



An unbiased C_p criterion for multivariate ridge regression

Hirokazu Yanagihara^{a,*}, Kenichi Satoh^b

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

^b Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan

ARTICLE INFO

Article history:

Received 9 February 2009

Available online 7 October 2009

AMS 2000 subject classifications:

primary 62J07

secondary 62F07

Keywords:

Bias correction

Mallows' C_p statistic

Model selection

Multivariate linear regression model

Ridge regression

ABSTRACT

Mallows' C_p statistic is widely used for selecting multivariate linear regression models. It can be considered to be an estimator of a risk function based on an expected standardized mean square error of prediction. An unbiased C_p criterion for selecting multivariate linear regression models has been proposed. In this paper, that unbiased C_p criterion is extended to the case of a multivariate ridge regression. It is analytically proved that the proposed criterion has not only a smaller bias but also a smaller variance than the existing C_p criterion, and is the uniformly minimum variance unbiased estimator of the risk function. We show that the criterion has useful properties by means of numerical experiments.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Let \mathbf{Y} be an $n \times p$ observation matrix and \mathbf{X} be an $n \times k$ matrix of non-stochastic explanatory variables of full rank k . Suppose that j denotes a subset of $\omega = \{1, \dots, k\}$ containing k_j elements, and \mathbf{X}_j denotes the $n \times k_j$ matrix consisting of the columns of \mathbf{X} indexed by the elements of j . Then we consider the following candidate model with k_j explanatory variables:

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

where $\boldsymbol{\Xi}_j$ is a $k_j \times p$ unknown regression coefficient matrix and $\boldsymbol{\Sigma}_j$ is a $p \times p$ unknown variance–covariance matrix. In particular, we call the model with $\mathbf{X}_\omega = \mathbf{X}$ the full model.

The multivariate linear regression model in (1) is one of the basic models in multivariate analysis. It is introduced in many textbooks on applied multivariate statistical analysis (see e.g., [1,2]), and even now it is widely applied in chemometrics, engineering, econometrics, psychometrics and many other fields for the prediction of correlated multiple responses using a set of explanatory variables (e.g., [3–5]). In the same way as for the univariate linear regression model, the ordinary least square estimator (LSE) of $\boldsymbol{\Xi}_j$ has a large variance when the number of explanatory variables is large, or correlations among explanatory variables are high. It is a well-known fact that a large variance causes poor prediction. We avoid such a problem by applying a ridge regression procedure [6] to an estimation of the multivariate linear regression model. Unfortunately, ridge regression sometimes shrinks the LSE too much when the number of explanatory variables is large, mainly because ridge regression shrinks the LSE uniformly. However the number of explanatory variables can be reduced by a variable selection procedure. Hence we consider ridge regression and variable selection simultaneously.

* Corresponding author.

E-mail address: yanagi@math.sci.hiroshima-u.ac.jp (H. Yanagihara).

The regression coefficient matrix Ξ_j is estimated by ridge regression, i.e.,

$$\hat{\Xi}_{j,\theta} = \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j' \mathbf{Y}, \quad (2)$$

where $\mathbf{M}_{j,\theta} = \mathbf{X}_j' \mathbf{X}_j + \theta \mathbf{I}_{k_j}$ ($\theta \geq 0$). Notice that $\hat{\Xi}_{j,0}$ is the ordinary maximum likelihood estimator of Ξ_j (or the ordinary LSE of Ξ_j). As for the univariate case, the estimator in (2) is derived from a penalized least square estimation as

$$\hat{\Xi}_{j,\theta} = \arg \min_{\Xi_j} \{ \text{tr} \{ (\mathbf{Y} - \mathbf{X}_j \Xi_j)' (\mathbf{Y} - \mathbf{X}_j \Xi_j) \} + \theta \text{tr}(\Xi_j' \Xi_j) \}.$$

In multivariate ridge regression, optimization of the subset j and the ridge parameter θ is an important problem.

Choosing j and θ so as to minimize a risk function is a very well-known method for model selection. In this paper, we use the expected mean square error (MSE) of prediction as a risk function. This measures the discrepancy between a predictor of \mathbf{Y} and a future observation. However, we cannot directly use such a risk function in a real situation, because it includes unknown parameters. Instead of the risk function, we can use an estimator that is an information criterion. Obtaining an unbiased estimator of the risk function will allow us to correctly evaluate the discrepancy between the predictor of \mathbf{Y} and a future observation. This will further facilitate a remarkable selection of j and θ .

Mallows' C_p statistic [7,8] can be considered to be an estimator of the risk function based on the expected MSE of prediction. Hence, in this paper, we call an estimator of the risk function a C_p criterion. For a single response variable, the discrepancy between the predictor and the future observation is measured by the Euclidean distance. However, in the case of multiple responses, we need to use a distance measure modified by the correlation between response variables. Hence, we have to use the discrepancy function defined by the Mahalanobis distance. Namely, we define the risk function in terms of the expected MSE, but standardized by the true variance–covariance matrix of observation. Such a risk function was proposed in [9]. Since the true variance–covariance matrix is unknown, it must be replaced by its estimator. However, the replacement makes it hard to obtain an unbiased C_p criterion for the ridge regression because the residual sum of squares and the estimated variance–covariance matrix are not independent. Nevertheless, we can develop an unbiased C_p criterion by decomposing the residual sum of squares into two parts, where the first part depends on the estimated variance–covariance matrix and the other part is independent of the estimated variance–covariance matrix. Such a decomposition can be derived from the formula in [10].

The definition of our unbiased C_p criterion is very simple, and it is not necessary to carry out complicated calculations to obtain an unbiased criterion, such as in [11]. In addition, we are able to prove analytically that the proposed criterion has not only a smaller bias but also a smaller variance than the existing C_p criterion. We call it the modified C_p (MC_p) criterion, because our unbiased C_p includes the criterion in [9] as a special case. Recently, Davies, Neath and Cavanaugh [12] showed that Fujikoshi and Satoh's MC_p is the uniformly minimum variance unbiased estimator (UMVUE) of the risk function. We show that the MC_p criterion becomes the UMVUE, even for the case of multivariate ridge regression.

In this paper, we deal with the multivariate ridge regression procedure and the model selection procedure simultaneously. Such a mixed procedure has another advantage over the MSE for prediction. We can show that an optimal θ under fixed j , which minimizes the MSE of prediction, is not 0 for any candidate models. This implicitly suggests that using the mixed procedure can result in the MSE of prediction being less than that for the single procedure.

This paper is organized in the following way: In Section 2, we propose the MC_p criterion for multivariate ridge regression by using the formula in [10]. Several mathematical properties of our criterion are shown in Section 3. In Section 4, we examine the performance of the proposed criterion by conducting numerical simulations. Section 5 contains a discussion and our conclusions. Technical details are provided in the Appendix.

2. Unbiased C_p criterion

Suppose that the true model of \mathbf{Y} can be expressed as

$$\mathbf{Y} \sim N_{n \times p}(\mathbf{\Gamma}_*, \mathbf{\Sigma}_* \otimes \mathbf{I}_n). \quad (3)$$

Let \mathbf{P}_A be the projection matrix to the subspace spanned by the columns of \mathbf{A} , i.e., $\mathbf{P}_A = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. Then, we suppose that the following assumption is satisfied.

- **Assumption:** the full model includes at least the true model, i.e., $\mathbf{P}_{X_{\omega}} \mathbf{\Gamma}_* = \mathbf{\Gamma}_*$.

Let $\hat{\mathbf{Y}}_{j,\theta}$ be the predictor of \mathbf{Y} given by $\hat{\mathbf{Y}}_{j,\theta} = \mathbf{X}_j \hat{\Xi}_{j,\theta}$, where $\hat{\Xi}_{j,\theta}$ is the ridge regression estimator of Ξ_j given by (2), and let \mathbf{U} be an $n \times p$ random variable matrix which is independent of \mathbf{Y} and has the same distribution as \mathbf{Y} . The random variable matrix \mathbf{U} can be regarded as a future observation or imaginary new observation. As a criterion for the goodness of fit of the candidate model, we consider the underlying risk function based on the MSE of prediction, as is proposed in [9];

$$R(j, \theta) = E_{\mathbf{Y}}^* E_{\mathbf{U}}^* \left[\text{tr} \left\{ (\mathbf{U} - \hat{\mathbf{Y}}_{j,\theta}) \mathbf{\Sigma}_*^{-1} (\mathbf{U} - \hat{\mathbf{Y}}_{j,\theta})' \right\} \right],$$

where E^* denotes the expectation under the true model in (3). We regard the model with $j^{(r)}$ and $\theta^{(r)}$ which minimizes $R(j, \theta)$ as the principal best model.

Let \mathcal{E} be an $n \times p$ matrix defined by $\mathbf{Y} - \Gamma_*$. Notice that

$$E_Y^*[\mathcal{E}] = \mathbf{O}_{n,p}, \quad E_Y^*[\mathcal{E}'\mathbf{A}\mathcal{E}] = \text{tr}(\mathbf{A})\Sigma_*, \quad (4)$$

where $\mathbf{O}_{n,p}$ is an $n \times p$ matrix with all elements zero, and \mathbf{A} is an $n \times n$ constant matrix. These expectations imply that

$$E_U^* \left[\text{tr} \left\{ (\mathbf{U} - \hat{\mathbf{Y}}_{j,\theta}) \Sigma_*^{-1} (\mathbf{U} - \hat{\mathbf{Y}}_{j,\theta})' \right\} \right] = np + \text{tr} \left\{ (\Gamma_* - \hat{\mathbf{Y}}_{j,\theta})' (\Gamma_* - \hat{\mathbf{Y}}_{j,\theta}) \Sigma_*^{-1} \right\}.$$

Let $\mathbf{W}_{j,\theta}$ be the residual sum of squares matrix for the ridge regression, i.e.,

$$\mathbf{W}_{j,\theta} = (\mathbf{Y} - \hat{\mathbf{Y}}_{j,\theta})' (\mathbf{Y} - \hat{\mathbf{Y}}_{j,\theta}) = \mathbf{Y}' (\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{Y}. \quad (5)$$

By using \mathcal{E} and $\mathbf{W}_{j,\theta}$, we derive

$$\text{tr} \left\{ (\Gamma_* - \hat{\mathbf{Y}}_{j,\theta})' (\Gamma_* - \hat{\mathbf{Y}}_{j,\theta}) \Sigma_*^{-1} \right\} = \text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1}) + \text{tr}(\mathcal{E}' \mathcal{E} \Sigma_*^{-1}) - 2 \text{tr} \left\{ \mathcal{E}' (\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j') \mathbf{Y} \Sigma_*^{-1} \right\}.$$

The expectations in (4) yield

$$E_Y^* \left[\text{tr} \left\{ \mathcal{E}' (\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j') \mathbf{Y} \Sigma_*^{-1} \right\} \right] = np - p \text{tr}(\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0}).$$

From the results mentioned above, another expression of $R(j, \theta)$ is given by

$$R(j, \theta) = E_Y^* \left[\text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1}) \right] + 2p \text{tr}(\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0}). \quad (6)$$

Therefore we can propose an estimator for the risk function by using an estimator for $E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1})]$.

Let \mathbf{S} be an unbiased estimator of Σ_* under the full model. This is defined by $\mathbf{S} = \mathbf{W}_{\omega,0}/(n-k)$, where $\mathbf{W}_{\omega,0}$ is the residual sum of squares matrix in the full model with $\theta = 0$, i.e., $\mathbf{W}_{\omega,0} = \mathbf{Y}' (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_\omega}) \mathbf{Y}$. By replacing Σ_* in (6) with \mathbf{S} , a naive estimator of the risk function can be defined, i.e., the following C_p criterion:

$$C_p(j, \theta) = \text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1}) + 2p \text{tr}(\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0}). \quad (7)$$

$C_p(j, 0)$ coincides with the multivariate C_p in [13] which is the information criterion for selecting variables in a multivariate linear regression model. Unfortunately, $C_p(j, \theta)$ has a constant bias for $R(j, \theta)$. It is not negligible when the sample size is small or the number of explanatory variables is large. Hence we try to remove this bias completely; i.e., our goal is to derive an unbiased estimator of $E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1})]$.

It follows from (5) that

$$\mathbf{W}_{j,0} = \mathbf{Y}' (\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j') \mathbf{Y} = \mathbf{W}_{\omega,0} + \hat{\mathbf{\Sigma}}_{\omega,0}' \mathbf{M}_{\omega,0} \hat{\mathbf{\Sigma}}_{\omega,0} - \hat{\mathbf{\Sigma}}_{j,0}' \mathbf{M}_{j,0} \hat{\mathbf{\Sigma}}_{j,0}.$$

Therefore, it is easy to obtain an unbiased estimator of $E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1})]$ when $\theta = 0$, because $\hat{\mathbf{\Sigma}}_{j,0}$ and \mathbf{S} are independent, and $\hat{\mathbf{\Sigma}}_{\omega,0}$ and \mathbf{S} are also independent. However, when $\theta \neq 0$, the equation above cannot be used. Thus, we have to develop an alternative plan to obtain an unbiased estimator of $E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1})]$.

By using the formula in [10], $\mathbf{W}_{j,\theta}$ can be rewritten as

$$\mathbf{W}_{j,\theta} = \mathbf{W}_{j,0} + \theta^2 \mathbf{Y}' \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0} \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j' \mathbf{Y}. \quad (8)$$

Notice that $\mathbf{W}_{j,0}$ can be expressed as $\mathbf{W}_{\omega,0} + \mathbf{Y}' (\mathbf{P}_{\mathbf{X}_\omega} - \mathbf{P}_{\mathbf{X}_j}) \mathbf{Y}$. From this, $\mathbf{W}_{j,\theta}$ in (8) can also be decomposed as

$$\mathbf{W}_{j,\theta} = \mathbf{W}_{\omega,0} + \mathbf{Y}' (\mathbf{P}_{\mathbf{X}_\omega} - \mathbf{P}_{\mathbf{X}_j}) \mathbf{Y} + \theta^2 \mathbf{Y}' \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0} \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j' \mathbf{Y}. \quad (9)$$

From this decomposition, it follows that $\mathbf{W}_{j,\theta} - \mathbf{W}_{\omega,0}$ and \mathbf{S} are independent, because $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_\omega})(\mathbf{P}_{\mathbf{X}_\omega} - \mathbf{P}_{\mathbf{X}_j}) = \mathbf{O}_{n,n}$ and $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_\omega}) \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0} \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j' = \mathbf{O}_{n,n}$ are satisfied. Using these independence results, we have

$$\begin{aligned} E_Y^* [\text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1})] &= (n-k) E_Y^* \left[\text{tr} \left\{ (\mathbf{W}_{j,\theta} - \mathbf{W}_{\omega,0}) \mathbf{W}_{\omega,0}^{-1} + \mathbf{I}_p \right\} \right] \\ &= (n-k) \left\{ E_Y^* \left[\text{tr} \left\{ (\mathbf{W}_{j,\theta} - \mathbf{W}_{\omega,0}) \mathbf{\Lambda} \right\} \right] + p \right\} \\ &= (n-k) \left\{ E_Y^* \left[\text{tr}(\mathbf{W}_{j,\theta} \mathbf{\Lambda}) \right] - E_Y^* \left[\text{tr}(\mathbf{W}_{\omega,0} \mathbf{\Lambda}) \right] + p \right\}, \end{aligned} \quad (10)$$

where $\mathbf{\Lambda} = E_Y^*[\mathbf{W}_{\omega,0}^{-1}]$. Since $\mathbf{W}_{\omega,0} \sim W_p(n-k, \Sigma_*)$, we can see that $E_Y^*[\mathbf{W}_{\omega,0}] = (n-k)\Sigma_*$ and $E_Y^*[\mathbf{W}_{\omega,0}^{-1}] = \Sigma_*^{-1}/(n-k-p-1)(n-k > p+1)$ (see e.g., [14, p. 74, Theorem 2.4.6]). Substituting the two expectations into (10) yields

$$E_Y^* [\text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1})] = \left(1 - \frac{p+1}{n-k} \right)^{-1} \left\{ E_Y^* [\text{tr}(\mathbf{W}_{j,\theta} \Sigma_*^{-1})] - p(p+1) \right\}. \quad (11)$$

It follows immediately from the Eq. (11) that an unbiased estimator of $E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \boldsymbol{\Sigma}_*^{-1})]$ can be defined by $\{1 - (p+1)/(n-k)\} \text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1}) + p(p+1)$. Then, when $n-k > p+1$ holds, we propose the following unbiased estimator of $R(j, \theta)$, which is the modified C_p criterion:

$$MC_p(j, \theta) = \left(1 - \frac{p+1}{n-k}\right) \text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1}) + 2p \text{tr}(\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0}) + p(p+1). \quad (12)$$

Notice that $MC_p(j, 0)$ coincides with the modified C_p criterion in [9], which is the information criterion for selecting variables in a multivariate linear regression model. Hence, it can be seen that our MC_p is an extended version of Fujikoshi and Satoh's modified C_p .

3. Several mathematical properties

In this section, we investigate several mathematical properties of the MC_p and C_p criteria. Let $g(j, \theta)$ be a function of j and θ defined by

$$g(j, \theta) = \text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1}). \quad (13)$$

By using $g(j, \theta)$ and $C_p(j, \theta)$ in (7), $MC_p(j, \theta)$ in (12) can be rewritten as

$$MC_p(j, \theta) = C_p(j, \theta) - (1-a) \{g(j, \theta) - (n-k)p\}, \quad (14)$$

where the coefficient a is defined as

$$a = 1 - \frac{p+1}{n-k}. \quad (15)$$

Notice that the inequality $0 < a < 1$ is satisfied, because $n-k > p+1$ is true. Thus, the relation $0 < 1-a < 1$ also holds. By substituting this inequality and (A.3) in Appendix A.1 into (14), we obtain the following relation between MC_p and C_p :

Theorem 1. For any distribution of \mathbf{Y} , the following inequality is always satisfied:

$$MC_p(j, \theta) \leq C_p(j, \theta),$$

with equality if and only if $\theta = 0$ and $j = \omega$.

Theorem 1 shows that MC_p is always smaller than C_p except for the full model with $\theta = 0$. In particular, when the candidate model is the full model with $\theta = 0$, MC_p coincides with C_p , i.e., $MC_p(\omega, 0) = C_p(\omega, 0) = (n+k)p$.

Recall that the MC_p criterion is an unbiased estimator of the risk function in (6). This unbiasedness and Theorem 1 lead us to another relation between the MC_p and C_p criteria.

Theorem 2. When the distribution of \mathbf{Y} is normal and the assumption $\mathbf{P}_{X_\omega} \boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_*$ is satisfied, the following inequality holds:

$$E_Y^*[MC_p(j, \theta)] = R(j, \theta) \leq E_Y^*[C_p(j, \theta)],$$

with equality if and only if $\theta = 0$ and $j = \omega$.

Theorem 2 shows that $C_p(j, \theta)$ overestimates $R(j, \theta)$ except in the case of the full model, with $\theta = 0$.

Theorem 2 describes the biases of the criteria. However, since an information criterion is an estimator of the risk function, not only bias but also variance is an important consideration. Therefore, we now consider the variances of the MC_p and C_p criteria. Let $h(j, \theta)$ be a function of j and θ defined by

$$h(j, \theta) = \text{tr}(\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0}). \quad (16)$$

Using $h(j, \theta)$ and $C_p(j, \theta)$, we can rewrite $MC_p(j, \theta)$ as

$$MC_p(j, \theta) = aC_p(j, \theta) + 2p(1-a)h(j, \theta) + p(p+1), \quad (17)$$

where a is given by (15). Since p , a and $h(j, \theta)$ are non-stochastic, it seems that variances of MC_p and C_p criteria are related by

$$\text{Var}[MC_p(j, \theta)] = a^2 \text{Var}[C_p(j, \theta)].$$

Let us recall that $0 < a < 1$ and $MC_p(\omega, 0) = C_p(\omega, 0)$. Consequently, we can derive the following theorem.

Theorem 3. For any distribution of \mathbf{Y} , the following inequality is always satisfied:

$$\text{Var}[MC_p(j, \theta)] \leq \text{Var}[C_p(j, \theta)], \quad (18)$$

with equality if and only if $\theta = 0$ and $j = \omega$.

Theorem 3 yields the surprising result that MC_p not only removes the bias of C_p but also reduces its variance. Furthermore, the inequality (18) holds even if the distribution of \mathbf{Y} is not normal. In general, the variance of the bias-corrected estimator is larger than that of the original estimator (see e.g., [15, p. 138]). Even though MC_p is a bias-corrected C_p , the variance of our MC_p is smaller than that of C_p except in the case that $\theta = 0$ and $j = \omega$. It can be seen that our MC_p criterion therefore has a very desirable property.

As for the variance of the MC_p criterion, we can obtain a stronger result than **Theorem 3**. In the univariate linear regression model, Davies, Neath and Cavanaugh [12] showed that $MC_p(j, 0)$ is the UMVUE of $R(j, 0)$. We can extend their result to the multivariate ridge regression. By slightly modifying the result in [14, pp. 18–20], it is easily seen that $\hat{\Sigma}_{\omega,0}$ and $\mathbf{W}_{\omega,0}$ are complete sufficient statistics when the true model is (3). Notice that $\mathbf{P}_{\mathbf{X}_\omega} \mathbf{Y} = \mathbf{X}_\omega \hat{\Sigma}_{\omega,0}^{-1} \mathbf{X}_\omega' \mathbf{Y}$ and $\mathbf{P}_{\mathbf{X}_\omega} \mathbf{X}_j = \mathbf{X}_j$. It follows from Eq. (9) that

$$MC_p(j, \theta) = a(n-k) \text{tr} \left\{ (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_j} + \theta^2 \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0} \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j') \mathbf{X}_\omega \hat{\Sigma}_{\omega,0}^{-1} \mathbf{W}_{\omega,0}^{-1} \hat{\Sigma}_{\omega,0}' \mathbf{X}_\omega' \right\} \\ + p \left\{ a(n-k) + 2 \text{tr}(\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0}) + (p+1) \right\}.$$

This means that $MC_p(j, \theta)$ is a function of complete sufficient statistics. From the Lehman–Scheffé theorem, we obtain the following result.

Theorem 4. When the distribution of \mathbf{Y} is normal and the assumption $\mathbf{P}_{\mathbf{X}_\omega} \mathbf{\Gamma}_* = \mathbf{\Gamma}_*$ is satisfied, $MC_p(j, \theta)$ is the UMVUE of $R(j, \theta)$.

Previous theorems have described characteristics of our criterion as an estimator of the risk function. On the other hand, in model selection, it is also important which model is chosen using an information criterion. In particular, since we are correcting the bias in the criterion, we need to investigate changes in the selected ridge parameter and/or the selected subset of ω due to this correction of the bias. Firstly, we consider θ_j chosen by minimizing $R(j, \theta)$, $MC_p(j, \theta)$ and $C_p(j, \theta)$ for a fixed j .

Let $\dot{R}(j, \theta)$, $\dot{MC}_p(j, \theta)$ and $\dot{C}_p(j, \theta)$ be the first derivatives with respect to θ of $R(j, \theta)$ and $MC_p(j, \theta)$ and $C_p(j, \theta)$, respectively. Notice that the matrix $\mathbf{M}_{j,0}$ is positive definite. Hence, from Eqs. (A.5)–(A.7) in Appendix A.2, we have

$$\dot{R}(j, 0) = \dot{MC}_p(j, 0) = \dot{C}_p(j, 0) = -2 \text{tr}(\mathbf{M}_{j,0}^{-1}) < 0. \quad (19)$$

This equation means that $R(j, \theta)$, $MC_p(j, \theta)$ and $C_p(j, \theta)$ are decreasing when $\theta = 0$. Let $\theta_j^{(r)}$, $\hat{\theta}_j^{(m)}$ and $\hat{\theta}_j^{(c)}$ be ridge parameters minimizing $R(j, \theta)$, $MC_p(j, \theta)$ and $C_p(j, \theta)$, respectively, for a fixed j , i.e.,

$$R(j, \theta_j^{(r)}) = \min_{\theta \geq 0} R(j, \theta),$$

and

$$MC_p(j, \hat{\theta}_j^{(m)}) = \min_{\theta \geq 0} MC_p(j, \theta), \quad C_p(j, \hat{\theta}_j^{(c)}) = \min_{\theta \geq 0} C_p(j, \theta). \quad (20)$$

Then Eq. (19) implies the following theorem.

Theorem 5. For any distribution of \mathbf{Y} and combinations of \mathbf{X} , the inequalities $\theta_j^{(r)} > 0$, $\hat{\theta}_j^{(m)} > 0$ and $\hat{\theta}_j^{(c)} > 0$ are always satisfied.

Theorem 5 shows that the principal optimal θ obtained by minimizing $R(j, \theta)$ is not zero. Hence, it remains possible that the mixed procedure with the ridge regression and variable selection makes the risk function smaller than that of the single procedure. Moreover, the inequalities $\hat{\theta}_j^{(m)} > 0$ and $\hat{\theta}_j^{(c)} > 0$ show that the transformation $\theta = e^\lambda$ ($\lambda \in \mathbb{R}$) can be used to search for θ minimizing $MC_p(j, \theta)$ and $C_p(j, \theta)$. It facilitates an optimization for the ridge regression, because the search area is stretched into the whole set \mathbb{R} .

Next, we consider a magnitude relation between $\hat{\theta}_j^{(m)}$ and $\hat{\theta}_j^{(c)}$. Suppose that the inequality $\hat{\theta}_j^{(m)} < \hat{\theta}_j^{(c)}$ holds. Then, from (A.8) in the Appendix A.3, we have $h(j, \hat{\theta}_j^{(c)}) < h(j, \hat{\theta}_j^{(m)})$. Moreover, by applying (20), the relations $C_p(j, \hat{\theta}_j^{(c)}) \leq C_p(j, \hat{\theta}_j^{(m)})$ can be derived. Substituting the two inequalities into (17) yields

$$MC_p(j, \hat{\theta}_j^{(m)}) = aC_p(j, \hat{\theta}_j^{(m)}) + 2p(1-a)h(j, \hat{\theta}_j^{(m)}) + p(p+1) \\ > aC_p(j, \hat{\theta}_j^{(c)}) + 2p(1-a)h(j, \hat{\theta}_j^{(c)}) + p(p+1) \\ = MC_p(j, \hat{\theta}_j^{(c)}),$$

because $0 < a < 1$ and $0 < 1-a < 1$ are satisfied. However, this result is contradictory to the result that $MC_p(j, \hat{\theta}_j^{(m)}) \leq MC_p(j, \theta)$ for any θ . Consequently, by reductio ad absurdum, we obtain the following theorem, which characterizes the relation between the two ridge parameters determined by the C_p and MC_p criteria.

Theorem 6. For any distribution of \mathbf{Y} and combinations of \mathbf{X} , the inequality $\hat{\theta}_j^{(m)} \geq \hat{\theta}_j^{(c)}$ is always satisfied.

Theorem 6 shows that the optimal θ obtained using MC_p is not smaller than that determined from C_p for fixed j . In general, the best model obtained by minimizing the existing C_p criterion tends to overfit the principal best model. Many studies have verified this characteristic by conducting numerical simulations, e.g., [9,16,17]. For ridge regression, overfitting means choosing a smaller value of θ than the principal best θ . The $MC_p(j, \theta)$ has improved this weak point by correcting the bias.

Theorem 6 gives the relation between the two ridge parameters resulting from the MC_p and C_p criteria. From the same idea used to prove Theorem 6, we can also obtain other inequalities between the best models, as theorems resulting from using the MC_p and C_p criteria. (We present the proofs in Appendix A.4, because they are very similar to the proof of Theorem 6).

Theorem 7. Let $\hat{j}_\theta^{(m)}$ and $\hat{j}_\theta^{(c)}$ be subsets of ω obtained by minimizing $MC_p(j, \theta)$ and $C_p(j, \theta)$ respectively, for a fixed θ . Then, the relation $\hat{j}_\theta^{(c)} \not\subset \hat{j}_\theta^{(m)}$ is always satisfied for any distributions of \mathbf{Y} and ridge parameters. In particular, for a nested model, $\hat{j}_\theta^{(m)} \subseteq \hat{j}_\theta^{(c)}$ holds.

Theorem 8. Let $\hat{j}^{(m)}$ and $\hat{\theta}^{(m)}$ be j and θ obtained by minimizing $MC_p(j, \theta)$, and let $\hat{j}^{(c)}$ and $\hat{\theta}^{(c)}$ be j and θ obtained minimizing $C_p(j, \theta)$. Then, the inequality $\hat{\theta}^{(m)} \geq \hat{\theta}^{(c)}$ or the relation $\hat{j}^{(c)} \not\subset \hat{j}^{(m)}$ are always satisfied for any distributions of \mathbf{Y} .

4. Numerical study

We evaluate the proposed criterion applied numerically to the polynomial regression model, $\mathbf{Y} \sim N_{n \times p}(\mathbf{\Gamma}_*, \mathbf{\Sigma}_* \otimes \mathbf{I}_p)$, with $p = 2, n = 20, k = 12$ and $\omega = \{1, \dots, 12\}$ where $\mathbf{\Gamma}_* = \mathbf{X}_\omega \mathbf{\Xi}_*$,

$$\mathbf{\Xi}_* = \delta \begin{pmatrix} 1 & 2 & 3 & 0 & \cdots & 0 \\ 1 & 4 & 9 & 0 & \cdots & 0 \end{pmatrix}', \quad \mathbf{\Sigma}_* = \begin{pmatrix} 1 & 0.5^2 \\ 0.5^2 & 1 \end{pmatrix},$$

$$\mathbf{X}_\omega = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{pmatrix} z_1 & z_1^2 & \cdots & z_1^k \\ \vdots & \vdots & \vdots & \vdots \\ z_n & z_n^2 & \cdots & z_n^k \end{pmatrix}.$$

Each column vector of the design matrix \mathbf{X}_ω is obtained by standardization of the corresponding column vector of \mathbf{Z} . The first column vector of \mathbf{Z} is generated from the independent uniform distribution on $(-1, 1)$. Notice that the candidate models are nested and that \mathbf{X}_j is the submatrix consisting of the first j columns of \mathbf{X}_ω . In a sense, the subindex j is the degree of a polynomial here.

Since MC_p is derived as an estimator of the MSE of prediction, we compare the related four criteria: MC_p , C_p , the cross-validation (CV) criterion [18] and the generalized cross-validation (GCV) criterion [19], with respect to the following three points:

- (i) the probabilities or frequencies of selected models,
- (ii) the expectation value of the selected ridge parameter,
- (iii) the MSE of prediction.

Here, CV and GCV criteria can be formally defined by

$$CV(j, \theta) = \text{tr}\{(\mathbf{I}_n - \mathbf{H}_{j,\theta})^{-2}(\mathbf{Y} - \hat{\mathbf{Y}}_{j,\theta})\mathbf{S}^{-1}(\mathbf{Y} - \hat{\mathbf{Y}}_{j,\theta})'\}, \quad (21)$$

$$GCV(j, \theta) = \frac{\text{tr}(\mathbf{W}_{j,\theta}\mathbf{S}^{-1})}{\{1 - \text{tr}(\mathbf{M}_{j,\theta}^{-1}\mathbf{M}_{j,0})/n\}^2}, \quad (22)$$

where $\mathbf{H}_{j,\theta}$ is an $n \times n$ diagonal matrix whose i th diagonal element corresponds to the (i, i) th element of $\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j'$. We call the CV and GCV criteria defined by (21) and (22) the formal CV and GCV. Additionally, it turns out from (A.18) in Appendix A.5 that the formal GCV criterion has larger variances than C_p and MC_p criteria except for the case of $\theta = \infty$. We selected both the candidate model and the ridge parameter, and calculated the MSE of the prediction as $np + E_{\mathbf{Y}}[\text{tr}\{(\mathbf{\Gamma}_* - \hat{\mathbf{Y}}_{j,\hat{\theta}})\mathbf{\Sigma}_*^{-1}(\mathbf{\Gamma}_* - \hat{\mathbf{Y}}_{j,\hat{\theta}})'\}]$. Moreover, we consider the following three optimization procedures:

- (a) a mixed procedure with ridge regression and variable selection (the proposed procedure); we optimize the subset j and the ridge parameter θ simultaneously,
- (b) the procedure with only variable selection (the ordinary procedure); we optimize only the subset j under $\theta = 0$ ($\hat{\theta}$ is always 0),
- (c) the procedure with only ridge regression of the full model (the ordinary procedure); we optimize only the ridge parameter θ under $j = \omega$ (\hat{j} is always ω).

Table 1

The frequencies of selected models, the expectation value of selected ridge parameters, the MSE of prediction for 1,000 repetitions under the true model with $\delta = 0$, or a constant model. A parenthetic value shows the result without ridge parameter. MSE_ω shows the MSE by using full model without variable selection.

j	MC_p			C_p			CV			GCV		
	freq.	$E[\hat{\theta}_j]$	(freq.)	freq.	$E[\hat{\theta}_j]$	(freq.)	freq.	$E[\hat{\theta}_j]$	(freq.)	freq.	$E[\hat{\theta}_j]$	(freq.)
1	639	82.9	(779)	430	70.7	(548)	513	75.7	(689)	496	76.4	(645)
2	130	83.7	(76)	140	68.0	(101)	152	74.5	(135)	152	75.3	(117)
3	25	84.1	(27)	38	67.1	(49)	48	74.9	(65)	42	76.8	(55)
4	24	84.9	(27)	26	66.6	(44)	40	76.2	(47)	37	77.6	(46)
5	13	84.7	(11)	24	66.4	(27)	29	77.4	(23)	20	77.6	(24)
6	17	85.0	(6)	34	66.6	(21)	24	78.3	(8)	27	78.0	(14)
7	15	85.0	(11)	32	65.8	(31)	14	77.9	(7)	21	77.7	(23)
8	14	85.6	(9)	36	65.0	(23)	11	78.6	(6)	26	78.2	(10)
9	12	85.4	(6)	22	64.6	(18)	15	78.3	(3)	16	78.3	(7)
10	14	85.3	(9)	39	64.1	(31)	21	78.4	(4)	23	78.6	(15)
11	21	85.1	(15)	46	64.4	(38)	23	78.3	(4)	30	77.9	(18)
12	76	85.3	(24)	133	65.0	(69)	110	78.0	(9)	110	78.1	(26)
MSE		42.9	(45.0)		46.6	(49.0)		42.9	(44.9)		44.1	(46.5)
MSE _ω		42.9			46.3			42.3			43.4	

Table 2

The frequencies of selected models, the expectation value of selected ridge parameters, the MSE of prediction for 1,000 repetitions under the true model with $\delta = 2.0$, or a third degree polynomial model. A parenthetic value shows the result without ridge parameter. MSE_ω shows the MSE by using full model without variable selection.

j	MC_p			C_p			CV			GCV		
	freq.	$E[\hat{\theta}_j]$	(freq.)	freq.	$E[\hat{\theta}_j]$	(freq.)	freq.	$E[\hat{\theta}_j]$	(freq.)	freq.	$E[\hat{\theta}_j]$	(freq.)
1	0	0.01	(0)	0	0.01	(0)	0	0.00	(0)	0	0.15	(0)
2	0	0.02	(0)	0	0.01	(0)	0	0.00	(0)	0	0.21	(0)
3	757	0.01	(779)	520	0.01	(554)	614	0.02	(693)	612	0.01	(647)
4	80	0.04	(87)	98	0.02	(100)	153	0.03	(161)	109	0.03	(119)
5	36	0.06	(27)	56	0.03	(50)	85	0.08	(65)	58	0.03	(52)
6	18	0.07	(11)	48	0.03	(45)	50	0.12	(29)	45	0.04	(35)
7	15	0.05	(20)	32	0.02	(36)	27	0.09	(15)	25	0.03	(35)
8	15	0.06	(18)	31	0.03	(31)	14	0.13	(8)	24	0.03	(22)
9	9	0.06	(6)	28	0.03	(24)	10	0.12	(5)	18	0.03	(14)
10	13	0.06	(10)	38	0.03	(39)	6	0.15	(4)	24	0.03	(21)
11	23	0.07	(15)	57	0.03	(42)	12	0.16	(8)	36	0.04	(20)
12	34	0.07	(27)	92	0.03	(79)	29	0.17	(12)	49	0.03	(35)
MSE		48.8	(48.6)		52.0	(52.1)		48.9	(48.3)		50.3	(50.1)
MSE _ω		54.2			55.6			56.5			54.0	

In order to compare procedure (a) with (b) and (c), we calculated properties (i) and (iii) derived from the procedure (b) and a property (iii) derived from the procedure (c). These properties were evaluated by Monte Carlo simulation with 1,000 iterations under two types of true model, (1) $\delta = 0$, or a constant model, (2) $\delta = 2.0$, or a third degree polynomial model. In the former case, smaller degree polynomial models estimated by larger ridge parameters should be selected; conversely, the third degree polynomial model estimated by smaller ridge parameters should be selected in the latter case.

Tables 1 and 2 show the three properties in the cases of $\delta = 0$ and $\delta = 2$, respectively. Values in parentheses in Tables 1 and 2 denote the properties obtained from the procedure (b). In both tables, MSE_ω shows the MSE of prediction obtained from the procedure (c). As a result of the simulation study, our MC_p criterion was much improved, compared to the original Mallows' C_p criterion in the sense of the MSE of prediction. Although the MSE of the MC_p criterion was almost the same as that of the CV criteria, MC_p selected preferable candidate models more often than CV in both of the cases (1) when larger ridge parameters were required, and (2) when ridge parameters were not as necessary, or the usual least square estimator without ridge parameters was sufficient. The performance of the GCV criterion might thus be located between that of the CV and C_p criteria. Therefore we conclude that the MC_p criterion is the best criterion among those four criteria in the sense of MSE prediction and the probability of selecting the preferable candidate model for a ridge regression.

Comparing procedures (a) with (b), we can see that the procedure (b) selected the true model with higher frequency than the procedure (a). However, the MSE of prediction obtained from the procedure (a) was smaller than that obtained from the procedure (b) when $\delta = 0$. Even though the MSE obtained from the procedure (a) was larger than that obtained from the procedure (b) when δ was 2.0, the difference was very small. On the other hand, the MSE obtained from the procedure (a) was smaller than MSE_ω when $\delta = 2.0$. Even though the MSE obtained from the procedure (a) was larger than MSE_ω when $\delta = 0$, its difference was very small. Ultimately, these results show that the mixed procedure using ridge regression and variable selection improved on the individual procedures, while still retaining most of their advantages.

Table 3

Estimated regression coefficients of selected best models.

Variable	Optimized θ				Without θ			
	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
With variable selection								
X_1	−0.30	−0.24	−0.01	0.36	−0.33	−0.27	−0.02	0.40
X_2	−0.61	−0.45	0.00	0.63	−0.67	−0.49	−0.02	0.71
X_3	−0.29	−0.09	−0.41	0.52	−0.35	−0.12	−0.42	0.58
X_4	–	–	–	–	–	–	–	–
X_5	−0.37	−0.29	0.59	−0.19	−0.40	−0.31	0.62	−0.19
X_6	–	–	–	–	–	–	–	–
X_7	–	–	–	–	–	–	–	–
Without variable selection (full model)								
X_1	−0.21	−0.13	0.02	0.18	−0.43	−0.14	−0.02	0.28
X_2	−0.28	−0.22	0.11	0.16	−0.69	−0.23	0.00	0.35
X_3	−0.17	0.12	−0.34	0.24	−0.55	0.15	−0.43	0.38
X_4	0.11	0.24	0.09	−0.34	−0.21	0.31	0.00	−0.23
X_5	−0.33	−0.28	0.55	−0.16	−0.37	−0.31	0.62	−0.18
X_6	0.24	−0.05	0.00	−0.12	0.27	−0.11	0.03	−0.12
X_7	−0.15	−0.09	−0.02	0.11	−0.09	−0.11	−0.02	0.12

Next, an illustrative application to an actual data set is shown. We used the Scottish election data [20], which are the results of voting in 71 ($= n$) Scottish constituencies in the two British general elections held in February and October 1974. Variables headed W_1 , W_2 , W_3 and W_4 denote the Conservative, Labour, Liberal, and Nationalist parties in October 1974 as do those headed C , S , L and N which apply to February 1974 (in the same order). Variable E is an electorate figure (February and October figures differed insignificantly) and R is a categorical variable defining the region where

- 1 = Glasgow; 2 = Remainder of the Clydeside conurbation;
 3 = Edinburgh; 4 = Remainder of the industrial centers;
 5 = Highlands; 6 = Rest of Scotland.

According to [20], four response variables Y_1 , Y_2 , Y_3 and Y_4 ($p = 4$), and seven explanatory variables X_1 , X_2 , X_3 , X_4 , X_5 , X_6 and X_7 ($k = 7$) were calculated from the raw data as follows: Response variables were generated from $Y_j = W_j/E - X_j$ ($j = 1, 2, 3, 4$), where X_1 , X_2 , X_3 and X_4 were

$$X_1 = C/E, \quad X_2 = S/E, \quad X_3 = L/E, \quad X_4 = N/E.$$

The remaining explanatory variables were figured as

$$X_5 = \begin{cases} 0.5 & : \text{Liberal intervenes, i.e., } W_3 > 0 \text{ and } L = 0 \\ 0 & : \text{otherwise,} \end{cases}$$

$$X_6 = \begin{cases} 0.5 & : R = 5, 6 \\ 0 & : \text{otherwise,} \end{cases}$$

$$X_7 = \begin{cases} 0.5 & : \text{Labour or Nationalist top parties in February 1974 and } |X_2 - X_4| \leq 0.2 \\ 0 & : \text{otherwise.} \end{cases}$$

Table 3 shows the estimates of regression coefficients for the following four models:

- the model with the best ridge parameter and subset of variables chosen by MC_p criterion (we optimized the subset j and ridge parameter θ simultaneously by $MC_p(j, \theta)$),
- the model with the best subset of variables chosen by MC_p criterion (we optimized only the subset j by $MC_p(j, 0)$),
- the full model with the best ridge parameter chosen by MC_p criterion (we optimized only the ridge parameter θ in the full model by $MC_p(\omega, \theta)$),
- the full model without θ (LSE of the full model).

In both cases (a) and (b), the model having variables X_1 , X_2 , X_3 and X_5 was chosen as the best subset. From the results, we can see that ridge regression without variable selection shirked the LSE of the full model too much. Although ridge regression with variable selection did not shrink the LSE of the selected model as much, we can find a clear difference between estimated regression coefficients of the models (a) and (b).

5. Conclusion and discussion

In this paper, we have proposed an unbiased C_p criterion, denoted as MC_p . The MC_p criterion is the UMVUE of the risk function based on the expected standardized MSE of prediction when the distribution of \mathbf{Y} is normal and $\mathbf{P}_{\mathbf{X}_\omega} \mathbf{\Gamma}_* = \mathbf{\Gamma}_*$.

is satisfied. One of the advantages of the MC_p criterion is that its definition is very simple. Furthermore, we have proved analytically that the MC_p criterion has smaller variance than the C_p criterion. In addition, the optimal ridge parameter obtained using MC_p is always at least as large as that resulting from C_p for fixed j , and the best subset of ω obtained by using MC_p is not included in the best subset obtained from use of C_p . In numerical studies, we demonstrated that the MC_p criterion is more effective than the C_p criterion and the formal CV and GCV criteria.

From the asymptotic expansion of the bias in [16], it seems that the effect of non-normality on the bias of $MC_p(j, 0)$ is very small; its order is merely $O(n^{-2})$ even under non-normality, when $\mathbf{P}_{\mathbf{X}_j} \mathbf{\Gamma}_* = \mathbf{\Gamma}_*$ is satisfied. Hence, we can expect that our $MC_p(j, \theta)$ also has similar good properties.

In the case of a single response, the risk function does not need to be standardized by the true variance–covariance matrix, and in this situation, an unbiased estimator is easy to obtain. This unbiased estimator may almost be equivalent to the criterion proposed in [21] and [22], among other studies. However, in the case of multiple responses, the standardization results in difficulty deriving an unbiased estimator. For this case, there has hitherto been no unbiased estimator of the risk function based on the expected standardized MSE of prediction for multivariate ridge regression. On the other hand, we can guess that an estimator of the risk function may be able to be derived easily by use of the CV method as in (21). However, in the case of multiple responses, an estimated variance–covariance matrix for the standardization should also be constructed by means of the jackknife method, as well as by using the predictor of \mathbf{Y} . Then, the GCV criterion cannot be strictly defined and unfortunately the CV criterion will have constant bias (see [17]). In addition, although we can define the formal GCV criterion as in (22), this has larger variance than MC_p criterion, except in the case of $\theta = \infty$. Therefore, for model selection based on a multivariate ridge regression, MC_p will not be supplanted by other criteria at present.

There have been many studies concerned with correction of the bias of an information criterion. However, in almost all cases the resulting papers have reported only on the bias correction and have not presented any theoretical results on the differences between models selected using the original criterion and the improved version. In contrast, this paper does consider changes in the selected model due to correcting the bias.

From the many viewpoints mentioned above, we consider that the results in our paper are useful, and thus we can recommend use of the MC_p criterion instead of the C_p criterion for multivariate ridge regression.

Choosing the best ridge parameter and subset of variables simultaneously may be able to reduce the value of the MSE of prediction to a greater extent than choosing only the ridge parameter or subset of variables. However, many computational tasks are involved when we search for the best ridge parameter and subset of variables at the same time. Hence, we have to say that this is not a realistic method for data with many explanatory variables. When there are many explanatory variables, we recommend reducing the number of candidate models by the variable selection procedure first, and later choosing the best ridge parameter and subset of variables among the reduced number of candidate models.

Acknowledgments

The authors thank Professor H. Wakaki, Hiroshima University for the useful advice on the UMVUE of the risk function. Moreover, the authors also thank the associate editor and the two reviewers for their valuable comments.

The first author was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007–2010. The second author was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #21700306, 2009–2011.

Appendix

A.1. Properties of the function $g(j, \theta)$

Let $\mathbf{Q}_j = (\mathbf{q}_{j,1}, \dots, \mathbf{q}_{j,k_j})$ be a $k_j \times k_j$ orthogonal matrix such that

$$\mathbf{Q}_j' \mathbf{M}_{j,0} \mathbf{Q}_j = \mathbf{D}_j = \text{diag}(d_{j,1}, \dots, d_{j,k_j}), \quad (\text{A.1})$$

where $d_{j,\alpha}$ ($\alpha = 1, \dots, k_j$) are the eigenvalues of $\mathbf{M}_{j,0}$, and let $\mathbf{z}_{j,1}, \dots, \mathbf{z}_{j,k_j}$ be n -dimensional vectors such that $(\mathbf{z}_{j,1}, \dots, \mathbf{z}_{j,k_j})' = \mathbf{Q}_j' \mathbf{X}_j' \mathbf{Y} \mathbf{S}^{-1/2}$. By using $\mathbf{z}_{j,\alpha}$ and $d_{j,\alpha}$ ($\alpha = 1, \dots, k_j$), we can write $g(j, \theta)$ in (13) as

$$g(j, \theta) = \text{tr}(\mathbf{Y}' \mathbf{Y} \mathbf{S}^{-1}) - 2 \sum_{\alpha=1}^{k_j} \frac{\|\mathbf{z}_{j,\alpha}\|^2}{d_{j,\alpha} + \theta} + \sum_{\alpha=1}^{k_j} \frac{\|\mathbf{z}_{j,\alpha}\|^2 d_{j,\alpha}}{(d_{j,\alpha} + \theta)^2}.$$

Since $d_{j,\alpha} > 0$ and $\theta \geq 0$ hold, we have

$$\frac{\partial}{\partial \theta} g(j, \theta) = 2\theta \sum_{\alpha=1}^{k_j} \frac{\|\mathbf{z}_{j,\alpha}\|^2}{(d_{j,\alpha} + \theta)^3} \geq 0, \quad (\text{A.2})$$

with equality if and only if $\theta = 0$ or $\theta = \infty$. Therefore, we can see that $g(j, \theta)$ is a strictly monotonic increasing function of $\theta \in [0, \infty]$. This result implies that $g(j, \theta) \geq g(j, 0)$ with equality if and only if $\theta = 0$. Notice that $\mathbf{P}_{\mathbf{X}_\omega} - \mathbf{P}_{\mathbf{X}_j}$ is positive definite except when $j = \omega$. Therefore, we obtain

$$\begin{aligned} g(j, 0) &= \text{tr}(\mathbf{W}_{j,0} \mathbf{S}^{-1}) \\ &= (n-k)p + \text{tr}\{\mathbf{Y}'(\mathbf{P}_{\mathbf{X}_\omega} - \mathbf{P}_{\mathbf{X}_j})\mathbf{Y}\mathbf{S}^{-1}\} \geq (n-k)p = g(\omega, 0), \end{aligned}$$

with equality if and only if $j = \omega$. Results obtained in this subsection are summarized in the following theorem.

Theorem A.1. *The function $g(j, \theta)$ is a strictly monotonic increasing function of $\theta \in [0, \infty]$ for fixed j , and has the following lower bound:*

$$g(j, \theta) \geq (n-k)p, \quad (\text{A.3})$$

with equality if and only if $\theta = 0$ and $j = \omega$.

A.2. First derivatives of $R(j, \theta)$, $C_p(j, \theta)$ and $MC_p(j, \theta)$

Notice that the function $h(j, \theta)$ in (16) can be rewritten as

$$h(j, \theta) = \sum_{\alpha=1}^{k_j} \frac{d_{j,\alpha}}{d_{j,\alpha} + \theta}, \quad (\text{A.4})$$

where $d_{j,\alpha}$ ($\alpha = 1, \dots, k_j$) are the eigenvalues of $\mathbf{M}_{j,0}$, which are defined by (A.1). Recall that $C_p(j, \theta) = g(j, \theta) + 2ph(j, \theta)$, where the functions $g(j, \theta)$ are given by (13). From Eqs. (A.2) and (A.4), we obtain $\dot{C}_p(j, \theta) = \partial C_p(j, \theta)/\partial \theta$ as

$$\begin{aligned} \dot{C}_p(j, \theta) &= 2\theta \sum_{\alpha=1}^{k_j} \frac{\|\mathbf{z}_{j,\alpha}\|^2}{(d_{j,\alpha} + \theta)^3} - 2p \sum_{\alpha=1}^{k_j} \frac{d_{j,\alpha}}{(d_{j,\alpha} + \theta)^2} \\ &= 2 \left\{ \theta \text{tr}(\mathbf{X}_j' \mathbf{Y} \mathbf{S}^{-1} \mathbf{Y}' \mathbf{X}_j \mathbf{M}_{j,\theta}^{-3}) - p \text{tr}(\mathbf{M}_{j,\theta}^{-2} \mathbf{M}_{j,0}) \right\}. \end{aligned} \quad (\text{A.5})$$

By using Eqs. (14) and (A.5), $\dot{MC}_p(j, \theta) = \partial MC_p(j, \theta)/\partial \theta$ is derived as

$$\dot{MC}_p(j, \theta) = 2 \left\{ a \theta \text{tr}(\mathbf{X}_j' \mathbf{Y} \mathbf{S}^{-1} \mathbf{Y}' \mathbf{X}_j \mathbf{M}_{j,\theta}^{-3}) - p \text{tr}(\mathbf{M}_{j,\theta}^{-2} \mathbf{M}_{j,0}) \right\}, \quad (\text{A.6})$$

where the coefficient a is given by (15). On the other hand, the expectation $E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \mathbf{\Sigma}_*^{-1})]$ is calculated as

$$\begin{aligned} E_Y^*[\text{tr}(\mathbf{W}_{j,\theta} \mathbf{\Sigma}_*^{-1})] &= \text{tr} \left\{ \mathbf{\Gamma}'_*(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{\Gamma}_* \mathbf{\Sigma}_*^{-1} \right\} + E_Y^* \left\{ \text{tr} \left\{ \mathbf{\mathcal{E}}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{\mathcal{E}} \mathbf{\Sigma}_*^{-1} \right\} \right\} \\ &= \text{tr} \left\{ \mathbf{\Gamma}'_*(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{\Gamma}_* \mathbf{\Sigma}_*^{-1} \right\} + p \text{tr} \left\{ (\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \right\}. \end{aligned}$$

Thus, a detailed expression of $R(j, \theta)$ in (6) is given by

$$R(j, \theta) = \text{tr} \left\{ \mathbf{\Gamma}'_*(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{\Gamma}_* \mathbf{\Sigma}_*^{-1} \right\} + np + p \text{tr} \left\{ (\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0})^2 \right\}.$$

Recall that $g(j, \theta) = \text{tr}\{\mathbf{Y}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{Y} \mathbf{S}^{-1}\}$. Therefore, by replacing \mathbf{Y} and \mathbf{S} with $\mathbf{\Gamma}_*$ and $\mathbf{\Sigma}_*$, we have

$$\frac{\partial}{\partial \theta} \text{tr} \left\{ \mathbf{\Gamma}'_*(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')^2 \mathbf{\Gamma}_* \mathbf{\Sigma}_*^{-1} \right\} = 2\theta \text{tr}(\mathbf{X}_j' \mathbf{\Gamma}_* \mathbf{\Sigma}_*^{-1} \mathbf{\Gamma}_* \mathbf{X}_j \mathbf{M}_{j,\theta}^{-3}).$$

It follows from a similar transformation as in (A.4) that

$$\text{tr} \left\{ (\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0})^2 \right\} = \sum_{\alpha=1}^{k_j} \frac{d_{j,\alpha}^2}{(d_{j,\alpha} + \theta)^2}.$$

This implies that

$$\frac{\partial}{\partial \theta} \text{tr} \left\{ (\mathbf{M}_{j,\theta}^{-1} \mathbf{M}_{j,0})^2 \right\} = -2 \sum_{\alpha=1}^{k_j} \frac{d_{j,\alpha}^2}{(d_{j,\alpha} + \theta)^3} = -2 \text{tr}(\mathbf{M}_{j,\theta}^{-3} \mathbf{M}_{j,0}^2).$$

Hence, $\dot{R}(j, \theta) = \partial R(j, \theta)/\partial \theta$ is obtained as

$$\dot{R}(j, \theta) = 2 \left\{ \theta \text{tr}(\mathbf{X}_j' \mathbf{\Gamma}_* \mathbf{\Sigma}_*^{-1} \mathbf{\Gamma}_* \mathbf{X}_j \mathbf{M}_{j,\theta}^{-3}) - p \text{tr}(\mathbf{M}_{j,\theta}^{-3} \mathbf{M}_{j,0}^2) \right\}. \quad (\text{A.7})$$

A.3. Monotonicity of the function $h(j, \theta)$

From the Eq. (A.4), we have the following theorem.

Theorem A.2. The function $h(j, \theta)$ is a strictly decreasing function of $\theta \in [0, \infty]$ for fixed j . Therefore, we have the following relation:

$$h(j, \theta_2) < h(j, \theta_1), \quad (\text{when } \theta_1 < \theta_2). \quad (\text{A.8})$$

Let $\mathbf{X}_{j_1} = (\mathbf{X}_j \mathbf{x})$ be an $n \times (k_j + 1)$ matrix, where \mathbf{x} is an n -dimensional vector which is linearly independent of any columns of \mathbf{X}_j . From the formula for the inverse matrix (see e.g., [14, p. 592, Theorem A.2.3]), we have

$$\mathbf{M}_{j_1, \theta}^{-1} = \begin{pmatrix} \mathbf{M}_{j, \theta}^{-1} + \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j' \mathbf{x} \mathbf{x}' \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} / c_{j, \theta} & -\mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j' \mathbf{x} / c_{j, \theta} \\ -\mathbf{x}' \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} / c_{j, \theta} & 1 / c_{j, \theta} \end{pmatrix}, \quad (\text{A.9})$$

where $c_{j, \theta} = \theta + \mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j') \mathbf{x}$. Let $b_{j,1}, \dots, b_{j,k_j}$ be such that $(b_{j,1}, \dots, b_{j,k_j})' = \mathbf{P}_j' \mathbf{X}_j' \mathbf{x}$. By using $b_{j,\alpha}$ and $d_{j,\alpha}$ ($\alpha = 1, \dots, k_j$), we can write $c_{j,\theta}$ as

$$c_{j, \theta} = \theta + \mathbf{x}' \mathbf{x} - \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2}{d_{j,\alpha} + \theta}. \quad (\text{A.10})$$

By partially differentiating the Eq. (A.10), we can see that $c_{j,\theta}$ is a monotonic increasing function of θ . This implies that $c_{j,\theta} \geq c_{j,0} = \mathbf{x}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_j}) \mathbf{x}$. Notice that $\mathbf{x}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_j}) \mathbf{x} > 0$ because $\mathbf{P}_{\mathbf{X}_j} \mathbf{x} \neq \mathbf{x}$. Hence it follows that $c_{j,\theta} > 0$. Moreover, from (A.9), $h(j_1, \theta)$ is given by

$$h(j_1, \theta) = h(j, \theta) + \frac{1}{c_{j, \theta}} \mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j')^2 \mathbf{x}. \quad (\text{A.11})$$

By applying a similar expression in (A.10), we have

$$\mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j')^2 \mathbf{x} = \mathbf{x}' \mathbf{x} - 2 \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2}{d_{j,\alpha} + \theta} + \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2 d_{j,\alpha}}{(d_{j,\alpha} + \theta)^2}.$$

The above equation leads to

$$\frac{\partial}{\partial \theta} \mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j')^2 \mathbf{x} = 2\theta \sum_{\alpha=1}^{k_j} \frac{b_{j,\alpha}^2}{(d_{j,\alpha} + \theta)^3} \geq 0,$$

with equality if and only if $\theta = 0$ or $\theta = \infty$. Therefore, we can see that $\mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j')^2 \mathbf{x}$ is a strictly monotonic increasing function of $\theta \in [0, \infty]$. From this result, we obtain

$$\mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j, \theta}^{-1} \mathbf{X}_j')^2 \mathbf{x} \geq \mathbf{x}'(\mathbf{I}_n - \mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j')^2 \mathbf{x} = \mathbf{x}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_j}) \mathbf{x} > 0.$$

Recall that $c_{j,\theta} > 0$. Substituting the above inequality into (A.11) yields $h(j_1, \theta) > h(j, \theta)$ when $\mathbf{X}_{j_1} = (\mathbf{X}_j \mathbf{x})$. By means of similar calculations, we obtain the following theorem.

Theorem A.3. For fixed θ , the following relation on $h(j, \theta)$ can be derived:

$$h(j_1, \theta) < h(j_2, \theta), \quad (\text{when } j_1 \subset j_2). \quad (\text{A.12})$$

A.4. Proofs of Theorems 7 and 8

First, we give the proof of Theorem 7. Recall that $\hat{J}_\theta^{(m)}$ and $\hat{J}_\theta^{(c)}$ are minimizing $MC_p(j, \theta)$ and $C_p(j, \theta)$ over j respectively, for fixed θ . Then, we have

$$MC_p(\hat{J}_\theta^{(m)}, \theta) = \min_{j \subseteq \omega} MC_p(j, \theta), \quad C_p(\hat{J}_\theta^{(c)}, \theta) = \min_{j \subseteq \omega} C_p(j, \theta). \quad (\text{A.13})$$

Suppose that the inequality $\hat{J}_\theta^{(c)} \subset \hat{J}_\theta^{(m)}$ holds. Then, from (A.12) in Appendix A.3, we derive $h(\hat{J}_\theta^{(c)}, \theta) < h(\hat{J}_\theta^{(m)}, \theta)$. Moreover, by applying (A.13), $C_p(\hat{J}_\theta^{(c)}, \theta) \leq C_p(\hat{J}_\theta^{(m)}, \theta)$ can be obtained. Notice that $0 < a < 1$ and $0 < 1 - a < 1$. Substituting the two inequalities into (17) yields

$$\begin{aligned} MC_p(\hat{J}_\theta^{(m)}, \theta) &= aC_p(\hat{J}_\theta^{(m)}, \theta) + 2p(1-a)h(\hat{J}_\theta^{(m)}, \theta) + p(p+1) \\ &> aC_p(\hat{J}_\theta^{(c)}, \theta) + 2p(1-a)h(\hat{J}_\theta^{(c)}, \theta) + p(p+1) \\ &= MC_p(\hat{J}_\theta^{(c)}, \theta). \end{aligned}$$

However, this result is contradictory to $MC_p(\hat{j}_\theta^{(m)}, \theta) \leq MC_p(j, \theta)$ for all j . Consequently, by reductio ad absurdum, the statement in [Theorem 7](#) is proved.

Next, we give the proof of [Theorem 8](#). Let us recall that $\hat{\theta}^{(m)}$ and $\hat{j}^{(m)}$ are the values of θ and j that minimize $MC_p(j, \theta)$, and let $\hat{\theta}^{(c)}$ and $\hat{j}^{(c)}$ be the values of θ and j that minimize $C_p(j, \theta)$. Then, we have

$$MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) = \min_{j \leq \omega, \theta \geq 0} MC_p(j, \theta), \quad C_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}) = \min_{j \leq \omega, \theta \geq 0} C_p(j, \theta). \quad (\text{A.14})$$

Suppose that the inequalities $\hat{\theta}^{(m)} < \hat{\theta}^{(c)}$ and $\hat{j}^{(c)} < \hat{j}^{(m)}$ hold. Then, from [\(A.8\)](#) and [\(A.12\)](#) in [Appendix A.3](#), we have $h(\hat{j}^{(c)}, \hat{\theta}^{(c)}) < h(\hat{j}^{(c)}, \hat{\theta}^{(m)}) < h(\hat{j}^{(m)}, \hat{\theta}^{(m)})$. Moreover, by applying [\(A.13\)](#), the inequalities $C_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}) \leq C_p(\hat{j}^{(m)}, \hat{\theta}^{(m)})$ are obtained. Recall again that $0 < a < 1$ and $0 < 1 - a < 1$. Substituting the two inequalities into [\(17\)](#) yields

$$\begin{aligned} MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) &= aC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) + 2p(1-a)h(\hat{j}^{(m)}, \hat{\theta}^{(m)}) + p(p+1) \\ &> aC_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}) + 2p(1-a)h(\hat{j}^{(c)}, \hat{\theta}^{(c)}) + p(p+1) \\ &= MC_p(\hat{j}^{(c)}, \hat{\theta}^{(c)}). \end{aligned}$$

However, this result is contradictory to $MC_p(\hat{j}^{(m)}, \hat{\theta}^{(m)}) \leq MC_p(j, \theta)$ for all θ and j . Consequently, by reductio ad absurdum, it follows that the statement of [Theorem 8](#) has been proved.

A.5. On variance of the formal GCV criterion

Notice that $\text{Var}[C_p(j, \theta)] = \text{Var}[\text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1})]$. Hence we have

$$\text{Var}[\text{GCV}(j, \theta)] = \frac{\text{Var}[\text{tr}(\mathbf{W}_{j,\theta} \mathbf{S}^{-1})]}{\{1 - h(j, \theta)/n\}^4} = \frac{\text{Var}[C_p(j, \theta)]}{\{1 - h(j, \theta)/n\}^4}, \quad (\text{A.15})$$

where $h(j, \theta)$ is given by [\(16\)](#). Let $(\mathbf{A})_{ab}$ denote the (a, b) th element of the matrix \mathbf{A} . From a simple calculation, we have

$$1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} = 1 - \sum_{\alpha=1}^{k_j} \frac{\{(\mathbf{X}_j \mathbf{Q}_j)_{i\alpha}\}^2}{d_{j,\alpha} + \theta},$$

where \mathbf{Q}_j is a $k_j \times k_j$ orthogonal matrix given by [\(A.1\)](#). Since $\sum_{\alpha=1}^{k_j} \{(\mathbf{X}_j \mathbf{Q}_j)_{i\alpha}\}^2 / (d_{j,\alpha} + \theta)$ is a strictly monotonic decreasing function of $\theta \in [0, \infty]$, an upper bound of $1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii}$ can be derived as

$$1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} \leq 1 - \lim_{\theta \rightarrow \infty} \sum_{\alpha=1}^{k_j} \frac{\{(\mathbf{X}_j \mathbf{Q}_j)_{i\alpha}\}^2}{d_{j,\alpha} + \theta} = 1.$$

From elementary linear algebra, it can be easily seen that $0 \leq (\mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j')_{ii} \leq 1$, because $(\mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j')_{ii}$ is a diagonal element of a symmetric idempotent matrix. By using the range of $(\mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j')_{ii}$ and the equation $\sum_{\alpha=1}^{k_j} \{(\mathbf{X}_j \mathbf{Q}_j)_{i\alpha}\}^2 / d_{j,\alpha} = (\mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j')_{ii}$, we obtain

$$1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} \geq 1 - \sum_{\alpha=1}^{k_j} \frac{\{(\mathbf{X}_j \mathbf{Q}_j)_{i\alpha}\}^2}{d_{j,\alpha}} = 1 - (\mathbf{X}_j \mathbf{M}_{j,0}^{-1} \mathbf{X}_j')_{ii} \geq 0.$$

Recall that $1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} \neq 0$ because the CV criterion defined by [\(21\)](#) does not exist if $1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} = 0$. The existence of CV and the nonnegativity of $1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii}$ lead us to a lower bound as $1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} > 0$. Combining the lower and upper bounds yields

$$0 < 1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii} \leq 1, \quad (i = 1, \dots, n).$$

Notice that $1 - h(j, \theta)/n = n^{-1} \sum_{i=1}^n \{1 - (\mathbf{X}_j \mathbf{M}_{j,\theta}^{-1} \mathbf{X}_j')_{ii}\}$. Hence, we also have

$$0 < 1 - \frac{1}{n} h(j, \theta) \leq 1. \quad (\text{A.16})$$

Consequently, from the Eqs. [\(A.15\)](#) and [\(A.16\)](#), the following inequality can be derived:

$$\text{Var}[C_p(j, \theta)] \leq \text{Var}[\text{GCV}(j, \theta)], \quad (\text{A.17})$$

with equality if and only if $\theta = \infty$. By using [\(A.17\)](#) together with [Theorem 3](#), we derive the following theorem.

Theorem A.4. For any distribution of \mathbf{Y} , the following inequality is always satisfied:

$$\text{Var}[MC_p(j, \theta)] \leq \text{Var}[C_p(j, \theta)] \leq \text{Var}[\text{GCV}(j, \theta)], \quad (\text{A.18})$$

with the first equality if and only if $\theta = 0$ and $j = \omega$, and the second equality if and only if $\theta = \infty$.

References

- [1] M.S. Srivastava, *Methods of Multivariate Statistics*, John Wiley & Sons, New York, 2002.
- [2] N.H. Timm, *Applied Multivariate Analysis*, Springer-Verlag, New York, 2002.
- [3] C. Sârbu, C. Onişor, M. Posa, S. Kevresan, K. Kuhajda, Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods, *Talanta* 75 (2008) 651–657.
- [4] R. Saxén, J. Sundell, ¹³⁷Cs in freshwater fish in Finland since 1986 – a statistical analysis with multivariate linear regression models, *J. Environ. Radioactiv.* 87 (2006) 62–76.
- [5] A. Yoshimoto, H. Yanagihara, Y. Ninomiya, Finding factors affecting a forest stand growth through multivariate linear modeling, *J. Jpn. For. Soc.* 87 (2005) 504–512 (in Japanese).
- [6] A.E. Hoerl, R. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* 12 (1970) 55–67.
- [7] C.L. Mallows, Some comments on C_p , *Technometrics* 15 (1973) 661–675.
- [8] C.L. Mallows, More comments on C_p , *Technometrics* 37 (1995) 362–372.
- [9] Y. Fujikoshi, K. Satoh, Modified AIC and C_p in multivariate linear regression, *Biometrika* 84 (1997) 707–716.
- [10] N.R. Draper, A.M. Herzberg, A ridge-regression sidelight, *Amer. Statist.* 41 (1987) 282–283.
- [11] C.M. Hurvich, J.S. Simonoff, C.-L. Tsai, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *J. Roy. Statist. Soc. Ser. B* 60 (1998) 271–293.
- [12] S.J. Davies, A.A. Neath, J.E. Cavanaugh, Estimation optimality of corrected AIC and modified C_p in linear regression model, *Internat. Statist. Rev.* 74 (2006) 161–168.
- [13] R.S. Sparks, D. Coutourides, L. Troskie, The multivariate C_p , *Comm. Statist. A Theory Methods* 12 (1983) 1775–1793.
- [14] M. Siotani, T. Hayakawa, Y. Fujikoshi, *Modern Multivariate Statistical Analysis: A Graduate Course and Handbook*, American Sciences Press, Columbus, Ohio, 1985.
- [15] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CLC, New York, 1993.
- [16] Y. Fujikoshi, H. Yanagihara, H. Wakaki, Bias corrections of some criteria for selection multivariate linear regression models in a general case, *Amer. J. Math. Management Sci.* 25 (2005) 221–258.
- [17] Y. Fujikoshi, T. Noguchi, M. Ohtaki, H. Yanagihara, Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models, *Ann. Inst. Statist. Math.* 55 (2003) 537–553.
- [18] M. Stone, Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc. Ser. B* 36 (1974) 111–147.
- [19] P. Craven, G. Wahba, Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* 31 (1979) 377–403.
- [20] P.J. Brown, Aspects of multivariate regression, *Trabajos Investigación Oper.* 31 (1980) 249–265.
- [21] B. Efron, The estimation of prediction error: covariance penalties and cross-validation, *J. Amer. Statist. Assoc.* 99 (2004) 619–632.
- [22] K.-C. Li, Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing, *Ann. Statist.* 14 (1986) 1101–1112.