Asymptotic optimality of C_p-type criteria in high-dimensional multivariate linear regression models

(Last Modified: July 24, 2021)

Shinpei Imori

Graduate School of Advanced Science and Engineering, Hiroshima University

1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan

imori@hiroshima-u.ac.jp

Abstract

We study the asymptotic optimality of C_p-type criteria from the perspective of prediction in high-dimensional multivariate linear regression models, where the dimension of a response matrix is large but does not exceed the sample size. We derive conditions in order that the generalized C_p (GC_p) exhibits asymptotic loss efficiency (ALE) and asymptotic mean efficiency (AME) in such high-dimensional data. Moreover, we clarify that one of the conditions is necessary for GC_p to exhibit both ALE and AME. As a result, it is shown that the modified C_p can claim both ALE and AME but the original C_p cannot in high-dimensional data. The finite sample performance of GC_p with several tuning parameters is compared through a simulation study.

Key words: Asymptotic theory; High-dimensional statistical inference; Model selection/variable selection.

1 Introduction

Variable selection problems are crucial in statistical fields to improve prediction accuracy and/or interpretability of a resultant model. There is a burgeoning literature which has attempted to solve the variable selection problem, and many selection procedures and their theoretical properties have been studied.

For example, Mallows' C_p criterion (Mallows 1973) and Akaike information criterion (AIC) (Akaike 1974) are known as useful selection methods from a predictive point of view because these procedures are optimal in some predictive sense (see Shibata 1981, 1983, Li 1987, Shao 1997). On the other hand, Bayesian information criterion (BIC) proposed by Schwarz (1978) is consistent (Nishii 1984) under appropriate conditions; that is, the probability that a model selected by BIC coincides with the true model converges to 1 as the sample size n tends to infinity. In this sense, BIC would be a feasible method from the perspective of interpretability. However, C_p and AIC are inconsistent (Nishii 1984) under the same condition. Details of properties of selection procedures are well studied in Shao (1997) in the context of univariate linear regression models. However, here, our target is multivariate linear regression models.

Recently, high-dimensional data are often encountered where the dimension of a response matrix in multivariate linear regression models p_n is large, whereas p_n does not exceed the sample size n. Considering such high-dimensional multivariate linear regression models, one may presume that the properties of selection methods such as optimality and consistency are inherited from univariate models. However, interestingly, properties derived when p_n is fixed can be altered in high-dimensional situations. For example, Yanagihara, Wakaki and Fujikoshi (2015) showed that AIC acquires the consistency property and that BIC loses its consistency in high-dimensional data. Similar results for C_p-type criteria were reported by Fujikoshi, Sakurai and Yanagihara (2014). The reason why this inversion arises may be that a difference in risks between two over-specified models (i.e., models including the true model) diverges with n and p_n tending to infinity, and thus penalty terms of C_p and AIC are moderate but that of BIC is too strong. In addition to these studies, model selection criteria in high-dimensional data contexts and their consistency properties have been vigorously studied in various models and situations (e.g., Katayama and Imori 2014, Imori and von Rosen 2015, Yanagihara 2015, Fujikoshi and Sakurai 2016, Bai, Choi and Fujikoshi 2018).

Compared with the consistency property, asymptotic optimality for prediction in high-dimensional data contexts is under-researched. Conventional results derived from univariate models are no longer reliable in high-dimensional data contexts, and extension to such cases is not mathematically trivial. In the present paper, we focus on asymptotic loss efficiency (ALE) (Li 1987, Shao 1997) and asymptotic mean efficiency (AME) (Shibata 1983) as criteria for the asymptotic optimality of variable selection. We derive sufficient conditions in order that a generalization of C_p (GC_p) exhibits ALE and AME in high-dimensional data. We also show that one of the sufficient conditions is necessary for GC_p to exhibit both of these efficiencies. As a result, we can observe that the modified C_p (MC_p) introduced by Fujikoshi and Satoh (1997) exhibits ALE and AME assuming moderate conditions although the original C_p does not under the same conditions.

Recently, Yanagihara (2020) also studied ALE and AME of GC_p in high-dimensional multivariate linear regression models although its conditions and results are based on the consistency property. For example, Yanagihara (2020) supposes that the true model is included in a set of candidate models, which is not assumed in the present paper. It is worth mentioning that previous studies of variable selection in multivariate linear regression models use a common regression model among response variables. We mitigate this limitation and allow each response variable to have different models in order to consider more practical situations such as response variables have a group structure.

The remainder of this paper is composed as follows. In Section 2, we clarify the variable selection framework used in this paper. In Section 3, the sufficient conditions for ALE and AME of GC_p are given. In Section 4, we study the asymptotic inefficiency of GC_p. Section 5 illustrates the finite sample performances of some C_p-type criteria. Finally, conclusions are offered in Section 6.

2 Model selection framework

2.1 True and candidate models

Let $\boldsymbol{Y}$ be an $n \times p_n$ response variable matrix and $\boldsymbol{X}$ be an $n \times k_n$ explanatory variable matrix, where n is the sample size, p_n is the dimension of response and k_n is the number of explanatory variables. We assume $\boldsymbol{X}$ to be of full rank and non-stochastic. We allow k_n and p_n to diverge to infinity with n tending to infinity, although neither k_n nor p_n exceeds n. Specific conditions for n, k_n, and p_n are given later.

The true distribution of $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{p_n})$ is given by

$$\boldsymbol{Y} = \boldsymbol{\Gamma}_* + \boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2},$$

where $\boldsymbol{\Gamma}_* = (\gamma_1^*, \ldots, \gamma_{p_n}^*) = E(\boldsymbol{Y})$, $\boldsymbol{\mathcal{E}}$ is an $n \times p_n$ error matrix, of which all entries are independent and identically distributed as the standard normal distribution $N(0, 1)$ and $\boldsymbol{\Sigma}_*$ is the true covariance matrix of each row of $\boldsymbol{Y}$. The relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is represented by a multivariate linear regression model as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}^{1/2},$$

where $\boldsymbol{B}$ is a $k_n \times p_n$ matrix of unknown regression coefficients and $\boldsymbol{\Sigma}$ is a $p_n \times p_n$ unknown covariance matrix. Here, we distinguish the covariance parameter $\boldsymbol{\Sigma}$ from the true one $\boldsymbol{\Sigma}_*$. Let $M = (M_1, \ldots, M_{p_n})$, where $\emptyset \neq M_j \subset M_F = \{1, \ldots, k_n\}$ is a candidate model for the jth response variable $\boldsymbol{y}_j$, that is, we assume $\boldsymbol{y}_j$ is relevant to $\boldsymbol{X}_{M_j}$ that is an $n \times k_{M_j}$ sub-matrix of $\boldsymbol{X}$ corresponding to M_j, and k_{M_j} is the cardinality of M_j. This setting can take account of a group structure of response variables. For example, if we have two groups $\{1, \ldots, m\}$ and $\{m+1, \ldots, p_n\}$ with some integer m, a restriction $M_1 = \ldots = M_m$ and $M_{m+1} = \ldots = M_{p_n}$ will be imposed. Using only one regression model for response variables, i.e., $M_1 = \ldots = M_{p_n}$, we have a simple variable selection problem often considered in previous studies. Then, a candidate model M implies a multivariate linear regression model defined as follows:

$$\boldsymbol{y}_j = \boldsymbol{X}_{M_j}\boldsymbol{\beta}_{M_j} + \boldsymbol{\varepsilon}_j, \quad j = 1, \ldots, p_n,$$

where $\boldsymbol{\beta}_{M_j}$ is a k_{M_j}-dimensional vector of unknown regression coefficients and $\boldsymbol{\varepsilon}_j$ is the jth column of

$\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}$, i.e., $\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_{p_n})$. Thus, a set of candidate models is denoted by $\mathcal{M}_n$ that is a subset of a comprehensive set $\{M = (M_1, \ldots, M_{p_n}) | M_j \subset M_F, j = 1, \ldots, p_n\}$. Note that $\mathcal{M}_n$ does not have to include the full model, i.e., $M = (M_F, \ldots, M_F)$.

2.2 Loss and risk functions

Herein, the goodness of fit of a candidate model M is measured by a quadratic loss function L_n given by

$$L_n(M) = \mathrm{tr}\{(\boldsymbol{\Gamma}_* - \hat{\boldsymbol{\Gamma}}(M))\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\Gamma}_* - \hat{\boldsymbol{\Gamma}}(M))^\top\}, \tag{1}$$

where each column of $\hat{\boldsymbol{\Gamma}}(M)$ is obtained based on a least squares estimator, i.e.,

$$\hat{\boldsymbol{\Gamma}}(M) = (\boldsymbol{P}_{M_1}\boldsymbol{y}_1, \ldots, \boldsymbol{P}_{M_{p_n}}\boldsymbol{y}_{p_n}), \tag{2}$$

and $\boldsymbol{P}_{M_j} = \boldsymbol{X}_{M_j}(\boldsymbol{X}_{M_j}^\top \boldsymbol{X}_{M_j})^{-1}\boldsymbol{X}_{M_j}^\top$. By substituting (2) into (1), we have

$$L_n(M) = \mathrm{tr}\{\boldsymbol{\Delta}(M)\} - 2\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}}(M)\} + \mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\mathcal{E}}(M)^\top \boldsymbol{\mathcal{E}}(M)\} \tag{3}$$

where $\boldsymbol{\Delta}(M) = \boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))\boldsymbol{\Sigma}_*^{-1/2}$, $\boldsymbol{\Gamma}_*(M) = (\boldsymbol{P}_{M_1}\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{P}_{M_{p_n}}\boldsymbol{\gamma}_{p_n}^*)$ and $\boldsymbol{\mathcal{E}}(M) = (\boldsymbol{P}_{M_1}\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{P}_{M_{p_n}}\boldsymbol{\varepsilon}_{p_n})$. Then, a risk function R_n is obtained as

$$R_n(M) = E(L_n(M)) = \mathrm{tr}\{\boldsymbol{\Delta}(M)\} + \mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\}, \tag{4}$$

where $\boldsymbol{A}(M) = (\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{I}_n)\boldsymbol{P}(M)(\boldsymbol{\Sigma}_*^{1/2} \otimes \boldsymbol{I}_n)$, a symbol $\otimes$ denotes a Kronecker product and $\boldsymbol{P}(M) = \mathrm{diag}\{\boldsymbol{P}_{M_1}, \ldots, \boldsymbol{P}_{M_{p_n}}\}$. It is worth mentioning that $\boldsymbol{A}(M)$ is an idempotent matrix. Thus, from Householder and Carpenter (1963), $\sigma_j(\boldsymbol{A}(M)) \le \sigma_j(\boldsymbol{A}(M))^2$ for all $j = 1, \ldots, p_n$, where $\sigma_j(\cdot)$ denotes the jth largest singular value. This and Theorem 3.3.13 in Horn and Jornson (1994) indicate that

$$\mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\} = \sum_{j=1}^{p_n} \sigma_j(\boldsymbol{A}(M))^2 \ge \sum_{j=1}^{p_n} \sigma_j(\boldsymbol{A}(M)) \ge \mathrm{tr}\{\boldsymbol{A}(M)\}.$$

This implies that $R_n(M) \ge p_n$ because $\mathrm{tr}\{\boldsymbol{A}(M)\} = \sum_{j=1}^{p_n} k_{M_j}$.

The best models with respect to the loss and risk functions are denoted by M_L^* and M_R^*, which minimize (1) and (4) among $\mathcal{M}_n$, respectively, i.e.,

$$M_L^* = \arg\min_{M \in \mathcal{M}_n} L_n(M), \quad M_R^* = \arg\min_{M \in \mathcal{M}_n} R_n(M).$$

Note that M_L^* is a random variable, M_R^* is non-stochastic, and both of them depend on n although they are suppressed for brevity.

2.3 Selection method and asymptotic efficiency

To select the best model among $\mathcal{M}_n$, we use GC_p defined by

$$GC_p(M; \alpha_n) = n\alpha_n \mathrm{tr}\{\hat{\boldsymbol{\Sigma}}(M)\boldsymbol{S}^{-1}\} + 2\sum_{j=1}^{p_n} k_{M_j}. \tag{5}$$

where α_n is a positive sequence, $\hat{\boldsymbol{\Sigma}}(M) = (\boldsymbol{Y} - \hat{\boldsymbol{\Gamma}}(M))^\top(\boldsymbol{Y} - \hat{\boldsymbol{\Gamma}}(M))/n$, $\boldsymbol{S} = \boldsymbol{Y}^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{Y}/(n - k_n)$ and $\boldsymbol{P}_{M_F}^\perp = \boldsymbol{I}_n - \boldsymbol{P}_{M_F}$. For theoretical purposes, we use α_n satisfying

$$\lim_{n\to\infty} \alpha_n = a \in [0, \infty).$$

When $\alpha_n = 1$ and $p_n = 1$, GC_p indicates C_p proposed by Mallows (1973). When $\alpha_n = 1-(p_n+1)/(n-k_n)$ and $M_1 = \cdots = M_{p_n}$, selection results by GC_p coincide with the modified C_p (called MC_p) by Fujikoshi and Satoh (1997). If the full model includes the true model and we set $M_1 = \cdots = M_{p_n}$, then MC_p is an unbiased estimator (Fujikoshi and Satoh 1997). Note that Atkinson (1980) introduced a criterion equivalent to GC_p for univariate data, and Nagai, Yanagihara and Satoh (2012) proposed for multivariate generalized ridge regression models although they assumed $M_1 = \cdots = M_{p_n}$.

The best model selected by minimizing GC_p among $\mathcal{M}_n$ is denoted by $\hat{M}_n$, i.e.,

$$\hat{M}_n = \arg\min_{M\in\mathcal{M}_n} GC_p(M; \alpha_n).$$

Then, we state that GC_p exhibits ALE (Li 1987, Shao 1997) if

$$\frac{L_n(\hat{M}_n)}{L_n(M_L^*)} \xrightarrow{p} 1, \quad n \to \infty, \tag{6}$$

and exhibits AME (Shibata 1983) if

$$\lim_{n\to\infty} \frac{E(L_n(\hat{M}_n))}{R_n(M_R^*)} = 1. \tag{7}$$

Note that $L_n(\hat{M}_n)$ and $E(L_n(\hat{M}_n))$ are respectively referred to as loss and risk functions of the best model selected by GC_p.

3 Asymptotic efficiency of GC_p

In this section, we present ALE and AME of $\mathrm{GC}_p(M;\alpha_n)$. Hereafter, we may omit symbol "$n\to\infty$" for simplifying expressions.

Firstly, we assume the following conditions for ALE:

(C1) $\lim_{n\to\infty} k_n/n = c_k \in [0,1)$, $\lim_{n\to\infty} p_n/n = c_p \in [0,1)$, $1-c_k-c_p>0$ and $n-k_n-p_n>0$.

(C2) $\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2}) = o(n)$.

(C3) There exists a constant $C_A \geq 1$ such that for all $M \in \mathcal{M}_n$, $\sigma_1(\boldsymbol{A}(M)) \leq C_A$.

(C4) For all $\delta \in (0,1)$, $\lim_{n\to\infty}\sum_{M\in\mathcal{M}_n} \delta^{R_n(M)} = 0$.

(C5) Let $\#(\mathcal{M}_n)$ be the cardinality of $\mathcal{M}_n$, i.e., the number of candidate models. Then, $\log\#(\mathcal{M}_n) = o(n)$.

The first part of condition (C1) is weaker than a condition assumed in Shibata (1981, 1983) if the full model $(M_F,\ldots,M_F)$ is included in the set of candidate models $\mathcal{M}_n$. The second part of (C1) constructs our high-dimensional framework, which is also considered in previous studies (see e.g., Fujikoshi, Sakurai and Yanagihara 2014, Yanagihara, Wakaki and Fujikoshi 2015). The third part is used for evaluating the lowest singular values of a high-dimensional Gaussian random matrix. The final part of (C1) is required to guarantee regularity of $\boldsymbol{S}$, which can be satisfied asymptotically from the previous three conditions. Condition (C2) is used to ignore an effect of $\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2})$, which is satisfied when $\boldsymbol{\Gamma}_*$ is well approximated by a linear regression model $\boldsymbol{XB}$ although a set of candidate models does not need to include the true model. When $p_n=1$, (C2) corresponds to an assumption in Shao (1997). Condition (C3) is only considered when we do not use a common model for response variables. Actually, $M=(M_1,\ldots,M_1)$ with some $M_1\subset M_F$ indicates that $\boldsymbol{A}(M)=\boldsymbol{I}_{p_n}\otimes\boldsymbol{P}_{M_1}$, and thus (C3) holds. If there exists $\lambda\geq 1$ such that $\lambda^{-1}\leq\lambda_{\min}(\boldsymbol{\Sigma}_*)\leq\lambda_{\max}(\boldsymbol{\Sigma}_*)\leq\lambda$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues, then (C3) holds for any $\mathcal{M}_n$ because for all $\boldsymbol{x}\in\mathbb{R}^{np_n}$,

$$\boldsymbol{x}^\top\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\boldsymbol{x} \leq \frac{\lambda_{\max}(\boldsymbol{\Sigma}_*)}{\lambda_{\min}(\boldsymbol{\Sigma}_*)}\boldsymbol{x}^\top\boldsymbol{x}.$$

On the other hand, conditions (C4) and (C5) control the number of candidate models. When $p_n=1$, (C4) corresponds to a condition in Shibata (1981, 1983). Let G be a positive constant integer. Suppose that response variables has G groups and each group consists of at least g_n response variables, where g_n satisfies $p_n=O(g_n)$. Then, when $p_n\to\infty$, $\log k_n = o(p_n)$ is a sufficient condition for (C4) because this indicates that $\log k_n = o(g_n)$ and

$$\sum_{M \in \mathcal{M}_n} \delta^{R_n(M)} \leq \left\{ \sum_{j=1}^{k_n} \binom{k_n}{j} \delta^{jg_n} \right\}^G \leq \left\{ \sum_{j=1}^{k_n} (k_n \delta^{g_n})^j \right\}^G \leq \left(\frac{k_n \delta^{g_n}}{1 - k_n \delta^{g_n}} \right)^G.$$

Hence, this may suggest that as p_n grows, the upper bound the number of candidate models (or the number of explanatory variables) for satisfying (C4) becomes large. Note that when $c_p > 0$, (C4) always holds due to (C5). Condition (C5) would be satisfied in actual use because violation of (C5) induces a huge computational burden.

Then, we can derive sufficient conditions for ALE of GC_p as the following theorem, of which a proof is given in Supplementary Materials.

Theorem 3.1. Suppose that conditions (C1)–(C5) hold. If $\alpha_n \to a = 1 - c_p/(1 - c_k)$ as $n \to \infty$, then $GC_p(M; \alpha_n)$ exhibits ALE, i.e.,

$$\frac{L_n(\hat{M}_n)}{L_n(M_L^*)} \xrightarrow{p} 1, \quad n \to \infty.$$

Next, we show AME of GC_p. Besides conditions (C1)–(C5), we assume the following condition:

(C6) There exists $\gamma_0 \in (0, 1)$ such that

$$\max_{M \in \mathcal{M}_n} \frac{R_n(M)}{R_n(M_R^*)} = O(\exp(n^{\gamma_0})).$$

Condition (C6) sets an upper bound of the risk ratio $R_n(M)/R_n(M_R^*)$, which prevents the maximum risk from being too large. Let us show that if there exist constants $C \geq 1$ and $\gamma \in [0, 1)$ such that $\lambda_{\min}(\boldsymbol{\Sigma}_*) \geq C \exp(-n^\gamma) > 0$ and $(\boldsymbol{\Gamma}_*)_{ij}^2 \leq C$ for all $1 \leq i \leq n$ and $1 \leq j \leq p_n$, then (C6) holds under (C1) and (C3). Conditions (C1) and (C3) indicates that

$$\begin{aligned}
R_n(M) &= \mathrm{tr}\{\boldsymbol{\Delta}(M)\} + \mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\} \\
&\leq \mathrm{vec}(\boldsymbol{\Gamma}_*)^\top (\boldsymbol{I}_{np_n} - \boldsymbol{P}(M))(\boldsymbol{\Sigma}_*^{-1} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np_n} - \boldsymbol{P}(M))\mathrm{vec}(\boldsymbol{\Gamma}_*) + C_A^2 np_n \\
&\leq np_n \{\lambda_{\min}(\boldsymbol{\Sigma}_*)^{-1} \max\{(\boldsymbol{\Gamma}_*)_{ij}^2 | 1 \leq i \leq n, 1 \leq j \leq p_n\} + C_A^2\} \\
&= O(n^2 \exp(n^\gamma)).
\end{aligned}$$

We have shown that for all $M \in \mathcal{M}_n$, $R_n(M) \geq p_n$ and especially, $R_n(M_R^*) \geq p_n$. Thus, by setting $\gamma_0 = (1 + \gamma)/2$, (C6) is satisfied.

Assuming (C1)–(C6), we have the following theorem:

Theorem 3.2. Suppose that conditions (C1)–(C6) hold. If $\alpha_n \to a = 1 - c_p/(1-c_k)$ as $n \to \infty$, then $GC_p(M; \alpha_n)$ exhibits AME, i.e.,

$$\lim_{n\to\infty} \frac{E(L_n(\hat{M}_n))}{R_n(M_R^*)} = 1.$$

A proof of this theorem is provided in Supplementary Materials. For both ALE and AME of GC_p, we assume $\alpha_n \to a = 1 - c_p/(1-c_k)$. Unless $c_p = 0$, this condition does not hold when $\alpha_n = 1$ (i.e., the original C_p). On the other hand, this condition is satisfied for all $c_k \in [0, 1)$ and $c_p \in [0, 1)$ as long as $1 - c_k - c_p > 0$, when $\alpha_n = 1 - (p_n + 1)/(n - k_n)$ (i.e., MC_p). Hence, MC_p is more reasonable for variable selection in high-dimensional data contexts from the perspective of prediction.

4 Asymptotic inefficiency of GC_p

As noted in the previous section, $\alpha_n \to a = 1 - c_p/(1-c_k)$ is a key condition for GC_p to acquire ALE and AME. In this section, we show that this is a necessary condition. Namely, when $\alpha_n \to a \neq 1 - c_p/(1-c_k)$, there is a situation such that

$$\lim_{n\to\infty} Pr\left(\frac{L_n(\hat{M}_n)}{L_n(M_L^*)} > 1\right) = 1, \quad \lim_{n\to\infty} \frac{E(L_n(\hat{M}_n))}{R_n(M_R^*)} > 1$$

even under conditions (C1)–(C6).

For expository purposes, let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, i.e., $k_n = 2$ such that $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}_2$, $\boldsymbol{\Gamma}_* = \sqrt{n}\boldsymbol{x}_2\boldsymbol{\beta}^\top$, where $\boldsymbol{\beta} \in \mathbb{R}^{p_n}$, $\boldsymbol{\Sigma}_* = \boldsymbol{I}_{p_n}$, and $\mathcal{M}_n = \{\{1\}^{p_n}, \{1,2\}^{p_n}\}$. Note that $M = \{1\}^{p_n}$ means $M_1 = \cdots M_{p_n} = \{1\}$ and $M = \{1,2\}^{p_n}$ is similarly defined. For brevity, we write $\{1\}$ and $\{1,2\}$ instead of $\{1\}^{p_n}$ and $\{1,2\}^{p_n}$, respectively. Suppose that $c_p \in (0,1)$ and $\boldsymbol{\beta}$ satisfies $\|\boldsymbol{\beta}\|^2 \to b \in (0, \infty)$, where $\|\cdot\|$ is the Euclidean norm. Then, because $\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{\Gamma}_* \boldsymbol{\Sigma}_*^{-1/2}) = 0$, $R_n(\{1\}) = n\|\boldsymbol{\beta}\|^2 + p_n$, and $R_n(\{1,2\}) = 2p_n$, conditions (C1)–(C6) are satisfied for sufficiently large n. Note that $c_k = 0$ in this situation because k_n is fixed.

From the definition of GC_p,

$$\begin{aligned}
&GC_p(\{1,2\}; \alpha_n) - GC_p(\{1\}; \alpha_n) \\
&= n\alpha_n \mathrm{tr}\{(\hat{\boldsymbol{\Sigma}}(\{1,2\}) - \hat{\boldsymbol{\Sigma}}(\{1\}))\boldsymbol{S}^{-1}\} + 2p_n \\
&= -(n-2)\alpha_n \boldsymbol{x}_2^\top \boldsymbol{Y}\boldsymbol{Y}^\top \boldsymbol{x}_2 \frac{\boldsymbol{x}_2^\top \boldsymbol{Y}\{\boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{x}_1\boldsymbol{x}_1^\top - \boldsymbol{x}_2\boldsymbol{x}_2^\top)\boldsymbol{Y}\}^{-1}\boldsymbol{Y}^\top \boldsymbol{x}_2}{\boldsymbol{x}_2^\top \boldsymbol{Y}\boldsymbol{Y}^\top \boldsymbol{x}_2} + 2p_n.
\end{aligned}$$

It follows from Theorem 3.2.12 in Muirhead (1982) that

$$\left(\frac{\boldsymbol{x}_2^\top \boldsymbol{Y}\{\boldsymbol{Y}^\top(\boldsymbol{I}_n - \boldsymbol{x}_1\boldsymbol{x}_1^\top - \boldsymbol{x}_2\boldsymbol{x}_2^\top)\boldsymbol{Y}\}^{-1}\boldsymbol{Y}^\top \boldsymbol{x}_2}{\boldsymbol{x}_2^\top \boldsymbol{Y}\boldsymbol{Y}^\top \boldsymbol{x}_2}\right)^{-1} \sim \chi^2_{n-p_n-1}.$$

On the other hand, because $\boldsymbol{Y}^\top \boldsymbol{x}_2 = \sqrt{n}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}^\top \boldsymbol{x}_2 \sim N_{p_n}(\sqrt{n}\boldsymbol{\beta}, \boldsymbol{I}_{p_n})$, $\boldsymbol{x}_2^\top \boldsymbol{Y}\boldsymbol{Y}^\top \boldsymbol{x}_2 \sim \chi^2_{p_n}(n\|\boldsymbol{\beta}\|^2)$, which denotes a non-central chi-square distribution with non-centrality parameter $n\|\boldsymbol{\beta}\|^2$. Note that $\chi^2_{n-p_n-1}/n = 1 - c_p + o_p(1)$ and $\chi^2_{p_n}(n\|\boldsymbol{\beta}\|^2)/n = c_p + b + o_p(1)$. Hence, it holds that

$$\frac{GC_p(\{1,2\};\alpha_n) - GC_p(\{1\};\alpha_n)}{n} = -\frac{a(c_p+b)}{1-c_p} + 2c_p + o_p(1). \tag{8}$$

Meanwhile, loss functions of models $\{1\}$ and $\{1,2\}$ are given as

$$L_n(\{1\}) = n\|\boldsymbol{\beta}\|^2 + \boldsymbol{x}_1^\top \boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^\top \boldsymbol{x}_1,$$
$$L_n(\{1,2\}) = \boldsymbol{x}_1^\top \boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^\top \boldsymbol{x}_1 + \boldsymbol{x}_2^\top \boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^\top \boldsymbol{x}_2.$$

Because $\boldsymbol{x}_i^\top \boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{E}}^\top \boldsymbol{x}_i \sim \chi^2_{p_n}$ $(i = 1,2)$, it follows that

$$\frac{L_n(\{1\})}{L_n(\{1,2\})} \xrightarrow{p} \frac{c_p+b}{2c_p} \in (0,\infty), \tag{9}$$
$$\lim_{n\to\infty} \frac{R_n(\{1\})}{R_n(\{1,2\})} = \frac{c_p+b}{2c_p} \in (0,\infty). \tag{10}$$

First, we consider a situation where $a > 0$. Let $b = c_p(1-c_p)/a$. It follows from (8) and (9) that

$$\frac{GC_p(\{1,2\};\alpha_n) - GC_p(\{1\};\alpha_n)}{n} \xrightarrow{p} \frac{c_p(1-c_p-a)}{1-c_p},$$
$$\frac{L_n(\{1\})}{L_n(\{1,2\})} \xrightarrow{p} \frac{a+1-c_p}{2a} = 1 + \frac{1-c_p-a}{2a}.$$

Hence, we have

$$\frac{L_n(\hat{M}_n)}{L_n(M_L^*)} \xrightarrow{p} \begin{cases} (a+1-c_p)/(2a) > 1, & a < 1-c_p, \\ (2a)/(a+1-c_p) > 1, & a > 1-c_p. \end{cases}$$

This implies that GC_p does not exhibit ALE when $0 < a < 1-c_p$ or $a > 1-c_p$.

On the other hand, (10) yields $M_R^* = \{1,2\}$ (resp. $\{1\}$) for sufficiently large n when $a < 1-c_p$ (resp.

$a > 1 - c_p$). Thus, by using $M_R^{**} = \mathcal{M}_n \setminus M_R^*$, we can see that

$$\begin{aligned}
\frac{E(L_n(\hat{M}_n))}{R_n(M_R^*)} &= \frac{E(L_n(M_R^*)I(\hat{M}_n = M_R^*))}{R_n(M_R^*)} + \frac{E(L_n(M_R^{**})I(\hat{M}_n = M_R^{**}))}{R_n(M_R^*)} \\
&= \frac{R_n(M_R^{**})}{R_n(M_R^*)} - \frac{E(\{L_n(M_R^{**}) - L_n(M_R^*)\}I(\hat{M}_n = M_R^*))}{R_n(M_R^*)} \\
&\geq \frac{R_n(M_R^{**})}{R_n(M_R^*)} - \frac{\sqrt{E(\{L_n(\{1\}) - L_n(\{1,2\})\}^2)}}{R_n(M_R^*)}\sqrt{Pr(\hat{M}_n = M_R^*)},
\end{aligned}$$

where $I(\cdot)$ is an indicator function and the last inequality follows from the Cauchy-Schwarz inequality. Note that

$$\begin{aligned}
\frac{\sqrt{E(\{L_n(\{1,2\}) - L_n(\{1\})\}^2)}}{R_n(M_R^*)} &= \sqrt{E((\chi^2_{p_n} - n\|\boldsymbol{\beta}\|^2)^2)}\max\left\{\frac{1}{2p_n}, \frac{1}{p_n + n\|\boldsymbol{\beta}\|^2}\right\} \\
&= \sqrt{2p_n + (p_n - n\|\boldsymbol{\beta}\|^2)^2}\max\left\{\frac{1}{2p_n}, \frac{1}{p_n + n\|\boldsymbol{\beta}\|^2}\right\} \\
&\to |a - (1 - c_p)|\max\left\{\frac{1}{2a}, \frac{1}{a + 1 - c_p}\right\} < \infty.
\end{aligned}$$

Because $\lim_{n\to\infty} Pr(\hat{M}_n = M_R^*) = 0$ and $R_n(M_R^{**})/R_n(M_R^*) > 1$, GC_p does not exhibit AME when $0 < a < 1 - c_p$ or $1 - c_p < a$.

Next, we consider a situation where $a = 0$. Then, (8) implies that $Pr(\hat{M}_n = \{1\}) \to 1$. However, when $b > c_p$, (9) and (10) yield $Pr(M_L^* = \{1,2\}) \to 1$ and $M_R^* = \{1,2\}$ for sufficiently large n, respectively. Hence, in the same manner as the argument when $a > 0$, we can appreciate that GC_p does not exhibit ALE or AME when $a = 0$.

Therefore, $\alpha_n \to a = 1 - c_p/(1 - c_k)$ is a necessary and sufficient condition for ALE and AME of GC_p under conditions (C1)–(C6).

5 Simulation study

This section provides details of a simulation study to compare GC_p among several α_n, where the goodness of criteria is measured by the loss function of the best model selected by each criterion. We prepare three parameters for α_n, that is, $\alpha_n = 1$ (i.e., C_p), $\alpha_n = 1 - (p_n + 1)/(n - k_n)$ (i.e., MC_p) and $\alpha_n = 2/\log n$ (i.e., BIC-type C_p, say BC_p). Because $2/\log n \leq 1 - (p_n + 1)/(n - k_n) \leq 1$ in our settings described below, the number of dimensions of the model selected by C_p (resp. BC_p) is larger (resp. smaller) than or equal to that by MC_p. Generally speaking, this inequality always holds for sufficiently large n.

Hereafter, we explain the simulation settings. Let the first column of $\boldsymbol{X}$ be a vector of ones in $\mathbb{R}^n$ and the other entries be independently generated from a uniform distribution $U(0,1)$. For all $1 \leq i \leq k_n$ and $1 \leq j \leq p_n$, let $(\boldsymbol{B}_*)_{ij} = u_{ij}d_i$, where u_{ij} are independently generated from $U(0, 1/2)$ and

Table 1: Average values of $L_n(\hat{M}_n)/L_n(M_L^*)$ and $L_n(\hat{M}_n)/R_n(M_R^*)$ of C_p, MC_p and BC_p among 1,000 repetitions for each (n, p_n, k_n). Standard deviations are shown in parentheses. Best values for $L_n(\hat{M}_n)/L_n(M_L^*)$ and $L_n(\hat{M}_n)/R_n(M_R^*)$ are emboldened for each (n, p_n, k_n). All values are rounded to 3 decimal places.

			$L_n(\hat{M}_n)/L_n(M_L^*)$			$L_n(\hat{M}_n)/R_n(M_R^*)$		
n	p_n	k_n	C_p	MC_p	BC_p	C_p	MC_p	BC_p
100	20	10	1.262	1.143	**1.115**	1.198	1.085	**1.056**
			(0.185)	(0.108)	(0.069)	(0.193)	(0.116)	(0.056)
200	40	20	1.139	**1.065**	1.169	1.125	**1.052**	1.153
			(0.079)	(0.048)	(0.046)	(0.089)	(0.059)	(0.016)
400	80	40	1.129	**1.027**	1.191	1.125	**1.023**	1.187
			(0.057)	(0.020)	(0.025)	(0.060)	(0.028)	(0.006)
800	160	80	1.117	**1.010**	1.182	1.114	**1.007**	1.178
			(0.033)	(0.007)	(0.012)	(0.035)	(0.012)	(0.002)
100	10	10	1.290	1.229	**1.153**	1.219	1.160	**1.085**
			(0.259)	(0.220)	(0.094)	(0.272)	(0.225)	(0.091)
200	10	20	1.167	**1.163**	1.191	1.110	**1.106**	1.127
			(0.116)	(0.110)	(0.088)	(0.131)	(0.119)	(0.033)
400	10	40	1.107	**1.107**	1.174	1.060	**1.060**	1.121
			(0.063)	(0.061)	(0.069)	(0.074)	(0.070)	(0.017)
800	10	80	1.065	**1.064**	1.233	1.049	**1.048**	1.213
			(0.045)	(0.043)	(0.050)	(0.057)	(0.054)	(0.009)

$d_i = 5\sqrt{k_n - i + 1}/k_n$. For comparative purposes, we examine a situation where $\boldsymbol{\Gamma}_* = \boldsymbol{X}\boldsymbol{B}_*$, which implies that the full model is the true model. Suppose that $\boldsymbol{\Sigma}_* = (0.7^{|i-j|})_{ij}$ for $1 \le i, j \le p_n$. We also suppose that there are two subsets $M^{(1)}, M^{(2)} \subset \{1, \ldots, p_n\}$ such that $M_1 = \cdots = M_{p_n/2} = M^{(1)}$ and $M_{p_n/2+1} = \cdots = M_{p_n} = M^{(2)}$, which implies that there are two groups of response variables. To reduce computational burden, we adopt a nested model set, i.e., we select $M^{(1)}$ and $M^{(2)}$ among $\{\{1\}, \ldots, \{1, \ldots, k_n\}\}$. It should be noted that the true (full) model is not always the best model from the perspective of prediction in our simulation study, because some coefficients are very small, so variable selection makes sense in this situation. This supposition is confirmed below.

We prepared two cases for p_n as high- and fixed-dimensional cases, where $p_n = n/5$ for the high-dimensional case, whereas $p_n = 10$ for the fixed case. The sample size n varies from 100 to 800, and we set $k_n = n/10$. Then, we generate $\boldsymbol{Y}$ and select the best subset of explanatory variables by each C_p-type criterion. After variable selection, we calculate the loss functions for each best model.

Table 1 provides average values of $L_n(\hat{M}_n)/L_n(M_L^*)$ and $L_n(\hat{M}_n)/R_n(M_R^*)$ of C_p, MC_p and BC_p based on 1,000 repetitions for each (n, p_n, k_n). Note that $L_n(\hat{M}_n)/L_n(M_L^*)$ and $L_n(\hat{M}_n)/R_n(M_R^*)$ are criteria for ALE and AME, respectively, and smaller is better. From this table, we can confirm that MC_p exhibits good performance regardless of p_n, and C_p works well when $p_n = 10$ but it does not work well when p_n is large. On the other hand, BC_p has higher values of $L_n(\hat{M}_n)/L_n(M_L^*)$ and $L_n(\hat{M}_n)/R_n(M_R^*)$

Table 2: Average dimensions of selected models by C_p, MC_p, and BC_p and loss minimizing models among 1,000 repetitions for each (n, p_n, k_n). Standard deviations are shown in parentheses. All values are rounded to 3 decimal places.

n	p_n	k_n	C_p	MC_p	BC_p	Loss
100	20	10	5.754 (1.848)	3.154 (1.507)	1.127 (0.314)	3.277 (1.145)
200	40	20	13.015 (2.066)	7.545 (2.161)	1.010 (0.083)	7.590 (1.222)
400	80	40	24.146 (2.803)	13.617 (2.185)	1.000 (0.000)	13.505 (1.171)
800	160	80	50.018 (3.448)	27.035 (2.811)	1.000 (0.000)	27.188 (1.930)
100	10	10	3.756 (1.959)	2.857 (1.562)	1.107 (0.289)	2.804 (0.900)
200	10	20	8.650 (3.499)	7.396 (3.444)	1.011 (0.097)	7.849 (2.430)
400	10	40	17.203 (6.020)	15.505 (6.064)	1.005 (0.071)	16.927 (5.135)
800	10	80	26.427 (8.229)	25.322 (8.077)	1.010 (0.093)	25.910 (5.655)

except when the sample size is small. These results concur with our theoretical exposition regarding efficiency and inefficiency.

Table 2 shows the average dimensions of models, i.e., $\#(M^{(1)})/2 + \#(M^{(2)})/2$ selected by each GC_p and loss minimizing models. This indicates that the number of dimensions of loss minimizing models varies depending on the sample size, and the full model is not (always) the best model in spite of the fact that the full model is true. Based on our simulation settings, BC_p tends to select much smaller models in comparison with models that have the smallest loss function while C_p often selects larger models when p_n is large. The average number of dimensions of models selected by MC_p is close to that of the loss minimizing models in both high- and fixed-dimensional situations. This implies that α_n substantially affects the dimensions of selected models as well as efficiency.

Hence, these results indicate that MC_p is a useful variable selection method regardless of p_n, and thus we recommend its use from the perspective of robust prediction.

6 Conclusions

We have derived sufficient conditions for ALE and AME of GC_p in high-dimensional multivariate linear regression models. It is shown that MC_p exhibits ALE and AME in high-dimensional data, while the original C_p, known as an asymptotically efficient criterion in univariate cases, does not exhibit ALE or AME under the same conditions. This is because a non-trivial bias term is omitted in the original C_p as an estimator of the risk function; this term plays an important role for adaptation to high-dimensional frameworks. Indeed, if the tuning parameter of GC_p, α_n, converges to $a \neq 1 - c_p/(1 - c_k)$ like in the case of C_p and BC_p, we showed that GC_p is asymptotically inefficient. Through a simulation study, the finite sample performances of C_p-type criteria are compared, and MC_p is better than C_p and BC_p in high-dimensional data.

Note that when p_n is large, MC_p works well even under the parametric scenario, where the true model is included in a set of candidate models. Unlike a univariate case, the risk of the true model always goes to infinity with $p_n \to \infty$. Thus, under the parametric scenario, it is possible that conditions (C1)–(C6) are satisfied, and then, the asymptotic efficiencies of MC_p hold. Moreover, assuming response variables to have a common model, i.e., $M_1 = \cdots = M_{p_n}$, MC_p has the consistency property as well under moderate conditions (Fujikoshi, Sakurai and Yanagihara 2014). Hence, MC_p can be regarded as a feasible method for variable selection from the perspective of both prediction and interpretability when p_n is large. This attractive property is only seen in high-dimensional situations, i.e., $p_n \to \infty$.

When p_n is greater than n, we cannot directly calculate $\boldsymbol{S}^{-1}$ and thus GC_p. Therefore, we need different approaches to estimate a covariance matrix $\boldsymbol{\Sigma}$ such as sparse or ridge estimation (e.g., Yamamura, Yanagihara and Srivastava 2010, Katayama and Imori 2014, Fujikoshi and Sakurai 2016). If we can estimate $\boldsymbol{\Sigma}$ accurately via these procedures, ALE and AME can be established by using it in place of $\boldsymbol{S}$. It should also be noted that our proof depends on the assumption that the response matrix follows a Gaussian distribution. Because we use some properties of the Gaussian distribution, this is not a trivial limitation from the perspective of generalizing the results. Another extension of this paper is to relax condition (C4) (see, Yang 1999). In Section 3, we gave a sufficient condition for (C4), that is, $\log k_n = o(p_n)$ assuming some group structure of response variables. Under this condition, even when the number of candidate models are exponentially large, i.e., $\#(\mathcal{M}_n) = 2^{k_n}$, (C4) holds. Although this condition is not restricted, when considering a situation where each response variable uses different models, it is still important to mitigate (C4). Yang (1999) proposed a criterion by using an additional penalty term, which can be used for model selection without the constraint on the number of candidate models. It may be possible to apply this idea to our setting. How best to navigate these issues represent fruitful terrain for future research.

Acknowledgements

This study is supported in part by JSPS KAKENHI Grant Number JP17K12650 and JP20K19757, and "Funds for the Development of Human Resources in Science and Technology" under MEXT, through the "Home for Innovative Researchers and Academic Knowledge Users (HIRAKU)" consortium.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, pp. 716–723.

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

Bai, Z., Choi, K. P. and Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics 46*, pp. 1050–1076.

Fujikoshi, Y. and Sakurai, T. (2016). High-dimensional consistency of rank estimation criteria in multivariate linear model. *Journal of Multivariate Analysis 149*, pp. 199–212.

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

Fujikoshi, Y., Sakurai, T. and Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *Journal of Multivariate Analysis 123*, pp. 184–200.

Horn, R. A. and Jornson, C. R. (1994). *Topics in matrix analysis.* Cambridge university press.

Householder, A. S. and Carpenter, J. A. (1963). The singular values of involutory and of idempotent matrices. *Numerische Mathematik 5*, pp. 234–237.

Imori, S. and von Rosen, D. (2015). Covariance components selection in high-dimensional growth curve model with random coefficients. *Journal of Multivariate Analysis 136*, pp. 86–94.

Katayama, S. and Imori, S. (2014). Lasso penalized model selection criteria for high-dimensional multivariate linear regression analysis. *Journal of Multivariate Analysis 132*, pp. 138–150.

Kollo, T. and von Rosen, D. (2005). *Advanced multivariate statistics with matrices.* Springer.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics 15*, pp. 958–975.

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

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* John Wiley & Sons.

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

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics 12*, pp. 758–765.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, pp. 461–464.

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

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

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics 35*, pp. 415–423.

Yamamura, M., Yanagihara, H. and Srivastava, M. S. (2010). Variable selection in multivariate linear regression models with fewer observations than the dimension. *Japanese Journal of Applied Statistics 39*, pp. 1–19.

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

Yanagihara, H. (2020). High-dimensionality-adjusted asymptotically loss and mean efficient GC_p criterion for normal multivariate linear regression models. *TR 20-03, Statistical Research Group*, Hiroshima University, Hiroshima.

Yanagihara, H., Wakaki, H. and Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electronic Journal of Statistics 9*, pp. 869–897.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica 9*, pp. 475–499.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge: Cambridge Series in Statistical and Probabilistic Mathematics.

A Supplementary materials

In this supplementary materials, we show ALE and AME of GC_p. Section A.1 provides four lemmas that are used for showing ALE and AME. Section A.2 gives a proof of Theorem 1, whereas a proof of Theorem 2 is obtained in Section A.3. An outline of the proofs of Theorems is based on Li (1987) and Shibata (1983) although some techniques are used to overcome the difficulties imposed by high-dimensionality.

A.1 Preliminaries

We introduce four lemmas that are used for showing ALE and AME.

Lemma A.1. Let X be a random variable distributed as $N(0,1)$. Then, for all $t > 0$,

$$Pr(X \geq t) \leq \exp\left(-\frac{t^2}{2}\right).$$

Lemma A.2. Let $\boldsymbol{X} \sim N_{n,p}(\boldsymbol{O}, \boldsymbol{I}_n, \boldsymbol{I}_p)$ and $n > p$. It holds that for all $t > 0$,

$$n^{1/2} - p^{1/2} - t \leq \sigma_p(\boldsymbol{X}) \leq \sigma_1(\boldsymbol{X}) \leq n^{1/2} + p^{1/2} + t$$

with probability at least $1 - 2\exp(-Ct^2)$, where C is a positive constant that does not depend on n and p.

Because Lemmas A.1 and A.2 can be found elsewhere (see e.g., Wainwright 2019: Example 2.1, Example 6.2), we omit their proofs for brevity.

Lemma A.3. Let $\boldsymbol{Z} \sim N_n(\boldsymbol{0}_n, \boldsymbol{I}_n)$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. If there exists constants $A_1, A_2 > 0$ such that $A_1 \geq \mathrm{tr}(\boldsymbol{A}^2)$ and $A_2 \geq \sigma_1(\boldsymbol{A})$, then for all $t \geq 0$,

$$Pr(|\boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z} - \mathrm{tr}(\boldsymbol{A})| \geq t) \leq 2\exp\left(-\frac{t}{8}\min\left\{\frac{t}{A_1}, \frac{1}{A_2}\right\}\right).$$

Proof. When $\boldsymbol{A} = \boldsymbol{O}_{n,n}$, the assertion is trivial. Thus, we assume $\boldsymbol{A} \neq \boldsymbol{O}_{n,n}$, which implies that $\mathrm{tr}(\boldsymbol{A}^2) > 0$ and $\sigma_1(\boldsymbol{A}) > 0$.

For a proof, we refer to Example 2.8 and Proposition 2.9 in Wainwright (2019). Let $X = \boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z}$. Note that $E(X) = \mathrm{tr}(\boldsymbol{A})$. At first, we attempt to show X is sub-exponentional, i.e., for all $|\theta| < 1/(4A_2)$,

$$E(\exp\{\theta(X - E(X))\}) \leq \exp(2A_1\theta^2).$$

Because $\boldsymbol{A}$ is symmetric, there exists an orthogonal matrix $\boldsymbol{Q}$ such that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$, where $\boldsymbol{D} = \mathrm{diag}\{\lambda_1(\boldsymbol{A}), \ldots, \lambda_n(\boldsymbol{A})\}$ and $\lambda_i(\cdot)$ denotes the ith eigenvalue. Let $\boldsymbol{Y} = \boldsymbol{Q}^\top\boldsymbol{Z} = (Y_1, \ldots, Y_n)^\top$ that follows $N_n(\mathbf{0}_n, \boldsymbol{I}_n)$. Then,

$$X = \boldsymbol{Y}^\top \boldsymbol{D} \boldsymbol{Y} = \sum_{i=1}^n \lambda_i(\boldsymbol{A}) Y_i^2.$$

For all $|\theta| < 1/(2\sigma_1(\boldsymbol{A}))$, $E(\exp\{\theta\lambda_i(\boldsymbol{A})Y_i^2\})$ exists because $2\theta\lambda_i(\boldsymbol{A}) \le 2|\theta|\sigma_1(\boldsymbol{A}) < 1$. Then, independence of $Y_1, \ldots, Y_n$ indicates that

$$\begin{aligned} E(\exp\{\theta(X - E(X))\}) &= E\left(\exp\left\{\sum_{i=1}^n \theta\lambda_i(\boldsymbol{A})(Y_i^2 - 1)\right\}\right) \\ &= \prod_{i=1}^n E(\exp\{\theta\lambda_i(\boldsymbol{A})(Y_i^2 - 1)\}) \\ &= \prod_{i=1}^n \exp\{-\theta\lambda_i(\boldsymbol{A})\}\{1 - 2\theta\lambda_i(\boldsymbol{A})\}^{-1/2} \\ &= \exp\left\{\sum_{i=1}^n \left\{-\theta\lambda_i(\boldsymbol{A}) - \frac{1}{2}\log(1 - 2\theta\lambda_i(\boldsymbol{A}))\right\}\right\}. \end{aligned}$$

Note that $2x^2 \ge -x - \log(1-2x)/2$ for all $|x| \le 1/4$. Hence, for all $|\theta| < 1/(4A_2)$, which implies that $\theta < 1/(2\sigma_1(\boldsymbol{A}))$ and $|\theta\lambda_i(\boldsymbol{A})| < 1/4$, we have

$$\begin{aligned} E(\exp\{\theta(X - E(X))\}) &\le \exp\left\{2\sum_{i=1}^n \theta^2\lambda_i(\boldsymbol{A})^2\right\} \\ &= \exp\{2\theta^2 \mathrm{tr}(\boldsymbol{A}^2)\} \\ &\le \exp(2A_1\theta^2). \end{aligned}$$

From Proposition 2.9 in Wainwright (2019), it follows that for all $t \ge 0$,

$$Pr(|\boldsymbol{Z}^\top \boldsymbol{A}\boldsymbol{Z} - \mathrm{tr}(\boldsymbol{A})| \ge t) \le \begin{cases} 2\exp\{-t^2/(8A_1)\} & 0 \le t \le A_1/A_2 \\ 2\exp\{-t/(8A_2)\} & t > A_1/A_2 \end{cases}$$

The right-hand side of the above inequality is bounded by $2\exp\{-t\min\{t/A_1, 1/A_2\}/8\}$. Hence, the proof is completed. □

Lemma A.3 yields the following lemma for chi-square distribution.

Lemma A.4. Let X_n follow chi-square distribution with n degrees of freedom and a_n be a sequence such that $a_n/n \to 1$. Then, for all $t > 0$ and for sufficiently large n,

$$Pr\left(\left|\left(\frac{X_n}{a_n}\right)^{-1} - 1\right| > t\right) \le 4\exp\left(-\frac{n\min\{4t^2, 1\}}{128}\right).$$

Proof. Because $a_n/n \to 1$, for sufficiently large n, $|a_n/n - 1| \le t/4$. Thus, we see that

$$
\begin{aligned}
Pr\left(\left|\left(\frac{X_n}{a_n}\right)^{-1} - 1\right| > t\right) &= Pr\left(\frac{|X_n - a_n|}{X_n} > t\right) \\
&\le Pr\left(\left\{\frac{|X_n - n| + |a_n - n|}{n - |X_n - n|} > t\right\} \cap \left\{|X_n - n| \le \frac{n}{4}\right\}\right) \\
&\quad + Pr\left(|X_n - n| > \frac{n}{4}\right) \\
&\le Pr\left(|X_n - n| > \frac{nt}{2}\right) + Pr\left(|X_n - n| > \frac{n}{4}\right) \\
&\le 2Pr\left(|X_n - n| > \frac{n\min\{2t, 1\}}{4}\right).
\end{aligned}
$$

Applying Lemma A.3 with $\boldsymbol{A} = \boldsymbol{I}_n$, $A_1 = n$ and $A_2 = 1$, we have

$$
Pr\left(|X_n - n| > \frac{n\min\{2t, 1\}}{4}\right) \le 2\exp\left(-\frac{n\min\{4t^2, 1\}}{128}\right).
$$

Thus, the proof is completed. □

A.2 Proof of Theorem 1

From the definition of $L_n(M)$ and $\mathrm{GC}_p(M; \alpha_n)$, these difference can be separated as

$$
\begin{aligned}
GC_p(M; \alpha_n) - L_n(M) &= \alpha_n \mathrm{tr}(\boldsymbol{Y}^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{Y} \boldsymbol{S}^{-1}) + \mathrm{tr}(\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F} \boldsymbol{\mathcal{E}}) - \mathrm{tr}(\boldsymbol{\Gamma}_*^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{\Gamma}_* \boldsymbol{\Sigma}_*^{-1}) \\
&\quad + \mathrm{tr}\{\boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F} \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}_*^{1/2} (\alpha_n \boldsymbol{S}^{-1} - \boldsymbol{\Sigma}_*^{-1})\} \\
&\quad + 2\mathrm{tr}\{(\boldsymbol{P}_{M_F} \boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}_*^{-1/2}\} \\
&\quad - 2\left\{\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F} \boldsymbol{\mathcal{E}}(M) \boldsymbol{\Sigma}_*^{-1/2}\} - \sum_{j=1}^{p_n} k_{M_j}\right\} \\
&\quad + \mathrm{tr}\{(\boldsymbol{P}_{M_F} \boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top (\boldsymbol{P}_{M_F} \boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))(\alpha_n \boldsymbol{S}^{-1} - \boldsymbol{\Sigma}_*^{-1})\} \\
&\quad + 2\mathrm{tr}\{(\boldsymbol{P}_{M_F} \boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top (\boldsymbol{P}_{M_F} \boldsymbol{\mathcal{E}} \boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\mathcal{E}}(M))(\alpha_n \boldsymbol{S}^{-1} - \boldsymbol{\Sigma}_*^{-1})\} \\
&\quad - 2\mathrm{tr}\{\boldsymbol{\Sigma}_*^{1/2} \boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F} \boldsymbol{\mathcal{E}}(M)(\alpha_n \boldsymbol{S}^{-1} - \boldsymbol{\Sigma}_*^{-1})\} \\
&\quad + \mathrm{tr}\{\boldsymbol{\mathcal{E}}(M)^\top \boldsymbol{\mathcal{E}}(M)(\alpha_n \boldsymbol{S}^{-1} - \boldsymbol{\Sigma}_*^{-1})\}.
\end{aligned}
$$

Let $\boldsymbol{T}_1(M) = (\boldsymbol{P}_{M_F}\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))\boldsymbol{\Sigma}_*^{-1/2}$ and $\boldsymbol{T}_2 = \alpha_n\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{I}_{p_n}$. Then, we only consider the terms that depend on M defined as follows:

$$
\begin{aligned}
B_n(M) = 2\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{\mathcal{E}}\} - 2&\left\{\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} - \sum_{j=1}^{p_n} k_{M_j}\right\} \\
&+ \mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)\boldsymbol{T}_2\} + 2\mathrm{tr}\{\boldsymbol{T}_1(M)^\top(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2})\boldsymbol{T}_2\} \\
&- 2\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\} + \mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\}.
\end{aligned}
$$

We can discern from the definition of $\hat{M}_n$ that

$$
L_n(\hat{M}_n) + GC_p(\hat{M}_n;\alpha_n) - L_n(\hat{M}_n) \le L_n(M_L^*) + GC_p(M_L^*;\alpha_n) - L_n(M_L^*),
$$

and thus

$$
L_n(\hat{M}_n)\left\{1 + \frac{B_n(\hat{M}_n)}{L_n(\hat{M}_n)}\right\} \le L_n(M_L^*)\left\{1 + \frac{B_n(M_L^*)}{L_n(M_L^*)}\right\}.
$$

This implies that GC_p exhibits ALE if we can show that $\max_{M\in\mathcal{M}_n}|B_n(M)|/L_n(M)$ converges to zero in probability.

Let $\boldsymbol{G} \in \mathbb{R}^{n\times k_n}$ and $\boldsymbol{H} \in \mathbb{R}^{n\times(n-k_n)}$ be matrices such that $\boldsymbol{P}_{M_F} = \boldsymbol{G}\boldsymbol{G}^\top$, $\boldsymbol{P}_{M_F}^\perp = \boldsymbol{H}\boldsymbol{H}^\top$ and $(\boldsymbol{G}, \boldsymbol{H})$ is orthogonal. It is worth mentioning that $\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}$ and $\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}$ are independent and normally distributed. Here, we define the following event:

$$
E_{n,\gamma} : (n-k_n)^{1/2} - p_n^{1/2} - n^{\gamma/2} \le \sigma_{p_n}(\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}) \le \sigma_1(\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}) \le (n-k_n)^{1/2} + p_n^{1/2} + n^{\gamma/2},
$$

where $\gamma \in (0,1)$ is a constant. Note that it follows from Lemma A.2 that there exists a positive constant $C > 0$ such that

$$
Pr(E_{n,\gamma}^c) \le 2\exp(-Cn^\gamma), \tag{11}
$$

where $E_{n,\gamma}^c$ means a complement set of $E_{n,\gamma}$. This implies $E_{n,\gamma}$ occurs with high probability. On $E_{n,\gamma}$, we have the following lemma:

Lemma A.5. Let $\gamma \in (0,1)$ be a constant. Suppose that conditions (C1) and (C2) hold. Then, on $E_{n,\gamma}$

$$
\sigma_1(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - (n-k_n)(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})^{-1}) = o(1), \quad \sigma_1(\boldsymbol{T}_2) = O(1).
$$

Proof. From basic linear algebra, it follows that

$$
\begin{aligned}
&(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2})^{-1}-(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})^{-1}\\
&=-(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2})^{-1}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}-\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})^{-1}.
\end{aligned}
$$

This equation implies that

$$
\begin{aligned}
\sigma_1((\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2})^{-1}-(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})^{-1}) \le\ & \frac{\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2})}{\sigma_{p_n}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2})\sigma_{p_n}(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})}\\
&+\frac{2\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})}{\sigma_{p_n}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2})\sigma_{p_n}(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})}.
\end{aligned}
\tag{12}
$$

Note that $\{(n-k_n)^{1/2}\pm p_n^{1/2}\pm n^{\gamma/2}\}/n^{1/2}$ converges to $(1-c_k)^{1/2}\pm c_p^{1/2}\in(0,\infty)$ under (C1). Thus, there exists a positive constant $C_1\ge 1$ such that for sufficiently large n, on $E_{n,\gamma}$,

$$
\frac{1}{C_1}\le\frac{\sigma_{p_n}(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})}{n}\le\frac{\sigma_1(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})}{n}\le C_1. \tag{13}
$$

On the other hand, due to (C2) and (13), we have

$$
\begin{aligned}
\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}}) &\le \sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2})^{1/2}\sigma_1(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})^{1/2}\\
&\le C_1^{1/2}n^{1/2}\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2})^{1/2}\\
&=o(n).
\end{aligned}
\tag{14}
$$

Furthermore, on the event $E_{n,\gamma}$, it follows from (C2), (13) and (14) that

$$
\begin{aligned}
\sigma_{p_n}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}) &\ge \sigma_{p_n}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\Gamma}_*\boldsymbol{\Sigma}_*^{-1/2})\\
&\quad+\sigma_{p_n}(\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})-2\sigma_1(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\Gamma}_*^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}})\\
&\ge\frac{n}{C_1}+o(n).
\end{aligned}
\tag{15}
$$

Hence, by substituting (13), (14) and (15) into (12), the first assertion is obtained.

Next, we show the second assertion. It follows from (15) that there exists a positive constant $C_2>0$

such that for sufficiently large n,

$$
\begin{aligned}
\sigma_1(\boldsymbol{T}_2) &\le 1 + \alpha_n \sigma_1(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2}) \\
&= 1 + \alpha_n (n-k_n)\sigma_{p_n}(\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2})^{-1} \\
&\le 1 + \frac{C_2\alpha_n(n-k_n)}{n}.
\end{aligned}
$$

Note that $\alpha_n \to a \in [0,\infty)$. Hence, the proof is completed. □

On $E_{n,\gamma}$, we can derive a convergence rate of $|B_n(M)|/R_n(M)$ as follows:

Lemma A.6. Let $\gamma \in (0,1)$ be a constant. Suppose that conditions (C1)–(C3) hold. If $\alpha_n \to a = 1 - c_p/(1-c_k)$ as $n \to \infty$, then there exist positive constants $C_j > 0$ ($j = 1,2,3$) and positive non-decreasing functions g_1, g_2 such that for all $\varepsilon > 0$, for sufficiently large n, for all $M \in \mathcal{M}_n$,

$$
Pr\left(\left\{\frac{|B_n(M)|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right) \le C_1 \exp\{-C_2 g_1(\varepsilon) n\} + C_1 \exp\{-C_3 g_2(\varepsilon)\varepsilon R_n(M)\}.
$$

The following lemma enables us to consider $|B_n(M)|/R_n(M)$ instead of $|B_n(M)|/L_n(M)$.

Lemma A.7. Suppose that condition (C3) holds. Then, for all $\varepsilon > 0$ and $M \in \mathcal{M}_n$,

$$
Pr\left(\left|\frac{L_n(M)}{R_n(M)} - 1\right| > \varepsilon\right) \le 4\exp\left\{-\frac{\min\{\varepsilon, 2\}\varepsilon R_n(M)}{32C_A^2}\right\}.
$$

Proof. Let

$$
\begin{aligned}
\xi(M) &= \frac{L_n(M) - R_n(M)}{R_n(M)} \\
&= \frac{-2\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\}}{R_n(M)} \\
&\quad + \frac{\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} - \mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\}}{R_n(M)}.
\end{aligned}
$$

Note that $\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M) = (\boldsymbol{P}_{M_1}^\perp \boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{P}_{M_{p_n}}^\perp \boldsymbol{\gamma}_{p_n}^*)$ and $\boldsymbol{\mathcal{E}}(M) = (\boldsymbol{P}_{M_1}\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{P}_{M_{p_n}}\boldsymbol{\varepsilon}_{p_n})$. This implies that

$$
\begin{aligned}
\mathrm{vec}((\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))\boldsymbol{\Sigma}_*^{-1/2}) &= (\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np_n} - \boldsymbol{P}(M))\mathrm{vec}(\boldsymbol{\Gamma}_*), \\
\mathrm{vec}(\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}) &= \boldsymbol{A}(M)\mathrm{vec}(\boldsymbol{\mathcal{E}}),
\end{aligned}
$$

where $\boldsymbol{P}(M) = \mathrm{diag}\{\boldsymbol{P}_{M_1}, \dots, \boldsymbol{P}_{M_{p_n}}\}$ and $\boldsymbol{A}(M) = (\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{I}_n)\boldsymbol{P}(M)(\boldsymbol{\Sigma}_*^{1/2} \otimes \boldsymbol{I}_n)$. Thus, it follows that

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} &= \mathrm{vec}((\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))\boldsymbol{\Sigma}_*^{-1/2})^\top \mathrm{vec}(\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}) \\ &= \mathrm{vec}(\boldsymbol{\Gamma}_*)^\top (\boldsymbol{I}_{np_n} - \boldsymbol{P}(M))(\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{I}_n)\boldsymbol{A}(M)\mathrm{vec}(\boldsymbol{\mathcal{E}}).\end{aligned}$$

Hence, $\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\}$ follows Gaussian distribution with mean 0 and variance $v(M)^2$, where

$$\begin{aligned}v(M)^2 &= \|\boldsymbol{A}(M)^\top (\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np_n} - \boldsymbol{P}(M))\mathrm{vec}(\boldsymbol{\Gamma}_*)\|_2^2 \\ &\le \sigma_1(\boldsymbol{A}(M))^2 \|(\boldsymbol{\Sigma}_*^{-1/2} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np_n} - \boldsymbol{P}(M))\mathrm{vec}(\boldsymbol{\Gamma}_*)\|_2^2 \\ &= \sigma_1(\boldsymbol{A}(M))^2 \mathrm{tr}\{\boldsymbol{\Delta}(M)\} \\ &\le C_A^2 R_n(M),\end{aligned}$$

where the last inequality follows from (C3) and $\mathrm{tr}\{\boldsymbol{\Delta}(M)\} \le R_n(M)$. Let Z be a random variable that follows $N(0,1)$. Then, from Lemma A.1, we can see that

$$\begin{aligned}Pr\left(\frac{|2\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}(\boldsymbol{\Gamma}_* - \boldsymbol{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\}|}{R_n(M)} > \varepsilon\right) &\le Pr\left(|Z| > \frac{\varepsilon R_n(M)^{1/2}}{2C_A}\right) \\ &\le 2\exp\left\{-\frac{\varepsilon^2 R_n(M)}{8C_A^2}\right\}. \qquad (16)\end{aligned}$$

On the other hand, it follows that

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} &= \mathrm{vec}(\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2})^\top \mathrm{vec}(\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}) \\ &= \mathrm{vec}(\boldsymbol{\mathcal{E}})^\top \boldsymbol{A}(M)^\top \boldsymbol{A}(M)\mathrm{vec}(\boldsymbol{\mathcal{E}}).\end{aligned}$$

Under (C3), $\sigma_1(\boldsymbol{A}(M)^\top \boldsymbol{A}(M)) \le C_A^2$ and $\mathrm{tr}\{(\boldsymbol{A}(M)^\top \boldsymbol{A}(M))^2\} \le C_A^2 \mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\} \le C_A^2 R_n(M)$. Thus, from Lemma A.3, we have

$$Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} - \mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\}|}{R_n(M)} > \varepsilon\right) \le 2\exp\left\{-\frac{\min\{\varepsilon, 1\}\varepsilon R_n(M)}{8C_A^2}\right\}. \quad (17)$$

By combining (16) and (17), it holds that

$$
\begin{aligned}
Pr(|\xi(M)| > \varepsilon) \leq Pr&\left(\frac{|2\mathrm{tr}\{\mathbf{\Sigma}_*^{-1/2}(\mathbf{\Gamma}_* - \mathbf{\Gamma}_*(M))^\top \boldsymbol{\mathcal{E}}(M)\mathbf{\Sigma}_*^{-1/2}\}|}{R_n(M)} > \frac{\varepsilon}{2}\right) \\
&+ Pr\left(\frac{|\mathrm{tr}\{\mathbf{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top \boldsymbol{\mathcal{E}}(M)\mathbf{\Sigma}_*^{-1/2}\} - \mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\}|}{R_n(M)} > \frac{\varepsilon}{2}\right) \\
&\leq 4\exp\left\{-\frac{\min\{\varepsilon, 2\}\varepsilon R_n(M)}{32C_A^2}\right\}.
\end{aligned}
$$

Thus, the proof is completed. □

By combining Lemmas A.6 and A.7 with (11), for all $\varepsilon > 0$, there exist positive constants $C_j > 0$ $(j = 1, \ldots, 4)$ such that for sufficiently large n,

$$
\begin{aligned}
&Pr\left(\max_{M\in\mathcal{M}_n}\frac{|B_n(M)|}{L_n(M)} > \varepsilon\right) \\
&\leq Pr\left(\left\{\max_{M\in\mathcal{M}_n}\frac{|B_n(M)|}{L_n(M)} > \varepsilon\right\} \cap \left\{\max_{M\in\mathcal{M}_n}\left|\frac{L_n(M)}{R_n(M)} - 1\right| \leq \frac{1}{2}\right\} \cap E_{n,\gamma}\right) \\
&\quad + Pr\left(\max_{M\in\mathcal{M}_n}\left|\frac{L_n(M)}{R_n(M)} - 1\right| > \frac{1}{2}\right) + Pr(E_{n,\gamma}^c) \\
&= Pr\left(\left\{\max_{M\in\mathcal{M}_n}\frac{|B_n(M)|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right) + Pr\left(\max_{M\in\mathcal{M}_n}\left|\frac{L_n(M)}{R_n(M)} - 1\right| > \frac{1}{2}\right) \\
&\quad + Pr(E_{n,\gamma}^c) \\
&\leq C_1\#(\mathcal{M}_n)\exp(-C_2 n) + C_1\sum_{M\in\mathcal{M}_n}\exp\{-C_3 R_n(M)\} + C_1\exp(-C_4 n^\gamma),
\end{aligned}
$$

which goes to zero under condition (C4) and (C5). Because $L_n(M_L^*) \leq L_n(M)$ for all $M \in \mathcal{M}_n$, this yields Theorem 1. Hence, hereafter, we attempt to show Lemma A.6. At first, we evaluate the first term of $B_n(M)$.

Lemma A.8. For all $\varepsilon > 0$ and $M \in \mathcal{M}_n$,

$$
Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{\mathcal{E}}\}|}{R_n(M)} > \varepsilon\right) \leq 2\exp\left\{-\frac{\varepsilon^2 R_n(M)}{2}\right\}.
$$

Proof. It is seen that $\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{\mathcal{E}}\}$ follows $N(0, \tau(M)^2)$, where $\tau(M)^2 = \mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)\} \leq \mathrm{tr}\{\mathbf{\Delta}(M)\} \leq R_n(M)$. Lemma A.1 yields that

$$
\begin{aligned}
Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{\mathcal{E}}\}|}{R_n(M)} > \varepsilon\right) &\leq Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{\mathcal{E}}\}|}{\tau(M)} > \varepsilon R_n(M)^{1/2}\right) \\
&\leq 2\exp\left\{-\frac{\varepsilon^2 R_n(M)}{2}\right\}.
\end{aligned}
$$

The proof is completed. □

The next lemma provides an evaluation of the second term of $B_n(M)$.

Lemma A.9. Suppose that condition (C3) holds. For all $\varepsilon > 0$ and for all $M \in \mathcal{M}_n$

$$Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} - \sum_{j=1}^{p_n} k_{M_j}|}{R_n(M)} > \varepsilon\right) \leq 2\exp\left\{-\frac{\varepsilon R_n(M)}{8C_A}\min\{\varepsilon C_A, 1\}\right\}.$$

Proof. It is easy to see that

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} &= \mathrm{vec}(\boldsymbol{\mathcal{E}})^\top \boldsymbol{A}(M)\mathrm{vec}(\boldsymbol{\mathcal{E}}) \\ &= \frac{\mathrm{vec}(\boldsymbol{\mathcal{E}})^\top(\boldsymbol{A}(M)+\boldsymbol{A}(M)^\top)\mathrm{vec}(\boldsymbol{\mathcal{E}})}{2}.\end{aligned}$$

Note that $\mathrm{tr}\{\boldsymbol{A}(M)\} = \mathrm{tr}\{\boldsymbol{A}(M)+\boldsymbol{A}(M)^\top\}/2 = \sum_{j=1}^{p_n} k_{M_j}$. In order to apply Lemma A.3, we check the conditions. From (C3), $\sigma_1(\boldsymbol{A}(M)+\boldsymbol{A}(M)^\top)/2 \leq \sigma_1(\boldsymbol{A}(M)) \leq C_A$. Recall that

$$\mathrm{tr}\{\boldsymbol{A}(M)\} \leq \sum_{j=1}^{np_n}\sigma_j(\boldsymbol{A}(M)) \leq \mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\}. \tag{18}$$

Hence, considering $\mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\} \leq R_n(M)$, Lemma A.3 implies that

$$Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} - \sum_{j=1}^{p_n} k_{M_j}|}{R_n(M)} > \varepsilon\right) \leq 2\exp\left\{-\frac{\varepsilon R_n(M)}{8C_A}\min\{\varepsilon C_A, 1\}\right\}.$$

□

Next, the third term of $B_n(M)$ is evaluated as follows:

Lemma A.10. Let $\gamma \in (0,1)$ be a constant. Suppose that conditions (C1) and (C2) hold. If $\alpha_n \to a = 1 - c_p/(1-c_k)$, then for all $\varepsilon > 0$, for sufficiently large n, for all $M \in \mathcal{M}_n$,

$$Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)\boldsymbol{T}_2\}|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right) \leq 4n\exp\left\{-\frac{(n-k_n-p_n+1)\min\{\varepsilon^2,1\}}{128}\right\}.$$

Proof. Let $\boldsymbol{\Omega} = (n-k_n)(\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{\mathcal{E}})^{-1}$. For all $M \in \mathcal{M}_n$, it follows from a triangle inequality that

$$\begin{aligned}|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)\boldsymbol{T}_2\}| \leq{}& \alpha_n|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega})\}| \\ &+ |\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\}|.\end{aligned}$$

Hence,

$$
\begin{aligned}
&Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)\boldsymbol{T}_2\}|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right) \\
&\leq Pr\left(\left\{\frac{\alpha_n|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega})\}|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right) \\
&\quad + Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\}|}{R_n(M)} > \frac{\varepsilon}{2}\right).
\end{aligned}
$$

Because $\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)\} \leq R_n(M)$,

$$
\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega})\}|}{R_n(M)} \leq \sigma_1(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega}).
$$

Lemma A.5 yields that on $E_{n,\gamma}$, $\sigma_1(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega})$ converges to zero. Note that $\alpha_n \to 1 - c_p/(1-c_k) > 0$. These results indicate that

$$
Pr\left(\left\{\frac{\alpha_n|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega})\}|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right) = 0,
$$

for sufficiently large n.

On the other hand, let $\boldsymbol{e}_i$ is the ith column of $\boldsymbol{I}_n$ and if $\boldsymbol{T}_1(M)^\top \boldsymbol{e}_i \neq \boldsymbol{0}_{p_n}$, $\boldsymbol{a}_i(M) = (\boldsymbol{e}_i^\top \boldsymbol{T}_1(M)\boldsymbol{T}_1(M)^\top \boldsymbol{e}_i)^{-1/2}\boldsymbol{T}_1(M)^\top \boldsymbol{e}_i$, otherwise $\boldsymbol{a}_i(M) = \boldsymbol{0}_{p_n}$. Then, it can be seen that

$$
\begin{aligned}
|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_n)\}| &\leq \sum_{i=1}^n |\boldsymbol{e}_i^\top \boldsymbol{T}_1(M)(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_n)\boldsymbol{T}_1(M)^\top \boldsymbol{e}_i| \\
&\leq \sum_{i=1}^n \boldsymbol{e}_i^\top \boldsymbol{T}_1(M)\boldsymbol{T}_1(M)^\top \boldsymbol{e}_i \max_{1\leq i\leq n} |\boldsymbol{a}_i(M)^\top (\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\boldsymbol{a}_i(M)| \\
&= \mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)\} \max_{1\leq i\leq n} |\boldsymbol{a}_i(M)^\top (\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\boldsymbol{a}_i(M)|.
\end{aligned}
$$

It is worth mentioning that when $\boldsymbol{a}_i(M) \neq \boldsymbol{0}_{p_n}$, $(\boldsymbol{a}_i(M)^\top (\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}^\perp \boldsymbol{\mathcal{E}})^{-1}\boldsymbol{a}_i(M))^{-1}$ follows chi-square distribution with $n - k_n - p_n + 1$ degrees of freedom (see, Theorem 3.2.11, Muirhead 1982). Because $(n-k_n)\alpha_n/(n-k_n-p_n+1) \to 1$, Lemma A.4 yields that for sufficiently large n, for all $M \in \mathcal{M}_n$

$$
\begin{aligned}
Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top \boldsymbol{T}_1(M)(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_n)\}|}{R_n(M)} > \frac{\varepsilon}{2}\right) &\leq Pr\left(\max_{1\leq i\leq n} |\boldsymbol{a}_i(M)^\top (\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\boldsymbol{a}_i(M)| > \frac{\varepsilon}{2}\right) \\
&\leq nPr\left(\left|\left(\frac{\chi^2_{n-k_n-p_n+1}}{(n-k_n)\alpha_n}\right)^{-1} - 1\right| > \frac{\varepsilon}{2}\right) \\
&\leq 4n\exp\left\{-\frac{(n-k_n-p_n+1)\min\{\varepsilon^2, 1\}}{128}\right\}.
\end{aligned}
$$

Hence, the proof is completed. □

An evaluation of the sixth term of $B_n(M)$ is obtained in a similar manner to Lemma A.10.

Lemma A.11. Let $\gamma \in (0,1)$ be a constant. Suppose that conditions (C1)–(C3) hold. If $\alpha_n \to a = 1 - c_p/(1-c_k)$, for all $\varepsilon > 0$, for sufficiently large n, for all $M \in \mathcal{M}_n$

$$Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\}|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right)$$
$$\leq 2\exp\left\{-\frac{\min\{\varepsilon,1\}\varepsilon R_n(M)}{8C_A^2}\right\} + 4n\exp\left\{-\frac{(n-k_n-p_n+1)\varepsilon^2}{128(1+\varepsilon)^2}\right\}.$$

Proof. Let $\boldsymbol{b}_i(M) = (\boldsymbol{e}_i^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{e}_i)^{-1/2}\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{e}_i$ if $\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{e}_i \neq \boldsymbol{0}_{p_n}$, and $\boldsymbol{b}_i(M) = \boldsymbol{0}_{p_n}$ otherwise. Note that $\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{e}_i = \boldsymbol{0}_{p_n}$ if and only if $\boldsymbol{e}_i^\top\boldsymbol{P}_{M_j}\boldsymbol{e}_i = 0$ for all $j = 1,\ldots,p_n$. Then, it can be seen that

$$|\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\}| \leq \max_{1\leq i\leq n}|\boldsymbol{b}_i(M)^\top\boldsymbol{T}_2\boldsymbol{b}_i(M)|\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\}.$$

Because $\mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\} \leq R_n(M)$, it follows from (17) that

$$Pr\left(\frac{\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\}}{R_n(M)} > 1+\varepsilon\right)$$
$$\leq Pr\left(\frac{|\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\} - \mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\}|}{R_n(M)} > \varepsilon\right)$$
$$\leq 2\exp\left\{-\frac{\min\{\varepsilon,1\}\varepsilon R_n(M)}{8C_A^2}\right\}.$$

Next, $(1+\varepsilon)\max_{1\leq i\leq n}|\boldsymbol{b}_i(M)^\top\boldsymbol{T}_2\boldsymbol{b}_i(M)|$ is evaluated. We see that

$$|\boldsymbol{b}_i(M)^\top\boldsymbol{T}_2\boldsymbol{b}_i(M)| \leq \alpha_n\sigma_1(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega}) + \max_{1\leq i\leq n}|\boldsymbol{b}_i(M)^\top(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\boldsymbol{b}_i(M)|.$$

Lemma A.5 and $\alpha_n \to 1 - c_p/(1-c_k) > 0$ under (C1) imply that

$$\alpha_n\sigma_1(\boldsymbol{\Sigma}_*^{1/2}\boldsymbol{S}^{-1}\boldsymbol{\Sigma}_*^{1/2} - \boldsymbol{\Omega}) = o(1),$$

on $E_{n,\gamma}$. Moreover, because $\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{E}}^\top\boldsymbol{P}_{M_F}^\perp\boldsymbol{\mathcal{E}}$ are independent, when $\boldsymbol{b}_i(M) \neq \boldsymbol{0}_{p_n}$, $(n-k_n)\{\boldsymbol{b}_i(M)^\top\boldsymbol{\Omega}\boldsymbol{b}_i(M)\}^{-1}$ follows chi-square distribution with $n-k_n-p_n+1$ degrees of freedom (see, Theorem 3.2.12, Muirhead 1982). Hence, a similar argument of the proof of Lemma A.10 yields that for sufficiently large n, for all $M \in \mathcal{M}_n$, it holds that

$$Pr\left(\max_{1\leq i\leq n}|\boldsymbol{b}_i(M)^\top(\alpha_n\boldsymbol{\Omega} - \boldsymbol{I}_{p_n})\boldsymbol{b}_i(M)| > \frac{\varepsilon}{2(1+\varepsilon)}\right) \leq 4n\exp\left\{-\frac{(n-k_n-p_n+1)\varepsilon^2}{128(1+\varepsilon)^2}\right\}.$$

By combining these results, we complete the proof. □

Next, we give an evaluation of the fourth term of $B_n(M)$.

Lemma A.12. Let $\gamma \in (0,1)$ be a constant. Suppose that conditions (C1)–(C3) hold. There exists a constant $C_T > 0$ such that for sufficiently large n, for all $\varepsilon > 0$ and $M \in \mathcal{M}_n$,

$$Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{T}_1(M)^\top(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2})\boldsymbol{T}_2\}|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right) \le 2\exp\left\{-\frac{\varepsilon^2 R_n(M)}{2C_T^2(1+C_A)^2}\right\}.$$

Proof. From a simple matrix transformation, we obtain

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{T}_1(M)^\top(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2})\boldsymbol{T}_2\} &= \mathrm{vec}(\boldsymbol{T}_1(M))^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\mathrm{vec}(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}) \\ &= \mathrm{vec}(\boldsymbol{T}_1(M))^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np_n} - \boldsymbol{A}(M))(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G})\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}).\end{aligned}$$

Because $\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}$ and $\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}$ are independent, given $\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}$,

$$\mathrm{vec}(\boldsymbol{T}_1(M))^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)(\boldsymbol{I}_{np_n} - \boldsymbol{A}(M))(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G})\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}) \sim N(0, \kappa(M)^2),$$

where $\kappa(M) = \|(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G}^\top)(\boldsymbol{I}_{np_n} - \boldsymbol{A}(M)^\top)(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\mathrm{vec}(\boldsymbol{T}_1(M))\|_2$. Note that

$$\begin{aligned}\kappa(M) &\le \{1 + \sigma_1(\boldsymbol{A}(M))\}\sigma_1(\boldsymbol{T}_2)\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)\}^{1/2} \\ &\le (1 + C_A)R_n(M)^{1/2}\sigma_1(\boldsymbol{T}_2).\end{aligned}$$

We use condition (C3) and $\mathrm{tr}\{\boldsymbol{T}_1(M)^\top\boldsymbol{T}_1(M)\} \le R_n(M)$ in the last inequality. Lemma A.5 indicates that on $E_{n,\gamma}$, there exists a positive constant $C_T > 0$ such that for sufficiently large n, $\sigma_1(\boldsymbol{T}_2) \le C_T$, which yields that

$$\frac{\kappa(M)^2}{R_n(M)} \le C_T^2(1 + C_A)^2.$$

Therefore, Lemma A.1 completes the proof. □

Finally, the fifth term is evaluated.

Lemma A.13. Let $\gamma \in (0,1)$ be a constant. Suppose that conditions (C1)–(C3) hold. There exists a constant $C_T > 0$ such that for all $\varepsilon > 0$, for sufficiently large n, for all $M \in \mathcal{M}_n$

$$
\begin{aligned}
&Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\}|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right)\\
&\leq 2\exp\left\{-\frac{\varepsilon R_n(M)}{32C_T}\min\left\{\frac{\varepsilon}{C_T}, \frac{2}{C_A}\right\}\right\} + 4nk_np_n\exp\left\{-\frac{(n-k_n-p_n+1)\min\{\varepsilon^2,4\}}{512}\right\}.
\end{aligned}
$$

Proof. Let $\boldsymbol{T}_3(M) = (\boldsymbol{I}_{p_n} \otimes \boldsymbol{G}^\top)\{(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\boldsymbol{A}(M) + \boldsymbol{A}(M)^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\}(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G})/2$. Then, because $\boldsymbol{P}_{M_F} = \boldsymbol{G}\boldsymbol{G}^\top$,

$$
\begin{aligned}
&\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\}\\
&= \mathrm{vec}(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}})^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\mathrm{vec}(\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2})\\
&= \mathrm{vec}(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}})^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\boldsymbol{A}(M)\mathrm{vec}(\boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}})\\
&= \mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})^\top(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G}^\top)(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\boldsymbol{A}(M)(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G})\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})\\
&= \mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})^\top\boldsymbol{T}_3(M)\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}).
\end{aligned}
$$

Because $\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}$ and $\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}$ are independent, a conditional expectation of $\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})^\top\boldsymbol{T}_3(M)\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})$ given $\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}$ is $\mathrm{tr}\{\boldsymbol{T}_3(M)\}$. Taking account of this point, we divide the probability into two parts as follows:

$$
\begin{aligned}
&Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{\mathcal{E}}^\top \boldsymbol{P}_{M_F}\boldsymbol{\mathcal{E}}(M)\boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{T}_2\}|}{R_n(M)} > \varepsilon\right\} \cap E_{n,\gamma}\right)\\
&\leq Pr\left(\left\{\frac{|\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})^\top\boldsymbol{T}_3(M)\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}) - \mathrm{tr}\{\boldsymbol{T}_3(M)\}|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right)\\
&\quad + Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{T}_3(M)\}|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right).
\end{aligned}
$$

Here, from Lemma A.5, on $E_{n,\gamma}$, there exists a positive constant $C_T > 0$ such that for sufficiently large

n, $\sigma_1(\boldsymbol{T}_2) \leq C_T$. Then,

$$
\begin{aligned}
\sigma_1(\boldsymbol{T}_3(M)) &\leq \sigma_1(\boldsymbol{T}_2)\sigma_1(\boldsymbol{A}(M)) \leq C_T C_A, \\
\mathrm{tr}\{\boldsymbol{T}_3(M)^2\} &\leq \frac{1}{4}\mathrm{tr}\{\{(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\boldsymbol{A}(M) + \boldsymbol{A}(M)^\top(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\}^2\} \\
&= \frac{1}{2}\mathrm{tr}\{\boldsymbol{A}(M)(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\boldsymbol{A}(M)(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\} \\
&\quad + \frac{1}{2}\mathrm{tr}\{\boldsymbol{A}(M)^\top(\boldsymbol{T}_2^2 \otimes \boldsymbol{I}_n)\boldsymbol{A}(M)\} \\
&= \frac{1}{2}\mathrm{vec}(\boldsymbol{A}(M)^\top)^\top\{(\boldsymbol{T}_2 \otimes \boldsymbol{I}_n) \otimes (\boldsymbol{T}_2 \otimes \boldsymbol{I}_n)\}\mathrm{vec}(\boldsymbol{A}(M)) \\
&\quad + \frac{1}{2}\mathrm{vec}(\boldsymbol{A}(M))^\top\{\boldsymbol{I}_{np_n} \otimes (\boldsymbol{T}_2^2 \otimes \boldsymbol{I}_n)\}\mathrm{vec}(\boldsymbol{A}(M)) \\
&\leq \frac{C_T^2}{2}\sqrt{\mathrm{vec}(\boldsymbol{A}(M)^\top)^\top\mathrm{vec}(\boldsymbol{A}(M)^\top)}\sqrt{\mathrm{vec}(\boldsymbol{A}(M))^\top\mathrm{vec}(\boldsymbol{A}(M))} \\
&\quad + \frac{C_T^2}{2}\mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\}\} \\
&= \frac{C_T^2}{2}\{\mathrm{tr}\{\boldsymbol{A}(M)\} + \mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\}\} \\
&\leq C_T^2 R_n(M),
\end{aligned}
$$

where the last inequality follows from (18) and $\mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\} \leq R_n(M)$. Thus, by considering a conditional probability given $\boldsymbol{H}^\top\boldsymbol{\mathcal{E}}$, Lemma A.3 yields

$$
\begin{aligned}
&Pr\left(\left\{\frac{|\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}})^\top\boldsymbol{T}_3(M)\mathrm{vec}(\boldsymbol{G}^\top\boldsymbol{\mathcal{E}}) - \mathrm{tr}\{\boldsymbol{T}_3(M)\}|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right) \\
&\leq 2\exp\left\{-\frac{\varepsilon R_n(M)}{32C_T}\min\left\{\frac{\varepsilon}{C_T}, \frac{2}{C_A}\right\}\right\}. \qquad (19)
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\mathrm{tr}\{\boldsymbol{T}_3(M)\} &= \frac{1}{2}\mathrm{tr}\{(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G})(\boldsymbol{T}_2 \otimes \boldsymbol{I}_{k_n})(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G}^\top)(\boldsymbol{A}(M) + \boldsymbol{A}(M)^\top)\} \\
&= \frac{1}{2}\sum_{i=1}^{np_n} d_i(M)\boldsymbol{v}_i(M)^\top(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G})(\boldsymbol{T}_2 \otimes \boldsymbol{I}_{k_n})(\boldsymbol{I}_{p_n} \otimes \boldsymbol{G}^\top)\boldsymbol{v}_i(M),
\end{aligned}
$$

where $d_i(M)$ is the ith eigenvalue of $\boldsymbol{A}(M)+\boldsymbol{A}(M)^\top$ and $\boldsymbol{v}_i(M)$ is its eigenvector with $\boldsymbol{v}_i(M)^\top\boldsymbol{v}_i(M) = 1$. By using a commutation matrix $\boldsymbol{K}_{p_n k_n}$, we can express $\boldsymbol{T}_2 \otimes \boldsymbol{I}_{k_n} = \boldsymbol{K}_{p_n k_n}(\boldsymbol{I}_{k_n} \otimes \boldsymbol{T}_2)\boldsymbol{K}_{p_n k_n}^\top$, where $\boldsymbol{K}_{p_n k_n}\boldsymbol{K}_{p_n k_n}^\top = \boldsymbol{I}_{np_n}$ (see, e.g., sections 1.3.2 and 1.3.3, Kollo and von Rosen 2005). Let $\boldsymbol{u}_i(M) = \boldsymbol{K}_{p_n k_n}^\top(\boldsymbol{I}_{p_n}\otimes\boldsymbol{G}^\top)\boldsymbol{v}_i(M) = (\boldsymbol{u}_{i1}(M)^\top, \ldots, \boldsymbol{u}_{ik_n}(M)^\top)^\top$, where $\boldsymbol{u}_{ij}(M) \in \mathbb{R}^{p_n}$. Define $\boldsymbol{z}_{ij}(M) = (\boldsymbol{u}_{ij}(M)^\top\boldsymbol{u}_{ij}(M))^{-1/2}\boldsymbol{u}_{ij}($

if $\boldsymbol{u}_{ij}(M) \neq \mathbf{0}_{p_n}$ and $\boldsymbol{z}_{ij}(M) = \mathbf{0}$ otherwise. Then,

$$\begin{aligned}
\mathrm{tr}\{\boldsymbol{T}_3(M)\} &= \frac{1}{2}\sum_{i=1}^{np_n} d_i(M) \sum_{j=1}^{k_n} \boldsymbol{u}_{ij}(M)^\top \boldsymbol{T}_2 \boldsymbol{u}_{ij}(M) \\
&= \frac{1}{2}\sum_{i=1}^{np_n} d_i(M) \sum_{j=1}^{k_n} \boldsymbol{u}_{ij}(M)^\top \boldsymbol{u}_{ij}(M) \boldsymbol{z}_{ij}(M)^\top \boldsymbol{T}_2 \boldsymbol{z}_{ij}(M).
\end{aligned}$$

Note that from Corollary 3.4.3 in Horn and Jornson (1994), the following is established:

$$\frac{1}{2}\sum_{i=1}^{np_n} |d_i(M)| = \frac{1}{2}\sum_{i=1}^{np_n} \sigma_i(\boldsymbol{A}(M) + \boldsymbol{A}(M)^\top) \leq \sum_{i=1}^{np_n} \sigma_i(\boldsymbol{A}(M)).$$

Because $\boldsymbol{u}_i(M)^\top \boldsymbol{u}_i(M) = \sum_{j=1}^{k_n} \boldsymbol{u}_{ij}(M)^\top \boldsymbol{u}_{ij}(M) \leq 1$, we see that

$$\begin{aligned}
|\mathrm{tr}\{\boldsymbol{T}_3(M)\}| &\leq \frac{1}{2}\sum_{i=1}^{np_n} |d_i(M)| \sum_{j=1}^{k_n} \boldsymbol{u}_{ij}(M)^\top \boldsymbol{u}_{ij}(M) \max_{\substack{1\leq i\leq np_n \\ 1\leq j\leq k_n}} |\boldsymbol{z}_{ij}(M)^\top \boldsymbol{T}_2 \boldsymbol{z}_{ij}(M)| \\
&\leq \sum_{i=1}^{np_n} \sigma_i(\boldsymbol{A}(M)) \max_{\substack{1\leq i\leq np_n \\ 1\leq j\leq k_n}} |\boldsymbol{z}_{ij}(M)^\top \boldsymbol{T}_2 \boldsymbol{z}_{ij}(M)| \\
&\leq R_n(M) \max_{\substack{1\leq i\leq np_n \\ 1\leq j\leq k_n}} |\boldsymbol{z}_{ij}(M)^\top \boldsymbol{T}_2 \boldsymbol{z}_{ij}(M)|,
\end{aligned}$$

where for the last inequality, we use (18) and $\mathrm{tr}\{\boldsymbol{A}(M)^\top \boldsymbol{A}(M)\} \leq R_n(M)$. As seen in the proof of Lemma A.10, we obtain for sufficiently large n, for all $M \in \mathcal{M}_n$,

$$\begin{aligned}
Pr\left(\left\{\frac{|\mathrm{tr}\{\boldsymbol{T}_3(M)\}|}{R_n(M)} > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right) &\leq Pr\left(\left\{\max_{\substack{1\leq i\leq np_n \\ 1\leq j\leq k_n}} |\boldsymbol{z}_{ij}(M)^\top \boldsymbol{T}_2 \boldsymbol{z}_{ij}(M)| > \frac{\varepsilon}{2}\right\} \cap E_{n,\gamma}\right) \\
&\leq 4nk_n p_n \exp\left\{-\frac{(n-k_n-p_n+1)\min\{\varepsilon^2, 4\}}{512}\right\}. \qquad (20)
\end{aligned}$$

From (19) and (20), the proof is completed. □

Thus, by combining there results, Lemma A.6 is obtained.

A.3 Proof of Theorem 2

To show AME, the following lemma plays an important role.

Lemma A.14. Let X_n and Y_n be random variables such that $Y_n \geq 1$. If $E((X_n/Y_n - 1)^2) \to 0$ and $E(Y_n^2) \to 1$, then

$$\lim_{n\to\infty} E(X_n) = 1.$$

Proof. Because $X_n - 1$ can be decomposed as $X_n - 1 = (X_n/Y_n - 1)Y_n + (Y_n - 1)$, by applying a triangle inequality and Cauchy-Schwarz inequality, it follows that

$$
\begin{aligned}
|E(X_n - 1)| &\le \left| E\left(\left(\frac{X_n}{Y_n} - 1\right) Y_n\right)\right| + |E(Y_n - 1)| \\
&\le E\left(\left(\frac{X_n}{Y_n} - 1\right)^2\right)^{1/2} E(Y_n^2)^{1/2} + E((Y_n - 1)^2)^{1/2}.
\end{aligned}
$$

Because $Y_n \ge 1$ from the assumption, we can see that $1 \le E(Y_n) \le E(Y_n^2) \to 1$. Hence, the right-hand side of the above inequality goes to 0. □

Showing the conditions of Lemma A.14 with

$$
X_n = \frac{L_n(\hat{M}_n)}{R_n(M_R^*)}, \quad Y_n = \frac{R_n(\hat{M}_n)}{R_n(M_R^*)},
$$

i.e., $E(Y_n^2) \to 1$ and $E((X_n/Y_n - 1)^2) \to 0$, then we can show that GC_p has AME.

Lemma A.15. Under conditions (C1)–(C6), if $\alpha_n \to a = 1 - c_p/(1 - c_k)$, then

$$
\lim_{n\to\infty} \frac{E(R_n(\hat{M}_n)^2)}{R_n(M_R^*)^2} = 1.
$$

Proof. As seen in Shibata (1983), for all $\eta > 0$, a set of candidate models $\mathcal{M}_n$ is separated into $\mathcal{M}_n^{(1)}$ and $\mathcal{M}_n^{(2)}$ such that

$$
\begin{aligned}
\mathcal{M}_n^{(1)} &= \left\{ M \in \mathcal{M}_n \,\middle|\, \frac{R_n(M)}{R_n(M_R^*)} \le (1 + \eta) \right\}, \\
\mathcal{M}_n^{(2)} &= \left\{ M \in \mathcal{M}_n \,\middle|\, \frac{R_n(M)}{R_n(M_R^*)} > (1 + \eta) \right\}.
\end{aligned}
$$

Then, $E(R_n(\hat{M}_n)^2)/R_n(M_R^*)^2 - 1$ can be decomposed as follows:

$$
\begin{aligned}
\frac{E(R_n(\hat{M}_n)^2)}{R_n(M_R^*)^2} - 1 = &\sum_{M\in\mathcal{M}_n^{(1)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\hat{M}_n = M) - 1 \\
&+ \sum_{M\in\mathcal{M}_n^{(2)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\hat{M}_n = M).
\end{aligned}
$$

If the last term of the right-hand side of the above equation converges to zero, it holds that $Pr(\hat{M}_n \in \mathcal{M}_n^{(2)}) \to 0$ because $R_n(M)^2/R_n(M_R^*)^2 > (1 + \eta)^2$ for all $M \in \mathcal{M}_n^{(2)}$. This then leads to $Pr(\hat{M}_n \in \mathcal{M}_n^{(1)}) \to 1$. On the other hand, $1 \le R_n(M)^2/R_n(M_R^*)^2 \le (1 + \eta)^2$ for all $M \in \mathcal{M}_n^{(1)}$. Hence, for sufficiently large n,

$$\left|\sum_{M\in\mathcal{M}_n^{(1)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\hat{M}_n = M) - 1\right| \leq (1+\eta)^2 Pr(\hat{M}_n \in \mathcal{M}_n^{(1)}) - 1$$
$$\leq (1+\eta)^2 - 1.$$

Because η is an arbitrary positive constant, it follows that

$$\lim_{n\to\infty} \frac{E(R_n(\hat{M}_n)^2)}{R_n(M_R^*)^2} = 1$$

if it holds that

$$\lim_{n\to\infty} \sum_{M\in\mathcal{M}_n^{(2)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\hat{M}_n = M) = 0.$$

We can see that

$$\begin{aligned}
&\sum_{M\in\mathcal{M}_n^{(2)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\hat{M}_n = M) \\
\leq &\sum_{M\in\mathcal{M}_n^{(2)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\{\hat{M}_n = M\} \cap E_{n,\gamma}) \\
&+ \max_{M\in\mathcal{M}_n} \frac{R_n(M)^2}{R_n(M_R^*)^2} \sum_{M\in\mathcal{M}_n^{(2)}} Pr(\{\hat{M}_n = M\} \cap E_{n,\gamma}^c) \\
\leq &\sum_{M\in\mathcal{M}_n^{(2)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\{\hat{M}_n = M\} \cap E_{n,\gamma}) + \max_{M\in\mathcal{M}_n} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(E_{n,\gamma}^c).
\end{aligned}$$

Hence, it is shown from (11) that there exists a positive constant $C > 0$ such that

$$\max_{M\in\mathcal{M}_n} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(E_{n,\gamma}^c) \leq 2 \max_{M\in\mathcal{M}_n} \frac{R_n(M)^2}{R_n(M_R^*)^2} \exp(-Cn^\gamma).$$

From condition (C6), there exists $\gamma \in (0,1)$ such that this goes to zero with n tending to infinity.

On the other hand, for all $M \in \mathcal{M}_n^{(2)}$,

$$\begin{aligned}
\hat{M}_n = M &\Rightarrow GC_p(M_R^*; \alpha_n) - GC_p(M; \alpha_n) \geq 0 \\
&\Leftrightarrow \frac{GC_p(M_R^*; \alpha_n) - GC_p(M; \alpha_n)}{R_n(M)} - \frac{R_n(M_R^*) - R_n(M)}{R_n(M)} \geq 1 - \frac{R_n(M_R^*)}{R_n(M)} \\
&\Rightarrow \frac{R_n(M) - GC_p(M; \alpha_n)}{R_n(M)} - \frac{R_n(M_R^*) - GC_p(M_R^*; \alpha_n)}{R_n(M)} \geq \frac{\eta}{1+\eta} \\
&\Rightarrow \left| \frac{L_n(M)}{R_n(M)} - 1 \right| + \left| \frac{L_n(M_R^*) - R_n(M_R^*)}{R_n(M)} \right| + \frac{|B_n(M)|}{R_n(M)} + \frac{|B_n(M_R^*)|}{R_n(M)} \geq \frac{\eta}{1+\eta}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
Pr(\{\hat{M}_n = M\} \cap E_{n,\gamma}) \leq & Pr\left(\left| \frac{L_n(M)}{R_n(M)} - 1 \right| \geq \frac{\eta}{4(1+\eta)} \right) \\
&+ Pr\left(\left| \frac{L_n(M_R^*)}{R_n(M_R^*)} - 1 \right| \geq \frac{\eta R_n(M)}{4(1+\eta) R_n(M_R^*)} \right) \\
&+ Pr\left(\left\{ \frac{|B_n(M)|}{R_n(M)} \geq \frac{\eta}{4(1+\eta)} \right\} \cap E_{n,\gamma} \right) \\
&+ Pr\left(\left\{ \frac{|B_n(M_R^*)|}{R_n(M_R^*)} \geq \frac{\eta R_n(M)}{4(1+\eta) R_n(M_R^*)} \right\} \cap E_{n,\gamma} \right).
\end{aligned}$$

It follows from Lemmas A.6 and A.7 that under conditions (C1)–(C6), for all $\varepsilon > 0$, for sufficiently large n,

$$\sum_{M \in \mathcal{M}_n^{(2)}} \frac{R_n(M)^2}{R_n(M_R^*)^2} Pr(\{\hat{M}_n = M\} \cap E_{n,\gamma}) < \varepsilon.$$

Therefore, the proof is completed. □

Lemma A.16. Under condition (C3), it can be seen that

$$\lim_{n \to \infty} E\left(\left(\frac{L_n(\hat{M}_n)}{R_n(\hat{M}_n)} - 1 \right)^2 \right) = 0.$$

Proof. Like in the case of Lemma A.7, define

$$\xi(M) = \frac{L_n(M) - R_n(M)}{R_n(M)}.$$

For all $\varepsilon > 0$,

$$
\begin{aligned}
E(\xi(\hat{M}_n)^2) &= E(\xi(\hat{M}_n)^2 \mathrm{I}(|\xi(\hat{M}_n)| > \varepsilon)) + E(\xi(\hat{M}_n)^2 \mathrm{I}(|\xi(\hat{M}_n)| \le \varepsilon)) \\
&\le E(\xi(\hat{M}_n)^2 \mathrm{I}(|\xi(\hat{M}_n)| > \varepsilon)) + \varepsilon^2 \\
&\le \sum_{M \in \mathcal{M}_n} E(\xi(M)^2 \mathrm{I}(|\xi(M)| > \varepsilon) \mathrm{I}(\hat{M}_n = M)) + \varepsilon^2 \\
&\le \sum_{M \in \mathcal{M}_n} E(\xi(M)^2 \mathrm{I}(|\xi(M)| > \varepsilon)) + \varepsilon^2 \\
&\le \sum_{M \in \mathcal{M}_n} E(\xi(M)^4)^{1/2} Pr(|\xi(M)| > \varepsilon)^{1/2} + \varepsilon^2.
\end{aligned}
$$

The last inequality follows from Cauchy-Schwarz inequality. From the proof of Lemma A.7, for all $M \in \mathcal{M}_n$,

$$
\xi(M) = \frac{2v(M)Z}{R_n(M)} + \sum_{i=1}^{np_n} \frac{\lambda_i(\boldsymbol{A}(M)^\top \boldsymbol{A}(M))}{R_n(M)} (Z_i^2 - 1),
$$

where $v(M)$ is defined in Lemma A.7, which satisfies $v(M)^2 \le C_A^2 R_n(M)$, and $Z, Z_1, \ldots, Z_{np_n}$ are independent and identically distributed as $N(0,1)$. Because it holds with a constant $C > 0$ that for sufficiently large n, for all $M \in \mathcal{M}_n$

$$
E\left(\left\{ \sum_{i=1}^{np_n} \frac{\lambda_i(\boldsymbol{A}(M)^\top \boldsymbol{A}(M))}{R_n(M)} (Z_i^2 - 1) \right\}^4 \right) \le \frac{C}{R_n(M)^2},
$$

$\max_{M \in \mathcal{M}_n} E[|\xi(M)|^4]$ is bounded. Thus, there exists a positive constant C_ξ such that

$$
\max_{M \in \mathcal{M}_n} E(\xi(M)^4) \le C_\xi^2.
$$

From Lemma A.7, we can see that there exists a positive constant C_ε such that for sufficiently large n,

$$
\begin{aligned}
\sum_{M \in \mathcal{M}_n} E(\xi(M)^4)^{1/2} Pr(|\xi(M)| > \varepsilon)^{1/2} &\le 2C_\xi \sum_{M \in \mathcal{M}_n} \exp(-C_\varepsilon R_n(M)) \\
&\le \varepsilon.
\end{aligned}
$$

Hence, we have

$$
E(\xi(\hat{M}_n)^2) \le \varepsilon + \varepsilon^2.
$$

Because ε can be arbitrary small, the proof is completed. □

By combining Lemmas A.15 and A.16 with A.14, Theorem 2 is established.