TR-No. 20-03, Hiroshima Statistical Research Group, 1–22

High-dimensionality-adjusted Asymptotically Loss and Mean Efficient GC_p Criterion for Normal Multivariate Linear Regression Models

Hirokazu Yanagihara*

Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan

(Last Modified: February 15, 2020)

Abstract

A variable selection method is put forward for multivariate linear regression models which obey normality. This method hinges on minimizing a generalized C_p (GC_p) criterion which is defined by adding a positive constant value (the product of α and the number of parameters in the mean structure) to the minimum value of the multivariate residual sum of squares. The paper seeks to clarify the sufficient condition for α to simultaneously satisfy asymptotically loss and mean efficient properties in an asymptotic framework such that the sample size always goes to ∞, but the dimension of the vector of response variables can be either fixed or infinite. Based on this, we propose an asymptotically loss and mean efficient GC_p criterion by using α which satisfies the obtained sufficient condition even under high dimensionality of the vector of response variables.

Key words: Loss efficiency, Generalized C_p criterion, High-dimensionality, Mean efficiency, Selection probability, Variable selection.

E-mail address: yanagi-hiro@hiroshima-u.ac.jp (Hirokazu Yanagihara)

1. Introduction

The multivariate linear regression model is central to theoretical and applied inferential analysis. This model is introduced in many statistical textbooks (see, e.g., Srivastava, 2002, chap. 9; Timm, 2002, chap. 4), and is widely used in chemometrics, engineering, econometrics, psychometrics, and many other fields, for the prediction of response variables to a set of explanatory variables. Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots. \boldsymbol{y}_n)'$ be an $n \times p$ matrix of p response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times k$ matrix of non-stochastic k explanatory variables, where n is the sample size. To ensure both the possibility of estimating the model and the existence of variable selection criteria, we assume that $\mathrm{rank}(\boldsymbol{X}) = k\ (< n)$ and $n - p - k - 1 > 0$. Let j denote a subset of $\omega = \{1, \ldots, k\}$ containing k_j elements and $\boldsymbol{X}_j$ denote the $n \times k_j$ matrix consisting of the columns of $\boldsymbol{X}$ indexed by the elements of j, where k_A denotes the number of elements of a set A, i.e., $k_A = \#(A)$. For example, if $j = \{1, 2, 4\}$, then $\boldsymbol{X}_j$ consists of the first, second, and fourth columns of $\boldsymbol{X}$. For convenience, elements of j are arranged in ascending order. We then consider the following multivariate linear regression model with k_j explanatory variables as the candidate model:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_j\boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ unknown matrix of regression coefficients, and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown covariance matrix with $\mathrm{rank}(\boldsymbol{\Sigma}_j) = p$. We identify the candidate model by set j and denote the candidate model in (1) as model j. Importantly, the model with $\boldsymbol{X}_\omega$ (namely $\boldsymbol{X}$) is called the full model. We will assume that the data are generated from the following true model:

$$\boldsymbol{Y} \sim N_{n\times p}(\boldsymbol{X}_{j_*}\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n), \tag{2}$$

where j_* is a set of integers indicating the subset of explanatory variables in the true model, $\boldsymbol{\Theta}_*$ is a $k_{j_*} \times p$ matrix of true regression coefficients, and $\boldsymbol{\Sigma}_*$ is a $p \times p$ true covariance matrix with $\mathrm{rank}(\boldsymbol{\Sigma}_*) = p$. We denote the model in (2) as the true model j_*. Henceforth, for expository purposes, we simply represent $\boldsymbol{X}_{j_*}$ and k_{j_*} as $\boldsymbol{X}_*$ and k_*, respectively.

In what follows, we focus on a variable selection method by minimizing a generalized C_p (GC_p) criterion, when p may be large but still smaller than n. The GC_p criterion is defined by adding a positive constant value to the minimum value of the multivariate residual sum of squares. This positive value is the product of α and the number of parameters in the mean structure, where α expresses a penalty for the complexity of the candidate model. The GC_p criterion in univariate and multivariate linear regression model contexts was proposed by Atkinson (1980) and Nagai *et al.* (2012), respectively. The family of GC_p criteria contains many widely known variable selection criteria, e.g., Mallows C_p criterion proposed by Sparks *et al.* (1983) (the original C_p was proposed by Mallows, 1973, under the univariate linear regression model) and the modified C_p (MC_p) criterion proposed by Fujikoshi and Satoh (1997), which is a completely bias-corrected version of the C_p criterion. Given that our focus is on multivariate linear regression models where the dimension p may be large, the following asymptotic framework is used for assessing the asymptotic property of the variable-selection method:

$$n \to \infty \text{ and } p/n \to c_0 \in [0, 1). \tag{3}$$

For simplicity, we will use "$n \to \infty, p/n \to c_0$" to refer to this asymptotic framework. It should be emphasized that we are not concerned whether p goes to ∞ or not in this asymptotic framework. If p increases, the columns of $\boldsymbol{\Theta}_*$ and the rows and columns of $\boldsymbol{\Sigma}_*$ increase, so it is necessary to clarify how true parameters increase. We assume that as p increases by 1, a new column is added to the right of the current $\boldsymbol{\Theta}_*$, and a new column and row are added to the right of and the bottom of the current $\boldsymbol{\Sigma}_*$, respectively.

There are two important properties of variable selection criteria. One is a consistency property whereby the selection probability of the true model by a variable selection criterion converges to 1 asymptotically. The second is an efficiency property which can be divided into two sub-properties. The first is asymptotic loss efficiency whereby the ratio of the minimum loss function to the loss function of the model selected by a variable selection criterion goes to 1 asymptotically. The second is asymptotic mean efficiency whereby the ratio of the minimum expected loss function to the expected loss function of the model selected by a variable selection criterion goes to 1 asymptotically (see, e.g., Shibata, 1980, 1981; Shao, 1997).

Yanagihara (2016a) clarified the sufficient condition for α in GC_p criterion to satisfy a consistency property under the asymptotic framework in (3), and proposed the high-dimensionality-adjusted consistent GC_p ($HCGC_p$) by using the sufficient condition. Recently An aim of this paper is to derive the sufficient condition for α in the GC_p criterion to simultaneously satisfy loss and mean efficiency properties under the asymptotic framework in (3). Then, we propose the asymptotically loss and mean efficient GC_p criterion by using α which satisfies the obtained sufficient condition even under high-dimensionality of the vector of response variables. In our setting, the true model is included in the set of all possible candidate models. For asymptotic loss efficiency under settings different from ours, see Imori (2020).

The remainder of the paper is organized as follows. In Section 2, we present the necessary notation and assumptions for assessing the loss and mean efficient properties of the GC_p criterion in model j (1). The main results are shown in Section 3. In Section 4, we present the results of numerical experiments and compare the variable selection performance of the proposed criterion with that of an existing variable selection criterion. Technical details are provided in the Appendix.

2. Formulation of Loss and Mean Efficiency Properties

First, we describe several classes of j which express subsets of $\boldsymbol{X}$ in the candidate model. Let $\mathcal{J}$ be a set of all possible candidate models denoted by $\mathcal{J} = \wp(\omega)$, where $\wp(A)$ is the powerset of set A. We assume the following regarding $\mathcal{J}$:

A1. The true model is included in the set of candidate models, i.e., $j_* \in \mathcal{J}$.

Moreover, we separate $\mathcal{J}$ into two sets, one a set of overspecified models wherein the explanatory variables contain all the explanatory variables of the true model j_* in (2) and the other a set of underspecified models, i.e.,

$$\mathcal{J}_+ = \{j \in \mathcal{J} | j_* \subseteq j\}, \quad \mathcal{J}_- = \mathcal{J}_+^c \cap \mathcal{J}, \tag{4}$$

where A^c denotes the compliment of a set A. We use the terminology "overspecified model" and "underspecified model" in the same sense as Fujikoshi and Satoh (1997).

Let $\boldsymbol{S}_j$ be the unbiased estimator of $\boldsymbol{\Sigma}_j$ in model j (1), i.e.,

$$\boldsymbol{S}_j = \frac{1}{n-k}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{Y}, \tag{5}$$

where $\boldsymbol{P}_j$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}_j$, i.e., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$. Let $d(\boldsymbol{A}, \boldsymbol{B})$ and $\hat{d}(\boldsymbol{A}, \boldsymbol{B})$ be the squared Mahalanobis' distances between two $n \times p$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as

$$d(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\left\{(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{A} - \boldsymbol{B})'\right\} = \|\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{A} - \boldsymbol{B})\|^2,$$
$$\hat{d}(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\left\{(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{S}_\omega^{-1}(\boldsymbol{A} - \boldsymbol{B})'\right\} = \|\boldsymbol{S}_\omega^{-1/2}(\boldsymbol{A} - \boldsymbol{B})\|^2.$$

The matrix of regression coefficients $\boldsymbol{\Theta}_j$ in model j (1) is estimated using the least squares approach; i.e., $\boldsymbol{\Theta}_j$ is estimated by minimizing the multivariate RSS, and then, the estimator of $\boldsymbol{\Theta}_j$ is given by

$$\hat{\boldsymbol{\Theta}}_j = (\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'\boldsymbol{Y} = \arg\min_{\boldsymbol{\Theta}_j \in \mathbb{M}(k_j, p)} \hat{d}(\boldsymbol{Y}, \boldsymbol{X}_j\boldsymbol{\Theta}_j),$$

where $\mathbb{M}(k, p)$ denotes a set of $k \times p$ real matrices.

Let a $p \times p$ noncentrality matrix be denoted by

$$\boldsymbol{\Delta}_j = \frac{1}{n}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Theta}_*'\boldsymbol{X}_*'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_*\boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1/2},$$

where $\boldsymbol{\Theta}_*$, $\boldsymbol{\Sigma}_*$, and $\boldsymbol{X}_*$ are the matrix of the true regression coefficients, the true covariance matrix, and the matrix of the true explanatory variables, respectively, given by (2). It is straightforward to discern from the definition of $\mathcal{J}_+$ that $\boldsymbol{\Delta}_j = \boldsymbol{O}_{p,p}$ if and only if $j \in \mathcal{J}_+$, where $\boldsymbol{O}_{n,p}$ is an $n \times p$ matrix of zeros. For the noncentrality matrix, we assume that

A2. $\boldsymbol{R}_0 = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ exists and is positive definite.

Let

$$\delta_j = \mathrm{tr}(\boldsymbol{\Delta}_j). \tag{6}$$

If Assumption A2 holds, then we have

$$^\forall j \in \mathcal{J}_-, \quad \inf_{n > p+k, p \geq 1} \delta_j > 0.$$

Even though δ_j diverges, we do not assume a specific order of the divergence. Instead, we use the following assumption:

A3. $^\exists\epsilon \in \mathbb{N} \quad s.t. \quad ^\forall j \in \mathcal{J}_-, \quad \lim_{n\to\infty, p/n \to c_0} \frac{1}{n^\epsilon}\, \boldsymbol{\Theta}_*\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Theta}_*' = \boldsymbol{O}_{k_*,k_*}.$

If Assumptions A2 and A3 hold, then we have

$$^\exists\epsilon \in \mathbb{N} \quad s.t. \quad ^\forall j \in \mathcal{J}_-, \quad \lim_{n\to\infty, p/n \to c_0} \frac{\delta_j}{n^\epsilon} = 0.$$

We determine the best subset of ω by minimizing a GC_p criterion. The GC_p criterion under model j is defined by

$$GC_p(j|\alpha) = (n - k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\omega^{-1}) + \alpha p k_j, \tag{7}$$

where α is a positive value expressing a penalty for the complexity of the candidate model. A variable selection criterion included in the family of GC_p criteria is specified by an individual α. This family contains the C_p, MC_p, and $HCGC_p$ criteria as special cases, i.e.,

$$\alpha = \begin{cases} 2 & (C_p) \\ 2c_{\mathrm{M}}^{-1} & (MC_p) \\ a_{\mathrm{C}}(\beta) & (HCGC_p) \end{cases},$$

where $c_{\rm M}$ is a positive value given by

$$c_{\rm M} = \frac{m-2}{n-k} = 1 - \frac{p+1}{n-k} \quad (m = n-k-p+1), \tag{8}$$

$a_{\rm C}(\beta)$ is a positive value expressed as a function of β and defined by

$$a_{\rm C}(\beta) = \frac{n}{n-p} + \beta, \tag{9}$$

and β is a positive value satisfying

$$\lim_{n\to\infty, p/n\to c_0} \sqrt{p}\beta = \infty, \qquad \lim_{n\to\infty, p/n\to c_0} \frac{p}{n}\beta = 0. \tag{10}$$

The best subset of ω, which is chosen by minimizing the GC_p criterion, is written as

$$\hat{j}_\alpha = \arg\min_{j\in\mathcal{J}} GC_p(j|\alpha). \tag{11}$$

Herein, the selected model is denoted as $\hat{j}_\alpha$.

The squared loss function between the true mean and the fitted value, i.e., $\boldsymbol{X}_*\boldsymbol{\Theta}_*$ and $\boldsymbol{X}_j\hat{\boldsymbol{\Theta}}_j$, is defined by

$$\mathcal{L}(j) = d(\boldsymbol{X}_*\boldsymbol{\Theta}_*, \boldsymbol{X}_j\hat{\boldsymbol{\Theta}}_j) = n\delta_j + \mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}), \tag{12}$$

where $\boldsymbol{\mathcal{E}} = (\boldsymbol{Y} - \boldsymbol{X}_*\boldsymbol{\Theta}_*)\boldsymbol{\Sigma}_*^{-1/2} \sim N_{n\times p}(\boldsymbol{O}_{n,p}, \boldsymbol{I}_p \otimes \boldsymbol{I}_n)$. Notice that $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}})] = pk_j$. Hence, the expectation of $\mathcal{L}(j)$, denoted by $\mathcal{R}(j)$, is given by

$$\mathcal{R}(j) = E\left[\mathcal{L}(j)\right] = n\delta_j + k_j p. \tag{13}$$

Subsets of ω minimizing $\mathcal{L}(j)$ and $\mathcal{R}(j)$ are written as

$$\hat{j}_{\rm LE} = \arg\min_{j\in\mathcal{J}} \mathcal{L}(j), \quad j_{\rm ME} = \arg\min_{j\in\mathcal{J}} \mathcal{R}(j), \tag{14}$$

Next, $\hat{j}_{\rm LE}$ and $j_{\rm ME}$ are the loss optimal model and the mean optimal model in the sense of minimizing the loss function and the mean of the loss function, respectively. By using the loss function, we formulate the asymptotically loss efficient property.

Definition 1 *If the following equation holds, we say that the GC_p criterion is asymptotically loss efficient:*

$$\operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{\mathcal{L}(\hat{j}_\alpha)}{\mathcal{L}(\hat{j}_{\rm LE})} = 1,$$

where plim *is a probability limit operator, i.e., denotes a convergence in probability.*

It should be kept in mind that

$$\mathcal{L}(\hat{j}_\alpha) = \sum_{j\in\mathcal{J}} \mathcal{L}(j) I(\hat{j}_\alpha = j), \quad \mathcal{L}(\hat{j}_{\rm LE}) = \sum_{j\in\mathcal{J}} \mathcal{L}(j) I(\hat{j}_{\rm LE} = j), \tag{15}$$

where $I(A)$ is the indicator function, i.e., $I(A) = 1$ if A is true and $I(A) = 0$ if A is not true. By using the mean of the loss function, we formulate the asymptotically mean efficient property.

Definition 2 *If the following equation holds, we say that the GC_p criterion is asymptotically mean efficient:*

$$\lim_{n\to\infty, p/n\to c_0} \frac{E[\mathcal{L}(\hat{j}_\alpha)]}{\mathcal{R}(j_{\rm ME})} = 1.$$

3. Main Results

3.1. Asymptotically Loss Efficient Property

To clarify the sufficient condition for α in (7) for loss efficiency, we consider asymptotic behaviors of the loss function $\mathcal{L}(j)$, the loss optimal model $\hat{j}_{\mathrm{LE}}$, and the selected model $\hat{j}_\alpha$, given by (12), (14), and (11), respectively.

First, we prepare three conditions for α which play an important role in determining whether the GC_p criterion is efficient.

C1. $\displaystyle\lim_{n\to\infty, p/n\to c_0} \sqrt{p}\left(-\frac{n}{n-p}+\alpha\right)=\infty.$

C2. ${}^\forall j \in \mathcal{J}_-$,

$$\lim_{n\to\infty, p/n\to c_0} \frac{p}{n\delta_j}\left(-\frac{n}{n-p}+\alpha\right)=0. \tag{16}$$

C3. $\displaystyle\lim_{n\to\infty, p/n\to c_0} \frac{p}{n}\alpha=\frac{2c_0}{1-c_0}.$

Let $\boldsymbol{\theta}_a^*$ $(a \in j_*)$ be the p-dimensional vector of regression coefficients for the ath explanatory variable. Then we prepare the following subsets of j_* and $\mathcal{J}_-$:

$$\xi_* = \left\{a \in j_* \middle| \boldsymbol{\theta}_a^{*\prime}\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\theta}_a^* = O(1) \text{ as } p\to\infty\right\}, \quad \mathcal{S}_- = \left\{j \in \mathcal{J}_- \middle| j \supseteq (\xi_*^c \cap j_*)\right\}, \tag{17}$$

where $\mathcal{J}_-$ is the set of underspecified models given by (4). By using the subsets, the following lemma concerning the convergence or divergence of δ_j given by (6) is derived (the proof is given in Appendix A.1).

Lemma 1 *Suppose that Assumption A2 holds.*

(*1*) *When $j \in \mathcal{S}_-$, δ_j converges to a positive value as $n\to\infty$ and $p/n\to c_0 \in [0,1)$, i.e.,*

$${}^\forall j \in \mathcal{S}_-, \quad {}^\exists \delta_{j,0} > 0 \quad s.t.\ \delta_{j,0} = \lim_{n\to\infty, p/n\to c_0} \delta_j.$$

(*2*) *When $j \in \mathcal{S}_-^c \cap \mathcal{J}_-$, δ_j diverges to ∞ as $n\to\infty$ and $p/n\to c_0\in[0,1)$, i.e.,*

$${}^\forall j \in \mathcal{S}_-^c \cap \mathcal{J}_-, \quad \lim_{n\to\infty, p/n\to c_0} \delta_j = \infty.$$

Lemma 1 indicates that δ_j in $j \in \mathcal{J}_-$ is not convergent if $\xi_* = \emptyset$ because $\mathcal{S}_- = \emptyset$ when $\xi_* = \emptyset$. Here, let $\delta_{j,0} = 0$ for $j \in \mathcal{J}_+$, where $\mathcal{J}_+$ is the set of overspecified models given by (4). By using Lemma 1 and the definition of $\mathcal{L}(j)$, we show that $\mathcal{L}(j)$ can converge to a positive value as follows (the proof is given in Appendix A.2).

Lemma 2 *Suppose that Assumption A2 holds. If $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$, $\mathcal{L}(j)$ for any $j \in \mathcal{S}_- \cup \mathcal{J}_+$ converges to a positive value as $n\to\infty$ and $p/n\to c_0\in[0,1)$, i.e.,*

$${}^\forall j \in \mathcal{S}_- \cup \mathcal{J}_+, \quad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{\mathcal{L}(j)}{n} = \delta_{j,0} + c_0 k_j = \lambda(j) > 0. \tag{18}$$

Let j_0 be the model defined by

$$j_0 = \begin{cases} \arg\min_{j \in \{j_*\} \cup \mathcal{S}_-} \lambda(j) & (\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}) \\ j_* & (\{\xi_* = \emptyset\} \cup \{c_0 = 0\}) \end{cases} . \tag{19}$$

By using j_0 and Lemma 2, the following lemma concerning behaviors of the loss optimal model is obtained (the proof is given in Appendix A.3).

Lemma 3 *Suppose that Assumptions A1 and A2 hold.*

(*1*) *The loss optimal model is the true model or an underspecified model with probability 1, i.e.,*

$$P(\hat{j}_{\mathrm{LE}} \in \{j_*\} \cup \mathcal{J}_-) = 1.$$

(*2*) *The probability of* $\hat{j}_{\mathrm{LE}} = j_0$ *converges to 1 as* $n \to \infty$ *and* $p/n \to c_0 \in [0, 1)$*, i.e.,*

$$\lim_{n\to\infty, p/n\to c_0} P(\hat{j}_{\mathrm{LE}} = j_0) = 1. \tag{20}$$

Since equation (20) holds, we call j_0 in (19) the asymptotically loss optimal model.

The following key lemma to prove the efficient property of the GC_p criterion is obtained (the proof is given in Appendix A.4).

Lemma 4 *Suppose that*

$$^{\exists}\ell \in \mathcal{J} \quad s.t. \quad \lim_{n\to\infty, p/n\to c_0} P(\hat{j}_{\mathrm{LE}} = \ell) = 1, \qquad \lim_{n\to\infty, p/n\to c_0} P(\hat{j}_\alpha = \ell) = 1.$$

Then, the GC_p *criterion is asymptotically loss efficient when* $n \to \infty$ *and* $p/n \to c_0 \in [0, 1)$.

From Lemmas 3 and 4, we can discern that the GC_p criterion is asymptotically loss efficient if $P(\hat{j}_\alpha = j_0) \to 1$ as $n \to \infty$ and $p/n \to c_0$. Probability convergence is ensured by the following lemma (the proof is given in Appendix A.5).

Lemma 5 *Suppose that Assumptions A1 and A2 hold. The selection probability of* j_0 *by the* GC_p *criterion converges to 1 as* $n \to \infty$ *and* $p/n \to c_0 \in [0, 1)$*, i.e.,*

$$\lim_{n\to\infty, p/n\to c_0} P(\hat{j}_\alpha = j_0) = 1,$$

if α *satisfies Conditions C1 and C2 when* $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$*, or Condition C3 when* $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$*, where* j_0 *is the asymptotically loss optimal model given by* (19).

From Lemmas 3, 4, and 5, we derive the following main theorem concerning the asymptotically loss efficient property.

Theorem 1 *Suppose that Assumptions A1 and A2 hold. The* GC_p *criterion is asymptotically loss efficient when* $n \to \infty$ *and* $p/n \to c_0 \in [0, 1)$ *if* α *satisfies Conditions C1 and C2 when* $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$*, or Condition C3 when* $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$.

3.2. Asymptotically Mean Efficient Property

To clarify the sufficient condition for α in (7) for mean efficiency, asymptotic behaviors of the mean of the loss function $\mathcal{R}(j)$, the mean optimal model j_{ME}, and the selected model $\hat{j}_\alpha$, given by (13), (14), and (11), respectively, are considered.

First, we present a lemma concerning convergence of $\mathcal{R}(j_{\mathrm{ME}})/p$ (the proof is given in Appendix A.6).

Lemma 6 *Suppose that Assumptions A1 and A2 hold. The minimum of $\mathcal{R}(j)/p$ is expanded as*

$$\frac{1}{p}\mathcal{R}(j_{\mathrm{ME}}) = \frac{1}{p}\mathcal{R}(j_0) + o(1), \quad \text{as } n \to \infty,\ \ p/n \to c_0,$$

where j_0 is the asymptotically loss optimal model given by (19).

We derive the order of the selection probability of some underspecified model by the GC_p criterion (the proof is given in Appendix A.7).

Lemma 7 *Suppose that Assumptions A1, A2, and A3 hold. If α satisfies Condition C2, $np^{-1}\delta_j P(\hat{j}_\alpha = j)$ for any $j \in \mathcal{J}_-$ converges to 0 as $n \to \infty$ and $p/n \to c_0 \in [0, 1)$. If α satisfies Condition C3, $np^{-1}\delta_j P(\hat{j}_\alpha = j)$ for any $j \in \mathcal{S}_-^c \cap \mathcal{J}_-$ converges to 0 as $n \to \infty$ and $p/n \to c_0 \in [0, 1)$.*

By using Lemma 7, the following theorem concerning convergence of $E[\mathcal{L}(\hat{j}_\alpha)]/p$ is derived (the proof is given in Appendix A.8).

Lemma 8 *Suppose that Assumptions A1, A2, and A3 hold. Then $E[\mathcal{L}(\hat{j}_\alpha)]/p$ is expanded as*

$$\frac{1}{p}E\left[\mathcal{L}(\hat{j}_\alpha)\right] = \frac{1}{p}\mathcal{R}(j_0) + o(1), \quad \text{as } n \to \infty,\ \ p/n \to c_0$$

if α satisfies Conditions C1 and C2 when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$, or Condition C3 when $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$, where j_0 is the asymptotically loss optimal model given by (19).

Finally, from Lemmas 6 and 8, we obtain the following main theorem concerning the asymptotically mean efficient property.

Theorem 2 *Suppose that Assumptions A1, A2, and A3 hold. The GC_p criterion is asymptotically mean efficient when $n \to \infty$ and $p/n \to c_0 \in [0, 1)$ if α satisfies Conditions C1 and C2 when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$ or Condition C3 when $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$.*

3.3. A High-dimensionality-adjusted Efficient GC_p Criterion

From Theorems 1 and 2, we can discern that sufficient conditions for α to satisfy asymptotic loss efficiency are equal to those to satisfy asymptotic mean efficiency. By using the α conditions, an asymptotically loss and mean efficient GC_p criterion can be proposed. Unfortunately, the conditions for α in Theorems 1 and 2 are not verifiable in practical terms because they depend on the true model. Hence, we attempt to derive sufficient conditions for α that do not depend on the true model. This is achieved by invoking the following corollary (the proof is given in Appendix A.9).

Corollary 1 *Suppose that Assumptions A1, A2, and A3 hold. The GC_p criterion is asymptotically loss and mean efficient when $n \to \infty$ and $p/n \to c_0 \in [0, 1)$ if α satisfies Conditions C1 and C3.*

Let $a_{\mathrm{E}}(\beta)$ be a positive value expressed as a function of β in (10), which is defined by

$$a_{\mathrm{E}}(\beta) = \frac{2n}{n-p} + \beta, \tag{21}$$

where β is a positive value given by (10). The conditions of β imply that

$$\lim_{n\to\infty, p/n\to c_0} \sqrt{p}\left(-\frac{n}{n-p} + a_{\mathrm{E}}(\beta)\right) = \lim_{n\to\infty, p/n\to c_0} \left(\frac{\sqrt{p}n}{n-p} + \sqrt{p}\beta\right) = \infty,$$
$$\lim_{n\to\infty, p/n\to c_0} \frac{p}{n} a_{\mathrm{E}}(\beta) = \lim_{n\to\infty, p/n\to c_0} \left(\frac{2p}{n-p} + \frac{p}{n}\beta\right) = \frac{2c_0}{1-c_0}.$$

The above equations indicate that $a_{\mathrm{E}}(\beta)$ satisfies Conditions C1 and C3. This means that the GC_p criterion with $\alpha = a_{\mathrm{E}}(\beta)$ is asymptotically loss and mean efficient under the asymptotic framework in (3). We call the GC_p criterion with $\alpha = a_{\mathrm{E}}(\beta)$ a high-dimensionality-adjusted efficient GC_p ($HEGC_p$) criterion.

4. Numerical Study

In this section, we present the results of a simulation study that compared the performance of variable selection using the proposed $HEGC_p$ criterion as well as four existing criteria. Let b_{E} be a positive value defined by

$$b_{\mathrm{E}} = \sqrt{\frac{n}{n-p}\frac{\log\log n}{\sqrt{p}}}.$$

Then, the GC_p criterion with $\alpha = 2c_{\mathrm{M}}^{-1} + b_{\mathrm{E}}$ is asymptotically loss and mean efficient. This is because $2c_{\mathrm{M}}^{-1} + b_{\mathrm{C}} = a_{\mathrm{E}}(2r_{\mathrm{M}} + b_{\mathrm{E}})$ and $2r_{\mathrm{M}} + b_{\mathrm{E}}$ satisfies conditions in (10), where r_{M} is a positive value defined by

$$r_{\mathrm{M}} = c_{\mathrm{M}}^{-1} - \frac{n}{n-p} = \frac{n+kp}{(n-p)(n-k-p-1)}.$$

Recall that the GC_p criterion with $\alpha = 2c_{\mathrm{M}}^{-1}$ is the MC_p criterion. Hence, we specifically refer to the GC_p criterion with $\alpha = 2c_{\mathrm{M}}^{-1} + b_{\mathrm{E}}$ as a high-dimensionality-adjusted efficient MC_p ($HEMC_p$) criterion. The Bayesian information criterion (BIC)-type C_p (BC_p) criterion is the GC_p criterion with $\alpha = \log n$ because the penalty term in the BIC is the product of $\log n$ and the number of parameters. Then, we compared the ratios of loss function expectations in selected models according to the five GC_p criteria:

Criterion 1 ($HEMC_p$): $\alpha = 2c_{\mathrm{M}}^{-1} + b_{\mathrm{E}}$, Criterion 2 ($C_p$): $\alpha = 2$,
Criterion 3 (MC_p) : $\alpha = 2c_{\mathrm{M}}^{-1}$, Criterion 4 ($BC_p$) : $\alpha = \log n$,
Criterion 5 ($HCGC_p$): $\alpha = a_{\mathrm{C}}(b_{\mathrm{C}})$,

where $a_{\mathrm{C}}(\beta)$ is given by (9) and b_{C} is a positive value defined by

Table 1. Efficiency results

Type of $\boldsymbol{\Sigma}_*$	Type of p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$
Auto-	$p < \infty$	○			○	○
correlation	$p \to \infty$	○	△ ($c_0 < 0.5$)	○	○	○
Uniform	$p < \infty$	○			○	○
correlation	$p \to \infty$	○	△ ($c_0 = 0$)	○	△ ($c_0 = 0$)	△ ($c_0 = 0$)

Note) The symbol ○ denotes an efficient criterion, and the symbol △ denotes that the criterion is efficient if and only if the equation within the brackets is satisfied.

$$b_{\mathrm{C}} = \sqrt{\frac{n}{n-p}\frac{\log n}{\sqrt{p}}}.$$

We conducted Monte Carlo simulations based on 10,000 replications, using several values of n and p. The set of candidate models was $\mathcal{J} = \{j_1, \ldots, j_{15}\}$, where $j_a = \{1, \ldots, a\}$ $(a = 1, \ldots, 15)$. We generated $z_1, \ldots, z_n$ independently from $U(-1, 1)$. Using $z_1, \ldots, z_n$, we constructed an $n \times k$ matrix of explanatory variables $\boldsymbol{X}$, in which the (a, b)th element was given by z_a^{b-1} $(a = 1, \ldots, n; b = 1, \ldots, 15)$. The true model was determined by $\boldsymbol{\Theta}_* = \boldsymbol{\eta}_* \mathbf{1}_p'$, $j_* = \{1, 2, 3, 4, 5\}$, and $\boldsymbol{\Sigma}_* = 0.4\boldsymbol{\Psi}_*(0.8)$, where $\boldsymbol{\eta}_* = (1, -2, 3, -4, 5)'$ and $\mathbf{1}_p$ is the p-dimensional vector of ones. Thus, $j_1, \ldots, j_4$ were underspecified models, $j_5, \ldots, j_{15}$ were overspecified models, and $j_* = j_5$. To specify the form of $\boldsymbol{\Psi}_*(\rho)$, two types of correlation matrices were prepared as follows.

Type 1. Autocorrelation matrix, i.e., the (a, b)th element of $\boldsymbol{\Psi}_*(\rho)$ is $\rho^{|a-b|}$.

Type 2. Uniform correlation matrix, i.e., $\boldsymbol{\Psi}_*(\rho) = (1 - \rho)\boldsymbol{I}_p + \rho \mathbf{1}_p \mathbf{1}_p'$.

It follows from the equation $\boldsymbol{\Theta}_* = \boldsymbol{\eta}_* \mathbf{1}_p'$ that for $a = 1, \ldots, 5$,

$$\boldsymbol{\theta}_a^{*\prime} \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_a^* = \frac{a^2}{0.4} \mathbf{1}_p' \boldsymbol{\Psi}_*(\rho)^{-1} \mathbf{1}_p = \begin{cases} \dfrac{a^2\{p(1-\rho) + 2\rho\}}{0.4(1+\rho)} = O(p) & (\boldsymbol{\Psi}_*(\rho) \text{ is Type 1}) \\ \dfrac{a^2 p}{0.4\{1 + \rho(p-1)\}} = O(1) & (\boldsymbol{\Psi}_*(\rho) \text{ is Type 2}) \end{cases}.$$

Hence, in the simulation models, ξ_* in (17) was

$$\xi_* = \begin{cases} \emptyset & (\boldsymbol{\Psi}_*(\rho) \text{ is Type 1}) \\ \{1, 2, 3, 4, 5\} & (\boldsymbol{\Psi}_*(\rho) \text{ is Type 2}) \end{cases}.$$

From the conditions stated in Theorems 1 and 2 and ξ_*, we can determine whether the GC_p criterion is efficient. Table 1 shows which criteria have been shown to have the properties of asymptotic loss and mean efficiency in various simulation studies. The symbol ○ denotes efficiency, and the symbol △ indicates that the criterion is efficient if the equation within the brackets is satisfied.

Tables 2 and 3 give the probabilities that each criterion will select $\hat{j}_{\mathrm{LE}}$ when the true covariance matrices are Types 1 and 2, respectively. Tables 4 and 5 give the relative expected loss functions selected by each criterion, i.e., $E[\mathcal{L}(\hat{j}_\alpha)]/R(j_{\mathrm{ME}})$, when the true covariance matrices are Types 1 and 2, respectively. In each table, the left-hand side shows the results when p is fixed, the right-hand side shows the results when p increases with n while maintaining a fixed ratio p/n, and the bottom

Table 2. **Probabilities of selecting $\hat{j}_{\mathrm{LE}}$ by each of the five criteria when the true variance-covariance matrix is an autocorrelation matrix**

		Selection Probability (%)						Selection Probability (%)				
n	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$
100	5	97.18	86.66	90.04	98.27	98.19	5	97.18	86.66	90.04	98.27	98.19
200	5	97.98	89.31	90.76	99.99	99.34	10	99.30	94.53	96.29	100.00	99.48
500	5	98.25	90.00	90.39	100.00	99.87	25	99.95	99.30	99.61	100.00	99.90
1000	5	98.34	90.39	90.61	100.00	99.89	50	100.00	99.98	99.99	100.00	99.98
100	10	98.79	90.50	95.58	96.02	98.27	10	98.79	90.50	95.58	96.02	98.27
200	10	99.31	94.95	96.58	100.00	99.51	20	99.79	97.40	99.12	100.00	99.37
500	10	99.41	95.90	96.36	100.00	99.88	50	100.00	99.91	100.00	100.00	99.92
1000	10	99.53	96.45	96.68	100.00	99.95	100	100.00	100.00	100.00	100.00	99.99
100	30	99.53	72.51	99.09	99.11	92.83	30	99.53	72.51	99.09	99.11	92.83
200	30	99.97	97.56	99.65	100.00	99.23	60	100.00	90.84	99.96	100.00	97.53
500	30	100.00	99.68	99.86	100.00	99.94	150	100.00	99.16	100.00	100.00	99.42
1000	30	99.98	99.72	99.79	100.00	99.97	300	100.00	99.97	100.00	100.00	99.82
100	50	99.36	0.77	99.01	98.07	59.26	50	99.36	0.77	99.01	98.07	59.26
200	50	100.00	95.39	99.97	100.00	98.49	100	99.97	2.66	99.96	100.00	85.09
500	50	100.00	99.88	99.99	100.00	99.92	250	100.00	6.24	100.00	100.00	94.94
1000	50	100.00	100.00	100.00	100.00	100.00	500	100.00	8.09	100.00	100.00	97.87
100	70	94.77	0.00	94.51	1.77	0.55	70	94.77	0.00	94.51	1.77	0.55
200	70	99.99	81.38	99.97	100.00	96.14	140	99.78	0.00	99.77	80.75	18.80
500	70	100.00	99.87	99.99	100.00	99.83	350	100.00	0.00	100.00	100.00	61.90
1000	70	100.00	99.98	100.00	100.00	99.97	700	100.00	0.00	100.00	100.00	78.76
Avg.		99.12	84.05	96.94	94.66	92.05		99.42	57.43	98.65	93.70	84.09

row denotes the average value of that column. From these tables, it is clear that variable selection methods based on efficient criteria performed very well. In particular, the selection probability of the $HEMC_p$ criterion was always high, and the relative expected loss function selected by the $HEMC_p$ criterion was always small, under a moderately sized n. By contrast, it is intuitive that the performance of variable selection methods based on criteria that are not efficient was inferior. The performance of the variable selection method based on the MC_p criterion was comparable to those based on the $HEMC_p$ criterion. However, since the MC_p criterion is not asymptotically loss and mean efficient when p is fixed, the performance of the variable selection method based on the MC_p criterion was slightly inferior to those based on the $HEMC_p$ criterion when p was small.

From the results of these simulation studies, we confirmed the superior performance of variable selection based on the $HEMC_p$ criterion; this held under any true model, regardless of the size of p. Therefore, we anticipate that the $HEMC_p$ criterion will perform well in empirical contexts.

Acknowledgment The author is grateful to Prof. Hirofumi Wakaki, Hiroshima University, for his helpful comments on the proof of Lemma A.2. This research was partially supported by the Ministry of Education, Science, Sports and Culture, and a Grant-in-Aid for Scientific Research (C), #18K03415, 2018-2021.

Table 3. Probabilities of selecting $\hat{j}_{\mathrm{LE}}$ by each of the five criteria when the true variance-covariance matrix is a uniform correlation matrix

		Selection Probability (%)						Selection Probability (%)				
n	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$
100	5	96.83	86.35	89.94	94.24	97.69	5	96.83	86.35	89.94	94.24	97.69
200	5	97.97	89.62	90.85	99.95	99.34	10	99.39	94.79	96.53	99.05	99.51
500	5	98.00	90.14	90.64	100.00	99.68	25	99.95	99.32	99.70	99.91	99.90
1000	5	98.60	90.95	91.12	100.00	99.94	50	100.00	99.99	100.00	100.00	99.99
100	10	94.31	90.24	94.47	64.00	95.00	10	94.31	90.24	94.47	64.00	95.00
200	10	99.11	94.36	96.00	99.35	99.41	20	99.78	97.30	99.14	65.21	99.48
500	10	99.54	96.18	96.75	100.00	99.86	50	100.00	99.87	100.00	36.43	99.89
1000	10	99.66	96.74	96.95	100.00	99.93	100	100.00	100.00	100.00	11.03	100.00
100	30	37.98	57.56	48.48	13.10	66.70	30	37.98	57.56	48.48	13.10	66.70
200	30	98.96	97.81	99.46	21.78	99.10	60	65.98	89.53	78.84	2.35	94.80
500	30	99.99	99.65	99.88	98.83	99.95	150	86.84	99.24	94.27	0.00	99.46
1000	30	99.98	99.82	99.89	100.00	99.98	300	98.46	99.99	99.74	0.00	99.88
100	50	35.50	0.06	37.27	37.13	9.80	50	35.50	0.06	37.27	37.13	9.80
200	50	82.97	95.23	92.04	3.01	97.92	100	57.47	0.33	53.83	53.18	12.83
500	50	100.00	99.81	99.99	36.14	99.85	250	76.30	0.36	67.13	47.51	6.61
1000	50	100.00	99.95	100.00	100.00	99.96	500	74.62	0.70	64.35	18.37	7.84
100	70	42.37	0.00	40.74	0.10	0.04	70	42.37	0.00	40.74	0.10	0.04
200	70	46.47	73.04	58.79	6.65	84.41	140	45.25	0.00	47.82	15.10	0.56
500	70	100.00	99.90	100.00	2.05	99.89	350	85.06	0.00	86.43	68.66	0.18
1000	70	100.00	99.99	100.00	93.32	99.99	700	97.82	0.00	97.27	97.09	0.00
Avg.		86.41	82.87	86.16	63.48	87.42		79.70	55.78	79.80	46.12	59.51

References

Atkinson, A C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413–418.

Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

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

Imori, S. (2020). Asymptotic optimality of C_p-type criteria in high-dimensional multivariate linear regression models. *TR No.* 20-02, *Statistical Research Group*, Hiroshima University, Hiroshima.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Nagai, I., Yanagihara, H. & Satoh, K. (2012). Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods. *Hiroshima Math. J.*, **42**, 301–324.

Oda, R. & Yanagihara, H. (2019). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. TR No. 19-01, *Statistical Research Group*, *Hiroshima University*, Hiroshima.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, **7**, 221–264.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.*, **8**, 147–164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Sparks, R. S., Coutsourides, D. & Troskie, L. (1983). The multivariate C_p. *Comm. Statist. A - Theory Methods*, **12**, 1775–1793.

Srivastava, M. S. (2002). *Methods of Multivariate Statistics*. John Wiley & Sons, New York.

Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer-Verlag, New York.

Yanagihara H. (2015). Conditions for consistency of a log-likelihood-based information criterion in normal multivariate linear regression models under the violation of normality assumption. *J. Japan Statist. Soc.*, **45**, 21–56.

Yanagihara, H. (2016a). A high-dimensionality-adjusted consistent C_p-type statistic for selecting variables in a normality-

Table 4. **Relative expected loss function selected by each of the five criteria when the true variance-covariance matrix is an autocorrelation matrix**

		Relative Expected Loss						Relative Expected Loss				
n	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$
100	5	1.021	1.097	1.070	1.026	1.015	5	1.021	1.097	1.070	1.026	1.015
200	5	1.017	1.073	1.063	1.003	1.008	10	1.005	1.028	1.019	1.001	1.004
500	5	1.017	1.070	1.067	1.004	1.005	25	1.000	1.003	1.001	1.000	1.000
1000	5	1.013	1.066	1.064	1.001	1.002	50	0.998	0.998	0.998	0.998	0.998
100	10	1.009	1.051	1.024	1.042	1.011	10	1.009	1.051	1.024	1.042	1.011
200	10	1.001	1.022	1.014	0.997	1.000	20	1.001	1.010	1.004	1.000	1.003
500	10	1.002	1.020	1.018	0.999	1.000	50	0.998	0.999	0.998	0.998	0.999
1000	10	1.000	1.015	1.014	0.997	0.997	100	1.001	1.001	1.001	1.001	1.001
100	30	1.000	1.158	1.002	1.004	1.026	30	1.000	1.158	1.002	1.004	1.026
200	30	1.000	1.009	1.001	1.000	1.003	60	0.999	1.030	0.999	0.999	1.007
500	30	1.000	1.002	1.001	1.000	1.001	150	1.000	1.002	1.000	1.000	1.002
1000	30	1.000	1.001	1.001	1.000	1.000	300	1.000	1.000	1.000	1.000	1.001
100	50	1.002	2.825	1.003	1.005	1.305	50	1.002	2.825	1.003	1.005	1.305
200	50	1.001	1.015	1.001	1.001	1.005	100	1.000	2.660	1.000	1.000	1.056
500	50	1.000	1.000	1.000	1.000	1.000	250	0.999	2.451	0.999	0.999	1.013
1000	50	1.000	1.000	1.000	1.000	1.000	500	1.000	2.333	1.000	1.000	1.005
100	70	1.018	2.998	1.019	2.747	2.840	70	1.018	2.998	1.019	2.747	2.840
200	70	1.001	1.078	1.001	1.001	1.012	140	1.002	3.002	1.002	1.082	2.045
500	70	0.999	1.000	0.999	0.999	1.000	350	1.000	3.000	1.000	1.000	1.215
1000	70	1.000	1.000	1.000	1.000	1.000	700	1.000	3.000	1.000	1.000	1.079
Avg.		1.005	1.225	1.018	1.091	1.111		1.003	1.732	1.007	1.095	1.181

assumed linear regression with multiple responses. *Procedia Computer Science*, **96**, 1096–1105.

Yanagihara, H. (2016b). Convergence order of the probability of selecting the true set of variables by consistent C_p-type criterion in normal multivariate linear regression models. *Sūrikaisekikenkyūsho Kōkyūroku*, **1999**, 72–85 (in Japanese).

Appendix

A. Proofs of Lemmas and Theorems

A.1. Proof of Lemma 1

Let $\mathbf{\Theta}_j^{*\prime}$ for $j \subseteq j_*$ be a $p \times k_j$ matrix of which the ith column vector consists of $\boldsymbol{\theta}_{a_i}^*$, where a_i is the ith element of j, and $\mathbf{\Phi}_j^*$ for $j \subseteq j_*$ be a $k_j \times k_j$ matrix defined by $\mathbf{\Phi}_j^* = \mathbf{\Theta}_j^* \mathbf{\Sigma}_*^{-1} \mathbf{\Theta}_j^{*\prime}$. By using this notation, for any $j \in \mathcal{J}_-$, δ_j is rewritten as

$$\delta_j = \frac{1}{n} \mathrm{tr}\left\{ \mathbf{\Phi}_{j^c \cap j_*}^* \boldsymbol{X}_{j^c \cap j_*}' (\boldsymbol{I}_n - \boldsymbol{P}_j) \boldsymbol{X}_{j^c \cap j_*} \right\}. \quad \text{(A.1)}$$

To assess which elements of $\mathbf{\Phi}_j^*$ are convergent or divergent for $j \subseteq j_*$, we prepare the following lemma.

Lemma A.1 *Let* $\boldsymbol{a}_p = (a_1, \ldots, a_p)'$ *and* $\boldsymbol{b}_p = (b_1, \ldots, b_p)'$ *be p-dimensional real vectors and let* $\boldsymbol{S}_p$ *be a* $p \times p$ *positive definite real matrix. Moreover, we define sequences* $\{c_p\}$, $\{d_p\}$, *and* $\{h_p\}$ *as*

Table 5. Relative expected loss function selected by each of the five criteria when the true variance-covariance matrix is a uniform correlation matrix

		Relative Expected Loss						Relative Expected Loss				
n	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$	p	$HEMC_p$	C_p	MC_p	BC_p	$HCGC_p$
100	5	1.024	1.099	1.071	1.069	1.020	5	1.024	1.099	1.071	1.069	1.020
200	5	1.014	1.070	1.061	1.001	1.005	10	1.006	1.028	1.019	1.016	1.005
500	5	1.010	1.062	1.058	0.996	0.999	25	1.001	1.003	1.002	1.002	1.001
1000	5	1.014	1.065	1.064	1.005	1.005	50	1.001	1.001	1.001	1.001	1.001
100	10	1.028	1.050	1.026	1.213	1.024	10	1.028	1.050	1.026	1.213	1.024
200	10	1.006	1.028	1.020	1.010	1.004	20	1.000	1.010	1.002	1.229	1.001
500	10	1.002	1.018	1.015	0.999	1.000	50	1.000	1.000	1.000	1.428	1.000
1000	10	1.005	1.019	1.018	1.002	1.003	100	1.000	1.000	1.000	1.633	1.000
100	30	1.060	1.161	1.040	1.152	1.037	30	1.060	1.161	1.040	1.152	1.037
200	30	1.003	1.007	1.001	1.303	1.002	60	1.032	1.028	1.020	1.220	1.008
500	30	1.001	1.002	1.001	1.015	1.001	150	1.014	1.002	1.006	1.303	1.002
1000	30	1.000	1.001	1.000	1.000	1.000	300	1.001	1.000	1.000	1.466	1.000
100	50	1.045	2.970	1.043	1.045	1.375	50	1.045	2.970	1.043	1.045	1.375
200	50	1.028	1.016	1.014	1.261	1.007	100	1.026	2.740	1.022	1.049	1.086
500	50	1.000	1.001	1.000	1.430	1.001	250	1.006	2.527	1.010	1.105	1.041
1000	50	1.001	1.001	1.001	1.001	1.001	500	1.004	2.381	1.007	1.194	1.022
100	70	1.032	3.353	1.035	3.070	3.170	70	1.032	3.353	1.035	3.070	3.170
200	70	1.027	1.075	1.019	1.168	1.012	140	1.033	3.255	1.033	1.162	2.231
500	70	0.999	0.999	0.999	1.411	0.999	350	1.015	3.245	1.013	1.027	1.321
1000	70	0.999	0.999	0.999	1.073	0.999	700	1.002	3.225	1.003	1.003	1.155
Avg.		1.015	1.250	1.024	1.211	1.133		1.016	1.804	1.018	1.269	1.225

$$c_p = \boldsymbol{a}_p' \boldsymbol{S}_p^{-1} \boldsymbol{a}_p, \quad d_p = \boldsymbol{b}_p' \boldsymbol{S}_p^{-1} \boldsymbol{b}_p, \quad h_p = \boldsymbol{a}_p' \boldsymbol{S}_p^{-1} \boldsymbol{b}_p.$$

Then, if $c_p = O(1)$ and $d_p = O(1)$ hold as $p \to \infty$, sequences $\{c_p\}$, $\{d_p\}$ and $\{h_p\}$ are convergent, and sequences $\{c_p\}$ and $\{d_p\}$ converge to positive values.

(Proof of Lemma A.1) From the definitions of c_p and d_p, we can observe that $\{c_p\}$ and $\{d_p\}$ are positive sequences. Let

$$\boldsymbol{a}_{p+1} = \begin{pmatrix} \boldsymbol{a}_p \\ a_{p+1} \end{pmatrix}, \quad \boldsymbol{b}_{p+1} = \begin{pmatrix} \boldsymbol{b}_p \\ b_{p+1} \end{pmatrix}, \quad \boldsymbol{S}_{p+1} = \begin{pmatrix} \boldsymbol{S}_p & \boldsymbol{r}_{p+1} \\ \boldsymbol{r}_{p+1}' & s_{p+1} \end{pmatrix}. \tag{A.2}$$

It follows from the general formula for the inverse of a block matrix, e.g., th. 8.5.11 in Harville (1997), that

$$\boldsymbol{S}_{p+1}^{-1} = \begin{pmatrix} \boldsymbol{S}_p^{-1} & \boldsymbol{0}_p \\ \boldsymbol{0}_p' & 0 \end{pmatrix} + \frac{1}{w_{p+1}} \boldsymbol{u}_{p+1} \boldsymbol{u}_{p+1}', \tag{A.3}$$

where $\boldsymbol{0}_p$ is the p-dimensional vector of zeros, and w_{p+1} and $\boldsymbol{u}_{p+1}$ are given by

$$w_{p+1} = s_{p+1} - \boldsymbol{r}_{p+1}' \boldsymbol{S}_p^{-1} \boldsymbol{r}_{p+1}, \quad \boldsymbol{u}_{p+1} = \begin{pmatrix} \boldsymbol{S}_p^{-1} \boldsymbol{r}_{p+1} \\ -1 \end{pmatrix}.$$

It should kept in mind that $w_{p+1} > 0$ because we assume that $\boldsymbol{S}_{p+1}$ is also a positive definite real matrix. By using (A.2) and (A.3), we derive

$$c_{p+1} = \boldsymbol{a}'_{p+1}\boldsymbol{S}^{-1}_{p+1}\boldsymbol{a}_{p+1} = \boldsymbol{a}'_p\boldsymbol{S}^{-1}_p\boldsymbol{a}_p + \frac{1}{w_{p+1}}\left(\boldsymbol{a}'_{p+1}\boldsymbol{u}_{p+1}\right)^2 = c_p + \frac{1}{w_{p+1}}\left(\boldsymbol{a}'_{p+1}\boldsymbol{u}_{p+1}\right)^2. \quad \text{(A.4)}$$

This means that $\{c_p\}$ is a monotonically increasing sequence. On the other hand, the assumption indicates that sequence $\{c_p\}$ is bounded from above. Hence, $\{c_p\}$ converges to a positive value. By a similar method, $\{d_p\}$ also converges to a positive value. Let us define $\boldsymbol{u}_1 = -1$ and $w_1 = s_1$. From (A.4), we can see that sequences $\{c_p\}$, $\{d_p\}$, and $\{h_p\}$ are expressed as

$$c_p = \sum_{i=1}^{p}\left(\frac{\boldsymbol{a}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)^2, \quad d_p = \sum_{i=1}^{p}\left(\frac{\boldsymbol{b}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)^2, \quad h_p = \sum_{i=1}^{p}\left(\frac{\boldsymbol{a}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)\left(\frac{\boldsymbol{b}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right),$$

It follows from the relationship between the arithmetic mean and geometric mean that

$$h^+_p = \sum_{i=1}^{p}\left|\left(\frac{\boldsymbol{a}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)\left(\frac{\boldsymbol{b}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)\right| = \sum_{i=1}^{p}\sqrt{\left(\frac{\boldsymbol{a}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)^2\left(\frac{\boldsymbol{b}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)^2} \le \frac{1}{2}\sum_{i=1}^{p}\left\{\left(\frac{\boldsymbol{a}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)^2 + \left(\frac{\boldsymbol{b}'_i\boldsymbol{u}_i}{\sqrt{w_i}}\right)^2\right\} = \frac{1}{2}(c_p + d_p).$$

It is straightforward to see that $\{h^+_p\}$ is a monotonically increasing sequence. Since $\{c_p\}$ and $\{d_p\}$ are bounded from above, sequence $\{h^+_p\}$ is also bounded from above. Hence, $\{h^+_p\}$ converges to a positive value. This implies that $\{h^+_p\}$ converges to a positive value. Since h^+_p is the absolute convergent series, h_p is the convergent series. Hence, h_p is also convergent. ■

It follows from Lemma A.1 that $\boldsymbol{\Phi}^*_{\xi_*}$ converges to some positive semidefinite matrix, where ξ_* is given in (17). Since $\boldsymbol{X}'\boldsymbol{X}/n$ is convergent, $\boldsymbol{X}'_{j^c\cap j_*}(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{X}_{j^c\cap j_*}/n$ is also convergent. Therefore, δ_j is convergent for any $j \in \mathcal{S}_-$. When $j \in \mathcal{S}^c_- \cap \mathcal{J}_-$, we have

$$^\exists a \in \xi_* \cap j_* \;\; s.t. \;\; a \notin j.$$

Then, by using the above result and $\boldsymbol{\theta}^{*\prime}_a\boldsymbol{\Sigma}^{-1}_*\boldsymbol{\theta}^*_a \to \infty$ as $p \to \infty$, δ_j in (A.1) is rewritten as

$$\delta_j \ge \frac{1}{n}\text{tr}\left\{\boldsymbol{\Phi}^*_{j^c\cap j_*}\boldsymbol{X}'_{j^c\cap j_*}(\boldsymbol{I}_n - \boldsymbol{P}_{\omega\backslash\{a\}})\boldsymbol{X}_{j^c\cap j_*}\right\} = \frac{1}{n}\boldsymbol{\theta}^{*\prime}_a\boldsymbol{\Sigma}^{-1}_*\boldsymbol{\theta}^*_a\boldsymbol{X}'_{\{a\}}(\boldsymbol{I}_n - \boldsymbol{P}_{\omega\backslash\{a\}})\boldsymbol{X}_{\{a\}} \to \infty.$$

Hence, δ_j is divergent for any $j \in \mathcal{S}^c_- \cap \mathcal{J}_-$. Consequently, Lemma A.1 is proved.

A.2. Proof of Lemma 2

When $c_0 \ne 0$, $p \to \infty$ holds. Notice that $\text{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}})/p \xrightarrow{p} k_j$ as $p \to \infty$ because $\text{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}})$ is distributed according to a chi-square distribution with pk_j degrees of freedom. From the result and Lemma 1, Lemma 2 is proved.

A.3. Proof of Lemma 3

Suppose that $j_1 \subset j_2$ for $j_1, j_2 \in \mathcal{J}_+$. Then, $P(\text{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{j_2}\boldsymbol{\mathcal{E}}) - \text{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{j_1}\boldsymbol{\mathcal{E}}) > 0) = 1$ holds because $\boldsymbol{P}_{j_2} - \boldsymbol{P}_{j_1}$ is positive semidefinite and the distribution of $\boldsymbol{\mathcal{E}}$ is continuous. This implies that $P(\hat{j}_{\text{LE}} \in \mathcal{J}_+\backslash\{j_*\}) = 0$. Since $(\mathcal{J}_+\backslash\{j_*\})^c = \{j_*\} \cup \mathcal{J}_-$, Lemma 3 (1) is proved.

Next, we show the proof of Lemma 3 (2). Notice that

$$\frac{1}{n}\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}) = \frac{p}{n}\cdot\frac{1}{p}\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}) \xrightarrow{p} k_j c_0,$$

as $n \to \infty$ and $p/n \to c_0$. By using this result, we show the following basic results used for proving Lemma 3 (2):

$$\begin{aligned} {}^{\forall} j \in \mathcal{J}_-, \quad \frac{1}{n\delta_j}\{\mathcal{L}(j) - \mathcal{L}(j_*)\} &= 1 + \frac{1}{\delta_j}\left\{\frac{1}{n}\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}) - \frac{1}{n}\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_{j_*}\boldsymbol{\mathcal{E}})\right\} \\ &= 1 + \frac{1}{\delta_j}(k_j - k_*)c_0 + o_p(1), \text{ as } n \to \infty,\ p/n \to c_0. \end{aligned} \tag{A.5}$$

Recall that $P(\hat{j}_{\mathrm{LE}} \in \mathcal{J}_+\backslash\{j_*\}) = 0$. Hence, to prove Lemma 3 (2), it is sufficient to show that the following equation is satisfied:

$${}^{\forall} j \in \{\{j_*\} \cup \mathcal{J}_-\}\backslash\{j_0\}, \qquad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{h_j}\{\mathcal{L}(j) - \mathcal{L}(j_0)\} = \tau_j > 0, \tag{A.6}$$

(see, e.g., the result in Yanagihara, 2015), where j_0 is the asymptotically loss optimal model given by (19), and h_j is some positive constant depending on model j, which does not converge to 0. First, we give the proof when $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$. From Lemma 2 (1) and (19), we have

$${}^{\forall} j \in \{\{j_*\} \cup \mathcal{S}_-\}\backslash\{j_0\}, \qquad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n}\{\mathcal{L}(j) - \mathcal{L}(j_0)\} = \lambda(j) - \lambda(j_0) > 0, \tag{A.7}$$

where $\lambda(j)$ is given by (18). On the other hand, from Lemma 2, for any $j \in \mathcal{S}_- \cap \mathcal{J}_-$, δ_j goes to ∞ as $n \to \infty$ and $p/n \to c_0$. Then, from (A.7), we can see that $\{\mathcal{L}(j_*) - \mathcal{L}(j_0)\}/(n\delta_j)$ is 0 or converges to 0 in probability as $n \to \infty$ and $p/n \to c_0$. It follows from this result, the divergence of δ_j and (A.5) that ${}^{\forall} j \in \mathcal{S}_-^c \cap \mathcal{J}_-$,

$$\begin{aligned} &\operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n\delta_j}\{\mathcal{L}(j) - \mathcal{L}(j_0)\} \\ &= \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n\delta_j}\{\mathcal{L}(j) - \mathcal{L}(j_*)\} + \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n\delta_j}\{\mathcal{L}(j_*) - \mathcal{L}(j_0)\} = 1 > 0. \end{aligned}$$

Therefore, (A.6) is proved when $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$. Next, we give the proof when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$. Recall that $\xi_* = \emptyset \Leftrightarrow \delta_j \to \infty$ for any $j \in \mathcal{J}_-$. Hence, substituting $\delta_j \to \infty$ or $c_0 = 0$ into (A.5) yields

$${}^{\forall} j \in \mathcal{J}_-, \qquad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n\delta_j}\{\mathcal{L}(j) - \mathcal{L}(j_*)\} = 1 > 0.$$

Recall that $j_0 = j_*$ when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$. Hence, (A.6) is proved when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$. Consequently, Lemma 3 (2) is proved.

A.4. Proof of Lemma 4

We first give the lemma required for proving Lemma 4.

Lemma A.2 *Let X be a Bernoulli random variable with probability of success θ, and let h be a positive value such that*

$$\lim_{n\to\infty, p/n\to c_0} h = \infty.$$

Suppose that

$$\lim_{n\to\infty, p/n\to c_0} \theta = 0.$$

Then, without depending on the order of divergence of h, $hX \xrightarrow{p} 0$ *as* $n \to \infty$ *and* $p/n \to c_0 \in [0, 1)$.

(Proof of Lemma A.2) Notice that hX is a non-negative random variable. For any $\eta > 0$, we have

$$P(hX > \eta) = \begin{cases} \theta & (\text{when } \eta \leq h) \\ 0 & (\text{when } \eta > h) \end{cases}.$$

From the above result and the θ assumption, it follows that

$$P(hX > \eta) \leq \theta \to 0,$$

as $n \to \infty$ and $p/n \to c_0$. This means that $hX \xrightarrow{p} 0$ as $n \to \infty$ and $p/n \to c_0 \in [0, 1)$. Hence, Lemma A.2 is proved. ■

To permit a unified description of proofs for $\mathcal{L}(\hat{j}_{\mathrm{LE}})$ and $\mathcal{L}(\hat{j}_\alpha)$, we write the optimal model and the selected model as $\hat{j}$. Notice that $I(\hat{j} = j)$ is a Bernoulli random variable. It follows from the assumption that

$$^\forall j \in \mathcal{J}\backslash\{\ell\}, \quad \lim_{n\to\infty, p/n\to c_0} P(I(\hat{j} = j) = 1) = 0. \tag{A.8}$$

This indicates that for any $j \in \mathcal{J}\backslash\{\ell\}$, $I(\hat{j} = j) = o_p(1)$ as $n \to \infty$ and $p/n \to c_0$. From (A.8) and Lemma A.2, we have

$$^\forall j \in \mathcal{J}\backslash\{\ell\}, \ \ p \cdot I(\hat{j} = j) \xrightarrow{p} 0, \quad ^\forall j \in \mathcal{J}_-\backslash\{\ell\}, \ n\delta_j \cdot I(\hat{j} = j) \xrightarrow{p} 0, \tag{A.9}$$

as $n \to \infty$ and $p/n \to c_0$. Notice that $\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}})/p = O_p(1)$ as $n \to \infty$ and $p/n \to c_0$. It follows from this equation and (A.9) that for any $j \in \mathcal{J}\backslash\{\ell\}$

$$\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}) \cdot I(\hat{j} = j) = \frac{1}{p}\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}}) \cdot \left\{p \cdot I(\hat{j} = j)\right\} \xrightarrow{p} 0, \tag{A.10}$$

as $n \to \infty$ and $p/n \to c_0$. By using (A.9) and (A.10), we have

$$^\forall j \in \mathcal{J}\backslash\{\ell\}, \ \mathcal{L}(j)I(\hat{j} = j) \xrightarrow{p} 0,$$

as $n \to \infty$ and $p/n \to c_0$. Substituting the above equation into (15) yields

$$\mathcal{L}(\hat{j}) = \mathcal{L}(\ell)I(\hat{j} = \ell) + o_p(1) = \mathcal{L}(\ell) - \mathcal{L}(\ell)I(\hat{j} \neq \ell) + o_p(1), \tag{A.11}$$

as $n \to \infty$ and $p/n \to c_0$. Notice that $I(\hat{j} \neq \ell)$ is also a Bernoulli random variable having

$$\lim_{n\to\infty, p/n\to c_0} P(I(\hat{j} \neq \ell) = 1) = 0.$$

Therefore, from Lemma A.2, we can see that

$$\mathcal{L}(\ell)I(\hat{j} \neq \ell) \xrightarrow{p} 0,$$

as $n \to \infty$ and $p/n \to c_0$. Substituting the above equations into (A.11) yields

$$\mathcal{L}(\hat{j}) = \mathcal{L}(\ell) + o_p(1),$$

as $n \to \infty$ and $p/n \to c_0$. It follows from the above equation that $\mathcal{L}(\hat{j}_{\rm LE}) = \mathcal{L}(\ell) + o_p(1)$ and $\mathcal{L}(\hat{j}_\alpha) = \mathcal{L}(\ell) + o_p(1)$ as $n \to \infty$ and $p/n \to c_0$. These equations imply that $\mathcal{L}(\hat{j}_{\rm LE})/\mathcal{L}(\hat{j}_\alpha) = 1 + o_p(1)$ as $n \to \infty$ and $p/n \to c_0$. Hence, Lemma 4 is proved.

A.5. Proof of Lemma 5

We first give lemmas required for proving Lemma 5 (see Yanagihara, 2016a, for the proof).

Lemma A.3 *Suppose that Assumptions A1 and A2 hold. The GC_p criterion is consistent when $n \to \infty$ and $p/n \to c_0 \in [0, 1)$ if α satisfies Conditions C1 and C2.*

Lemma A.4 *Suppose that Assumptions A1 and A2 hold. When $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$, statistics based on transformed $\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\omega^{-1})$ converge to positive values in probability as $n \to \infty$ and $p/n \to c_0 \in [0, 1)$, i.e.,*

$$\forall j \in \mathcal{S}_- \cup \mathcal{J}_+, \quad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n}\left\{(n-k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\omega^{-1}) - (n-k)p\right\} = \frac{\delta_{j,0} + (k-k_j)c_0}{1-c_0},$$

$$\forall j \in \mathcal{S}_-^c \cap \mathcal{J}_-, \quad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n\delta_j}\left\{(n-k_j)\mathrm{tr}(\boldsymbol{S}_j\boldsymbol{S}_\omega^{-1}) - (n-k)p\right\} = 1,$$

where $\boldsymbol{S}_j$ is given by (5).

To prove Lemma 5, it suffices to show that the following equation is satisfied:

$$\forall j \in \{\{j_*\} \cup \mathcal{J}_-\}\backslash\{j_0\}, \quad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{h_j}\left\{GC_p(j|\alpha) - GC_p(j_0|\alpha)\right\} = \tau_j > 0, \tag{A.12}$$

(see, e.g., the result in Yanagihara, 2015), where h_j is some positive value depending on the model j, which does not converge to 0. Recall that $j_0 = j_*$ when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$, $j_0 = j_*$. Hence, when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$, the equations in (A.12) hold if α satisfies the conditions for consistency given in Lemma A.3, i.e., Conditions C1 and C2.

Next, we consider the case of $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$. Although it is stated in (19) that $\lambda(j_0)$ is the minimum among all $j \in \{j_*\} \cup \mathcal{S}_-$, we know that

$$\forall j \in \{\mathcal{S}_- \cup \mathcal{J}_+\}\backslash\{j_0\}, \quad \lambda(j_0) < \lambda(j), \tag{A.13}$$

where $\lambda(j)$ is given by (18). This is because $\lambda(j) = k_j c_0$ and $k_* < k_j$ holds for any $j \in \mathcal{J}_+\backslash\{j_*\}$. It follows from the above result, (A.13), Lemma A.4, and Condition C3 that

$$\forall j \in \{\mathcal{S}_- \cup \mathcal{J}_+\}\backslash\{j_0\}, \quad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n}\left\{GC_p(j|\alpha) - GC_p(j_0|\alpha)\right\} = \frac{\lambda(j) - \lambda(j_0)}{1-c_0} > 0,$$

$$\forall j \in \mathcal{S}_-^c \cap \mathcal{J}_-, \quad \operatorname*{plim}_{n\to\infty, p/n\to c_0} \frac{1}{n\delta_j}\left\{GC_p(j|\alpha) - GC_p(j_0|\alpha)\right\} = 1 > 0.$$

Consequently, Lemma 5 is proved.

A.6. Proof of Lemma 6

We first give a lemma required for proving Lemmas 6 and 8.

Lemma A.5 *Let X be a Bernoulli random variable with probability of success θ satisfying*

$$\lim_{n\to\infty, p/n\to c_0} \theta = 0.$$

Then, for any $j \in \mathcal{J}$, $E[\mathcal{L}(j)X]/p$ can be expanded as

$$\frac{1}{p}E[\mathcal{L}(j)X] = \frac{n}{p}\delta_j\theta + o(1),$$

as $n \to \infty$ and $p/n \to c_0$.

(Proof of Lemma A.5) Notice that

$$\begin{aligned} &E\left[p^{-1}\mathcal{L}(j)\right] = p^{-1}n\delta_j + k_j, && E[X] = \theta, \\ &Var\left[p^{-1}\mathcal{L}(j)\right] = Var\left[p^{-1}\mathrm{tr}(\boldsymbol{\mathcal{E}}'\boldsymbol{P}_j\boldsymbol{\mathcal{E}})\right] = 2p^{-1}k_j, && Var[X] = \theta(1-\theta). \end{aligned}$$

From the above equations and the θ assumption, it follows that

$$\begin{aligned} &\left|E\left[p^{-1}\mathcal{L}(j)X\right] - E\left[p^{-1}\mathcal{L}(j)\right]E[X]\right| = \left|Cov\left[p^{-1}\mathcal{L}(j), X\right]\right| \\ &\leq \sqrt{Var\left[p^{-1}\mathcal{L}(j)\right]Var[X]} = \sqrt{2p^{-1}k_j\theta(1-\theta)} \to 0, \end{aligned}$$

as $n \to \infty$ and $p/n \to c_0$. Hence, we derive

$$\frac{1}{p}E\left[\mathcal{L}(j)X\right] = \frac{1}{p}E\left[\mathcal{L}(j)\right]E[X] + o(1) = \left(\frac{n}{p}\delta_j + k_j\right)\theta + o(1) = \frac{n}{p}\delta_j\theta + o(1),$$

as $n \to \infty$ and $p/n \to c_0$. Consequently, Lemma A.5 is proved. ■

Since $\mathcal{R}(j_{\mathrm{ME}})$ is the minimum value of $\mathcal{R}(j)$ among $j \in \mathcal{J}$, $\mathcal{R}(j_{\mathrm{ME}}) \leq \mathcal{R}(j_0)$ holds. On the other hand, $\mathcal{L}(\hat{j}_{\mathrm{LE}})$ is the minimum value of $\mathcal{L}(j)$ among $j \in \mathcal{J}$, $\mathcal{L}(\hat{j}_{\mathrm{LE}}) \leq \mathcal{L}(j_{\mathrm{ME}})$ holds. This implies that $E[\mathcal{L}(\hat{j}_{\mathrm{LE}})] \leq E[\mathcal{L}(j_{\mathrm{ME}})] = \mathcal{R}(j_{\mathrm{ME}})$. Therefore, we have

$$E\left[\mathcal{L}(\hat{j}_{\mathrm{LE}})\right] \leq \mathcal{R}(j_{\mathrm{ME}}) \leq \mathcal{R}(j_0). \tag{A.14}$$

By using the equation in (15) and j_0 in (19), we derive

$$\mathcal{L}(j_0) \geq \mathcal{L}(\hat{j}_{\mathrm{LE}}) = \sum_{j\in\mathcal{J}} \mathcal{L}(j)I(\hat{j}_{\mathrm{LE}} = j).$$

This implies that

$$\begin{aligned} &\mathcal{L}(j_0)\left\{1 - I(\hat{j}_{\mathrm{LE}} = j_0)\right\} \geq \sum_{j\in\mathcal{J}\backslash\{j_0\}} \mathcal{L}(j)I(\hat{j}_{\mathrm{LE}} = j) \\ \iff\ &\mathcal{L}(j_0)I(\hat{j}_{\mathrm{LE}} \neq j_0) \geq \sum_{j\in\mathcal{J}\backslash\{j_0\}} \mathcal{L}(j)I(\hat{j}_{\mathrm{LE}} = j). \end{aligned} \tag{A.15}$$

It follows from Theorem 3 that $P(\hat{j}_{\rm LE} \neq j_0) \to 0$ as $n \to \infty$ and $p/n \to c_0$. Since $I(\hat{j}_{\rm LE} \neq j_0) \sim B(1, P(\hat{j}_{\rm LE} \neq j_0))$, by using Lemma A.5, we have

$$\frac{1}{p}E\left[\mathcal{L}(j_0)I(\hat{j}_{\rm LE} \neq j_0)\right] = \frac{n}{p}\delta_{j_0}P(\hat{j}_{\rm LE} \neq j_0) + o(1),$$

as $n \to \infty$ and $p/n \to c_0$. Notice that

$$\frac{n}{p}\delta_{j_0} = \begin{cases} c_0^{-1}\delta_{j_0,0} + o(1) & (\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}) \\ 0 & (\{\xi_* = \emptyset\} \cup \{c_0 = 0\}) \end{cases}. \tag{A.16}$$

Since $0 < c_0^{-1}\delta_{j_0,0} < \infty$, we have

$$\frac{1}{p}E\left[\mathcal{L}(j_0)I(\hat{j}_{\rm LE} \neq j_0)\right] \to 0, \tag{A.17}$$

as $n \to \infty$ and $p/n \to c_0$. It follows from (A.15) and (A.17) that

$$\frac{1}{p}\sum_{j \in \mathcal{J}\backslash\{j_0\}} E\left[\mathcal{L}(j)I(\hat{j}_{\rm LE} = j)\right] \to 0,$$

as $n \to \infty$ and $p/n \to c_0$. This implies that

$$\begin{aligned} \frac{1}{p}E\left[\mathcal{L}(\hat{j}_{\rm LE})\right] &= \frac{1}{p}E\left[\mathcal{L}(j_0)I(\hat{j}_{\rm LE} = j_0)\right] + o(1) \\ &= \frac{1}{p}E\left[\mathcal{L}(j_0)\right] - \frac{1}{p}E\left[\mathcal{L}(j_0)I(\hat{j}_{\rm LE} \neq j_0)\right] + o(1) \\ &= \frac{1}{p}\mathcal{R}(j_0) + o(1), \end{aligned}$$

as $n \to \infty$ and $p/n \to c_0$. By using the above result and (A.14), Lemma 6 is proved.

A.7. Proof of Lemma 7

We first give a lemma required for proving Lemma 7.

Lemma A.6 *Let $\mathcal{W}_-$ be a set of underspecified models which satisfy equation* (16)*, i.e.,*

$$\mathcal{W}_- = \left\{ j \in \mathcal{J}_- \,\middle|\, \lim_{n\to\infty, p/n\to c_0} \frac{p}{n\delta_j}\left(-\frac{n}{n-p} + \alpha\right) = 0 \right\}.$$

Suppose that Assumptions A1 and A2 hold. The order of the selection probability of $j \in \mathcal{W}_-$ by the GC_p is given by

$$P(\hat{j}_\alpha = j) = O(n^{-r}),$$

where r is any positive integer.

(Proof of Lemma A.6) Let $q_j = k_{j^c \cap j_*}$ for any $j \in \mathcal{J}_-$, and let $V, U_1, \ldots, U_{q_j}$ be mutually independent random variables defined by $V \sim \chi_m^2$ and $U_i \sim \chi^2(n\delta_{j,i})$, where $\delta_{j,i}$ is a positive value satisfying $\delta_{j,1} + \cdots + \delta_{j,q_j} = \delta_j$ and $\delta_{i,j} \geq \delta_j/q_j^2$ for all $i \in \{1, \ldots, q_j\}$. Here, m is given in (8) and δ_j is given by (6). Suppose that Assumption A1 holds. Then, from Yanagihara (2016b) or Oda and Yanagihara

(2019), the probability of selecting $j \in \mathcal{J}_-$ by a GC_p is bounded as

$$P(\hat{j}_\alpha = j) \leq \sum_{i=1}^{q_j} P(U_i/V \leq p\alpha/(n-k)). \tag{A.18}$$

Let $\gamma_{j,i}$ be a constant defined by

$$\gamma_{j,i} = \frac{1}{\delta_{j,i}}\left(\frac{p+n\delta_{j,i}}{m-2} - \frac{p\alpha}{n-k}\right) \quad (j \in \mathcal{J}_-, i = 1,\ldots,q_j).$$

From a simple calculation, we have

$$\gamma_{j,i} = \frac{n}{m-2} + \frac{p(n+kp)}{(m-2)(n-k)(n-p)\delta_{j,i}} - \frac{p}{(n-k)\delta_{j,i}}\left(-\frac{n}{n-p} + \alpha\right). \tag{A.19}$$

It follows from the relations $\delta_{j,i} \geq \delta_j/q_j^2$ and $q_j \leq k_*$, the boundedness of $1/\delta_j$, and the definition of $\mathcal{W}_-$ that

$$\lim_{n\to\infty, p/n\to c_0} \frac{p(n+kp)}{(m-2)(n-k)(n-p)\delta_{j,i}} = 0, \quad \lim_{n\to\infty, p/n\to c_0} \frac{p}{(n-k)\delta_{j,i}}\left(-\frac{n}{n-p} + \alpha\right) = 0. \tag{A.20}$$

Using equations (A.19) and (A.20) yields

$$\lim_{n\to\infty, p/n\to c_0} \gamma_{j,i} = \frac{1}{1-c_0} > 0.$$

Hence, $\gamma_{j,i}$ is positive for sufficiently large n or large n and p. This implies that for sufficiently large n or large n and p

$$\begin{aligned} P\left(\frac{U_i}{V} \leq \frac{p\alpha}{n-k}\right) &= P\left(\frac{1}{\delta_{j,i}}\left(\frac{U_i}{V} - \frac{p+n\delta_{j,i}}{m-2}\right) \leq -\gamma_{j,i}\right) \\ &\leq P\left(\frac{1}{\delta_{j,i}}\left|\frac{U_i}{V} - \frac{p+n\delta_{j,i}}{m-2}\right| \geq \gamma_{j,i}\right) \leq \frac{1}{(\gamma_{j,i}\delta_{j,i})^{2r}} E\left[\left(\frac{U_i}{V} - \frac{p+n\delta_{j,i}}{m-2}\right)^{2r}\right], \end{aligned} \tag{A.21}$$

where r is any positive integer. The results in Oda and Yanagihara (2019) indicate that

$$E\left[\left(\frac{U_i}{V} - \frac{p+n\delta_{j,i}}{m-2}\right)^{2r}\right] = O\left(n^{-r}\delta_{j,i}^{2r}\right). \tag{A.22}$$

Substituting (A.21) and (A.22) into (A.18) yields $P(\hat{j}_\alpha = j) = O(n^{-r})$. Consequently, Lemma A.6 is proved. ■

If α satisfies Condition C2, $\mathcal{W}_- = \mathcal{J}_-$ holds. On the other hand, when α satisfies Condition C3, for equation (16) to hold, it must be the case that $\delta_j = \infty$. Hence, if α satisfies Condition C3, $\mathcal{W}_- = \mathcal{S}_-^c \cap \mathcal{J}_-$ holds. Recall that r in Lemma A.6 is any positive integer. From Lemma A.6, the selection probability of $j \in \mathcal{W}_-$ can be expressed as

$$P(\hat{j}_\alpha = j) = \frac{M_j}{n^{\epsilon+1}},$$

where M_j is a constant independent of n and p, and ϵ is the positive integer defined in Assumption A3. This implies that

$$\frac{n}{p}\delta_j P(\hat{j}_\alpha = j) = \frac{1}{p}\frac{\delta_j}{n^\epsilon} M_j = 0,$$

as $n \to \infty$ and $p/n \to c_0$. Consequently, Lemma 7 is proved.

A.8. Proof of Lemma 8

From Lemma 5, if α satisfies Conditions C1 and C2 when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$, or Condition C3 when $\{\xi_* \neq \emptyset\} \cap \{c_0 \neq 0\}$, the following equations are satisfied:

$$\lim_{n\to\infty, p/n\to c_0} P(\hat{j}_\alpha \neq j_0) = 0, \qquad \lim_{n\to\infty, p/n\to c_0} P(\hat{j}_\alpha = j) = 0 \ ({}^\forall j \in \mathcal{J}\backslash\{j_0\}),$$

where j_0 is the asymptotically loss optimal model given by (19). Notice that

$$I(\hat{j}_\alpha \neq j_0) \sim B\left(1, P(\hat{j}_\alpha \neq j_0)\right), \quad I(\hat{j}_\alpha = j) \sim B\left(1, P(\hat{j}_\alpha = j)\right), \ ({}^\forall j \in \mathcal{J}\backslash\{j_0\}).$$

By using Lemma A.5 and (A.16), we have

$$\frac{1}{p}E\left[\mathcal{L}(j_0)I(\hat{j}_\alpha \neq j_0)\right] = \frac{n}{p}\delta_{j_0}P(\hat{j}_\alpha \neq j_0) + o(1) = o(1) \text{ as } n \to \infty, \ p/n \to c_0. \tag{A.23}$$

Recall that $\delta_j = 0$ holds for any $j \in \mathcal{J}_+$, or $np^{-1}\delta_j = O(1)$ as $n \to \infty$ and $p/n \to c_0$ for any $j \in \mathcal{S}_-$ when $c_0 \neq 0$, From these equations and Lemma 7, we can see that

$$\begin{aligned}\frac{n}{p}\delta_j P(\hat{j}_\alpha = j) &= \begin{cases} 0 & ({}^\forall j \in \mathcal{J}_+\backslash\{j_0\}) \\ o(1) & \begin{pmatrix} {}^\forall j \in \mathcal{S}_-\backslash\{j_0\} & \text{when } c_0 \neq 0 \\ {}^\forall j \in \mathcal{S}_-^c \cap \mathcal{J}_- & \text{when } c_0 \neq 0 \\ {}^\forall j \in \mathcal{J}_- & \text{when } c_0 = 0 \end{pmatrix} \end{cases} \\ &= \begin{cases} 0 & ({}^\forall j \in \mathcal{J}_+\backslash\{j_0\}) \\ o(1) & ({}^\forall j \in \mathcal{J}_-\backslash\{j_0\}) \end{cases},\end{aligned}$$

as $n \to \infty$ and $p/n \to c_0$. It follows from the equation and Lemma A.5 that for any $j \in \mathcal{J}\backslash\{j_0\}$

$$\frac{1}{p}E\left[\mathcal{L}(j)I(\hat{j}_\alpha = j)\right] = \frac{n}{p}\delta_j P(\hat{j}_\alpha = j) + o(1) = o(1), \tag{A.24}$$

as $n \to \infty$ and $p/n \to c_0$. Notice that

$$\begin{aligned}\frac{1}{p}E\left[\mathcal{L}(\hat{j}_\alpha)\right] &= \frac{1}{p}\sum_{j\in\mathcal{J}} E\left[\mathcal{L}(j)I(\hat{j}_\alpha = j)\right] \\ &= \frac{1}{p}\mathcal{R}(j_0) - \frac{1}{p}E\left[\mathcal{L}(j_0)I(\hat{j}_\alpha \neq j_0)\right] + \frac{1}{p}\sum_{j\in\mathcal{J}\backslash\{j_0\}} E\left[\mathcal{L}(j)I(\hat{j}_\alpha = j)\right].\end{aligned}$$

By substituting (A.23) and (A.24) into the above equation, Lemma 8 is proved.

A.9. Proof of Corollary 1

To prove Corollary 1, it is sufficient to show that α satisfying Condition C3 satisfies Condition C2 when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$. Recall that $\xi_* = \emptyset \Leftrightarrow \delta_j \to \infty$ for any $j \in \mathcal{J}_-$. Notice that $p/(n\delta_j)$ and $p\alpha/(n\delta_j)$ converge to 0 when $\delta_j \to \infty$ or $c_0 = 0$. These convergences to 0 and $n/(n-p) \to 1/(1-c_0)$ imply that

$$\lim_{n\to\infty, p/n\to c_0} \frac{p}{n\delta_j}\left(-\frac{n}{n-p} + \alpha\right) = 0.$$

This indicates that Condition C2 holds when $\{\xi_* = \emptyset\} \cup \{c_0 = 0\}$. Consequently, Corollary 1 is proved.