Optimization of Ridge Parameters in Multivariate Generalized Ridge Regression by Plug-in Methods

(Last Modified: February 14, 2010)

Isamu Nagai[1*], Hirokazu Yanagihara[1] and Kenichi Satoh[2]

[1] *Department of Mathematics, Graduate School of Science, Hiroshima University*
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

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

Abstract

Generalized ridge (GR) regression for a 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 square error (MSE) of a predicted value, many authors have proposed 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 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 illustration.

AMS 2000 *subject classifications*: Primary 62J07; Secondary 62F07.

Key words: Generalized ridge regression; Mallows' C_p statistic; Model selection; Multivariate linear regression model; Non-iterative estimation; Plug-in method.

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

[*] Corresponding author, E-mail: *d093481@hiroshima-u.ac.jp*

$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$, $\boldsymbol{X}$ and $\boldsymbol{\mathcal{E}}$ be the $n \times p$ matrix of response variables, the $n \times k$ matrix of non-stochastic standardized explanatory variables ($\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$) 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 $\boldsymbol{\mathcal{E}}$ are independently and identically distributed according to a distribution with mean $\boldsymbol{0}_p$ and unknown covariance matrix $\boldsymbol{\Sigma}$. The matrix form of the multivariate linear regression model is expressed as

$$\boldsymbol{Y} = \boldsymbol{1}_n \boldsymbol{\mu}' + \boldsymbol{X}\Xi + \boldsymbol{\mathcal{E}}, \tag{1.1}$$

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

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

$$\hat{\Xi} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{1.2}$$

respectively. For simplicity, and because $\hat{\Xi}$ is unbiased, it is widely used in actual data analysis, see e.g., Dien *et al.* (2006), Sárbu *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 Ξ is not a good estimator in the sense of having a large variance. The ridge regression for a 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 the ridge parameter. Since estimates of a ridge estimator depends heavily on the value of θ, optimization of θ is a very important problem. Choosing θ so that the mean square error (MSE) of a predictor of $\boldsymbol{Y}$ becomes small is a common procedure. However, an optimal value of θ cannot be obtained without an iterative computational algorithm.

However, 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, ($\theta_i \geq 0$, $i = 1, \ldots, k$). 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 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 GR regression is to be able to obtain a shrinkage estimate for regression coefficients without the use of an iterative optimization algorithm on $\boldsymbol{\theta}$. It also has other advantages, namely, whereas ridge regression shrinks uniformly all coefficients of the LS estimator by a single ridge parameter, for GR regression, the amount of shrinkage is different for each explanatory variable. Thus GR regression is more flexible than ridge regression. From this viewpoint, we deal not with ridge regression but GR regression. We refer to GR regression for a multivariate linear model as a multivariate GR (MGR) regression.

Methods for optimizing $\boldsymbol{\theta}$ in 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 optimization methods in GR regression, formulas for obtaining optimal $\boldsymbol{\theta}$ in the MGR regression will be derived.

By extending the formulas for a GR estimator with 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 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 (2009) as special cases. In this paper, we consider the generalized C_p (GC_p) criterion proposed by Atkinson (1980) for 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 MGR estimators with 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 univariate GR regression to 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 MGR estimators with optimized ridge parameters. In Section 4, we discuss relationships between test statistics and optimized values of $\boldsymbol{\theta}$, and give the magnitude relation among optimized $\boldsymbol{\theta}$s. In Section 5, we compare derived MGR estimators with existing MGR estimators by conducting numerical studies. Technical details are provided in an Appendix.

2. MGR Estimator and Target MSE

2.1. Preliminaries

By naturally extending the GR estimator, we derive the MGR estimator for (1.1) as

$$\hat{\Xi}_{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\Theta\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{2.1}$$

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.,

$$\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{Q} = \mathrm{diag}(d_1, \dots, d_k) = \boldsymbol{D}. \tag{2.2}$$

Here $d_1, \dots, d_k$ are eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and we note that the d_i are always positive. We can check that the estimator in (2.1) corresponds to the ordinary LS estimator in (1.2) when $\boldsymbol{\theta} = \boldsymbol{0}_k$. This means that the estimator in (2.1) includes the ordinary LS estimator. If $p = 1$, then the estimator in (2.1) 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 $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}} = \boldsymbol{1}_n\bar{\boldsymbol{y}}' + \boldsymbol{X}\hat{\Xi}_{\boldsymbol{\theta}}$. 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 $\boldsymbol{B}$:

$$r(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\left\{(\boldsymbol{A} - \boldsymbol{B})\,\boldsymbol{\Sigma}^{-1}\,(\boldsymbol{A} - \boldsymbol{B})'\right\}. \tag{2.3}$$

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

$$\boldsymbol{S} = \frac{1}{n-k-1}(\boldsymbol{Y} - \boldsymbol{1}_n\bar{\boldsymbol{y}}' - \boldsymbol{X}\hat{\boldsymbol{\Xi}})'(\boldsymbol{Y} - \boldsymbol{1}_n\bar{\boldsymbol{y}}' - \boldsymbol{X}\hat{\boldsymbol{\Xi}}), \tag{2.4}$$

where $\hat{\boldsymbol{\Xi}}$ is given in (1.2). By replacing $\boldsymbol{\Sigma}$ with (2.4), we can estimate (2.3) by

$$\hat{r}(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{tr}\left\{(\boldsymbol{A} - \boldsymbol{B})\,\boldsymbol{S}^{-1}\,(\boldsymbol{A} - \boldsymbol{B})'\right\}. \tag{2.5}$$

These two functions in (2.3) and (2.5) correspond to summations of the Mahalanobis distances and the sample Mahalanobis distances between rows of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. By using (2.3), the MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}$ is defined as

$$\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}] = E[r(E[\boldsymbol{Y}], \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})]. \tag{2.6}$$

In this paper, we choose $\boldsymbol{\theta}$ that minimizes the MSE in (2.6) as the principal optimum.

2.2. Model Transformation

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

$$(\boldsymbol{X}, \boldsymbol{1}_n) = \boldsymbol{P}_1\boldsymbol{L}\boldsymbol{P}_2', \tag{2.7}$$

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

$$(\boldsymbol{X}, \boldsymbol{1}_n)'(\boldsymbol{X}, \boldsymbol{1}_n) = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{0}_k \\ \boldsymbol{0}_k' & n \end{pmatrix}. \tag{2.8}$$

Since the orthogonal matrix $\boldsymbol{P}_2$ diagonalizes (2.8), from (2.2), $\boldsymbol{P}_2$ and $\boldsymbol{L}$ can be expressed as

$$\boldsymbol{P}_2 = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0}_k \\ \boldsymbol{0}_k' & 1 \end{pmatrix}, \tag{2.9}$$

and

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

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

Let

$$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)' = \boldsymbol{P}_1'\boldsymbol{Y},\ \boldsymbol{\Gamma} = (\gamma_1, \ldots, \gamma_k)' = \boldsymbol{Q}'\Xi,\ \boldsymbol{\mathcal{V}} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n)' = \boldsymbol{P}_1'\boldsymbol{\mathcal{E}}. \tag{2.10}$$

By using (2.7) and (2.9), $\boldsymbol{Z}$ is calculated as

$$\boldsymbol{Z} = \boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)\begin{pmatrix} \boldsymbol{\Xi} \\ \boldsymbol{\mu}' \end{pmatrix} + \boldsymbol{P}_1'\boldsymbol{\mathcal{E}} = \boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)\boldsymbol{P}_2\begin{pmatrix} \boldsymbol{Q}'\boldsymbol{\Xi} \\ \boldsymbol{\mu}' \end{pmatrix} + \boldsymbol{\mathcal{V}} = \boldsymbol{L}\begin{pmatrix} \boldsymbol{\Gamma} \\ \boldsymbol{\mu}' \end{pmatrix} + \boldsymbol{\mathcal{V}}. \tag{2.11}$$

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 $\mathrm{Cov}[\boldsymbol{z}_i] = \boldsymbol{\Sigma}$ $(i = 1, \ldots, n)$. Thus, from this result and (2.11), the following equation is obtained:

$$\boldsymbol{z}_i = \begin{cases} \sqrt{d_i}\boldsymbol{\gamma}_i + \boldsymbol{\nu}_i & (i = 1, \ldots, k) \\ \sqrt{n}\boldsymbol{\mu} + \boldsymbol{\nu}_i & (i = k+1) \\ \boldsymbol{\nu}_i & (i = k+2, \ldots, n) \end{cases}, \ (E[\boldsymbol{\nu}_i] = \boldsymbol{0}_p,\ \mathrm{Cov}[\boldsymbol{\nu}_i] = \boldsymbol{\Sigma}). \tag{2.12}$$

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

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

$$\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{C}'\boldsymbol{Z}. \tag{2.13}$$

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:

$$\boldsymbol{Q}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{Q}\boldsymbol{C}'\boldsymbol{P}_1\boldsymbol{Y} = \hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}. \tag{2.14}$$

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

$$\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{P}_1'\boldsymbol{P}_1\boldsymbol{L}\boldsymbol{P}_2'\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{0}_k \\ \boldsymbol{0}_k' & 1 \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\mu}}' \end{pmatrix} = \boldsymbol{P}_1'(\boldsymbol{X}, \boldsymbol{1}_n)\begin{pmatrix} \hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\mu}}' \end{pmatrix} = \boldsymbol{P}_1'\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}. \tag{2.15}$$

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

$$\begin{aligned} \mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}] &= E[\mathrm{tr}\{(E[\boldsymbol{Y}] - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}^{-1}(E[\boldsymbol{Y}] - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})'\boldsymbol{P}_1\boldsymbol{P}_1'\}] \\ &= E[r(E[\boldsymbol{Z}], \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})] = \mathrm{MSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}]. \end{aligned} \tag{2.16}$$

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 Optimal $\boldsymbol{\theta}$

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

$$r(E[\boldsymbol{Z}], \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}) = \mathrm{tr}\left\{ \boldsymbol{L} \begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix}' \boldsymbol{L}' \right\}. \tag{2.17}$$

By elementary linear algebra,

$$\boldsymbol{L} \begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix} = \begin{pmatrix} \mathrm{diag}(\sqrt{d_1}, \dots \sqrt{d_k}, \sqrt{n}) \\ \boldsymbol{O}_{n-k-1,k+1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}' - \hat{\boldsymbol{\mu}}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}^{1/2} \left(\boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \right) \\ \sqrt{n}\, (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \\ \boldsymbol{O}_{n-k-1,p} \end{pmatrix}. \tag{2.18}$$

Notice that

$$\begin{aligned} \boldsymbol{D}^{1/2}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} &= \boldsymbol{D}^{1/2}(\boldsymbol{D} + \Theta)^{-1}\boldsymbol{C}'\boldsymbol{Z} \\ &= (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}(\boldsymbol{D}, \boldsymbol{O}_{k,n-k})\boldsymbol{Z} = \left(\frac{d_1}{d_1 + \theta_1}\boldsymbol{z}_1, \dots, \frac{d_k}{d_k + \theta_k}\boldsymbol{z}_k \right)'. \end{aligned} \tag{2.19}$$

This equation implies that

$$\begin{aligned} \boldsymbol{D}^{1/2} \left(\boldsymbol{\Gamma} - \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \right) &= \boldsymbol{D}^{1/2}\boldsymbol{\Gamma} - (\boldsymbol{D} + \boldsymbol{\Theta})^{-1}(\boldsymbol{D}, \boldsymbol{O}_{k,n-k})\boldsymbol{Z} \\ &= \left(\sqrt{d_1}\boldsymbol{\gamma}_1 - \frac{d_1}{d_1 + \theta_1}\boldsymbol{z}_1, \dots, \sqrt{d_k}\boldsymbol{\gamma}_k - \frac{d_k}{d_k + \theta_k}\boldsymbol{z}_k \right)'. \end{aligned} \tag{2.20}$$

By using equations (2.17), (2.18) and (2.20), we can derive another expression for MSE$[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}]$ as

$$\begin{aligned} \mathrm{MSE}[\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}] &= E[r(E[\boldsymbol{Z}], \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})] \\ &= \sum_{i=1}^{k} E\left[\left(\sqrt{d_i}\boldsymbol{\gamma}_i - \frac{d_i}{d_i + \theta_i}\boldsymbol{z}_i \right)' \boldsymbol{\Sigma}^{-1} \left(\sqrt{d_i}\boldsymbol{\gamma}_i - \frac{d_i}{d_i + \theta_i}\boldsymbol{z}_i \right) \right] \\ &\quad + nE[(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})]. \end{aligned} \tag{2.21}$$

Recall that $\hat{\boldsymbol{\mu}} = \boldsymbol{z}_{k+1}/\sqrt{n}$. It follows from (2.12) that

$$\begin{aligned} nE[(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})] &= E[(\sqrt{n}\boldsymbol{\mu} - \boldsymbol{z}_{k+1})'\boldsymbol{\Sigma}^{-1}(\sqrt{n}\boldsymbol{\mu} - \boldsymbol{z}_{k+1})] \\ &= \mathrm{tr}(\mathrm{Cov}[\boldsymbol{z}_{k+1}]\boldsymbol{\Sigma}^{-1}) = p. \end{aligned} \tag{2.22}$$

Moreover, by using the results that $E[\boldsymbol{z}_i] = \sqrt{d_i}\boldsymbol{\gamma}_i$ and $E[\boldsymbol{z}_i\boldsymbol{z}_i'] = \boldsymbol{\Sigma}+d_i\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i'$ $(i = 1, \dots, k)$, we calculate that

$$E\left[\left(\sqrt{d_i}\boldsymbol{\gamma}_i - \frac{d_i}{d_i+\theta_i}\boldsymbol{z}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\sqrt{d_i}\boldsymbol{\gamma}_i - \frac{d_i}{d_i+\theta_i}\boldsymbol{z}_i\right)\right] = \varphi(\theta_i|d_i, \boldsymbol{\gamma}_i), \tag{2.23}$$

where

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

Substituting (2.22) and (2.23) into (2.21) yields

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

The above equation indicates that the principal optimal value of θ_i can be obtained by minimizing $\varphi(\theta_i|d_i, \boldsymbol{\gamma}_i)$ individually. Let $\theta_i^* \geq 0$ $(i = 1, \dots, k)$ be the principal optimal value of θ_i. The first partial derivative of $\varphi(\theta_i|d_i, \boldsymbol{\gamma}_i)$ with respect to θ_i is calculated as

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

The above equation yields the principal optimal value of θ_i as

$$\theta_i^* = \frac{p}{\boldsymbol{\gamma}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}_i}, \quad (i = 1, \dots, k). \tag{2.24}$$

3. MGR Estimators with Optimized Ridge Parameters

For the case of a univariate linear model, many authors have provided formulas for GR estimators with optimized ridge parameters. By extending their methods for optimizing $\boldsymbol{\theta}$ to the multivariate case, we derive formulas for MGR estimators with optimized ridge parameters. Since the MGR estimator $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}$ in (2.1) is obtained by using the equation $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}} = \boldsymbol{Q}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}$ in (2.14), we deal with $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}}$ in (2.13) instead of $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta}}$. Let $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \dots, \hat{\boldsymbol{\gamma}}_k)'$ be the ordinary LS estimator of $\boldsymbol{\Gamma}$, i.e., $\hat{\boldsymbol{\Gamma}} = \boldsymbol{D}^{-1}\boldsymbol{C}'\boldsymbol{Z}$. Then, we have

$$\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} = (\boldsymbol{D}+\boldsymbol{\Theta})^{-1}\boldsymbol{C}'\boldsymbol{Z} = (\boldsymbol{D}+\boldsymbol{\Theta})^{-1}\boldsymbol{D}\hat{\boldsymbol{\Gamma}}. \tag{3.1}$$

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 $\hat{\boldsymbol{\gamma}}_i(\hat{\theta}_i)$ 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 (3.1), we can see that $\hat{\boldsymbol{\gamma}}_i(\hat{\theta}_i)$ is expressed as

$$\hat{\boldsymbol{\gamma}}_i(\hat{\theta}_i) = \frac{d_i}{d_i+\hat{\theta}_i}\hat{\boldsymbol{\gamma}}_i, \quad (i = 1, \dots, k). \tag{3.2}$$

It is easy to obtain that $\hat{\boldsymbol{\gamma}}_i = \hat{\boldsymbol{\gamma}}_i(0)$. Let

$$t_i = \boldsymbol{z}_i' \boldsymbol{S}^{-1} \boldsymbol{z}_i, \quad (i = 1, \dots, k). \tag{3.3}$$

Since $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{z}_i / \sqrt{d_i}$, t_i in (3.3) can be rewritten as

$$t_i = d_i \hat{\boldsymbol{\gamma}}_i' \boldsymbol{S}^{-1} \hat{\boldsymbol{\gamma}}_i, \quad (i = 1, \dots, k). \tag{3.4}$$

If $\hat{\theta}_i$ is a function of t_i, then we can express $\hat{\boldsymbol{\gamma}}_i(\hat{\theta}_i)$ in (3.2) as

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

where $w(t_i)$ is a function of t_i. From (3.2), 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 optimized ridge parameters. According to his notation, we specify the individual MGR estimator with an optimized value of $\boldsymbol{\theta}$ using the weight function.

3.1. Plug-in Methods

In this subsection, we consider optimization methods based on the plug-in method. The plug-in estimation is specified by estimators of $\boldsymbol{\gamma}_i$.

3.1.1. Once Plug-in Method

Since the principal optimal value of $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_k^*)'$ is obtained as (2.24), we estimate θ_i^* by replacing $\boldsymbol{\gamma}_i$ and $\boldsymbol{\Sigma}$ with $\hat{\boldsymbol{\gamma}}_i$ and $\boldsymbol{S}$. Hence we obtain the following optimal $\boldsymbol{\theta}$ by single plug-in estimation:

$$\hat{\theta}_i^{[1]} = \frac{p}{\hat{\boldsymbol{\gamma}}_i' \boldsymbol{S}^{-1} \hat{\boldsymbol{\gamma}}_i} = \frac{d_i p}{t_i}, \quad (i = 1, \dots, k). \tag{3.5}$$

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

3.1.2. Multiple Plug-in Method

If multicollinearity occurs, the PI method does not yield a good estimate, since $\hat{\gamma}_i$ depends on the ordinary LS estimator. Hence using the MGR estimator instead of $\hat{\gamma}_i$ yields the following optimal value of $\boldsymbol{\theta}$:

$$\hat{\theta}_i^{[s]} = \frac{p}{\hat{\boldsymbol{\gamma}}_i^{[s-1]\prime} \boldsymbol{S}^{-1} \hat{\boldsymbol{\gamma}}_i^{[s-1]}}, \ (s = 1, 2, \ldots ; i = 1, \ldots , k), \tag{3.6}$$

where $\hat{\boldsymbol{\gamma}}_i^{[s]} = d_i \hat{\boldsymbol{\gamma}}_i / (d_i + \hat{\theta}_i^{[s]})$, $(s = 0, 1, \ldots)$ and $\hat{\theta}_i^{[0]} = 0$. Notice that $\hat{\boldsymbol{\gamma}}_i^{[1]}$ is equal to the estimator obtained using the PI method. Equation (3.6) implies that

$$\hat{\theta}_i^{[s]} = \left(1 + \frac{\hat{\theta}_i^{[s-1]}}{d_i} \right)^2 \hat{\theta}_i^{[1]}, \ (s = 1, 2, \ldots ; i = 1, \ldots , k). \tag{3.7}$$

In the case of $p = 1$, the value of (3.6) was proposed by Hoerl and Kennard (1970), and they used $\hat{\boldsymbol{\gamma}}_i^{[2]}$ to estimate the regression coefficient. Hence we also use $\hat{\boldsymbol{\gamma}}_i^{[2]}$ which is obtained by using $\hat{\theta}_i^{[2]}$. We denote this plug-in twice method as PI_2. The optimal value of θ_i derived using the PI_2 method is given by

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

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

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

3.1.3. Infinite Plug-in Method

For the case of $p = 1$, Hemmerle (1975) showed that the value of (3.6) converges as $s \to \infty$. By extending the proof in Hemmerle (1975) to the multivariate case, we obtain the following limiting value of (3.6) as $s \to \infty$:

$$\hat{\theta}_i^{[\infty]} = \begin{cases} \dfrac{d_i \{ t_i - 2p - \sqrt{t_i (t_i - 4p)} \}}{2p} & (t_i \geq 4p) \\ \infty & (t_i < 4p) \end{cases}, \quad (i = 1, \ldots , k), \tag{3.8}$$

(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

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

3.2. IC-based Method

Yanagihara, Nagai and Satoh (2009) proposed C_p-type criteria for optimizing $\boldsymbol{\theta}$. By omitting constant terms, their criteria are included in the following GC_p criterion:

$$GC_p(\boldsymbol{\theta}|\lambda) = \lambda^{-1}\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}) + 2p\mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\Theta\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{X}\}, \tag{3.9}$$

where the function $\hat{r}$ is given by (2.5). The optimal value of θ_i which minimizes (3.9) is obtained as

$$\hat{\theta}_i^{(\mathrm{G})}(\lambda) = \begin{cases} \dfrac{\lambda p d_i}{t_i - \lambda p} & (t_i > \lambda p) \\ \infty & (t_i \leq \lambda p) \end{cases}, \quad (i = 1, \dots, k), \tag{3.10}$$

(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

$$w^{(\mathrm{G})}(t_i|\lambda) = \begin{cases} 1 - \dfrac{\lambda p}{t_i} & (t_i > \lambda p) \\ 0 & (t_i \leq \lambda p) \end{cases}. \tag{3.11}$$

3.2.1. Optimization by Minimizing the C_p Criterion

Yanagihara, Nagai and Satoh (2009) proposed a crude C_p criterion whose main term corresponds to $GC_p(\boldsymbol{\theta}|1)$. From (3.10), $\hat{\theta}_i^{(\mathrm{C})}$ that minimizes the C_p criterion is $\hat{\theta}_i^{(\mathrm{C})} = \hat{\theta}_i^{(\mathrm{G})}(1)$ $(i = 1, \dots, k)$. Then equation (3.11) 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.

3.2.2. Optimization by Minimizing the MC_p criterion

If $\boldsymbol{\mathcal{E}} \sim N_{n\times p}(\boldsymbol{O}_{n,p}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$ and $n - k - p - 2 > 0$, Yanagihara, Nagai and Satoh (2009) proposed the MC_p criterion, whose main term corresponds to $GC_p(\boldsymbol{\theta}|c_{\mathrm{M}})$ where $c_{\mathrm{M}} = (n - k - 1)/(n - k - p - 2)$. Hence $\hat{\theta}_i^{(\mathrm{M})}$ 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.

3.2.3. James-Stein Estimator

Kubokawa (1991) proposed an improved James-Stein estimator which is a shrinkage estimator when $p \geq 3$. Suppose that $\boldsymbol{\mathcal{E}} \sim N_{n\times p}(\boldsymbol{O}_{n,p}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Since $\hat{\boldsymbol{\gamma}}_i \sim N_p(\boldsymbol{\gamma}_i, \boldsymbol{\Sigma}/d_i)$

$(i = 1, \ldots, k)$, $(n-k-1)\boldsymbol{S} \sim W_p(n-k-1, \boldsymbol{\Sigma})$ and $\boldsymbol{S} \perp\!\!\!\perp \hat{\boldsymbol{\gamma}}_i$ $(i = 1, \ldots, k)$ are satisfied, the James-Stein estimator of $\boldsymbol{\gamma}_i$ is obtained as

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

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

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

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

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

From (3.10), 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.

3.3. Other 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 $\boldsymbol{\theta}$:

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

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

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

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

4. Properties of Optimized Ridge Parameters

4.1. Relationship with Hypothesis Testing

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 $\hat{\boldsymbol{\gamma}}_i$	shrinking $\hat{\boldsymbol{\gamma}}_i$
PI$_\infty$	$4p$	shrinking $\hat{\boldsymbol{\gamma}}_i$	$\mathbf{0}_p$
C_p	p	shrinking $\hat{\boldsymbol{\gamma}}_i$	$\mathbf{0}_p$
MC_p	$c_{\mathrm{M}}p$	shrinking $\hat{\boldsymbol{\gamma}}_i$	$\mathbf{0}_p$
JS	$c_{\mathrm{J}}p$	shrinking $\hat{\boldsymbol{\gamma}}_i$	$\mathbf{0}_p$
PC	$2p$	$\hat{\boldsymbol{\gamma}}_i$	$\mathbf{0}_p$

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

$$H_0 : \boldsymbol{\gamma}_i = \mathbf{0}_p \text{ vs. } H_1 : \boldsymbol{\gamma}_i \neq \mathbf{0}_p. \tag{4.1}$$

In this subsection, we discuss the relationship between each method for optimizing $\boldsymbol{\theta}$ and the hypothesis test of (4.1). Since $\mathrm{Cov}[\hat{\boldsymbol{\gamma}}_i] = \boldsymbol{\Sigma}/d_i$, the test statistic for (4.1) is t_i in (3.4). Suppose that $\boldsymbol{\mathcal{E}} \sim N_{n\times p}(\boldsymbol{O}_{n,p}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$. Then the test statistic t_i is distributed according to 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$_\infty$, C_p, MC_p, JS and PC methods, the MGR estimators with optimized ridge parameters of $\boldsymbol{\gamma}_i$ 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 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$_\infty$, 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 optimized ridge parameters becoming $\mathbf{0}_p$. The MGR estimators with 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$_\infty$), p (C_p), $c_{\mathrm{M}}p$ (MC_p),

Table 2. The significance levels in several cases

k	n	PI_∞	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

$c_{\mathrm{J}}p$ (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 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.

4.2. Magnitude Relations Among Optimized $\boldsymbol{\theta}$

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

It follows from (3.7) that $\hat{\theta}_i^{[s]} > 0$, $(s = 1, 2, \dots)$, 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:*

$$0 < \hat{\theta}_i^{[1]} < \hat{\theta}_i^{[2]} < \cdots < \hat{\theta}_i^{[\infty]}, \quad (i = 1, \dots, k). \tag{4.2}$$

For $\boldsymbol{\theta}$ optimized by the IC-based method, we obtain the following theorem from (3.10):

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

$$\hat{\theta}_i^{(G)}(\lambda_1) \leq \hat{\theta}_i^{(G)}(\lambda_2), \quad (i = 1, \dots, k), \tag{4.3}$$

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

From theorem 2, we have

$$\hat{\theta}_i^{(\mathrm{C})} \le \hat{\theta}_i^{(\mathrm{M})}, \quad \hat{\theta}_i^{(\mathrm{J})} \le \hat{\theta}_i^{(\mathrm{M})}, \quad (i = 1, \ldots, k),$$

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

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

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:*

$$\begin{cases} \hat{\theta}_i^{[1]} < \hat{\theta}_i^{(G)}(\lambda), & (\textit{when } \lambda \ge 1), \\ \hat{\theta}_i^{(G)}(\lambda) \le \hat{\theta}_i^{[\infty]}, & (\textit{when } 0 < \lambda \le 1), \end{cases} \quad (i = 1, \ldots, k), \tag{4.4}$$

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

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

$$\hat{\theta}_i^{[1]} < \hat{\theta}_i^{(\mathrm{C})} \le \hat{\theta}_i^{[\infty]}, \quad (i = 1, \ldots, k),$$

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

4.3. Magnitude Relations Among Weight Functions

The shrinkage ratio of each method corresponds to the weight function $w(t_i)$. A method with smaller $w(t_i)$ shrinks $\hat{\boldsymbol{\gamma}}_i$ to a greater extent. When $w(t_i)$ is nearly equal to one, the method shrinks $\hat{\boldsymbol{\gamma}}_i$ 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 $\hat{\boldsymbol{\gamma}}_i$ 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

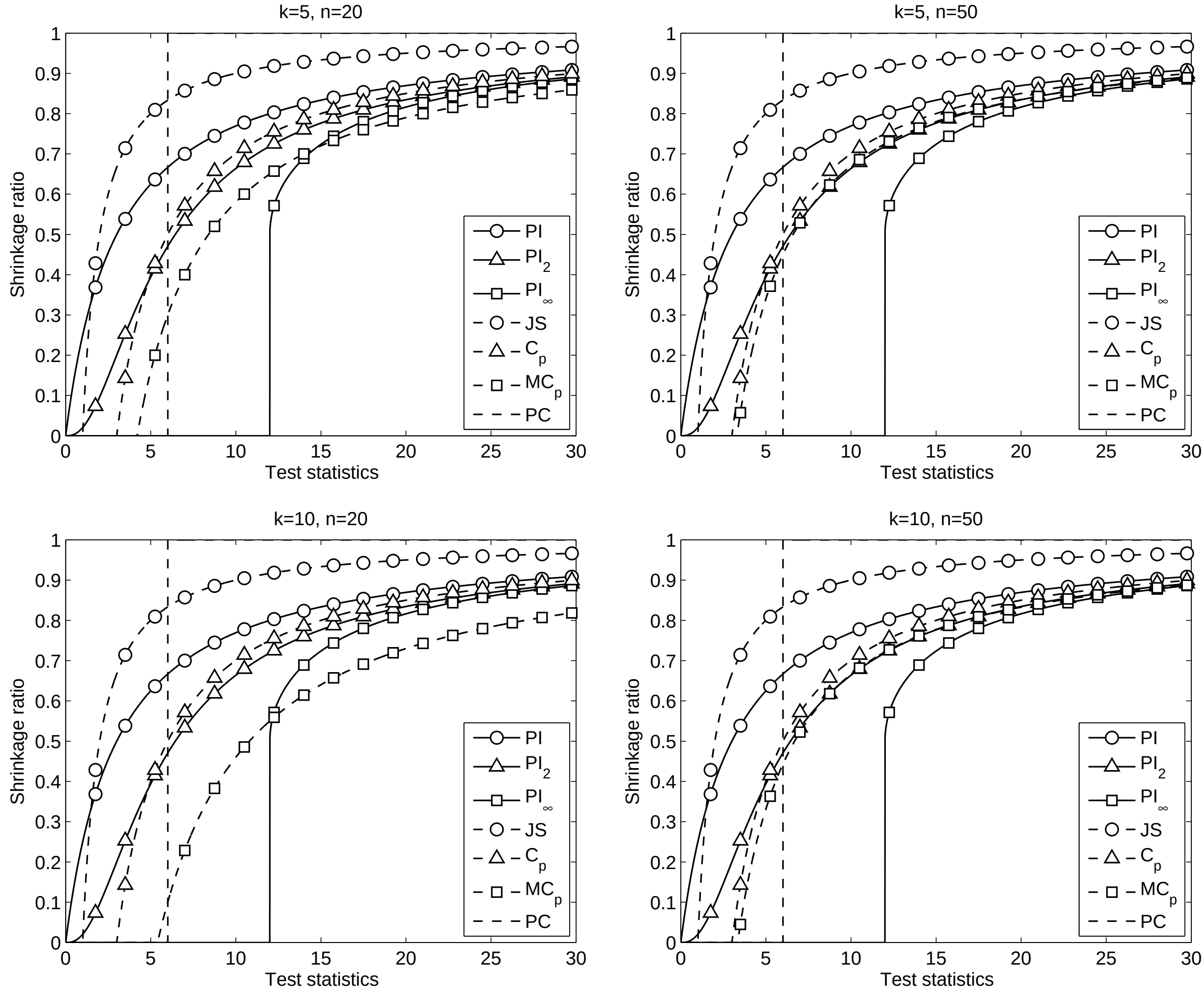

Figure 1. Shrinkage ratio (value of weight function) for each optimization method in several cases.

$w(t) = d_i/(d_i + \hat{\theta}_i)$. 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:

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

$$\begin{aligned}
&w^{[\infty]}(t) < \cdots < w^{[2]}(t) < w^{[1]}(t),\\
&w^{(M)}(t) \le \begin{cases} w^{(J)}(t) \le w^{(C)}(t) & (p < \{3 + \sqrt{9 + 8(n-k-1)}\}/2),\\ w^{(C)}(t) \le w^{(J)}(t) & (p \ge \{3 + \sqrt{9 + 8(n-k-1)}\}/2),\end{cases}\\
&w^{[\infty]}(t) \le w^{(C)}(t) < w^{[1]}(t).
\end{aligned}$$

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

5. Numerical Study

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

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

The explanatory matrix $\boldsymbol{X}$ was generated from $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{\Psi}^{1/2}$ where $\boldsymbol{\Psi} = \boldsymbol{R}_k \boldsymbol{\Delta}_k(\rho_x)\boldsymbol{R}_k$ 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 Ξ was defined by $\Xi = \delta \boldsymbol{F} \Xi_0$, where δ is constant, and $\boldsymbol{F}$ and Ξ are defined as

$$\boldsymbol{F} = \begin{pmatrix} \boldsymbol{I}_\kappa & \boldsymbol{O}_{\kappa,10-\kappa} \\ \boldsymbol{O}_{k-\kappa} & \boldsymbol{O}_{k-\kappa,10-\kappa} \end{pmatrix},$$

$$\Xi_0 = \begin{pmatrix} 0.8501 & -0.2753 & -0.3193 & 0.2754 & 0.2693 & -0.0676 & 0.2239 & -0.0352 & 0.3240 & -0.3747 \\ 0.6571 & -0.2432 & -0.2926 & 0.2608 & 0.2164 & -0.0663 & 0.2197 & -0.0346 & 0.3199 & -0.3727 \\ 0.2159 & -0.1187 & -0.1671 & 0.1766 & 0.2066 & -0.0561 & 0.1880 & -0.0305 & 0.2868 & -0.3554 \end{pmatrix}'.$$

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 Ξ_0, which is an essential regression coefficient matrix, are the same as in Lawless (1981). Simulated data values $\boldsymbol{Y}$ were generated by $N_{n\times 3}(\boldsymbol{X}\Xi, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$ repeatedly under several selections of n, k, κ, δ, ρ_x and ρ_y, where $\boldsymbol{\Sigma} = \boldsymbol{R}_3 \boldsymbol{\Delta}_3(\rho_y)\boldsymbol{R}_3$ and the number of repetition was $10,000$. At each repetition, we evaluated $r(\boldsymbol{X}\Xi, \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}})$, where $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{1}_n \bar{\boldsymbol{y}}' + \boldsymbol{X}\hat{\Xi}_{\hat{\boldsymbol{\theta}}}$ which is the predicted value of $\boldsymbol{Y}$ obtained from each method. The average of $r(\boldsymbol{X}\Xi, \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}})$ 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 a predictor of $\boldsymbol{Y}$ derived by considering the LS estimator of Ξ. We observe that the method can improve the LS estimation when values in the table 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 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$_\infty$ 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$_\infty$ has the most shrinkage of the LS estimators. On average, C_p was the best method in all cases if we except PI$_2$ and MC_p. One of the reasons is that the shape of 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 all cases although the ratios of improvement are not as great. We summarize the results of the numerical study in Table 7 which shows the best method and additionally the second best method in several cases.

Please insert Tables 3, 4, 5, 6 and 7 around here

Appendix

A.1. The Proof of Equation (3.8)

In this subsection, we show that the $\hat{\theta}_i^{[s]}$ in (3.6) converge to $\hat{\theta}_i^{[\infty]}$ in (3.8) 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. Hence, firstly, we show that $\hat{\theta}_i^{[s]}$ is bounded above when $t_i \geq 4p$ is satisfied, where t_i is given by (3.3) or (3.4). Recall that $\hat{\theta}_i^{[1]} = d_i p/t_i$, where d_i is an eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$, which is defined by (2.2). Thus, we have $\hat{\theta}_i^{[1]} \leq d_i/4$ when $t_i \geq 4p$ holds. By using this bound of $\hat{\theta}_i^{[1]}$ and (3.7), the following inequality can be derived:

$$\hat{\theta}_i^{[s]} \leq \frac{d_i}{4}\left(1 + \frac{\hat{\theta}_i^{[s-1]}}{d_i}\right)^2, \tag{A.1}$$

with equality if and only if $t_i = 4p$. From (A.1) and the bound of $\hat{\theta}_i^{[1]}$, an inequality for

$\hat{\theta}_i^{[s]}$ with $s = 2$ is obtained as

$$\hat{\theta}_i^{[2]} \leq \frac{d_i}{4}\left(1 + \frac{\hat{\theta}_i^{[1]}}{d_i}\right)^2 \leq \frac{d_i}{4}\left(1 + \frac{1}{4}\right)^2 = d_i\left(\frac{5}{8}\right)^2, \tag{A.2}$$

with equality if and only if $t_i = 4p$. Suppose that the following inequality holds:

$$\hat{\theta}_i^{[s]} \leq d_i\left(1 - \frac{3}{2^{s+1}}\right)^2. \tag{A.3}$$

Equation (A.2) states that (A.3) holds when $s = 2$. By using (A.1), we have the following inequality when (A.3) holds:

$$\hat{\theta}_i^{[s+1]} \leq \frac{d_i}{4}\left\{1 + \left(1 - \frac{3}{2^{s+1}}\right)^2\right\}^2 = d_i\left(1 - \frac{3}{2^{s+1}} + \frac{18}{4^{s+2}}\right)^2. \tag{A.4}$$

On the other hand, for any positive integer s, we have

$$1 - \frac{3}{2^{s+2}} - \left(1 - \frac{3}{2^{s+1}} + \frac{18}{4^{s+2}}\right) = \frac{3}{2^{s+2}} - \frac{18}{4^{s+2}} = \frac{3}{2^{s+2}}\left(1 - \frac{3}{2^{s+1}}\right) \geq 0, \tag{A.5}$$

with equality if and only if $s \to \infty$. From (A.4), it is easy to see that $1 - 3/2^{s+1} + 18/4^{s+2} > 0$ always holds. Moreover, we can see that $1 - 3/2^{s+2} > 0$ is satisfied for any positive integer s. These results together with (A.5) imply that

$$\left(1 - \frac{3}{2^{s+2}}\right)^2 \geq \left(1 - \frac{3}{2^{s+1}} + \frac{18}{4^{s+2}}\right)^2, \tag{A.6}$$

with equality if and only if $s \to \infty$. Combining (A.4) and (A.6) yields

$$\hat{\theta}_i^{[s+1]} \leq d_i\left(1 - \frac{3}{2^{s+2}}\right)^2.$$

Consequently, by mathematical induction, it follows that the inequality (A.3) holds for $s \geq 2$. The equality of (A.3) holds if and only if $(t_i = 4p, s = 2)$ or $(t_i = 4p, s \to \infty)$. Since $\{\hat{\theta}_i^{[s]}\}$ is a monotonic increasing sequence, an upper bound of $\hat{\theta}_i^{[s]}$ is obtained by letting s to ∞ on the right hand side of (A.3). Notice that $\lim_{s\to\infty}(1 - 3/2^{s+1}) = 1$. Therefore, we can see that $\hat{\theta}_i^{[s]} \leq d_i$ is always satisfied for any integer s when $t_i \geq 4p$ holds. The equality of the bound holds if and only if $t_i = 4p$ and $s \to \infty$.

Next, we assume that $\hat{\theta}_i^{[s]}$ converges to some value, i.e., $\lim_{s\to\infty} \hat{\theta}_i^{[s]} = a_i < \infty$. Then, from (3.7), we can see that a_i satisfies the following equation:

$$a_i = \left(1 + \frac{a_i}{d_i}\right)^2 \frac{d_i p}{t_i}.$$

By solving the above quadratic equation for a_i, we have

$$a_i = d_i b_{\mathrm{U}}(t_i) \text{ or } d_i b_{\mathrm{L}}(t_i), \tag{A.7}$$

where $b_{\mathrm{U}}(t_i)$ and $b_{\mathrm{L}}(t_i)$ are functions of t_i, which are given by

$$b_{\mathrm{U}}(t_i) = \frac{t_i - 2p + \sqrt{t_i(t_i - 4p)}}{2p}, \quad b_{\mathrm{L}}(t_i) = \frac{t_i - 2p - \sqrt{t_i(t_i - 4p)}}{2p}.$$

If $t_i < 4p$ holds, a_i does not exist. This result is contradictory to the assumption that a_i exists. Hence, by reductio ad absurdum, we can see that $\hat{\theta}_i^{[s]}$ does not converge when $t_i < 4p$ holds. Recall that $\{\hat{\theta}_i^{[s]}\}$ is a monotonic increasing sequence. Hence, if $t_i < 4p$ holds, $\lim_{s\to\infty} \hat{\theta}_i^{[s]} = \infty$ is satisfied.

Finally, we study which of the two values in (A.7) is suitable for the limiting value of $\hat{\theta}_i^{[s]}$ as $s \to \infty$. It is clearly known that $b_{\mathrm{U}}(t_i)$ is a monotonic increasing positive-valued function of t_i when $t_i \geq 4p$. Hence, we have $d_i b_{\mathrm{U}}(t_i) \geq d_i b_{\mathrm{U}}(4p) = d_i$. However, the limiting value of $\hat{\theta}_i^{[s]}$ must not exceed d_i. Therefore, $d_i b_{\mathrm{U}}(t_i)$ is not appropriate for the limiting value of $\hat{\theta}_i^{[s]}$. On the other hand, we have $d_i b_{\mathrm{L}}(t_i) = d_i / b_{\mathrm{U}}(t_i)$. Since $b_{\mathrm{U}}(t_i)$ is a monotonic increasing positive-valued function of t_i when $t_i \geq 4p$, $d_i b_{\mathrm{L}}(t_i)$ is a monotonic decreasing positive-valued function of t_i when $t_i \geq 4p$. Hence, we have $0 < d_i b_{\mathrm{L}}(t_i) \leq d_i b_{\mathrm{L}}(4p) = d_i$. This leads us to the conclusion that $d_i b_{\mathrm{L}}(t_i)$ is the appropriate value for the limit of $\hat{\theta}_i^{[s]}$.

A.2. The Proof of Equation (3.10)

From (2.2), the second part of $GC_p(\boldsymbol{\theta}|\lambda)$ in (3.9) can be rewritten as

$$\mathrm{tr}\{(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{X}'\boldsymbol{X}\} = \mathrm{tr}\{(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{D}\} = \sum_{i=1}^{k} \frac{d_i}{d_i + \theta_i}. \tag{A.8}$$

Moreover, from (2.10) and (2.15), the first part of $GC_p(\boldsymbol{\theta}|\lambda)$ can be rewritten as

$$\begin{aligned} \hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}}) &= \mathrm{tr}\left\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})\boldsymbol{S}^{-1}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta}})'\right\} \\ &= \mathrm{tr}\left\{\boldsymbol{P}_1(\boldsymbol{Z} - \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})\boldsymbol{S}^{-1}(\boldsymbol{Z} - \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}})'\boldsymbol{P}_1'\right\} = \hat{r}(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}). \end{aligned} \tag{A.9}$$

By using (2.15) and (2.18), we have

$$\hat{\boldsymbol{Z}}_{\boldsymbol{\theta}} = \boldsymbol{L}\begin{pmatrix} \hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\mu}}' \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}^{1/2}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\theta}} \\ \sqrt{n}\hat{\boldsymbol{\mu}}' \\ \boldsymbol{O}_{n-k-1,p} \end{pmatrix}. \tag{A.10}$$

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 (2.19) and (A.10) into (A.9) yields

$$\hat{r}(\boldsymbol{Z}, \hat{\boldsymbol{Z}}_{\boldsymbol{\theta}}) = \sum_{i=1}^{k} \left(\frac{\theta_i}{d_i + \theta_i} \right)^2 t_i + \sum_{i=k+2}^{n} \boldsymbol{z}_i' \boldsymbol{S}^{-1} \boldsymbol{z}_i, \tag{A.11}$$

where t_i is given by (3.3) or (3.4). 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 (A.9) and (A.10), 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, (A.8), (A.9) and (A.11), $GC_p(\boldsymbol{\theta}|\lambda)$ can be rewritten as

$$GC_p(\boldsymbol{\theta}|\lambda) = \sum_{i=1}^{k} f(\theta_i|d_i, t_i, \lambda) + \lambda^{-1} p(n-k-1), \tag{A.12}$$

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

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

Hence in order to obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda) = (\hat{\theta}_1^{(\mathrm{G})}(\lambda), \ldots, \hat{\theta}_k^{(\mathrm{G})}(\lambda))'$, $(\hat{\theta}_i^{(\mathrm{G})}(\lambda) \geq 0, \ i = 1, \ldots, k)$ making $GC_p(\boldsymbol{\theta}|\lambda)$ a 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

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

This derivative indicates that $f(\theta_i|d_i, t_i, \lambda)$ becomes a minimum at $\theta_i = \lambda p d_i/(t_i - \lambda p)$ 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, Equation (3.10) follows.

A.3. The Proof of Equation (4.4)

Firstly, we show the proof of the first inequality of Equation (4.4). 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 $\hat{\theta}_i^{[i]} < \infty$ are satisfied when $t_i \leq \lambda p$. When $t_i > \lambda p$, from (3.5) and (3.10), we can see that

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

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 (4.4). 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

$$1 - \frac{4p}{t_i} < \left(1 - \frac{2p}{t_i - p}\right)^2 < \left(1 - \frac{2p}{t_i - \lambda p}\right)^2. \tag{A.13}$$

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 (A.13) that

$$\sqrt{1 - \frac{4p}{t_i}} < 1 - \frac{2p}{t_i - \lambda p}. \tag{A.14}$$

By multiplying both sides of (A.14) by t_i after calculation, we have

$$\frac{t_i}{t_i - \lambda p} < \frac{t_i - \sqrt{t_i(t_i - 4p)}}{2p}. \tag{A.15}$$

Subtracting 1 from both sides of (A.15) yields

$$\frac{\lambda p}{t_i - \lambda p} < \frac{t_i - 2p - \sqrt{t_i(t_i - 4p)}}{2p}. \tag{A.16}$$

Thus, when $t_i > 4p$, $\hat{\theta}_i^{(\mathrm{G})}(\lambda) < \hat{\theta}_i^{[\infty]}$ can be derived by multiplying both sides of (A.16) by d_i. Consequently, $\hat{\theta}_i^{(\mathrm{G})}(\lambda) \leq \hat{\theta}_i^{[\infty]}$ is obtained when $0 < \lambda \leq 1$.

Acknowledgment

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.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[19] Yanagihara, H. & Satoh, K. (2010). An unbiased C_p criterion for multivariate ridge regression. *J. Multivariate Anal.* (in press).

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

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

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

κ	δ	ρ_x	ρ_ε	PI	PI$_2$	PI$_\infty$	C_p	MC_p	JS	PC
0	0.0	0.2	0.2	51.02	36.75	**23.35**	37.66	29.98	66.47	53.99
		0.2	0.9	51.29	36.92	**23.50**	37.82	30.11	66.85	54.21
		0.9	0.2	50.64	36.41	**23.27**	37.28	29.70	65.99	53.18
		0.9	0.9	50.95	36.72	**23.50**	37.59	29.99	66.34	53.59
3	1.0	0.2	0.2	69.60	**66.19**	82.01	68.71	67.77	81.22	93.44
		0.2	0.9	**79.42**	79.77	109.0	81.22	83.49	87.54	101.1
		0.9	0.2	58.56	48.94	**45.28**	51.02	45.73	73.00	72.41
		0.9	0.9	64.35	57.85	63.00	60.45	**57.07**	77.71	84.58
	3.0	0.2	0.2	**89.05**	89.89	108.6	90.26	93.18	93.06	100.5
		0.2	0.9	**92.68**	93.24	104.2	93.25	95.48	94.93	98.44
		0.9	0.2	77.46	**76.09**	99.57	76.83	78.36	85.33	92.64
		0.9	0.9	81.69	**80.25**	94.04	80.33	81.74	87.42	89.63
5	1.0	0.2	0.2	74.68	**70.79**	80.31	71.42	70.88	83.18	84.22
		0.2	0.9	80.80	**78.70**	86.50	80.01	80.10	87.91	94.47
		0.9	0.2	71.99	**68.85**	85.12	70.62	70.18	82.35	91.05
		0.9	0.9	79.14	**78.36**	97.96	79.93	81.19	87.13	98.65
	3.0	0.2	0.2	87.42	**87.39**	103.8	88.22	89.79	92.20	99.84
		0.2	0.9	**93.40**	94.28	106.5	94.53	96.61	95.81	100.7
		0.9	0.2	**88.17**	88.64	105.8	88.89	91.35	91.99	98.01
		0.9	0.9	90.66	90.52	100.1	**90.49**	92.20	93.24	95.37
Average				74.15	69.83	78.27	70.83	**69.74**	82.98	85.50

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

κ	δ	ρ_x	ρ_ε	PI	PI_2	PI_∞	C_p	MC_p	JS	PC
0	0.0	0.2	0.2	47.37	32.20	**19.24**	32.73	30.49	62.86	46.09
		0.2	0.9	47.47	32.31	**19.45**	32.82	30.58	62.98	46.17
		0.9	0.2	47.27	32.16	**19.48**	32.66	30.44	62.69	45.88
		0.9	0.9	47.59	32.32	**19.31**	32.87	30.61	63.22	46.27
3	1.0	0.2	0.2	**78.14**	78.68	109.8	80.03	80.52	85.94	99.61
		0.2	0.9	83.00	**82.83**	103.6	82.91	83.31	88.22	92.53
		0.9	0.2	64.16	**58.78**	74.87	60.64	60.06	76.19	82.61
		0.9	0.9	70.81	**67.76**	92.12	68.71	68.57	80.47	85.60
	3.0	0.2	0.2	90.80	90.72	102.7	**90.66**	91.01	93.32	95.28
		0.2	0.9	90.69	89.63	94.62	**89.56**	89.63	93.17	92.65
		0.9	0.2	78.64	**75.96**	83.84	77.25	77.11	85.92	91.76
		0.9	0.9	85.30	**85.26**	105.6	86.02	86.36	90.28	98.39
5	1.0	0.2	0.2	81.77	79.53	86.84	**79.53**	79.54	87.14	86.77
		0.2	0.9	83.13	79.77	**78.51**	80.45	80.13	88.44	88.24
		0.9	0.2	77.12	**76.35**	101.3	77.64	77.92	84.94	95.79
		0.9	0.9	82.19	**82.01**	104.7	82.27	82.64	87.69	92.54
	3.0	0.2	0.2	**93.49**	95.82	116.0	97.17	97.78	96.48	109.8
		0.2	0.9	**95.54**	98.89	124.6	99.90	100.8	97.72	111.8
		0.9	0.2	89.18	**89.02**	101.9	89.53	89.83	92.56	97.96
		0.9	0.9	91.22	**90.66**	98.21	90.69	90.83	93.76	94.66
Average				76.24	**72.53**	82.83	73.20	72.91	83.70	85.02

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

κ	δ	ρ_x	ρ_ε	PI	PI$_2$	PI$_\infty$	C_p	MC_p	JS	PC
0	0.0	0.2	0.2	50.13	35.55	**21.25**	36.76	22.78	66.67	56.01
		0.2	0.9	49.90	35.31	**21.07**	36.50	22.60	66.42	55.58
		0.9	0.2	49.81	35.21	**20.89**	36.41	22.47	66.36	55.65
		0.9	0.9	50.03	35.55	**21.43**	36.73	22.88	66.46	55.60
3	1.0	0.2	0.2	60.22	51.35	51.62	53.38	**45.64**	74.74	76.44
		0.2	0.9	65.44	59.34	67.36	61.21	**57.57**	77.98	83.03
		0.9	0.2	54.75	43.17	35.28	45.03	**33.81**	70.44	66.84
		0.9	0.9	58.66	49.00	46.45	51.11	**42.13**	73.71	74.16
	3.0	0.2	0.2	75.62	**72.89**	85.71	74.42	75.38	84.77	92.67
		0.2	0.9	82.27	**81.45**	95.52	82.70	86.06	89.04	98.04
		0.9	0.2	68.70	63.60	72.75	65.42	**63.10**	80.36	86.54
		0.9	0.9	74.19	**70.89**	83.21	72.24	72.92	83.72	90.46
5	1.0	0.2	0.2	69.66	**65.33**	77.04	67.21	65.84	81.02	88.64
		0.2	0.9	75.57	**73.35**	88.38	74.92	76.73	84.70	93.74
		0.9	0.2	59.90	49.72	43.38	51.31	**42.20**	74.05	70.87
		0.9	0.9	62.96	54.49	52.73	56.34	**48.96**	76.26	76.95
	3.0	0.2	0.2	**87.24**	87.71	103.8	88.70	93.75	92.43	101.7
		0.2	0.9	**91.45**	93.02	110.1	93.73	100.1	94.82	104.1
		0.9	0.2	72.44	**68.56**	79.17	70.17	69.18	82.49	89.04
		0.9	0.9	77.89	**75.79**	88.85	77.34	78.43	86.45	94.74
10	1.0	0.2	0.2	**86.66**	89.14	119.2	89.87	99.76	91.88	104.7
		0.2	0.9	**90.62**	92.68	113.2	92.98	101.9	94.05	103.0
		0.9	0.2	67.46	61.17	65.16	62.82	**58.86**	79.17	82.27
		0.9	0.9	71.54	66.84	74.24	68.49	**66.40**	82.10	87.73
	3.0	0.2	0.2	**96.75**	97.45	104.6	97.32	102.3	97.67	99.76
		0.2	0.9	**96.58**	96.62	98.42	96.69	99.36	97.50	98.80
		0.9	0.2	81.29	**79.51**	91.85	80.32	82.82	87.92	93.23
		0.9	0.9	84.92	**83.51**	92.59	84.29	86.50	90.55	95.14
Average				71.88	66.72	72.33	68.02	**65.73**	81.92	84.84

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

κ	δ	ρ_x	ρ_ε	PI	PI$_2$	PI$_\infty$	C_p	MC_p	JS	PC
0	0.0	0.2	0.2	42.79	26.21	**12.02**	26.81	24.03	59.80	41.75
		0.2	0.9	42.79	26.25	**12.06**	26.87	24.09	59.76	41.97
		0.9	0.2	42.46	25.97	**12.01**	26.56	23.79	59.32	41.22
		0.9	0.9	42.94	26.38	**12.14**	27.00	24.21	59.97	41.98
3	1.0	0.2	0.2	62.08	**55.58**	64.92	57.69	56.74	75.28	80.24
		0.2	0.9	70.87	**68.33**	91.41	70.07	69.99	81.17	91.29
		0.9	0.2	53.87	**43.98**	47.69	45.80	44.24	68.83	67.55
		0.9	0.9	59.24	**51.77**	64.15	53.23	52.20	72.52	73.74
	3.0	0.2	0.2	**81.15**	81.21	106.1	81.87	82.41	87.24	96.20
		0.2	0.9	85.59	84.41	93.91	**84.40**	84.58	89.89	90.73
		0.9	0.2	67.47	61.39	70.04	62.02	**61.32**	77.55	75.63
		0.9	0.9	71.75	**66.40**	68.81	67.47	66.75	81.07	81.08
5	1.0	0.2	0.2	**77.58**	78.72	112.0	80.78	81.45	86.19	104.6
		0.2	0.9	**86.73**	90.28	128.8	90.96	92.24	91.46	106.8
		0.9	0.2	58.58	**49.81**	51.26	51.83	50.42	72.39	72.52
		0.9	0.9	65.06	**59.72**	72.16	61.88	61.13	77.25	84.08
	3.0	0.2	0.2	**94.18**	95.45	107.2	95.48	96.16	95.91	100.8
		0.2	0.9	95.96	96.06	99.77	**95.92**	96.27	96.81	97.64
		0.9	0.2	78.67	**78.30**	101.4	79.83	80.16	86.49	98.38
		0.9	0.9	**84.79**	85.90	108.0	86.85	87.50	90.07	100.7
10	1.0	0.2	0.2	**89.95**	90.31	100.0	90.97	91.42	93.51	99.98
		0.2	0.9	**91.22**	91.64	101.5	92.29	92.72	94.04	100.3
		0.9	0.2	77.31	**76.52**	99.90	77.89	78.13	85.45	95.74
		0.9	0.9	82.02	**81.93**	100.0	83.25	83.61	88.72	99.19
	3.0	0.2	0.2	99.53	101.7	115.4	101.4	102.1	**99.42**	103.1
		0.2	0.9	99.97	101.9	115.1	101.6	102.2	**99.80**	103.0
		0.9	0.2	**94.12**	96.55	119.9	96.33	97.30	95.80	102.2
		0.9	0.9	**94.89**	96.17	108.2	95.75	96.39	95.92	98.03
Average				74.77	**71.03**	81.99	71.89	71.56	82.92	85.37

Table 7. Comparison of ridge parameter optimizations by dimension of the true model and sample size

			Average	
	κ		n	
	Small	Large	Small	Large
Best	PI_∞	PI_2	MC_p	PI_2
Second	MC_p	PI, MC_p or C_p	PI_2	MC_p