Statistica Sinica **33** (2023), 1233-1248
doi:https://doi.org/10.5705/ss.202020.0425

ASYMPTOTIC OPTIMALITY OF C_P-TYPE CRITERIA IN HIGH-DIMENSIONAL MULTIVARIATE LINEAR REGRESSION MODELS

Shinpei Imori

Hiroshima University

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, we show 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 the GC_p with several tuning parameters is compared by means of a simulation study.

Key words and phrases: Asymptotic theory, high-dimensional statistical inference, model selection/variable selection.

1. Introduction

Variable selection problems are crucial in statistical fields for improving the prediction accuracy and/or interpretability of a resultant model. There is a burgeoning body of literature that 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 the Akaike information criterion (AIC) (Akaike (1974)) are 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, the Bayesian information criterion (BIC) proposed by Schwarz (1978) is consistent (Nishii (1984)) under appropriate conditions; that is, the probability that a model selected by the BIC coincides with the true model converges to one as the sample size n tends to infinity. In this sense, the BIC is a feasible method from the

Corresponding author: Shinpei Imori, Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan 739-8526. E-mail: imori@hiroshima-u.ac.jp.

perspective of interpretability. However, the C_p and AIC are inconsistent (Nishii (1984)) under the same condition. The properties of the selection procedures are well studied in Shao (1997) in the context of univariate linear regression models. In this study, we focus on multivariate linear regression models.

High-dimensional data are often encountered, where the dimension of a response matrix in multivariate linear regression models p_n is large, but does not exceed the sample size n. In 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, the properties derived when p_n is fixed can be altered in high-dimensional situations. For example, Yanagihara, Wakaki and Fujikoshi (2015) showed that the AIC acquires the consistency property and the 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 the difference in the risks between two over-specified models (i.e., models including the true model) diverges with n and p_n tending to infinity, and thus the penalty terms of the C_p and AIC are moderate, but that of the 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, few studies have examined the asymptotic optimality for prediction in high-dimensional data contexts. Conventional results derived from univariate models are no longer reliable in high-dimensional data contexts, and an extension to such cases is not mathematically trivial. In this study, we focus on the 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 generalized C_p (GC_p) exhibits ALE and AME in high-dimensional data. We also show that one of the sufficient conditions is necessary for the GC_p to exhibit both of these efficiencies. As a result, we show 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) studied the ALE and AME of the GC_p in high-dimensional multivariate linear regression models, although the 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 we do not assume here. Note that previous studies on variable selection in multivariate linear regression models usually use a common regression model among the response variables. We mitigate this limitation, and allow each response variable to have a different model in order to consider more practical situations, such as response variables having a group structure.

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

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 the response, and k_n is the number of the 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 $\boldsymbol{\Gamma}_* = (\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_{p_n}^*) = E(\boldsymbol{Y})$, $\boldsymbol{\mathcal{E}}$ is an $n \times p_n$ error matrix, of which all entries are independent and identically distributed (i.i.d.) 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 $\boldsymbol{X}_{M_j}$, which 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 into account 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 $M_1=\cdots=M_m$ and $M_{m+1}=\cdots=M_{p_n}$ is imposed. Using only one regression model for the response variables, that is, $M_1=\cdots=M_{p_n}$, 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 $\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}$, that is, $\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}_*^{1/2}=(\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_{p_n})$. Thus, a set of candidate models denoted by $\mathcal{M}_n$ 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, that is, $M=(M_F,\ldots,M_F)$.

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

$$L_n(M)=\mathrm{tr}\{(\boldsymbol{\Gamma}_*-\hat{\boldsymbol{\Gamma}}(M))\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\Gamma}_*-\hat{\boldsymbol{\Gamma}}(M))^\top\},\tag{2.1}$$

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

$$\hat{\boldsymbol{\Gamma}}(M)=(\boldsymbol{P}_{M_1}\boldsymbol{y}_1,\ldots,\boldsymbol{P}_{M_{p_n}}\boldsymbol{y}_{p_n}),\tag{2.2}$$

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 (2.2) into (2.1), we have

$$\begin{aligned}L_n(M)=&\,\mathrm{tr}\{\boldsymbol{\Delta}(M)\}-2\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\Gamma}_*-\boldsymbol{\Gamma}_*(M))^\top\boldsymbol{\mathcal{E}}(M)\}\\&+\mathrm{tr}\{\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\mathcal{E}}(M)^\top\boldsymbol{\mathcal{E}}(M)\},\end{aligned}\tag{2.3}$$

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

$$R_n(M)=E(L_n(M))=\mathrm{tr}\{\boldsymbol{\Delta}(M)\}+\mathrm{tr}\{\boldsymbol{A}(M)^\top\boldsymbol{A}(M)\},\tag{2.4}$$

where $\boldsymbol{A}(M)=(\boldsymbol{\Sigma}_*^{-1/2}\otimes\boldsymbol{I}_n)\boldsymbol{P}(M)(\boldsymbol{\Sigma}_*^{1/2}\otimes\boldsymbol{I}_n)$, the symbol $\otimes$ denotes a Kronecker product, and $\boldsymbol{P}(M)=\mathrm{diag}\{\boldsymbol{P}_{M_1},\ldots,\boldsymbol{P}_{M_{p_n}}\}$. Note that $\boldsymbol{A}(M)$ is an idempotent matrix. Thus, from Householder and Carpenter (1963), $\sigma_j(\boldsymbol{A}(M))\leq\sigma_j(\boldsymbol{A}(M))^2$, 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

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

This implies that $R_n(M) \geq p_n$, 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^*, respectively, which minimize (2.1) and (2.4), respectively among $\mathcal{M}_n$, that is,

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

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 the GC_p defined by

$$GC_p(M; \alpha_n) = n\alpha_n \mathrm{tr}\{\hat{\boldsymbol{\Sigma}}(M)\boldsymbol{S}^{-1}\} + 2\sum_{j=1}^{p_n} k_{M_j}, \tag{2.5}$$

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

The best model selected by minimizing the GC_p among $\mathcal{M}_n$ is denoted by $\hat{M}_n$, that is,

$$\hat{M}_n = \operatorname*{argmin}_{M \in \mathcal{M}_n} GC_p(M; \alpha_n).$$

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

$$\frac{L_n(\hat{M}_n)}{L_n(M_L^*)} \xrightarrow{p} 1, \quad n \to \infty, \tag{2.6}$$

and exhibits AME (Shibata (1983)) if

$$\lim_{n\to\infty} \frac{E(L_n(\hat{M}_n))}{R_n(M_R^*)} = 1. \tag{2.7}$$

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 the GC_p.

3. Asymptotic Efficiency of the GC_p

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

First, we assume the following conditions for the 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 $\mathcal{M}_n$, that is, the number of candidate models. Then, $\log \#(\mathcal{M}_n) = o(n)$.

The first part of condition (C1) is weaker than the 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 to evaluate the lowest singular values of a high-dimensional Gaussian random matrix. The final part of (C1) is required to guarantee the regularity of $\boldsymbol{S}$, which can be satisfied asymptotically from the previous three conditions. Condition (C2) is used to ignore the 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 the 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, respectively, 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 the response variables have 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

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

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

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

Theorem 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, that is,*

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

Next, we show the AME of the GC_p. In addition to 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. We 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) indicate that

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

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

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

Theorem 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, that is,*

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

A proof of this theorem is provided in Supplementary Material. For both the ALE and the AME of the 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 the GC_p to acquire ALE and AME. In this section, we show that this is a necessary condition. That is, when $\alpha_n \to a \ne 1 - c_p/(1-c_k)$, there is a situation such that

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

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

even under conditions (C1)–(C6).

For expository purposes, let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$, that is, $k_n = 2$, 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 the 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

$$\frac{GC_p(\{1,2\};\alpha_n) - GC_p(\{1\};\alpha_n)}{n} = -\frac{a(c_p+b)}{1-c_p} + 2c_p + o_p(1). \tag{4.1}$$

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

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

respectively, 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

$$\frac{L_n(\{1\})}{L_n(\{1,2\})} \xrightarrow{p} \frac{c_p+b}{2c_p} \in (0,\infty), \tag{4.2}$$

$$\lim_{n\to\infty} \frac{R_n(\{1\})}{R_n(\{1,2\})} = \frac{c_p+b}{2c_p} \in (0,\infty). \tag{4.3}$$

First, we consider a situation where $a > 0$. Let $b = c_p(1-c_p)/a$. It follows from (4.1) and (4.2) 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

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

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

On the other hand, (4.3) 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 have 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 $R_n(M_R^{**})/R_n(M_R^*) > 1$, the 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, (4.1) implies that $Pr(\hat{M}_n = \{1\}) \to 1$. However, when $b > c_p$, (4.2) and (4.3) yield $Pr(M_L^* = \{1,2\}) \to 1$ and $M_R^* = \{1,2\}$, respectively, for sufficiently large n. Hence, in the same manner as the argument when $a > 0$, the 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 the ALE and AME of the GC_p under conditions (C1)–(C6).

5. Simulation Study

This section presents a simulation study to compare the GC_p among several α_n, where the goodness of the 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 setting, the number of dimensions of the model selected by C_p (resp. BC_p) is larger (resp. smaller) than or equal to that selected by the MC_p. In general, 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 $d_i = 5\sqrt{k_n - i + 1}/k_n$. 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 $1 \leq i, j \leq p_n$. 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 the computational burden, we adopt a nested model set, that is, we select $M^{(1)}$ and $M^{(2)}$ from among $\{\{1\}, \ldots, \{1, \ldots, k_n\}\}$. Note 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, and $p_n = 10$ 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 the variable selection, we calculate the loss functions for each best model.

Table 1 provides the 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, where smaller values are 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 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 $L_n(\hat{M}_n)/R_n(M_R^*)$, except when the sample size is

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. The best values for $L_n(\hat{M}_n)/L_n(M_L^*)$ and $L_n(\hat{M}_n)/R_n(M_R^*)$ are shown in bold for each (n, p_n, k_n). All values are rounded to three 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)

small. These results concur with our theoretical exposition regarding efficiency and inefficiency.

Table 2 shows the average dimensions of the models, that is, $\#(M^{(1)})/2 + \#(M^{(2)})/2$ selected by each GC_p and the loss minimizing models. This indicates that the number of dimensions of the 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, whereas C_p often selects larger models when p_n is large. The average number of dimensions of the models selected by MC_p is close to that of the loss minimizing models in both the high- and the fixed-dimensional situations. This implies that α_n substantially affects the dimensions and the efficiency of selected models.

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

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 three decimal places.

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

6. Conclusion

We have derived sufficient conditions for the ALE and the AME of the GC_p in high-dimensional multivariate linear regression models. We have shown that MC_p exhibits ALE and AME in high-dimensional data, whereas 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 nontrivial bias term is omitted in the original C_p as an estimator of the risk function. This term plays an important role in the adaptation to high-dimensional frameworks. Indeed, if the tuning parameter of the GC_p, α_n, converges to $a \neq 1 - c_p/(1 - c_k)$, as in the case of the C_p and the BC_p, we showed that the GC_p is asymptotically inefficient. We compared the finite-sample performance of C_p-type criteria using simulations, and showed that the MC_p is better than the C_p and BC_p in high-dimensional data.

Note that when p_n is large, the 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, in which case the asymptotic efficiencies of the MC_p hold. Moreover, assuming the response variables have a common model, that is, $M_1 = \cdots =$

M_{p_n}, the MC_p has the consistency property as well under moderate conditions (Fujikoshi, Sakurai and Yanagihara (2014)). Hence, the 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, that is, $p_n \to \infty$.

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 a 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 using these procedures, the ALE and AME can be established by using it in place of $\boldsymbol{S}$. Note 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 study would be 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 the response variables. Under this condition, even when the number of candidate models is exponentially large, that is, $\#(\mathcal{M}_n) = 2^{k_n}$, (C4) holds. Although this condition is not restricted, when considering a situation in which 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. These topics are left for future research.

Supplementary Material

The online Supplementary Material provides the proofs of Theorems 1 and 2.

Acknowledgments

The author would like to thank the associate editor and the two reviewers for their valuable comments and suggestions. This study was supported in part by JSPS KAKENHI Grant Numbers 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**, 716–723.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika* **67**, 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**, 1050–1076.

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

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 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**, 184–200.

Horn, R. A. and Jornson, C. R. (1994). *Topics in Matrix Analysis.* Cambridge University Press, New York.

Householder, A. S. and Carpenter, J. A. (1963). The singular values of involutory and of idempotent matrices. *Numerische Mathematik* **5**, 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**, 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**, 138–150.

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**, 958–975.

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

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

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**, 301–324.

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

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

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

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

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics* **35**, 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**, 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**, 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**, 869–897.

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

Shinpei Imori

Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan 739-8526.

E-mail: imori@hiroshima-u.ac.jp

(Received February 2020; accepted September 2021)