recent years, regularization methods have been studied intensively. In Gaussian concentration graph selection problem, it is able to select conditional independence of variables by glasso proposed by Friedmann et al. (2008). Generally, glasso is powerful tool to estimate models. However, it needs heavy computations through algorithms. On the other hand, it may be pointed that our KOO method needs to compute $p(p-1)/2$ selecting criteria. Computational speed compared to such a regularized model selection methods is also reported in Oda and Yanagiraha (2020). In addition, our method has consistency for high-dimensional asymptotic frameworks above.

The present paper is organized as follows. In Section 2, we state the main results of this paper, which include the probability that the selected model is identical to the true model converges to 1. Small scale simulation studies are carried out, and the results are in Section 3. Concluding the paper is given in Section 4. Proofs of the main theorems and the an example of the real data application are provided in the Appendix for interested readers.

# 2. KOO method

## 2.1. Explicit form of KOO statistic

In this subsection, we present the explicit form of the KOO statistic underlying our method. The partial correlation $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})}$ of $X_{j_1}$ and $X_{j_2}$ given $X_{(-\boldsymbol{j})}$ is expressed as follows:

$$\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{\sigma_{j_1 j_2 \cdot (-\boldsymbol{j})}}{\sqrt{\sigma_{j_1 j_1 \cdot (-\boldsymbol{j})}}\sqrt{\sigma_{j_2 j_2 \cdot (-\boldsymbol{j})}}},$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{j_1 j_1 \cdot (-\boldsymbol{j})} &:= \begin{pmatrix} \sigma_{j_1 j_1 \cdot (-\boldsymbol{j})} & \sigma_{j_1 j_2 \cdot (-\boldsymbol{j})} \\ \sigma_{j_2 j_1 \cdot (-\boldsymbol{j})} & \sigma_{j_2 j_2 \cdot (-\boldsymbol{j})} \end{pmatrix} \\ &:= \begin{pmatrix} \sigma_{j_1 j_1} & \sigma_{j_1 j_2} \\ \sigma_{j_2 j_1} & \sigma_{j_2 j_2} \end{pmatrix} - \boldsymbol{\sigma}_{j_1 j_2 \cdot (-\boldsymbol{j})} \boldsymbol{\Sigma}^{-1}_{(-\boldsymbol{j})(-\boldsymbol{j})} \boldsymbol{\sigma}'_{j_1 j_2 \cdot (-\boldsymbol{j})}, \end{aligned} \tag{2.1}$$

$\boldsymbol{\sigma}_{j_1 j_2 \cdot (-\boldsymbol{j})}$ is the partition matrix of $\boldsymbol{\Sigma}$ consisting of the $(j_1, j_2)$ rows after removing the $(j_1, j_2)$ columns, and $\boldsymbol{\Sigma}_{(-\boldsymbol{j})(-\boldsymbol{j})}$ is the partition matrix of $\boldsymbol{\Sigma}$ after removing the $(j_1, j_2)$ columns and $(j_1, j_2)$ rows. Let $\mathbf{S} = (s_{j_1 j_2})$ be the sample covariance matrix based on a sample of size $n + 1$. Then, using partition matrices of $\boldsymbol{S}$ similar to $\boldsymbol{\Sigma}$, the sample partial correlation is obtained as

$$r_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{s_{j_1 j_2 \cdot (-\boldsymbol{j})}}{\sqrt{s_{j_1 j_1 \cdot (-\boldsymbol{j})}}\sqrt{s_{j_2 j_2 \cdot (-\boldsymbol{j})}}}.$$

It is well known that there is a close relationship between the partial correlation coefficients and the coefficients of $\boldsymbol{\Sigma}^{-1} = (\sigma^{j_1 j_2})$, in fact that

$$\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = (-1)^{\delta_{j_1 j_2}+1} \frac{\rho^{j_1 j_2}}{\sqrt{\rho^{j_1 j_1}}\sqrt{\rho^{j_2 j_2}}} = (-1)^{\delta_{j_1 j_2}+1} \frac{\sigma^{j_1 j_2}}{\sqrt{\sigma^{j_1 j_1}}\sqrt{\sigma^{j_2 j_2}}}.$$

Here, $\delta_{j_1 j_2}$ denotes the Kronecker delta and $\rho^{j_1 j_2}$ denotes $(j_1, j_2)$ element of the inverted correlation matrix. Thus, the case of $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = 0$ is equivalent to the one of $\sigma^{j_1 j_2} = 0$.

Our KOO method is specifically as the following way. Let

$$T_{j_1 j_2,d} = \mathrm{GIC}_{\boldsymbol{\omega}\backslash \boldsymbol{j},d} - \mathrm{GIC}_{\boldsymbol{\omega},d} \quad (\boldsymbol{j} = (j_1, j_2)), \tag{2.2}$$

where GIC is defined by (1.1). Following the issue in Nishii et al. (1988) and Zhao et al. (1986), we define the chosen model by KOO method as

$$\widehat{J}_{G,d} = \{(j_1, j_2) \mid T_{j_1 j_2,d} > 0, 1 \le j_1 < j_2 \le p\}. \tag{2.3}$$

We call $T_{j_1 j_2,d}$ as KOO statistics. It is noted that Bai et al. (2018) define slightly modified statistics as KOO statistics in multivariate regression model selection. We can state the selection procedure as follows: if $T_{j_1 j_2,d}$ is positive, $(j_1, j_2)$ is selected, and if $T_{j_1 j_2,d}$ is not positive, $(j_1, j_2)$ is not selected.

Now, we give an explicit form of $T_{j_1 j_2,d}$. It is considered that $T_{j_1 j_2,d}$ in (2.2) is related to the likelihood ratio criterion for the hypothesis $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = 0$. In fact, we can express it as

$$\begin{aligned} T_{j_1 j_2,d} &= -2 \log L(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\backslash \boldsymbol{j},d}) + d g_{\boldsymbol{\omega}\backslash \boldsymbol{j}} \\ &\qquad - \left\{ -2 \log L(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega},d}) + d g_{\boldsymbol{\omega}} \right\} \\ &= -2 \log \mathrm{LRT} - d, \end{aligned}$$

where $\mathrm{LRT} = L(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}\backslash \boldsymbol{j},d}) / L(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega},d})$ is a likelihood ratio statistic for testing the hypothesis $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = 0$. It can be seen that the term $d\,(> 0)$ in the statistics $T_{j_1 j_2,d}$ is like a penalty term to suppress the overestimation due to LRT. We now consider the term $L_{j_1 j_2} = -2 \log \mathrm{LRT}$. Using Fujikoshi et al. (2010, Theorem 4.3.2), we have

$$-2 \log \mathrm{LRT} = -n \log(1 - r^2_{j_1 j_2 \cdot (-\boldsymbol{j})}).$$

Thus, an explicit form of $T_{j_1, j_2, d}$ and its selection procedure are obtained as

$$T_{j_1, j_2, d} = -n \log(1 - r^2_{j_1 j_2 \cdot (-\boldsymbol{j})}) - d > 0 \;\Leftrightarrow\; (j_1, j_2) \in \widehat{J}_d. \tag{2.4}$$

## 2.2. Consistencies of KOO methods based on BIC

In this subsection we deduce consistencies of KOO methods based on BIC and its modifications for selecting Gaussian graphical model, i.e., we prove the consistencies for selecting the method (2.4) with $d = \log n$ and its modification. Before stating these results, we show the needed condition below.

A3: $\inf_{p \in \mathbb{N}} \min_{j_1 j_2 \in J_*} \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})} > 0$.

Note that A3 guarantees that the squared partial correlations in the true model $J_*$ are positive uniformly in $p \in \mathbb{N}$. Our main results of this paper are stated below. The proofs will be given in the Appendix.

First we give the case of $d = \log n$. Under the assumptions A1 and A3, we have proved the consistency of BIC with KOO method, i.e. the limiting probability that $\widehat{J}_{\log n}$ coincides with $J_*$ becomes 1, which is given in the following theorem.

**Theorem 2.1.** *Under assumptions* A1 *and* A3*,*

$$\lim_{\substack{n,p \to \infty \\ (n,p) \in \mathfrak{I}_1}} \Pr\left(\widehat{J}_{\log n} = J_*\right) = 1.$$

Remember that the condition $n > p^{4p/(n-p+1)}$ in $\mathfrak{I}_1$. This is a sufficient condition that the triangular array

$$a_{n,p} = \frac{1}{4} \frac{n-1}{n-p+1} \frac{\log n}{\log p} \quad ((n,p) \in \mathfrak{I}_1).$$

is greater than 1, under which Theorem 2.1 is shown. It is noted that the condition $n > p^{4p/(n-p+1)}$ requires quite large sample size $n$ compared to the dimensionality $p$. For example, $n \geq 160,228$ when $p = 20$, $n \geq 810,395$ when $p = 30$, and so on.

Next, we give an improvement of BIC penalty term and relax the condition for consistency of the KOO method. Through the proof of Theorem

2.1, if we set

$$d = 4\frac{n}{n-p}\log n,$$

then the triangular array $a_{n,p}$ in the proof of Theorem 2.1 is given by

$$a_{n,p} = \frac{\log n}{\log p} > 1 \quad ((n,p) \in \mathfrak{I}_2).$$

Using the same reduction of the proof of Theorem 2.1, we can show a consistency of the model selection (2.4) with $d = \{4n/(n-p)\}\log n$, which is given as the following theorem.

**Theorem 2.2.** *Under assumptions* A2 *and* A3*,*

$$\lim_{\substack{n,p\to\infty \\ (n,p)\in\mathfrak{I}_2}} \Pr\left(\widehat{J}_{4\{n/(n-p)\}\log n} = J_*\right) = 1.$$

We mention some remarks for Theorem 2.2. We do not make any assumption for the order of $q = |J_*| \leq p(p-1)/2$, the size of the true model $J_*$, under the assumption A2 and A3. On the other hand, "$q$ is fixed" or "$\gamma_n = q/n \to \gamma < c$" is assumed in the previous studies (see, e.g., Bai et al. (2018) and Fujikoshi (2022)). We refer to the model selection by $\widehat{J}_{4\{n/(n-p)\}\log n}$ as the KOO method based on modified BIC.

# 3. Simulation Studies

In this section, we numerically examine the performance of the KOO method based on modified BIC in large-dimensional framework with different settings. In Section 2, we proposed an improvement of $d = \log n$ in (2.4) as $d = \{4n/(n-p)\}\log n$. The KOO method chooses the model

$$\hat{j}^{\,\mathrm{MB}} = \widehat{J}_{4\{n/(n-p)\}\log n} = \{(j_1, j_2) \mid L_{j_1,j_2} - 4\{n/(n-p)\}\log n > 0\}.$$

For comparison, we also report for KOO methods with AIC, i.e., $d = 2$, BIC, i.e., $d = \log n$, and GIC for $d = \sqrt{n}$. Note that the consistency for GIC for

$d = \sqrt{n}$ has been shown under the asymptotic framework that $n, p \to \infty$ while $p/n \to c \in (0, 1)$ by following the proof of the consistency given in Fujikoshi and Sakurai (2019). We do not write them due to the redundant derivations, but they have written in our working paper (Fujikoshi et al., 2022). Specifically, the KOO methods with AIC, BIC and GIC for $d = \sqrt{n}$ choose the model

$$\begin{aligned}\hat{j}^{\,\mathrm{A}} &= \widehat{J}_2 = \{(j_1, j_2) \mid L_{j_1,j_2} - 2 > 0\}\},\\ \hat{j}^{\,\mathrm{B}} &= \widehat{J}_{\log n} = \{(j_1, j_2) \mid L_{j_1,j_2} - \log n > 0\}\},\\ \hat{j}^{\,\mathrm{G}} &= \widehat{J}_{\sqrt{n}} = \{(j_1, j_2) \mid L_{j_1,j_2} - \sqrt{n} > 0\}\}.\end{aligned}$$

In addition, we computed the probabilities of selecting the true model using glasso proposed by Friedmann et al. (2008) for the case in which $p \in \{10, 50\}$. Following Friedmann et al. (2008), we consider sparse and dense setting. The sparse setting is the bidiagonal matrix: $(\boldsymbol{\Sigma}^{-1})_{ii} = 1$, $(\boldsymbol{\Sigma}^{-1})_{i,i-1} = (\boldsymbol{\Sigma}^{-1})_{i-1,i} = -0.5$, and zero otherwise. In the dense setting, $\boldsymbol{\Sigma}^{-1} = \mathbf{I} + \mathbf{1}_p\mathbf{1}_p'$. Here, $(\boldsymbol{\Sigma}^{-1})_{ij}$ denotes $(i, j)$th entry in the matrix, $i, j \in \{1, \ldots, p\}$ and $\mathbf{1}_p$ denotes a $p$-dimensional vector of 1's. In addition, we consider the identity setting. We take the sample size as $n \in \{200, 350, 500, 700, 1000\}$ and the dimension $p \in \{10, 50\}$.

Table 1 presents the averaging the true selection times over $10,000$ simulations. The columns in the table represent the case, with first four columns for KOO methods (modified BIC, AIC, BIC and GIC with $d = \sqrt{n}$) and the last for glasso. Here, the result by glasso is given by using the R package "CVglasso", which is obtained in `https://github.com/MGallow/CVglasso`. The upper panel is reserved for sparse case, the middle panel for dense case and the lower panel for identity case.

To begin with, we observe a high performance of selecting the true model in all cases of $p$ and $n$ for KOO methods with modified BIC and GIC. Comparing the case in $(n, p) = (200, 50)$ shows that the performance of modified BIC is superior than the one of the GIC under the identity setting. The dense setting depicts opposite exceptions compared to the identity setting.

The performance under dense setting appears to perform well except for the case in which $n = 200$. In our dense setting, although all partial correlations are non-zero, the performance remains high.

We mention for the performance of KOO method based on AIC and BIC. Although BIC with KOO method has been proven to be consistent, the performance of selecting the true model is not so good. Even for the case that $(n, p) = (1000, 10)$ and the sparse setting, the proportion of the true model is 0.7. It is expected that the proportion gets close to 1 if $n$ becomes large for fixed $p\,(= 10)$. The performance for $p = 50$ is not acceptable for BIC under sparse and identity settings. The performance of AIC is not acceptable under our settings.

The performance of selecting the true model is poor for glasso in our settings. The proportion of the true model is 0 in the sparse and the dense settings, and is about 0.4 in the identity case.

Next, we concentrate on the comparisons of performance among KOO methods base on modified BIC and GIC for $p > 50$, especially examine simulation for some $r = p/n \geq 1/3$. We set $p = \lfloor nr \rfloor$ for $r \in \{1/3, 1/2, 3/5, 3/4\}$ and $n \in \{200, 350, 500, 700, 1000, 1500, 2000\}$. Here, the notation "$\lfloor\ \rfloor$" represents the floor function. The results given in Table 2 is the same as that in Table 1. The columns in the table represent the case for $n$. The rows are reserved for $r$ in each of sparse, dense and identity cases.

The results are very similar to that in Table 1. First, it requires large $n$ to obtain high performance as $r$ gets close to 1. In the sparse setting, modified BIC with KOO method select the true model with higher probability than GIC, although it needs relatively larger sample size. Its highly performance is particularly remarkable under identity setting. On the other hand, KOO method with modified BIC seems to be inferior than that with GIC under the dense setting. Focus on the case of the combination of $(n, p)$ that the proportion of true model is close to 1 (e.g., $\geq 0.95$), if GIC is close to 1 then the modified BIC is also close. However the converse does not hold. In this sense, we recommend to use the modified BIC with KOO method if we are

Table 1: Proportion of selecting all nonzero partial correlations when $p \in \{10, 50\}$.

Sparse

| $p$ | $n$ | $\hat{j}^{\text{MB}}$ | $\hat{j}^{\text{A}}$ | $\hat{j}^{\text{B}}$ | $\hat{j}^{\text{G}}$ | glasso |
|---|---|---|---|---|---|---|
| 10 | 200 | 0.99 | 0.01 | 0.48 | 0.99 | 0.00 |
| | 350 | 1.00 | 0.01 | 0.60 | 1.00 | 0.00 |
| | 500 | 1.00 | 0.01 | 0.60 | 1.00 | 0.00 |
| | 700 | 1.00 | 0.01 | 0.71 | 1.00 | 0.00 |
| | 1000 | 1.00 | 0.01 | 0.75 | 1.00 | 0.00 |
| 50 | 200 | 0.39 | 0.00 | 0.00 | 0.31 | 0.00 |
| | 350 | 1.00 | 0.00 | 0.00 | 0.93 | 0.00 |
| | 500 | 1.00 | 0.00 | 0.00 | 0.99 | 0.00 |
| | 700 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| | 1000 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |

Dense

| $p$ | $n$ | $\hat{j}^{\text{MB}}$ | $\hat{j}^{\text{A}}$ | $\hat{j}^{\text{B}}$ | $\hat{j}^{\text{G}}$ | glasso |
|---|---|---|---|---|---|---|
| 10 | 200 | 0.94 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 350 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 500 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 700 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 1000 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| 50 | 200 | 0.09 | 1.00 | 1.00 | 0.84 | 0.00 |
| | 350 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 500 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 700 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |
| | 1000 | 1.00 | 1.00 | 1.00 | 1.00 | 0.00 |

Identity

| $p$ | $n$ | $\hat{j}^{\text{IB}}$ | $\hat{j}^{\text{A}}$ | $\hat{j}^{\text{B}}$ | $\hat{j}^{\text{G}}$ | glasso |
|---|---|---|---|---|---|---|
| 10 | 200 | 1.00 | 0.00 | 0.33 | 0.99 | 0.40 |
| | 350 | 1.00 | 0.00 | 0.47 | 1.00 | 0.38 |
| | 500 | 1.00 | 0.00 | 0.55 | 1.00 | 0.38 |
| | 700 | 1.00 | 0.00 | 0.61 | 1.00 | 0.38 |
| | 1000 | 1.00 | 0.00 | 0.67 | 1.00 | 0.37 |
| 50 | 200 | 1.00 | 0.00 | 0.00 | 0.26 | 0.48 |
| | 350 | 1.00 | 0.00 | 0.00 | 0.93 | 0.44 |
| | 500 | 1.00 | 0.00 | 0.00 | 0.99 | 0.43 |
| | 700 | 1.00 | 0.00 | 0.00 | 1.00 | 0.41 |
| | 1000 | 1.00 | 0.00 | 0.00 | 1.00 | 0.40 |

going to select the true model since one cannot know whether the true model is sparse or dense.

# 4. Discussion and conclusions

Fast and consistent variable selection method is proposed for conditional independence under Gaussian case. Our primary proposed method is based on applying the idea of the selection in Nishii et al. (1988) and Zhao et al. (1986) to modified BIC, and is composed of simple, computationally efficient. Computation of our method is remarkably faster than glasso. Details are reported in Oda and Yanagiraha (2020) for the problem of selecting multivariate regression.

We also compare our proposed methods with glasso for the precision of selecting true model. The primary proposed method generally outperforms in most cases. It may be mentioned that glasso can apply the high-dimensional data whose the dimensionality is larger than the sample size, however proposed method cannot. Improving our proposed method for high-dimensional case is left as a future work.

# A Proofs of Theorem 2.1-2.2

In this Appendix, we present the proofs of Theorem 2.1-2.2. Before stating them, we give preliminary results. In subsections A2 and A3 we state the proof of consistencies.

Hereafter, we use the notation $\chi^2_f$ as the random variable distributed as chi-square distribution with $f$ degrees of freedom.

Table 2: Proportion of selecting all nonzero partial correlations.

| Sparse | | | | | $n$ | | | |
|---|---|---|---|---|---|---|---|---|
| $r$ | | 200 | 350 | 500 | 700 | 1000 | 1500 | 2000 |
| 1/3 | $\hat{j}^{\mathrm{MB}}$ | 0.05 | 0.97 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | $\hat{j}^{\mathrm{G}}$ | 0.03 | 0.09 | 0.24 | 0.50 | 0.79 | 0.95 | 0.99 |
| 1/2 | $\hat{j}^{\mathrm{MB}}$ | 0.00 | 0.22 | 0.95 | 0.98 | 0.98 | 0.98 | 0.98 |
| | $\hat{j}^{\mathrm{G}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.32 |
| 3/5 | $\hat{j}^{\mathrm{MB}}$ | 0.00 | 0.00 | 0.39 | 0.96 | 0.98 | 0.97 | 0.98 |
| | $\hat{j}^{\mathrm{G}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3/4 | $\hat{j}^{\mathrm{MB}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.74 | 0.96 | 0.96 |
| | $\hat{j}^{\mathrm{G}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Dense | | | | | $n$ | | | |
| $r$ | | 200 | 350 | 500 | 700 | 1000 | 1500 | 2000 |
| 1/3 | $\hat{j}^{\mathrm{MB}}$ | 0.01 | 0.65 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | $\hat{j}^{\mathrm{G}}$ | 0.65 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 1/2 | $\hat{j}^{\mathrm{MB}}$ | 0.00 | 0.02 | 0.49 | 0.99 | 1.00 | 1.00 | 1.00 |
| | $\hat{j}^{\mathrm{G}}$ | 0.22 | 0.89 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 3/5 | $\hat{j}^{\mathrm{MB}}$ | 0.00 | 0.00 | 0.03 | 0.60 | 1.00 | 1.00 | 1.00 |
| | $\hat{j}^{\mathrm{G}}$ | 0.07 | 0.59 | 0.97 | 1.00 | 1.00 | 1.00 | 1.00 |
| 3/4 | $\hat{j}^{\mathrm{MB}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.96 | 1.00 |
| | $\hat{j}^{\mathrm{G}}$ | 0.01 | 0.09 | 0.43 | 0.90 | 1.00 | 1.00 | 1.00 |
| Identity | | | | | $n$ | | | |
| $r$ | | 200 | 350 | 500 | 700 | 1000 | 1500 | 2000 |
| 1/2 | $\hat{j}^{\mathrm{MB}}$ | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| | $\hat{j}^{\mathrm{G}}$ | 0.02 | 0.08 | 0.22 | 0.49 | 0.79 | 0.95 | 0.99 |
| 1/3 | $\hat{j}^{\mathrm{MB}}$ | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| | $\hat{j}^{\mathrm{G}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.33 |
| 3/5 | $\hat{j}^{\mathrm{MB}}$ | 0.97 | 0.97 | 0.97 | 0.98 | 0.97 | 0.98 | 0.97 |
| | $\hat{j}^{\mathrm{G}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3/4 | $\hat{j}^{\mathrm{MB}}$ | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| | $\hat{j}^{\mathrm{G}}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

## A1. Preliminary Results

We consider the distribution of the statistic

$$T_{j_1,j_2,d} + d = -n\log(1 - r^2_{j_1,j_2\cdot(-\boldsymbol{j})}).$$

Using $r^2_{j_1j_2\cdot(-\boldsymbol{j})} = s^2_{j_1j_2\cdot(-\boldsymbol{j})} \cdot \{s_{j_1j_1\cdot(-\boldsymbol{j})}s_{j_2j_2\cdot(-\boldsymbol{j})}\}^{-1}$, we can write

$$-n\log(1 - r^2_{j_1,j_2\cdot(-\boldsymbol{j})}) = n\log(1 + Q_{j_1j_2\cdot(-\boldsymbol{j})}),$$

where

$$Q_{j_1j_2\cdot(-\boldsymbol{j})} = \frac{s^2_{j_1j_2\cdot(-\boldsymbol{j})}}{s_{j_1j_1\cdot(-\boldsymbol{j})}s_{j_2j_2\cdot(-\boldsymbol{j})} - s^2_{j_1j_2\cdot(-\boldsymbol{j})}}. \tag{A.1}$$

Thus, it is necessary to study the distribution of $Q_{j_1j_2\cdot(-\boldsymbol{j})}$ in order to obtain the distribution of $-n\log(1 - r^2_{j_1,j_2\cdot(-\boldsymbol{j})})$. For this purpose, we have the following theorem.

**Theorem A1.** *Let* $Q_{j_1j_2\cdot(-\boldsymbol{j})}$ *be the statistic defined by* (A.1). *Then we can express it as a ratio of independent chi-square variates:*

$$Q_{j_1j_2\cdot(-\boldsymbol{j})} = \frac{\left(Z + \frac{\rho_{j_1j_2\cdot(-\boldsymbol{j})}}{\sqrt{1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}}\sqrt{\chi^2_m}\right)^2}{\chi^2_{m-1}} \tag{A.2}$$

*where* $Z$ *denotes the variate distributed as the standard normal distribution,* $\chi^2_m$ *denotes chi-square variate with* $m = n - (p-2)$ *degrees of freedom,* $\chi^2_{m-1}$ *denotes chi-square variate with* $m - 1 = n - p + 1$ *degrees of freedom, and* $Z$*,* $\chi^2_m$ *and* $\chi^2_{m-1}$ *are independent.*

Hereafter, we abbreviate the formula $(A.2)$ as

$$Q_{j_1j_2\cdot(-\boldsymbol{j})} = \chi^2_1\left(\tau^2\right)\left\{\chi^2_{m-1}\right\}^{-1}, \tag{A.3}$$

where $\chi^2_1(\lambda)$ is noncentral chi-square variate with 1 degree of freedom and the noncentrality parameter $\lambda$,

$$\tau^2 = \rho^2_{j_1j_2\cdot(-\boldsymbol{j})}(1 - \rho^2_{j_1j_2\cdot(-\boldsymbol{j})})^{-1}\chi^2_m,$$

and $\chi^2_1(\cdot)$, $\chi^2_m$ and $\chi^2_{m-1}$ are independent.

**Proof of Theorem A1**: First, note that

$$n \begin{pmatrix} s_{j_1 j_1 \cdot (-\boldsymbol{j})} & s_{j_1 j_2 \cdot (-\boldsymbol{j})} \\ s_{j_2 j_1 \cdot (-\boldsymbol{j})} & s_{j_2 j_2 \cdot (-\boldsymbol{j})} \end{pmatrix} \sim \mathrm{W}_2(m, \boldsymbol{\Sigma}_{j_1 j_2 \cdot (-\boldsymbol{j})}),$$

where $\mathrm{W}_2(m, \boldsymbol{\Sigma}_{j_1 j_2 \cdot (-\boldsymbol{j})})$ denotes the two-dimensional Wishart distribution with $m = n - (p-2)$ degrees of freedom and covariance matrix $\boldsymbol{\Sigma}_{j_1 j_2 \cdot (-\boldsymbol{j})}$ defined in (2.1). We can express $Q_{j_1 j_2 \cdot (-\boldsymbol{j})}$ as

$$Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{w^2_{j_1 j_2 \cdot (-\boldsymbol{j})}}{w_{j_1 j_1 \cdot (-\boldsymbol{j})} w_{j_2 j_2 \cdot (-\boldsymbol{j})} - w^2_{j_1 j_2 \cdot (-\boldsymbol{j})}}, \tag{A.4}$$

where $w_{j_1 j_2 \cdot (-\boldsymbol{j})} = n s_{j_1 j_2 \cdot (-\boldsymbol{j})} \left\{ \sigma_{j_1 j_1 \cdot (-\boldsymbol{j})} \cdot \sigma_{j_2 j_2 \cdot (-\boldsymbol{j})} \right\}^{-1/2}$. We simply write $\mathbf{W}$ to indicate the two-dimensional random matrix $\mathbf{W}_{j_1 j_2 \cdot (-\boldsymbol{j})}$, which is defined as follows:

$$\mathbf{W}_{j_1 j_2 \cdot (-\boldsymbol{j})} = \begin{pmatrix} w_{j_1 j_1 \cdot (-\boldsymbol{j})} & w_{j_1 j_2 \cdot (-\boldsymbol{j})} \\ w_{j_2 j_1 \cdot (-\boldsymbol{j})} & w_{j_2 j_2 \cdot (-\boldsymbol{j})} \end{pmatrix}.$$

Then

$$\mathbf{W} \sim \mathrm{W}_2 \left( m, \begin{pmatrix} 1 & \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \\ \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} & 1 \end{pmatrix} \right).$$

From the definition of the Wishart distribution, we can assert $\mathbf{W} = \mathbf{U}'\mathbf{U}$, where

$$\mathbf{U} = (\boldsymbol{u}_1 \;\; \boldsymbol{u}_2) \sim \mathrm{N}_{m \times 2} \left( \mathbf{O}, \mathbf{I}_m \otimes \begin{pmatrix} 1 & \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \\ \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} & 1 \end{pmatrix} \right),$$

in which $\mathbf{A} \otimes \mathbf{B}$ means the Kronecker product of the two matrices $\mathbf{A}$ and $\mathbf{B}$ (see, e.g., Muirhead, 2009). Then, we can write $Q_{j_1 j_2 \cdot (-\boldsymbol{j})}$ in (A.4) as follows:

$$Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{\boldsymbol{u}_2' \frac{1}{\boldsymbol{u}_1' \boldsymbol{u}_1} \boldsymbol{u}_1 \boldsymbol{u}_1' \boldsymbol{u}_2}{\boldsymbol{u}_2' \left( \mathbf{I}_m - \frac{1}{\boldsymbol{u}_1' \boldsymbol{u}_1} \boldsymbol{u}_1 \boldsymbol{u}_1' \right) \boldsymbol{u}_2}.$$

The conditional distribution of $\boldsymbol{u}_2$ given $\boldsymbol{u}_1$ is

$$\boldsymbol{u}_2 | \boldsymbol{u}_1 \sim \mathrm{N}_m(\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \boldsymbol{u}_1, (1 - \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})}) \mathbf{I}_m).$$

Using this conditional distribution, we can claim that

$$\boldsymbol{u}_2' \left( \mathbf{I}_m - \frac{1}{\boldsymbol{u}_1' \boldsymbol{u}_1} \boldsymbol{u}_1 \boldsymbol{u}_1' \right) \boldsymbol{u}_2 \sim (1 - \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})}) \chi^2_{m-1},$$

and this is independent to $\boldsymbol{u}_1$. In general, the numerator and the denominator are conditionally independent. The conditional distribution of the numerator $\boldsymbol{u}_2' \frac{1}{\boldsymbol{u}_1'\boldsymbol{u}_1} \boldsymbol{u}_1 \boldsymbol{u}_1' \boldsymbol{u}_2$ given $\boldsymbol{u}_1$ is a noncentral chi-squared distribution such that the number of degrees of freedom is 1 and the noncentral parameter is $\tau^2$, where

$$\tau^2 = \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})} \left\{1 - \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})}\right\}^{-1} \boldsymbol{u}_1' \boldsymbol{u}_1.$$

These imply Theorem A1. □

Next, we show the lower bound of the probability that $\widehat{J}_d = J_*$. It holds that

$$\begin{aligned}
\Pr(\widehat{J}_d = J_*) &= \Pr\left(\left(\bigcap_{(j_1,j_2)\in J_*} \text{“}T_{j_1 j_2, d} > 0\text{”}\right) \cap \left(\bigcap_{(j_1,j_2)\notin J_*} \text{“}T_{j_1 j_2, d} \leq 0\text{”}\right)\right) \\
&= 1 - \Pr\left(\left(\bigcup_{(j_1,j_2)\in J_*} \text{“}T_{j_1 j_2, d} \leq 0\text{”}\right) \cup \left(\bigcup_{(j_1,j_2)\notin J_*} \text{“}T_{j_1 j_2, d} > 0\text{”}\right)\right) \\
&\geq 1 - \sum_{(j_1,j_2)\in J_*} \Pr(T_{j_1, j_2, d} \leq 0) - \sum_{(j_1,j_2)\notin J_*} \Pr(T_{j_1, j_2, d} > 0).
\end{aligned}$$

Thus, $\Pr(\widehat{J}_d = J_*)$ converges to 1, i.e. variable selection method $\widehat{J}_d$ has consistency if the following [F1] and [F2] hold.

$$\begin{aligned}
\text{[F1]}: \ \text{P1} &\equiv \sum_{(j_1,j_2)\in J_*} \Pr(T_{j_1 j_2, d} \leq 0) \to 0. \\
\text{[F2]}: \ \text{P2} &\equiv \sum_{(j_1,j_2)\notin J_*} \Pr(T_{j_1 j_2, d} > 0) \to 0.
\end{aligned}$$

This approach has been used in Fujikoshi and Sakurai (2019), Oda and Yanagihara (2021), and Fujikoshi (2022), as well as other studies.

## A2. Proof of Theorem 2.1

This subsection treats the proof of Theorem 2.1. In Section A1, we mentioned that Theorem 2.1 is proved by showing [F1] and [F2]; thus we deduce them in Section A2.1-A2.2.

### A2.1. Proof of [F2]

Firstly, we give a bound of $\Pr(T_{j_1j_2,d} > 0)$. When $(j_1, j_2) \notin J_*$, from Theorem A1 we can write $T_{j_1j_2,d} = n\log\left(1+\chi_1^2/\chi_{m-1}^2\right) - d$, where $\chi_1^2$ denotes chi-square variate with 1 degree of freedom which is independent to $\chi_{m-1}^2$, and therefore we have

$$\Pr(T_{j_1j_2,d} > 0) = \Pr\left(n\log\left(1+\frac{\chi_1^2}{\chi_{m-1}^2}\right) - d > 0\right).$$

It is observed that

$$n\log\left(1+\frac{\chi_1^2}{\chi_{m-1}^2}\right) - d > 0 \iff \frac{\chi_1^2}{\chi_{m-1}^2} > e^{d/n} - 1.$$

Then we can deduce that

$$\Pr(T_{j_1j_2,d} > 0) = \Pr\left(\frac{\chi_1^2}{\chi_{m-1}^2} > e^{d/n} - 1\right) \le \Pr\left(\frac{\chi_1^2}{\chi_{m-1}^2} > \frac{d}{n}\right). \tag{A.5}$$

Define

$$a_{n,p} = \frac{1}{4}\frac{m-1}{n}\frac{\log n}{\log p}. \tag{A.6}$$

We find that

$$a_{n,p} > 1 \qquad \iff \qquad n > p^{4n/(m-1)} = p^{4n/(n-p+1)} \qquad \Longleftarrow \qquad (n,p) \in \mathfrak{I}_1. \tag{A.7}$$

Suppose that $E_{n,p} = \{\chi_{m-1}^2 \ge (m-1)/a_{n,p}\}$ for $(n,p) \in \mathfrak{I}_1$. Then, it follows that

$$\begin{aligned}
\Pr\left(\frac{\chi_1^2}{\chi_{m-1}^2} > \frac{d}{n}\right) &= \Pr\left(\left[\left\{\frac{\chi_1^2}{\chi_{m-1}^2} > \frac{d}{n}\right\} \cap E_{n,p}\right] \cup \left[\left\{\frac{\chi_1^2}{\chi_{m-1}^2} > \frac{d}{n}\right\} \cap E_{n,p}^c\right]\right) \\
&= \Pr\left(\left\{\frac{\chi_1^2}{\chi_{m-1}^2} > \frac{d}{n}\right\} \cap E_{n,p}\right) + \Pr\left(\left\{\frac{\chi_1^2}{\chi_{m-1}^2} > \frac{d}{n}\right\} \cap E_{n,p}^c\right) \\
&\le \Pr\left(\chi_1^2 > \frac{m-1}{n}\frac{d}{a_{n,p}}\right) + \Pr\left(\chi_{m-1}^2 < \frac{m-1}{a_{n,p}}\right),
\end{aligned}$$

and so

$$\Pr(T_{j_1j_2,d} > 0) \le \Pr\left(\chi_1^2 > \frac{m-1}{n}\frac{d}{a_{n,p}}\right) + \Pr\left(\chi_{m-1}^2 < \frac{m-1}{a_{n,p}}\right). \tag{A.8}$$

By showing that the right-hand side of inequality $(A.8)$ is $o(p^{-2})$, we claim that $\Pr(T_{j_1 j_2,d} > 0) = o(p^{-2})$. First, we consider the order of the first term in right-hand side of the inequality $(A.8)$. Let $Z_0$ be the standard normal variate. Then we have

$$\begin{aligned}
p^2 \Pr\left(\chi_1^2 > \frac{m-1}{n}\frac{d}{a_{n,p}}\right) &= p^2 \Pr\left(|Z_0| > \sqrt{\frac{m-1}{n}\frac{d}{a_{n,p}}}\right) \\
&\leq p^2 \sqrt{\frac{2}{\pi}} \left\{\frac{m-1}{n}\frac{d}{a_{n,p}}\right\}^{-1/2} \exp\left(-\frac{1}{2}\frac{m-1}{n}\frac{d}{a_{n,p}}\right) \\
&= \sqrt{\frac{2}{\pi}} \left\{\frac{m-1}{n}\frac{d}{a_{n,p}}\right\}^{-1/2} \\
&\quad \cdot \exp\left\{-\frac{1}{2}\frac{m-1}{n}\frac{d}{a_{n,p}}\left(1 - 4\frac{n}{m-1}\frac{a_{n,p}}{d}\log p\right)\right\} \qquad \text{(A.9)} \\
&= \sqrt{\frac{2}{\pi}} \left(4\log p\right)^{-1/2},
\end{aligned}$$

where the last equality follows from the setting of $d = \log n$ and the definition of $a_{n,p}$ given in $(A.6)$. This implies that

$$\Pr\left(\chi_1^2 > \frac{m-1}{n}\frac{d}{a_{n,p}}\right) = o(p^{-2}). \qquad \text{(A.10)}$$

Next, we consider the order of the second term in right-hand side of the inequality $(A.8)$. It can be expressed that

$$\begin{aligned}
\Pr\left(\chi_{m-1}^2 < \frac{m-1}{a_{n,p}}\right) &= \Pr\left(\chi_{m-1}^2 - (m-1) < -\left(1 - a_{n,p}^{-1}\right)(m-1)\right) \\
&= \Pr\left(-\chi_{m-1}^2 - \{-(m-1)\} > \left(1 - a_{n,p}^{-1}\right)(m-1)\right) \\
&= \Pr\left(\sum_{i=1}^{m-1}\{-Z_i^2 - (-1)\} > \left(1 - a_{n,p}^{-1}\right)(m-1)\right), \qquad \text{(A.11)}
\end{aligned}$$

where $Z_1, Z_2, \ldots, Z_{m-1}$ are independent standard normal variates. Now, it holds that

$$\frac{1}{a_{n,p}} = \frac{\log p^{4n/(n-p+1)}}{\log n} > 0 \quad ((n,p) \in \mathfrak{I}_1).$$

From $(A.7)$ we find that $a_{n,p}^{-1} < 1$ for $(n,p) \in \mathfrak{I}_1$, and thus we claim that $0 < a_{n,p}^{-1} < 1$. It follows from Bernstein's inequality that the right-hand side of the equality in $(A.11)$ is dominated in the following way:

$$\begin{aligned}
&\Pr\left(\sum_{i=1}^{m-1}\{-Z_i^2 - (-1)\} > \left(1 - a_{n,p}^{-1}\right)(m-1)\right)\\
&\quad \le \exp\left(-\frac{\{(1 - a_{n,p}^{-1})(m-1)\}^2/W}{2\{1 + (1 - a_{n,p}^{-1})(m-1)B/W\}}\right), \qquad \text{(A.12)}
\end{aligned}$$

where $B$ and $W$ are positive constants satisfying $\sum_{i=1}^{m-1}\mathrm{E}[|-Z_i^2+1|^2] \le W$ and

$$\mathrm{E}(|-Z_i^2+1|^k) \le \frac{1}{2}\mathrm{E}(|-Z_i^2+1|^2)B^{k-2}k! \quad (k \in \{2,3,\ldots\}).$$

Such constants can be taken as $W = 2(m-1) = \sum_{i=1}^{m-1}\mathrm{E}[(Z_i^2-1)^2]$ and $B = 12$. Here, the latter case is shown in the following way: For $Z \sim N(0,1)$ and $k \ge 3$,

$$\begin{aligned}
\mathrm{E}(|Z^2-1|^k) &\le \mathrm{E}(|Z^2+1|^k) = 2^k\mathrm{E}\left[\left|\frac{Z^2}{2}+\frac{1}{2}\right|^k\right]\\
&\le 2^k\mathrm{E}\left[\frac{|Z^2|^k}{2}+\frac{1}{2}\right]\\
&= 2^{k-1}\{\mathrm{E}(Z^{2k})+1\} = 2^{k-1}\left\{\prod_{r=0}^{k-1}(1+2r)+1\right\}\\
&= 2^{k-1}\left\{\frac{\prod_{r=0}^{k-1}(1+2r)}{2^{k-1}k!}+\frac{1}{2^{k-1}k!}\right\}2^{k-1}k!\\
&\le \left\{\frac{3\cdot 5}{4\cdot 6}+\frac{1}{4\cdot 6}\right\}4^{k-1}k! = \frac{8}{3}\cdot 4^{k-2}k! < 3\cdot 4^{k-2}k!\\
&\le 3^{k-2}4^{k-2}k! = \frac{1}{2}\mathrm{E}[|Z^2-1|^2]12^{k-2}k!,
\end{aligned}$$

where the second inequality follows from Jensen's inequality. From the inequality given in $(A.12)$ we obtain

$$\begin{aligned}
&p^2\mathrm{Pr}\left(\chi^2_{m-1} < \frac{m-1}{a_{n,p}}\right) \\
&\leq p^2\mathrm{exp}\left(-\frac{\{(1-a_{n,p}^{-1})(m-1)\}^2/W}{2\{1+(1-a_{n,p}^{-1})(m-1)B/W\}}\right) \\
&= \exp\left(-\frac{m-1}{2}\left\{\frac{(1-a_{n,p}^{-1})^2}{W/(m-1)+(1-a_{n,p}^{-1})B} - 4\frac{\log p}{m-1}\right\}\right). \qquad \text{(A.13)}
\end{aligned}$$

Under A5, $a_{n,p}^{-1} \to a_0 \in [0,1)$; thus the right-hand side of the equality in $(A.13)$ converges to 0. This implies that

$$\mathrm{Pr}\left(\chi^2_{m-1} < \frac{m-1}{a_{n,p}}\right) = o(p^{-2}). \qquad \text{(A.14)}$$

It follows that

$$\begin{aligned}
\mathrm{P2} &= \sum_{(j_1,j_2)\notin J_*} \mathrm{Pr}(T_{j_1,j_2,d} > 0) \\
&\leq p^2\mathrm{Pr}\left(\frac{\chi^2_1}{\chi^2_{m-1}} > e^{d/n}-1\right) \\
&\leq p^2\{o(p^{-2})+o(p^{-2})\} = o(1),
\end{aligned}$$

where the first inequality holds from $(A.5)$ and the second inequality holds from $(A.13)$, $(A.10)$ and $(A.14)$; thus P2 → 0.

### A2.2. Proof of [F1]

Firstly, we give a bound of $\mathrm{Pr}(T_{j_1j_2,d} \leq 0)$. Let

$$h = h(\chi^2_1, \chi^2_{m-1}, \chi^2_m) = \frac{\chi^2_1\left(\frac{\rho^2_{j_1j_2\cdot(-j)}}{1-\rho^2_{j_1j_2\cdot(-j)}}\chi^2_m\right)}{\chi^2_{m-1}}.$$

When $(j_1, j_2) \in J_*$, from Theorem A1 we have

$$\begin{aligned}
\Pr(T_{j_1,j_2,d} \le 0) &= \Pr\left(n \log(1+h) - d \le 0\right) \\
&= \Pr\left((1+h))\, e^{-d/n} \le 1\right) \\
&= \Pr\left((1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h)\, e^{-d/n} - 1 \le -\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}\right) \\
&\le \Pr\left(\left|(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h)\, e^{-d/n} - 1\right| \ge \rho^2_{j_1j_2\cdot(-\boldsymbol{j})}\right) \\
&\le \frac{1}{(\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})^5} \mathrm{E}\left[\left|(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h)\, e^{-d/n} - 1\right|^5\right] \\
&\le \frac{1}{(\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})^5} \left(\mathrm{E}\left[\left\{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h)\, e^{-d/n} - 1\right\}^6\right]\right)^{5/6},
\end{aligned} \tag{A.15}$$

where the second inequality follows from the result that $\Pr(|X| \ge \varepsilon) \le \varepsilon^{-5}\mathrm{E}(|X|^5)$ for random variable $X$ and $\varepsilon > 0$, and the last inequality holds from Hölder's inequality. It is noted that the moment in the right-hand side of inequality $(A.15)$ exists from the assumption that $m-1-12 = n-p-11 > 0$. It follows that

$$\begin{aligned}
&\mathrm{E}\left[\left\{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h)\, e^{-d/n} - 1\right\}^6\right] \\
&= \mathrm{E}\left(\left[\left\{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h) - 1\right\} e^{-d/n} + (e^{-d/n} - 1)\right]^6\right) \\
&\le 2^5 \left(e^{-6d/n} \mathrm{E}\left[\left\{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h) - 1\right\}^6\right] + (e^{-d/n} - 1)^6\right),
\end{aligned}$$

where the inequality follows from Jensen's inequality observed as $\mathrm{E}[(X/2 + Y/2)^6] \le (1/2)\mathrm{E}(X^6) + (1/2)\mathrm{E}(Y^6)$ for random variables $X$ and $Y$. This implies that

$$\begin{aligned}
&\Pr(T_{j_1,j_2,d} \le 0) \\
&\le \frac{2^5}{(\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})^5} \left(e^{-6d/n} \mathrm{E}\left[\left\{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,(1+h) - 1\right\}^6\right] + (e^{-d/n} - 1)^6\right)^{5/6}.
\end{aligned} \tag{A.16}$$

It is noted that

$$e^{-d/n} - 1 = -e^{-\theta d/n} \frac{d}{n}, \quad 0 < {}^{\exists}\theta < d/n. \tag{A.17}$$

Thus we obtain

$$p^{12/5}|e^{-d/n}-1|^6 < p^{12/5}\left(\frac{d}{n}\right)^6 = \left(\frac{p}{n}\right)^{12/5}\left(\frac{5}{3}\frac{\log n^{3/5}}{n^{3/5}}\right)^6 \to 0. \qquad \text{(A.18)}$$

To evaluate the order of the moment in $(A.16)$ we propose a lemma, of which the proof is simple but tedious so we omit.

**Lemma A1.** *Suppose that $\chi_1^2(\cdot)$, $\chi_{m-1}^2$ and $\tau^2$ are the same definition as in $(A.3)$. Then,*

$$\begin{aligned}\mathrm{E}\left[\{(1-\rho_{j_1j_2\cdot(-\boldsymbol{j})}^2)(1+h)-1\}^6\right] &= \mathrm{E}\left[\left\{(1-\rho_{j_1j_2\cdot(-\boldsymbol{j})}^2)\left(1+\frac{\chi_1^2(\tau^2)}{\chi_{m-1}^2}\right)-1\right\}^6\right] \\ &= O(m^{-3}).\end{aligned}$$

Applying the result in $(A.18)$ and Lemma A1 to $(A.16)$, we obtain

$$\begin{aligned}\mathrm{P1} &= \sum_{(j_1,j_2)\in J_*}\Pr(T_{j_1,j_2,d}\le 0) \\ &\le \frac{2^5}{\min_{(j_1,j_2)\in J_*}(\rho_{j_1j_2\cdot(-\boldsymbol{j})}^2)^5}p^2\left[O(m^{-3})+o(p^{-12/5})\right]^{5/6},\end{aligned}$$

which converges to 0 under A1 and A2.

## A3. Proof of Theorem 2.2

In this subsection, we only sketch the outline of the proof of Theorem 2.2 since the proof uses the same descriptions as the one of Theorem 2.1. Set the triangular array $a_{n,p}$ as

$$a_{n,p} = \frac{\log n}{\log p} > 1 \quad ((n,p)\in\mathfrak{J}_2).$$

We show [F1] and [F2] in Section A3.1-A3.2.

### A3.1. Proof of [F2]

It is found that the descriptions from the beginning of Section A2.1 to $(A.9)$ hold. The right-hand side of the equality in $(A.9)$ is dominated by

$$\sqrt{\frac{2}{\pi}}\left(4\frac{n-p+1}{n-p}\log p\right)^{-1/2}\exp\left(-2\frac{\log p}{n-p}\right),$$

which converges to 0 under A2. This implies that

$$\Pr\left(\chi_1^2 > \frac{m-1}{n}\frac{d}{a_{n,p}}\right) = o(p^{-2}).$$

On the other hand, by Bernstein's inequality, we have

$$p^2 \Pr\left(\chi_{m-1}^2 < \frac{m-1}{a_{n,p}}\right)$$
$$\leq \exp\left(-\frac{m-1}{2}\left\{\frac{(1-a_{n,p}^{-1})^2}{W/(m-1)+(1-a_{n,p}^{-1})B} - 4\frac{\log p}{m-1}\right\}\right). \qquad \text{(A.19)}$$

Note that $0 < a_{n,p}^{-1} < 1$. Under A2, if $p/n \to 0$ and $(\log p)/\log n \to \ell \in [0,1)$ then $a_{n,p}^{-1} \to \ell$; thus the right-hand side of inequality $(A.19)$ converges to 0. Now we consider the case that $p/n \to c \in (0,1)$. Variate part of the exponential function in the right-hand side of inequality $(A.19)$ can be described as

$$-\frac{1}{8}\left(\frac{\log n^{1/2}}{n^{1/2}}\right)^{-2}\frac{m-1}{n}\left\{\frac{\left(\log\frac{p}{n}\right)^2}{W/(m-1)+(1-a_{n,p}^{-1})B}\right.$$
$$\left.-3^3\cdot 4\frac{n^{2/3}p^{1/3}}{m-1}\frac{\log p^{1/3}}{p^{1/3}}\left(\frac{\log n^{1/3}}{n^{1/3}}\right)^2\right\},$$

which diverges to $-\infty$ as $n, p \to \infty$ while $p/n \to c \in (0,1)$; thus we observe that the right-hand side of inequality $(A.19)$ converges to 0. From the inequality $(A.8)$ we observe that $\Pr(T_{j_1,j_2,d} > 0) = o(p^{-2})$ uniformly in $(j_1, j_2) \notin J_*$; thus [F2] holds.

### A3.2. Proof of [F1]

It is found that the descriptions from the beginning of Section A2.2 to $(A.16)$ hold. Using the same derivation as $(A.17)$ we observe that $(e^{-4(\log n)/(n-p)} - 1)^6 = o(p^{-12/5})$; thus [F1] holds.

# B Real Data Application

The Kaggle housing dataset, which was obtained at “https://www.kaggle.com/c/house-prices-advanced-regression-techniques”, consists of $p = 29$ observations for $n = 2274$ houses. Here, observations are, for example, “SalePrice”, “MS.SubClass”, and “Lot.Frontage”. We applied our model selection method $\widehat{J}_{4\{n/(n-p)\}\log n}$ to this dataset. Of $p(p-1)/2 = 406$ partial correlations, 50 partial correlations were observed to be nonzero. For example of chosen variables, “Garage.Cars and Garage.Area”, “Year.Built and Garage.Yr.Blt” and “Bedroom.AbvGr and TotRms.AbvGrd” are the set of top three in order of absolute value of the partial correlation coefficients. Table 3 is the list of chosen variables by $\hat{j}^{\,\mathrm{IB}}$. We also write the absolute value of partial correlation coefficients (Abs. of PCC) in this table.

Table 3: List for choosen set of variables by $\hat{j}^{\mathrm{IB}}$ for Kaggle housing data

| Variables | | Abs. of PCC |
|---|---|---|
| Garage.Cars | Garage.Area | 0.68 |
| Year.Built | Garage.Yr.Blt | 0.56 |
| Bedroom.AbvGr | TotRms.AbvGrd | 0.56 |
| Overall.Qual | SalePrice | 0.48 |
| Overall.Cond | Year.Remod.Add | 0.39 |
| Overall.Cond | Year.Built | 0.38 |
| MS.SubClass | Lot.Frontage | 0.36 |
| Lot.Frontage | Lot.Area | 0.30 |
| Kitchen.AbvGr | TotRms.AbvGrd | 0.29 |
| Full.Bath | Half.Bath | 0.28 |
| Garage.Yr.Blt | Garage.Area | 0.27 |
| MS.SubClass | Half.Bath | 0.26 |
| Year.Built | Enclosed.Porch | 0.24 |
| Mas.Vnr.Area | SalePrice | 0.24 |
| TotRms.AbvGrd | SalePrice | 0.24 |
| Year.Built | TotRms.AbvGrd | 0.23 |
| Bsmt.Full.Bath | Full.Bath | 0.23 |
| MS.SubClass | Kitchen.AbvGr | 0.23 |
| Bsmt.Full.Bath | SalePrice | 0.22 |
| Year.Built | Year.Remod.Add | 0.22 |
| Full.Bath | Bedroom.AbvGr | 0.22 |
| Half.Bath | TotRms.AbvGrd | 0.21 |
| Year.Remod.Add | Garage.Yr.Blt | 0.21 |
| MS.SubClass | Full.Bath | 0.18 |
| Bsmt.Full.Bath | Bsmt.Half.Bath | 0.17 |
| Year.Remod.Add | Full.Bath | 0.17 |
| Mo.Sold | Yr.Sold | 0.17 |
| Year.Built | Half.Bath | 0.17 |
| Overall.Qual | Year.Built | 0.17 |
| MS.SubClass | SalePrice | 0.16 |
| Misc.Val | SalePrice | 0.15 |
| Year.Remod.Add | Bedroom.AbvGr | 0.15 |
| Full.Bath | TotRms.AbvGrd | 0.15 |
| Year.Built | Garage.Area | 0.15 |
| Bsmt.Full.Bath | Half.Bath | 0.14 |
| Lot.Area | SalePrice | 0.14 |
| Year.Built | Bsmt.Full.Bath | 0.14 |
| Lot.Frontage | Garage.Area | 0.14 |
| Garage.Area | SalePrice | 0.14 |
| Full.Bath | SalePrice | 0.14 |
| Lot.Frontage | Pool.Area | 0.13 |
| Half.Bath | Bedroom.AbvGr | 0.13 |
| Half.Bath | Kitchen.AbvGr | 0.13 |
| MS.SubClass | Overall.Qual | 0.13 |
| Fireplaces | SalePrice | 0.13 |
| MS.SubClass | Bsmt.Full.Bath | 0.13 |
| Overall.Cond | Bedroom.AbvGr | 0.12 |
| Fireplaces | Screen.Porch | 0.12 |
| Overall.Cond | TotRms.AbvGrd | 0.12 |
| TotRms.AbvGrd | Misc.Val | 0.12 |

# References

[1] Bai, Z., Fujikoshi, Y. and Hu, J. (2018). Strong consistency of the AIC, BIC, Cp and KOO methods in high-dimensional multivariate linear regression. *Hiroshima Statistical Research Group*, TR; 18-09.

[2] Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencencies; Models, Analysis and Interpretation.* London, Chapman and Hall.

[3] Dempster, A. P. (1982). Covariance selection . *Biometrika*, **32**, 95-108.

[4] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with graphical lasso. *Biostatistics*, **9**, 432-441.

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

[6] Fujikoshi, Y. and Sakurai, T. (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**, 155–171.

[7] Fujikoshi, Y., Sakurai, T. and Yamada, T. (2022). High-Dimensional Consistencies of KOO Methods for Selecting Graphical Models. *Hiroshima Statistical Research Group*, TR; 22-06.

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

[9] Muirhead, R. J. (2009). *Aspects of Multivariate Statistical Theory.* John Wiley & Sons.

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

[11] 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. *Electronic Journal of Statistics*, **14**, 1386–1412.

[12] ODA, R. and YANAGIHARA, H. (2021). A consistent likelihood-based variable selection method in normal multivariate linear regression. In Intelligent Decision Technologies, 391-401, I. Czarnowski et al. (eds.)

[13] ODA, R., SUZUKI, Y., YANAGIHARA, H. and FUJIKOSHI, Y. (2020). A consistent variable selection method in high-dimensional canonical discriminant analysis. *Journal of Multivariate Analysis*, **175**, 104561.

[14] SAKURAI, T. and FUJIKOSHI, Y. (2020). Exploring consistencies of information criterion and test-based criterion for high-dimensional multivariate regression models under three covariance structures. In Festschrift in honor of Professr Dietrich von Rosen's 65th birthday (eds, T. Holgerson and M. Singnull). Springer.

[15] YANAGIHARA, H., WAKAKI, H. and FUJIKOSHI, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics*, **9**, 869–897.

[16] 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.