Hiroshima Math. J.
42 (2012), 301–324

Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods

Isamu Nagai, Hirokazu Yanagihara and Kenichi Satoh

(Received May 11, 2011)
(Revised November 10, 2011)

Abstract. Generalized ridge (GR) regression for an univariate linear model was proposed simultaneously with ridge regression by Hoerl and Kennard (1970). In this paper, we deal with a GR regression for a multivariate linear model, referred to as a multivariate GR (MGR) regression. From the viewpoint of reducing the mean squared error (MSE) of a predicted value, many authors have proposed several GR estimators consisting of ridge parameters optimized by non-iterative methods. By expanding their optimizations of ridge parameters to the multiple response case, we derive some MGR estimators with ridge parameters optimized by the plug-in method. We analytically compare obtained MGR estimators with existing MGR estimators, and numerical studies are also given for an illustration.

1. Introduction

We consider a multivariate linear regression model with n observations of a p-dimensional vector of response variables and a k-dimensional vector of regressors (for more detailed information, see for example, Srivastava, 2002, Chapter 9; Timm, 2002, Chapter 4). Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \dots, \boldsymbol{y}_n)'$, $\boldsymbol{X}$ and $\mathscr{E}$ be the $n \times p$ matrix of response variables, the $n \times k$ matrix of non-stochastic centerized explanatory variables (i.e., $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$) of $\mathrm{rank}(\boldsymbol{X}) = k$ $(< n)$, and the $n \times p$ matrix of error variables, respectively, where n is the sample size, $\boldsymbol{1}_n$ is an n-dimensional vector of ones and $\boldsymbol{0}_k$ is a k-dimensional vector of zeros. Suppose that the row vectors of $\mathscr{E}$ are independently and identically distributed according to a distribution with mean $\boldsymbol{0}_p$ and an unknown covariance matrix $\boldsymbol{\Sigma}$. The matrix form of the multivariate linear regression model is expressed as

$$\boldsymbol{Y} = \boldsymbol{1}_n \boldsymbol{\mu}' + \boldsymbol{X}\boldsymbol{\Xi} + \mathscr{E}, \tag{1}$$

2010 *Mathematics Subject Classification*. Primary 62J07; Secondary 62H12.
Key words and phrases. Generalized ridge regression; Mallows' C_p statistic; Model selection; Multivariate linear regression model; Non-iterative estimation; Plug-in method.

where $\boldsymbol{\mu}$ is a p-dimensional unknown vector and $\boldsymbol{\Xi}$ is a $k \times p$ unknown regression coefficient matrix.

Since $\boldsymbol{X}$ is centerized, the maximum likelihood estimators under normality or least squares (LS) estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Xi}$ are given by $\bar{\boldsymbol{y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{y}_i$ and

$$\hat{\boldsymbol{\Xi}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{2}$$

respectively. Since $\bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{\Xi}}$ are simple and the unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Xi}$, it is widely used in actual data analysis, see e.g., Dien *et al.* (2006), Sârbu *et al.* (2008), Saxén and Sundell (2006), Skagerberg, Macgregor and Kiparissides (1992), Yoshimoto, Yanagihara and Ninomiya (2005). However, when multicollinearity occurs in $\boldsymbol{X}$, the LS estimator of $\boldsymbol{\Xi}$ is not a good estimator in the sense of having a large variance. The ridge regression for an univariate linear model proposed by Hoerl and Kennard (1970) is one of the ways of avoiding such problems that arise from multicollinearity. The ridge estimator is defined by adding $\theta\boldsymbol{I}_k$ to $\boldsymbol{X}'\boldsymbol{X}$ in the LS estimator, where θ (≥ 0) is called a ridge parameter. Since estimates of the ridge estimator depend heavily on the value of θ, optimization of θ is a very important problem. Choosing θ so that the mean squared error (MSE) of a predictor of $\boldsymbol{Y}$ becomes small is a common procedure. However, the optimal value of θ cannot be obtained without any iterative computational algorithm.

Hoerl and Kennard (1970) also proposed a generalized ridge (GR) regression for the univariate linear model simultaneously with the ridge regression. The GR estimator is defined not by a single ridge parameter but by multiple ridge parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$, $(\theta_i \geq 0,\ i = 1, \dots, k)$. Even though the number of parameters has increased, we can obtain an explicit solution for $\boldsymbol{\theta}$ to the minimization problem of the MSE of a predictor of $\boldsymbol{Y}$. By using such closed forms for the solutions, many authors have proposed several GR estimators such that $\boldsymbol{\theta}$ can be obtained by non-iterative optimization methods (see e.g., Lawless, 1981).

It is well known that the ridge estimator is a shrinkage estimator of regression coefficients towards the origin. One of the advantages of the GR regression is to be able to obtain a shrinkage estimate for regression coefficients without the use of any iterative optimization algorithm on $\boldsymbol{\theta}$. It also has other advantages, namely, whereas the ridge regression shrinks uniformly all coefficients of the LS estimator by a single ridge parameter, for the GR regression, the amount of shrinkage is different for each explanatory variable. Thus the GR regression is more flexible than the ridge regression. From this viewpoint, we deal not with the ridge regression but the GR regression. We refer to the GR regression for a multivariate linear model as the multivariate GR (MGR) regression.

Methods for optimizing $\boldsymbol{\theta}$ in the GR regression can be roughly divided into the following types:

- We obtain the optimal $\boldsymbol{\theta}$ by replacing unknown parameters with their estimators in the explicit solution of $\boldsymbol{\theta}$ to the minimization problem for the MSE of a predictor of $\boldsymbol{Y}$;
- We choose an optimal value of $\boldsymbol{\theta}$ that makes the estimator of the MSE of a predicted value of $\boldsymbol{Y}$ a minimum.

In this paper, the first type of method is referred to as a plug-in method. Since the second method corresponds to a determination of $\boldsymbol{\theta}$ by minimizing an information criterion (IC), i.e., the C_p criterion proposed by Mallows (1973; 1995) (for the multivariate case, see Sparks, Coutsourides and Troskie (1983)), the second type of method is called an IC-based method. For each of the above two types of the optimization methods in the GR regression, formulas for obtaining optimal $\boldsymbol{\theta}$ in the MGR regression will be derived.

By extending the formulas for a GR estimator with the optimized ridge parameters from the plug-in method to the multivariate case, we are able to propose several MGR estimators with ridge parameters optimized by a non-iterative method. As for the C_p criterion for the MGR regression, Yanagihara, Nagai and Satoh (2009) considered the C_p criterion and proposed a bias-corrected C_p criterion called a modified C_p (MC_p) criterion. Their MC_p criterion includes criteria proposed by Fujikoshi and Satoh (1997) and Yanagihara and Satoh (2010) as special cases. In this paper, we consider the generalized C_p (GC_p) criterion (originally GC_p for selecting variables in the univariate regression was proposed by Atkinson (1980)) for the MGR regression, which includes C_p and MC_p criteria omitting constant terms, as special cases. By using the GC_p criterion, we can deal systematically with the optimization of $\boldsymbol{\theta}$ when using an IC-based method. In particular, a family of the MGR estimators with the optimal $\boldsymbol{\theta}$ obtained using the IC-based framework contains the James-Stein estimator proposed by Kubokawa (1991).

This paper is organized in the following way: In Section 2, we extend the univariate GR regression to the MGR regression. Then we illustrate a target MSE of a predictor of $\boldsymbol{Y}$ and derive $\boldsymbol{\theta}$ so that the MSE is minimized. In Section 3, we consider the MGR estimators with the optimized ridge parameters and propose plug-in method for the MGR estimator by extending the method for the GR estimator. In Section 4, we consider the GC_p criterion and optimized method based on IC-based method and another method. In Section 5, we discuss relationships between test statistics and the optimized values of $\boldsymbol{\theta}$, and give the magnitude relation among optimized $\boldsymbol{\theta}$s. In Section 6, we compare derived MGR estimators with existing MGR estimators by conducting numerical studies. Technical details are provided in Appendix.

2. MGR estimator and target MSE

2.1. Preliminaries. By naturally extending the GR estimator, we derive the MGR estimator for (1) as

$$\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{3}$$

where $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta})$ and $\boldsymbol{Q}$ is the $k \times k$ orthogonal matrix which diagonalizes $\boldsymbol{X}'\boldsymbol{X}$, i.e.,

$$\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{Q} = \mathrm{diag}(d_1, \dots, d_k) = \boldsymbol{D}. \tag{4}$$

Here $d_1, \dots, d_k$ are eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$. We note that the d_i's are always positive. We can check that the estimator in (3) corresponds to the ordinary LS estimator in (2) when $\boldsymbol{\theta} = \boldsymbol{0}_k$. This means that the estimator in (3) includes the ordinary LS estimator. If $p = 1$, then the estimator in (3) corresponds to the GR estimator proposed by Hoerl and Kennard (1970).

Let $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$ be a predictor of $\boldsymbol{Y}$, given by $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}} = \boldsymbol{1}_n\bar{\boldsymbol{y}}' + \boldsymbol{X}\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}$. In order to define the MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$, we define the following discrepancy function for measuring the distance between $n \times p$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$r(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\{(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{A} - \boldsymbol{B})'\}. \tag{5}$$

Since $\boldsymbol{\Sigma}$ is an unknown covariance matrix, we use the following unbiased estimator instead of $\boldsymbol{\Sigma}$:

$$\boldsymbol{S} = \frac{1}{n-k-1}(\boldsymbol{Y} - \boldsymbol{1}_n\bar{\boldsymbol{y}}' - \boldsymbol{X}\hat{\boldsymbol{\Xi}})'(\boldsymbol{Y} - \boldsymbol{1}_n\bar{\boldsymbol{y}}' - \boldsymbol{X}\hat{\boldsymbol{\Xi}}), \tag{6}$$

where $\hat{\boldsymbol{\Xi}}$ is given in (2). By replacing $\boldsymbol{\Sigma}$ with (6), we can estimate (5) by

$$\hat{r}(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\{(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{S}^{-1}(\boldsymbol{A} - \boldsymbol{B})'\}. \tag{7}$$

These two functions in (5) and (7) correspond to summations of the Mahalanobis distance and the sample Mahalanobis distance between rows of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. By using (5), the MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$ is defined as

$$\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}] = E[r(E[\boldsymbol{Y}], \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})]. \tag{8}$$

In this paper, we choose $\boldsymbol{\theta}$ that minimizes the MSE in (8) as the principal optimum.

2.2. Model transformation. In this subsection, we consider an orthogonal transformation of $\boldsymbol{Y}$ in order to simplify the calculation of $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}]$. Such a transformation with $p = 1$ was used in Goldstein and Smith (1974) and Walker and Page (2001), etc. We extend their transformation to the multivariate regression case. By using the singular value decomposition, we can determine an $n \times n$ orthogonal matrix $\boldsymbol{P}_1$ and a $(k+1) \times (k+1)$ orthogonal matrix $\boldsymbol{P}_2$ such that

$$(\boldsymbol{X}, \boldsymbol{1}_n) = \boldsymbol{P}_1 \boldsymbol{L} \boldsymbol{P}_2', \tag{9}$$

where $\boldsymbol{L}$ is an $n \times (k+1)$ matrix. Recall that $\boldsymbol{X}$ is centerized. Therefore, we have

$$(\boldsymbol{X}, \boldsymbol{1}_n)'(\boldsymbol{X}, \boldsymbol{1}_n) = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{0}_k \\ \boldsymbol{0}_k' & n \end{pmatrix}. \tag{10}$$

Since the orthogonal matrix $\boldsymbol{P}_2$ diagonalizes (10), from (4), $\boldsymbol{P}_2$ and $\boldsymbol{L}$ can be expressed as

$$\boldsymbol{P}_2 = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0}_k \\ \boldsymbol{0}_k' & 1 \end{pmatrix}, \tag{11}$$

and

$$\boldsymbol{L} = (\mathrm{diag}(\sqrt{d_1}, \dots, \sqrt{d_k}, \sqrt{n}), \boldsymbol{O}_{k+1, n-k-1})',$$

where $\boldsymbol{O}_{n,k}$ is an $n \times k$ matrix of zeros.

Let

$$\boldsymbol{Z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)' = \boldsymbol{P}_1'\boldsymbol{Y}, \qquad \boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_k)' = \boldsymbol{Q}'\boldsymbol{\Xi},$$
$$\mathscr{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)' = \boldsymbol{P}_1'\mathscr{E}. \tag{12}$$

By using (9) and (11), $\boldsymbol{Z}$ is calculated as

$$\boldsymbol{Z} = \boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)\begin{pmatrix} \boldsymbol{\Xi} \\ \boldsymbol{\mu}' \end{pmatrix} + \boldsymbol{P}_1'\mathscr{E} = \boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)\boldsymbol{P}_2\begin{pmatrix} \boldsymbol{Q}'\boldsymbol{\Xi} \\ \boldsymbol{\mu}' \end{pmatrix} + \mathscr{V} = \boldsymbol{L}\begin{pmatrix} \boldsymbol{\Gamma} \\ \boldsymbol{\mu}' \end{pmatrix} + \mathscr{V}. \tag{13}$$

Since $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{Y})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ holds, we have

$$\mathrm{Cov}[\mathrm{vec}(\boldsymbol{Z})] = (\boldsymbol{I}_p \otimes \boldsymbol{P}_1')\,\mathrm{Cov}[\mathrm{vec}(\boldsymbol{Y})](\boldsymbol{I}_p \otimes \boldsymbol{P}_1) = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n.$$

This equation means that $\mathrm{Cov}[\boldsymbol{z}_i] = \boldsymbol{\Sigma}$, $(i = 1, \dots, n)$. Thus, from this result and (13), the following equation is obtained:

$$\boldsymbol{z}_i = \begin{cases} \sqrt{d_i}\boldsymbol{\gamma}_i + \boldsymbol{v}_i & (i = 1, \dots, k) \\ \sqrt{n}\boldsymbol{\mu} + \boldsymbol{v}_i & (i = k+1) \\ \boldsymbol{v}_i & (i = k+2, \dots, n) \end{cases}, \qquad (E[\boldsymbol{v}_i] = \boldsymbol{0}_p, \mathrm{Cov}[\boldsymbol{v}_i] = \boldsymbol{\Sigma}). \tag{14}$$

2.3. Equivalence of MSE$[\hat{\boldsymbol{Y}}_\theta]$ and MSE$[\hat{\boldsymbol{Z}}_\theta]$.

By a simple calculation, we can determine that the LS estimator of $(\boldsymbol{\Gamma}', \boldsymbol{\mu})'$ is $(\boldsymbol{L}'\boldsymbol{L})^{-1}\boldsymbol{L}'\boldsymbol{Z}$. Hence, the LS estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{\mu}$ can be expressed as $\hat{\boldsymbol{\Gamma}} = \boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{Z}$ and $\hat{\boldsymbol{\mu}} = \boldsymbol{z}_{k+1}/\sqrt{n}$, respectively, where $\boldsymbol{C} = (\boldsymbol{D}^{1/2}, \boldsymbol{O}_{k,n-k})'$. By replacing $\boldsymbol{D}$ in $\hat{\boldsymbol{\Gamma}}$ with $\boldsymbol{D} + \boldsymbol{\Theta}$, the MGR estimator of $\boldsymbol{\Gamma}$ can be determined as

$$\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{C}'\boldsymbol{Z}. \tag{15}$$

Notice that $\boldsymbol{P}_1'\boldsymbol{X}\boldsymbol{Q} = \boldsymbol{C}$. Hence, the relation between the MGR estimators of $\boldsymbol{\Xi}$ and $\boldsymbol{\Gamma}$ is as follows:

$$\boldsymbol{Q}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{Q}\boldsymbol{C}'\boldsymbol{P}_1'\boldsymbol{Y} = \hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}. \tag{16}$$

Let $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}$ be a predictor of $\boldsymbol{Z}$, i.e., $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{L}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}', \hat{\boldsymbol{\mu}})'$. The relation between $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{P}_1'\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{P}_2'\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0}_k \\ \boldsymbol{0}_k' & 1 \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\mu}}' \end{pmatrix} = \boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)\begin{pmatrix} \hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\mu}}' \end{pmatrix} = \boldsymbol{P}_1'\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}. \tag{17}$$

Notice that $E[\boldsymbol{Z}] = \boldsymbol{P}_1'E[\boldsymbol{Y}]$. Thus $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}]$ can be rewritten as

$$\begin{aligned} \mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}] &= E[\mathrm{tr}\{(E[\boldsymbol{Y}] - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}^{-1}(E[\boldsymbol{Y}] - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})'\boldsymbol{P}_1\boldsymbol{P}_1'\}] \\ &= E[r(E[\boldsymbol{Z}], \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})] = \mathrm{MSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}]. \end{aligned} \tag{18}$$

The above equation implies that the MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$ is equivalent to the MSE of $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}$. Therefore it appears that we can search for $\boldsymbol{\theta}$ minimizing the MSE of $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}$ instead of the MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$.

2.4. Principal optimum $\boldsymbol{\theta}$. Recall that $E[\boldsymbol{Z}] = \boldsymbol{L}(\boldsymbol{\Gamma}', \boldsymbol{\mu})'$ and $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{L}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}', \hat{\boldsymbol{\mu}})'$. Then $r(E[\boldsymbol{Z}], \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})$ can be rewritten as

$$r(E[\boldsymbol{Z}], \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}) = \mathrm{tr}\left\{\boldsymbol{L}\begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix}\boldsymbol{\Sigma}^{-1}\begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix}'\boldsymbol{L}'\right\}. \tag{19}$$

By elementary linear algebra,

$$\boldsymbol{L}\begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix} = \begin{pmatrix} \mathrm{diag}(\sqrt{d_1}, \dots \sqrt{d_k}, \sqrt{n}) \\ \boldsymbol{O}_{n-k-1,k+1} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}^{1/2}(\boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}) \\ \sqrt{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \\ \boldsymbol{O}_{n-k-1,p} \end{pmatrix}. \tag{20}$$

Notice that

$$\begin{aligned} \boldsymbol{D}^{1/2}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} &= \boldsymbol{D}^{1/2}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{C}'\boldsymbol{Z} \\ &= (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}(\boldsymbol{D}, \boldsymbol{O}_{k,n-k})\boldsymbol{Z} = \left(\frac{d_1}{d_1 + \theta_1}z_1, \dots, \frac{d_k}{d_k + \theta_k}z_k\right)'. \end{aligned} \tag{21}$$

This equation implies that

$$\boldsymbol{D}^{1/2}(\boldsymbol{\Gamma}-\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}) = \boldsymbol{D}^{1/2}\boldsymbol{\Gamma}-(\boldsymbol{D}+\boldsymbol{\Theta})^{-1}(\boldsymbol{D},\boldsymbol{O}_{k,n-k})\boldsymbol{Z}$$
$$= \left(\sqrt{d_1}\gamma_1-\frac{d_1}{d_1+\theta_1}z_1,\ldots,\sqrt{d_k}\gamma_k-\frac{d_k}{d_k+\theta_k}z_k\right)'. \tag{22}$$

By using equations (19), (20) and (22), we can derive another expression for the MSE$[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}]$ as

$$\begin{aligned}\mathrm{MSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}] &= E[r(E[\boldsymbol{Z}],\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})]\\ &= \sum_{i=1}^{k}E\left[\left(\sqrt{d_i}\gamma_i-\frac{d_i}{d_i+\theta_i}z_i\right)'\boldsymbol{\Sigma}^{-1}\left(\sqrt{d_i}\gamma_i-\frac{d_i}{d_i+\theta_i}z_i\right)\right]\\ &\quad+nE[(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})].\end{aligned} \tag{23}$$

Recall that $\hat{\boldsymbol{\mu}}=z_{k+1}/\sqrt{n}$. It follows from (14) that

$$\begin{aligned}nE[(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})] &= E[(\sqrt{n}\boldsymbol{\mu}-z_{k+1})'\boldsymbol{\Sigma}^{-1}(\sqrt{n}\boldsymbol{\mu}-z_{k+1})]\\ &= \mathrm{tr}(\mathrm{Cov}[z_{k+1}]\boldsymbol{\Sigma}^{-1})=p.\end{aligned} \tag{24}$$

Moreover, by using the results that $E[z_i]=\sqrt{d_i}\gamma_i$ and $E[z_iz_i']=\boldsymbol{\Sigma}+d_i\gamma_i\gamma_i'$, $(i=1,\ldots,k)$, we calculate that

$$E\left[\left(\sqrt{d_i}\gamma_i-\frac{d_i}{d_i+\theta_i}z_i\right)'\boldsymbol{\Sigma}^{-1}\left(\sqrt{d_i}\gamma_i-\frac{d_i}{d_i+\theta_i}z_i\right)\right]=\varphi(\theta_i|d_i,\gamma_i), \tag{25}$$

where

$$\varphi(\theta_i|d_i,\gamma_i)=d_i\gamma_i'\boldsymbol{\Sigma}^{-1}\gamma_i-\frac{2d_i^2}{d_i+\theta_i}\gamma_i'\boldsymbol{\Sigma}^{-1}\gamma_i+\left(\frac{d_i}{d_i+\theta_i}\right)^2(p+d_i\gamma_i'\boldsymbol{\Sigma}^{-1}\gamma_i).$$

Substituting (24) and (25) into (23) yields

$$\mathrm{MSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}]=\sum_{i=1}^{k}\varphi(\theta_i|d_i,\gamma_i)+p.$$

The above equation indicates that the principal optimum value of θ_i can be obtained by minimizing $\varphi(\theta_i|d_i,\gamma_i)$ individually. Let $\theta_i^*\ge 0$, $(i=1,\ldots,k)$ be the principal optimum value of θ_i. The first partial derivative of $\varphi(\theta_i|d_i,\gamma_i)$ with respect to θ_i is calculated as

$$\frac{\partial}{\partial\theta_i}\varphi(\theta_i|d_i,\gamma_i)=\frac{2d_i^2}{(d_i+\theta_i)^3}(\theta_i\gamma_i'\boldsymbol{\Sigma}^{-1}\gamma_i-p).$$

The above equation yields the principal optimum value of θ_i as

$$\theta_i^* = \frac{p}{\gamma_i' \boldsymbol{\Sigma}^{-1} \gamma_i}, \qquad (i = 1, \dots, k). \tag{26}$$

3. Plug-in method

3.1. MGR estimators with the optimized ridge parameters. For the case of an univariate linear model, many authors have provided formulas for the GR estimators with the optimized ridge parameters. By extending their methods for optimizing $\boldsymbol{\theta}$ to the multivariate case, we derive formulas for the MGR estimators with the optimized ridge parameters. Since the MGR estimator $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}$ in (3) is obtained by using the equation $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}} = \boldsymbol{Q}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}$ in (16), we deal with $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}$ in (15) instead of $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}$. Let $\hat{\boldsymbol{\Gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_k)'$ be the ordinary LS estimator of $\boldsymbol{\Gamma}$, i.e., $\hat{\boldsymbol{\Gamma}} = \boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{Z}$. This implies that $\hat{\gamma}_i = z_i/\sqrt{d_i}$. Then, we have

$$\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{C}'\boldsymbol{Z} = (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{D}\hat{\boldsymbol{\Gamma}}. \tag{27}$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)'$, $(\hat{\theta}_i \geq 0,\ i = 1, \dots, k)$ be the value of $\boldsymbol{\theta}$ optimized by such a method, and let $\hat{\gamma}_i(\hat{\theta}_i)$ be the ith row vector of $\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{\theta}}}$, which is defined by substituting $\hat{\boldsymbol{\theta}}$ into $\boldsymbol{\theta}$ in $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}$. From equation (27), we can see that $\hat{\gamma}_i(\hat{\theta}_i)$ is expressed as

$$\hat{\gamma}_i(\hat{\theta}_i) = \frac{d_i}{d_i + \hat{\theta}_i}\hat{\gamma}_i, \qquad (i = 1, \dots, k). \tag{28}$$

It is clearly that $\hat{\gamma}_i(0) = \hat{\gamma}_i$. Let

$$t_i = z_i' \boldsymbol{S}^{-1} z_i, \qquad (i = 1, \dots, k). \tag{29}$$

Since $\hat{\gamma}_i = z_i/\sqrt{d_i}$, t_i in (29) can be rewritten as

$$t_i = d_i \hat{\gamma}_i' \boldsymbol{S}^{-1} \hat{\gamma}_i, \qquad (i = 1, \dots, k). \tag{30}$$

If $\hat{\theta}_i$ is a function of t_i, then we can express $\hat{\gamma}_i(\hat{\theta}_i)$ in (28) as

$$\hat{\gamma}_i(\hat{\theta}_i) = w(t_i)\hat{\gamma}_i, \qquad (i = 1, \dots, k),$$

where $w(t_i)$ is a function of t_i. From (28), it is clearly the case that $0 \leq w(t_i) \leq 1$, because $d_i > 0$ and $\hat{\theta}_i \geq 0$. Hence $w(t_i)$ is called the weight function. By using such a weight function, Lawless (1981) expressed several GR estimators with the optimized ridge parameters. According to his notation, we specify the individual MGR estimator with an optimized value of $\boldsymbol{\theta}$ using the weight function.

3.2. Once plug-in method. Since the principal optimum value $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_k^*)'$ is obtained as (26), we estimate θ_i^* by replacing γ_i and $\boldsymbol{\Sigma}$ with $\hat{\gamma}_i$ and $\boldsymbol{S}$. Then we obtain the following optimal $\boldsymbol{\theta}$ by single plug-in estimation:

$$\hat{\theta}_i^{[1]} = \frac{p}{\hat{\gamma}_i' \boldsymbol{S}^{-1} \hat{\gamma}_i} = \frac{d_i p}{t_i}, \qquad (i = 1, \dots, k). \tag{31}$$

Since $w(t_i) = d_i/(d_i + \hat{\theta}_i)$, the weight function corresponding to $\hat{\theta}_i^{[1]}$ is given by

$$w^{[1]}(t_i) = \frac{t_i}{t_i + p}.$$

We refer to this once plug-in method as PI. In the case of $p = 1$, the above results coincide with the result in Hoerl and Kennard (1970).

3.3. Multiple plug-in method. We will avoid problems that arise from multicollinearity by the once plug-in method. However, we find that $\hat{\theta}_i^{[1]}$ is made by the ordinary LS estimator of γ_i. Hence, if multicollinearity occurs, $\hat{\theta}_i^{[1]}$ tends to small beyond necessity because $\hat{\gamma}_i$ tends to have large variance. Such an under evaluation problem of θ_i may be improved by using the MGR estimator instead of $\hat{\gamma}_i$ in the optimal θ_i because the MGR estimator tends to have smaller variance than the ordinary LS estimator. Therefore, we obtain the following optimal $\boldsymbol{\theta}$ by multiple plug-in estimation:

$$\hat{\theta}_i^{[s]} = \frac{p}{\hat{\gamma}_i^{[s-1]'} \boldsymbol{S}^{-1} \hat{\gamma}_i^{[s-1]}}, \qquad (s = 1, 2, \dots; i = 1, \dots, k), \tag{32}$$

where $\hat{\gamma}_i^{[s]} = d_i \hat{\gamma}_i/(d_i + \hat{\theta}_i^{[s]})$, $(s = 0, 1, \dots)$ and $\hat{\theta}_i^{[0]} = 0$. Notice that $\hat{\gamma}_i^{[1]}$ is equal to the estimator obtained using the PI method. Equation (32) implies that

$$\hat{\theta}_i^{[s]} = \left(1 + \frac{\hat{\theta}_i^{[s-1]}}{d_i}\right)^2 \hat{\theta}_i^{[1]}, \qquad (s = 1, 2, \dots; i = 1, \dots, k). \tag{33}$$

In the case of $p = 1$, the value of (32) was proposed by Hoerl and Kennard (1970), and they used $\hat{\gamma}_i^{[2]}$ to estimate the regression coefficient. Hence we also use $\hat{\gamma}_i^{[2]}$ which is obtained by using $\hat{\theta}_i^{[2]}$. We denote this twice plug-in method as PI$_2$. The optimal value of θ_i derived using the PI$_2$ method is given by

$$\hat{\theta}_i^{[2]} = \frac{d_i p (t_i + p)^2}{t_i^3}, \qquad (i = 1, \dots, k),$$

and the weight function corresponding to $\hat{\theta}_i^{[2]}$ is given by

$$w^{[2]}(t_i) = \frac{t_i^3}{t_i^3 + p(t_i + p)^2}.$$

3.4. Infinite plug-in method. For the case of $p = 1$, Hemmerle (1975) showed that the value of (32) converges as $s \to \infty$. By extending the proof in Hemmerle (1975) to the multivariate case, we obtain the following limiting value of (32) as $s \to \infty$:

$$\hat{\theta}_i^{[\infty]} = \begin{cases} \dfrac{d_i\{t_i - 2p - \sqrt{t_i(t_i - 4p)}\}}{2p} & (t_i \geq 4p) \\ \infty & (t_i < 4p) \end{cases}, \qquad (i = 1, \dots, k), \tag{34}$$

(the proof is given in Appendix A.1). We refer to this infinite plug-in method as PI_∞. The weight function $w^{[\infty]}(t_i)$ corresponding to $\hat{\theta}_i^{[\infty]}$ is given by

$$w^{[\infty]}(t_i) = \begin{cases} \dfrac{2p}{t_i(1 - \sqrt{1 - 4p/t_i})} & (t_i \geq 4p) \\ 0 & (t_i < 4p) \end{cases}.$$

4. Alternative methods

4.1. IC-based method. Yanagihara, Nagai and Satoh (2009) proposed C_p criterion for optimizing $\boldsymbol{\theta}$ and its bias-corrected C_p (Modified C_p; MC_p) criterion. By omitting constant terms, their criteria are included in a class of criteria specified by λ. The class is expressed by the generalized C_p (GC_p) criterion as

$$GC_p(\boldsymbol{\theta}|\lambda) = \lambda^{-1}\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}) + 2p\,\mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{X}\}, \tag{35}$$

where the function $\hat{r}$ is given by (7). It notes that $GC_p(\boldsymbol{\theta}|1)$ and $GC_p(\boldsymbol{\theta}|c_{\mathrm{M}})$ are corresponding the main terms with respect to $\boldsymbol{\theta}$ in the C_p and MC_p criteria where $c_{\mathrm{M}} = (n - k - 1)/(n - k - p - 2)$. By using the GC_p criterion, we can deal systematically with the optimization of $\boldsymbol{\theta}$ when we use IC-based method.

The optimal value of θ_i which minimizes (35) is obtained as

$$\hat{\theta}_i^{(\mathrm{G})}(\lambda) = \begin{cases} \dfrac{\lambda p d_i}{t_i - \lambda p} & (t_i > \lambda p) \\ \infty & (t_i \leq \lambda p) \end{cases}, \qquad (i = 1, \dots, k), \tag{36}$$

(the proof is given in Appendix A.2). Then the weight function $w^{(\mathrm{G})}(t_i|\lambda)$ corresponding to $\hat{\theta}_i^{(\mathrm{G})}(\lambda)$ is given by

$$w^{(\mathrm{G})}(t_i|\lambda) = \begin{cases} 1 - \dfrac{\lambda p}{t_i} & (t_i > \lambda p) \\ 0 & (t_i \leq \lambda p) \end{cases}. \tag{37}$$

From (36), $\hat{\theta}_i^{(\mathrm{C})}$ that minimizes the C_p criterion is $\hat{\theta}_i^{(\mathrm{C})} = \hat{\theta}_i^{(\mathrm{G})}(1)$, $(i = 1, \dots, k)$. Then equation (37) yields the weight function of this estimator as $w^{(\mathrm{C})}(t_i) = w^{(\mathrm{G})}(t_i|1)$. This optimization method is referred to as C_p.

Moreover, $\hat{\theta}_i^{(\mathrm{M})}$ minimizing the MC_p criterion is given by $\hat{\theta}_i^{(\mathrm{M})} = \hat{\theta}_i^{(\mathrm{G})}(c_{\mathrm{M}})$, $(i = 1, \dots, k)$, and the weight function is $w^{(\mathrm{M})}(t_i) = w^{(\mathrm{G})}(t_i|c_{\mathrm{M}})$. This optimization method is referred to as MC_p.

Kubokawa (1991) proposed an improved James-Stein estimator which is a shrinkage estimator when $p \geq 3$. Suppose that $\mathscr{E} \sim N_{n\times p}(\boldsymbol{O}_{n,p}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Since $\hat{\gamma}_i \sim N_p(\gamma_i, \boldsymbol{\Sigma}/d_i)$, $(i = 1, \dots, k)$, $(n-k-1)\boldsymbol{S} \sim W_p(n-k-1, \boldsymbol{\Sigma})$ and $\boldsymbol{S} \perp\!\!\!\perp \hat{\gamma}_i$, $(i = 1, \dots, k)$ are satisfied, the James-Stein estimator of γ_i is obtained as

$$\hat{\gamma}_i^{(\mathrm{J})} = \begin{cases} \left(1 - \dfrac{c_{\mathrm{J}} p}{t_i}\right)\hat{\gamma}_i & (t_i > c_{\mathrm{J}} p) \\ \boldsymbol{0}_p & (t_i \leq c_{\mathrm{J}} p) \end{cases},$$

where $c_{\mathrm{J}} = (n-k-1)(p-2)/\{p(n-k-p+2)\}$. Hence, the weight function for this optimization is obtained as

$$w^{(\mathrm{J})}(t_i) = \begin{cases} 1 - \dfrac{c_{\mathrm{J}} p}{t_i} & (t_i > c_{\mathrm{J}} p) \\ 0 & (t_i \leq c_{\mathrm{J}} p) \end{cases}.$$

Since $w^{(\mathrm{J})}(t_i) = d_i/(d_i + \hat{\theta}_i^{(\mathrm{J})})$, we have

$$\hat{\theta}_i^{(\mathrm{J})} = \begin{cases} \dfrac{c_{\mathrm{J}} p d_i}{t_i - c_{\mathrm{J}} p} & (t_i > c_{\mathrm{J}} p) \\ \infty & (t_i \leq c_{\mathrm{J}} p) \end{cases}, \qquad (i = 1, \dots, k).$$

From (36), we can see that $\hat{\theta}_i^{(\mathrm{J})} = \hat{\theta}_i^{(\mathrm{G})}(c_{\mathrm{J}})$ holds. This implies that $\hat{\theta}_i^{(\mathrm{J})}$ is also obtained by minimizing $GC_p(\boldsymbol{\theta}|c_{\mathrm{J}})$. This optimization method is referred to as JS.

4.2. Another method. In the case of $p = 1$, there is a method for optimizing $\boldsymbol{\theta}$ which does not correspond to either a plug-in method or an IC-based method. Such a method was proposed by Lott (1973). By extending this method to the multivariate case, we obtain the following optimal $\boldsymbol{\theta}$:

$$\hat{\theta}_i^{(\mathrm{P})} = \begin{cases} 0 & (t_i > 2p) \\ \infty & (t_i \leq 2p) \end{cases}, \qquad (i = 1, \dots, k),$$

and the weight function $w^{(\mathrm{P})}(t_i)$ corresponding to $\hat{\theta}_i^{(\mathrm{P})}$ is given by

$$w^{(\mathrm{P})}(t_i) = \begin{cases} 1 & (t_i > 2p) \\ 0 & (t_i \leq 2p) \end{cases}.$$

According to the notation in Lawless (1981), this optimization method is referred to as PC (principal component).

5. Properties of the optimized ridge parameters

5.1. Relationship with hypothesis testing. Sometimes, an estimate of the MGR estimator of γ_i becomes $\mathbf{0}_p$ after optimizing. This result can be considered from the viewpoint that we estimate γ_i as $\mathbf{0}_p$ when the null hypothesis in the following hypothesis test is accepted:

$$H_0 : \gamma_i = \mathbf{0}_p \qquad \text{vs.} \qquad H_1 : \gamma_i \neq \mathbf{0}_p. \tag{38}$$

In this subsection, we discuss the relationship between each method for optimizing $\boldsymbol{\theta}$ and the hypothesis test of (38). Since $\mathrm{Cov}[\hat{\gamma}_i] = \boldsymbol{\Sigma}/d_i$, the test statistic for (38) is t_i in (30). Suppose that $\mathscr{E} \sim N_{n\times p}(\boldsymbol{O}_{n,p}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Then the test statistic t_i is distributed according to the Hotelling's T^2 distribution with p and $n-k-1$ degrees of freedom when the null hypothesis H_0 is true (see e.g., Siotani, Hayakawa and Fujikoshi, 1985, p. 190). For the PI_∞, C_p, MC_p, JS and PC methods, the MGR estimators of γ_i with the optimized ridge parameters become $\mathbf{0}_p$ if the test statistic t_i is smaller than a threshold value a, i.e., $4p$, p, $c_{\mathrm{M}}p$, $c_{\mathrm{J}}p$ and $2p$, respectively. This indicates that the MGR estimator with the optimized ridge parameter becomes $\mathbf{0}_p$ when the hypothesis H_0 is accepted. The significance level of the above test is determined by the particular threshold value a. When the hypothesis H_0 is rejected, the MGR estimators with the ridge parameter optimized by PI_∞, C_p, MC_p and JS methods are shrinkage estimators of the ordinary LS estimator of $\boldsymbol{\Gamma}$. These shrinkage ratios become small as t_i increases and eventually approach 1. On the other hand, the PC method does not shrink the ordinary LS estimator of $\boldsymbol{\Gamma}$ even when the hypothesis H_0 is rejected. The PI and PI_2 methods do not result in the MGR estimators with the optimized ridge parameters becoming $\mathbf{0}_p$. The MGR estimators with the ridge parameters optimized by the PI and PI_2 methods are always shrinkage estimators of the ordinary LS estimator of $\boldsymbol{\Gamma}$. These shrinkage ratios also become small as t_i increases and eventually approach 1. The relations between hypothesis testing and estimation are shown in Table 1.

Table 2 shows the significance levels $P(t_i > a)$ with $a = 4p$ (PI_∞), p (C_p), $c_{\mathrm{M}}p$ (MC_p), $c_{\mathrm{J}}p$ (JS) and $2p$ (PC) when $(k,n) = (5,20), (5,50), (10,20), (10,50)$ and $p = 3$. From Table 2, we can see that the significance level of PI_∞ is the smallest among the five methods in all cases. This means that the PI_∞ method most frequently makes the MGR estimator with the optimized ridge parameter into $\mathbf{0}_p$. We note that the significance level of the JS method is greater than that of the C_p method and that the significance level of the C_p method is greater than that of the MC_p method.

Table 1. Relationship between hypothesis testing and shrinkage of the estimator

Method	a	H_0 is rejected	H_0 is accepted
PI, PI$_2$	—	shrinking $\hat{\gamma}_i$	shrinking $\hat{\gamma}_i$
PI$_\infty$	$4p$	shrinking $\hat{\gamma}_i$	$\mathbf{0}_p$
C_p	p	shrinking $\hat{\gamma}_i$	$\mathbf{0}_p$
MC_p	$c_{\mathrm{M}}p$	shrinking $\hat{\gamma}_i$	$\mathbf{0}_p$
JS	$c_{\mathrm{J}}p$	shrinking $\hat{\gamma}_i$	$\mathbf{0}_p$
PC	$2p$	$\hat{\gamma}_i$	$\mathbf{0}_p$

Table 2. The significance levels in several cases

k	n	PI$_\infty$	C_p	MC_p	JS	PC
5	20	0.0524	0.4895	0.3515	0.8348	0.2170
	50	0.0166	0.4231	0.3805	0.8121	0.1428
10	20	0.0978	0.5426	0.3204	0.8526	0.2832
	50	0.0181	0.4271	0.3790	0.8135	0.1470

5.2. Magnitude relations among optimized $\boldsymbol{\theta}$'s. In this subsection, we obtain magnitude relations among $\boldsymbol{\theta}$ optimized by each method.

It follows from (33) that $\hat{\theta}_i^{[s]} > 0$, $(s = 1, 2, \ldots)$, because $\hat{\theta}_i^{[1]} > 0$. When $s = 2$, we have

$$\hat{\theta}_i^{[2]} = \left(1 + \frac{\hat{\theta}_i^{[1]}}{d_i}\right)^2 \hat{\theta}_i^{[1]} > \hat{\theta}_i^{[1]}.$$

Suppose that $\hat{\theta}_i^{[m]} > \hat{\theta}_i^{[m-1]}$ is satisfied. Then, we derive

$$\hat{\theta}_i^{[m+1]} = \left(1 + \frac{\hat{\theta}_i^{[m]}}{d_i}\right)^2 \hat{\theta}_i^{[1]} > \left(1 + \frac{\hat{\theta}_i^{[m-1]}}{d_i}\right)^2 \hat{\theta}_i^{[1]} = \hat{\theta}_i^{[m]}.$$

Consequently, by mathematical induction, we obtain the following theorem:

Theorem 1. *The following relationships among the optimized $\boldsymbol{\theta}$ always hold:*

$$0 < \hat{\theta}_i^{[1]} < \hat{\theta}_i^{[2]} < \cdots < \hat{\theta}_i^{[\infty]}, \qquad (i = 1, \ldots, k). \tag{39}$$

For $\boldsymbol{\theta}$ optimized by the IC-based method, we obtain the following theorem from (36):

Theorem 2. *When $\lambda_1 < \lambda_2$ holds, the optimized value of $\boldsymbol{\theta}$ always satisfies:*

$$\hat{\theta}_i^{(G)}(\lambda_1) \le \hat{\theta}_i^{(G)}(\lambda_2), \qquad (i = 1, \ldots, k), \tag{40}$$

with equality if and only if $t_i \le \lambda_1 p$.

From theorem 2, we have

$$\hat{\theta}_i^{(\mathrm{C})} \leq \hat{\theta}_i^{(\mathrm{M})}, \qquad \hat{\theta}_i^{(\mathrm{J})} \leq \hat{\theta}_i^{(\mathrm{M})}, \qquad (i = 1, \dots, k),$$

because $1 < c_{\mathrm{M}}$ and $c_{\mathrm{J}} < c_{\mathrm{M}}$ are satisfied. Notice that $c_{\mathrm{J}} \geq 1$ holds when $p \geq \{3 + (9 + 8(n-k-1))^{1/2}\}/2$ and $c_{\mathrm{J}} < 1$ holds when $p < \{3 + (9 + 8(n-k-1))^{1/2}\}/2$. Hence, we have

$$\begin{cases} \hat{\theta}_i^{(\mathrm{C})} \leq \hat{\theta}_i^{(\mathrm{J})} & (p \geq \{3 + \sqrt{9 + 8(n-k-1)}\}/2), \\ \hat{\theta}_i^{(\mathrm{J})} \leq \hat{\theta}_i^{(\mathrm{C})} & (p < \{3 + \sqrt{9 + 8(n-k-1)}\}/2), \end{cases} \qquad (i = 1, \dots, k).$$

The magnitude relations with $\hat{\boldsymbol{\theta}}$ optimized by the plug-in method and IC-based methods are shown as follows (the proof is given in Appendix A.3):

THEOREM 3. *The following relationships among the optimized values of $\boldsymbol{\theta}$ hold:*

$$\begin{cases} \hat{\theta}_i^{[1]} < \hat{\theta}_i^{(G)}(\lambda), & (when\ \lambda \geq 1), \\ \hat{\theta}_i^{(G)}(\lambda) \leq \hat{\theta}_i^{[\infty]}, & (when\ 0 < \lambda \leq 1), \end{cases} \qquad (i = 1, \dots, k), \qquad (41)$$

with equality if and only if $t_i \leq \lambda p$.

It follows from $\hat{\theta}_i^{(\mathrm{G})}(1) = \hat{\theta}_i^{(\mathrm{C})}$ and theorem 3 that

$$\hat{\theta}_i^{[1]} < \hat{\theta}_i^{(\mathrm{C})} \leq \hat{\theta}_i^{[\infty]}, \qquad (i = 1, \dots, k),$$

with equality if and only if $t_i \leq p$.

5.3. Magnitude relations among weight funstions. The shrinkage ratio of each method corresponds to the weight function $w(t_i)$. A method with smaller $w(t_i)$ shrinks $\hat{\gamma}_i$ to a greater extent. When $w(t_i)$ is nearly equal to one, the method shrinks $\hat{\gamma}_i$ hardly at all. Figure 1 shows the weight functions associated with each method when $(k, n) = (5, 20), (5, 50), (10, 20), (10, 50)$ and $p = 3$. From these figures, we can see that the weight function of MC_p is always smaller than those of PI, PI$_2$, C_p and JS. Thus the MC_p method always shrinks $\hat{\gamma}_i$ to a greater extent than do the PI, PI$_2$, C_p and JS methods. The weight functions of PI$_2$ and C_p are always smaller than that of PI. The weight function of PI$_\infty$ is always smaller than those of C_p, PI, PI$_2$ and PC.

The above magnitude relations among the weight functions are satisfied only when $(k, n) = (5, 20), (5, 50), (10, 20), (10, 50)$ and $p = 3$. Notice that the weight function $w(t_i) = d_i/(d_i + \hat{\theta}_i)$. Hence, we can obtain the magnitude relations among the weight functions by using theorems 1, 2 and 3. General magnitude relations among the weight functions are given by the following theorem:

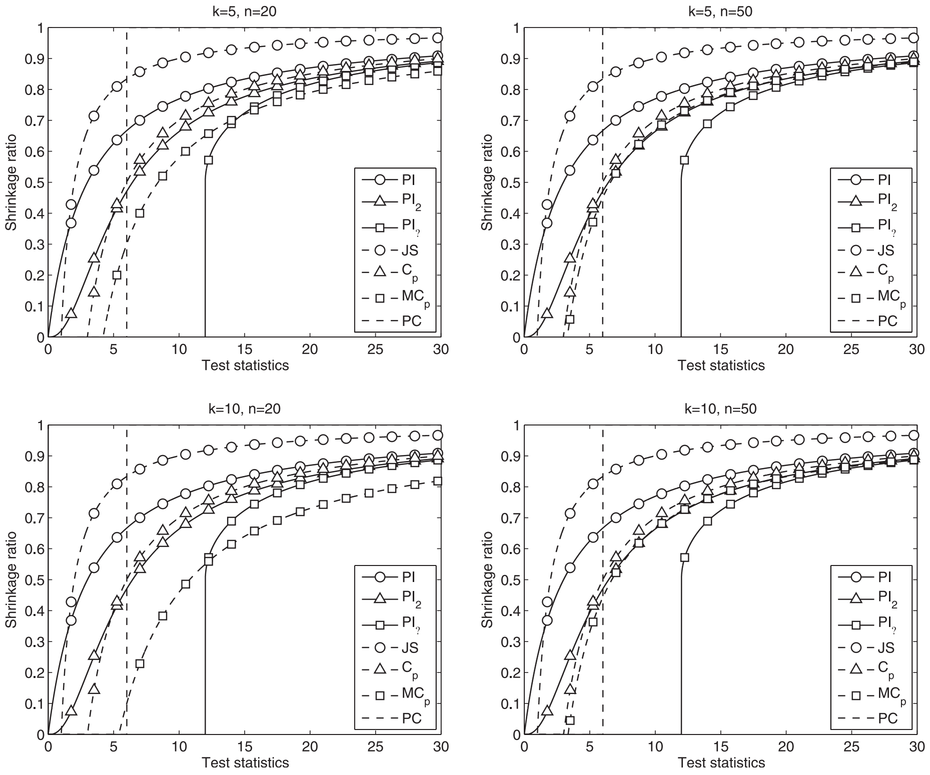

Fig. 1. Shrinkage ratio (value of weight function) for each optimization method in several cases.

Theorem 4. *The following relationships among the weight functions hold:*

$$w^{[\infty]}(t_i) < \cdots < w^{[2]}(t_i) < w^{[1]}(t_i),$$

$$w^{(M)}(t_i) \le \begin{cases} w^{(J)}(t_i) \le w^{(C)}(t_i) & (p \ge \{3 + \sqrt{9 + 8(n-k-1)}\}/2), \\ w^{(C)}(t_i) \le w^{(J)}(t_i) & (p < \{3 + \sqrt{9 + 8(n-k-1)}\}/2), \end{cases}$$

$$w^{[\infty]}(t_i) \le w^{(C)}(t_i) < w^{[1]}(t_i).$$

Notice that these relationships among the methods correspond to the relationships among the significance levels of the various methods.

6. Numerical studies

In this section, we conduct numerical studies to compare the MSEs of predictors of $\boldsymbol{Y}$ consisting of the MGR estimators with the optimized ridge parameters. Let $\boldsymbol{R}_q$ and $\boldsymbol{\Delta}_q(\rho)$ be $q \times q$ matrices defined by

$$\boldsymbol{R}_q = \mathrm{diag}(1, \dots, q), \qquad \varDelta_q(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{q-1} \\ \rho & 1 & \rho & \cdots & \rho^{q-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{q-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{q-1} & \rho^{q-2} & \rho^{q-3} & \cdots & 1 \end{pmatrix}.$$

The explanatory matrix $\boldsymbol{X}$ was generated from $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{\varPsi}^{1/2}$ where $\boldsymbol{\varPsi} = \boldsymbol{R}_k^{1/2}\varDelta_k(\rho_x)\boldsymbol{R}_k^{1/2}$ and $\boldsymbol{W}$ is an $n \times k$ matrix whose elements were generated independently from the uniform distribution on $(-1, 1)$. The $k \times p$ unknown regression coefficient matrix $\boldsymbol{\varXi}$ was defined by $\boldsymbol{\varXi} = \delta\boldsymbol{F}\boldsymbol{\varXi}_0$, where δ is a constant, and $\boldsymbol{F}$ and $\boldsymbol{\varXi}$ are defined as

$$\boldsymbol{F} = \begin{pmatrix} \boldsymbol{I}_\kappa & \boldsymbol{O}_{\kappa, 10-\kappa} \\ \boldsymbol{O}_{k-\kappa} & \boldsymbol{O}_{k-\kappa, 10-\kappa} \end{pmatrix}, \qquad \boldsymbol{\varXi}_0 = \begin{pmatrix} 0.8501 & 0.6571 & 0.2159 \\ -0.2753 & -0.2432 & -0.1187 \\ -0.3193 & -0.2926 & -0.1671 \\ 0.2754 & 0.2608 & 0.1766 \\ 0.2693 & 0.2164 & 0.2066 \\ -0.0676 & -0.0663 & -0.0561 \\ 0.2239 & 0.2197 & 0.1880 \\ -0.0352 & -0.0346 & -0.0305 \\ 0.3240 & 0.3199 & 0.2868 \\ -0.3747 & -0.3727 & -0.3554 \end{pmatrix}.$$

Here δ controls the scale of the regression coefficient matrix and $\boldsymbol{F}$ controls the number of non-zero regression coefficients via κ (dimension of the true model). Values of elements of $\boldsymbol{\varXi}_0$, which is an essential regression coefficient matrix, are the same as in Lawless (1981). Simulated data values $\boldsymbol{Y}$ were generated by $N_{n\times 3}(\boldsymbol{X}\boldsymbol{\varXi}, \boldsymbol{\varSigma} \otimes \boldsymbol{I}_n)$ repeatedly under several selections of n, k, κ, δ and ρ_x, where $\boldsymbol{\varSigma} = \boldsymbol{R}_3^{1/2}\varDelta_3(0.8)\boldsymbol{R}_3^{1/2}$ and the number of repetition was $10,000$. At each repetition, we evaluated $r(\boldsymbol{X}\boldsymbol{\varXi}, \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}})$, where $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{1}_n\bar{\boldsymbol{y}}' + \boldsymbol{X}\hat{\boldsymbol{\varXi}}_{\hat{\boldsymbol{\theta}}}$ which is the predicted value of $\boldsymbol{Y}$ obtained from each method. The average of $r(\boldsymbol{X}\boldsymbol{\varXi}, \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}})$ across $10,000$ repetition was regarded as the MSE of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}}$. In the simulation, a standardized $\boldsymbol{X}$ was used for estimating regression coefficients. Tables 3, 4, 5 and 6 depict $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}}]/\{3(k+1)\} \times 100$ in the case of $(k, n) = (5, 20), (5, 50), (10, 20)$ and $(10, 50)$, respectively, where $3(k+1)$ is the MSE of the predictor of $\boldsymbol{Y}$ derived by using the LS estimator of $\boldsymbol{\varXi}$.

In tables 3, 4, 5 and 6, we observe that the method can improve the LS estimation when values in the tables do not exceed 100. In each table, the average of $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}}]/\{3(k+1)\} \times 100$ across all cases is also depicted in the bottom line of the table. From the tables, we can see that all methods improve the ordinary LS method in almost all cases. The PI_2 method

Table 3. MSE of each method ($k = 5$, $n = 20$)

κ	δ	ρ_x	PI	PI_2	PI_∞	C_p	MC_p	JS	PC
0	0.0	0.2	50.84	36.51	**23.08**	37.42	29.72	66.34	53.80
		0.8	50.91	36.59	**23.25**	37.47	29.82	66.38	53.56
		0.99	51.03	36.75	**23.43**	37.65	30.00	66.48	53.98
3	1.0	0.2	67.30	**62.39**	69.92	65.59	62.74	80.68	92.45
		0.8	57.06	46.37	**40.11**	48.28	42.30	71.90	68.89
		0.99	51.81	37.91	**25.15**	38.94	31.43	67.28	55.82
	3.0	0.2	**96.60**	103.34	148.20	103.14	110.90	97.90	113.59
		0.8	75.36	**74.01**	97.23	75.60	76.52	84.66	95.46
		0.99	56.42	45.43	**38.36**	47.29	41.13	71.37	67.53
5	1.0	0.2	74.10	**72.98**	96.22	75.55	76.16	84.56	100.71
		0.8	67.01	**62.10**	72.66	64.74	62.50	79.63	89.26
		0.99	59.84	50.45	52.24	51.48	**47.25**	72.81	69.32
	3.0	0.2	**94.69**	98.22	121.30	98.30	103.00	96.21	105.66
		0.8	**90.12**	93.12	125.76	93.21	97.92	93.66	104.29
		0.99	64.97	56.03	**49.79**	57.41	52.71	76.81	72.99
Average			67.20	60.81	67.11	62.14	**59.61**	78.44	79.82

improved on the ordinary LS method more than the PI method in almost all cases when $n = 20$. When κ is small, it is necessary to shrink the LS estimator to a greater extent. On the other hand, it is not necessary to shrink the LS estimator when κ is large. Thus PI_∞ works well when κ is small but does not work well when κ is large since κ controls the number of non-zero elements in the true regression coefficient matrix Ξ and PI_∞ has the most shrinkage of the LS estimators. These estimation methods more improve when δ is small than δ is large since δ means the scale of the true regression coefficient matrix and these estimate methods shrink the LS estimator. When ρ_x is 0.99, these estimate methods improve the LS estimator even if κ and δ are large since the LS estimator is unstable. On average, C_p was the best method in almost cases except PI_2 and MC_p. One of the reasons is that the shape of the weight function of C_p is near to that of PI_2, which is shown in Figure 1. Furthermore, because the MC_p criterion is the bias corrected C_p criterion, the results from the MC_p and C_p methods become similar when n is large. The PI and JS methods improve the ordinary LS method in almost cases although the ratios of improvement are not as great. On average, PI_∞ is the best method to obtain stable estimator except $(k, n) = (10, 20)$. When $(k, n) = (10, 20)$, MC_p is the best method on average to obtain stable estimator.

Table 4. MSE of each method ($k = 5$, $n = 50$)

κ	δ	ρ_x	PI	PI$_2$	PI$_\infty$	C_p	MC_p	JS	PC
0	0.0	0.2	47.29	32.07	**19.19**	32.57	30.34	62.83	45.92
		0.8	47.61	32.39	**19.36**	32.92	30.67	63.18	46.59
		0.99	47.45	32.29	**19.47**	32.80	30.57	62.92	46.10
3	1.0	0.2	72.23	**70.58**	98.32	71.75	71.91	81.26	90.43
		0.8	60.71	**53.64**	61.60	55.78	54.80	73.95	77.87
		0.99	49.18	35.14	**24.04**	36.08	33.98	64.64	51.12
	3.0	0.2	83.96	**81.75**	83.95	82.55	82.42	89.09	91.32
		0.8	**82.78**	84.85	119.82	86.15	86.97	88.85	104.87
		0.99	60.54	**53.46**	58.64	56.16	55.09	74.47	80.63
5	1.0	0.2	**80.15**	81.29	114.58	82.15	82.80	86.98	99.23
		0.8	71.57	**69.36**	96.08	70.84	70.83	81.53	91.18
		0.99	59.20	48.89	**45.25**	49.25	47.84	71.47	61.47
	3.0	0.2	91.53	90.87	99.02	**90.73**	90.93	94.01	94.40
		0.8	**87.83**	88.93	115.44	88.97	89.63	91.33	98.13
		0.99	66.36	**59.40**	61.12	61.10	60.12	77.51	78.26
Average			67.23	**60.99**	69.06	61.99	61.26	77.60	77.17

A. Appendix

A.1. The proof of equation (34). In this subsection, we show that the $\hat{\theta}_i^{[s]}$ in (32) converges to $\hat{\theta}_i^{[\infty]}$ in (34) as $s \to \infty$ by extending the technique in Hemmerle (1975).

Theorem 1 shows that $\{\hat{\theta}_i^{[s]}\}$ is a monotonic increasing sequence. If $\hat{\theta}_i^{[s]}$ is bounded above, $\hat{\theta}_i^{[s]}$ surely converges. To prove the convergence, we prove the following lemma:

LEMMA 1. *Let a_1 be a positive number. Define a sequence of real numbers by*

$$a_{s+1} = (1 + a_s)^2 a_1, \qquad (s = 1, 2, \dots).$$

Then a_s converges to some number if and only if $a_1 \le 1/4$. If $a_1 \le 1/4$, we obtain

$$\lim_{s \to \infty} a_s = \frac{1}{2a_1} - 1 - \frac{\sqrt{1 - 4a_1}}{2a_1}. \tag{42}$$

Table 5. MSE of each method $(k = 10, n = 20)$

κ	δ	ρ_x	PI	PI$_2$	PI$_\infty$	C_p	MC_p	JS	PC
0	0.0	0.2	49.77	35.22	**21.05**	36.40	22.57	66.23	55.34
		0.8	50.27	35.80	**21.64**	36.99	23.11	66.70	55.96
		0.99	50.11	35.51	**21.29**	36.69	22.76	66.68	55.82
3	1.0	0.2	57.73	47.42	43.32	49.39	**39.69**	72.94	71.97
		0.8	53.27	40.64	30.17	42.34	**29.95**	69.36	63.52
		0.99	50.56	36.23	**22.42**	37.50	23.74	67.05	57.10
	3.0	0.2	73.25	**69.84**	82.01	71.63	71.33	83.63	92.06
		0.8	65.95	60.01	69.20	61.94	**58.46**	78.63	84.43
		0.99	53.57	41.00	30.90	42.71	**30.45**	69.70	63.87
5	1.0	0.2	60.81	52.30	54.38	54.04	**47.44**	74.61	75.55
		0.8	57.84	47.66	44.80	49.29	**40.49**	72.46	70.34
		0.99	54.96	42.70	35.20	43.71	**33.31**	69.89	61.94
	3.0	0.2	76.32	**74.00**	86.35	76.20	76.35	85.99	97.57
		0.8	69.44	63.82	71.08	65.27	**62.56**	80.46	83.88
		0.99	56.84	44.89	**34.23**	46.26	34.90	71.42	64.57
10	1.0	0.2	67.46	62.71	74.94	64.92	**62.67**	79.84	88.92
		0.8	60.78	51.81	50.95	53.62	**45.99**	74.93	74.79
		0.99	55.46	43.27	35.17	44.26	**33.86**	70.22	61.84
	3.0	0.2	**86.68**	87.82	109.28	88.68	95.67	91.73	102.46
		0.8	79.91	**79.81**	101.50	81.68	85.70	88.25	102.05
		0.99	58.76	47.86	39.12	49.51	**38.88**	73.12	69.11
Average			61.41	52.40	51.38	53.95	**46.66**	74.94	73.96

Proof. It can be easily checked that $a_{s+1} \geq a_s$ by the inductive method. Suppose that the $\alpha = \lim_{s \to \infty} a_s$ exists. Then α is one of the solutions of the following quadratic equation with respect to x;

$$x = (1+x)^2 a_1 \qquad \Leftrightarrow \qquad x^2 + 2\left(1 - \frac{1}{2a_1}\right)x + 1 = 0. \tag{43}$$

The discriminant shows (43) has real roots if and only if $a_1 \leq 1/4$.

If $a_1 = 1/4$, the equation (43) has the multiple root $\alpha = 1$. Suppose $a_1 \leq 1/4$ and $a_k \leq 1$ for some k. Then

$$a_{k+1} \leq (1+1)^2/4 = 1.$$

Hence the sequence $\{a_s\}$ is bounded by the induction. Since $\{a_s\}$ is an

Table 6. MSE of each method ($k = 10$, $n = 50$)

κ	δ	ρ_x	PI	PI$_2$	PI$_\infty$	C_p	MC_p	JS	PC
0	0.0	0.2	42.87	26.28	**12.11**	26.88	24.10	59.87	41.73
		0.8	42.68	26.17	**12.10**	26.74	23.99	59.61	41.42
		0.99	42.99	26.49	**12.40**	27.09	24.32	59.93	42.03
3	1.0	0.2	63.01	**58.23**	76.41	60.78	60.16	76.18	86.79
		0.8	50.32	38.31	**35.46**	39.91	37.91	66.16	59.75
		0.99	43.62	27.86	**14.99**	28.69	26.04	60.40	44.43
	3.0	0.2	**84.48**	85.35	107.83	86.11	86.78	89.75	99.43
		0.8	67.10	**62.97**	79.81	65.18	64.72	78.80	88.53
		0.99	50.32	38.50	**35.38**	40.26	38.29	66.20	61.03
5	1.0	0.2	69.15	**67.09**	92.34	69.52	69.48	80.63	95.30
		0.8	56.52	47.06	49.43	48.49	**47.03**	70.50	67.56
		0.99	49.48	35.26	**23.94**	35.79	33.46	64.67	49.55
	3.0	0.2	**91.12**	94.28	123.88	94.49	95.72	93.81	105.41
		0.8	69.83	**64.95**	68.86	66.82	66.14	80.19	84.84
		0.99	53.86	42.07	**34.85**	43.63	41.58	68.77	61.46
10	1.0	0.2	**79.11**	81.30	112.51	84.22	84.96	88.14	111.74
		0.8	63.54	**57.32**	68.44	59.06	58.23	75.96	79.99
		0.99	50.43	36.79	**25.88**	37.60	35.34	65.54	52.49
	3.0	0.2	99.63	103.28	121.89	102.29	103.58	**98.83**	102.91
		0.8	**82.53**	83.03	101.04	85.10	85.48	89.54	104.24
		0.99	59.89	**51.95**	55.35	54.25	52.94	73.63	76.64
Average			62.50	**54.98**	60.23	56.33	55.25	74.62	74.16

increasing sequence, it has the limiting value which is not greater than 1. If $a_1 \leq 1/4$, the equation (43) has the roots

$$\alpha_1 = \frac{1}{2a_1} - 1 - \frac{\sqrt{1-4a_1}}{2a_1} \quad \text{and} \quad \alpha_2 = \frac{1}{2a_1} - 1 + \frac{\sqrt{1-4a_1}}{2a_1}.$$

Since $\alpha_2 = \alpha_2(a_1)$ is a strict decreasiong function of a_1, $\alpha_2(a_1) > \alpha_2(1/4) = 1$ for any $a_1 < 1/4$. Hence the limiting value is α_1 because it can not be greater than 1.

We consider a_s in the above lemma as $\hat{\theta}_i^{[s]}/d_i$. Then $a_1 \leq 1$ when $t_i \geq 4p$. By using this lemma, we obtain the $\hat{\theta}_i^{[\infty]}$ when $t_i \geq 4p$. On the other hand, from Theorem 1, we note $\{a_s\}$ is also monotone increasing sequence. Hence, if $t_i < 4p$ holds, $\lim_{s\to\infty} \hat{\theta}_i^{[s]} = \infty$ is satisfied.

A.2. The proof of equation (36). From (4), the second part of $GC_p(\boldsymbol{\theta}|\lambda)$ in (35) can be rewritten as

$$\mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{X}\} = \mathrm{tr}\{(\boldsymbol{D}+\boldsymbol{\Theta})^{-1}\boldsymbol{D}\} = \sum_{i=1}^{k}\frac{d_i}{d_i+\theta_i}. \tag{44}$$

Moreover, from (12) and (17), the first part of $GC_p(\boldsymbol{\theta}|\lambda)$ can be rewritten as

$$\begin{aligned}\hat{r}(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}) &= \mathrm{tr}\{(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})\boldsymbol{S}^{-1}(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})'\}\\ &= \mathrm{tr}\{\boldsymbol{P}_1(\boldsymbol{Z}-\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})\boldsymbol{S}^{-1}(\boldsymbol{Z}-\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})'\boldsymbol{P}_1'\} = \hat{r}(\boldsymbol{Z},\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}). \end{aligned}\tag{45}$$

By using (17) and (20), we have

$$\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{L}\begin{pmatrix}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}\\ \hat{\boldsymbol{\mu}}'\end{pmatrix} = \begin{pmatrix}\boldsymbol{D}^{1/2}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}\\ \sqrt{n}\hat{\boldsymbol{\mu}}'\\ \boldsymbol{O}_{n-k-1,p}\end{pmatrix}. \tag{46}$$

Notice that $\hat{\boldsymbol{\mu}} = \boldsymbol{z}_{k+1}/\sqrt{n}$ and $\boldsymbol{z}_i - \{d_i/(d_i+\theta_i)\}\boldsymbol{z}_i = \{\theta_i/(d_i+\theta_i)\}\boldsymbol{z}_i$. Substituting (21) and (46) into (45) yields

$$\hat{r}(\boldsymbol{Z},\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}) = \sum_{i=1}^{k}\left(\frac{\theta_i}{d_i+\theta_i}\right)^2 t_i + \sum_{i=k+2}^{n}\boldsymbol{z}_i'\boldsymbol{S}^{-1}\boldsymbol{z}_i, \tag{47}$$

where t_i is given by (29) or (30). Let $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Z}}$ be $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} = \boldsymbol{0}_k$, respectively. Then, from similar calculations with (45) and (46), we derive

$$(\boldsymbol{Y}-\hat{\boldsymbol{Y}})'(\boldsymbol{Y}-\hat{\boldsymbol{Y}}) = (\boldsymbol{Z}-\hat{\boldsymbol{Z}})'(\boldsymbol{Z}-\hat{\boldsymbol{Z}}) = \sum_{i=k+2}^{n}\boldsymbol{z}_i\boldsymbol{z}_i'.$$

This equation implies that $(n-k-1)\boldsymbol{S} = \sum_{i=k+2}^{n}\boldsymbol{z}_i\boldsymbol{z}_i'$. Consequently, by using this result, (44), (45) and (47), $GC_p(\boldsymbol{\theta}|\lambda)$ can be rewritten as

$$GC_p(\boldsymbol{\theta}|\lambda) = \sum_{i=1}^{k} f(\theta_i|d_i,t_i,\lambda) + \lambda^{-1}p(n-k-1), \tag{48}$$

where the function $f(\theta_i|d_i,t_i,\lambda)$ is defined by

$$f(\theta_i|d_i,t_i,\lambda) = \lambda^{-1}\left(\frac{\theta_i}{d_i+\theta_i}\right)^2 t_i + \frac{2pd_i}{d_i+\theta_i}, \qquad (i=1,\dots,k).$$

Hence in order to obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda) = (\hat{\theta}_1^{(\mathrm{G})}(\lambda),\dots,\hat{\theta}_k^{(\mathrm{G})}(\lambda))'$, $(\hat{\theta}_i^{(\mathrm{G})}(\lambda) \geq 0$, $i=1,\dots,k)$ making $GC_p(\boldsymbol{\theta}|\lambda)$ the minimum, we can see that it is necessary only to minimize $f(\theta_i|d_i,t_i,\lambda)$ individually. The first partial derivative of $f(\theta_i|d_i,t_i,\lambda)$ with respect to θ_i is calculated as

$$\frac{\partial}{\partial\theta_i}f(\theta_i|d_i,t_i,\lambda)=\frac{2d_i}{\lambda(d_i+\theta_i)^3}\{\theta_i(t_i-\lambda p)-\lambda pd_i\}.$$

This derivative indicates that $f(\theta_i|d_i,t_i,\lambda)$ becomes a minimum at $\theta_i=\lambda pd_i/(t_i-\lambda p)$ when $t_i-\lambda p>0$ holds. On the other hand, $f(\theta_i|d_i,t_i,\lambda)$ is a monotonic decreasing function of θ_i when $t_i-\lambda p\leq 0$ holds. Thus, $f(\theta_i|d_i,t_i,\lambda)$ converges to the minimum value as $\theta_i\to\infty$ when $t_i-\lambda p\leq 0$ holds. Consequently, from the above two results, the equation (36) follows.

A.3. The proof of equation (41). Firstly, we show the proof of the first inequality of equation (41). It is easy to obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda)>\hat{\theta}_i^{[1]}$ when $t_i\leq\lambda p$, because $\hat{\theta}_i^{(\mathrm{G})}(\lambda)=\infty$ and $\hat{\theta}_i^{[1]}<\infty$ are satisfied when $t_i\leq\lambda p$. When $t_i>\lambda p$, from (31) and (36), we can see that

$$\hat{\theta}_i^{(\mathrm{G})}(\lambda)-\hat{\theta}_i^{[1]}=\frac{d_ip\{(\lambda-1)t_i+\lambda p\}}{t_i(t_i-\lambda p)}.$$

Since $t_i>0$ holds, the right side of the above equation becomes positive when $\lambda\geq 1$. Thus, $\hat{\theta}_i^{(\mathrm{G})}(\lambda)>\hat{\theta}_i^{[1]}$ holds when $\lambda\geq 1$.

Next, we show the proof of the second inequality of equation (41). Suppose that $0<\lambda\leq 1$. It is easy to obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda)\leq\hat{\theta}_i^{[\infty]}$ when $t_i\leq 4p$, because $\hat{\theta}_i^{[\infty]}=\infty$ and $\hat{\theta}_i^{(\mathrm{G})}(\lambda)\leq\infty$ are satisfied when $t_i\leq 4p$. Notice that

$$\left(1-\frac{2p}{t_i-p}\right)^2-\left(1-\frac{4p}{t_i}\right)=\frac{4p^3}{t_i(t_i-p)^2}>0.$$

The above equation and the inequality $t_i-p\leq t_i-\lambda p$ imply that

$$1-\frac{4p}{t_i}<\left(1-\frac{2p}{t_i-p}\right)^2<\left(1-\frac{2p}{t_i-\lambda p}\right)^2. \tag{49}$$

Since $t_i\geq 4p$ is assumed, we obtain $1-2p/(t_i-p)=(t_i-3p)/(t_i-p)>0$. Hence, $1-2p/(t_i-\lambda p)>0$ can also be derived. It follows from this result and the inequality (49) that

$$\sqrt{1-\frac{4p}{t_i}}<1-\frac{2p}{t_i-\lambda p}. \tag{50}$$

By multiplying both sides of (50) by t_i after calculation, we have

$$\frac{t_i}{t_i-\lambda p}<\frac{t_i-\sqrt{t_i(t_i-4p)}}{2p}. \tag{51}$$

Subtracting 1 from both sides of (51) yields

$$\frac{\lambda p}{t_i - \lambda p} < \frac{t_i - 2p - \sqrt{t_i(t_i - 4p)}}{2p}. \tag{52}$$

Thus, when $t_i > 4p$, $\hat{\theta}_i^{(G)}(\lambda) < \hat{\theta}_i^{[\infty]}$ can be derived by multiplying both sides of (52) by d_i. Consequently, $\hat{\theta}_i^{(G)}(\lambda) \leq \hat{\theta}_i^{[\infty]}$ is obtained when $0 < \lambda \leq 1$.

Acknowledgement

A part of this research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007–2010. The authors would like to express his gratitude to the referee for his kind criticisms and suggestions.

References

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

[2] S. J. V. Dien, S. Iwatani, Y. Usuda and K. Matsui, Theoretical analysis of amino acid-producing Eschenrichia coli using a stoixhiometrix model and multivariate linear regression, J. Biosci. Bioeng., **102** (2006), 34–40.

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

[4] M. Goldstein and A. F. M. Smith, Ridge-type estimators for regression analysis, J. Roy. Statist. Soc. Ser. B, **36** (1974), 284–291.

[5] W. J. Hemmerle, An explicit solution for generalized ridge regression, Technometrics, **17** (1975), 309–314.

[6] A. E. Hoerl and R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, **12** (1970), 55–67.

[7] T. Kubokawa, An approach to improving the James-Stein estimator, J. Multivariate Anal., **36** (1991), 121–126.

[8] J. F. Lawless, Mean squared error properties of generalized ridge estimators, J. Amer. Statist. Assoc., **76** (1981), 462–466.

[9] W. F. Lott, The optimal set of principal component restrictions on a least squares regression, Comm. Statist., **2** (1973), 449–464.

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

[11] C. L. Mallows, More comments on C_p, Technometrics, **37** (1995), 362–372.

[12] C. Sârbu, C. Onisor, M. Posa, S. Kevresan and K. Kuhajda, Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods, Talanta, **75** (2008), 651–657.

[13] R. Saxén and J. Sundell, ^{137}Cs in freshwater fish in Finland since 1986—a statistical analysis with multivariate linear regression models, J. Environ. Radioactiv., **87** (2006), 62–76.

[14] M. Siotani, T. Hayakawa, and Y. Fujikoshi, Modern Multivariate Statistical Analysis: A Graduate Course and Handbook, American Sciences Press, Columbus, Ohio, 1985.

[15] B. Skagerberg, J. MacGregor and C. Kiparissides, Multivariate data analysis applied to low-density polyethylene reactors, Chemometr. Intell. Lab. Syst., **14** (1992), 341–356.

[16] R. S. Sparks, D. Coutsourides and L. Troskie, The multivariate C_p, Comm. Statist. A—Theory Methods, **12** (1983), 1775–1793.

[17] M. S. Srivastava, Methods of Multivariate Statistics, John Wiley & Sons, New York, 2002.

[18] N. H. Timm, Applied Multivariate Analysis, Springer-Verlag, New York, 2002.

[19] S. G. Walker and C. J. Page, Generalized ridge regression and a generalization of the C_p statistics, J. Appl. Statist., **28** (2001), 911–922.

[20] H. Yanagihara and K. Satoh, An unbiased C_p criterion for multivariate ridge regression, J. Multivariate Anal., **101** (2010), 1226–1238.

[21] H. Yanagihara, I. Nagai and K. Satoh, A bias-corrected C_p criterion for optimizing ridge parameters in multivariate generalized ridge regression, Japanese J. Appl. Statist., **38** (2009), 151–172 (in Japanese).

[22] A. Yoshimoto, H. Yanagihara and Y. Ninomiya, Finding factors affecting a forest stand growth through multivariate linear modeling, J. Jpn. For. Res., **87** (2005), 504–512 (in Japanese).

Isamu Nagai
Department of Mathematics
Graduate School of Science
Hiroshima University
Higashi-Hiroshima 739-8526, *Japan*
E-mail: inagai@hiroshima-u.ac.jp
URL: http://home.hiroshima-u.ac.jp/inagai/

Hirokazu Yanagihara
Department of Mathematics
Graduate School of Science
Hiroshima University
Higashi-Hiroshima 739-8526, *Japan*
E-mail: yanagi@math.sci.hiroshima-u.ac.jp

Kenichi Satoh
Department of Environmetrics and Biometrics
Research Institute for Radiation Biology and Medicine
Hiroshima University
Hiroshima 734-8553, *Japan*
E-mail: ksatoh@hiroshima-u.ac.jp