A high-dimensional bias-corrected AIC for selecting response variables in multivariate calibration

Ryoya Oda*, Yoshie Mima†,
Hirokazu Yanagihara and Yasunori Fujikoshi

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

(Last Modified: December 21, 2018)

Abstract

In a multivariate linear regression with a p-dimensional response vector $\boldsymbol{y}$ and a q-dimensional explanatory vector $\boldsymbol{x}$, we consider a multivariate calibration problem requiring the estimation of an unknown explanatory vector $\boldsymbol{x}_0$ corresponding to a response vector $\boldsymbol{y}_0$, based on $\boldsymbol{y}_0$ and n-samples of $\boldsymbol{x}$ and $\boldsymbol{y}$. We propose a high-dimensional bias-corrected Akaike's information criterion (HAIC$_\text{C}$) for selecting response variables. To correct the bias between a risk function and its estimator, we use a hybrid-high-dimensional asymptotic framework such that n tends to ∞ but p/n does not exceed 1. Through numerical experiments, we verify that the HAIC$_\text{C}$ performs better than a formal AIC.

1 Introduction

Multivariate linear regression is very widely used in various fields. Let $\boldsymbol{y} = (y_1, \ldots, y_p)'$ and $\boldsymbol{x} = (x_1, \ldots, x_q)'$ be a p-dimensional response vector and a q-dimensional nonstochastic explanatory vector, respectively. A multivariate linear regression is written as

$$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\mathcal{B}}'\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\alpha}$ is a p-dimensional unknown vector of intercept coefficients, $\boldsymbol{\mathcal{B}}$ is a $q \times p$ unknown matrix of regression coefficients, and $\boldsymbol{\varepsilon}$ is a p-dimensional error vector. We assume that $\boldsymbol{\varepsilon}$ is distributed according to the p-dimensional normal distribution with a mean vector $\mathbf{0}_p$, which is a p-dimensional vector of zeros, and a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. Further, assume that there are n observations $(\boldsymbol{y}_1, \boldsymbol{x}_1), \ldots, (\boldsymbol{y}_n, \boldsymbol{x}_n)$, which are expressed as $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$: $n \times p$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$: $n \times q$ in matrix notation, and $\boldsymbol{X}$ is centralized ($\boldsymbol{X}'\mathbf{1}_n = \mathbf{0}_q$), where $\mathbf{1}_n$ is the n-dimensional vector of ones. Our model selection criterion is defined for the case that $\text{rank}(\boldsymbol{X}) = q \leq p$ and $n - p - q - 2 > 0$.

*Corresponding author. Email: oda.stat@gmail.com

†Current address: Institute of educational foundation Fukuyama Akenohoshi, 3-4-1 Nishifukatsucho, Fukuyama-shi, Hiroshima 721-8545, Japan.

In actual empirical contexts, calibration is often needed and is widely applied (e.g., Alves and Poppi, 2016; Bro, 2003; Carvalho *et al.*, 2015). A typical calibration problem can be described as follows. Suppose that a value $\boldsymbol{y}_0$ of $\boldsymbol{y}$ in (1) has been observed, but the corresponding value $\boldsymbol{x}_0$ of $\boldsymbol{x}$ is unknown. Then the problem is to make an inference about an unknown vector $\boldsymbol{x}_0$, based on $\boldsymbol{y}_0$, $\boldsymbol{Y}$, and $\boldsymbol{X}$. Brown (1982) summarized various aspects of this problem. There is considerable literature on point estimators of $\boldsymbol{x}_0$ and confidence regions for $\boldsymbol{x}_0$ in the controlled calibration model. Another calibration problem involves removing redundant response variables in estimating $\boldsymbol{x}_0$. Generally, estimating a reduced model on a subset of the available data leads to an inferior fit. This will be the case when all the parameters are known. However, it is understood that when some response variables are redundant, we can expect better prediction by neglecting these redundant response variables. In this vein, two approaches are salient: one based on test procedures for a redundancy hypothesis and the other based on model selection procedures for selecting response variables. Brown (1982) proposed a procedure based on a test for redundancy of response variables. In the context of model selection, Fujikoshi and Nishii (1986) proposed a formal Akaike's information criterion (AIC) for the redundancy of response variables in estimating $\boldsymbol{x}_0$.

In the AIC, the goodness of fit of a model is measured by the Kullback-Leibler (KL) discrepancy function. The best model is defined as that where the risk function defined by the expected KL loss function is lowest among all the redundancy models. The AIC is defined by adding an estimator of the bias between the risk function and the expectation of a negative twofold maximum log-likelihood to the negative twofold maximum log-likelihood, and the AIC is regarded as an estimator of the risk function. Following Akaike (1973), Fujikoshi and Nishii (1986) formally regarded the twofold number of parameters in the redundancy model as an estimator of bias. However, it is known that when the sample size n is small or the number of redundancy models is large, a formal AIC does not estimate the risk function well and tends to choose a model such that many response variables are not redundant. Therefore, it is important to correct the bias in order to better estimate the risk function. For selecting explanatory variables in an ordinary linear regression, Bedrick and Tsai (1994), Hurvich and Tsai (1989), and Sugiura (1978) corrected the bias for overspecified models. However, these corrections did not consider the selection of response variables in multivariate calibration.

In recent years, it has become commonplace to analyze high-dimensional data such that not only the sample size n but also the dimension p are large. Bedrick and Tsai (1994), Hurvich and Tsai (1989) and Sugiura (1978) used a large sample (LS) asymptotic framework such that n only tends to ∞ in order to correct bias. However, corrections using the LS asymptotic framework become inadequate when the response dimension p is large. It is known that corrections using the high-dimensional (HD) asymptotic framework such that both n and p tend to ∞ are better than those using the LS asymptotic framework. Fujikoshi *et al.* (2014) offered a bias correction using the HD asymptotic framework in the context of selecting explanatory variables in a multivariate linear regression. It can be expected that such a correction performs well when both n and p are large but is inadequate when p is small. However, it can be difficult for analysts to decide whether p is large or small; thus, when p is moderate, it can be burdensome to determine whether corrected AICs by the LS or HD asymptotic framework should be applied.

In this paper, we correct the bias under a hybrid-high-dimensional (HHD) asymptotic framework:

$$n \to \infty, \ \frac{p}{n} \to c \in [0, 1).$$

It should be noted that the LS and HD asymptotic frameworks are included in the HHD asymptotic framework as special cases. Thus, it is expected that the bias-corrected AIC determined by using the HHD asymptotic framework is a better estimator of the risk function regardless of the size of p. We refer to this bias-corrected AIC as the high-dimensional bias-corrected AIC ($\mathrm{HAIC_C}$).

The remainder of the paper is organized as follows. In section 2, we introduce a framework for selecting response variables in multivariate calibration. In section 3, the $\mathrm{HAIC_C}$ is proposed and we obtain an asymptotic property of the $\mathrm{HAIC_C}$. In section 4, we explore, and verify, the performance of the $\mathrm{HAIC_C}$ through conducting numerical simulations. Technical details are relegated to the Appendix.

2 A framework for selecting response variables

2.1 A redundancy hypothesis of response variables

Let $\boldsymbol{Z} = (\boldsymbol{1}_n, \boldsymbol{X})$ and $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\mathcal{B}}')'$ be $n \times (q+1)$ and $(q+1) \times p$ matrices. Then, the multivariate linear regression for $\boldsymbol{Y}$ and $\boldsymbol{X}$ is written as

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\Theta} + \boldsymbol{\mathcal{E}}, \tag{2}$$

where $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ is the $n \times p$ error matrix and $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are mutually independent and identically distributed according to $N_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$ assuming that $\boldsymbol{\Sigma}$ is positive definite. To introduce the redundancy hypothesis of response variables by Fujikoshi and Nishii (1986), we prepare a classical estimator of $\boldsymbol{x}_0$. If the population parameters $\boldsymbol{\alpha}$, $\boldsymbol{\mathcal{B}}$, and $\boldsymbol{\Sigma}$ are known, then the classical estimator $\widehat{\boldsymbol{x}}_0$ is written as

$$\widehat{\boldsymbol{x}}_0 = \arg\min_{\boldsymbol{x}_0}\{(\boldsymbol{y}_0 - \boldsymbol{\alpha} - \boldsymbol{\mathcal{B}}'\boldsymbol{x}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_0 - \boldsymbol{\alpha} - \boldsymbol{\mathcal{B}}'\boldsymbol{x}_0)\} = (\boldsymbol{\mathcal{B}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{B}}')^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_0 - \boldsymbol{\alpha}). \tag{3}$$

We partition $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2')'$. Without loss of generality, let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be the p_1-dimensional and $(p-p_1)$-dimensional vectors of the candidate redundant variables and non-redundant variables, respectively. To define the redundancy hypothesis, we consider only the case $q \le p_1 < p$. We express the partitions of $\boldsymbol{y}_0$, $\boldsymbol{\alpha}$, $\boldsymbol{\mathcal{B}}$, and $\boldsymbol{\Sigma}$ corresponding to the division of $\boldsymbol{y} = (\boldsymbol{y}_1', \boldsymbol{y}_2')'$ as follows:

$$\boldsymbol{y}_0 = (\boldsymbol{y}_{0,1}', \boldsymbol{y}_{0,2}')', \ \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')', \ \boldsymbol{\mathcal{B}} = (\boldsymbol{\mathcal{B}}_1, \boldsymbol{\mathcal{B}}_2), \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \tag{4}$$

where $\boldsymbol{y}_{0,1}$ and $\boldsymbol{y}_{0,2}$ are p_1-dimensional and $(p-p_1)$-dimensional vectors, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are p_1-dimensional and $(p-p_1)$-dimensional vectors, $\boldsymbol{\mathcal{B}}_1$ and $\boldsymbol{\mathcal{B}}_2$ are $q \times p_1$ and $q \times (p-p_1)$ matrices, $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are the $p_1 \times p_1$ and $(p-p_1) \times (p-p_1)$ covariance matrices of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, and $\boldsymbol{\Sigma}_{12}$ is the $p_1 \times (p-p_1)$ covariance matrix of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Let $\boldsymbol{C} = (\boldsymbol{\mathcal{B}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{B}}')^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{\Sigma}^{-1}$. We also partition the classical estimator $\widehat{\boldsymbol{x}}_0$ in (3) as

$$\widehat{\boldsymbol{x}}_0 = \boldsymbol{C}(\boldsymbol{y}_0 - \boldsymbol{\alpha}) = \boldsymbol{C}_1(\boldsymbol{y}_{0,1} - \boldsymbol{\alpha}_1) + \boldsymbol{C}_2(\boldsymbol{y}_{0,2} - \boldsymbol{\alpha}_2),$$

where $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are the $q \times p_1$ and $q \times (p - p_1)$ submatrices of $\boldsymbol{C} = (\boldsymbol{C}_1, \boldsymbol{C}_2)$. From Fujikoshi and Nishii (1986), the hypothesis such that $\boldsymbol{y}_2$ is redundant for estimating $\boldsymbol{x}_0$ is written as

$$\boldsymbol{C}_2 = \boldsymbol{\mathcal{B}}_2 - \boldsymbol{\mathcal{B}}_1 \boldsymbol{\Gamma} = \boldsymbol{O}_{q,p-p_1}, \tag{5}$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ and $\boldsymbol{O}_{q,p-p_1}$ is the $q \times (p - p_1)$ matrix of zeros.

2.2 Models, Risk, and Bias

Let $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ be the $n \times p_1$ and $n \times (p - p_1)$ partitioned matrices of $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$, and let $\boldsymbol{\Theta}_1 = (\boldsymbol{\alpha}_1, \boldsymbol{\mathcal{B}}_1')'$, $\boldsymbol{\Theta}_2 = (\boldsymbol{\alpha}_2, \boldsymbol{\mathcal{B}}_2')'$, and $\widetilde{\boldsymbol{\Theta}}_2 = \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1\boldsymbol{\Gamma}$. From a property of a conditional distribution of a multivariate normal distribution (e.g., Srivastava and Khatri, 1979), we can express (2) as follows:

$$\boldsymbol{Y}_1 \sim N_{n\times p_1}(\boldsymbol{Z}\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_{11} \otimes \boldsymbol{I}_n),\ \boldsymbol{Y}_2|\boldsymbol{Y}_1 \sim N_{n\times(p-p_1)}(\boldsymbol{Z}\widetilde{\boldsymbol{\Theta}}_2 + \boldsymbol{Y}_1\boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{22\cdot 1} \otimes \boldsymbol{I}_n),$$

where $\boldsymbol{\Sigma}_{ab\cdot c} = \boldsymbol{\Sigma}_{ab} - \boldsymbol{\Sigma}_{ac}\boldsymbol{\Sigma}_{cc}^{-1}\boldsymbol{\Sigma}_{ca}$. Under hypothesis (5), it is straightforward to observe that $\boldsymbol{Z}\widetilde{\boldsymbol{\Theta}}_2 = \boldsymbol{1}_n\boldsymbol{\delta}'$, where $\boldsymbol{\delta} = \boldsymbol{\alpha}_2 - \boldsymbol{\Gamma}'\boldsymbol{\alpha}_1$. Therefore, the candidate model M such that $\boldsymbol{y}_2$ is redundant is expressed as

$$\boldsymbol{Y}_1 \sim N_{n\times p_1}(\boldsymbol{Z}\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_{11} \otimes \boldsymbol{I}_n),\ \boldsymbol{Y}_2|\boldsymbol{Y}_1 \sim N_{n\times(p-p_1)}(\boldsymbol{1}_n\boldsymbol{\delta}' + \boldsymbol{Y}_1\boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{22\cdot 1} \otimes \boldsymbol{I}_n). \tag{6}$$

Let $\boldsymbol{S} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{X})'(\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{X})$ be n times as large as the sample covariance matrix of $(\boldsymbol{y}', \boldsymbol{x}')'$, where $\boldsymbol{J}_n = n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n'$. We partition $\boldsymbol{S}$ as follows:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \boldsymbol{S}_{1x} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} & \boldsymbol{S}_{2x} \\ \boldsymbol{S}_{x1} & \boldsymbol{S}_{x2} & \boldsymbol{S}_{xx} \end{pmatrix},$$

where the sizes of $\boldsymbol{S}_{11}$, $\boldsymbol{S}_{22}$, and $\boldsymbol{S}_{xx}$ are $p_1 \times p_1$, $(p - p_1) \times (p - p_1)$, and $q \times q$, respectively. Let $f(\boldsymbol{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma})$ be the probability density function of $N_{n\times p}(\boldsymbol{Z}\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Then, the log-likelihood function is derived as

$$\begin{aligned} \ell(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{Y}, \boldsymbol{X}) &= -2 \log f(\boldsymbol{Y}; \boldsymbol{\Theta}, \boldsymbol{\Sigma}) \\ &= -\frac{1}{2}\left[np\log 2\pi + n\log|\boldsymbol{\Sigma}| + \mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta})\}\right]. \end{aligned}$$

By maximizing $\ell(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{Y}, \boldsymbol{X})$ under model (6), we can obtain the maximum likelihood estimators (MLEs) of $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22\cdot 1}$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\mathcal{B}}_1$, $\boldsymbol{\mathcal{B}}_2$, $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\delta}$ as follows (the proof is given in Appendix A):

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{11} &= \frac{1}{n}\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Y}_1, & \widehat{\boldsymbol{\Sigma}}_{22\cdot 1} &= \frac{1}{n}\boldsymbol{Y}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{1}_n, \boldsymbol{Y}_1)})\boldsymbol{Y}_2, \\ \widehat{\boldsymbol{\alpha}}_1 &= \bar{\boldsymbol{y}}_1 - \boldsymbol{S}_{1x}\boldsymbol{S}_{xx}^{-1}\bar{\boldsymbol{x}}, & \widehat{\boldsymbol{\alpha}}_2 &= \bar{\boldsymbol{y}}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{1x}\boldsymbol{S}_{xx}^{-1}\bar{\boldsymbol{x}}, \\ \widehat{\boldsymbol{\mathcal{B}}}_1 &= \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x1}, & \widehat{\boldsymbol{\mathcal{B}}}_2 &= \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x1}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}, \\ \widehat{\boldsymbol{\Theta}}_1 &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1 = \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_1' \\ \widehat{\boldsymbol{\mathcal{B}}}_1 \end{pmatrix}, & \widehat{\boldsymbol{\Theta}}_2 &= \begin{pmatrix} \bar{\boldsymbol{y}}_2' - \bar{\boldsymbol{x}}'\boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x1}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} \\ \boldsymbol{S}_{xx}^{-1}\boldsymbol{S}_{x1}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} \end{pmatrix}, \\ \widehat{\boldsymbol{\Gamma}} &= \boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}, & \widehat{\boldsymbol{\delta}} &= \bar{\boldsymbol{y}}_2 - \widehat{\boldsymbol{\Gamma}}'\bar{\boldsymbol{y}}_1, \end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ for a matrix $\boldsymbol{A}$, $\bar{\boldsymbol{y}}_1 = n^{-1}\mathbf{Y}_1'\mathbf{1}_n$, $\bar{\boldsymbol{y}}_2 = n^{-1}\mathbf{Y}_2'\mathbf{1}_n$, and $\bar{\boldsymbol{x}} = n^{-1}\boldsymbol{X}'\mathbf{1}_n$. We assume that the data are generated from the following true model M_*:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{\Delta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n),\ \boldsymbol{y}_0 \sim N_p(\boldsymbol{\xi}_*, \boldsymbol{\Sigma}_*),$$

where $\boldsymbol{\Delta}_*$ is an $n \times p$ true mean matrix, $\boldsymbol{\xi}_*$ is a p-dimensional true mean vector, and $\boldsymbol{\Sigma}_*$ is a $p \times p$ true covariance matrix assuming that $\boldsymbol{\Sigma}_*$ is positive definite. Under (6), we partition the true covariance matrix $\boldsymbol{\Sigma}_*$ in the same way as $\boldsymbol{\Sigma}$, as follows:

$$\boldsymbol{\Sigma}_* = \begin{pmatrix} \boldsymbol{\Sigma}_{11*} & \boldsymbol{\Sigma}_{12*} \\ \boldsymbol{\Sigma}_{21*} & \boldsymbol{\Sigma}_{22*} \end{pmatrix}.$$

Let $\boldsymbol{\Gamma}_* = \boldsymbol{\Sigma}_{11*}^{-1}\boldsymbol{\Sigma}_{12*}$. We state that the candidate model M is an overspecified model when M satisfies

$$\boldsymbol{\Delta}_* = \boldsymbol{Z}\boldsymbol{\Theta}_*,\ \boldsymbol{\mathcal{B}}_{2*} - \boldsymbol{\mathcal{B}}_{1*}\boldsymbol{\Gamma}_* = \boldsymbol{O}_{q,p-p_1}, \tag{7}$$

where $\boldsymbol{\Theta}_* = (\boldsymbol{\alpha}_*, \boldsymbol{\mathcal{B}}_*')'$ is a $(q+1)\times p$ true unknown matrix, $\boldsymbol{\alpha}_*$ is the p-dimensional true unknown vector of intercept coefficients, $\boldsymbol{\mathcal{B}}_*$ is the $q \times p$ true unknown matrix of regression coefficients, and $\boldsymbol{\mathcal{B}}_{1*}$ and $\boldsymbol{\mathcal{B}}_{2*}$ are $q \times p_1$ and $q \times (p - p_1)$ submatrices of $\boldsymbol{\mathcal{B}}_* = (\boldsymbol{\mathcal{B}}_{1*}, \boldsymbol{\mathcal{B}}_{2*})$. Let $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$ be the expected negative twofold log-likelihood function as follows:

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Sigma}) = E_{\boldsymbol{Y}}^*[-2\ell(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{Y}, \boldsymbol{X})] = -2\ell(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Delta}_*, \boldsymbol{X}) + n\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_*), \tag{8}$$

where $E_{\boldsymbol{Y}}^*$ is the expectation with respect to $\boldsymbol{Y}$ under the true model M_*. Then, we define the risk function R_{KL} as

$$R_{\mathrm{KL}} = E_{\boldsymbol{Y}}^*[\mathcal{L}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}})].$$

This type of risk function is essentially the same as the expected KL discrepancy function between the true model and a redundancy model. Although the best model is defined as the candidate model with the smallest risk function, we need to estimate the risk function because the risk function includes unknown parameters. Therefore, we usually estimate R_{KL} as $-2\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}; \boldsymbol{Y}, \boldsymbol{X})$, which is expressed as

$$-2\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}; \boldsymbol{Y}, \boldsymbol{X}) = np\{\log(2\pi) + 1\} + n(\log|\widehat{\boldsymbol{\Sigma}}_{11}| + \log|\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|), \tag{9}$$

under a candidate model M (the proof of (9) is given in Appendix A). However, when R_{KL} is estimated as (9), there is the following bias between the risk function and $-2\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}; \boldsymbol{Y}, \boldsymbol{X})$:

$$B_{\mathrm{KL}} = R_{\mathrm{KL}} - E_{\boldsymbol{Y}}^*[-2\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}; \boldsymbol{Y}, \boldsymbol{X})]. \tag{10}$$

Although $-2\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}; \boldsymbol{Y}, \boldsymbol{X}) + B_{\mathrm{KL}}$ may be considered as a criterion, note that the bias B_{KL} also includes unknown parameters. Therefore, we usually consider estimating B_{KL} by using its estimator.

3 Main results

3.1 High-Dimensional Bias-Corrected AIC

Let the number of parameters in the candidate model M be $h_M = \{p + qp_1 + p(p+1)/2\}$. Following Akaike (1973), by approximating B_{KL} as $2h_M$, the formal AIC is given by

$$\mathrm{AIC} = \begin{cases} np\{\log(2\pi)+1\} + n(\log|\widehat{\boldsymbol{\Sigma}}_{11}| + \log|\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|) + 2h_M & (q \le p_1 < p) \\ np\{\log(2\pi)+1\} + n\log|\widehat{\boldsymbol{\Sigma}}_{\omega}| + 2h_M & (p_1 = p) \end{cases}, \tag{11}$$

where $\widehat{\boldsymbol{\Sigma}}_{\omega} = n^{-1}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Y}$ is the MLE of $\boldsymbol{\Sigma}$ in the model where none of the response variables $\boldsymbol{y}$ are redundant. This formal AIC has been given by Fujikoshi and Nishii (1986) and the model which minimizes the AIC should be selected.

However, since the AIC cannot approximate the risk function well when n is small or when n and p are both large, the AIC may not perform adequately in general terms. Therefore, it is important to consider a new criterion that has a better approximation than the AIC in such cases. Correcting the bias B_{KL} in the HHD asymptotic framework, we propose a high-dimensional bias-corrected AIC ($\mathrm{HAIC_C}$) as follows:

$$\mathrm{HAIC_C} = \begin{cases} np\{\log(2\pi)+1\} + n(\log|\widehat{\boldsymbol{\Sigma}}_{11}| + \log|\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|) + \widehat{m}(n,p) & (q \le p_1 < p) \\ np\log(2\pi) + n\log|\widehat{\boldsymbol{\Sigma}}_{\omega}| + \dfrac{np(n+q+1)}{n-p-q-2} & (p_1 = p) \end{cases}. \tag{12}$$

Here, $\widehat{m}(n,p)$ is defined by

$$\begin{aligned} \widehat{m}(n,p) &= m_1(n,p) + \widehat{m}_2(n,p), \\ m_1(n,p) &= -np + \frac{np_1(n+q+1)}{n-p_1} + \frac{np_1(q+2)(n+q+1)}{(n-p_1)^2} + \frac{n^2p_1(q+2)^2}{(n-p_1)^3} \\ &\quad + \frac{n(p-p_1)(n+1)}{n-p} + \frac{2n(p-p_1)(n+1)}{(n-p)^2} + \frac{4n^2(p-p_1)}{(n-p)^3} \\ &\quad + \frac{n(p-p_1)q}{(n-p)(n-p_1)}\left[n\left\{1 + 2\left(\frac{1}{n-p_1} + \frac{1}{n-p}\right)\right\} - (q+2)\right] \\ &\quad + \frac{n(n+1)(p-p_1)(p_1-q)}{(n-p)(n-p_1)}\left\{1 + 2\left(\frac{1}{n-p_1} + \frac{1}{n-p}\right)\right\} \\ &\quad + \frac{4n\{n(p_1-q)+p_1\}(p-p_1)}{(n-p)(n-p_1)}\left\{\frac{1}{(n-p_1)^2} + \frac{1}{(n-p)^2} + \frac{1}{(n-p)(n-p_1)}\right\} \\ &\quad + \frac{8n^2(p-p_1)p_1}{(n-p)(n-p_1)}\left\{\frac{1}{(n-p_1)^3} + \frac{1}{(n-p)^3} + \frac{1}{(n-p)(n-p_1)^2} + \frac{1}{(n-p)^2(n-p_1)}\right\}, \end{aligned} \tag{13}$$

$$\widehat{m}_2(n,p) = \frac{n^2}{(n-p)(n-p_1)}\{2(q+1)\widehat{\tau}_1 - \widehat{\tau}_2 - \widehat{\tau}_3\},$$

and $\widehat{\tau}_1$, $\widehat{\tau}_2$ and $\widehat{\tau}_3$ are given by

$$\widehat{\tau}_1 = \frac{p-p_1}{n-p_1}\left\{\mathrm{tr}\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\} - (p_1-q)\right\},\ \widehat{\tau}_2 = \frac{n}{n-p_1}\widehat{\tau}_1^2, \tag{14}$$

$$\widehat{\tau}_3 = \frac{n(p-p_1)}{(n-p_1)^2}\left\{\mathrm{tr}[\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}^2] - (p_1-q)\right\} \tag{15}$$

where $\boldsymbol{S}_e = n\hat{\boldsymbol{\Sigma}}_{11} = \boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Y}_1$ and $\boldsymbol{S}_h = \boldsymbol{Y}_1'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}_1$. Practically, to propose the HAIC$_{\mathrm{C}}$ when $q \le p_1 < p$, we took the following step. We expanded the bias B_{KL} in overspecified models (the result is given in Appendix B, Proposition B.2). Since there are unknown parameters included in the expanded term, we replaced the unknown parameters as $\widehat{\tau}_1$, $\widehat{\tau}_2$ and $\widehat{\tau}_3$. Therefore, the HAIC$_{\mathrm{C}}$ is proposed using $\widehat{m}(n,p)$ instead of B_{KL}. When $p_1 = p$, the redundancy model means the usual multivariate linear regression such that all response variables $\boldsymbol{y}$ are not redundant. Therefore, from Bedrick and Tsai (1994), the bias for $p_1 = p$ is exactly equal to $-np + \{np(n + q + 1)\}/(n - p - q - 2)$.

3.2 Asymptotic property of HAIC$_{\mathrm{C}}$

To present an asymptotic property of the HAIC$_{\mathrm{C}}$, we start by offering some notation and delineating assumptions. Let $\boldsymbol{\Xi}_*$ be the $p_1 \times p_1$ constant matrix given by

$$\boldsymbol{\Xi}_* = \boldsymbol{\mathcal{B}}_{1*}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\mathcal{B}}_{1*}. \tag{16}$$

Note that $\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{\Xi}_*\boldsymbol{\Sigma}_{11*}^{-1/2}$ is the $p_1 \times p_1$ symmetric matrix and $\mathrm{rank}(\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{\Xi}_*\boldsymbol{\Sigma}_{11*}^{-1/2}) \le \min\{p_1, q\} = q$. Therefore, its spectral decomposition can be expressed as

$$\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{\Xi}_*\boldsymbol{\Sigma}_{11*}^{-1/2} = \boldsymbol{H}\begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{O}_{q,p_1-q} \\ \boldsymbol{O}_{p_1-q,q} & \boldsymbol{O}_{p_1-q,p_1-q} \end{pmatrix}\boldsymbol{H}', \tag{17}$$

where $\boldsymbol{H}$ is a $p_1 \times p_1$ orthogonal matrix, and $\boldsymbol{\Lambda}$ is a $q \times q$ diagonal matrix of which the a-th diagonal element is a singular value λ_a, i.e., $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ with $\lambda_1 \ge \cdots \ge \lambda_q \ge 0$. Let $\boldsymbol{G}$ be a $q \times q$ random matrix distributed according to the non-central Wishart distribution with $(n-1)$ degrees of freedom, covariance matrix $\boldsymbol{\Phi}^{-1}$, and non-central parameter matrix $\boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1}$, that is,

$$\boldsymbol{G} \sim W_q(n-1, \boldsymbol{\Phi}^{-1}; \boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1}), \tag{18}$$

where $\boldsymbol{\Phi}$ is defined by

$$\boldsymbol{\Phi} = (n-1)\boldsymbol{I}_q + \boldsymbol{\Lambda}. \tag{19}$$

We prepare the following assumptions for the moments of $\boldsymbol{G}$:

Assumption A1. $E[\mathrm{tr}(\boldsymbol{G}^{-2})] = O(1)$.
Assumption A2. $E[\mathrm{tr}(\boldsymbol{G}^{-4})] = O(1),\ E[\mathrm{tr}(\boldsymbol{G}^{-2})^2] = O(n + \lambda_q),\ E[\mathrm{tr}(\boldsymbol{G}^{-8})] = O((n+\lambda_q)^3)$.

Note that $\boldsymbol{\Phi}^{-1} = O((n+\lambda_q)^{-1})$ and $\boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1} = O(1)$ because

$$||\boldsymbol{\Phi}^{-1}||^2 = \sum_{i=1}^{q} \frac{1}{(n-1+\lambda_i)^2} \le \frac{q}{(n-1+\lambda_q)^2} = O((n+\lambda_q)^{-2}),$$

$$||\boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1}||^2 = \sum_{i=1}^{q} \left(\frac{\lambda_i}{n-1+\lambda_i}\right)^2 \le \frac{q\lambda_1^2}{(n-1+\lambda_1)^2} = O(1),$$

where $||\cdot||$ is the Frobenius norm. Therefore, Assumptions A1 and A2 are natural. We obtain the asymptotic unbiasedness of the HAIC$_{\mathrm{C}}$ for R_{KL} as Theorem 3.1. Theorem 3.1 is directly derived from (10), Proposition B.2 in Appendix B, and Proposition C.2 in Appendix C.

Theorem 3.1. *Suppose that $q \leq p_1 < p$, and Assumptions A1 and A2 hold. Then, under overspecified models* (7)*, we have*

$$R_{\mathrm{KL}} - E_{\boldsymbol{Y}}^{*}[\mathrm{HAIC_C}] = O(pn^{-1}(n+\lambda_q)^{-1/2}),$$

as $n \to \infty$ and $p/n \to c \in [0,1)$, where λ_q is the q-th singular value of $\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{\Xi}_*\boldsymbol{\Sigma}_{11*}^{-1/2}$.

From Theorem 3.1, the $\mathrm{HAIC_C}$ approximates the risk function well in cases where n is not large and when n and p are both large. Thus, it is expected that the $\mathrm{HAIC_C}$ will work well.

4 Numerical simulations

We conduct numerical simulations to show that the $\mathrm{HAIC_C}$ in (12) works better than the AIC in (11). The p redundancy models j ($j = 1, \ldots, p$) were prepared for Monte Carlo simulations with 10,000 iterations. Here, model j denotes a model such that $y_1, \ldots, y_j$ are not redundant and $y_{j+1}, \ldots, y_p$ are redundant. Suppose that the true model is the model such that p_*-response variables $y_1, \ldots, y_{p_*}$ are not redundant. The data $\boldsymbol{Y}$ were generated from the true model $N_{n\times p}(\boldsymbol{Z}\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n)$, where $\boldsymbol{Z} = (\boldsymbol{1}_n, \boldsymbol{X})$, $\boldsymbol{\Theta}_* = (\boldsymbol{\alpha}_*, \boldsymbol{\mathcal{B}}_*')'$, and we gave $\boldsymbol{\alpha}_* = \boldsymbol{1}_p$ and $\boldsymbol{\Sigma}_* = (1-0.8)\boldsymbol{I}_p + 0.8\boldsymbol{1}_p\boldsymbol{1}_p'$. We constructed $\boldsymbol{X}$ and $\boldsymbol{\mathcal{B}}_*$ as follows. First, we independently generated $u_1, \ldots, u_n$ from $U(-1,1)$, where $U(a,b)$ denotes the uniform distribution with the range (a,b). Using $u_1, \ldots, u_n$, we constructed $\boldsymbol{X}$ as $\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}_0$, where (i,j)-th element of $\boldsymbol{X}_0$ is defined by u_i^j ($1 \leq i \leq n, 1 \leq j \leq q$). Second, let $\boldsymbol{\mathcal{B}}_{1*}$ and $\boldsymbol{\mathcal{B}}_{2*}$ be submatrices of $\boldsymbol{\mathcal{B}}_*$ in (4) when $p_1 = p_*$. Similarly, $\boldsymbol{\Sigma}_{11*}$ and $\boldsymbol{\Sigma}_{12*}$ are submatrices of $\boldsymbol{\Sigma}_*$. Using the above notation, we constructed $\boldsymbol{\mathcal{B}}_{2*}$ as $\boldsymbol{\mathcal{B}}_{2*} = \boldsymbol{\mathcal{B}}_{1*}\boldsymbol{\Sigma}_{11*}^{-1}\boldsymbol{\Sigma}_{12*}$, where the ℓ ($1 \leq \ell \leq p_* - 1$)-th column vectors and p_*-th column vector of $\boldsymbol{\mathcal{B}}_{1*}$ are defined by $(1, 2, \ldots, 2)'$ and $(1, 2p_*, \ldots, 2p_*)'$, respectively. Next, we set $q = 2$, and data were generated for various combinations of n, p, and p_*. In order to compare the performances of the $\mathrm{HAIC_C}$ and AIC, the following three properties were considered:

- Ratio of the information criterion (IC) average to the risk function R_{KL}: $E_{\boldsymbol{Y}}^{*}[\mathrm{IC}]/R_{\mathrm{KL}}$.
- The probability of selecting the true model: the frequency with which the true model is chosen as the best model.
- The KL information of the predicted values of the best model chosen by the information criterion, which is defined by

$$\mathrm{KL} = \frac{1}{np}\{E_{\boldsymbol{Y}}^{*}[\mathcal{L}(\widehat{\boldsymbol{\Theta}}_{\mathrm{best}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{best}})] - \mathcal{L}(\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_*)\},$$

  where $\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$ is given by (8) and $\widehat{\boldsymbol{\Theta}}_{\mathrm{best}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{best}}$ are the MLEs of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$, respectively, under the best model.

A high-performance model selector is considered to be a model selection criterion where the ratio of the information criterion average to the risk function is close to 1, where there is a high probability of selecting the true model, and where there is a small amount of KL information. Figures 1-3 show the ratio of the average of each criterion, i.e., the $\mathrm{HAIC_C}$ and the AIC to R_{KL}. From Figures 1-3, it can be observed that the values for the $\mathrm{HAIC_C}$ are closer to 1 than those for the AIC. Therefore, the $\mathrm{HAIC_C}$ can approximate the risk exactly, but the AIC underestimates this risk. Moreover, as the number of dimensions p increases, the bias of the AIC increases. Table 1 shows the probabilities of selecting the true model and the amount of KL information for the $\mathrm{HAIC_C}$ and AIC. Therein, it is clear that the probability for the $\mathrm{HAIC_C}$ is higher than that for the AIC. On the other hand, the KL information for the $\mathrm{HAIC_C}$ nearly equals that for the AIC, except when $p = 80$ and $p = 160$, when it is significantly smaller than that for the AIC.

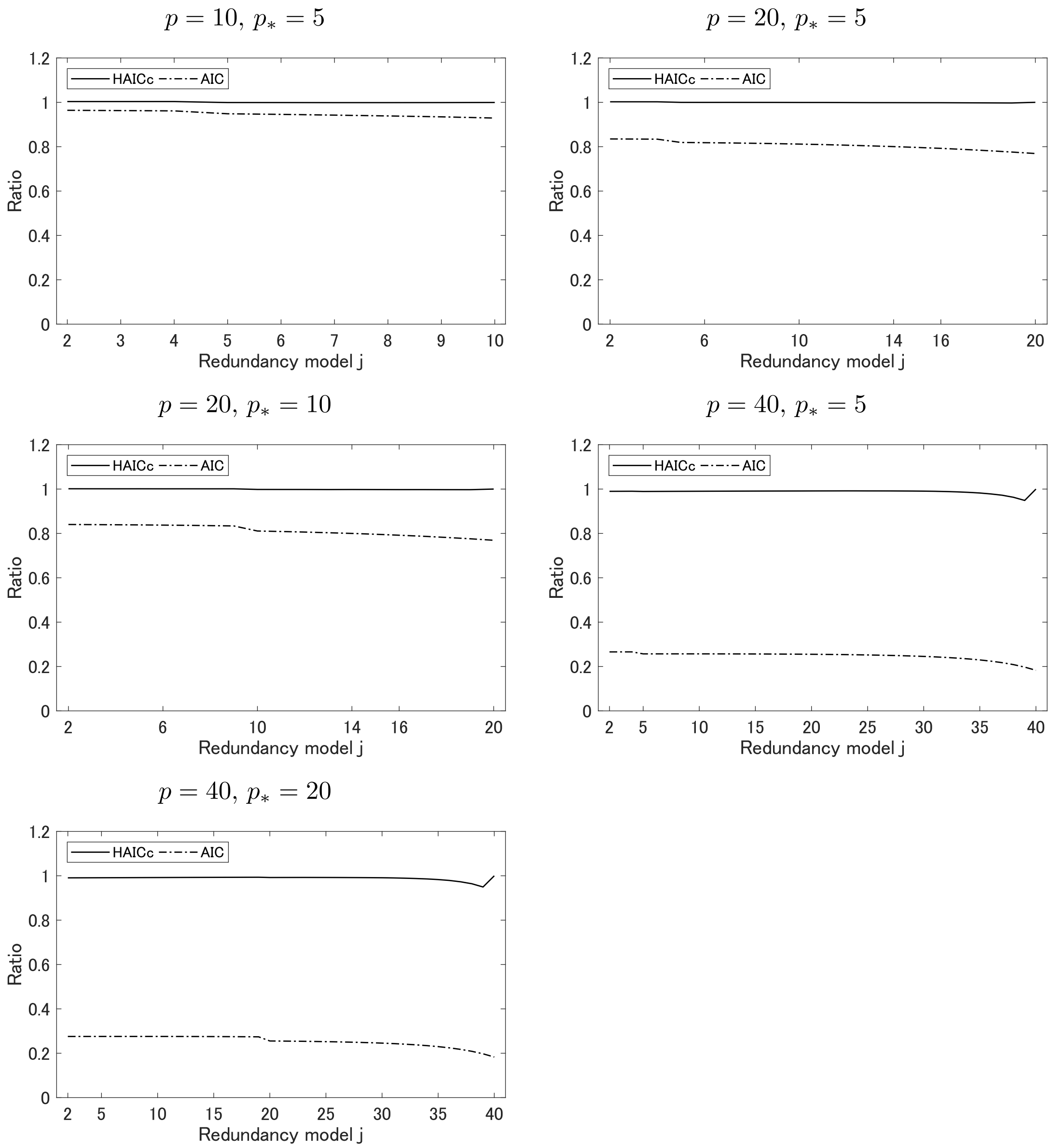

Figure 1: Ratio of the average of each criterion to the risk function with $n = 50$

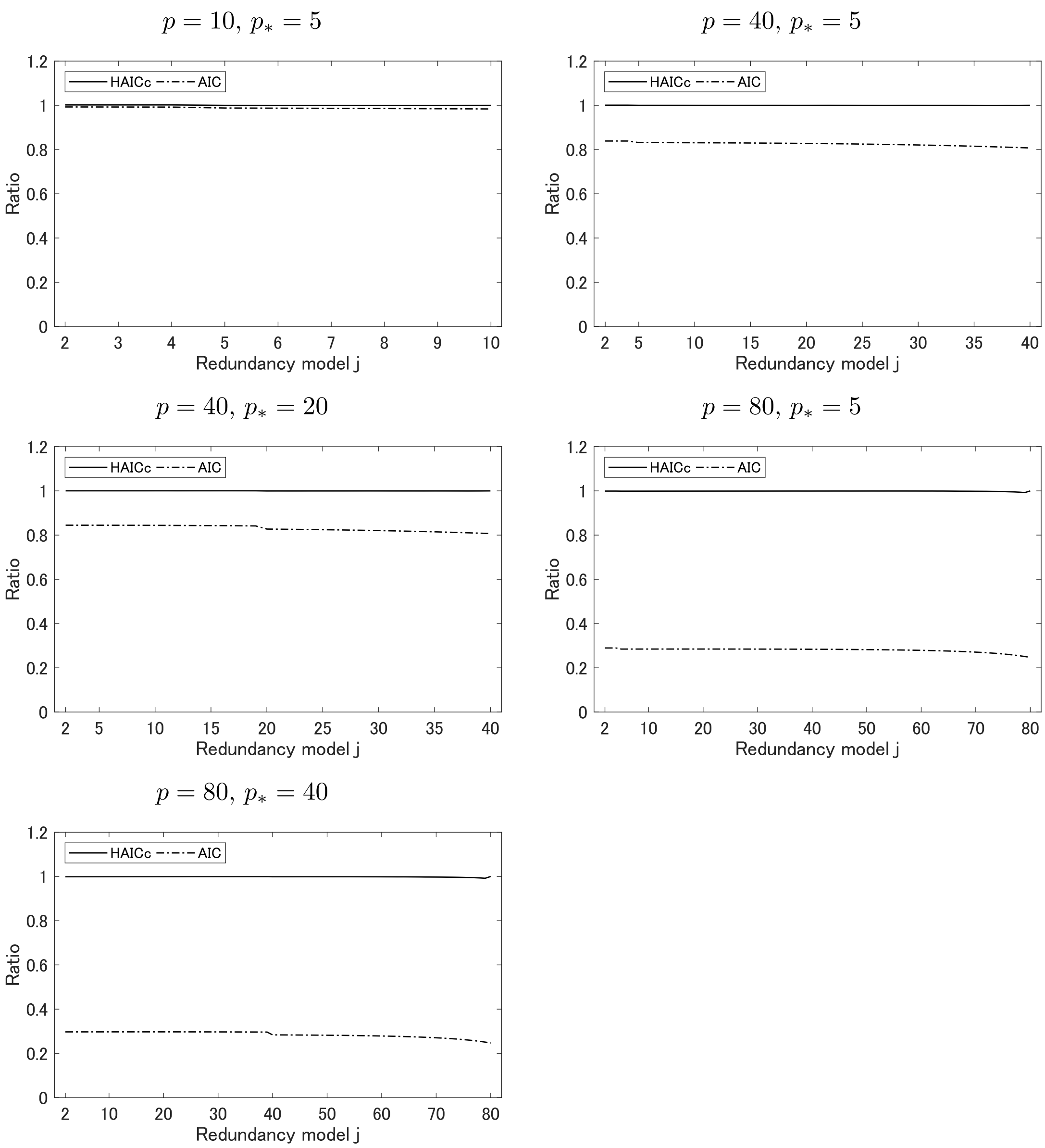

Figure 2: Ratio of the average of each criterion to the risk function with $n = 100$

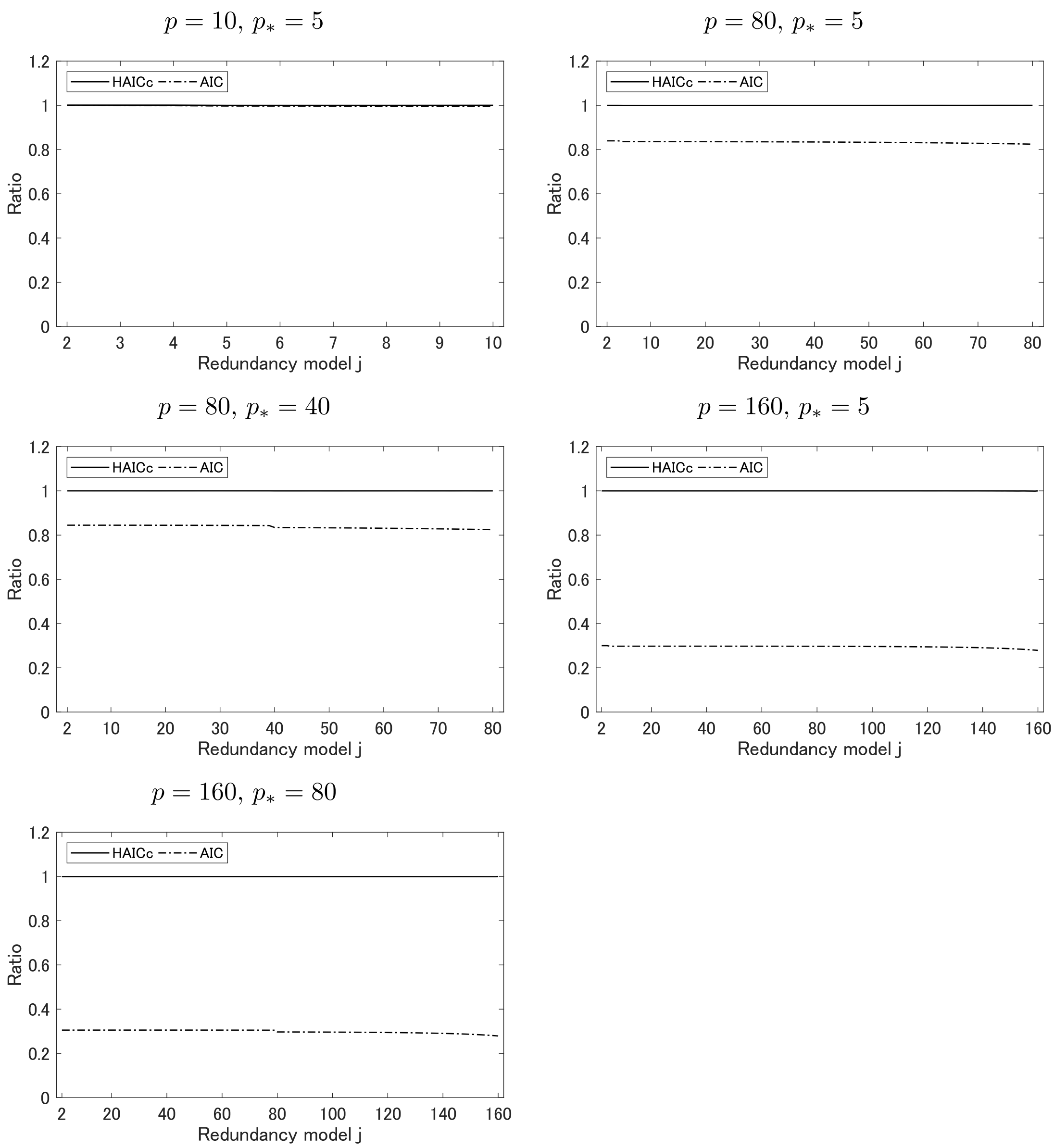

Figure 3: Ratio of the average of each criterion to the risk function with $n = 200$

Table 1: Probabilities of selecting the true model and the KL information of the predicted values of the best model

			Probabilities (%)		KL	
n	p	p_*	HAIC$_\mathrm{C}$	AIC	HAIC$_\mathrm{C}$	AIC
50	10	5	92.48	68.86	0.2290	0.2329
50	20	5	92.58	61.51	0.5596	0.5715
50	20	10	96.95	52.14	0.5888	0.6075
50	40	5	96.23	10.29	4.7150	6.5024
50	40	20	99.87	1.64	4.7853	6.7014
100	10	5	87.58	75.58	0.0916	0.0921
100	40	5	86.90	72.05	0.4903	0.4909
100	40	20	96.18	56.16	0.5068	0.5109
100	80	5	89.86	16.72	3.9800	4.5527
100	80	40	99.81	0.56	4.0173	4.6931
200	10	5	84.28	77.84	0.0419	0.0419
200	80	5	83.51	76.72	0.4609	0.4610
200	80	40	95.44	60.14	0.4689	0.4696
200	160	5	84.36	22.35	3.6758	3.9010
200	160	80	99.61	0.13	3.6953	3.9898

Acknowledgments

Ryoya Oda was supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. Hirokazu Yanagihara and Yasunori Fujikoshi were both partially supported by the Ministry of Education, Science, Sports, and Culture through Grants-in-Aid for Scientific Research (C), #18K03415 and #16K00047, respectively.

Appendix

A Deriving MLEs and proof of (9)

To derive MLEs, we use the following lemma (e.g., Fujikoshi *et al.*, 2010, A.2.11):

Lemma A.1. *Let $\boldsymbol{Y}$ be an $n \times p$ known matrix and $\boldsymbol{A}$ be an $n \times k$ known matrix of rank k. Consider a function of $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$ and $k \times p$ matrix $\boldsymbol{\Theta}$ given by*

$$g(\boldsymbol{\Theta}, \boldsymbol{\Sigma}) = m \log |\boldsymbol{\Sigma}| + \mathrm{tr}\{(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\Theta})'\},$$

where $m > 0$ and all elements of $\boldsymbol{\Theta}$ are bounded. Then, $g(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$ takes the minimum at

$$\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}, \ \boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}} = \frac{1}{m}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{Y},$$

and the minimum value is given by $m \log |\widehat{\boldsymbol{\Sigma}}| + mp$.

First, we derive MLEs of $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22\cdot1}$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\mathcal{B}}_1$, $\boldsymbol{\mathcal{B}}_2$, $\boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_2$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\delta}$. From (6), we have

$$\begin{aligned}-2\ell(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{Y}, \boldsymbol{X}) = np \log 2\pi \\ &+ n \log |\boldsymbol{\Sigma}_{11}| + \mathrm{tr}\{(\boldsymbol{Y}_1 - \boldsymbol{Z}\boldsymbol{\Theta}_1)\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_1 - \boldsymbol{Z}\boldsymbol{\Theta}_1)'\} \\ &+ n \log |\boldsymbol{\Sigma}_{22\cdot1}| + \mathrm{tr}\{(\boldsymbol{Y}_2 - \mathbf{1}_n\boldsymbol{\delta}' - \boldsymbol{Y}_1\boldsymbol{\Gamma})\boldsymbol{\Sigma}_{22\cdot1}^{-1}(\boldsymbol{Y}_2 - \mathbf{1}_n\boldsymbol{\delta}' - \boldsymbol{Y}_1\boldsymbol{\Gamma})'\}.\end{aligned}$$

Let $g_1(\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_{11}) = n \log |\boldsymbol{\Sigma}_{11}| + \mathrm{tr}\{(\boldsymbol{Y}_1 - \boldsymbol{Z}\boldsymbol{\Theta}_1)\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{Y}_1 - \boldsymbol{Z}\boldsymbol{\Theta}_1)'\}$ and $g_2(\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{22\cdot1}) = n \log |\boldsymbol{\Sigma}_{22\cdot1}| + \mathrm{tr}\{(\boldsymbol{Y}_2 - \mathbf{1}_n\boldsymbol{\delta}' - \boldsymbol{Y}_1\boldsymbol{\Gamma})\boldsymbol{\Sigma}_{22\cdot1}^{-1}(\boldsymbol{Y}_2 - \mathbf{1}_n\boldsymbol{\delta}' - \boldsymbol{Y}_1\boldsymbol{\Gamma})'\}$. The set of unknown parameters $\{\boldsymbol{\Theta}, \boldsymbol{\Sigma}\}$ exhibits one-to-one correspondence with the set $\{\boldsymbol{\alpha}_1, \boldsymbol{\mathcal{B}}_1, \boldsymbol{C}_2, \boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22\cdot1}\}$, and each set of parameters for g_1 and g_2 is separated. Thus, we only consider each minimization of g_1 and g_2. Note that $\mathrm{rank}(\boldsymbol{Z}) = q + 1$ and $\mathrm{rank}\{(\mathbf{1}_n, \boldsymbol{Y}_1)\} = p_1 + 1$. Then, from Lemma A.1, we can state that

$$\begin{aligned}\min_{\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_{11}} g_1(\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_{11}) &= n \log |\widehat{\boldsymbol{\Sigma}}_{11}| + \mathrm{tr}\{(\boldsymbol{Y}_1 - \boldsymbol{Z}\widehat{\boldsymbol{\Theta}}_1)\widehat{\boldsymbol{\Sigma}}_{11}^{-1}(\boldsymbol{Y}_1 - \boldsymbol{Z}\widehat{\boldsymbol{\Theta}}_1)'\} \\ &= n \log |\widehat{\boldsymbol{\Sigma}}_{11}| + np_1, \\ \min_{\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{22\cdot1}} g_2(\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_{22\cdot1}) &= n \log |\widehat{\boldsymbol{\Sigma}}_{22\cdot1}| + \mathrm{tr}\{(\boldsymbol{Y}_2 - \mathbf{1}_n\widehat{\boldsymbol{\delta}}' - \boldsymbol{Y}_1\widehat{\boldsymbol{\Gamma}})\widehat{\boldsymbol{\Sigma}}_{22\cdot1}^{-1}(\boldsymbol{Y}_2 - \mathbf{1}_n\widehat{\boldsymbol{\delta}}' - \boldsymbol{Y}_1\widehat{\boldsymbol{\Gamma}})'\} \\ &= n \log |\widehat{\boldsymbol{\Sigma}}_{22\cdot1}| + n(p - p_1),\end{aligned}$$

where $\widehat{\boldsymbol{\Theta}}_1 = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1$,$\widehat{\boldsymbol{\Sigma}}_{11} = n^{-1}\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Y}_1$, $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')' = \{(\mathbf{1}_n, \boldsymbol{Y}_1)'(\mathbf{1}_n, \boldsymbol{Y}_1)\}^{-1}(\mathbf{1}_n, \boldsymbol{Y}_1)'\boldsymbol{Y}_2$ and $\widehat{\boldsymbol{\Sigma}}_{22\cdot1} = n^{-1}\boldsymbol{Y}_2'(\boldsymbol{I}_n - \boldsymbol{P}_{(\mathbf{1}_n, \boldsymbol{Y}_1)})\boldsymbol{Y}_2$. Therefore, the MLEs of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, and $\boldsymbol{\Theta}_2$ can also be derived from $\widehat{\boldsymbol{\Theta}}_1$ and $\widehat{\boldsymbol{\delta}}$.

Next, we show (9). From Lemma A.1, it is clearly that

$$-2\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}; \boldsymbol{Y}, \boldsymbol{X}) = np\{\log(2\pi) + 1\} + n(\log |\widehat{\boldsymbol{\Sigma}}_{11}| + \log |\widehat{\boldsymbol{\Sigma}}_{22\cdot1}|).$$

□

B Calculating and expanding bias

Here, we present propositions for calculating and expanding the bias B_{KL} in (10).

B.1 Calculating bias and its proof

Proposition B.1. *Suppose that* $q \le p_1 < p$. *In overspecified models* (7), *we obtain the exact expression of the bias as follows:*

$$\begin{aligned}B_{\mathrm{KL}} = -np + \frac{np_1(n+q+1)}{n-p_1-q-2} + \frac{n(n+1)(p-p_1)}{n-p-2} \\ + \frac{n(n+1)(p-p_1)}{n-p-2} E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{S}_{11}^{-1}\boldsymbol{\Sigma}_{11*})] + \frac{n(p-p_1)}{n-p-2} E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{S}_{11}^{-1}\boldsymbol{\Xi}_*)],\end{aligned} \quad \text{(B.1)}$$

where $\boldsymbol{\Xi}_*$ *is defined in* (16).

The proof of Proposition B.1 is presented as follows. To calculate the bias, we describe a lemma for distributions of some statistics (the proof is given in Appendix D).

Lemma B.1. *Suppose that $q \le p_1 < p$. Let $\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ be the MLE of $\boldsymbol{\Sigma}_{22\cdot 1}$ under model* (6)*, and let*

$$\boldsymbol{T} = \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{1/2}\left[\widehat{\boldsymbol{\Gamma}} - \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1}\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\Omega}_* - \boldsymbol{\Gamma}_*\right]\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2}, \tag{B.2}$$

$$\boldsymbol{U} = \{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{1/2}\left[\begin{pmatrix}\widehat{\boldsymbol{\delta}}' \\ \widehat{\boldsymbol{\Gamma}}\end{pmatrix} - \{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{-1}(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{\Omega}_* + \boldsymbol{Y}_1\boldsymbol{\Gamma}_*)\right]\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2}, \tag{B.3}$$

where $\boldsymbol{\Omega}_* = \boldsymbol{\Delta}_{2*} - \boldsymbol{\Delta}_{1*}\boldsymbol{\Gamma}_*$, and $\boldsymbol{\Delta}_{1*}$ and $\boldsymbol{\Delta}_{2*}$ are the $n \times p_1$ and $n \times (p - p_1)$ partitioned matrices of $\boldsymbol{\Delta}_* = (\boldsymbol{\Delta}_{1*}, \boldsymbol{\Delta}_{2*})$. Then, $\boldsymbol{T}$ and $\boldsymbol{U}$ are independent of $\boldsymbol{Y}_1$, and $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ are mutually conditional independent under $\boldsymbol{Y}_1$, and

$$\begin{aligned} n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|\boldsymbol{Y}_1 &\sim W_{p-p_1}(n - p_1 - 1, \boldsymbol{\Sigma}_{22\cdot 1*}; \widetilde{\boldsymbol{\Omega}}_*), \\ \boldsymbol{T} &\sim N_{p_1 \times (p-p_1)}(\boldsymbol{O}_{p,p-p_1}, \boldsymbol{I}_{p-p_1} \otimes \boldsymbol{I}_{p_1}), \\ \boldsymbol{U} &\sim N_{(p_1+1) \times (p-p_1)}(\boldsymbol{O}_{p_1+1,p-p_1}, \boldsymbol{I}_{p-p_1} \otimes \boldsymbol{I}_{p_1+1}), \end{aligned}$$

where $\widetilde{\boldsymbol{\Omega}}_* = \boldsymbol{\Omega}_*'(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{1}_n, \boldsymbol{Y}_1)})\boldsymbol{\Omega}_*$. Moreover, if the model is overspecified model (7)*, then $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $\boldsymbol{Y}_1$ are mutually independent, $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ are also mutually independent, and*

$$\begin{aligned} n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1} &\sim W_{p-p_1}(n - p_1 - 1, \boldsymbol{\Sigma}_{22\cdot 1*}), \\ \boldsymbol{T} &= \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{1/2}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_*)\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2}, \\ \boldsymbol{U} &= \{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{1/2}\begin{pmatrix}\widehat{\boldsymbol{\delta}}' - \boldsymbol{\delta}_*' \\ \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_*\end{pmatrix}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2}, \end{aligned}$$

where $\boldsymbol{\delta}_* = \boldsymbol{\alpha}_{2*} - \boldsymbol{\Gamma}_*'\boldsymbol{\alpha}_{1*}$ and $\boldsymbol{\alpha}_{1*}$ and $\boldsymbol{\alpha}_{2*}$ are the p_1- and $(p - p_1)$-dimensional subvectors of $\boldsymbol{\alpha}_* = (\boldsymbol{\alpha}_{1*}', \boldsymbol{\alpha}_{2*}')'$.

From the general formula of the determinant of a partitioned matrix (e.g., Lütkepohl, 1997, 4.2.2 (6)), $\log|\widehat{\boldsymbol{\Sigma}}| = \log|\widehat{\boldsymbol{\Sigma}}_{11}| + \log|\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|$ holds. Under overspecified models (7), $\mathcal{L}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}})$ can be expressed as

$$\begin{aligned} &\mathcal{L}(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\Sigma}}) \\ &= np\log 2\pi + n(\log|\widehat{\boldsymbol{\Sigma}}_{11}| + \log|\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|) + n\mathrm{tr}(\boldsymbol{\Sigma}_*\widehat{\boldsymbol{\Sigma}}^{-1}) + \mathrm{tr}\{\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)\}. \end{aligned}$$

Thus, from the above equation and (9), the bias B_{KL} in (10) can be expressed as

$$B_{\mathrm{KL}} = -np + E_{\boldsymbol{Y}}^*\left[n\mathrm{tr}(\boldsymbol{\Sigma}_*\widehat{\boldsymbol{\Sigma}}^{-1}) + \mathrm{tr}\{\widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)\}\right].$$

We separate the second and third terms in the above equation into four terms in order to calculate the bias. It follows from the general formula for the inverse of a block matrix (e.g., Harville, 1997, Theorem 8.5.11) that

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{11}^{-1} & \boldsymbol{O}_{p_1,p-p_1} \\ \boldsymbol{O}_{p-p_1,p_1} & \boldsymbol{O}_{p-p_1,p-p_1}\end{pmatrix} + \begin{pmatrix}-\widehat{\boldsymbol{\Gamma}} \\ \boldsymbol{I}_{p-p_1}\end{pmatrix}\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}^{-1}(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1}).$$

Then, we can derive another expression of the bias B_{KL} as follows:

$$B_{\mathrm{KL}} = -np + nE_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Sigma}_{11*}\widehat{\boldsymbol{\Sigma}}_{11}^{-1})] + nE_{\boldsymbol{Y}}^*\left[\mathrm{tr}\left\{\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}^{-1}(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})\boldsymbol{\Sigma}_*\begin{pmatrix}-\widehat{\boldsymbol{\Gamma}} \\ \boldsymbol{I}_{p-p_1}\end{pmatrix}\right\}\right]$$

$$
\begin{aligned}
&+ E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\{\widehat{\boldsymbol{\Sigma}}_{11}^{-1}(\widehat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{1*})'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\Theta}}_1 - \boldsymbol{\Theta}_{1*})\}\right] \\
&+ E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}^{-1}(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)\begin{pmatrix} -\widehat{\boldsymbol{\Gamma}} \\ \boldsymbol{I}_{p-p_1}\end{pmatrix}\right\}\right] \\
&= -np + (i) + (ii) + (iii) + (iv)\ (say),
\end{aligned} \tag{B.4}
$$

where $\boldsymbol{\Theta}_{1*}$ and $\boldsymbol{\Theta}_{2*}$ are the $(q+1)\times p_1$ and $(q+1)\times(p-p_1)$ submatrices of $\boldsymbol{\Theta}_* = (\boldsymbol{\Theta}_{1*}, \boldsymbol{\Theta}_{2*})$. Now, we calculate (i), (ii), (iii), and (iv) in (B.4). Starting with (i) in (B.4), it can be stated that $n\boldsymbol{\Sigma}_{11*}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{11}\boldsymbol{\Sigma}_{11*}^{-1/2}$ is distributed according to $W_{p_1}(n-q-1, \boldsymbol{I}_{p_1})$. Thus, by using the expectation of an inverted Wishart distribution (see, Watamori, 1990), we have

$$
(i) = nE^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{\Sigma}_{11*}\widehat{\boldsymbol{\Sigma}}_{11}^{-1})] = \frac{n^2 p_1}{n - p_1 - q - 2}. \tag{B.5}
$$

Now we calculate (ii) in (B.4). From Lemma B.1, by using the expectation of an inverted Wishart distribution (e.g., Fujikoshi *et al.*, 2010, Theorem 2.2.7), the expectation of $E^*_{\boldsymbol{Y}}[\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}^{-1}]$ can be calculated as

$$
E^*_{\boldsymbol{Y}}[\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}^{-1}] = nE^*_{\boldsymbol{Y}}[(n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1})^{-1}] = \frac{n}{n-p-2}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1}.
$$

From the above equation and the fact that $\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ is independent of $\widehat{\boldsymbol{\Gamma}}$, (ii) can be expressed as

$$
(ii) = \frac{n^2}{n-p-2}E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{\boldsymbol{\Sigma}_*\begin{pmatrix} -\widehat{\boldsymbol{\Gamma}} \\ \boldsymbol{I}_{p-p_1}\end{pmatrix}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1}(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})\right\}\right].
$$

By the definition of $\boldsymbol{T}$ in (B.2), we have

$$
\begin{aligned}
&\boldsymbol{\Sigma}_*\begin{pmatrix} -\widehat{\boldsymbol{\Gamma}} \\ \boldsymbol{I}_{p-p_1}\end{pmatrix}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1}(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1}) \\
&= \boldsymbol{\Sigma}_*\begin{pmatrix} -\boldsymbol{\Gamma}_* \\ \boldsymbol{I}_{p-p_1}\end{pmatrix}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1}(-\boldsymbol{\Gamma}'_*, \boldsymbol{I}_{p-p_1}) \\
&\quad + \boldsymbol{\Sigma}_*\begin{pmatrix} -\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1/2}\boldsymbol{T}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2}\boldsymbol{\Gamma}'_* & -\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1/2}\boldsymbol{T}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2} \\ \boldsymbol{O}_{p-p_1,p_1} & \boldsymbol{O}_{p-p_1,p-p_1}\end{pmatrix} \\
&\quad + \boldsymbol{\Sigma}_*\begin{pmatrix} -\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1/2}\boldsymbol{T}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2}\boldsymbol{\Gamma}'_* & -\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1/2}\boldsymbol{T}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2} \\ \boldsymbol{O}_{p-p_1,p_1} & \boldsymbol{O}_{p-p_1,p-p_1}\end{pmatrix}' \\
&\quad + \boldsymbol{\Sigma}_*\begin{pmatrix} \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1/2}\boldsymbol{T}\boldsymbol{T}'\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1/2} & \boldsymbol{O}_{p_1,p-p_1} \\ \boldsymbol{O}_{p-p_1,p_1} & \boldsymbol{O}_{p-p_1,p-p_1}\end{pmatrix} \\
&= (ii.a) + (ii.b) + (ii.c) + (ii.d)\ (say).
\end{aligned}
$$

From the general formula for the inverse of a block matrix, it is straightforward to observe that

$$
\mathrm{tr}\{(ii.a)\} = p - p_1.
$$

Since $\boldsymbol{T}$ and $\boldsymbol{Y}_1$ are mutually independent, we can calculate the expectations of $\mathrm{tr}\{(ii.b)\}$ and $\mathrm{tr}\{(ii.c)\}$ as follows:

$$
E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(ii.b)\}] = 0,\ E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(ii.c)\}] = 0.
$$

Note that $E^*_{\boldsymbol{Y}}[\boldsymbol{T}\boldsymbol{T}'] = (p - p_1)\boldsymbol{I}_{p_1}$. Then, from the independence of $\boldsymbol{T}$ and $\boldsymbol{Y}_1$, we can calculate $\mathrm{tr}\{(ii.d)\}$ as

$$E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(ii.d)\}] = (p - p_1)E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{\boldsymbol{\Sigma}_{11*}\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1}\right\}\right].$$

Hence, (ii) can be expressed as

$$(ii) = \frac{n^2(p - p_1)}{n - p - 2}\left\{1 + E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\left\{\boldsymbol{\Sigma}_{11*}\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1}\right\}\right]\right\}. \tag{B.6}$$

Now we move on to calculating (iii) in (B.4). Recall that $n\widehat{\boldsymbol{\Sigma}}_{11} = \boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{Y}_1$ and $\widehat{\boldsymbol{\Theta}}_1 = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}_1$. Since it is straightforward to observe that $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}) = \boldsymbol{O}_{q+1,n}$, $n\widehat{\boldsymbol{\Sigma}}_{11}$ and $\widehat{\boldsymbol{\Theta}}_1$ are mutually independent from Cochran's Theorem (e.g., Fujikoshi *et al.*, 2010, Theorem 2.4.2). Thus, by using a property of conditional expectations, we obtain

$$\begin{aligned}
(iii) &= \frac{n}{n - p_1 - q - 2}E^*_{\boldsymbol{Y}}\left[\mathrm{tr}\{\boldsymbol{\Sigma}_{11*}^{-1}(\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_1 - \boldsymbol{Z}\boldsymbol{\Theta}_{1*})'(\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Y}_1 - \boldsymbol{Z}\boldsymbol{\Theta}_{1*})\}\right] \\
&= \frac{n}{n - p_1 - q - 2}\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{Z}})\mathrm{tr}(\boldsymbol{I}_{p_1}) \\
&= \frac{np_1(q + 1)}{n - p_1 - q - 2}.
\end{aligned} \tag{B.7}$$

Finally, we calculate (iv) in (B.4). Recall that $\boldsymbol{\Theta}_*$ is expressed by $\boldsymbol{\alpha}_{1*}$, $\boldsymbol{\alpha}_{2*}$, $\boldsymbol{\mathcal{B}}_{1*}$, and $\boldsymbol{\mathcal{B}}_{2*}$. Then, we can express $(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)'$ as

$$\begin{aligned}
(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)' &= (-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1*} & (\widehat{\boldsymbol{\mathcal{B}}}_1 - \boldsymbol{\mathcal{B}}_{1*})' \\ \widehat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_{2*} & (\widehat{\boldsymbol{\mathcal{B}}}_2 - \boldsymbol{\mathcal{B}}_{2*})' \end{pmatrix} \\
&= \left(\widehat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_{2*} - \widehat{\boldsymbol{\Gamma}}'(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1*}), (\widehat{\boldsymbol{\mathcal{B}}}_2 - \boldsymbol{\mathcal{B}}_{2*})' - \widehat{\boldsymbol{\Gamma}}'(\widehat{\boldsymbol{\mathcal{B}}}_1 - \boldsymbol{\mathcal{B}}_{1*})'\right)'.
\end{aligned} \tag{B.8}$$

From the definitions of $\widehat{\boldsymbol{\delta}}$, $\widehat{\boldsymbol{\alpha}}_1$, and $\widehat{\boldsymbol{\alpha}}_2$, we have $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\alpha}}_2 - \widehat{\boldsymbol{\Gamma}}'\widehat{\boldsymbol{\alpha}}_1$. Thus,

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_{2*} - \widehat{\boldsymbol{\Gamma}}'(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1*}) &= \widehat{\boldsymbol{\alpha}}_2 - \widehat{\boldsymbol{\Gamma}}'\widehat{\boldsymbol{\alpha}}_1 - (\boldsymbol{\alpha}_{2*} - \widehat{\boldsymbol{\Gamma}}'\boldsymbol{\alpha}_{1*}) \\
&= \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_* + \boldsymbol{\alpha}_{2*} - \boldsymbol{\Gamma}_*'\boldsymbol{\alpha}_{1*} - (\boldsymbol{\alpha}_{2*} - \widehat{\boldsymbol{\Gamma}}'\boldsymbol{\alpha}_{1*}) \\
&= \widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_* + (\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_*)'\boldsymbol{\alpha}_{1*} \\
&= \left\{(1, \boldsymbol{\alpha}_{1*}')\begin{pmatrix} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_*)' \\ \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_* \end{pmatrix}\right\}'.
\end{aligned} \tag{B.9}$$

Since it is straightforward that $\widehat{\boldsymbol{\mathcal{B}}}_2 - \widehat{\boldsymbol{\mathcal{B}}}_1\widehat{\boldsymbol{\Gamma}} = \boldsymbol{O}_{q,p-p_1}$ and it is prudent that $\boldsymbol{\mathcal{B}}_{2*} - \boldsymbol{\mathcal{B}}_{1*}\boldsymbol{\Gamma}_* = \boldsymbol{O}_{q,p-p_1}$ under overspecified models, we obtain

$$\begin{aligned}
(\widehat{\boldsymbol{\mathcal{B}}}_2 - \boldsymbol{\mathcal{B}}_{2*}) - (\widehat{\boldsymbol{\mathcal{B}}}_1 - \boldsymbol{\mathcal{B}}_{1*})\widehat{\boldsymbol{\Gamma}} &= \widehat{\boldsymbol{\mathcal{B}}}_2 - \widehat{\boldsymbol{\mathcal{B}}}_1\widehat{\boldsymbol{\Gamma}} - (\boldsymbol{\mathcal{B}}_{2*} - \boldsymbol{\mathcal{B}}_{1*}\boldsymbol{\Gamma}_*) + \boldsymbol{\mathcal{B}}_{1*}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_*) \\
&= (\boldsymbol{0}_q, \boldsymbol{\mathcal{B}}_{1*})\begin{pmatrix} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_*)' \\ \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_* \end{pmatrix}.
\end{aligned} \tag{B.10}$$

By using (B.9) and (B.10), we can express (B.8) as

$$(-\widehat{\boldsymbol{\Gamma}}', \boldsymbol{I}_{p-p_1})(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_*)' = \begin{pmatrix} (\widehat{\boldsymbol{\delta}} - \boldsymbol{\delta}_*)' \\ \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_* \end{pmatrix}'\begin{pmatrix} 1 & \boldsymbol{\alpha}_{1*}' \\ \boldsymbol{0}_q & \boldsymbol{\mathcal{B}}_{1*} \end{pmatrix}'$$

$$= \boldsymbol{\Sigma}_{22\cdot 1*}^{1/2}\boldsymbol{U}'\{(\mathbf{1}_n, \boldsymbol{Y}_1)'(\mathbf{1}_n, \boldsymbol{Y}_1)\}^{-1/2}\begin{pmatrix} 1 & \boldsymbol{\alpha}_{1*}' \\ \mathbf{0}_q & \boldsymbol{\mathcal{B}}_{1*} \end{pmatrix}',$$

where $\boldsymbol{U}$ is defined in (B.3). Thus, by using the above equation and the independence of $\boldsymbol{U}$ and $\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$, (iv) can be expressed as follows:

$$(iv) = \frac{n(p-p_1)}{n-p-2}E_{\boldsymbol{Y}}^*\left[\mathrm{tr}\left\{\{(\mathbf{1}_n, \boldsymbol{Y}_1)'(\mathbf{1}_n, \boldsymbol{Y}_1)\}^{-1}\begin{pmatrix} n & n\boldsymbol{\alpha}_{1*}' \\ n\boldsymbol{\alpha}_{1*} & n\boldsymbol{\alpha}_{1*}\boldsymbol{\alpha}_{1*}' + \boldsymbol{\mathcal{B}}_{1*}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\mathcal{B}}_{1*} \end{pmatrix}\right\}\right].$$

From the general formula for the inverse of a block matrix, $\{(\mathbf{1}_n, \boldsymbol{Y}_1)'(\mathbf{1}_n, \boldsymbol{Y}_1)\}^{-1}$ is expressed as

$$\{(\mathbf{1}_n, \boldsymbol{Y}_1)'(\mathbf{1}_n, \boldsymbol{Y}_1)\}^{-1} = \begin{pmatrix} n^{-1} + \bar{\boldsymbol{y}}_1'\boldsymbol{S}_{11}^{-1}\bar{\boldsymbol{y}}_1 & -\bar{\boldsymbol{y}}_1'\boldsymbol{S}_{11}^{-1} \\ -\boldsymbol{S}_{11}^{-1}\bar{\boldsymbol{y}}_1 & \boldsymbol{S}_{11}^{-1} \end{pmatrix}.$$

Note that $\boldsymbol{S}_{11}$ and $\bar{\boldsymbol{y}}_1$ are mutually independent and $E_{\boldsymbol{Y}}^*[\bar{\boldsymbol{y}}_1] = \boldsymbol{\alpha}_{1*}$. Thus, we can calculate (iv) as follows:

$$\begin{aligned}(iv) &= \frac{n(p-p_1)}{n-p-2}\left[1 + E_{\boldsymbol{Y}}^*\left[\mathrm{tr}\left\{\{n(\bar{\boldsymbol{y}}_1 - \boldsymbol{\alpha}_{1*})(\bar{\boldsymbol{y}}_1 - \boldsymbol{\alpha}_{1*})' + \boldsymbol{\mathcal{B}}_{1*}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\mathcal{B}}_{1*}\}\boldsymbol{S}_{11}^{-1}\right\}\right]\right] \\ &= \frac{n(p-p_1)}{n-p-2}\left[1 + E_{\boldsymbol{Y}}^*\left[\mathrm{tr}\left\{\{\boldsymbol{\Sigma}_{11*} + \boldsymbol{\mathcal{B}}_{1*}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\mathcal{B}}_{1*}\}\boldsymbol{S}_{11}^{-1}\right\}\right]\right]. \qquad \text{(B.11)}\end{aligned}$$

Therefore, from (B.4), (B.5), (B.6), (B.7), and (B.11), Proposition B.1 can be derived. □

B.2 Results for expanding bias and its proof

Proposition B.2. *Suppose that $q \le p_1 < p$ and Assumption A1 both hold. Then, under overspecified models* (7)*, we have*

$$B_{\mathrm{KL}} = m(n,p) + O(p \cdot \max\{n^{-2}, (n+\lambda_q)^{-3/2}\}), \qquad \text{(B.12)}$$

as $n \to \infty$, $p/n \to c \in [0,1)$, where λ_q is the q-th diagonal element of $\boldsymbol{\Lambda}$ defined in (17)*. Here, $m(n,p)$ is the constant given by*

$$m(n,p) = m_1(n,p) + m_2(n,p),$$

where $m_1(n,p)$ is defined in (13)*, and $m_2(n,p)$ is given by*

$$m_2(n,p) = \frac{n^2(p-p_1)}{(n-p)(n-p_1)}\left\{2(q+1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) - n\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 - n\mathrm{tr}(\boldsymbol{\Phi}^{-2})\right\},$$

in which $\boldsymbol{\Phi}$ is defined by (19)*.*

The proof of Proposition B.2 is presented as follows. We use the following lemma (the proof is given in Appendix D).

Lemma B.2. *Suppose that $n - q - 2 > 0$. Let $\widetilde{\boldsymbol{G}}$ be random matrices defined by*

$$\widetilde{\boldsymbol{G}} = \sqrt{n+\lambda_q}(\boldsymbol{G} - \boldsymbol{I}_q),$$

where $\boldsymbol{G}$ *is defined by* (18), λ_q *are given by* (17), *and* $\boldsymbol{\Phi} = (n-1)\boldsymbol{I}_q + \boldsymbol{\Lambda}$ *is defined by* (19). *Then, using the HHD asymptotic framework, we have* $\widetilde{\boldsymbol{G}} = O_p(1)$, *and* $\boldsymbol{G}^{-1}$ *can be expanded as*

$$\boldsymbol{G}^{-1} = \boldsymbol{I}_q - \frac{1}{\sqrt{n+\lambda_q}}\widetilde{\boldsymbol{G}} + \frac{1}{n+\lambda_q}\widetilde{\boldsymbol{G}}^2 + \boldsymbol{R}_{\boldsymbol{G}}, \tag{B.13}$$

where $\boldsymbol{R}_{\boldsymbol{G}} = (n+\lambda_q)^{-3/2}\boldsymbol{G}^{-1}\widetilde{\boldsymbol{G}}^3 = O_p((n+\lambda_q)^{-3/2})$. *Further, when Assumption A1 holds, we have*

$$E[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}\boldsymbol{R}_{\boldsymbol{G}})] = O(n^{-1}(n+\lambda_q)^{-3/2}), \tag{B.14}$$

where $\boldsymbol{D}_{11} = (1+n^{-1})\boldsymbol{I}_q + n^{-1}\boldsymbol{\Lambda}$ *and* $\boldsymbol{\Lambda}$ *is given in* (17).

We focus on the case that $q < p_1 < p$ because the proof is simpler where $p_1 = q$. Let

$$\boldsymbol{V} = \boldsymbol{H}'\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{S}_{11}\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{H}, \tag{B.15}$$

where $\boldsymbol{H}$ is the $p_1 \times p_1$ orthogonal matrix given by (17). It is straightforward that $\boldsymbol{V}$ is distributed according to $W_{p_1}(n-1, \boldsymbol{I}_{p_1}; \widetilde{\boldsymbol{\Lambda}})$, where $\widetilde{\boldsymbol{\Lambda}}$ is given by

$$\widetilde{\boldsymbol{\Lambda}} = \begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{O}_{q,p_1-q} \\ \boldsymbol{O}_{p_1-q,q} & \boldsymbol{O}_{p_1-q,p_1-q} \end{pmatrix}. \tag{B.16}$$

Further, we partition $\boldsymbol{V}$ as

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix},$$

where the sizes of $\boldsymbol{V}_{11}$ and $\boldsymbol{V}_{22}$ are $q \times q$ and $(p_1-q) \times (p_1-q)$, respectively. Let $\widetilde{\boldsymbol{V}}_{11}$ be the $p_1 \times p_1$ random matrix given by

$$\widetilde{\boldsymbol{V}}_{11} = \boldsymbol{\Phi}^{-1/2}\boldsymbol{V}_{11}\boldsymbol{\Phi}^{-1/2}. \tag{B.17}$$

Since $\boldsymbol{\Phi} = E^*_{\boldsymbol{Y}}[\boldsymbol{V}_{11}]$, it is straightforward that $\widetilde{\boldsymbol{V}}_{11}$ is distributed according to $W_q(n-1, \boldsymbol{\Phi}^{-1}; \boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1})$. Now, we calculate the parts of the expectations in (B.1). Using the definitions in (17) and (B.15), we have

$$\frac{n(n+1)(p-p_1)}{n-p-2}E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{S}_{11}^{-1}\boldsymbol{\Sigma}_{11*})] + \frac{n(p-p_1)}{n-p-2}E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{S}_{11}^{-1}\boldsymbol{\Xi}_*)] = \frac{n^2(p-p_1)}{n-p-2}E^*_{\boldsymbol{Y}}\left[\mathrm{tr}(\boldsymbol{D}\boldsymbol{V}^{-1})\right], \tag{B.18}$$

where $\boldsymbol{D}$ and the partitioned expression are given by

$$\boldsymbol{D} = (1+n^{-1})\boldsymbol{I}_{p_1} + n^{-1}\widetilde{\boldsymbol{\Lambda}} = \begin{pmatrix} \boldsymbol{D}_{11} & \boldsymbol{O}_{q,p_1-q} \\ \boldsymbol{O}_{p_1-q,q} & \boldsymbol{D}_{22} \end{pmatrix}.$$

By applying the general formula for the inverse of a block matrix to $\boldsymbol{V}$, $\mathrm{tr}(\boldsymbol{D}\boldsymbol{V}^{-1})$ can be expressed as

$$\mathrm{tr}(\boldsymbol{D}\boldsymbol{V}^{-1}) = \mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1}) + \mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}\boldsymbol{V}_{22\cdot 1}^{-1}\boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1}) + \mathrm{tr}(\boldsymbol{D}_{22}\boldsymbol{V}_{22}^{-1}).$$

Recall that $\boldsymbol{V} \sim W_{p_1}(n-1, \boldsymbol{I}_{p_1}; \widetilde{\boldsymbol{\Lambda}})$. By using the properties of a partitioned non-central Wishart matrix (see Kabe, 1964), we can state that $\boldsymbol{V}_{11} \sim W_q(n-1, \boldsymbol{I}_q; \boldsymbol{\Lambda})$, $\boldsymbol{V}_{22\cdot 1} \sim W_{p_1-q}(n-q-$

$1, \boldsymbol{I}_{p_1-q})$, $\boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1/2} \sim N_{(p_1-q)\times p_1}(\boldsymbol{O}_{p_1-q,p_1}, \boldsymbol{I}_{p_1} \otimes \boldsymbol{I}_{p_1-q})$, and $\boldsymbol{V}_{11}$, $\boldsymbol{V}_{22\cdot 1}$ and $\boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1/2}$ are mutually independent. Thus, we can calculate the expectations of $\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}\boldsymbol{V}_{22\cdot 1}^{-1}\boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1})$ and $\mathrm{tr}(\boldsymbol{D}_{22}\boldsymbol{V}_{22}^{-1})$ as follows:

$$\begin{aligned} E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1}\boldsymbol{V}_{12}\boldsymbol{V}_{22\cdot 1}^{-1}\boldsymbol{V}_{21}\boldsymbol{V}_{11}^{-1})] &= \frac{p_1-2}{n-p_1-2}E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1})], \\ E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{22}\boldsymbol{V}_{22}^{-1})] &= \frac{1}{n-p_1-2}\mathrm{tr}(\boldsymbol{D}_{22}). \end{aligned}$$

Hence, from the above equations, $E_{\boldsymbol{Y}}^*\left[\mathrm{tr}(\boldsymbol{D}\boldsymbol{V}^{-1})\right]$ is expressed as

$$E_{\boldsymbol{Y}}^*\left[\mathrm{tr}(\boldsymbol{D}\boldsymbol{V}^{-1})\right] = \frac{n-q-2}{n-p_1-2}E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1})] + \frac{1}{n-p_1-2}\mathrm{tr}(\boldsymbol{D}_{22}). \tag{B.19}$$

From the definition of $\widetilde{\boldsymbol{V}}_{11}$ in (B.17), $E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1})]$ is expressed as $E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}\widetilde{\boldsymbol{V}}_{11}^{-1})]$. Let

$$\boldsymbol{L} = \sqrt{n+\lambda_q}(\widetilde{\boldsymbol{V}}_{11} - \boldsymbol{I}_q).$$

Then, by using (B.13) and (B.14) in Lemma B.2, we can expand $E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1})]$ as

$$E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1})] = \mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}) + \frac{1}{n+\lambda_q}E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^2)] + O(n^{-1}(n+\lambda_q)^{-3/2}). \tag{B.20}$$

From the expectation of a non-central Wishart distribution, $(n+\lambda_q)^{-1}E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^2)]$ can be calculated as

$$\begin{aligned} &\frac{1}{n+\lambda_q}E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}\boldsymbol{L}^2)] \\ &= -2\left(1-\frac{1}{n}\right)\mathrm{tr}(\boldsymbol{\Phi}^{-3}) - \left(1-\frac{5}{n}\right)\mathrm{tr}(\boldsymbol{\Phi}^{-2}) - 2\left(1-\frac{1}{n}\right)\mathrm{tr}(\boldsymbol{\Phi}^{-1})\mathrm{tr}(\boldsymbol{\Phi}^{-2}) \\ &\quad - \left(1-\frac{5}{n}\right)\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + \frac{2(q+1)}{n}\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + \frac{q}{n} - \mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{\Phi}^{-1}). \end{aligned} \tag{B.21}$$

From (B.21), we can express (B.20) as

$$E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{D}_{11}\boldsymbol{V}_{11}^{-1})] = -\mathrm{tr}(\boldsymbol{\Phi}^{-2}) - \mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + \frac{2(q+1)}{n}\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + \frac{q}{n} + O(n^{-1}(n+\lambda_q)^{-3/2}). \tag{B.22}$$

Thus, by using (B.18), (B.19), (B.22), and Proposition B.1, we can expand the bias as follows:

$$\begin{aligned} B_{\mathrm{KL}} &= -np + \frac{np_1(n+q+1)}{n-p_1-q-2} + \frac{n(n+1)(p-p_1)}{n-p-2} \\ &\quad + \frac{n^2(n-q-2)(p-p_1)}{(n-p-2)(n-p_1-2)}\left\{-\mathrm{tr}(\boldsymbol{\Phi}^{-2}) - \mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + \frac{2(q+1)}{n}\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + \frac{q}{n}\right\} \\ &\quad + \frac{n^2(p-p_1)}{(n-p-2)(n-p_1-2)}\left(1+\frac{1}{n}\right)(p_1-q) + O(p(n+\lambda_q)^{-3/2}). \end{aligned}$$

Note that

$$\begin{aligned} \frac{1}{n-p_1-q-2} &= \frac{1}{n-p_1}\left\{1 + \frac{q+2}{n-p_1} + \frac{(q+2)^2}{(n-p_1)^2} + O(n^{-3})\right\}, \\ \frac{1}{n-p-2} &= \frac{1}{n-p}\left\{1 + \frac{2}{n-p} + \frac{4}{(n-p)^2} + O(n^{-3})\right\}, \end{aligned}$$

$$\frac{1}{n-p_1-2} = \frac{1}{n-p_1}\left\{1+\frac{2}{n-p_1}+\frac{4}{(n-p_1)^2}+O(n^{-3})\right\},$$

$$\begin{aligned}\frac{1}{(n-p-2)(n-p_1-2)} = \frac{1}{(n-p)(n-p_1)}\Big\{&1+\frac{2}{n-p}+\frac{2}{n-p_1}+\frac{4}{(n-p)^2}+\frac{4}{(n-p_1)^2}\\&+\frac{4}{(n-p)(n-p_1)}+\frac{8}{(n-p)^3}+\frac{8}{(n-p_1)^3}\\&+\frac{8}{(n-p)^2(n-p_1)}+\frac{8}{(n-p)(n-p_1)^2}+O(n^{-4})\Big\}.\end{aligned}$$

Therefore, by using the above equations, we can derive (B.12). □

C Asymptotic properties of $\widehat{\tau}_1$, $\widehat{\tau}_2$ and $\widehat{\tau}_3$

Here we present the results for expanding $\widehat{\tau}_1$, $\widehat{\tau}_2$ and $\widehat{\tau}_3$ defined by (14) and (15).

C.1 Expanding $\widehat{\tau}_1$, $\widehat{\tau}_2$, and $\widehat{\tau}_3$ and the proof

Proposition C.1. *Suppose that $q \leq p_1 < p$. Then, we have*

$$\begin{aligned}\widehat{\tau}_1 &= (p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + O_p(pn^{-1/2}(n+\lambda_q)^{-1}),\\ \widehat{\tau}_2 &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + O_p(pn^{1/2}(n+\lambda_q)^{-2}),\\ \widehat{\tau}_3 &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-2}) + O_p(pn^{1/2}(n+\lambda_q)^{-2}),\end{aligned}$$

as $n \to \infty$, $p/n \to c \in [0,1)$, where $\boldsymbol{\Phi}$ is defined by (19).

The proof of Proposition C.1 is presented as follows. To expand $\widehat{\tau}_1$, $\widehat{\tau}_2$ and $\widehat{\tau}_3$ by using a HHD asymptotic framework, we use the following lemma which was essentially obtained in Wakaki *et al.* (2014):

Lemma C.1. *Let $\boldsymbol{F}_h = \boldsymbol{F}'\boldsymbol{F}$ and $\boldsymbol{F}_e$ be independently distributed according to $W_p(q, \boldsymbol{I}_p; \boldsymbol{M}'\boldsymbol{M})$ and $W_p(n, \boldsymbol{I}_p)$, respectively. Here, $\boldsymbol{F}$ is a $q\times p$ random matrix distributed according to $N_{q\times p}(\boldsymbol{M}, \boldsymbol{I}_p \otimes \boldsymbol{I}_q)$. Put*

$$\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}',\ \boldsymbol{W} = \boldsymbol{B}^{1/2}(\boldsymbol{F}\boldsymbol{F}_e^{-1}\boldsymbol{F}')^{-1}\boldsymbol{B}^{1/2}.$$

Then, $\boldsymbol{B}$ and $\boldsymbol{W}$ are independently distributed according to $W_q(p, \boldsymbol{I}_q; \boldsymbol{M}\boldsymbol{M}')$ and $W_q(n-p+q, \boldsymbol{I}_q)$, respectively. Further, the nonzero eigenvalues of $\boldsymbol{F}_h\boldsymbol{F}_e^{-1}$ are the same as those of $\boldsymbol{B}\boldsymbol{W}^{-1}$, and we have

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{F}_e(\boldsymbol{F}_e+\boldsymbol{F}_h)^{-1}\} &= \mathrm{tr}\{\boldsymbol{W}(\boldsymbol{W}+\boldsymbol{B})^{-1}\} + (p-q),\\ \mathrm{tr}\left[\{\boldsymbol{F}_e(\boldsymbol{F}_e+\boldsymbol{F}_h)^{-1}\}^2\right] &= \mathrm{tr}\left[\{\boldsymbol{W}(\boldsymbol{W}+\boldsymbol{B})^{-1}\}^2\right] + (p-q).\end{aligned}$$

We consider the distributions of $\boldsymbol{S}_e$ and $\boldsymbol{S}_h$ defined in (15). Since $\boldsymbol{Y}_1$ is distributed according to $N_{n\times p_1}(\boldsymbol{Z}\boldsymbol{\Theta}_{1*}, \boldsymbol{\Sigma}_{11*} \otimes \boldsymbol{I}_n)$, it is straightforward that $\boldsymbol{S}_e$ and $\boldsymbol{S}_h$ are mutually independent and

$$\boldsymbol{S}_e \sim W_{p_1}(n-q-1, \boldsymbol{\Sigma}_{11*}),\ \boldsymbol{S}_h \sim W_{p_1}(q, \boldsymbol{\Sigma}_{11*}; \boldsymbol{\Xi}_*),$$

where $\boldsymbol{\Xi}_*$ is given by (16). Let

$$\boldsymbol{T}_e = \boldsymbol{H}'\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{S}_e\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{H},\ \boldsymbol{T}_h = \boldsymbol{H}'\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{S}_h\boldsymbol{\Sigma}_{11*}^{-1/2}\boldsymbol{H},$$

where $\boldsymbol{H}$ is the $p_1 \times p_1$ orthogonal matrix defined in (17). It follows from the properties of non-central Wishart matrices that

$$\boldsymbol{T}_e \sim W_{p_1}(n-q-1, \boldsymbol{I}_{p_1}),\ \boldsymbol{T}_h \sim W_{p_1}(q, \boldsymbol{I}_{p_1}; \widetilde{\boldsymbol{\Lambda}}),$$

where $\widetilde{\boldsymbol{\Lambda}}$ is defined by (B.16). Then, we can express $\mathrm{tr}\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}$ and $\mathrm{tr}[\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}^2]$ as

$$\mathrm{tr}\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\} = \mathrm{tr}\{\boldsymbol{T}_e(\boldsymbol{T}_e+\boldsymbol{T}_h)^{-1}\},\ \mathrm{tr}[\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}^2] = \mathrm{tr}\left[\{\boldsymbol{T}_e(\boldsymbol{T}_e+\boldsymbol{T}_h)^{-1}\}^2\right].$$

First, we express $\mathrm{tr}\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}$ and $\mathrm{tr}[\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}^2]$ as functions of $q \times q$ matrices in order to examine their asymptotic behaviors. From the definition of the non-central Wishart distribution, a different expression for $\boldsymbol{T}_h$ is given by $\boldsymbol{T}_h = \boldsymbol{U}_h'\boldsymbol{U}_h$, where $\boldsymbol{U}_h \sim N_{q\times p_1}(\widetilde{\boldsymbol{\Lambda}}_1, \boldsymbol{I}_{p_1}\otimes\boldsymbol{I}_q)$ and $\tilde{\boldsymbol{\Lambda}}_1$ is the $q \times p_1$ partitioned matrix of $\widetilde{\boldsymbol{\Lambda}} = (\widetilde{\boldsymbol{\Lambda}}_1', \boldsymbol{O}_{p_1,p_1-q})'$ satisfying $\widetilde{\boldsymbol{\Lambda}}_1'\widetilde{\boldsymbol{\Lambda}}_1 = \boldsymbol{\Lambda}$. Let

$$\boldsymbol{W}_1 = \boldsymbol{U}_h\boldsymbol{U}_h',\ \boldsymbol{W}_2 = \boldsymbol{W}_1^{1/2}(\boldsymbol{U}_h\boldsymbol{T}_e^{-1}\boldsymbol{U}_h')^{-1}\boldsymbol{W}_1^{1/2}. \tag{C.1}$$

Then, from Lemma C.1, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are independently distributed according to $W_q(p_1, \boldsymbol{I}_q; \boldsymbol{\Lambda})$ and $W_q(n-p_1-1, \boldsymbol{I}_q)$, respectively, and we have

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\} &= \mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\} + (p_1-q),\\ \mathrm{tr}[\{\boldsymbol{S}_e(\boldsymbol{S}_e+\boldsymbol{S}_h)^{-1}\}^2] &= \mathrm{tr}\left[\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2\right] + (p_1-q).\end{aligned}$$

Next, we expand $\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}$ and $\mathrm{tr}\left[\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2\right]$. Let

$$\boldsymbol{W}_3 = \boldsymbol{\Phi}^{-1/2}(\boldsymbol{W}_1+\boldsymbol{W}_2)\boldsymbol{\Phi}^{-1/2}. \tag{C.2}$$

Since $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are mutually independent, we can state that $\boldsymbol{W}_3 \sim W_q(n-1, \boldsymbol{\Phi}^{-1}; \boldsymbol{\Lambda}\boldsymbol{\Phi}^{-1})$. Therefore, using (B.13) in Lemma B.2, $\boldsymbol{W}_3^{-1}$ can be expanded as

$$\boldsymbol{W}_3^{-1} = \boldsymbol{I}_q + O_p((n+\lambda_q)^{-1/2}). \tag{C.3}$$

By calculating the expectation and variance of $\boldsymbol{W}_2$, we can also expand $\boldsymbol{W}_2$ as

$$\boldsymbol{W}_2 = (n-p_1-1)\boldsymbol{I}_q + O_p(n^{1/2}).$$

Thus, $\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}$ is expressed as

$$\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2} = (n-p_1)\boldsymbol{\Phi}^{-1} + O_p(n^{1/2}(n+\lambda_q)^{-1}). \tag{C.4}$$

From (C.3) and (C.4), the following equations can be derived:

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\} &= \mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_3^{-1})\\ &= (n-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + O_p(n^{1/2}(n+\lambda_q)^{-1}),\\ \mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2 &= (n-p_1)^2\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + O_p(n^{3/2}(n+\lambda_q)^{-2}),\end{aligned}$$

$$\begin{aligned}\mathrm{tr}\left[\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2\right] &= \mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_3^{-1})^2\} \\ &= (n-p_1)^2\mathrm{tr}(\boldsymbol{\Phi}^{-2}) + O_p(n^{3/2}(n+\lambda_q)^{-2}).\end{aligned}$$

Thus, we can derive the following equations:

$$\begin{aligned}\frac{p-p_1}{n-p_1}\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\} &= (p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + O_p(pn^{-1/2}(n+\lambda_q)^{-1}), \\ \frac{n(p-p_1)}{(n-p_1)^2}\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2 &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + O_p(pn^{1/2}(n+\lambda_q)^{-2}), \\ \frac{n(p-p_1)}{(n-p_1)^2}\mathrm{tr}\left[\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2\right] &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-2}) + O_p(pn^{1/2}(n+\lambda_q)^{-2}).\end{aligned}$$

Therefore, we can derive the result of Proposition C.1. □

C.2 Expanding the expectations of $\widehat{\tau}_1$, $\widehat{\tau}_2$, and $\widehat{\tau}_3$

Proposition C.2. *Suppose that $q \le p_1 < p$ and Assumptions A1 and A2 hold. Then, we have*

$$\begin{aligned}E_{\boldsymbol{Y}}^*[\widehat{\tau}_1] &= (p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + O(pn^{-1}(n+\lambda_q)^{-1/2}), \\ E_{\boldsymbol{Y}}^*[\widehat{\tau}_2] &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + O(pn^{-1}(n+\lambda_q)^{-1/2}), \\ E_{\boldsymbol{Y}}^*[\widehat{\tau}_3] &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-2}) + O(pn^{-1}(n+\lambda_q)^{-1/2}),\end{aligned}$$

as $n \to \infty$, $p/n \to c \in [0,1)$, where $\boldsymbol{\Phi}$ is defined by (19).

The proof of Proposition C.2 is presented as follows. Let $\widetilde{\boldsymbol{W}}_3 = \sqrt{n+\lambda_q}(\boldsymbol{W}_3 - \boldsymbol{I}_q)$, where $\boldsymbol{W}_3$ is given by (C.2). Then, from Lemma B.2, we can observe that $\boldsymbol{W}_3^{-1}$ is expanded as $\boldsymbol{W}_3^{-1} = \boldsymbol{I}_q - \boldsymbol{R}_{\boldsymbol{W}_3}$, where $\boldsymbol{R}_{\boldsymbol{W}_3} = (n+\lambda_q)^{-1/2}\boldsymbol{W}_3^{-1}\widetilde{\boldsymbol{W}}_3$. We consider $E_{\boldsymbol{Y}}^*[\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}]$, $E_{\boldsymbol{Y}}^*[\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2]$ and $E_{\boldsymbol{Y}}^*[\mathrm{tr}[\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2]]$, where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are given by (C.1). By using $\boldsymbol{R}_{\boldsymbol{W}_3}$, we have

$$E_{\boldsymbol{Y}}^*[\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}] = E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})] - E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})], \tag{C.5}$$

$$\begin{aligned}E_{\boldsymbol{Y}}^*[\mathrm{tr}\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2] &= E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^2] \\ &\quad - 2E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})] \\ &\quad + E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})^2], \end{aligned}\tag{C.6}$$

$$\begin{aligned}E_{\boldsymbol{Y}}^*[\mathrm{tr}[\{\boldsymbol{W}_2(\boldsymbol{W}_1+\boldsymbol{W}_2)^{-1}\}^2]] &= E_{\boldsymbol{Y}}^*[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^2\}] - 2E_{\boldsymbol{Y}}^*[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^2\boldsymbol{R}_{\boldsymbol{W}_3}\}] \\ &\quad + E_{\boldsymbol{Y}}^*[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})^2\}]. \end{aligned}\tag{C.7}$$

Since $\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}$ is distributed according to $W_q(n-p_1-1, \boldsymbol{\Phi}^{-1})$, we can calculate the expectation as follows:

$$E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})] = (n-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + O((n+\lambda_q)^{-1}), \tag{C.8}$$

$$E_{\boldsymbol{Y}}^*[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^2] = (n-p_1)^2\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + O(n(n+\lambda_q)^{-2}), \tag{C.9}$$

$$E_{\boldsymbol{Y}}^*[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^2\}] = (n-p_1)^2\mathrm{tr}(\boldsymbol{\Phi}^{-2}) + O(n(n+\lambda_q)^{-2}). \tag{C.10}$$

By using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})] \\
&= \frac{1}{\sqrt{n+\lambda_q}} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_3^{-1}\widetilde{\boldsymbol{W}}_3)] \\
&\le \frac{1}{\sqrt{n+\lambda_q}} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^4\}]^{1/2} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_3^4)]^{1/2} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_3^{-2})]^{1/2} \\
&= O((n+\lambda_q)^{-1/2}), \qquad \text{(C.11)} \\
&E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})] \\
&\le \frac{1}{\sqrt{n+\lambda_q}} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^8\}]^{1/4} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_3^4)]^{1/4} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_3^{-2})]^{1/2} \\
&= O((n+\lambda_q)^{-1/2}), \qquad \text{(C.12)} \\
&E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})^2] \\
&\le \frac{1}{n+\lambda_q} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^4\}^2]^{1/4} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\widetilde{\boldsymbol{W}}_3^4)\}^2]^{1/4} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{W}_3^{-2})\}^2]^{1/2} \\
&= O((n+\lambda_q)^{-1/2}), \qquad \text{(C.13)} \\
&E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^2\boldsymbol{R}_{\boldsymbol{W}_3}\}] \\
&\le \frac{1}{\sqrt{n+\lambda_q}} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^4\}]^{1/2} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_3^8)]^{1/4} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_3^{-4})]^{1/4} \\
&= O((n+\lambda_q)^{-1/2}), \qquad \text{(C.14)} \\
&E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2}\boldsymbol{R}_{\boldsymbol{W}_3})^2\}] \\
&\le \frac{1}{(n+\lambda_q)^2} E^*_{\boldsymbol{Y}}[\mathrm{tr}\{(\boldsymbol{\Phi}^{-1/2}\boldsymbol{W}_2\boldsymbol{\Phi}^{-1/2})^4\}]^{1/2} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\widetilde{\boldsymbol{W}}_3^8)]^{1/2} E^*_{\boldsymbol{Y}}[\mathrm{tr}(\boldsymbol{W}_3^{-8})]^{1/2} \\
&= O((n+\lambda_q)^{-1/2}). \qquad \text{(C.15)}
\end{aligned}$$

Therefore, from (C.5)-(C.15), we have

$$\begin{aligned}
E^*_{\boldsymbol{Y}}[\widehat{\tau}_1] &= (p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1}) + O(pn^{-1}(n+\lambda_q)^{-1/2}), \\
E^*_{\boldsymbol{Y}}[\widehat{\tau}_2] &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + O(pn^{-1}(n+\lambda_q)^{-1/2}), \\
E^*_{\boldsymbol{Y}}[\widehat{\tau}_3] &= n(p-p_1)\mathrm{tr}(\boldsymbol{\Phi}^{-2}) + O(pn^{-1}(n+\lambda_q)^{-1/2}).
\end{aligned}$$

□

D Proofs of Lemma B.1 and B.2

D.1 Proof of Lemma B.1

From a property of a conditional distribution of a multivariate normal distribution (e.g., Srivastava and Khatri, 1979), we have

$$\boldsymbol{Y}_2|\boldsymbol{Y}_1 \sim N_{n\times(p-p_1)}(\boldsymbol{\Omega}_* + \boldsymbol{Y}_1\boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_{22\cdot 1*} \otimes \boldsymbol{I}_n).$$

Then, another expression of the true model M_* is given as follows:

$$\boldsymbol{Y}_1 \sim N_{n\times p_1}(\boldsymbol{\Delta}_{1*}, \boldsymbol{\Sigma}_{11*} \otimes \boldsymbol{I}_n),\ \boldsymbol{Y}_2|\boldsymbol{Y}_1 \sim N_{n\times(p-p_1)}(\boldsymbol{\Omega}_* + \boldsymbol{Y}_1\boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_{22\cdot 1*} \otimes \boldsymbol{I}_n).$$

On the other hand, we can express $\boldsymbol{Y}_2$ as

$$\boldsymbol{Y}_2 = \boldsymbol{\Omega}_* + \boldsymbol{Y}_1\boldsymbol{\Gamma}_* + \boldsymbol{A}\boldsymbol{\Sigma}_{22\cdot 1*}^{1/2}, \tag{D.1}$$

where $\boldsymbol{A}$ and $\boldsymbol{Y}_1$ are mutually independent, and $\boldsymbol{A}$ is distributed according to $N_{n\times(p-p_1)}(\boldsymbol{O}_{n,p-p_1}, \boldsymbol{I}_{p-p_1} \otimes \boldsymbol{I}_n)$. By using (D.1), $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$, $\widehat{\boldsymbol{\Gamma}}$, and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ are expressed as

$$\begin{aligned}
n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1} &= (\boldsymbol{\Omega}_* + \boldsymbol{A}\boldsymbol{\Sigma}_{22\cdot 1*}^{1/2})'(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{1}_n,\boldsymbol{Y}_1)})(\boldsymbol{\Omega}_* + \boldsymbol{A}\boldsymbol{\Sigma}_{22\cdot 1*}^{1/2}). \qquad \text{(D.2)}\\
\widehat{\boldsymbol{\Gamma}} &= \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1}\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{\Omega}_* + \boldsymbol{A}\boldsymbol{\Sigma}_{22\cdot 1*}^{1/2}) + \boldsymbol{\Gamma}_*,\\
\begin{pmatrix} \widehat{\boldsymbol{\delta}}' \\ \widehat{\boldsymbol{\Gamma}} \end{pmatrix} &= \{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{-1}(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{\Omega}_* + \boldsymbol{Y}_1\boldsymbol{\Gamma}_* + \boldsymbol{A}\boldsymbol{\Sigma}_{22\cdot 1*}^{1/2}).
\end{aligned}$$

Hence, the conditional distribution of $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$, $\widehat{\boldsymbol{\Gamma}}$, and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ under $\boldsymbol{Y}_1$ are

$$\begin{aligned}
&n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}|\boldsymbol{Y}_1 \sim W_{p-p_1}(n - p_1 - 1, \boldsymbol{\Sigma}_{22\cdot 1*}; \widetilde{\boldsymbol{\Omega}}_*),\\
&\widehat{\boldsymbol{\Gamma}}|\boldsymbol{Y}_1 \sim N_{p_1\times(p-p_1)}(\{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1}\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\Omega}_* + \boldsymbol{\Gamma}_*, \boldsymbol{\Sigma}_{22\cdot 1*} \otimes \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{-1}),\\
&\begin{pmatrix} \widehat{\boldsymbol{\delta}}' \\ \widehat{\boldsymbol{\Gamma}} \end{pmatrix} |\boldsymbol{Y}_1\\
&\sim N_{(p_1+1)\times(p-p_1)}(\{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{-1}(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{\Omega}_* + \boldsymbol{Y}_1\boldsymbol{\Gamma}_*), \boldsymbol{\Sigma}_{22\cdot 1*} \otimes \{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{-1}).
\end{aligned}$$

From the above equations, we can state that $\boldsymbol{T}$ and $\boldsymbol{U}$ are distributed according to $N_{p_1\times(p-p_1)}$ $(\boldsymbol{O}_{p,p-p_1}, \boldsymbol{I}_{p-p_1} \otimes \boldsymbol{I}_{p_1})$ and $N_{(p_1+1)\times(p-p_1)}(\boldsymbol{O}_{p_1+1,p-p_1}, \boldsymbol{I}_{p-p_1} \otimes \boldsymbol{I}_{p_1+1})$, respectively, and are independent of $\boldsymbol{Y}_1$.

Next, we show that $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ are mutually conditionally independent under $\boldsymbol{Y}_1$. It is straightforward that

$$\{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{-1}(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{1}_n,\boldsymbol{Y}_1)}) = \boldsymbol{O}_{p_1+1,n}. \tag{D.3}$$

Hence, from Cochran's Theorem, the conditional independence of $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ is obtained.

Finally, we consider the case that the model is overspecified. By the definition of overspecified models in (7), the equation $\boldsymbol{\Omega}_* = \boldsymbol{1}_n\boldsymbol{\delta}_*'$ holds. Thus, from (D.2), we can state that

$$n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1} = \boldsymbol{\Sigma}_{22\cdot 1*}^{1/2}\boldsymbol{A}'(\boldsymbol{I}_n - \boldsymbol{P}_{(\boldsymbol{1}_n,\boldsymbol{Y}_1)})\boldsymbol{A}\boldsymbol{\Sigma}_{22\cdot 1*}^{1/2} \sim W_{p-p_1}(n - p_1 - 1, \boldsymbol{\Sigma}_{22\cdot 1*}),$$

and $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $\boldsymbol{Y}_1$ are mutually independent. Also, $\boldsymbol{T}$ and $\boldsymbol{U}$ are expressed as

$$\begin{aligned}
\boldsymbol{T} &= \{\boldsymbol{Y}_1'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_1\}^{1/2}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_*)\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2},\\
\boldsymbol{U} &= \{(\boldsymbol{1}_n, \boldsymbol{Y}_1)'(\boldsymbol{1}_n, \boldsymbol{Y}_1)\}^{1/2}\begin{pmatrix} \widehat{\boldsymbol{\delta}}' - \boldsymbol{\delta}_*' \\ \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_* \end{pmatrix}\boldsymbol{\Sigma}_{22\cdot 1*}^{-1/2},
\end{aligned}$$

and $n\widehat{\boldsymbol{\Sigma}}_{22\cdot 1}$ and $(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\Gamma}}')'$ are mutually independent from (D.3). □

D.2 Proof of Lemma B.2

This proof is based on an idea in Hashiyama *et al.* (2014). From the expectation of a non-central Wishart distribution, we can calculate the expectations as $E[\boldsymbol{G}] = \boldsymbol{I}_q$ and

$$E\left[||\boldsymbol{G} - \boldsymbol{I}_q||^2\right] = -(n-1)\mathrm{tr}(\boldsymbol{\Phi}^{-2}) - (n-1)\mathrm{tr}(\boldsymbol{\Phi}^{-1})^2 + 2(q+1)\mathrm{tr}(\boldsymbol{\Phi}^{-1})$$

$$= O((n + \lambda_q)^{-1}).$$

The above equation implies that $\widetilde{\boldsymbol{G}} = O_p(1)$. By applying Taylor expansion to $\boldsymbol{G}^{-1}$, we derive

$$\boldsymbol{G}^{-1} = \left(\boldsymbol{I}_q + \frac{1}{\sqrt{n + \lambda_q}} \widetilde{\boldsymbol{G}} \right)^{-1} = \boldsymbol{I}_q - \frac{1}{\sqrt{n + \lambda_q}} \widetilde{\boldsymbol{G}} + \frac{1}{n + \lambda_q} \widetilde{\boldsymbol{G}}^2 + \boldsymbol{R}_{\boldsymbol{G}}, \tag{D.4}$$

where $\boldsymbol{R}_{\boldsymbol{G}}$ is a remainder term in the Taylor expansion. First, we calculate an explicit form of $\boldsymbol{R}_{\boldsymbol{G}}$. From (D.4), the following equation can be derived:

$$\begin{aligned} \boldsymbol{I}_q &= \boldsymbol{G}\boldsymbol{G}^{-1} \\ &= \left(\boldsymbol{I}_q + \frac{1}{\sqrt{n + \lambda_q}} \widetilde{\boldsymbol{G}} \right) \left(\boldsymbol{I}_q - \frac{1}{\sqrt{n + \lambda_q}} \widetilde{\boldsymbol{G}} + \frac{1}{n + \lambda_q} \widetilde{\boldsymbol{G}}^2 + \boldsymbol{R}_{\boldsymbol{G}} \right) \\ &= \boldsymbol{I}_q + \boldsymbol{G}\boldsymbol{R}_{\boldsymbol{G}} - \frac{1}{(n + \lambda_q)^{3/2}} \widetilde{\boldsymbol{G}}^3. \end{aligned}$$

From the above equation, $\boldsymbol{R}_{\boldsymbol{G}}$ can be expressed as

$$\boldsymbol{R}_{\boldsymbol{G}} = \frac{1}{(n + \lambda_q)^{3/2}} \boldsymbol{G}^{-1} \widetilde{\boldsymbol{G}}^3.$$

Next, we show (B.14). By using the Cauchy-Schwarz inequality, we have

$$E[|\mathrm{tr}(\boldsymbol{D}_{11} \boldsymbol{\Phi}^{-1} \boldsymbol{R}_{\boldsymbol{G}})|] \leq (n + \lambda_q)^{-3/2} \left\{ E[\mathrm{tr}(\boldsymbol{D}_{11}^2 \boldsymbol{\Phi}^{-2} \widetilde{\boldsymbol{G}}^6)] E[\mathrm{tr}(\boldsymbol{G}^{-2})] \right\}^{1/2}.$$

We can observe that $\boldsymbol{D}_{11} \boldsymbol{\Phi}^{-1} = O(n^{-1})$ because

$$||n \boldsymbol{D}_{11} \boldsymbol{\Phi}^{-1}||^2 = \sum_{i=1}^{q} \left(\frac{n + 1 + \lambda_i}{n - 1 + \lambda_i} \right)^2 = O(1).$$

By using results for expectations of the multivariate normal random vector (e.g., Gupta and Nagar, 2000), we can calculate $E[\mathrm{tr}(\boldsymbol{D}_{11}^2 \boldsymbol{\Phi}^{-2} \widetilde{\boldsymbol{G}}^6)] = O(n^{-2})$. Therefore, we derive (B.14). □

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), pp. 995–1010. Akadémiai Kiadó, Budapest.

[2] Alves, J. C. L. & Poppi, R. J. (2016). Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration. *Fuel*, **165**, 379–388.

[3] Bedrick, E. J. & Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, **50**, 226–231.

[4] Bro, R. (2003). Multivariate calibration: What is in chemometrics for the analytical chemist? *Anal. Chim. Acta*, **500**, 185–194.

[5] Brown, P. J. (1982). Multivariate calibration. *J. R. Statist. Soc.*, **B**, **44** 287–321.

[6] Carvalho, B. M. A., Carvalho, L. M., Coimbra, J. S. R., Minim, L. A., Barcellos, E. S., Júnior, W. F. S., Detmann, E. & Carvalho, G. G. P. (2015). Rapid detection of whey in milk powder samples by spectrophotometric and multivariate calibration. *Food Chem.*, **174**, 1–7.

[7] Fujikoshi, Y. and Nishii, R. (1986). Selection of variables in multivariate regression. *Hiroshima Math. J.*, **13**, 269–277.

[8] Fujikoshi, Y., Sakurai, T. & Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *J. Multivariate Anal.*, **123**, 184–200.

[9] Fujikoshi, Y., Ulyanov, V. V. & Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* John Wiley & Sons, Inc., Hoboken, New Jersey.

[10] Gupta, A. & Nagar, D. (2000). *Matrix Variate Distributions.* Chapman and Hall/CRC, Boca Raton, FL.

[11] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York.

[12] Hashiyama, Y., Yanagihara, Y. & Fujikoshi, Y. (2014). Jackknife bias correction of the AIC for selecting variables in canonical correlation analysis under model misspecification. *Linear Algebra Appl.*, **455**, 82–106.

[13] Hurvich, C. M. & Tsai, C.-L. (1999). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

[14] Kabe, D. G. (1964). A note on the Bartlett decomposition of a Wishart matrix. *J. Roy. Statist. Soc. Ser. B*, **26**, 270–273.

[15] Lütkepohl, H. (1997). *Handbook of Matrices.* Wiley, Chichester.

[16] Srivastava, M. S. & Khatri, C. G. (1979). *An Introduction to Multivariate Statistics.* North-Holland, New York.

[17] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Comm. Statist. Theory Methods,* **A7**, 13–26.

[18] Wakaki, H., Fujikoshi, Y. & Ulyanov, V. V. (2014). Asymptotic expansions of the distributions of MANOVA test statistics when the dimension is large. *Hiroshima Math. J.*, **44**, 247–259.

[19] Watamori, Y. (1990). On the moments of traces of Wishart and inverted Wishart matrices. *South African Statist. J.*, **24**, 153–176.