High-Dimensional Consistencies of KOO Methods for Selecting Graphical Models

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

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

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

****Department of Mathematical Sciences, Interdisciplinary Faculty of Science and Engineering, Shimane University, 1060 Nishikawatsu-cho, Matsue-shi, Shimane 690-8504, Japan*

Abstract

This paper considers a covariance selection problem model which estimates the set of nonzero partial correlations. First, we propose a knock-one-out (KOO) method based on a general information criterion. Next, two KOO methods based on two new model selection criteria are introduced. It is shown that our KOO methods have high-dimensional consistency under appropriate assumptions. The proposed model selection methods are examined for two real datasets. Some simulation results are also given.

AMS 2000 Subject Classification: primary 62H15; secondary 62H10

Key Words and Phrases: Covariance selection problem, High-dimensional consistency, General information criteria, KOO methods, New model selection criteria, Partial correlations

Abbreviated title: KOO Methods for Selecting Graphical Models

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 (Cox and Wermuth (1996), Yuan and Lin (2007)). 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 $\sigma_{j_1 j_2}$. Let $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})}$ be 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 such that it contains all pairs $\boldsymbol{j} = (j_1, j_2)$ satisfying $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0$. Suppose we are interested in finding the true model defined by

$$J_* = \{(j_1, j_2) \mid \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0, j_1, j_2 \in \{1, 2, \ldots, p\}, j_1 < j_2\}. \tag{1.1}$$

Then, 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). These candidate models are denoted by M_J or simply J, which is a subset of $\boldsymbol{\omega}$.

Let $\mathbf{S}$ be the sample covariance matrix based on a sample of size $n+1$ from a p-variate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, $n\mathbf{S}$ follows the Wishart distribution $\mathrm{W}_p(n, \boldsymbol{\Sigma})$. We use AIC and BIC to find the model which minimizes

$$GIC_{J,d} = -2 \log L(\widehat{\boldsymbol{\Sigma}}_J) + d g_J, \tag{1.2}$$

where $L(\widehat{\boldsymbol{\Sigma}}_J)$ 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.

However, these direct approaches will not be feasible when p is large, since the possible number of models becomes large. However, we can use a knock-one-out (KOO) method based on these model selection approaches. This idea goes back to Nishii et al. (1988) and Zhao et al. (1986). The term "KOO" was introduced by Bai et al. (2018). For a review of the KOO method, see, e.g., Fujikoshi (2022).

The KOO method is specifically as follows. Let $M_{\boldsymbol{\omega}}$ or $\boldsymbol{\omega}$ be a model such that all of the partial correlations are nonzero. Further, let $M_{\boldsymbol{\omega}\backslash \boldsymbol{j}}$ or $\boldsymbol{\omega}\backslash \boldsymbol{j}$ be a model such that the partial correlation $\rho_{j_1 j_2\cdot(-\boldsymbol{j})}$ is zero but the other partial correlations are nonzero. Let

$$T_{j_1 j_2,d} = GIC_{\boldsymbol{\omega}\backslash \boldsymbol{j},d} - GIC_{\boldsymbol{\omega},d}. \tag{1.3}$$

Then, our KOO method chooses the model given by

$$\widehat{J}_{G,d} = \{(j_1, j_2) \mid T_{j_1 j_2,d} > 0, 1 \le j_1 < j_2 \le p\}. \tag{1.4}$$

The selection procedure may be stated as follows: if $T_{j_1 j_2}$ is positive, (j_1, j_2) is selected, and if $T_{j_1 j_2}$ is not positive, (j_1, j_2) is not selected. In this paper, under a high-dimensional framework, we study consistency of $\widehat{J}_{G,d}$. Further, we introduce two new model selection criteria, DIC and ZIC. In addition to $\widehat{J}_{G,d}$, we consider the two other KOO methods $\widehat{J}_{D,d}$ and $\widehat{J}_{Z,d}$ based on model selection criteria DIC and ZIC.

For high-dimensional data such that $p > n$, Lasso and other regularization methods have been extended. In the case of the Gaussian concentration graph selection problem, see, e.g., Yuan and Lin (2007), Friedman et al. (2007), and Hirose et al. (2017).

The present paper is organized as follows. In Section 2, we give a distributional reduction for a key statistic $T_{j_1 j_2,d}$ or $L_{j_1 j_2} = T_{j_1 j_2,d} - d$. In Section 3, we present high-dimensional consistency of $\widehat{J}_{G,d}$. In Section 4, we propose the new two model selection criteria: DIC and ZIC. These are constructed based on the same idea as AIC and PEC (Fujikoshi et al. (2011)) by starting from the sets of partial correlations and their z-transformations. In Section 5, a KOO method based on DIC ($\widehat{J}_{D,d}$) and one based on ZIC ($\widehat{J}_{Z,d}$) are proposed. High-dimensional consistencies for $\widehat{J}_{D,d}$ and $\widehat{J}_{Z,d}$ are also shown. Simulation results are given in Section 6. Numerical examples are given in Section 7. In Section 8, we briefly discuss our selection criteria. Technical details are provided in Appendices.

2. Distribution of Key Statistics

The partial correlation $\rho_{j_1j_2\cdot(-\boldsymbol{j})}$ of X_{j_1} and X_{j_2} given $X_{(-\boldsymbol{j})}$ is defined as follows:

$$\rho_{j_1j_2\cdot(-\boldsymbol{j})} = \frac{\sigma_{j_1j_2\cdot(-\boldsymbol{j})}}{\sqrt{\sigma_{j_1j_1\cdot(-\boldsymbol{j})}}\sqrt{\sigma_{j_2j_2\cdot(-\boldsymbol{j})}}},$$

where

$$\begin{pmatrix} \sigma_{j_1j_1\cdot(-\boldsymbol{j})} & \sigma_{j_1j_2\cdot(-\boldsymbol{j})} \\ \sigma_{j_2j_1\cdot(-\boldsymbol{j})} & \sigma_{j_2j_2\cdot(-\boldsymbol{j})} \end{pmatrix} = \begin{pmatrix} \sigma_{j_1j_1} & \sigma_{j_1j_2} \\ \sigma_{j_2j_1} & \sigma_{j_2j_2} \end{pmatrix} - \boldsymbol{\sigma}_{j_1j_2\cdot(-\boldsymbol{j})}\boldsymbol{\Sigma}_{(-\boldsymbol{j})(-\boldsymbol{j})}^{-1}\boldsymbol{\sigma}'_{j_1j_2\cdot(-\boldsymbol{j})},$$

$\boldsymbol{\sigma}_{j_1j_2\cdot(-\boldsymbol{j})}$ 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 $\mathbf{S} = (s_{j_1j_2})$ 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 given as

$$r_{j_1j_2\cdot(-\boldsymbol{j})} = \frac{s_{j_1j_2\cdot(-\boldsymbol{j})}}{\sqrt{s_{j_1j_1\cdot(-\boldsymbol{j})}}\sqrt{s_{j_2j_2\cdot(-\boldsymbol{j})}}}. \tag{2.1}$$

It is well known that there is a close relationship between the partial correlation coefficients and the coefficients of $\boldsymbol{\Sigma}^{-1} = (\sigma^{j_1j_2})$, in fact that

$$\rho_{j_1j_2\cdot(-\boldsymbol{j})} = (-1)^{\delta_{j_1j_2}+1}\frac{\rho^{j_1j_2}}{\sqrt{\rho^{j_1j_1}}\sqrt{\rho^{j_2j_2}}}. \tag{2.2}$$

Here, $\delta_{j_1j_2}$ is the Kronecker delta. Thus, the zero of $\rho_{j_1j_2\cdot(-\boldsymbol{j})}$ is equivalent to the zero of the (j_1, j_2) component of $\boldsymbol{\Sigma}^{-1}$.

Here, we note that $T_{j_1 j_2,d}$ is related to the Likelihood Ratio Criterion (LRC) for the hypothesis $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = 0$. In fact, from (1.4) we can express it as

$$\begin{aligned} T_{j_1 j_2,d} &= -2 \log L(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega} \setminus \boldsymbol{j},d}) + d g_{\boldsymbol{\omega} \setminus \boldsymbol{j}} \\ &\qquad - \left\{ -2 \log L(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\omega},d}) + d g_{\boldsymbol{\omega}} \right\} \\ &= -2 \log \mathrm{LRT} - d, \end{aligned} \tag{2.3}$$

where LRT is a likelihood ratio statistic for testing the hypothesis $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})} = 0$. It should be noted that the LRC is based on the likelihood of $\mathbf{S}$. We now consider the term $L_{j_1 j_2} = -2 \log \mathrm{LRT}$. From Fujikoshi et al. (2010, Theorem 4.3.2), we have

$$L_{j_1 j_2} = -n \log(1 - r^2_{j_1 j_2 \cdot (-\boldsymbol{j})}). \tag{2.4}$$

Thus, our KOO method can be expressed as

$$-n \log(1 - r^2_{j_1 j_2 \cdot (-\boldsymbol{j})}) - d > 0 \;\Leftrightarrow\; (j_1, j_2) \in \widehat{J}_{G,d}. \tag{2.5}$$

Next, we consider the distribution of $L_{j_1 j_2}$. Using $r^2_{j_1 j_2 \cdot (-\boldsymbol{j})} = s^2_{j_1 j_2 \cdot (-\boldsymbol{j})} \cdot \{ s_{j_1 j_1 \cdot (-\boldsymbol{j})} s_{j_2 j_2 \cdot (-\boldsymbol{j})} \}^{-1}$, we can use the expression

$$L_{j_1 j_2} = n \log(1 + Q_{j_1 j_2 \cdot (-\boldsymbol{j})}), \tag{2.6}$$

where

$$Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{s^2_{j_1 j_2 \cdot (-\boldsymbol{j})}}{s_{j_1 j_1 \cdot (-\boldsymbol{j})} s_{j_2 j_2 \cdot (-\boldsymbol{j})} - s^2_{j_1 j_2 \cdot (-\boldsymbol{j})}}. \tag{2.7}$$

Thus, it is necessary to study the distribution of $Q_{j_1 j_2 \cdot (-\boldsymbol{j})}$ in order to obtain the distribution of $L_{j_1 j_2}$. For this purpose, we have the following theorem.

Theorem 2.1. *Let $Q_{j_1 j_2 \cdot (-\boldsymbol{j})}$ be the statistic defined by* (2.7)*. Then we can express it as*

$$Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \chi^2_1 \left(\tau^2 \right) \left\{ \chi^2_{m-1} \right\}^{-1}, \tag{2.8}$$

where $m = n - (p - 2)$, and $\tau^2 = \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})} (1 - \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})})^{-1} \chi^2_m$. Here, a noncentral chi-square variate $\chi^2_1(\cdot)$ and two chi-square variables χ^2_{m-1} and χ^2_m are mutually independent. If $(j_1, j_2) \notin J_*$, then we can write $Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \chi^2_1 \left\{ \chi^2_{m-1} \right\}^{-1}$.

Proof of Theorem 2.1. 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 $\boldsymbol{\Sigma}_{j_1 j_2 \cdot (-\boldsymbol{j})}$. We can express $Q_{j_1 j_2 \cdot (-\boldsymbol{j})}$ as

$$Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{w^2_{j_1 j_2 \cdot (-\boldsymbol{j})}}{w_{j_1 j_1 \cdot (-\boldsymbol{j})} w_{j_2 j_2 \cdot (-\boldsymbol{j})} - w^2_{j_1 j_2 \cdot (-\boldsymbol{j})}}, \tag{2.9}$$

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, 1982). Then, we can write $Q_{j_1 j_2 \cdot (-\boldsymbol{j})}$ in (2.9) as follows:

$$Q_{j_1 j_2 \cdot (-\boldsymbol{j})} = \frac{\boldsymbol{u}_2' \frac{1}{\boldsymbol{u}_1' \boldsymbol{u}_1} \boldsymbol{u}_1 \boldsymbol{u}_1' \boldsymbol{u}_2}{\boldsymbol{u}_2' \left(\mathbf{I}_m - \frac{1}{\boldsymbol{u}_1' \boldsymbol{u}_1} \boldsymbol{u}_1 \boldsymbol{u}_1' \right) \boldsymbol{u}_2}. \tag{2.10}$$

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

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

is independent of $\boldsymbol{u}_1$. 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 τ^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 2.1. □

3. Consistency of KOO Method

3.1. Outline of Proof

In this section, we show the high-dimensional consistency of the KOO method $\widehat{J}_d$ in (1.4). Our consistency will be obtained by showing the following two properties:

$$[\mathrm{F1}]: \ \mathrm{P1} \equiv \sum_{(j_1, j_2) \in J_*} \Pr(T_{j_1 j_2, d} \le 0) \to 0. \tag{3.1}$$

$$[\mathrm{F2}]: \ \mathrm{P2} \equiv \sum_{(j_1, j_2) \notin \boldsymbol{j}_*} \Pr(T_{j_1 j_2, d} > 0) \to 0. \tag{3.2}$$

Here, P1 denotes the sum of probabilities that a partial correlation is identified as zero despite not being zero and P2 denotes the sum of probabilities that a partial correlation is identified as nonzero despite being zero. Conditions [F1] and [F2] are sufficient to show the consistency, which can be seen from the following inequality:

$$
\begin{aligned}
\Pr(\widehat{J}_d = J_*) &= \Pr\left(\left(\bigcap_{(j_1,j_2)\in J_*} \text{“}T_{j_1 j_2,d} > 0\text{”}\right) \cap \left(\bigcap_{(j_1,j_2)\notin J_*} \text{“}T_{j_1 j_2,d} \leq 0\text{”}\right)\right) \\
&= 1 - \Pr\left(\left(\bigcup_{(j_1,j_2)\in J_*} \text{“}T_{j_1 j_2,d} \leq 0\text{”}\right) \cup \left(\bigcup_{(j_1,j_2)\notin J_*} \text{“}T_{j_1 j_2,d} > 0\text{”}\right)\right) \\
&\geq 1 - \sum_{(j_1,j_2)\in J_*} \Pr(T_{j_1,j_2,d} \leq 0) - \sum_{(j_1,j_2)\notin J_*} \Pr(T_{j_1,j_2,d} > 0). \quad (3.3)
\end{aligned}
$$

If [F1] and [F2] hold, then $\Pr(\widehat{J}_d = J_*)$ converges to 1, i.e. variable selection method $\widehat{J}_d$ has consistency. This approach has been used in Fujikoshi and Sakurai (2019), Oda and Yanagihara (2021), and Fujikoshi (2022), as well as other studies.

We make the following assumptions:

- A1: The high-dimensional asymptotic framework: the sample size n and the dimensionality p diverge together under the restriction that $p/n \to c_1 \in (0, 1)$.
- A2: The true subset J_* is included in the full set $\boldsymbol{\Omega}$, i.e., $J_* \subset \boldsymbol{\Omega}$, and the size of J_*, i.e., $\#J_*$, does not depend on the dimensionality p. That is, $\#J_*$ is finite on p.
- A3: There exist positive constants $\overline{c}, \underline{c} \in (0, 1)$ so that whenever $\boldsymbol{j} = (j_1, j_2) \in J_*$, $\lim_{p\to\infty} \max_{j_1 j_2 \in J_*} \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})} = \overline{c}$ and $\lim_{p\to\infty} \min_{j_1 j_2 \in J_*} \rho^2_{j_1 j_2 \cdot (-\boldsymbol{j})} = \underline{c}$.
- A4: The threshold is set as $d = n^{\delta}$, $1/4 < \delta < 1$.

Assumption A1 requires that c_1 be larger than 0 and smaller than 1, but when p is finite or very small, we need to consider the case $c_1 = 0$. However, this case is not considered here. In our proof, we use the condition $n - p > 17$, but this will be satisfied under A1. Assumption A2 means that the number of nonzero partial correlations is fixed instead of growing with p. From a practical point of view, this case will be important since it makes interpretation simple. Further, it can be expected that our model selection shall be quite accurate in our target situation that only a few partial correlations, relative to the total number of variables, are significant. Related to Assumption A2, we assume that the limits of nonzero partial correlations are not 0 or 1.

Note that under A2, P1 is a finite sum, whereas P2 is an infinite sum. These properties will be used in our proofs of asymptotic consistency.

3.2. Proof of [F2]

When $(j_1, j_2) \notin J_*$, from Theorem 2.1 we can write $T_{j_1 j_2, d} = n \log\left(1 + \chi_1^2/\chi_{m-1}^2\right) - d$, and therefore we have

$$\Pr(T_{j_1 j_2, d} > 0) = \Pr\left(n \log\left(1 + \frac{\chi_1^2}{\chi_{m-1}^2}\right) - d > 0\right).$$

It is observed that

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

From this result, by letting $U = \chi_1^2/\chi_{m-1}^2$, we have

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

Further, the following inequalities hold.

$$\begin{aligned}\Pr(U > e^{d/n} - 1) &\leq \Pr(|U| > d/n) \\ &\leq (d/n)^{-2\ell} \mathrm{E}(U^{2\ell}), \quad \ell \in \{1, 2, \ldots\},\end{aligned}$$

where the first inequality follows from the fact that $e^{d/n} - 1 > d/n > 0$, and the second inequality is derived as follows: $^{\forall} h > 0$,

$$\begin{aligned}\mathrm{E}(U^{2\ell}) &= \int u^{2\ell} f(u) du \\ &\geq \int_{|u| \geq h} u^{2\ell} f(u) du \\ &\geq h^{2\ell} \int_{|u| \geq h} f(u) du = h^{2\ell} \Pr(|U| \geq h).\end{aligned}$$

Above, $f(\cdot)$ is the probability density function of U. Now, we find that

$$\mathrm{E}(U^{2\ell}) = \mathrm{E}[(\chi_1^2/\chi_{m-1}^2)^{2\ell}] = O(m^{-2\ell}) = O(n^{-2\ell}).$$

It can be deduced from the assumption $d = n^\delta$ that $(d/n)^{-2\ell}\mathrm{E}(U^{2\ell}) = O(n^{-2\ell\delta})$. From these results, by letting $\ell = 4$, we obtain that

$$\Pr(T_{j_1,j_2,d} > 0) = \Pr(U > e^{d/n} - 1) < (d/n)^{-8}\mathrm{E}(U^8) = O(n^{-8\delta}).$$

Consequently,

$$\mathrm{P2} = \sum_{(j_1,j_2)\notin J_*} \Pr(T_{j_1,j_2,d} > 0) < p^2 O(n^{-8\delta}) = O(n^{2-8\delta}) \to 0$$

when $1/4 < \delta$. Related to setting $\ell = 4$, it is necessary that $m = n - p > 17$ for the moment $\mathrm{E}(U^{2\ell}) = \mathrm{E}(U^8)$ to exist. However, as mentioned previously, this condition is satisfied asymptotically under A1.

3.3. Proof of [F1]

When $(j_1, j_2) \in J_*$, P1 is a finite sum. Looking term-wise, we shall see the following result:

$$\Pr\left(n\log\left(1 + A_{j_1j_2\cdot(-\boldsymbol{j})}\right) - d \le 0\right) \to 0,$$

where

$$A_{j_1j_2\cdot(-\boldsymbol{j})} = \chi_1^2\left(\frac{\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}{1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}\chi_m^2\right)\left\{\chi_{m-1}^2\right\}^{-1}. \tag{3.4}$$

Then we can write

$$\begin{aligned} 1 + A_{j_1j_2\cdot(-\boldsymbol{j})} &= 1 + \frac{\boldsymbol{u}_2'\frac{1}{\boldsymbol{u}_1'\boldsymbol{u}_1}\boldsymbol{u}_1\boldsymbol{u}_1'\boldsymbol{u}_2}{\boldsymbol{u}_2'\left(\mathbf{I}_m - \frac{1}{\boldsymbol{u}_1'\boldsymbol{u}_1}\boldsymbol{u}_1\boldsymbol{u}_1'\right)\boldsymbol{u}_2} \\ &= \frac{\boldsymbol{u}_2'\boldsymbol{u}_2}{\boldsymbol{u}_2'\left(\mathbf{I}_m - \frac{1}{\boldsymbol{u}_1'\boldsymbol{u}_1}\boldsymbol{u}_1\boldsymbol{u}_1'\right)\boldsymbol{u}_2}. \end{aligned}$$

Further, the numerator and denominator of the last expression are as follows:

$$\begin{aligned}
&\boldsymbol{u}_2'\boldsymbol{u}_2 \sim \chi_m^2 \\
&\boldsymbol{u}_2'\left(\mathbf{I}_m - \frac{1}{\boldsymbol{u}_1'\boldsymbol{u}_1}\boldsymbol{u}_1\boldsymbol{u}_1'\right)\boldsymbol{u}_2 \sim (1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\chi^2_{m-1}.
\end{aligned}$$

These imply

$$\frac{\boldsymbol{u}_2'\boldsymbol{u}_2}{\boldsymbol{u}_2'\left(\mathbf{I}_m - \frac{1}{\boldsymbol{u}_1'\boldsymbol{u}_1}\boldsymbol{u}_1\boldsymbol{u}_1'\right)\boldsymbol{u}_2} \xrightarrow{p} \frac{1}{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})}.$$

From the assumption $d = n^\delta$, $1/4 < \delta < 1$, we find that $d/n \to 0$ as $n \to \infty$, and so

$$\begin{aligned}
&\frac{1}{n}\{n\log\left(1 + A_{j_1j_2\cdot(-\boldsymbol{j})}\right) - d\} \\
&\quad \xrightarrow{p} \log\frac{1}{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})} = -\log(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}) > 0. \qquad (3.5)
\end{aligned}$$

Probability convergence in (3.5) implies that the probability that $n\log\left(1 + A_{j_1j_2\cdot(-\boldsymbol{j})}\right) - d$ is negative or equals zero approaches zero.

4. Derivation of Model Selection Criteria DIC and ZIC

In the previous section, we considered a KOO criterion based on the GIC criterion for selecting the set of nonzero partial correlations. This section gives two model selection criteria in the predictive sense. We name the first one as DIC and the second one as ZIC. The respective derivations are given in separate subsections. These criteria were developed based on an idea similar to AIC and Cp, but our basic statistics are the partial correlations themselves and the distances between the basic statistics are based on a Frobenius norm. Thus, these criteria also correspond to PEC (Fujikoshi et al. (2011)) in multivariate regression models.

4.1. Derivation of DIC

Let $\mathfrak{R}$ be the $p \times p$ matrix of population partial correlations, i.e., whose (j_1, j_2) entry is $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})}$. In addition, let $\mathbf{R}$ be the matrix of sample partial correlations which corresponds to $\mathfrak{R}$. We measure the goodness of fit between $\mathbf{R}$ and $\mathfrak{R}$ by the Frobenius norm given by

$$D(\mathbf{R}, \mathfrak{R}) = \frac{1}{2}\mathrm{tr}(\mathbf{R} - \mathfrak{R})^2 = \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} (r_{j_1 j_2 \cdot (-\boldsymbol{j})} - \rho_{j_1 j_2 \cdot - \boldsymbol{j}})^2.$$

Let M_J be the model corresponding to $J = \{(j_1, j_2) \mid \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0,\ j_1, j_2 \in \{1, 2, \ldots, p\},\ j_1 < j_2\}$. Then, we consider the minimum distance estimator under M_J such that

$$\min_{\mathfrak{R} \in M_J} D(\mathbf{R}, \mathfrak{R}) = D(\mathbf{R}, \widehat{\mathfrak{R}}_{M_J}).$$

Noting that

$$\begin{aligned}\min_{\mathfrak{R} \in M_J} D(\mathbf{R}, \mathfrak{R}) &= \min_{\mathfrak{R} \in M_J} \left\{ \sum_{(j_1, j_2) \in J} (r_{j_1 j_2 \cdot (-\boldsymbol{j})} - \rho_{j_1 j_2 \cdot (-\boldsymbol{j})})^2 + \sum_{(j_1, j_2) \in J^c} r^2_{j_1 j_2 \cdot (-\boldsymbol{j})} \right\} \\ &= \sum_{(j_1, j_2) \in J^c} r^2_{j_1 j_2 \cdot (-\boldsymbol{j})},\end{aligned}$$

we can see that $\widehat{\mathfrak{R}}_{M_J} = \mathbf{A} = (a_{j_1 j_2})$, where

$$a_{j_1 j_2} = \begin{cases} r_{j_1 j_2 \cdot (-\boldsymbol{j})}, & (j_1, j_2) \in J, \\ 0, & (j_1, j_2) \notin J. \end{cases}$$

As a criterion for choosing a model M_J based on the point of predictive inference, we consider

$$\mathrm{Risk}_{M_J} = \mathrm{E}_z^* \mathrm{E}_x^* [D(\mathbf{R}_z, \widehat{\mathfrak{R}}_{M_J})]. \tag{4.1}$$

along to AIC and Cp as in Fujikoshi and Satoh (1997). In (4.1), z denotes the variate for future data, and $\mathbf{R}_z$ is a copy of $\mathbf{R}$, i.e., has the same distribution as $\mathbf{R}$ but is independent of $\mathbf{R}$. Further, E^* denotes the expectation with respect to the true model M^*. Our construction method of a model selection criterion is similar to that of AIC, but we start from $\mathbf{R}$, not the sample covariance matrix $\mathbf{S}$.

Now we propose a model selection criterion by considering an estimator for Risk_{M_J} given in (4.1). Consider a naive estimator $D(\mathbf{R}, \widehat{\mathfrak{R}}_{M_J})$ which is obtained from $D(\mathbf{R}_z, \widehat{\mathfrak{R}}_{M_J})$ by replacing $\mathbf{R}_z$, by $\mathbf{R}$, and consider further modifying it. More precisely, we can write Risk_{M_J} as

$$\mathrm{Risk}_{M_J} = \mathrm{E}_z^* \mathrm{E}_x^* [D(\mathbf{R}, \widehat{\mathfrak{R}}_{M_J})] + B_{M_J}.$$

and consider estimating B_{M_J}, where

$$B_{M_J} = \mathrm{E}_z^* \mathrm{E}_x^* [D(\mathbf{R}_z, \widehat{\mathfrak{R}}_{M_J}) - D(\mathbf{R}, \widehat{\mathfrak{R}}_{M_J})].$$

From Appendix A1,

$$B_{M_J} = \frac{k_J}{n-p+1} - \sum_{(j_1,j_2)\in J\cap J_*} \frac{(2-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}{n-p+1} + O(k_J n^{-2}),$$

where k_J is the number of elements in candidate model M_J. Now assume that the true model is included in a candidate model. Then, $J_* \cap J = J_*$. For any model M_J including M_{J_*},

$$B_2 = \sum_{(j_1,j_2)\in J\cap J_*} \frac{(2-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}{n-p+1}$$

takes a definite value. Neglecting B_2, which is not effected by a change in models, and the remainder term, which is $O(k_J n^{-2})$, we can make the approximation $B_{M_J} \approx k_J/(n-p+1)$. Therefore, we propose the model selection criterion

$$\begin{aligned}\mathrm{DIC}_{M_J} &= D(\mathbf{R}, \widehat{\mathfrak{R}}_{M_J}) + \frac{k_J}{n-p+1} \\ &= \sum_{(j_1,j_2)\in J^c} r^2_{j_1j_2\cdot(-\boldsymbol{j})} + \frac{k_J}{n-p+1}. \end{aligned} \tag{4.2}$$

4.2. Derivation of ZIC

In this section, we use the Fisher's z-transforms instead of the sample partial correlations. Let $\zeta_{j_1 j_2}$ and $z_{j_1 j_2}$ be the Fisher's z-transforms of the population partial correlation $\rho_{j_1 j_2 \cdot (-\boldsymbol{j})}$ and the sample partial correlation $r_{j_1 j_2 \cdot (-\boldsymbol{j})}$, i.e.,

$$\zeta_{j_1, j_2} = \frac{1}{2} \log \frac{1 + \rho_{j_1 j_2 \cdot (-\boldsymbol{j})}}{1 - \rho_{j_1 j_2 \cdot (-\boldsymbol{j})}}, \quad z_{j_1, j_2} = \frac{1}{2} \log \frac{1 + r_{j_1 j_2 \cdot (-\boldsymbol{j})}}{1 - r_{j_1 j_2 \cdot (-\boldsymbol{j})}}.$$

Further, let us denote $\mathfrak{Z} = (\zeta_{j_1 j_2})$ and $\mathbf{Z} = (z_{j_1 j_2})$. We measure the goodness of fit between $\mathfrak{Z}$ and $\mathbf{Z}$ by the Frobenius norm given by

$$\begin{aligned} D(\mathbf{Z}, \mathfrak{Z}) &= \frac{1}{2} \mathrm{tr}(\mathbf{Z} - \mathfrak{Z})^2 \\ &= \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} (z_{j_1 j_2} - \zeta_{j_1 j_2})^2 \\ &= \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \left[\frac{1}{2} \log \frac{1 + r_{j_1 j_2 \cdot (-\boldsymbol{j})}}{1 - r_{j_1 j_2 \cdot (-\boldsymbol{j})}} - \frac{1}{2} \log \frac{1 + \rho_{j_1 j_2 \cdot (-\boldsymbol{j})}}{1 - \rho_{j_1 j_2 \cdot (-\boldsymbol{j})}} \right]^2 . \end{aligned}$$

Let M_J be the model corresponding to $J = \{(j_1, j_2) \mid \rho_{j_1 j_2 \cdot (-\boldsymbol{j})} \neq 0,\ j_1, j_2 \in \{1, 2, \ldots, p\},\ j_1 < j_2\}$. As an estimator under M_J, we consider the minimum distance estimator

$$\min_{\mathfrak{Z} \in M_J} D(\mathbf{Z}, \mathfrak{Z}) = D(\mathbf{Z}, \widehat{\mathfrak{Z}}_{M_J}).$$

Here, the estimator takes the following form:

$$\begin{aligned} \min_{\mathfrak{Z} \in M_J} D(\mathbf{Z}, \mathfrak{Z}) &= \min_{\mathfrak{Z} \in M_J} \left\{ \sum_{(j_1, j_2) \in J} (z_{j_1 j_2} - \zeta_{j_1 j_2})^2 + \sum_{(j_1, j_2) \in J^c} z_{j_1 j_2}^2 \right\} \\ &= \sum_{(j_1, j_2) \in J^c} z_{j_1 j_2}^2 . \end{aligned}$$

Letting $\widehat{\mathfrak{Z}}_{M_J} = \mathbf{B} = (b_{j_1 j_2})$, we have

$$b_{j_1 j_2} = \begin{cases} z_{j_1 j_2} = \dfrac{1}{2} \log \dfrac{1 + r_{j_1 j_2 \cdot (-\boldsymbol{j})}}{1 - r_{j_1 j_2 \cdot (-\boldsymbol{j})}}, & (j_1, j_2) \in J, \\ 0, & (j_1, j_2) \notin J. \end{cases}$$

By the same consideration as in Section 4.1., we measure the goodness of M_J from the point of predictive inference. More specifically, we consider

$$\text{Risk}_{M_J} = \text{E}_z^* \text{E}_x^* [D(\mathbf{Z}_z, \widehat{\mathfrak{Z}}_{M_J})]. \tag{4.3}$$

Here, we use the same notation as in (4.1).

We propose a model selection criterion by considering an estimator for Risk_{M_J} given in (4.3). Consider the naive estimator $D(\mathbf{Z}, \widehat{\mathfrak{Z}}_{M_J})$ which is obtained by replacing $\mathbf{Z}_z$, by $\mathbf{Z}$, and consider further modifying it. More precisely, we write Risk_{M_J} as

$$\text{Risk}_{M_J} = \text{E}_z^* \text{E}_x^* [D(\mathbf{Z}, \widehat{\mathfrak{Z}}_{M_J})] + B_{M_J},$$

and consider estimating B_{M_J}. Here,

$$B_{M_J} = \text{E}_z^* \text{E}_x^* [D(\mathbf{Z}_z, \widehat{\mathfrak{Z}}_{M_J}) - D(\mathbf{Z}, \widehat{\mathfrak{Z}}_{M_J})].$$

From Appendix A2,

$$B_{M_J} = \frac{k_J}{n - p + 1} + O(k_J n^{-2}).$$

It follows that we can approximate the bias term B_{M_J} by $k_J/(n - p + 1)$, omitting the terms of $O(k_J n^{-2})$. Based on this approximation, we propose the following model selection criterion:

$$\begin{aligned} \text{ZIC}_{M_J} &= D(\mathbf{Z}, \widehat{\mathbf{Z}}_{M_J}) + \frac{k_J}{n - p + 1} \\ &= \sum_{(j_1, j_2) \in J^c} \left(\frac{1}{2} \log \frac{1 + r_{j_1 j_2 \cdot (-\boldsymbol{j})}}{1 - r_{j_1 j_2 \cdot (-\boldsymbol{j})}} \right)^2 + \frac{k_J}{n - p + 1}. \end{aligned} \tag{4.4}$$

5. Consistency of KOO Methods based on DIC and ZIC

Define two generalization criteria for DIC and ZIC, which include the threshold term $d = n^{\delta}$, as follows:

$$\mathrm{DIC}_{J,d} = \sum_{(j_1,j_2)\in J^c} r^2_{j_1j_2\cdot(-\boldsymbol{j})} + \frac{dk_J}{m},$$
$$\mathrm{ZIC}_{J,d} = \sum_{(j_1,j_2)\in J^c} z^2_{j_1j_2\cdot(-\boldsymbol{j})} + \frac{dk_J}{m},$$

where $m = n-p$. By letting $d = 1$ and neglecting the term of $o(m^{-1})$, $\mathrm{DIC}_{J,d}$ and $\mathrm{ZIC}_{J,d}$ coincide with DIC_{M_J} and ZIC_{M_J}, respectively. Let the statistics $U_{j_1j_2,d}$ and $V_{j_1j_2,d}$ be defined as follows:

$$\begin{aligned} U_{j_1j_2,d} &= \mathrm{DIC}_{\boldsymbol{\omega}\backslash\boldsymbol{j},d} - \mathrm{DIC}_{\boldsymbol{\omega},d} \\ &= r^2_{j_1j_2\cdot(-\boldsymbol{j})} - \frac{d}{m}, \\ V_{j_1j_2,d} &= \mathrm{ZIC}_{\boldsymbol{\omega}\backslash\boldsymbol{j},d} - \mathrm{ZIC}_{\boldsymbol{\omega},d} \\ &= z^2_{j_1j_2\cdot(-\boldsymbol{j})} - \frac{d}{m}. \end{aligned}$$

Then, our KOO methods choose the model by

$$\widehat{J}_{D,d} = \{(j_1, j_2) \mid U_{j_1j_2,d} > 0, 1 \le j_1 < j_2 \le p\},$$
$$\widehat{J}_{Z,d} = \{(j_1, j_2) \mid V_{j_1j_2,d} > 0, 1 \le j_1 < j_2 \le p\}.$$

Thus, we find that

$$(n-p)r^2_{j_1j_2\cdot(-\boldsymbol{j})} - d > 0 \ \Leftrightarrow\ (j_1, j_2) \in \widehat{J}_{D,d}, \tag{5.1}$$
$$(n-p)\left(\frac{1}{2}\log\frac{1+r_{j_1j_2\cdot(-\boldsymbol{j})}}{1-r_{j_1j_2\cdot(-\boldsymbol{j})}}\right)^2 - d > 0 \ \Leftrightarrow\ (j_1, j_2) \in \widehat{J}_{Z,d}. \tag{5.2}$$

We show the high-dimensional consistencies of KOO methods $\widehat{J}_{D,d}$ and $\widehat{J}_{Z,d}$ by using the same derivation as in Section 3.1.. Since the proofs for $\widehat{J}_{D,d}$ and $\widehat{J}_{Z,d}$ are quite similar, we only give that for $\widehat{J}_{Z,d}$. High-dimensional consistency for $\widehat{J}_{Z,d}$ holds if the following two properties are satisfied.

$$[\mathrm{F3}]:\ \mathrm{P3} \equiv \sum_{(j_1,j_2)\in J_*} \Pr(V_{j_1j_2,d} \le 0) \to 0.$$
$$[\mathrm{F4}]:\ \mathrm{P4} \equiv \sum_{(j_1,j_2)\notin \boldsymbol{j}_*} \Pr(V_{j_1j_2,d} > 0) \to 0.$$

5.1. **Proof of** [F4]

The following equivalences hold:

$$\begin{aligned}
&\infty > z^2_{j_1j_2} > d/m \\
&\iff 1 > r^2_{j_1j_2\cdot(-\boldsymbol{j})} > \tanh^2(\sqrt{d/m}) \\
&\iff \frac{1}{1-\tanh^2(\sqrt{d/m})} < \frac{1}{1-r^2_{j_1j_2\cdot(-\boldsymbol{j})}} < \infty,
\end{aligned}$$

where $\tanh^2(x) = (\tanh(x))^2$ and $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. When $(j_1, j_2) \notin J_*$, we can write $(1 - r^2_{j_1j_2\cdot(-\boldsymbol{j})})^{-1} = 1 + \chi^2_1/\chi^2_{m-1}$ from Theorem 2.1, and therefore we have

$$\begin{aligned}
\Pr\left(z^2_{j1j_2} > d/m\right) &= \Pr\left(\frac{1}{1-\tanh^2(\sqrt{d/m})} < \frac{1}{1-r^2_{j_1j_2\cdot(-\boldsymbol{j})}}\right) \\
&= \Pr\left(\frac{1}{1-\tanh^2(\sqrt{d/m})} < 1 + \frac{\chi^2_1}{\chi^2_{m-1}}\right) \\
&= \Pr\left(\frac{\chi^2_1}{\chi^2_{m-1}} > \frac{\tanh^2(\sqrt{d/m})}{1-\tanh^2(\sqrt{d/m})}\right).
\end{aligned}$$

Using the same derivation as in Section 3.2., we have

$$\Pr\left(z^2_{j1j_2} > d/m\right) < \left\{\frac{\tanh^2(\sqrt{d/m})}{1-\tanh^2(\sqrt{d/m})}\right\}^{-2\ell} \mathrm{E}(U^{2\ell}),$$

where $U = \chi^2_1/\chi^2_{m-1}$. Note that $\tanh(x)/x \to 1$ as $x \to 0$. If $0 < \delta < 1$, then $d/m = O(n^{\delta-1}) \to 0$ as $n \to \infty$, and so $\tanh(\sqrt{d/m})/\sqrt{d/m} \to 1$. This implies that

$$\left\{\frac{\tanh^2(\sqrt{d/m})}{1-\tanh^2(\sqrt{d/m})}\right\}^{-2\ell} = O((d/m)^{-2\ell}) = O(n^{-2\ell(\delta-1)}).$$

Recalling that $\mathrm{E}(U^{2\ell}) = O(n^{-2\ell})$, we find that

$$\Pr\left(z^2_{j1j_2} > d/m\right) < \left\{\frac{\tanh^2(\sqrt{d/m})}{1-\tanh^2(\sqrt{d/m})}\right\}^{-2\ell} \mathrm{E}(U^{2\ell}) = O(n^{-2\ell\delta}).$$

Therefore, by letting $\ell = 4$,

$$\text{P4} = \sum_{(j_1,j_2)\notin J_*} \Pr(V_{j_1,j_2,d} > 0) \le p^2 O(n^{-8\delta}) = O(n^{2-8\delta}) \to 0$$

for the case $1/4 < \delta$.

5.2. Proof of [F3]

When $(j_1, j_2) \in J_*$, P3 is a finite sum, and so we shall check the convergence term-wise as follows:

$$\Pr\left(z_{j1j_2}^2 - d/m \le 0\right) \to 0. \tag{5.3}$$

From Section 5.1., we find that $z_{j1j_2}^2 - d/m \le 0$ is equivalent to $\{1 - \tanh^2(\sqrt{d/m})\}^{-1} \ge (1 - r^2_{j_1 j_2\cdot(-\boldsymbol{j})})^{-1}$, and so the probability in (5.3) is equal to

$$\Pr\left(\frac{1}{1 - r^2_{j_1 j_2\cdot(-\boldsymbol{j})}} - \frac{1}{1 - \tanh^2(\sqrt{d/m})} \le 0\right).$$

By virtue of Theorem 2.1, the convergence (5.3) holds if

$$\Pr\left(1 + A_{j_1 j_2\cdot(-\boldsymbol{j})} - \frac{1}{1 - \tanh^2(\sqrt{d/m})} \le 0\right) \to 0,$$

where $A_{j_1 j_2\cdot(-\boldsymbol{j})}$ is defined by (3.4). From Section 3.3., we have that

$$1 + A_{j_1 j_2\cdot(-\boldsymbol{j})} \xrightarrow{p} \frac{1}{1 - \rho^2_{j_1 j_2\cdot(-\boldsymbol{j})}} > 1.$$

From the assumption $d = n^\delta$, $1/4 < \delta < 1$, we obtain that $(1 - \tanh^2(\sqrt{d/m}))^{-1} \to 1$ as $n \to \infty$, and thus

$$1 + A_{j_1 j_2\cdot(-\boldsymbol{j})} - \frac{1}{1 - \tanh^2(\sqrt{d/m})} \xrightarrow{p} \frac{\rho^2_{j_1 j_2\cdot(-\boldsymbol{j})}}{1 - \rho^2_{j_1 j_2\cdot(-\boldsymbol{j})}} > 0. \tag{5.4}$$

Consequently, the probability convergence in (5.4) implies that the probability that $1 + A_{j_1 j_2\cdot(-\boldsymbol{j})} - \{1 - \tanh^2(\sqrt{d/m})\}^{-1}$ is negative or equals zero approaches zero.

6. Simulation Results

In this section, we look at the actual performance of our methods with regards to finding the sets of nonzero partial correlations. The methods are given by (2.5), (5.1), and (5.2). We have shown that each of these methods is asymptotically consistent within a high-dimensional asymptotic framework (see Section 3 and Section 5).

Our simulation dataset has been constructed as follows. Suppose that $\boldsymbol{Z}$ is distributed as $\mathrm{N}_p(\mathbf{0}, \mathbf{\Sigma})$. The covariance matrix is set to be

$$\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_k, \mathbf{I}_{p-3k}), \quad \mathbf{\Sigma}_i = \begin{pmatrix} 1 & a & b \\ a & a^2+1 & ab \\ b & ab & b^2+1 \end{pmatrix}. \tag{6.1}$$

Then, the partial correlation matrix of $\mathbf{\Sigma}_i$ is obtained as

$$\begin{pmatrix} 1 & \rho_{12\cdot 3} & \rho_{13\cdot 2} \\ \rho_{12\cdot 3} & 1 & 0 \\ \rho_{13\cdot 2} & 0 & 1 \end{pmatrix},$$

where

$$\rho_{12\cdot 3} = \frac{a}{\sqrt{a^2+b^2+1}}, \quad \rho_{13\cdot 2} = \frac{b}{\sqrt{a^2+b^2+1}},$$

and the partial correlation $\rho_{23\cdot 1}$ is zero. Such structure is, for example, given by the following relation:

$$\begin{cases} Z_1 = e_1, & e_1 \sim \mathrm{N}(0,1), \\ Z_2 = aZ_1 + e_2, & e_2 \sim \mathrm{N}(0,1), \quad e_2 \perp e_1, \quad e_2 \perp e_3, \\ Z_3 = bZ_1 + e_3, & e_3 \sim \mathrm{N}(0,1), \quad e_3 \perp e_1, \quad e_3 \perp e_2. \end{cases}$$

In general, it is known that

$$\rho_{ij\cdot\mathrm{rest}} = 0 \Longleftrightarrow \rho^{ij} = 0 \tag{6.2}$$

(see, e.g., Fujikoshi et al. (2010), p.84, (4.3.7)). Therefore, the number of nonzero partial correlations is equal to the number of nonzero off-diagonal elements of $\mathbf{\Sigma}^{-1} = \mathrm{diag}(\mathbf{\Sigma}_1^{-1}, \ldots, \mathbf{\Sigma}_k^{-1}, \mathbf{I}_{p-3k})$. Note that

$$\mathbf{\Sigma}_i^{-1} = \begin{pmatrix} a^2+b^2+1 & -a & -b \\ -a & 1 & 0 \\ -b & 0 & 1 \end{pmatrix}.$$

Therefore, the number of nonzero off-diagonal elements is equal to $4k$. In this case, it is assumed that all components of $\boldsymbol{Z} = (Z_1, \ldots, Z_p)'$ are independent with mean 0 and variance 1, and $\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$.

We carried out the simulation with 10^4 repetitions. In accordance with assumption A4, the threshold is set to $d = n^{\delta}$, where δ is set as $\delta = 1/2, 1/3, 1/4$. Note that $\delta \in \{1/2, 1/3\}$ satisfies assumption A4, whereas $\delta = 1/4$ does not. We referred to Fujikoshi (2022) for the candidates of δ. Keeping the simulation settings simple, we set $k = 1$ and $a = b$, so that all nonzero partial correlations are equal and are $\tilde{\rho} \in \{0.3, 0.5, 0.696\}$. The settings of the other parameter are as follows: $n \in \{100, 200, 500, 1000, 10000\}$, $p_1 = 6$, and $p_2 = p - 6 \in \{10, 30, 60\}$. The results of the simulation are shown in Table 1, from which we made the following observations:

- The proportion of the true model, i.e., the proportion of selection for sets of all nonzero partial correlations, approaches 1 as the sample size increases for each of the variable selection criteria based on KOO, ((2.5), (5.1), and (5.2)).
  - In order to get a good estimator for the set of nonzero partial correlations, it seems that a very large sample size n is required, in comparison to the dimensionality p.
  - The smaller p_2 is, the faster the proportion converges to 1.
  - The closer the nonzero partial correlation is to 1, the faster the proportion converges to 1.
- For the case that the proportion is less than 1, the proportion of the true model for each of (5.1) and (5.2) is larger than the one for (2.5). This can be checked, for example, for the case in which $\tilde{\rho} = 0.696$, $n = 200$, $p_2 = 60$.
- In our simulation setting, the accuracy in selecting the true model is good for the case in which $\delta = 1/2$ among the three settings of δ.

- In our simulation setting, consistency seems not to hold numerically for the case in which $\delta = 1/4$.

Table 1: Proportion of selecting all nonzero partial correlations

$\tilde{\rho} = 0.3$			$d = n^{1/2}$			$d = n^{1/3}$			$d = n^{1/4}$		
n	p_1	p_2	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$
100	6	10	0.01	0.00	0.00	0.07	0.07	0.07	0.02	0.03	0.03
200	6	10	0.13	0.08	0.11	0.42	0.47	0.46	0.11	0.14	0.13
500	6	10	0.91	0.89	0.91	0.81	0.83	0.83	0.31	0.34	0.33
1000	6	10	1.00	1.00	1.00	0.94	0.95	0.94	0.51	0.52	0.52
10000	6	10	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.95	0.94
100	6	30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	6	30	0.10	0.03	0.04	0.00	0.00	0.00	0.00	0.00	0.00
500	6	30	0.90	0.84	0.86	0.07	0.14	0.13	0.00	0.00	0.00
1000	6	30	1.00	1.00	1.00	0.46	0.52	0.51	0.00	0.00	0.00
10000	6	30	1.00	1.00	1.00	1.00	1.00	1.00	0.51	0.52	0.52
100	6	60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	6	60	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
500	6	60	0.87	0.75	0.79	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	60	1.00	1.00	1.00	0.02	0.07	0.07	0.00	0.00	0.00
10000	6	60	1.00	1.00	1.00	0.99	0.99	0.99	0.06	0.06	0.06

$\tilde{\rho} = 0.5$			$d = n^{1/2}$			$d = n^{1/3}$			$d = n^{1/4}$		
n	p_1	p_2	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$
100	6	10	0.83	0.82	0.83	0.26	0.39	0.35	0.05	0.10	0.09
200	6	10	0.99	1.00	0.99	0.53	0.60	0.58	0.14	0.19	0.18
500	6	10	1.00	1.00	1.00	0.82	0.84	0.83	0.35	0.38	0.37
1000	6	10	1.00	1.00	1.00	0.94	0.95	0.95	0.53	0.54	0.54
10000	6	10	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.94	0.94
100	6	30	0.05	0.36	0.33	0.00	0.00	0.00	0.00	0.00	0.00
200	6	30	0.80	0.95	0.92	0.00	0.00	0.00	0.00	0.00	0.00
500	6	30	1.00	1.00	1.00	0.08	0.15	0.13	0.00	0.00	0.00
1000	6	30	1.00	1.00	1.00	0.46	0.53	0.52	0.00	0.00	0.00
10000	6	30	1.00	1.00	1.00	1.00	1.00	1.00	0.52	0.53	0.53
100	6	60	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
200	6	60	0.08	0.80	0.71	0.00	0.00	0.00	0.00	0.00	0.00
500	6	60	0.98	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	60	1.00	1.00	1.00	0.02	0.07	0.07	0.00	0.00	0.00
10000	6	60	1.00	1.00	1.00	0.99	0.99	0.99	0.06	0.07	0.07

$\tilde{\rho} = 0.696$			$d = n^{1/2}$			$d = n^{1/3}$			$d = n^{1/4}$		
n	p_1	p_2	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$	$\widehat{J}_{G,d}$	$\widehat{J}_{D,d}$	$\widehat{J}_{Z,d}$
100	6	10	0.92	0.97	0.95	0.34	0.45	0.42	0.09	0.16	0.14
200	6	10	0.99	0.99	0.99	0.59	0.65	0.63	0.20	0.24	0.23
500	6	10	1.00	1.00	1.00	0.85	0.86	0.86	0.42	0.44	0.43
1000	6	10	1.00	1.00	1.00	0.95	0.95	0.95	0.58	0.60	0.59
10000	6	10	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.95	0.95
100	6	30	0.07	0.69	0.53	0.00	0.00	0.00	0.00	0.00	0.00
200	6	30	0.81	0.95	0.93	0.00	0.01	0.00	0.00	0.00	0.00
500	6	30	1.00	1.00	1.00	0.09	0.16	0.15	0.00	0.00	0.00
1000	6	30	1.00	1.00	1.00	0.47	0.54	0.53	0.00	0.00	0.00
10000	6	30	1.00	1.00	1.00	1.00	1.00	1.00	0.52	0.53	0.53
100	6	60	0.00	0.38	0.12	0.00	0.00	0.00	0.00	0.00	0.00
200	6	60	0.08	0.83	0.73	0.00	0.00	0.00	0.00	0.00	0.00
500	6	60	0.98	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
1000	6	60	1.00	1.00	1.00	0.03	0.08	0.07	0.00	0.00	0.00
10000	6	60	1.00	1.00	1.00	0.99	0.99	0.99	0.07	0.07	0.07

7. Applications

This section presents two real examples, and we apply our variable selection based on the KOO method. The first example regards the scores on an examination covering five subjects and comes from Mardia et al. (1979). The second example regards Kaggle housing data, which are available at "https://www.kaggle.com/c/house-prices-advanced-regression-techniques".

7.1. Scores on five subjects

Consider the scores on an examination covering five subjects for $n = 88$ students given in Mardia et al. (1979). The five scores are X_1 (mechanics), X_2 (vectors), X_3 (algebra), X_4 (analysis), and X_5 (statistics). The partial correlation coefficients between X_i and X_j given the remaining three variables are given in Table 2. From these coefficients, one may conjecture the following

Table 2: Partial correlation matrix of five subjects

X_1	$-$				
X_2	0.33	$-$			
X_3	0.23	0.28	$-$		
X_4	0.00	0.08	0.43	$-$	
X_5	0.02	0.02	0.36	0.25	$-$

model:

$$\text{M}: \rho_{14\cdot(-\{2,3,5\})} = \rho_{24\cdot(-\{1,3,5\})} = \rho_{15\cdot(-\{2,3,4\})} = \rho_{25\cdot(-\{1,3,4\})} = 0, \qquad (7.1)$$

which is equivalent to

$$X_1 \perp X_4 \mid (X_2, X_3, X_5), \qquad X_1 \perp X_5 \mid (X_2, X_3, X_4),$$
$$X_2 \perp X_4 \mid (X_1, X_3, X_5), \qquad X_2 \perp X_5 \mid (X_1, X_3, X_4).$$

Here, the notation $X_i \perp X_j \mid (X_{k_1}, X_{k_2}, X_{k_3})$ indicates the conditional independence of X_i and X_j under the condition that $(X_{k_1}, X_{k_2}, X_{k_3})$ is given.

Based on the simulation results, the methods based on GIC, DIC, and ZIC have almost the same precision for selecting the true model; thus, we use (5.1) for this dataset. For comparison, we perform a hypothesis test for the null hypothesis that $H_0 : \rho_{j_1 j_2\cdot(-\boldsymbol{j})} = 0$. The likelihood ratio statistic is based on

$$T = \frac{r_{j_1 j_2\cdot(-\boldsymbol{j})}\sqrt{n-p-2}}{\sqrt{1 - r^2_{j_1 j_2\cdot(-\boldsymbol{j})}}}.$$

Under null hypothesis H_0, T follows a t distribution with $n-p-2$ degrees of freedom, and so H_0 is rejected at significance level α if the observed value of $|T|$ is larger than the upper $\alpha/2$ percentile of the t distribution. We computed statistics for T and $U_{j_1 j_2, d}$ based on the values in Table 2 to construct Table 3. The p-values for testing H_0 are also shown. The δ columns give the values of $U_{j_1 j_2, d}$ for $d = n^{\delta}/(n-p)$. Values in bold are non-negative statistics with p-values less than 0.05. We observed that the result of variable selection with $\delta = 1/3$ matches the ones of testing H_0 with significance level 0.05.

Table 3: Variable selection and hypothesis testing results for scores on five subjects

	$r_{j_1 j_2 \cdot (-\boldsymbol{j})}$	T	p-value	$\delta = 1/2$	$\delta = 1/3$
X_3-X_4	0.43	4.313	**0.000**	**0.071**	**0.131**
X_3-X_5	0.36	3.494	**0.001**	**0.016**	**0.076**
X_1-X_2	0.33	3.166	**0.002**	-0.005	**0.055**
X_2-X_3	0.28	2.641	**0.010**	-0.035	**0.024**
X_4-X_5	0.25	2.338	**0.022**	-0.051	**0.008**
X_2-X_4	0.08	0.727	0.469	-0.107	-0.048
X_1-X_5	0.02	0.181	0.857	-0.113	-0.054
X_1-X_5	0.02	0.181	0.857	-0.113	-0.054
X_1-X_4	0.00	0.000	1.000	-0.114	-0.054

7.2. Kaggle housing data

The Kaggle housing dataset, which was obtained at "https://www.kaggle.com/c/house-prices-advanced-regression-techniques", consists of $p = 37$ observations for $n = 2930$ houses. Here, observations are, for example, "SalePrice", "MS.SubClass", and "Lot.Frontage". We applied our model selection method $\widehat{J}_{D,d}$ to this dataset. Of $p(p-1)/2 = 666$ partial correlations, 246 partial correlations were observed to be nonzero when $\delta = 1/2$ and 345 partial correlations were observed to be nonzero when $\delta = 1/3$.

8. Discussion

In this study, we considered the Gaussian concentration graph selection problem, i.e., the problem for selecting partial correlation under normality. We proposed a knock-one-out (KOO) method based on a general information criterion (GIC). In addition, we proposed two alternative KOO methods based on two distance criteria (DIC, ZIC). Consistency was shown for each of the KOO methods. Based on a simulation study, the KOO methods in this paper have good precision for selecting the true model.

In this paper, we do not consider how to select the δ in the threshold $d = n^{\delta}$. It is important to determine δ so that the true model shall be selected. It is also important to study high-dimensional consistency properties of the model selection criteria when the size of the true model is large. More precisely, we want to examine such problems when

$$p/n \to c_1 \in (0,1), \quad \#J_* = O(p).$$

These are left as future problems.

Appendix A1. Reduction of B_{M_J} in Section 4.1

The bias term B_{M_J} can be expressed as follows:

$$\begin{aligned}
B_{M_J} &= \mathrm{E}_z^*\mathrm{E}_x^*[D(\mathbf{R}_z, \widehat{\mathfrak{R}}_{M_J}) - D(\mathbf{R}, \widehat{\mathbf{R}}_{M_J})] \\
&= \mathrm{E}_z^*\mathrm{E}_x^*\left[\frac{1}{2}\mathrm{tr}(\mathbf{R}_z - \widehat{\mathfrak{R}}_{M_J})^2 - \frac{1}{2}\mathrm{tr}(\mathbf{R} - \widehat{\mathfrak{R}}_{M_J})^2\right] \\
&= \mathrm{E}_z^*\mathrm{E}_x^*\left[\frac{1}{2}\mathrm{tr}\mathbf{R}_z^2 - \mathrm{tr}\mathbf{R}_z\widehat{\mathfrak{R}}_{M_J} - \frac{1}{2}\mathrm{tr}\mathbf{R}^2 + \mathrm{tr}\mathbf{R}\widehat{\mathfrak{R}}_{M_J}\right] \\
&= \mathrm{E}_z^*\mathrm{E}_x^*\left[-\mathrm{tr}\mathbf{R}_z\widehat{\mathfrak{R}}_{M_J} + \mathrm{tr}\mathbf{R}\widehat{\mathfrak{R}}_{M_J}\right] \\
&= \mathrm{E}_x^*\left[\mathrm{tr}\widehat{\mathfrak{R}}_{M_J}(\mathbf{R} - \boldsymbol{\Psi}^*)\right].
\end{aligned}$$

Here, $\boldsymbol{\Psi}^* = (\psi_{j_1j_2}^*)$, where

$$\begin{aligned}
\psi_{j_1j_2}^* &= \mathrm{E}^*(r_{j_1j_2\cdot(-\boldsymbol{j})}) \\
&= \frac{2}{n-p+1}\left\{\frac{\Gamma(\frac{n-p+2}{2})}{\Gamma(\frac{n-p+1}{2})}\right\}^2 \rho_{j_1j_2\cdot(-\boldsymbol{j})} \\
&\quad \times {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; \frac{1}{2}(n-p+1)+1; \rho_{j_1j_2\cdot(-\boldsymbol{j})}^2\right).
\end{aligned}$$

Note that $\psi_{j_1j_2}^* = 0$ for $(j_1, j_2) \notin J_*$. In general, it is known that the distributional results on partial correlations are obtained from the ones on ordinal correlations by transforming n to $n-p+1$. For the above expression for

$\psi^*_{j_1j_2}$, see, for example, Muirhead (1982). Therefore, we have

$$\mathrm{tr}\widehat{\boldsymbol{\mathfrak{R}}}_{M_J}(\mathbf{R}-\boldsymbol{\Psi}^*) = \sum_{(j_1,j_2)\in J\cap J_*} r_{j_1j_2\cdot(-\boldsymbol{j})}(r_{j_1j_2\cdot(-\boldsymbol{j})} - \psi^*_{j_1j_2}) + \sum_{(j_1,j_2)\in J\cap J_*^c} r^2_{j_1j_2\cdot(-\boldsymbol{j})}.$$

Then, we have

$$\begin{aligned}
&\mathrm{E}^*_x\left[\mathrm{tr}\widehat{\boldsymbol{\mathfrak{R}}}_{M_J}(\mathbf{R}-\boldsymbol{\Psi}^*)\right] \\
&= \sum_{(j_1,j_2)\in J\cap J_*}\left[\mathrm{E}^*_x(r^2_{j_1j_2\cdot(-\boldsymbol{j})}) - (\psi^*_{j_1j_2})^2\right] + \sum_{(j_1,j_2)\in J\cap J_*^c}\mathrm{E}^*_x(r^2_{j_1j_2\cdot(-\boldsymbol{j})}) \\
&= \sum_{(j_1,j_2)\in J\cap J_*}\left[\mathrm{V}^*_x(r_{j_1j_2\cdot(-\boldsymbol{j})})\right] + \sum_{(j_1,j_2)\in J\cap J_*^c}\mathrm{E}^*_x(r^2_{j_1j_2\cdot(-\boldsymbol{j})}).
\end{aligned}$$

Note (see, e.g., Muirhead (1982)) that for $(j_1, j_2) \in J_*$,

$$\mathrm{E}(r^2_{j_1j_2\cdot(-\boldsymbol{j})}) = 1 - \frac{n-p}{n-p+1}(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\,{}_2F_1\left(1,1;\frac{1}{2}(n-p+1)+1;\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}\right).$$

These imply that when $n, p \to \infty, p/n \to c_1 \in (0,1)$,

$$\begin{aligned}
\mathrm{V}^*_x(r_{j_1j_2\cdot(-\boldsymbol{j})}) &= \frac{(1-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})^2}{n-p+1} + O(n^{-2}) \\
&= \frac{1-(2-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}{n-p+1} + O(n^{-2}).
\end{aligned}$$

On the other hand, when $(j_1, j_2) \notin J_*$, $\rho_{j_1j_2\cdot(-\boldsymbol{j})} = 0$ and hence

$$\mathrm{E}(r^2_{j_1j_2\cdot(-\boldsymbol{j})}) = \frac{1}{n-p+1}.$$

From the above, we have

$$\begin{aligned}
&\mathrm{E}^*_x\left[\mathrm{tr}\widehat{\boldsymbol{\mathfrak{R}}}_{M_J}(\mathbf{R}-\boldsymbol{\Psi}^*)\right] \\
&= \sum_{(j_1,j_2)\in J\cap J_*}\left[\mathrm{V}^*_x(r_{j_1j_2\cdot(-\boldsymbol{j})})\right] + \sum_{(j_1,j_2)\in J\cap J_*^c}\mathrm{E}^*_x(r^2_{j_1j_2\cdot(-\boldsymbol{j})}) \\
&= \frac{k_J}{n-p+1} - \sum_{(j_1,j_2)\in J\cap J_*}\frac{(2-\rho^2_{j_1j_2\cdot(-\boldsymbol{j})})\rho^2_{j_1j_2\cdot(-\boldsymbol{j})}}{n-p+1} + O(k_J n^{-2}),
\end{aligned}$$

where k_J is the number of elements in a candidate model M_J.

Appendix A2. Reduction of B_{M_J} in Section 4.2

The bias term B_{M_J} can be expressed as follows;

$$\begin{aligned} B_{M_J} &= \mathrm{E}_z^*\mathrm{E}_x^*[D(\mathbf{Z}_z, \widehat{\boldsymbol{\mathfrak{Z}}}_{M_J}) - D(\mathbf{Z}, \widehat{\boldsymbol{\mathfrak{Z}}}_{M_J})] \\ &= \mathrm{E}_z^*\mathrm{E}_x^*\left[\frac{1}{2}\mathrm{tr}(\mathbf{Z}_z - \widehat{\boldsymbol{\mathfrak{Z}}}_{M_J})^2 - \frac{1}{2}\mathrm{tr}(\mathbf{Z} - \widehat{\boldsymbol{\mathfrak{Z}}}_{M_J})^2\right] \\ &= \mathrm{E}_z^*\mathrm{E}_x^*\left[\frac{1}{2}\mathrm{tr}\mathbf{Z}_z^2 - \mathrm{tr}\mathbf{Z}_z\widehat{\boldsymbol{\mathfrak{Z}}}_{M_J} - \frac{1}{2}\mathrm{tr}\mathbf{Z}^2 + \mathrm{tr}\mathbf{Z}\widehat{\mathbf{Z}}_{M_J}\right] \\ &= \mathrm{E}_z^*\mathrm{E}_x^*\left[-\mathrm{tr}\mathbf{Z}_z\widehat{\boldsymbol{\mathfrak{Z}}}_{M_J} + \mathrm{tr}\mathbf{Z}\widehat{\boldsymbol{\mathfrak{Z}}}_{M_J}\right] \\ &= \mathrm{E}_x^*\left[\mathrm{tr}\widehat{\boldsymbol{\mathfrak{Z}}}_{M_J}(\mathbf{Z} - \boldsymbol{\mathfrak{M}}^*)\right]. \end{aligned}$$

Here, $\boldsymbol{\mathfrak{M}}^* = (\mu_{j_1j_2}^*)$, where

$$\mu_{j_1j_2}^* = \mathrm{E}_x^*(z_{j_1j_2}) = \frac{1}{2}\mathrm{E}_x^*\left[\log(1 + r_{j_1j_2\cdot(-\boldsymbol{j})}) - \log(1 - r_{j_1j_2\cdot(-\boldsymbol{j})})\right].$$

Therefore, we have

$$\mathrm{tr}\widehat{\boldsymbol{\mathfrak{Z}}}_{M_J}(\mathbf{Z} - \boldsymbol{\mathfrak{M}}^*) = \sum_{(j_1,j_2)\in J} z_{j_1j_2}(z_{j_1j_2} - \mu_{j_1j_2}^*).$$

It follows that

$$\begin{aligned} &\mathrm{E}_x^*\left[\mathrm{tr}\widehat{\boldsymbol{\mathfrak{Z}}}_{M_J}(\mathbf{Z} - \boldsymbol{\mathfrak{M}}^*)\right] \\ &= \sum_{(j_1,j_2)\in J}\left[\mathrm{E}_x^*(z_{j_1j_2}^2) - (\mu_{j_1j_2}^*)^2\right] \\ &= \sum_{(j_1,j_2)\in J}\mathrm{V}_x^*(z_{j_1j_2}). \end{aligned}$$

From Hotelling (1953), under assumption A1, we find that

$$\begin{aligned} \mu_{j_1j_2} = \mathrm{E}(z_{j_1j_2}) &= \zeta_{j_1j_2} + \frac{\rho_{j_1j_2\cdot(-\boldsymbol{j})}}{2(n-p+1)} + O(n^{-2}), \\ \mathrm{E}[(z_{j_1j_2} - \zeta_{j_1j_2})^2] &= \frac{1}{n-p+1} + \frac{8 - \rho_{j_1j_2\cdot(-\boldsymbol{j})}^2}{4(n-p+1)^2} + O(n^{-3}). \end{aligned}$$

These imply that

$$\begin{aligned}\mathrm{V}_x^*(z_{j_1j_2}) &= \mathrm{E}_x^*[(z_{j_1j_2} - \zeta_{j_1j_2}^*)^2] - (\mu_{j_1j_2}^* - \zeta_{j_1j_2}^*)^2 \\ &= \frac{1}{n-p+1} + \frac{4 - (\rho_{j_1j_2\cdot(-\boldsymbol{j})}^*)^2}{2(n-p+1)^2} + O(n^{-3}),\end{aligned}$$

and so it holds that

$$\mathrm{E}_x^*\left[\mathrm{tr}\widehat{\mathbf{\mathcal{Z}}}_{M_J}(\mathbf{Z} - \mathbf{\mathcal{M}}^*)\right] = \sum_{(j_1,j_2)\in J} \mathrm{V}_x^*(z_{j_1j_2}) = \frac{k_J}{n-p+1} + O(k_J n^{-2}),$$

where k_J is the number of elements in candidate model M_J.

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. (2007). 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 YANAGIHARA, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *Journal of Multivariate Analysis*, **123**, 184–200.

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

[9] FUJIKOSHI, Y., KAN, T. TAKAHASHI, S. and SAKURAI,T. (2011). Prediction error citerion for selecting variables in a linear regression model. *Ann. Inst. Statist. Math.*, **63**, 387-403.

[10] HIROSE, K., FUJISAWA, H. and SESE, J. (2017). *Journal of Multivariate Analysis*, **161**, 172–190.

[11] HOTELLING, H. (1953). New light on the correlation coefficient and its transforms (with discussion). *Journal of the Royal Statistical Society B*, **15**, 193–232.

[12] LENG, C. and TANG, C. Y. (2012). Sparse matrix graphical models. *J. Amer. Statist. Assoc.* , **107**, 1187–1200.

[13] MARDIA, K. V., KENT, J.T. and BIBBY, J. M. (1979), *Multivariate Analysis*. Academic Press, New York.

[14] 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.

[15] 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.)

[16] 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.

[17] Sakurai, T., Nakata, T. and Fujikoshi, Y. (2013). High-dimensional AICs for selection of variables in discriminant analysis. *Sankhya, Ser. A*, **75**, 139-170.

[18] 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. Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *J. Multiv. Anal.*, **107**, 119-140.

[19] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.

[20] 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.