TR-No. 20-07, Hiroshima Statistical Research Group, 1–49

# Ridge Parameters Optimization based on Minimizing Model Selection Criterion in Multivariate Generalized Ridge Regression

**Mineaki Ohishi**

Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan

### Abstract

A multivariate generalized ridge (MGR) regression provides a shrinkage estimator of the multivariate linear regression by multiple ridge parameters. Since the ridge parameters which adjust the amount of shrinkage of the estimator are unknown, their optimization is an important task to obtain a better estimator. For the univariate case, a fast algorithm has been proposed for optimizing ridge parameters based on minimizing a model selection criterion (MSC) and the algorithm can be applied to various MSCs. In this paper, we extend this algorithm to MGR regression. We also describe the relationship between the MGR estimator which is not sparse and a multivariate adaptive-Lasso estimator which is sparse, under orthogonal explanatory variables.

(Last Modified: October 27, 2020)

**Key words:** Adaptive-Lasso regression, Generalized ridge regression, Model selection criterion, Multivariate analysis, Ridge parameters optimization, Sparsity.

E-mail address: mineaki-ohishi@hiroshima-u.ac.jp

## 1. Introduction

We consider $n$ pairs of data $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$ $(i = 1, \ldots, n)$, where $\boldsymbol{y}_i$ is a $p$-dimensional vector of response variables, $\boldsymbol{x}_i$ is a $k$-dimensional vector of explanatory variables, and $n$ satisfies $n > \max\{p, k+1\}$. A multivariate linear regression model is a statistical model for multiple response variables (e.g., Srivastava, 2002, Chap. 9; Timm, 2002, Chap. 4). Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ be an $n \times p$ matrix of response variables, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be an $n \times k$ matrix of explanatory variables, and $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ be an $n \times p$ matrix of error variables. Then, the multivariate linear regression model is given by

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

where $\boldsymbol{1}_n$ is an $n$-dimensional vector of ones, $\boldsymbol{\mu}$ is a $p$-dimensional vector of location parameters, and $\Xi = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)'$ is a $k \times p$ matrix of regression coefficients. We assume that $\boldsymbol{X}$ is centralized and has full column rank, i.e., $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$ and $\text{rank}(\boldsymbol{X}) = k$, and that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are independently and identically distributed according to mean vector $\boldsymbol{0}_p$ and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{0}_k$ is a $k$-dimensional vector of zeros. One of the most basic methods for estimating the unknown parameters $\boldsymbol{\mu}$ and $\Xi$ in (1.1) is the least squares (LS) method. The LS estimators of $\boldsymbol{\mu}$ and $\Xi$ are given by

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}} = \frac{1}{n}\boldsymbol{Y}'\boldsymbol{1}_n, \quad \hat{\Xi} = \boldsymbol{M}^{-1}\boldsymbol{X}'\boldsymbol{Y}\ (\boldsymbol{M} = \boldsymbol{X}'\boldsymbol{X}). \tag{1.2}$$

These estimators are equal to the maximum likelihood estimators (MLEs) of $\boldsymbol{\mu}$ and $\Xi$ under normality, i.e., the assumption that

$$\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \sim i.i.d.\ N_p(\boldsymbol{0}_p, \boldsymbol{\Sigma}).$$

The LS estimators can be obtained as simple forms as per (1.2) regardless of having good theoretical properties, e.g., unbiasedness and asymptotic normality. Unfortunately, it cannot be said that $\hat{\Xi}$ is a good estimator, in the sense that the variance of the estimator becomes large when multicollinearity occurs.

For the univariate case, i.e., when $p = 1$, a generalized ridge (GR) regression was proposed by Hoerl & Kennard (1970) to avoid the problem posed by multicollinearity. The GR regression can be expected to overcome this problem by shrinking an estimator of regression coefficients. The GR estimator can be obtained as closed form and the amount of shrinkage of the estimator is adjusted by $k$ regularization parameters called ridge parameters. However, since the ridge parameters are unknown, to obtain a better estimator, we have a new problem to address, namely ridge parameters optimization. A model selection criterion (MSC) minimization method is one approach to solve the problem of ridge parameters optimization, which selects ridge parameters minimizing the MSC as the optimal ridge parameters. Most MSCs consist of a residual sum of squares (RSS) and generalized degrees of freedom (GDF). In other words, they account for model fit and model complexity. Salient examples include the $C_p$ criterion (Mallows, 1973), Akaike's information criterion (AIC; Akaike, 1973) under normality, and the generalized cross-validation (GCV) criterion (Craven & Wahba, 1979). Usually, the optimal parameters selected by an MSC minimization method cannot be obtained as closed forms and iterative calculation is often required. This presents difficulties in terms of the validity and applicability of such method. Fortunately, Nagai *et al.* (2012) showed that the optimal ridge parameters based on minimizing a generalized $C_p$ ($GC_p$) criterion (Atkinson, 1980) which is a generalization of the $C_p$ criterion can be obtained as closed forms and Yanagihara (2018) showed that the optimal ridge parameters based on minimizing the GCV criterion can be obtained as closed forms. There are various MSCs having wide class like the $GC_p$ criterion; for example, there are the generalized information criterion (GIC; Nishii, 1984), which includes AIC, and the extended GCV (EGCV) criterion (Ohishi *et al.*, 2020a), which includes the GCV criterion. All these criteria can be regarded as bivariate functions of the RSS and GDF. Ohishi *et al.* (2020a) defined a MSC having a wider class as the bivariate function and proposed an algorithm to minimize it rapidly. Since the ridge parameters can be easily optimized by using various MSCs, the GR regression is a useful method to avoid problems arising from multicollinearity.

Ohishi *et al.* (2020a) also clarified a class of ridge parameters optimized by the MSC minimization method. From the results, under orthogonal explanatory variables, the GR estimator which was previously non-sparse is now characterized by sparsity, i.e., includes 0, after the ridge parameters are optimized. On the other hand, Lasso regression (Tibshirani, 1996) and adaptive-Lasso (AL) regression (Zou, 2006) which is an extension of the Lasso regression are well-known methods for providing a sparse estimator. They also give shrinkage estimators like the GR regression. Although the amount of shrinkage and extent of sparsity of the AL estimator (including the Lasso estimator) are adjusted by a regularization parameter called a tuning parameter, since this parameter is unknown, its optimization is required. Moreover, the AL estimator cannot usually be obtained without iterative calculation. However, Ohishi *et al.* (2020b) showed that the AL estimator can be obtained as closed form under orthogonal explanatory variables and the GR and AL estimators are equivalent after regularization parameters are optimized by the MSC minimization method.

Yanagihara *et al.* (2009) and Nagai *et al.* (2012) naturally extended the GR regression to a multivariate GR (MGR) regression. The MGR estimator is also a shrinkage estimator by $k$ ridge parameters like the GR estimator and we have to consider the ridge parameters optimization. In the MSC minimization method for the MGR regression, although the ridge parameters optimized by the $GC_p$ criterion minimization method can be obtained as closed forms (Nagai *et al.*, 2012), whether this is the case for other criteria is unclear. Recently, Mori & Suzuki (2018) proposed $ZMC_p$ criterion and ZKLIC which are modified versions of the modified $C_p$ ($MC_p$) criterion (Fujikoshi & Satoh, 1997) and the bias-corrected AIC (AIC$_\mathrm{C}$; Hurvich & Tsai, 1989) for MGR regression. However, these MSCs are designed for selecting explanatory variables, not for optimizing ridge parameters. In this paper, we extend the algorithm proposed by Ohishi *et al.* (2020a) to MGR regression. Furthermore, we describe the relationship between MGR regression and multivariate AL (MAL) regression under orthogonal explanatory variables.

The remainder of the paper is organized as follows. In Section 2, we describe the MGR estimator and MSCs for optimizing ridge parameters, and define a MSC class. In Section 3, we extend the algorithm proposed by Ohishi *et al.* (2020a) to optimize ridge parameters in MGR regression by the MSC minimization method. In Section 4, the MSC class defined in Section 2 is extended, corresponding to various distances. Moreover, we propose an algorithm for minimizing the extended MSC. In Section 5, we propose a new method for optimizing ridge parameters by using MSCs. In Section 6, we describe the MAL estimator and an equivalence between the MGR and MAL estimators under the regularization parameters optimized by the MSC minimization method. In Section 7, the performance of the ridge parameters optimized by the MSC minimization methods is compared by simulation. Technical details are provided in the Appendix.

## 2. Preliminaries

By a singular value decomposition, $n \times n$ and $k \times k$ orthogonal matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ and a $k \times k$ diagonal matrix $\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_k)$ express $\boldsymbol{X}$ as

$$\boldsymbol{X} = \boldsymbol{P} \begin{pmatrix} \boldsymbol{D}^{1/2} \\ \boldsymbol{O}_{n-k,k} \end{pmatrix} \boldsymbol{Q}' = \boldsymbol{P}_1 \boldsymbol{D}^{1/2} \boldsymbol{Q}', \tag{2.1}$$

where $\boldsymbol{O}_{n,k}$ is an $n \times k$ matrix of zeros, $\boldsymbol{P}_1$ is an $n \times k$ matrix obtained from the partition $\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$, which satisfies $\boldsymbol{P}_1'\boldsymbol{1}_n = \boldsymbol{0}_k$ and $\boldsymbol{P}_1'\boldsymbol{P}_1 = \boldsymbol{I}_k$, and $d_1, \dots, d_k$ are eigenvalues of $\boldsymbol{M}$ $(= \boldsymbol{X}'\boldsymbol{X})$ satisfying $d_1 \geq \cdots \geq d_k > 0$. Then, the MGR estimators of $\boldsymbol{\mu}$ and $\Xi$ are given by

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}, \quad \hat{\Xi}_{\boldsymbol{\theta}} = \boldsymbol{M}_{\boldsymbol{\theta}}^{-1} \boldsymbol{X}'\boldsymbol{Y} \ (\boldsymbol{M}_{\boldsymbol{\theta}} = \boldsymbol{M} + \boldsymbol{Q}\Theta\boldsymbol{Q}'), \tag{2.2}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$, $\Theta = \mathrm{diag}(\theta_1, \dots, \theta_k)$ and $\theta_j \in \mathbb{R}_+ = \{\theta \in \mathbb{R} \mid \theta \geq 0\}$ $(j = 1, \dots, k)$ is a regularization parameter called a ridge parameter. Since $\boldsymbol{M}_{\boldsymbol{\theta}} = \boldsymbol{M}$ when $\boldsymbol{\theta} = \boldsymbol{0}_k$, $\hat{\Xi}_{\boldsymbol{\theta}}$ coincides with $\hat{\Xi}$ in (1.2) when $\boldsymbol{\theta} = \boldsymbol{0}_k$ and the MGR estimators coincide with the GR estimators when $p = 1$. The MGR estimators in (2.2) denote the minimizers of the following penalized RSS (PRSS):

$$\mathrm{tr}\left\{(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\Xi)'(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\Xi) + \Xi'\boldsymbol{Q}\Theta\boldsymbol{Q}'\Xi\right\}. \tag{2.3}$$

Although the ridge parameters adjust the amount of shrinkage of the MGR estimator of $\Xi$, since they are unknown, their optimization is an important task to obtain a better estimator. To simplify calculation, following Yanagihara (2018) and Ohishi *et al.* (2020a), we transform the ridge parameters as

$$\delta_j = \frac{\theta_j}{d_j + \theta_j} \in [0, 1] \quad (j = 1, \ldots, k).$$

Since this transformation is a one-to-one correspondence, the optimization of $\theta_j$ is equal to that of $\delta_j$. Hence, we optimize $\delta_j$ instead of $\theta_j$ and we also call $\delta_j$ a ridge parameter in this paper. Let $\boldsymbol{\delta}$ and $\boldsymbol{\Delta}$ be a $k$-dimensional vector and a $k \times k$ diagonal matrix of the ridge parameters defined by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)'$ and $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1, \ldots, \delta_k)$, respectively, and let $\boldsymbol{Z}$ be a $k \times p$ matrix defined by

$$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)' = \boldsymbol{P}_1'\boldsymbol{Y}. \tag{2.4}$$

Then, the MGR estimator of $\boldsymbol{\Xi}$ in (2.2) can be rewritten as

$$\hat{\boldsymbol{\Xi}}_\delta = \boldsymbol{Q}(\boldsymbol{I}_k - \boldsymbol{\Delta})\boldsymbol{D}^{-1/2}\boldsymbol{Z} = \hat{\boldsymbol{\Xi}} - \boldsymbol{Q}\boldsymbol{\Delta}\boldsymbol{D}^{-1/2}\boldsymbol{Z}. \tag{2.5}$$

In this paper, we optimize the ridge parameter $\boldsymbol{\delta}$ by using the MSC minimization method.

The MGR estimator in (2.5) gives a predictive matrix of $\boldsymbol{Y}$ as

$$\hat{\boldsymbol{Y}}_\delta = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}' + \boldsymbol{X}\hat{\boldsymbol{\Xi}}_\delta = \boldsymbol{H}_\delta\boldsymbol{Y}, \quad \boldsymbol{H}_\delta = \boldsymbol{J}_n + \boldsymbol{P}_1(\boldsymbol{I}_k - \boldsymbol{\Delta})\boldsymbol{P}_1',$$

where $\boldsymbol{J}_n = \boldsymbol{1}_n\boldsymbol{1}_n'/n$ and $\boldsymbol{H}_\delta$ is an $n \times n$ matrix called a hat matrix. Most MSCs consist of the predictive matrix and the hat matrix. The predictive matrix is used to evaluate model fit. We define an estimator and an unbiased estimator of the covariance matrix $\boldsymbol{\Sigma}$ as

$$\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta}) = \frac{1}{n}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\delta)'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_\delta), \quad \boldsymbol{S} = \frac{1}{b}\hat{\boldsymbol{\Sigma}}_0 \left(\hat{\boldsymbol{\Sigma}}_0 = \hat{\boldsymbol{\Sigma}}(\boldsymbol{0}_k),\ b = 1 - (k+1)/n\right). \tag{2.6}$$

Under normality, $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})$ is a penalized MLE of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_0$ is an MLE of $\boldsymbol{\Sigma}$. Then, model fit, i.e., the distance between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}_\delta$ is defined by

$$\mathrm{tr}\left\{\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})\boldsymbol{S}^{-1}\right\}.$$

On the other hand, the hat matrix is used to evaluate model complexity and it is defined by the following GDF:

$$\mathrm{df}(\boldsymbol{\delta}) = p\,\mathrm{tr}(\boldsymbol{H}_\delta). \tag{2.7}$$

The $GC_p$ and EGCV criteria for optimizing ridge parameters consist of $\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})\boldsymbol{S}^{-1}\}$ and $\mathrm{df}(\boldsymbol{\delta})$. Similar to Yanagihara (2018), we have the following lemma about $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})$ and $\mathrm{df}(\boldsymbol{\delta})$.

**Lemma 1.** *Let $\boldsymbol{B}_\delta$ and $\boldsymbol{W}$ be $p \times p$ matrices defined by*

$$\boldsymbol{B}_\delta = \boldsymbol{Z}'\boldsymbol{\Delta}^2\boldsymbol{Z}, \quad \boldsymbol{W} = n\hat{\boldsymbol{\Sigma}}_0.$$

*Then,* $\hat{\Sigma}(\boldsymbol{\delta})$ *and* $\mathrm{df}(\boldsymbol{\delta})$ *can be partitioned into terms which do and do not include* $\boldsymbol{\delta}$ *as follows:*

$$\hat{\Sigma}(\boldsymbol{\delta}) = \frac{1}{n}(\boldsymbol{W} + \boldsymbol{B}_{\delta}) = \hat{\Sigma}_0 + \frac{1}{n}\sum_{j=1}^{k} \boldsymbol{z}_j\boldsymbol{z}_j'\delta_j^2,$$

$$\mathrm{df}(\boldsymbol{\delta}) = p(1+k) - p\,\mathrm{tr}\,\boldsymbol{\Delta} = p\left\{(1+k) - \sum_{j=1}^{k}\delta_j\right\}.$$

From Lemma 1, we have

$$\mathrm{tr}\left\{\hat{\Sigma}(\boldsymbol{\delta})\boldsymbol{S}^{-1}\right\} = b\,\mathrm{tr}\left(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}\right) + bp, \quad \boldsymbol{B}_{\boldsymbol{\delta}}^{*} = \boldsymbol{W}^{-1/2}\boldsymbol{B}_{\boldsymbol{\delta}}\boldsymbol{W}^{-1/2}.$$

Then, the $GC_p$ and EGCV criteria for optimizing ridge parameters are defined by

$$GC_p(\boldsymbol{\delta}) = nb\,\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) + nbp + \alpha\,\mathrm{df}(\boldsymbol{\delta}),$$
$$\mathrm{EGCV}(\boldsymbol{\delta}) = \frac{b\,\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) + bp}{\{1 - \mathrm{df}(\boldsymbol{\delta})/np\}^{\alpha}},$$

where $\alpha$ is a positive value adjusting the strength of the penalty for model complexity. Existing criteria are expressed by changing the value of $\alpha$, for example, the $GC_p$ and EGCV criteria coincide with the $C_p$ and GCV criteria, respectively, when $\alpha = 2$ and the $GC_p$ criterion coincides with the $MC_p$ criterion (Yanagihara *et al.*, 2009) when $\alpha = 2\{1 + (p+1)/(n-k-p-2)\}$. From the above, MSCs for optimizing ridge parameters can be regarded as bivariate functions of $\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*})$ and $\mathrm{df}(\boldsymbol{\delta})$. Lemma 1 gives ranges of $\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*})$ and $\mathrm{df}(\boldsymbol{\delta})$.

**Lemma 2.** *The* $\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*})$ *and* $\mathrm{df}(\boldsymbol{\delta})$ *are included in the following ranges:*

$$\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) \in [0, \mathrm{tr}\,(\boldsymbol{Z}^{*}\boldsymbol{Z}^{*\prime})], \quad \mathrm{df}(\boldsymbol{\delta}) \in [p, p(1+k)],$$

*where* $\boldsymbol{Z}^{*} = \boldsymbol{Z}\boldsymbol{W}^{-1/2}$.

Moreover, let $f$ be a bivariate function defined by the following class.

**Definition 1. (Class of the bivariate function $f$)** For a positive value $r_+$, $f$ satisfies the following conditions:

(A1) For any $(r, u) \in [0, r_+] \times [p, np)$, $f(r, u)$ is continuous.

(A2) For any $(r, u) \in [0, r_+] \times [p, np)$, $f(r, u)$ is first order partially differentiable and its partial derivatives are positive.

We define MSC for optimizing ridge parameters by using $f$ in Definition 4 as

$$\mathrm{MSC}(\boldsymbol{\delta}) = f\left(\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}), \mathrm{df}(\boldsymbol{\delta})\right). \tag{2.8}$$

For the $GC_p$ and EGCV criteria, $f$ is given by

$$f(r,u) = \begin{cases} f_{GC_p}(r,u) = nb(r+p) + \alpha u & (GC_p \text{ criterion}) \\ f_{\mathrm{EGCV}}(r,u) = b(r+p)/(1-u/np)^{\alpha} & (\text{EGCV criterion}) \end{cases},$$

and $r_+$ is given by

$$r_+ = \mathrm{tr}\,(\boldsymbol{Z}^* \boldsymbol{Z}^{*\prime}).$$

Then, the optimal ridge parameters based on minimizing the MSC in (2.8) are given by

$$\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_k)' = \arg \min_{\boldsymbol{\delta} \in [0,1]^k} \mathrm{MSC}(\boldsymbol{\delta}).$$

## 3. Fast Optimization of Ridge Parameters

In this section, to obtain $\boldsymbol{\delta}$ minimizing the MSC in (2.8), we extend the algorithm for optimizing ridge parameters in the GR regression proposed by Ohishi *et al.* (2020a). First, we define the following class of ridge parameters.

**Definition 2.（Class of ridge parameters）** For $h \in \mathbb{R}_+$, a class of ridge parameters is defined by

$$\hat{\boldsymbol{\delta}}(h) = \left(\hat{\delta}_1(h), \ldots, \hat{\delta}_k(h)\right)', \quad \hat{\delta}_j(h) = 1 - \mathrm{soft}\left(1, h/\boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j\right),$$

where $\boldsymbol{z}_j$ is the $p$-dimensional vector defined by (2.4). Furthermore, $\mathrm{soft}(x, a)$ is a soft-thresholding operator (e.g., Donoho & Johnstone, 1994), i.e., $\mathrm{soft}(x, a) = \mathrm{sign}(x)(|x| - a)_+$, and $(x)_+ = \max\{x, 0\}$.

When $\boldsymbol{S} = \boldsymbol{I}_p$ and $p = 1$, the class of ridge parameters in Definition 2 corresponds to that for the GR regression defined by Ohishi *et al.* (2020a). Using this class, the MGR estimator in (2.5) is given as a function of $h$:

$$\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\delta}}(h)} = \boldsymbol{Q}\boldsymbol{V}(h)\boldsymbol{Q}'\hat{\boldsymbol{\Xi}},$$

where $\boldsymbol{Q}$ is the $k \times k$ orthogonal matrix defined by (2.1) and $\boldsymbol{V}(h)$ is a $k \times k$ diagonal matrix which has the following diagonal elements:

$$v_j(h) = 1 - \hat{\delta}_j(h) = \mathrm{soft}\left(1, h/\boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j\right) \quad (j = 1, \ldots, k).$$

The $\boldsymbol{V}(h)$ rewrites the predictive matrix of $\boldsymbol{Y}$ as

$$\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}(h)} = \{\boldsymbol{J}_n + \boldsymbol{P}_1 \boldsymbol{V}(h) \boldsymbol{P}_1'\} \boldsymbol{Y},$$

where $\boldsymbol{P}_1$ is the $n \times k$ matrix defined by (2.1). Then, the ridge parameters optimized by the MSC minimization method are given by the following theorem (the proof is given in Appendix A.1).

**Theorem 1.** *We define $r_+$ as*

$$r_+ = \mathrm{tr}\,(\boldsymbol{Z}^* \boldsymbol{Z}^{*\prime})\,.$$

*For $f$ with the class in Definition 4, let $\phi(h)$ $(h \in \mathbb{R}_+ \backslash \{0\})$ be a function defined by*

$$\phi(h) = \mathrm{MSC}(\hat{\boldsymbol{\delta}}(h)),$$

*and suppose that $\exists v > 0$ s.t. $\phi(v) < \lim_{h \to 0} \phi(h)$. Then, the ridge parameters optimized by the MSC minimization method are given by $\hat{\boldsymbol{\delta}}(\hat{h})$ and $\hat{h}$ is given by*

$$\hat{h} = \arg \min_{h \in \mathbb{R}_+ \backslash \{0\}} \phi(h).$$

From this theorem, the class of ridge parameters in Definition 2 is the class of the "optimal" ridge parameters.

Let $t_j$ $(j = 1, \ldots, k)$ be the $j$th order statistic of $\boldsymbol{z}_1' \boldsymbol{S}^{-1} \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k' \boldsymbol{S}^{-1} \boldsymbol{z}_k$ and $R_j$ $(j = 0, 1, \ldots, k)$ be a range defined by

$$R_j = \begin{cases} (0, t_1] & (j = 0) \\ (t_j, t_{j+1}] & (j = 1, \ldots, k-1) \\ (t_k, \infty] & (j = k) \end{cases}. \tag{3.1}$$

Then, similar to Ohishi *et al.* (2020a), we have the following proposition.

**Proposition 1.** *The $\phi(h)$ in Theorem 1 satisfies the following properties:*

(P1) *For all $h \in \mathbb{R}_+ \backslash \{0\}$, $\phi(h)$ is continuous.*

(P2) *For all $h \geq t_k$, $\phi(h) = f(r_+, p)$.*

(P3) *The $\phi(h)$ can be expressed as the following piecewise function:*

$$\phi(h) = \phi_a(h) = f\left((c_{1,a} + c_{2,a} h^2)/nb, p(1 + k - a - c_{2,a} h)\right) \quad (h \in R_a;\ a = 0, 1, \ldots, k),$$

*where $c_{1,a}$ and $c_{2,a}$ are nonnegative constants given by*

$$c_{1,a} = \begin{cases} 0 & (a = 0) \\ \displaystyle\sum_{j=1}^{a} t_j & (a = 1, \ldots, k) \end{cases}, \quad c_{2,a} = \begin{cases} \displaystyle\sum_{j=a+1}^{k} \frac{1}{t_j} & (a = 0, 1, \ldots, k-1) \\ 0 & (a = k) \end{cases}.$$

From the results, the MSC minimization problem for optimizing ridge parameters in the MGR regression can be solved by applying the fast algorithm for the GR regression proposed by Ohishi *et al.* (2020a). That is, we have the following theorem.

**Theorem 2.** *Suppose that the derivative of $\phi_a(h)$ in Proposition 1 is expressed as*

$$\frac{d}{dh}\phi_a(h) = \chi_a(h)\psi_a(h) \quad (h \in R_a;\ a = 0, 1, \dots, k-1),$$

*and $\psi(h) = \psi_a(h)$ $(h \in R_a)$ is continuous for all $h \in \mathbb{R}_+ \backslash \{0\}$, where $\chi_a(h)$ is a positive function and $\psi_a(h)$ is a polynomial. Moreover, suppose that $\exists v > 0$ s.t. $\phi(v) < \lim_{h\to 0} \phi(h)$ and let $h_a$ be a root of $\psi_a(h) = 0$ satisfying*

$$\exists \epsilon_a > 0 \ s.t. \ \forall \epsilon \in (0, \epsilon_a),\ \psi_a(h_a - \epsilon) < 0. \tag{3.2}$$

*Then, minimizer candidates of $\phi(h)$ are given by*

$$\mathcal{S} = \left\{ \bigcup_{a \in \mathcal{A}} \{h_a\} \right\} \bigcup \mathcal{T},$$

$$\mathcal{A} = \{a \in \{0, 1, \dots, k-1\} \mid h_a \in R_a\}, \quad \mathcal{T} = \begin{cases} \{t_k\} & (\psi_{k-1}(t_k) < 0) \\ \emptyset & (\psi_{k-1}(t_k) \geq 0) \end{cases}.$$

*Hence, the ridge parameters optimized by the MSC minimization method are given by $\hat{\boldsymbol{\delta}}(\hat{h})$ and $\hat{h}$ is given by*

$$\hat{h} = \arg\min_{h \in \mathcal{S}} \phi(h).$$

Although the range of $h$ is a set of positive values, Theorem 2 can reduce a search range of $h$ to $\mathcal{S}$ which is a set of discrete points. Furthermore, each element of $\mathcal{S}$ is given as closed form and $\#(\mathcal{S}) \leq k + 1$; hence we can quickly optimize the ridge parameters. In the theorem, although $\psi_a(h)$ is implicitly supposed as a linear or quadratic function, the theorem can naturally be extended to higher order polynomial functions. In particular, roots of $\psi_a(h) = 0$ can be obtained as closed forms when $\psi_a(h)$ is a cubic or a quartic function, by using Cardano's formula (e.g., David, 2004, Chap. 1) or Ferrari's method (e.g., Tignol, 2001, Chap. 3). Hence, if the degree of $\psi_a(h)$ is four or less, we can quickly optimize the MSC.

### 3.1. Examples

In this subsection, we provide specific examples of the MSC minimization methods for optimizing ridge parameters in the MGR regression. To emphasize that the optimal ridge parameters depend on $\alpha$, we specify that $\alpha$ is given.

### 3.1.1. The $GC_p$ criterion

Although the ridge parameters optimized by the $GC_p$ criterion minimization method have already been given by Nagai *et al.* (2012), here we show how to derive them by applying Theorem 2. The $GC_p$ criterion for optimizing ridge parameters is given by

$$GC_p(\boldsymbol{\delta} \mid \alpha) = f_{GC_p}\left(\mathrm{tr}(\boldsymbol{B}^*_{\boldsymbol{\delta}}), \mathrm{df}(\boldsymbol{\delta}) \,\middle|\, \alpha\right).$$

When $h \in R_a$ $(a = 0, 1, \ldots, k)$, $\phi$ and its derivative are given by

$$\begin{aligned} \phi(h \mid \alpha) = \phi_a(h \mid \alpha) &= c_{2,a}h^2 - \alpha p c_{2,a} h + nbp + c_{1,a} + \alpha p(1 + k - a), \\ \frac{d}{dh}\phi_a(h \mid \alpha) &= c_{2,a}(2h - \alpha p). \end{aligned}$$

Hence, the ridge parameters optimized by the $GC_p$ criterion minimization method are given as the following closed form:

$$\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\hat{h}_\alpha), \quad \hat{h}_\alpha = \frac{\alpha p}{2}.$$

### 3.1.2. The EGCV criterion

The EGCV criterion for optimizing ridge parameters is given by

$$\mathrm{EGCV}(\boldsymbol{\delta} \mid \alpha) = f_{\mathrm{EGCV}}\left(\mathrm{tr}(\boldsymbol{B}^*_{\boldsymbol{\delta}}), \mathrm{df}(\boldsymbol{\delta}) \,\middle|\, \alpha\right).$$

When $h \in R_a$ $(a = 0, 1, \ldots, k)$, $\phi$ and its derivative are given by

$$\begin{aligned} \phi(h \mid \alpha) = \phi_a(h \mid \alpha) &= \frac{bp + (c_{1,a} + c_{2,a}h^2)/n}{\{b + (a + c_{2,a}h)/n\}^\alpha}, \\ \frac{d}{dh}\phi_a(h \mid \alpha) &= \frac{c_{2,a}}{n^2\{b + (a + c_{2,a}h)/n\}^{\alpha+1}}\psi_a(h \mid \alpha), \\ \psi_a(h \mid \alpha) &= -(\alpha - 2)c_{2,a}h^2 + 2(a + nb)h - \alpha(nbp + c_{1,a}). \end{aligned}$$

When $\alpha = 2$, i.e., using the GCV criterion minimization method, we have

$$\psi_a(h \mid 2) = 2\{(a + nb)h - nbp - c_{1,a}\},$$

and a root of $\psi_a(h \mid 2) = 0$ is

$$h_a = \frac{nbp + c_{1,a}}{a + nb}.$$

Moreover, similar to Yanagihara (2018), the following statement is true:

$$\exists! a^* \in \{0, 1, \ldots, k - 1\} \ s.t. \ h_{a^*} \in R_{a^*}.$$

Hence, the ridge parameters optimized by the GCV criterion minimization method are given by the following closed forms: $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(h_{a^*})$.

When $\alpha > 2$, since $\psi_a(h \mid \alpha)$ is a concave quadratic function, a root of $\psi_a(h \mid \alpha) = 0$ satisfying the condition (3.2) is given by

$$h_{\alpha,a} = \frac{(a + nb) - \sqrt{(a + nb)^2 - \alpha(\alpha - 2)c_{2,a}(nbp + c_{1,a})}}{(\alpha - 2)c_{2,a}}.$$

Therefore, candidates of $\hat{h}_\alpha$ are given by

$$\mathcal{S}_\alpha = \left\{ \bigcup_{a \in \mathcal{A}_\alpha} \{h_{\alpha,a}\} \right\} \bigcup \mathcal{T}_\alpha,$$

where $\mathcal{A}_\alpha$ and $\mathcal{T}_\alpha$ are sets given by

$$\mathcal{A}_\alpha = \{a \in \{0, 1, \dots, k-1\} \mid h_{\alpha,a} \in R_a\}, \quad \mathcal{T}_\alpha = \begin{cases} \{t_k\} & \left(r_+ > 2(1 - n^{-1})t_k/\alpha b - p\right) \\ \emptyset & \left(r_+ \le 2(1 - n^{-1})t_k/\alpha b - p\right) \end{cases}.$$

Hence, the ridge parameters optimized by the EGCV criterion minimization method are given by

$$\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}(\hat{h}_\alpha), \quad \hat{h}_\alpha = \arg\min_{h \in \mathcal{S}_\alpha} \phi(h \mid \alpha).$$

In the EGCV criterion minimization method, the number of minimizer candidates is only $k + 1$ at most.

### 3.2. Relationships between the Optimal Ridge Parameters

This subsection provides some theoretical properties concerning the relationships between the optimal ridge parameters. The class of the optimal ridge parameters satisfies

$$\forall h_1, h_2 \in \mathbb{R}_+, h_1 < h_2 \Rightarrow \hat{\delta}_j(h_1) \le \hat{\delta}_j(h_2) \quad (j = 1, \dots, k),$$

with equality only when $h_1 \ge t_k$. This fact yields some relationships concerning the ridge parameters optimized by the $GC_p$ and EGCV criteria minimization methods. Immediately, we have the following result which is similar to Nagai *et al.* (2012).

**Proposition 2.** *For positive values $\alpha_1$ and $\alpha_2$, we define the ridge parameters optimized by the $GC_p$ criterion minimization method as*

$$\hat{\delta}_{1,j} = \hat{\delta}_j(\hat{h}_{\alpha_1}), \quad \hat{\delta}_{2,j} = \hat{\delta}_j(\hat{h}_{\alpha_2}) \quad (j = 1, \dots, k),$$

*where $\hat{h}_\alpha = \alpha p/2$. Then, we have*

$$\alpha_1 < \alpha_2 \Rightarrow \hat{\delta}_{1,j} \le \hat{\delta}_{2,j}.$$

This proposition states that the stronger the penalty for model complexity, the larger the amount of shrinkage of the estimator, when using the $GC_p$ criterion minimization method. Next, we consider the ridge parameters optimized by the $GC_p$ and the GCV criteria minimization methods. Similar to Yanagihara (2018), we have the following lemma.

**Lemma 3.** *The $h_{a^*}$ obtained by the GCV criterion minimization method satisfies $h_{a^*} \leq p$.*

This lemma leads to the following result which is similar to the case when $p = 1$ (Yanagihara, 2018).

**Proposition 3.** *Let $\hat{\delta}_{\alpha,j}^{GC_p}$ and $\hat{\delta}_j^{\mathrm{GCV}}$ $(j = 1, \ldots, k)$ be the ridge parameters optimized by the $GC_p$ and GCV criteria minimization methods, respectively. Then, we have*

$$\alpha \geq 2 \Rightarrow \hat{\delta}_j^{\mathrm{GCV}} \leq \hat{\delta}_{\alpha,j}^{GC_p}.$$

The value of $\alpha$ in the MSC is often 2 or more. This means that the ridge parameters optimized by the $GC_p$ criterion minimization method shrink the estimator more than the GCV criterion minimization method in most cases. Finally, we consider the ridge parameters optimized by the EGCV criterion minimization method. We express $\phi(h \mid \alpha) = \mathrm{EGCV}(\hat{\boldsymbol{\delta}}(h) \mid \alpha)$ as

$$\phi(h \mid \alpha) = \hat{\sigma}^2(h)\eta(h \mid \alpha),$$

where

$$\hat{\sigma}^2(h) = bp + b\,\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^*), \quad \eta(h \mid \alpha) = \frac{1}{\{1 - \mathrm{df}(h)/np\}^{\alpha}}, \quad \mathrm{df}(h) = \mathrm{df}(\hat{\boldsymbol{\delta}}(h)),$$

and let $\hat{h}_\alpha$ be the minimizer of $\phi(h \mid \alpha)$. Then, $\eta(h \mid \alpha)$ has the following property (the proof is given in Appendix A.2).

**Lemma 4.** *Suppose that $0 < h_1 < h_2$. Then, we have*

$$\eta(h_2 \mid \alpha) \leq \eta(h_1 \mid \alpha).$$

This lemma leads to the following proposition (the proof is given in Appendix A.3).

**Proposition 4.** *The EGCV criterion minimization method has the following properties:*

(1) *Suppose that $\alpha_1 < \alpha_2$. Then, we have*

$$\hat{h}_{\alpha_1} = t_k \Rightarrow \hat{h}_{\alpha_2} = t_k.$$

(2) *For positive values $\alpha_1$ and $\alpha_2$, we define the ridge parameters optimized by the EGCV criterion minimization method as*

$$\hat{\delta}_{1,j} = \hat{\delta}_j(\hat{h}_{\alpha_1}), \quad \hat{\delta}_{2,j} = \hat{\delta}_j(\hat{h}_{\alpha_2}) \quad (j = 1, \dots, k),$$

*and suppose that* $\hat{h}_{\alpha_2} \neq t_k$. *Then, we have*

$$\alpha_1 < \alpha_2 \Rightarrow \hat{\delta}_{1,j} \leq \hat{\delta}_{2,j},$$

*with equality only when* $\hat{h}_{\alpha_1} \geq \boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j$.

This proposition states that the stronger the penalty for model complexity, the larger the amount of shrinkage of the estimator, when using the EGCV criterion minimization method.

## 4. Extending the MSC Class

In the previous section, we showed that the algorithm for the GR regression can be applied to minimize the MSC in (2.8), where the distance between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{\delta}}$ is defined by $\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})\boldsymbol{S}^{-1}\}$ and the MSC is defined by using $\mathrm{tr}(\boldsymbol{B}_{\boldsymbol{\delta}}^*)$ obtained from the distance. In this section, we focus on how to measure the distance.

Let $g$ be a real-valued function defined by the following class.

**Definition 3. (Class of the function** $g$**)** For any $p \times p$ positive definite matrix $\boldsymbol{A}$, the $g$ satisfies the following conditions:

(A1) The $g(\boldsymbol{A})$ is positive.

(A2) The $\partial g(\boldsymbol{A})/\partial \boldsymbol{A}$ is a positive definite.

Using the function $g$, we extend the MSC in (2.8) to

$$\mathrm{MSC}(\boldsymbol{\delta} \mid g) = f\left(g(\boldsymbol{B}_{\boldsymbol{\delta}}^*), \mathrm{df}(\boldsymbol{\delta})\right), \tag{4.1}$$

where $f$ is the bivariate function given by Definition 4. For example, $g$ includes the following functions:

$$g(\boldsymbol{A}) = \begin{cases} g_{\mathrm{LH}}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}) & \text{(LH-distance)} \\ g_{\mathrm{LR}}(\boldsymbol{A}) = \log\left|\boldsymbol{I}_p + \boldsymbol{A}\right| & \text{(LR-distance)} \\ g_{\mathrm{BNP}}(\boldsymbol{A}) = \mathrm{tr}\left\{\boldsymbol{A}(\boldsymbol{I}_p + \boldsymbol{A})^{-1}\right\} & \text{(BNP-distance)} \\ g_{\mathrm{ML}}(\boldsymbol{A}) = \mathrm{tr}\left\{(\boldsymbol{I}_p + \boldsymbol{A})^{-1}\right\} + \log|\boldsymbol{I}_p + \boldsymbol{A}| - p & \text{(ML-distance)} \\ g_{\mathrm{GLS}}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^2)/2 & \text{(GLS-distance)} \end{cases}.$$

The MSC in (4.1) is equal to that in (2.8) when $g(\boldsymbol{A}) = g_{\mathrm{LH}}(\boldsymbol{A})$ and the following equation holds:

$$g_{\mathrm{LH}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) = \mathrm{tr}\left(\boldsymbol{B}_{\boldsymbol{\delta}}\boldsymbol{W}^{-1}\right).$$

Since we can regard $\boldsymbol{B}_{\boldsymbol{\delta}}$ as a between-group variation matrix and $\boldsymbol{W}$ as a within-group variation matrix, $g_{\mathrm{LH}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*})$ is a Lawley-Hotelling trace criterion (LH-statistic; e.g., Anderson, 2003, Chap. 8) which is a well-known statistic in multivariate analysis. That is, the MSC in (2.8) measures the distance between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}_{\boldsymbol{\delta}}$ based on the LH-statistic. Similarly, regarding the LR-distance and the BNP-distance, the following equations hold:

$$g_{\mathrm{LR}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) = \log\left|(\boldsymbol{W}+\boldsymbol{B}_{\boldsymbol{\delta}})\boldsymbol{W}^{-1}\right|, \quad g_{\mathrm{BNP}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) = \mathrm{tr}\left\{\boldsymbol{B}_{\boldsymbol{\delta}}(\boldsymbol{W}+\boldsymbol{B}_{\boldsymbol{\delta}})^{-1}\right\}.$$

They are a Likelihood-Ratio criterion and a Bartlett-Nanda-Pillai trace criterion, respectively, which are also well-known statistics (e.g., Anderson, 2003, Chap. 8). MSC based on the LR-distance includes the GIC and the $\mathrm{AIC_C}$ under normality. The above three distances based on the three statistics pertain to the mean structure of a model. In contrast, there are distances with respect to the covariance structure of a model, e.g., the ML-distance and the GLS-distance. Regarding these distances, the following equations hold:

$$\begin{aligned} g_{\mathrm{ML}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) &= \log\left|\hat{\Sigma}(\boldsymbol{\delta})\right| + \mathrm{tr}\left\{\hat{\Sigma}(\boldsymbol{\delta})^{-1}\hat{\Sigma}_0\right\} - \log\left|\hat{\Sigma}_0\right| - p,\\ g_{\mathrm{GLS}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) &= \frac{1}{2}\,\mathrm{tr}\left[\left\{\left(\hat{\Sigma}_0 - \hat{\Sigma}(\boldsymbol{\delta})\right)\hat{\Sigma}_0^{-1}\right\}^2\right]. \end{aligned}$$

They are distances between $\hat{\Sigma}(\boldsymbol{\delta})$ and $\hat{\Sigma}_0$ called a maximum likelihood fitting function and a generalized least square fitting function, respectively (e.g., Bollen, 1989, Chap. 4). Using $g(\boldsymbol{A})$, the $GC_p$ and EGCV criteria, and the GIC and the $\mathrm{AIC_C}$ under normality are given by

$$\begin{aligned} GC_p(\boldsymbol{\delta}) &= nbg_{\mathrm{LH}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) + nbp + \alpha\,\mathrm{df}(\boldsymbol{\delta}),\\ \mathrm{EGCV}(\boldsymbol{\delta}) &= \frac{bg_{\mathrm{LH}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) + bp}{\{1-\mathrm{df}(\boldsymbol{\delta})/np\}^{\alpha}},\\ \mathrm{GIC}(\boldsymbol{\delta}) &= ng_{\mathrm{LR}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) + np\log b + \alpha\,\mathrm{df}(\boldsymbol{\delta}),\\ \mathrm{AIC_C}(\boldsymbol{\delta}) &= ng_{\mathrm{LR}}(\boldsymbol{B}_{\boldsymbol{\delta}}^{*}) + np\log b + \frac{np\{n+\mathrm{df}(\boldsymbol{\delta})\}}{n-p-1-\mathrm{df}(\boldsymbol{\delta})}. \end{aligned}$$

Using the GIC, it is also possible to adjust the strength of the penalty for model complexity, and for example, the GIC coincides with the AIC when $\alpha = 2$, the HQC (Hannan & Quinn, 1979) when $\alpha = 2\log\log n$, and the BIC (Schwarz, 1978) when $\alpha = \log n$. For the GIC and $\mathrm{AIC_C}$, the bivariate function $f(r,u)$ is given by

$$f(r,u) = \begin{cases} f_{\mathrm{GIC}}(r,u) = n(r + p\log b) + \alpha u & (\mathrm{GIC})\\ f_{\mathrm{AIC_C}}(r,u) = n(r + p\log b) + \dfrac{np(n+u)}{n-p-1-u} & (\mathrm{AIC_C}) \end{cases}.$$

The following subsections describe two algorithms to minimize the MSC in (4.1).

### 4.1. Minimizing MSC via Iterative Method

This subsection describes an algorithm for solving the MSC minimization method via an iterative method with an iterative function. That is, we derive the iterative function. Notice that

$$\boldsymbol{B}_{\boldsymbol{\delta}}^* = \sum_{j=1}^{k} \boldsymbol{z}_j^* \boldsymbol{z}_j^{*\prime} \delta_j^2, \quad \boldsymbol{z}_j^* = \boldsymbol{W}^{-1/2} \boldsymbol{z}_j.$$

Therefore, the following partial derivatives can be obtained:

$$\frac{\partial}{\partial \delta_j} \boldsymbol{B}_{\boldsymbol{\delta}}^* = 2\boldsymbol{z}_j^* \boldsymbol{z}_j^{*\prime} \delta_j, \quad \frac{\partial}{\partial \delta_j} \mathrm{df}(\boldsymbol{\delta}) = -p.$$

We express the $(i, \ell)$ element of a matrix $\boldsymbol{A}$ as $a_{i\ell} = (\boldsymbol{A})_{i\ell}$ and define

$$\dot{g}_{i\ell}(\boldsymbol{B}) = \left. \frac{\partial}{\partial a_{i\ell}} g(\boldsymbol{A}) \right|_{\boldsymbol{A}=\boldsymbol{B}}, \quad \dot{\boldsymbol{G}}(\boldsymbol{B}) = \left. \frac{\partial}{\partial \boldsymbol{A}} g(\boldsymbol{A}) \right|_{\boldsymbol{A}=\boldsymbol{B}}.$$

$\boldsymbol{B}_{\boldsymbol{\delta}}^*$ is a symmetric matrix, thus we have

$$\frac{\partial}{\partial \delta_j} g(\boldsymbol{B}_{\boldsymbol{\delta}}^*) = \sum_{i=1}^{p} \sum_{\ell=i}^{p} \frac{\partial}{\partial \delta_j} (\boldsymbol{B}_{\boldsymbol{\delta}}^*)_{i\ell} \cdot \dot{g}_{i\ell}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) = 2\boldsymbol{z}_j^{*\prime} \dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) \boldsymbol{z}_j^* \delta_j.$$

Hence, a partial derivative of the MSC is given by

$$\frac{\partial}{\partial \delta_j} \mathrm{MSC}(\boldsymbol{\delta} \mid g) = 2\boldsymbol{z}_j^{*\prime} \dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) \boldsymbol{z}_j^* \dot{f}_r \left( g(\boldsymbol{B}_{\boldsymbol{\delta}}^*), \mathrm{df}(\boldsymbol{\delta}) \right) \delta_j - p \dot{f}_u \left( g(\boldsymbol{B}_{\boldsymbol{\delta}}^*), \mathrm{df}(\boldsymbol{\delta}) \right),$$

where

$$\dot{f}_r(x, y) = \left. \frac{\partial}{\partial r} f(r, u) \right|_{(r,u)=(x,y)}, \quad \dot{f}_u(x, y) = \left. \frac{\partial}{\partial u} f(r, u) \right|_{(r,u)=(x,y)}.$$

By solving $\partial \, \mathrm{MSC}(\boldsymbol{\delta} \mid g)/\partial \boldsymbol{\delta} = \boldsymbol{0}_k$, we can obtain the following iterative method:

$$\begin{aligned} \boldsymbol{\delta}^{(i+1)} &= \boldsymbol{\zeta}(\boldsymbol{\delta}^{(i)}) = \left( \zeta_1(\boldsymbol{\delta}^{(i)}), \ldots, \zeta_k(\boldsymbol{\delta}^{(i)}) \right)' \quad (i = 0, 1, \ldots), \\ \zeta_j(\boldsymbol{\delta}) &= 1 - \mathrm{soft}\left( 1, \tau(\boldsymbol{\delta}) / \boldsymbol{z}_j^{*\prime} \dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) \boldsymbol{z}_j^* \right), \end{aligned} \tag{4.2}$$

where $(i)$ is the iteration number, $\boldsymbol{\delta}^{(0)}$ is a given initial vector, and $\tau(\boldsymbol{\delta})$ is given by

$$\tau(\boldsymbol{\delta}) = \frac{p \dot{f}_u \left( g(\boldsymbol{B}_{\boldsymbol{\delta}}^*), \mathrm{df}(\boldsymbol{\delta}) \right)}{2 \dot{f}_r \left( g(\boldsymbol{B}_{\boldsymbol{\delta}}^*), \mathrm{df}(\boldsymbol{\delta}) \right)} > 0.$$

By repeating the update of $\boldsymbol{\delta}^{(i)}$ with the iterative function $\boldsymbol{\zeta}$, we can obtain the optimal $\boldsymbol{\delta}$. This iterative method has the following property (the proof is given in Appendix A.4).

**Proposition 5.** *For a k-dimensional vector $\boldsymbol{\epsilon}$ wherein all elements are nonnegative, suppose that*

$$\tau(\boldsymbol{\delta}) \leq \tau(\boldsymbol{\delta} + \boldsymbol{\epsilon}), \quad \boldsymbol{z}_j^{*\prime} \dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) \boldsymbol{z}_j^* \geq \boldsymbol{z}_j^{*\prime} \dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}+\boldsymbol{\epsilon}}^*) \boldsymbol{z}_j^*. \tag{4.3}$$

*Then, the iterative method with iterative function* (4.2) *converges if* $\forall j \in \{1, \ldots, k\}$, $\delta_j^{(1)} \geq \delta_j^{(0)}$. *Furthermore, the iterative method also converges if* $\forall j \in \{1, \ldots, k\}$, $\delta_j^{(1)} \leq \delta_j^{(0)}$.

From this proposition, when assumption (4.3) holds, the iterative method with iterative function (4.2) converges if the initial vector is $\boldsymbol{0}_k$ or $\boldsymbol{1}_k$.

#### 4.1.1. LR-distance

For the MSC based on the LR-distance, the following equation holds:

$$\frac{\partial}{\partial \boldsymbol{A}} g_{\mathrm{LR}}(\boldsymbol{A}) = \frac{\partial}{\partial \boldsymbol{A}} \log |\boldsymbol{I}_p + \boldsymbol{A}| = (\boldsymbol{I}_p + \boldsymbol{A})^{-1}.$$

Therefore, we have

$$\boldsymbol{W}^{-1/2} \dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) \boldsymbol{W}^{-1/2} = (\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})^{-1} = \frac{1}{n} \hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}.$$

Hence, the iterative function for solving the MSC minimization method based on the LR-distance is given by

$$\zeta_j(\delta) = 1 - \mathrm{soft}\left(1, n\tau(\boldsymbol{\delta}) / \boldsymbol{z}_j^{\prime} \hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1} \boldsymbol{z}_j\right). \tag{4.4}$$

Furthermore, from Lemma 1, for $\boldsymbol{\epsilon}$ in Proposition 5 and for any $p$-dimensional vector $\boldsymbol{a}$, the following equation holds:

$$\boldsymbol{a}^{\prime} \hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta}) \boldsymbol{a} \leq \boldsymbol{a}^{\prime} \hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta} + \boldsymbol{\epsilon}) \boldsymbol{a} \Longleftrightarrow \boldsymbol{a}^{\prime} \hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1} \boldsymbol{a} \geq \boldsymbol{a}^{\prime} \hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta} + \boldsymbol{\epsilon})^{-1} \boldsymbol{a}.$$

Let $\hat{\boldsymbol{\delta}}^{\mathrm{LR}}$ be a solution obtained by the iterative method with iterative function (4.4). Then,

$$\hat{\boldsymbol{\delta}}^{\mathrm{LR}} = \boldsymbol{\zeta}(\hat{\boldsymbol{\delta}}^{\mathrm{LR}}).$$

The ridge parameters optimized by the MSC minimization method based on the LR-distance are given by

$$\hat{\delta}_j^{\mathrm{LR}} = 1 - \mathrm{soft}\left(1, n\tau(\hat{\boldsymbol{\delta}}^{\mathrm{LR}}) / \boldsymbol{z}_j^{\prime} \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\delta}}^{\mathrm{LR}})^{-1} \boldsymbol{z}_j\right) \quad (j = 1, \ldots, k).$$

On the other hand, the ridge parameters optimized by the MSC minimization method based on the LH-distance are given by the following form:

$$\hat{\delta}_j^{\mathrm{LH}} = 1 - \mathrm{soft}\left(1, \hat{h} / \boldsymbol{z}_j^{\prime} \boldsymbol{S}^{-1} \boldsymbol{z}_j\right) \quad (j = 1, \ldots, k).$$

The $\hat{\delta}_j^{\mathrm{LH}}$ includes $\boldsymbol{S}^{-1}$ and $\boldsymbol{S}$ is an estimator of the covariance matrix for the full model. Thus, $\hat{\delta}_j^{\mathrm{LH}}$ has a disadvantage because $\boldsymbol{S}^{-1}$ is unstable when $k$ is large. Whereas, $\hat{\delta}_j^{\mathrm{LR}}$ does not include $\boldsymbol{S}^{-1}$, but rather $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\delta}}^{\mathrm{LR}})^{-1}$ and $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\delta}}^{\mathrm{LR}})$ is an estimator of the covariance matrix adjusted by $\hat{\boldsymbol{\delta}}^{\mathrm{LR}}$. Thus, $\hat{\delta}_j^{\mathrm{LR}}$ has an advantage because $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\delta}}^{\mathrm{LR}})^{-1}$ is stable even when $k$ is large.

**Example 1**

We derive an iterative function for solving the GIC minimization method. From $f(r,u) = f_{\mathrm{GIC}}(r,u)$, we have

$$\dot{f}_r(r,u) = n, \quad \dot{f}_u(r,u) = \alpha,$$

and therefore, $\tau(\boldsymbol{\delta}) = \alpha p/2n$. Hence, the iterative function for the GIC minimization method is given by

$$\zeta_j(\boldsymbol{\delta}) = 1 - \mathrm{soft}\left(1, \alpha p/2\boldsymbol{z}_j'\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\boldsymbol{z}_j\right) \quad (j = 1, \ldots, k). \tag{4.5}$$

Moreover, since $\tau(\boldsymbol{\delta})$ does not depend on $\boldsymbol{\delta}$, from Proposition 5, the iterative method for solving the GIC minimization method converges under an appropriate initial vector.

**Example 2**

We derive an iterative function for solving the $\mathrm{AIC_C}$ minimization method. From $f(r,u) = f_{\mathrm{AIC_C}}(r,u)$, we have

$$\dot{f}_r(r,u) = n, \quad \dot{f}_u(r,u) = \frac{np(2n-p-1)}{(n-p-1-u)^2},$$

and therefore, we have

$$\tau(\boldsymbol{\delta}) = \frac{p^2(2n-p-1)}{2\{n-p-1-\mathrm{df}(\boldsymbol{\delta})\}^2}.$$

Hence, the iterative function for the $\mathrm{AIC_C}$ minimization method is given by

$$\zeta_j(\boldsymbol{\delta}) = 1 - \mathrm{soft}\left(1, \frac{np^2(2n-p-1)}{2\{n-p-1-\mathrm{df}(\boldsymbol{\delta})\}^2\boldsymbol{z}_j'\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\boldsymbol{z}_j}\right) \quad (j = 1, \ldots, k).$$

Moreover, for $\boldsymbol{\epsilon}$ in Proposition 5, the following equation holds:

$$\mathrm{df}(\boldsymbol{\delta}) \geq \mathrm{df}(\boldsymbol{\delta} + \boldsymbol{\epsilon}).$$

Therefore

$$\tau(\boldsymbol{\delta}) \geq \tau(\boldsymbol{\delta} + \boldsymbol{\epsilon}),$$

and thus, the iterative method for solving the $\mathrm{AIC_C}$ minimization method does not satisfy Proposition 5.

### 4.1.2. BNP-distance

For the MSC based on the BNP-distance, the following equation holds:

$$\frac{\partial}{\partial \boldsymbol{A}} g_{\mathrm{BNP}}(\boldsymbol{A}) = \frac{\partial}{\partial \boldsymbol{A}} \mathrm{tr}\left\{\boldsymbol{A}(\boldsymbol{I}_p + \boldsymbol{A})^{-1}\right\} = -\frac{\partial}{\partial \boldsymbol{A}} \mathrm{tr}\left\{(\boldsymbol{I}_p + \boldsymbol{A})^{-1}\right\} = (\boldsymbol{I}_p + \boldsymbol{A})^{-2}.$$

Therefore, we have

$$\boldsymbol{W}^{-1/2}\dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*)\boldsymbol{W}^{-1/2} = (\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})^{-1}\boldsymbol{W}(\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})^{-1} = \frac{1}{n}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\hat{\boldsymbol{\Sigma}}_0\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}.$$

Hence, the iterative function for solving the MSC minimization method based on the BNP-distance is given by

$$\zeta_j(\boldsymbol{\delta}) = 1 - \mathrm{soft}\left(1, \frac{n\tau(\boldsymbol{\delta})}{\boldsymbol{z}_j'\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\hat{\boldsymbol{\Sigma}}_0\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\boldsymbol{z}_j}\right).$$

Accordingly, using the BNP-distance, the optimal ridge parameters are stable even when $k$ is large.

**Example**

As an example of MSC based on the BNP-distance, we consider the following criterion:

$$\mathrm{BNPC}(\boldsymbol{\delta}) = ng_{\mathrm{BNP}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) + \alpha\,\mathrm{df}(\boldsymbol{\delta}).$$

Then, since

$$\dot{f}_r(r, u) = n, \quad \dot{f}_u(r, u) = \alpha,$$

we have $\tau(\boldsymbol{\delta}) = \alpha p/2n$. Hence, the iterative function for solving the BNPC minimization method is given by

$$\zeta_j(\boldsymbol{\delta}) = 1 - \mathrm{soft}\left(1, \frac{\alpha p}{2\boldsymbol{z}_j'\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\hat{\boldsymbol{\Sigma}}_0\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\boldsymbol{z}_j}\right). \tag{4.6}$$

### 4.1.3. ML-distance

For the MSC based on the ML-distance, the following equation holds:

$$\frac{\partial}{\partial \boldsymbol{A}} g_{\mathrm{ML}}(\boldsymbol{A}) = \frac{\partial}{\partial \boldsymbol{A}} \left[\mathrm{tr}\left\{(\boldsymbol{I}_p + \boldsymbol{A})^{-1}\right\} + \log|\boldsymbol{I}_p + \boldsymbol{A}|\right] = -(\boldsymbol{I}_p + \boldsymbol{A})^{-2} + (\boldsymbol{I}_p + \boldsymbol{A})^{-1}.$$

Therefore, we have

$$\begin{aligned}\boldsymbol{W}^{-1/2}\dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*)\boldsymbol{W}^{-1/2} &= (\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})^{-1} - (\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})^{-1}\boldsymbol{W}(\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})^{-1}\\ &= \frac{1}{n}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1} - \frac{1}{n}\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}\hat{\boldsymbol{\Sigma}}_0\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})^{-1}.\end{aligned}$$

Hence, the iterative function for solving the MSC minimization method based on the ML-distance is given by

$$\zeta_j(\delta) = 1 - \mathrm{soft}\left(1, \frac{n\tau(\boldsymbol{\delta})}{\boldsymbol{z}_j'\{\hat{\Sigma}(\boldsymbol{\delta})^{-1} - \hat{\Sigma}(\boldsymbol{\delta})^{-1}\hat{\Sigma}_0\hat{\Sigma}(\boldsymbol{\delta})^{-1}\}\boldsymbol{z}_j}\right).$$

Accordingly, using the ML-distance, the optimal ridge parameters are stable even when $k$ is large.

#### 4.1.4. GLS-distance

For the MSC based on the GLS-distance, the following equation holds:

$$\frac{\partial}{\partial \boldsymbol{A}} g_{\mathrm{GLS}}(\boldsymbol{A}) = \frac{1}{2} \cdot \frac{\partial}{\partial \boldsymbol{A}} \mathrm{tr}\left(\boldsymbol{A}^2\right) = \boldsymbol{A}.$$

Therefore, we have

$$\boldsymbol{W}^{-1/2}\dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*)\boldsymbol{W}^{-1/2} = \boldsymbol{W}^{-1}\boldsymbol{B}_{\boldsymbol{\delta}}\boldsymbol{W}^{-1} = \frac{1}{n}\hat{\Sigma}_0^{-1}\{\hat{\Sigma}(\boldsymbol{\delta}) - \hat{\Sigma}_0\}\hat{\Sigma}_0^{-1}.$$

Hence, the iterative function for solving the MSC minimization method based on the GLS-distance is given by

$$\zeta_j(\delta) = 1 - \mathrm{soft}\left(1, \frac{n\tau(\boldsymbol{\delta})}{\boldsymbol{z}_j'\hat{\Sigma}_0^{-1}\{\hat{\Sigma}(\boldsymbol{\delta}) - \hat{\Sigma}_0\}\hat{\Sigma}_0^{-1}\boldsymbol{z}_j}\right).$$

Since $\hat{\Sigma}_0$ is an estimator of the covariance matrix for the full model, the optimal ridge parameters are unstable when $k$ is large.

### 4.2. Minimizing MSC via Coordinate Descent

In the previous subsection, we described an algorithm to minimize the MSC via the iterative method with an iterative function obtained by solving $\partial\,\mathrm{MSC}(\boldsymbol{\delta} \mid g)/\partial\boldsymbol{\delta} = \boldsymbol{0}_k$. In this subsection, we update $\delta_1, \ldots, \delta_k$ individually, not simultaneously. That is, we minimize the MSC via a coordinate descent algorithm.

#### 4.2.1. LR-distance

We partition $\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}}$ and $\mathrm{df}(\boldsymbol{\delta})$ into

$$\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}} = \boldsymbol{W}_j + \boldsymbol{z}_j\boldsymbol{z}_j'\delta_j^2, \quad \boldsymbol{W}_j = \boldsymbol{W} + \sum_{\ell \neq j}^{k} \boldsymbol{z}_\ell\boldsymbol{z}_\ell'\delta_\ell^2,$$

$$\mathrm{df}(\boldsymbol{\delta}) = q_{1,j} - p\delta_j, \quad q_{1,j} = p\left\{(1 + k) - \sum_{\ell \neq j}^{k} \delta_\ell\right\}.$$

Then, the following equations hold:

$$
\begin{aligned}
\boldsymbol{I}_p + \boldsymbol{B}^*_{\boldsymbol{\delta}} &= \boldsymbol{W}^{-1/2}(\boldsymbol{W} + \boldsymbol{B}_{\boldsymbol{\delta}})\boldsymbol{W}^{-1/2} = \boldsymbol{W}^{-1/2}(\boldsymbol{W}_j + \boldsymbol{z}_j\boldsymbol{z}_j'\delta_j^2)\boldsymbol{W}^{-1/2}, \qquad (4.7)\\
\left|\boldsymbol{W}_j + \boldsymbol{z}_j\boldsymbol{z}_j'\delta_j^2\right| &= \left|\boldsymbol{W}_j\right|\left(1 + \boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{z}_j\delta_j^2\right).
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
g_{\mathrm{LR}}(\boldsymbol{B}^*_{\boldsymbol{\delta}}) &= \log|\boldsymbol{I}_p + \boldsymbol{B}^*_{\boldsymbol{\delta}}| = \log\left(1 + \boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{z}_j\delta_j^2\right) + \log\left|\boldsymbol{W}_j\boldsymbol{W}^{-1}\right|\\
&= \log(1 + q_{2,j}\delta_j^2) + q_{3,j},
\end{aligned}
$$

where $q_{2,j}$ and $q_{3,j}$ are constants which do not depend on $\delta_j$ given by

$$
q_{2,j} = \boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{z}_j, \quad q_{3,j} = \log\left|\boldsymbol{W}_j\boldsymbol{W}^{-1}\right|.
$$

Hence, the following partial derivative is obtained:

$$
\frac{\partial}{\partial\delta_j}g_{\mathrm{LR}}(\boldsymbol{B}^*_{\boldsymbol{\delta}}) = \frac{2q_{2,j}\delta_j}{1 + q_{2,j}\delta_j^2}.
$$

**Example 1**

The partial derivative of the GIC is given by

$$
\dot{f}_j(\delta_j) = \frac{\partial}{\partial\delta_j}\,\mathrm{GIC}(\boldsymbol{\delta}) = \frac{1}{1 + q_{2,j}\delta_j^2}(-\alpha p q_{2,j}\delta_j^2 + 2nq_{2,j}\delta_j - \alpha p).
$$

An update equation of the coordinate descent algorithm for solving the GIC minimization method is given by the following theorem (the proof is given in Appendix A.5).

**Theorem 3.** *Let $f_j(\delta)$ be a function for $\delta \in [0,1]$ and suppose that the derivative of $f_j(\delta)$ is given by the following form:*

$$
\begin{aligned}
\dot{f}_j(\delta) &= \frac{1}{\dot{f}_{j,1}(\delta)}\dot{f}_{j,2}(\delta) \quad (\dot{f}_{j,1}(\delta) > 0),\\
\dot{f}_{j,2}(\delta) &= -c_{j,2}\delta^2 + 2c_{j,1}\delta - c_{j,0} \quad (c_{j,0}, c_{j,1}, c_{j,2} > 0),
\end{aligned}
$$

*and we define $\tilde{\delta}_j$ as*

$$
\tilde{\delta}_j = \frac{1 - \sqrt{1 - c_{j,2}c_{j,0}/c_{j,1}^2}}{c_{j,2}/c_{j,1}}.
$$

*Then, $\hat{\delta}_j = \arg\min_{\delta\in[0,1]} f_j(\delta)$ is given by*

(1) *Case of* $1 - c_{j,2}c_{j,0}/c_{j,1}^2 \geq 0$:

$$\hat{\delta}_j = \begin{cases} \tilde{\delta}_j & \left(c_{j,2} > c_{j,1} \text{ or } (c_{j,2} \leq c_{j,1} \text{ and } \tilde{\delta}_j < 1)\right) \\ 1 & \left(c_{j,2} \leq c_{j,1} \text{ and } \tilde{\delta}_j \geq 1\right) \end{cases}.$$

(2) *Case of* $1 - c_{j,2}c_{j,0}/c_{j,1}^2 < 0$:

$$\hat{\delta}_j = 1.$$

**Example 2**

The partial derivative of the $\mathrm{AIC_C}$ is given by

$$\dot{f}_j(\delta_j) = \frac{\partial}{\partial \delta_j}\mathrm{AIC_C}(\boldsymbol{\delta}) = \frac{1}{(1 + q_{2,j}\delta_j^2)(n - p - 1 - q_{1,j} + p\delta_j)^2/n}\dot{f}_{j,2}(\delta_j),$$

$$\dot{f}_{j,2}(\delta) = 2p^2 q_{2,j}\delta^3 + pq_{2,j}\{4(n - p - 1 - q_{1,j}) - p(2n - p - 1)\}\delta^2 \\ + 2q_{2,j}(n - p - 1 - q_{1,j})^2\delta - p^2(2n - p - 1).$$

An update equation of the coordinate descent algorithm for solving the $\mathrm{AIC_C}$ minimization method is given by the following theorem (the proof is given in Appendix A.6).

**Theorem 4.** *Let* $f_j(\delta)$ *be a function for* $\delta \in [0, 1]$ *and suppose that the derivative of* $f_j(\delta)$ *is given by the following form:*

$$\dot{f}_j(\delta) = \frac{1}{\dot{f}_{j,1}(\delta)}\dot{f}_{j,2}(\delta) \quad (\dot{f}_{j,1}(\delta) > 0),$$

$$\dot{f}_{j,2}(\delta) = c_{j,3}\delta^3 + c_{j,2}\delta^2 + c_{j,1}\delta - c_{j,0} \quad (c_{j,0} > 0),$$

*and let* $m$ $(0 \leq m \leq 3)$ *be the number of stationary points of* $\dot{f}_{j,2}(\delta)$ *which is included in* $(0, 1)$ *and* $\tilde{\delta}_{j,1}, \dots, \tilde{\delta}_{j,m}$ $(m \geq 1)$ *be the stationary points satisfying* $\tilde{\delta}_{j,1} < \cdots < \tilde{\delta}_{j,m}$. *Moreover, we define a set* $\mathcal{S}_j$ *as*

$$\mathcal{S}_j = \{1\}\ (m = 0);\ \{\tilde{\delta}_{j,1}\}\ (m = 1);\ \{\tilde{\delta}_{j,1}, 1\}\ (m = 2);\ \{\tilde{\delta}_{j,1}, \tilde{\delta}_{j,3}\}\ (m = 3).$$

*Then,* $\hat{\delta}_j = \arg\min_{\delta \in [0,1]} f_j(\delta)$ *is given by*

$$\hat{\delta}_j = \arg\min_{\delta \in \mathcal{S}_j} f_j(\delta).$$

### 4.2.2. BNP-distance

Equation (4.7) leads to

$$\left(\boldsymbol{I}_p + \boldsymbol{B}^*_{\boldsymbol{\delta}}\right)^{-1} = \boldsymbol{W}^{1/2}(\boldsymbol{W}_j + \boldsymbol{z}_j\boldsymbol{z}_j'\delta_j^2)^{-1}\boldsymbol{W}^{1/2},$$

and the following holds:

$$(\boldsymbol{W}_j + \boldsymbol{z}_j\boldsymbol{z}_j'\delta_j^2)^{-1} = \boldsymbol{W}_j^{-1} - \frac{\boldsymbol{W}_j^{-1}\boldsymbol{z}_j\boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\delta_j^2}{1 + \boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{z}_j\delta_j^2} = \boldsymbol{W}_j^{-1} - \frac{\boldsymbol{W}_j^{-1}\boldsymbol{z}_j\boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\delta_j^2}{1 + q_{2,j}\delta_j^2}.$$

Therefore, we have

$$\begin{aligned} g_{\mathrm{BNP}}(\boldsymbol{B}^*_{\boldsymbol{\delta}}) &= p - \mathrm{tr}\left\{(\boldsymbol{I}_p + \boldsymbol{B}^*_{\boldsymbol{\delta}})^{-1}\right\} = p - \mathrm{tr}\left\{\left(\boldsymbol{W}_j^{-1} - \frac{\boldsymbol{W}_j^{-1}\boldsymbol{z}_j\boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\delta_j^2}{1 + q_{2,j}\delta_j^2}\right)\boldsymbol{W}\right\} \\ &= \frac{q_{4,j}\delta_j^2}{1 + q_{2,j}\delta_j^2} + q_{5,j}, \end{aligned}$$

where $q_{4,j}$ and $q_{5,j}$ are constants which do not depend on $\delta_j$ given by

$$q_{4,j} = \boldsymbol{z}_j'\boldsymbol{W}_j^{-1}\boldsymbol{W}\boldsymbol{W}_j^{-1}\boldsymbol{z}_j,\ q_{5,j} = p - \mathrm{tr}\left(\boldsymbol{W}_j^{-1}\boldsymbol{W}\right).$$

Hence, the following partial derivative is obtained:

$$\frac{\partial}{\partial\delta_j}g_{\mathrm{BNP}}(\boldsymbol{B}_{\boldsymbol{\delta}}) = \frac{2q_{4,j}\delta_j}{(1 + q_{2,j}\delta_j^2)^2}.$$

**Example**

The partial derivative of the BNPC is given by

$$\dot{f}_j(\delta_j) = \frac{\partial}{\partial\delta_j}\,\mathrm{BNPC}(\boldsymbol{\delta}) = \frac{1}{(1 + q_{2,j}\delta_j^2)^2}(-\alpha p q_{2,j}^2\delta_j^4 - 2\alpha p q_{2,j}\delta_j^2 + 2nq_{4,j}\delta_j - \alpha p).$$

An update equation of the coordinate descent algorithm for solving the BNPC minimization method is given by the following theorem obtained which is similar to Theorem 4.

**Theorem 5.** *Let $f_j(\delta)$ be a function for $\delta \in [0, 1]$ and suppose that the derivative of $f_j(\delta)$ is given by the following form:*

$$\begin{aligned} \dot{f}_j(\delta) &= \frac{1}{\dot{f}_{j,1}(\delta)}\dot{f}_{j,2}(\delta) \quad (\dot{f}_{j,1}(\delta) > 0), \\ \dot{f}_{j,2}(\delta) &= c_{j,4}\delta^4 + c_{j,3}\delta^3 + c_{j,2}\delta^2 + c_{j,1}\delta - c_{j,0} \quad (c_{j,0} > 0), \end{aligned}$$

*and let $m$ $(0 \le m \le 4)$ be the number of stationary points of $\dot{f}_{j,2}(\delta)$ which is included in $(0, 1)$*

*and* $\tilde{\delta}_{j,1}, \ldots, \tilde{\delta}_{j,m}$ $(m \geq 1)$ *be the stationary points satisfying* $\tilde{\delta}_{j,1} < \cdots < \tilde{\delta}_{j,m}$. *Moreover, we define a set* $\mathcal{S}_j$ *as*

$$\mathcal{S}_j = \begin{cases} \{1\} & (m = 0); \quad \{\tilde{\delta}_{j,1}\} \qquad\quad (m = 1); \quad \{\tilde{\delta}_{j,1}, 1\} \quad (m = 2) \\ \{\tilde{\delta}_{j,1}, \tilde{\delta}_{j,3}\} & (m = 3); \quad \{\tilde{\delta}_{j,1}, \tilde{\delta}_{j,3}, 1\} \quad (m = 4) \end{cases}.$$

*Then,* $\hat{\delta}_j = \arg\min_{\delta \in [0,1]} f_j(\delta)$ *is given by*

$$\hat{\delta}_j = \arg\min_{\delta \in \mathcal{S}_j} f_j(\delta).$$

### 4.2.3. ML-distance

Notice that

$$g_{\mathrm{ML}}(\boldsymbol{A}) = g_{\mathrm{LR}}(\boldsymbol{A}) - g_{\mathrm{BNP}}(\boldsymbol{A}).$$

Hence, we have

$$g_{\mathrm{ML}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) = \log(1 + q_{2,j}\delta_j^2) - \frac{q_{4,j}\delta_j^2}{1 + q_{2,j}\delta_j^2} + q_{3,j} - q_{5,j},$$

and the following partial derivative is obtained:

$$\frac{\partial}{\partial \delta_j} g_{\mathrm{ML}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) = \frac{2q_{2,j}\delta}{1 + q_{2,j}\delta^2} - \frac{2q_{4,j}\delta}{(1 + q_{2,j}\delta^2)^2}.$$

### 4.2.4. GLS-distance

We have

$$\boldsymbol{B}_{\boldsymbol{\delta}}\boldsymbol{W}^{-1} = \boldsymbol{z}_j\boldsymbol{z}_j'\boldsymbol{W}^{-1}\delta_j^2 + \boldsymbol{W}_j\boldsymbol{W}^{-1} - \boldsymbol{I}_p,$$

and therefore

$$g_{\mathrm{GLS}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) = \frac{1}{2}\operatorname{tr}\left\{(\boldsymbol{B}_{\boldsymbol{\delta}}\boldsymbol{W}^{-1})^2\right\} = \frac{1}{2}(q_{6,j}\delta_j^4 + 2q_{7,j}\delta_j^2 + q_{8,j}),$$

where $q_{\ell,j}$ $(\ell = 6, 7, 8)$ are constants which do not depend on $\delta_j$ given by

$$q_{6,j} = (\boldsymbol{z}_j'\boldsymbol{W}^{-1}\boldsymbol{z}_j)^2, \quad q_{7,j} = \boldsymbol{z}_j'\boldsymbol{W}^{-1}(\boldsymbol{W}_j - \boldsymbol{W})\boldsymbol{W}^{-1}\boldsymbol{z}_j, \quad q_{8,j} = \operatorname{tr}\left\{\left(\boldsymbol{W}_j\boldsymbol{W}^{-1} - \boldsymbol{I}_p\right)^2\right\}.$$

Hence, the following partial derivative is obtained:

$$\frac{\partial}{\partial \delta_j} g_{\mathrm{GLS}}(\boldsymbol{B}_{\boldsymbol{\delta}}^*) = 2q_{6,j}\delta_j^3 + 2q_{7,j}\delta_j.$$

## 5. Plug-in Iteration

In the previous section, we described the minimization of MSC extended to general distance. For MSC based on the LH-distance, the class of the optimal ridge parameters is obtained and since the minimizer is given as closed form and is unique, or minimizer candidates are given as closed forms and finite points, MSC can be minimized quickly. In contrast, since the optimal ridge parameters include the inverse of the estimator of the covariance matrix for the full model, those parameters are unstable when $k$ is large. On the other hand, for MSC based on general distance, in particular the LR-distance, the BNP-distance, and the ML-distance, since the estimator of the covariance matrix which is included in the optimal ridge parameters is an adjusted estimator, the optimal ridge parameters are stable even when $k$ is large. In contrast, such MSC cannot be minimized quickly. As above, MSC based on the LH-distance and MSC based on another distance have contrasting properties. We propose a new approach, called the Plug-in Iteration Method (PIM) which is a hybrid method drawing on the merits of the various MSCs. The PIM optimizes ridge parameters by repeating the following procedure: first, the ridge parameters are optimized by the MSC minimization method based on the LH-distance; next, the ridge parameters are optimized again by using the ridge parameters optimized in the previous step.

The ridge parameters optimized by the MSC minimization method based on the LH-distance include $\boldsymbol{S}$, and this derives from the fact that the original distance $\mathrm{tr}\{\hat{\Sigma}(\boldsymbol{\delta})\boldsymbol{S}^{-1}\}$ includes $\boldsymbol{S}$. Although the MSC was hitherto defined by using $\mathrm{tr}(\boldsymbol{B}^*_{\boldsymbol{\delta}})$ obtained from the original distance, we now redefine it using the original distance. For any $p \times p$ positive definite matrix $\boldsymbol{A}$, we define

$$r_+(\boldsymbol{A}) = \mathrm{tr}\left(\hat{\Sigma}_0\boldsymbol{A}^{-1}\right) + \frac{1}{n}\,\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{A}^{-1}\boldsymbol{Z}'\right),$$

and let $f^\dagger$ be a bivariate function defined by the following class.

**Definition 4.** **(Class of the bivariate function $f^\dagger$)** The $f^\dagger$ satisfies the following conditions:

(A1') For any $(r,u) \in (0, r_+(\boldsymbol{A})] \times [p, np)$, $f^\dagger(r,u)$ is continuous.

(A2') For any $(r,u) \in (0, r_+(\boldsymbol{A})] \times [p, np)$, $f^\dagger(r,u)$ is positive.

(A3') For any $(r,u) \in (0, r_+(\boldsymbol{A})] \times [p, np)$, $f^\dagger(r,u)$ is first order partially differentiable and its partial derivatives are positive.

Using the bivariate function $f^\dagger$, we redefine the MSC based on the LH-distance as

$$\mathrm{MSC}^\dagger(\boldsymbol{\delta} \mid \boldsymbol{A}) = f^\dagger\left(\mathrm{tr}\left\{\hat{\Sigma}(\boldsymbol{\delta})\boldsymbol{A}^{-1}\right\}, \mathrm{df}(\boldsymbol{\delta})\right). \tag{5.1}$$

This MSC covers a wider class than the MSC in (2.8) and is equal to the MSC in (2.8) when $\boldsymbol{A} = \boldsymbol{S}$. For the $GC_p$ and EGCV criteria, $f^\dagger$ is given by

$$f^\dagger = \begin{cases} nr + \alpha u & (GC_p \text{ criterion}) \\ r/(1 - u/np)^\alpha & (\text{EGCV criterion}) \end{cases}.$$

Similar to Theorem 1, the optimal $\boldsymbol{\delta}$ minimizing the MSC in (5.1) is given by the following corollary.

**Corollary 1.** *We define a function* $\phi(h \mid \boldsymbol{A})$ $(h \in \mathbb{R}_+\backslash\{0\})$ *as*

$$\phi(h \mid \boldsymbol{A}) = \mathrm{MSC}^\dagger(\hat{\boldsymbol{\delta}}(h \mid \boldsymbol{A}) \mid \boldsymbol{A}),$$

*and suppose that* $\exists v > 0$ *s.t.* $\phi(v \mid \boldsymbol{A}) < \lim_{h\to 0} \phi(h \mid \boldsymbol{A})$*, where* $\hat{\boldsymbol{\delta}}(h \mid \boldsymbol{A}) = (\hat{\delta}_1(h \mid \boldsymbol{A}), \dots, \hat{\delta}_k(h \mid \boldsymbol{A}))'$ *is a class of ridge parameters given by*

$$\hat{\delta}_j(h \mid \boldsymbol{A}) = 1 - \mathrm{soft}\left(1, h/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right).$$

*Then, we have the following:*

(1) *The optimal ridge parameters based on minimizing* $\mathrm{MSC}^\dagger(\boldsymbol{\delta} \mid \boldsymbol{A})$ *are given by* $\hat{\boldsymbol{\delta}}(\hat{h}_{\boldsymbol{A}} \mid \boldsymbol{A})$ *and* $\hat{h}_{\boldsymbol{A}}$ *is given by*

$$\hat{h}_{\boldsymbol{A}} = \arg \min_{h \in \mathbb{R}_+\backslash\{0\}} \phi(h \mid \boldsymbol{A}).$$

(2) *The* $\phi(h \mid \boldsymbol{A})$ *has the following properties:*

(P1) *For all* $h \in \mathbb{R}_+\backslash\{0\}$, $\phi(h \mid \boldsymbol{A})$ *is continuous.*

(P2) *For all* $h \geq t_k$, $\phi(h \mid \boldsymbol{A}) = f^\dagger(r_+(\boldsymbol{A}), p)$.

(P3) *The* $\phi(h \mid \boldsymbol{A})$ *can be expressed as the following piecewise function:*

$$\begin{aligned} \phi(h \mid \boldsymbol{A}) &= \phi_a(h \mid \boldsymbol{A}) \quad (h \in R_a;\ a = 0, 1, \dots, k) \\ &= f^\dagger\left(\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_0\boldsymbol{A}^{-1}) + (c_{1,a} + c_{2,a}h^2)/n, p(1 + k - a - c_{2,a}h)\right), \end{aligned}$$

*where* $R_a$, $c_{1,a}$ *and* $c_{2,a}$ *are range and nonnegative constants similar to* (3.1) *and Proposition 1, respectively. However,* $t_j$ $(j = 1, \dots, k)$ *is the jth order statistic of* $\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j$ $(j = 1, \dots, k)$.

Corollary 1 is an extension of Theorem 1 and Proposition 1 and they are equivalent when $\boldsymbol{A} = \boldsymbol{S}$. Furthermore, $\hat{h}_{\boldsymbol{A}}$ can be obtained by applying Theorem 2.

Using Corollary 1, we describe the PIM algorithm. Let $\boldsymbol{S}^{(0)} = \boldsymbol{S}$ and we define $\hat{\boldsymbol{\delta}}^{(0)}(h) =$

$(\hat{\delta}_1^{(0)}(h), \ldots, \hat{\delta}_k^{(0)}(h))'$ as $\hat{\boldsymbol{\delta}}^{(0)}(h) = \hat{\boldsymbol{\delta}}(h \mid \boldsymbol{S}^{(0)})$ and define the optimal ridge parameters based on minimizing $\mathrm{MSC}^{\dagger}(\boldsymbol{\delta} \mid \boldsymbol{S}^{(0)})$ as

$$\begin{aligned} \hat{\boldsymbol{\delta}}^{(0)} &= \left(\hat{\delta}_1^{(0)}, \ldots, \hat{\delta}_k^{(0)}\right)', \quad \hat{\delta}_j^{(0)} = \hat{\delta}_j^{(0)}(\hat{h}^{(0)}), \\ \hat{h}^{(0)} &= \arg \min_{h \in \mathbb{R}_+ \backslash \{0\}} \phi^{(0)}(h), \quad \phi^{(0)}(h) = \mathrm{MSC}^{\dagger}(\hat{\boldsymbol{\delta}}^{(0)}(h) \mid \boldsymbol{S}^{(0)}). \end{aligned}$$

Therefore $\hat{\delta}_j^{(0)}$ is given by

$$\hat{\delta}_j^{(0)} = 1 - \mathrm{soft}\left(1, \hat{h}^{(0)}/\boldsymbol{z}_j'\{\boldsymbol{S}^{(0)}\}^{-1}\boldsymbol{z}_j\right). \tag{5.2}$$

Furthermore, by substituting $\hat{\boldsymbol{\delta}}^{(0)}$, we define $\boldsymbol{S}^{(1)}$ as

$$\boldsymbol{S}^{(1)} = \boldsymbol{W}^{1/2}\dot{\boldsymbol{G}}(\boldsymbol{B}^*_{\hat{\boldsymbol{\delta}}^{(0)}})^{-1}\boldsymbol{W}^{1/2},$$

and let $\hat{\boldsymbol{\delta}}^{(1)}(h)$ be a class of ridge parameters wherein the $j$th element $(j = 1, \ldots, k)$ is given by

$$\hat{\delta}_j^{(1)}(h) = 1 - \mathrm{soft}\left(1, h/\boldsymbol{z}_j'\{\boldsymbol{S}^{(1)}\}^{-1}\boldsymbol{z}_j\right).$$

Then, we optimize the ridge parameters again as

$$\begin{aligned} \hat{\boldsymbol{\delta}}^{(1)} &= \left(\hat{\delta}_1^{(1)}, \ldots, \hat{\delta}_k^{(1)}\right)', \quad \hat{\delta}_j^{(1)} = \hat{\delta}_j^{(1)}(\hat{h}^{(1)}), \\ \hat{h}^{(1)} &= \arg \min_{h \in \mathbb{R}_+ \backslash \{0\}} \phi^{(1)}(h), \quad \phi^{(1)}(h) = \mathrm{MSC}^{\dagger}(\hat{\boldsymbol{\delta}}^{(1)}(h) \mid \boldsymbol{S}^{(1)}). \end{aligned}$$

The $\hat{h}^{(1)}$ can be obtained quickly by applying Theorem2. Since the optimal ridge parameter $\hat{\boldsymbol{\delta}}^{(0)}$ includes $\boldsymbol{S}$, it is unstable when $k$ is large. Whereas, since $\boldsymbol{S}^{(1)}$ is adjusted by substituting $\hat{\boldsymbol{\delta}}^{(0)}$, $\hat{\boldsymbol{\delta}}^{(1)}$ is stable even when $k$ is large. The PIM algorithm is summarized as follows.

**PIM Algorithm**

Step 1. Let the initial vector $\hat{\boldsymbol{\delta}}^{(0)}$ be the ridge parameters optimized by the MSC minimization method based on the LH-distance and $i \leftarrow 0$.

Step 2. Define $\boldsymbol{S}^{(i+1)}$ and $\phi^{(i+1)}(h)$ as

$$\boldsymbol{S}^{(i+1)} = \boldsymbol{W}^{1/2}\dot{\boldsymbol{G}}(\boldsymbol{B}^*_{\hat{\boldsymbol{\delta}}^{(i)}})^{-1}\boldsymbol{W}^{1/2}, \quad \phi^{(i+1)}(h) = \mathrm{MSC}^{\dagger}\left(\hat{\boldsymbol{\delta}}^{(i+1)}(h) \mid \boldsymbol{S}^{(i+1)}\right),$$

where the class of ridge parameters is given by

$$\hat{\boldsymbol{\delta}}^{(i+1)}(h) = \left(\hat{\delta}_1^{(i+1)}(h), \ldots, \hat{\delta}_k^{(i+1)}(h)\right)', \quad \hat{\delta}_j^{(i+1)}(h) = 1 - \mathrm{soft}\left(1, h/\boldsymbol{z}_j'\{\boldsymbol{S}^{(i+1)}\}^{-1}\boldsymbol{z}_j\right).$$

Step 3. By using Theorem 2, update the ridge parameters as

$$\hat{\boldsymbol{\delta}}^{(i+1)} = \hat{\boldsymbol{\delta}}^{(i+1)}(\hat{h}^{(i+1)}), \quad \hat{h}^{(i+1)} = \arg \min_{h \in \mathbb{R}_+ \backslash \{0\}} \phi^{(i+1)}(h).$$

Step 4. If $\hat{\boldsymbol{\delta}}^{(i+1)}$ converges, the algorithm is complete. If not, let $i \leftarrow i + 1$ and return to Step 2.

Since the MSC minimized at each iteration is based on the LH-distance, the minimization is fast. Furthermore, an estimator of the covariance matrix which is included in $\hat{\boldsymbol{\delta}}^{(i)}$ is stable by substituting the ridge parameters optimized in the previous step. Thus, the PIM is a hybrid method which leverages the merits of the various MSCs.

The PIM algorithm is similar to the iterative method. In particular, when using the $GC_p$ criterion, for all $i$ $(= 0, 1, \ldots, )$, we have

$$\hat{h}^{(i)} = \frac{\alpha p}{2}.$$

Therefore, the PIM is the following iterative method:

$$\hat{\delta}_j^{(i+1)} = 1 - \mathrm{soft}\left(1, \alpha p / 2 \boldsymbol{z}_j^{*\prime} \dot{\boldsymbol{G}}(\boldsymbol{B}^*_{\hat{\boldsymbol{\delta}}^{(i)}}) \boldsymbol{z}_j^*\right),$$

and this is equal to the iterative method wherein the initial vector is the ridge parameters optimized by the $GC_p$ criterion minimization method, the iterative function is equation (4.2), and $\tau(\boldsymbol{\delta}) = \alpha p/2$. That is, when using the $GC_p$ criterion, the PIM with the GIC is equal to the GIC minimization method and the PIM with the BNPC is equal to the BNPC minimization method.

## 6. Relationship with Multivariate Adaptive-Lasso Regression

In this section, we describe a relationship between the MGR and MAL estimators after the regularization parameters are optimized by the MSC minimization method based on the LH-distance. The MAL estimator cannot usually be obtained as closed form. However, it can be obtained as closed form under orthogonal explanatory variables. Although we use general $\boldsymbol{X}$ until the previous section, this section deals with orthogonal explanatory variables. Furthermore, instead of using the transformed ridge parameters $\delta_1, \ldots, \delta_k$, we approach this via the original ridge parameters $\theta_1, \ldots, \theta_k$.

### 6.1. Estimators with Optimal Regularization Parameters under Orthogonality

The orthogonality of $\boldsymbol{X}$ means $\boldsymbol{Q} = \boldsymbol{I}_k$ in (2.1). Therefore, the LS and the MGR estimators of $\Xi$ in (1.2) and (2.2), respectively, are rewritten as

$$\hat{\Xi} = \left(\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_k\right)' = \boldsymbol{D}^{-1/2}\boldsymbol{Z}, \quad \hat{\boldsymbol{\xi}}_j = \frac{1}{\sqrt{d_j}}\boldsymbol{z}_j,$$

$$\hat{\Xi}^{\mathrm{R}}_{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\theta_1,1}, \ldots, \hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\theta_k,k}\right)' = \boldsymbol{D}^{1/2}(\boldsymbol{D} + \Theta)^{-1}\boldsymbol{Z}, \quad \hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\theta_j,j} = \frac{\sqrt{d_j}}{d_j + \theta_j}\boldsymbol{z}_j, \tag{6.1}$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_k)$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)'$ are the $k \times k$ diagonal matrix and the $k \times p$ matrix given by (2.1) and (2.4), respectively. The ridge parameters are optimized using the following MSC based on the LH-distance:

$$\mathrm{MSC}_{\mathrm{R}}(\boldsymbol{\theta} \mid \boldsymbol{A}) = f^{\dagger}\left(\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}_{\mathrm{R}}(\boldsymbol{\theta})\boldsymbol{A}^{-1}\}, \mathrm{df}_{\mathrm{R}}(\boldsymbol{\theta})\right), \tag{6.2}$$

where $\hat{\boldsymbol{\Sigma}}_{\mathrm{R}}(\boldsymbol{\theta})$ and $\mathrm{df}_{\mathrm{R}}(\boldsymbol{\theta})$ are given by transforming the parameter from $\boldsymbol{\delta}$ to $\boldsymbol{\theta}$ in $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\delta})$ and $\mathrm{df}(\boldsymbol{\delta})$, which are given by (2.6) and (2.7), respectively, as

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{R}}(\boldsymbol{\theta}) = \hat{\boldsymbol{\Sigma}}_0 + \frac{1}{n}\sum_{j=1}^{k} \boldsymbol{z}_j\boldsymbol{z}_j'\left(\frac{\theta_j}{d_j+\theta_j}\right)^2, \quad \mathrm{df}_{\mathrm{R}}(\boldsymbol{\theta}) = p(1+k) - p\sum_{j=1}^{k}\frac{\theta_j}{d_j+\theta_j}.$$

Thus, the MSC in (6.2) is the parameter-transformed version of the MSC in (5.1). Furthermore, since the transformation is a one-to-one correspondence, Corollary 1 gives the following class of ridge parameters optimized by minimizing $\mathrm{MSC}_{\mathrm{R}}(\boldsymbol{\theta} \mid \boldsymbol{A})$:

$$\hat{\boldsymbol{\theta}}(h \mid \boldsymbol{A}) = \left(\hat{\theta}_1(h \mid \boldsymbol{A}), \ldots, \hat{\theta}_k(h \mid \boldsymbol{A})\right)', \quad \hat{\theta}_j(h \mid \boldsymbol{A}) = \begin{cases} \dfrac{d_j h}{\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j - h} & (h < \boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j) \\ \infty & (h \geq \boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j) \end{cases}.$$

Notice that for all $x \in \mathbb{R}_+$,

$$\mathrm{MSC}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(x \mid \boldsymbol{A}) \mid \boldsymbol{A}) = \mathrm{MSC}^{\dagger}(\hat{\boldsymbol{\delta}}(x \mid \boldsymbol{A}) \mid \boldsymbol{A}).$$

Then, from Corollary 1, the optimal ridge parameters based on minimizing the MSC in (6.2) are given by

$$\begin{aligned} &\hat{\theta}_j = \hat{\theta}_j(\hat{h}_{\boldsymbol{A}} \mid \boldsymbol{A}) \quad (j = 1, \ldots, k), \\ &\hat{h}_{\boldsymbol{A}} = \arg\min_{h \in \mathbb{R}_+ \backslash \{0\}} \phi(h \mid \boldsymbol{A}), \quad \phi(h \mid \boldsymbol{A}) = \mathrm{MSC}^{\dagger}(\hat{\boldsymbol{\delta}}(h \mid \boldsymbol{A}) \mid \boldsymbol{A}), \end{aligned} \tag{6.3}$$

and using these optimal ridge parameters, the optimal MGR estimator based on minimizing the MSC in (6.2) is given by

$$\hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\hat{\theta}_j, j} = \frac{1}{\sqrt{d_j}}\,\mathrm{soft}\left(1, \hat{h}_{\boldsymbol{A}}/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right)\boldsymbol{z}_j. \tag{6.4}$$

Since $\hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\hat{\theta}_j, j} = \boldsymbol{0}_p$ when $\hat{h}_{\boldsymbol{A}} \geq \boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j$, we found that the non-sparse MGR estimator is sparse after the ridge parameters are optimized.

Next, we describe the MAL estimator of $\boldsymbol{\Xi}$. Ohishi *et al.* (2020b) derived the AL estimator as closed form under orthogonality of $\boldsymbol{X}$. As a natural extension of this result, the MAL estimator can be obtained as closed form. Let $\boldsymbol{L}_\lambda$ be a $k \times k$ diagonal matrix of which the $j$th diagonal element is given by

$$\ell_{\lambda,j} = \frac{1}{d_j}\,\mathrm{soft}\left(1, \lambda w_j/\sqrt{d_j}\|\boldsymbol{z}_j\|\right) \quad (j = 1, \ldots, k),$$

where $\lambda \in \mathbb{R}_+$ is a regularization parameter called a tuning parameter and $w_j$ is a weight. Then, the MAL estimator of $\Xi$ is given by

$$\begin{aligned} \hat{\Xi}^{\mathrm{L}}_{\lambda} &= \left(\hat{\boldsymbol{\xi}}^{\mathrm{L}}_{\lambda,1}, \ldots, \hat{\boldsymbol{\xi}}^{\mathrm{L}}_{\lambda,k}\right)' = \boldsymbol{L}_{\lambda}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{L}_{\lambda}\boldsymbol{D}^{1/2}\boldsymbol{Z}, \\ \hat{\boldsymbol{\xi}}^{\mathrm{L}}_{\lambda,j} &= \sqrt{d_j}\ell_{\lambda,j}\boldsymbol{z}_j = \frac{1}{\sqrt{d_j}}\,\mathrm{soft}\left(1, \lambda w_j/\sqrt{d_j}\|\boldsymbol{z}_j\|\right)\boldsymbol{z}_j. \end{aligned} \tag{6.5}$$

Since $\boldsymbol{L}_{\lambda} = \boldsymbol{D}^{-1}$ when $\lambda = 0$, the MAL estimator coincides with the LS estimator when $\lambda = 0$, and the MAL estimator coincides with the AL estimator given in Ohishi *et al.* (2020b) when $p = 1$. The MAL estimator is sparse in the sense that $\hat{\boldsymbol{\xi}}^{\mathrm{L}}_{\lambda,j} = \boldsymbol{0}_p$ when $\lambda w_j \geq \sqrt{d_j}\|\boldsymbol{z}_j\|$. The $\hat{\Xi}^{\mathrm{L}}_{\lambda}$ in (6.5) denotes the minimizer of the following PRSS:

$$\mathrm{tr}\,\{(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\Xi)'(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\Xi)\} + 2\lambda\sum_{j=1}^{k} w_j\|\boldsymbol{\xi}_j\|. \tag{6.6}$$

The MGR estimator in (6.1) depends on $k$ regularization parameters. Whereas, the MAL estimator in (6.5) depends on only one regularization parameter. Furthermore, although the MGR estimator is not sparse, the MAL estimator is characterized by sparsity. Hence, it can be stated that the MGR and MAL estimators have different properties.

The MAL estimator in (6.5) gives a predictive matrix of $\boldsymbol{Y}$ for the MAL regression as follows:

$$\hat{\boldsymbol{Y}}^{\mathrm{L}}_{\lambda} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}' + \boldsymbol{X}\hat{\Xi}^{\mathrm{L}}_{\lambda} = \boldsymbol{H}^{\mathrm{L}}_{\lambda}\boldsymbol{Y}, \quad \boldsymbol{H}^{\mathrm{L}}_{\lambda} = \boldsymbol{J}_n + \boldsymbol{X}\boldsymbol{L}_{\lambda}\boldsymbol{X}'.$$

Using $\hat{\boldsymbol{Y}}^{\mathrm{L}}_{\lambda}$ and $\boldsymbol{H}^{\mathrm{L}}_{\lambda}$, we define an estimator of $\boldsymbol{\Sigma}$ and a GDF as

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\mathrm{L}}(\lambda) &= \frac{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{\mathrm{L}}_{\lambda})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}^{\mathrm{L}}_{\lambda})}{n} = \frac{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{X}\boldsymbol{L}_{\lambda}\boldsymbol{X}')^2\boldsymbol{Y}}{n}, \\ \mathrm{df}_{\mathrm{L}}(\lambda) &= p\,\mathrm{tr}(\boldsymbol{H}^{\mathrm{L}}_{\lambda}). \end{aligned}$$

Similar to Ohishi *et al.* (2020b), we have the following lemma concerning $\hat{\boldsymbol{\Sigma}}_{\mathrm{L}}(\lambda)$ and $\mathrm{df}_{\mathrm{L}}(\lambda)$.

**Lemma 5.** *The* $\hat{\boldsymbol{\Sigma}}_{\mathrm{L}}(\lambda)$ *and* $\mathrm{df}_{\mathrm{L}}(\lambda)$ *are expressed as*

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{L}}(\lambda) = \hat{\boldsymbol{\Sigma}}_0 + \frac{1}{n}\boldsymbol{Z}'(\boldsymbol{I}_k - \boldsymbol{D}\boldsymbol{L}_{\lambda})^2\boldsymbol{Z} = \hat{\boldsymbol{\Sigma}}_0 + \frac{1}{n}\sum_{j=1}^{k}\left\{1 - \mathrm{soft}\left(1, \lambda w_j/\sqrt{d_j}\|\boldsymbol{z}_j\|\right)\right\}^2\boldsymbol{z}_j\boldsymbol{z}_j',$$

$$\mathrm{df}_{\mathrm{L}}(\lambda) = p + p\sum_{j=1}^{k}\mathrm{soft}\left(1, \lambda w_j/\sqrt{d_j}\|\boldsymbol{z}_j\|\right).$$

Then, the MSC for optimizing the tuning parameter in the MAL regression is given by

$$\mathrm{MSC_L}(\lambda \mid \boldsymbol{A}) = f^{\dagger}\left(\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathrm{L}}(\lambda)\boldsymbol{A}^{-1}), \mathrm{df_L}(\lambda)\right), \tag{6.7}$$

and the tuning parameter optimized by the MSC minimization method is given by

$$\hat{\lambda}_{\boldsymbol{A}} = \arg\min_{\lambda \in \mathbb{R}_+} \mathrm{MSC_L}(\lambda \mid \boldsymbol{A}).$$

Regarding the weight $w_j$, in general, an inverse of a norm of an estimator of $\boldsymbol{\xi}_j$ is used. When using the weight $w_j = 1/||\hat{\boldsymbol{\xi}}_j||$ based on the LS estimator, the optimal MAL estimator based on minimizing the MSC in (6.7) is given by

$$\hat{\boldsymbol{\xi}}^{\mathrm{L}}_{\hat{\lambda}_{\boldsymbol{A}},j} = \frac{1}{\sqrt{d_j}} \mathrm{soft}\left(1, \hat{\lambda}_{\boldsymbol{A}}/||\boldsymbol{z}_j||^2\right) \boldsymbol{z}_j. \tag{6.8}$$

## 6.2. Equivalence between MGR and MAL estimators

This subsection investigates a relationship between the MGR and MAL estimators under the regularization parameters optimized by the MSC minimization method. Although the optimal MGR estimator in (6.4) and the optimal MAL estimator in (6.8) have similar forms, the optimal MGR estimator does not include $||\boldsymbol{z}_j||^2$, but rather $\boldsymbol{z}_j\boldsymbol{A}^{-1}\boldsymbol{z}_j$ normalized by $\boldsymbol{A}$. First, we focus on the difference.

Let $\boldsymbol{T}$ be an $n \times p$ matrix defined by $\boldsymbol{T} = \boldsymbol{Y}\boldsymbol{A}^{-1/2}$, $\boldsymbol{U}$ and $\boldsymbol{\Gamma}$ be $k \times p$ matrices defined by $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)' = \boldsymbol{Z}\boldsymbol{A}^{-1/2} = \boldsymbol{P}_1'\boldsymbol{T}$ and $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k)' = \boldsymbol{\Xi}\boldsymbol{A}^{-1/2}$, respectively, and $\boldsymbol{v}$ be a $p$-dimensional vector defined by $\boldsymbol{v} = \boldsymbol{A}^{-1/2}\boldsymbol{\mu}$. Then, we normalize the PRSS for the MGR regression as

$$\begin{aligned}&\mathrm{tr}\left\{(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\boldsymbol{\Xi})'(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\boldsymbol{\Xi})\boldsymbol{A}^{-1} + \boldsymbol{\Xi}'\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'\boldsymbol{\Xi}\boldsymbol{A}^{-1}\right\} \\ &\quad = \mathrm{tr}\left\{(\boldsymbol{T} - \boldsymbol{1}_n\boldsymbol{v}' - \boldsymbol{X}\boldsymbol{\Gamma})'(\boldsymbol{T} - \boldsymbol{1}_n\boldsymbol{v}' - \boldsymbol{X}\boldsymbol{\Gamma}) + \boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'\boldsymbol{\Gamma}\right\}.\end{aligned}$$

This normalized PRSS provides the MGR estimator of $\boldsymbol{\gamma}_j$ as

$$\hat{\boldsymbol{\gamma}}^{\mathrm{R}}_{\theta_j,j} = \frac{\sqrt{d_j}}{d_j + \theta_j}\boldsymbol{u}_j.$$

Therefore, the MGR normalized estimator of $\boldsymbol{\xi}_j$ is given by

$$\hat{\boldsymbol{\xi}}^{\mathrm{R}\dagger}_{\theta_j,j} = \boldsymbol{A}^{1/2}\hat{\boldsymbol{\gamma}}^{\mathrm{R}}_{\theta_j,j} = \frac{\sqrt{d_j}}{d_j + \theta_j}\boldsymbol{z}_j,$$

and this is equal to the MGR estimator in (6.1). That is, the MGR estimator in (6.1) is a normalized estimator in spite of the fact that it is obtained from non-normalized PRSS in (2.3).

Thus, the optimal MGR normalized estimator is given by (6.4). On the other hand, based on Xin *et al.* (2017), we normalize the PRSS for the MAL regression as

$$\begin{aligned}&\mathrm{tr}\left\{(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\boldsymbol{\Xi})'(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}' - \boldsymbol{X}\boldsymbol{\Xi})\boldsymbol{A}^{-1}\right\} + 2\lambda\sum_{j=1}^{k} w_j\|\boldsymbol{A}^{-1/2}\boldsymbol{\xi}_j\|\\ &= \mathrm{tr}\left\{(\boldsymbol{T} - \mathbf{1}_n\boldsymbol{v}' - \boldsymbol{X}\boldsymbol{\Gamma})'(\boldsymbol{T} - \mathbf{1}_n\boldsymbol{v}' - \boldsymbol{X}\boldsymbol{\Gamma})\right\} + 2\lambda\sum_{j=1}^{k} w_j\|\boldsymbol{\gamma}_j\|.\end{aligned}$$

When using the general weight $w_j = 1/\|\hat{\boldsymbol{\gamma}}_j\|$ ($\hat{\boldsymbol{\gamma}}_j$ is the LS estimator of $\boldsymbol{\gamma}_j$), this normalized PRSS provides the MAL estimator of $\boldsymbol{\gamma}_j$ as

$$\hat{\boldsymbol{\gamma}}^{\mathrm{L}}_{\lambda,j} = \frac{1}{\sqrt{d_j}}\,\mathrm{soft}\left(1, \lambda/\|\boldsymbol{u}_j\|^2\right)\boldsymbol{u}_j.$$

Therefore, the MAL normalized estimator of $\boldsymbol{\xi}_j$ is given by

$$\hat{\boldsymbol{\xi}}^{\mathrm{L}\dagger}_{\lambda,j} = \boldsymbol{A}^{1/2}\hat{\boldsymbol{\gamma}}^{\mathrm{L}}_{\lambda,j} = \frac{1}{\sqrt{d_j}}\,\mathrm{soft}\left(1, \lambda/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right)\boldsymbol{z}_j,$$

and this is different from the MAL estimator in (6.5) obtained as the minimizer of the PRSS in (6.6) with the weight $w_j = 1/\|\hat{\boldsymbol{\xi}}_j\|$. Hence, the difference between the two optimal estimators (6.4) and (6.8) is whether the estimator is normalized or not. If $\hat{h}_{\boldsymbol{A}} = \hat{\lambda}_{\boldsymbol{A}}$, the two optimal normalized estimators are equivalent. The equivalence is given by the following theorem (the proof is given in Appendix A.7).

**Theorem 6.** *Suppose that $w_j = 1/\|\hat{\boldsymbol{\gamma}}_j\|$ and let $\hat{\theta}_j$ ($j = 1, \ldots, k$) and $\hat{\lambda}$ be the regularization parameters optimized by the MSC minimization method based on the LH-distance defined by*

$$\begin{aligned}&\hat{\theta}_j = \hat{\theta}_j(\hat{h}_{\boldsymbol{A}} \mid \boldsymbol{A}), \quad \hat{h}_{\boldsymbol{A}} = \arg\min_{h\in\mathbb{R}_+}\mathrm{MSC}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(h \mid \boldsymbol{A}) \mid \boldsymbol{A}),\\ &\hat{\lambda} = \hat{\lambda}_{\boldsymbol{A}} = \arg\min_{\lambda\in\mathbb{R}_+}\mathrm{MSC}_{\mathrm{L}}(\lambda \mid \boldsymbol{A}).\end{aligned}$$

*Then, the following equation holds:*

$$\hat{\boldsymbol{\xi}}^{\mathrm{R}\dagger}_{\hat{\theta}_j,j} = \hat{\boldsymbol{\xi}}^{\mathrm{L}\dagger}_{\hat{\lambda},j} \quad (j = 1, \ldots, k).$$

In Theorem 6, the normalized estimators derived the equivalence. Next, we focus on the MSC to investigate the equivalence. The optimal MAL estimator in (6.8) includes $\|\boldsymbol{z}_j\|^2$ and this originates from the non-normalized PRSS in (6.6). In contrast, $\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j$ which is included in the optimal MGR estimator in (6.4) originates from the distance $\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}_{\mathrm{MGR}}(\boldsymbol{\theta})\boldsymbol{A}^{-1}\}$ normalized by $\boldsymbol{A}$. This leads to the following equivalence (the proof is given in Appendix A.8).

**Theorem 7.** *Suppose that $w_j = 1/\|\hat{\boldsymbol{\xi}}_j\|$ and let $\hat{\theta}_j$ $(j = 1, \ldots, k)$ and $\hat{\lambda}$ be the regularization parameters optimized by the MSC minimization method based on the LH-distance defined by*

$$\hat{\theta}_j = \hat{\theta}_j(\hat{h}_{\boldsymbol{I}_p} \mid \boldsymbol{I}_p), \quad \hat{h}_{\boldsymbol{I}_p} = \arg\min_{h \in \mathbb{R}_+} \mathrm{MSC}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(h \mid \boldsymbol{I}_p) \mid \boldsymbol{I}_p),$$
$$\hat{\lambda} = \hat{\lambda}_{\boldsymbol{I}_p} = \arg\min_{\lambda \in \mathbb{R}_+} \mathrm{MSC}_{\mathrm{L}}(\lambda \mid \boldsymbol{I}_p).$$

*Then, the following equation holds:*

$$\hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\hat{\theta}_j, j} = \hat{\boldsymbol{\xi}}^{\mathrm{L}}_{\hat{\lambda}, j} \quad (j = 1, \ldots, k).$$

## 7. Numerical Studies

In this section, we explore the performance of the MSC minimization methods for optimizing ridge parameters by evaluating prediction accuracies of predictive matrices via simulation. This simulation is executed using R (ver. 3.6.0) on a computer with a Windows 10 Pro operating system, Intel (R) Core i7-7700 processor, and 16 GB of RAM. Let $\boldsymbol{R}_k = \mathrm{diag}(1, \ldots, k)$ and let $\boldsymbol{\Omega}_k(\rho)$ be a $k \times k$ matrix of which the $(i, j)$ element is given by $\rho^{|i-j|}$. Then, the simulation data are generated from the following model:

$$\boldsymbol{Y} \sim N_{n \times p}(\boldsymbol{X}\boldsymbol{\Xi}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n), \quad \boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}_0\boldsymbol{\Psi}(0.99)^{1/2}, \quad \boldsymbol{\Sigma} = \boldsymbol{R}_p^{1/2}\boldsymbol{\Omega}_p(\rho_y)\boldsymbol{R}_p^{1/2},$$

where $\boldsymbol{\Xi}$ and $\boldsymbol{X}_0$ are $k \times p$ and $n \times k$ matrices wherein all the elements are identically and independently distributed according to $U(-1, 1)$ and $\boldsymbol{\Psi}(\rho)$ is a correlation matrix of $\boldsymbol{X}$ defined by $\boldsymbol{\Psi}(\rho) = \boldsymbol{R}_k^{1/2}\boldsymbol{\Omega}_k(\rho)\boldsymbol{R}_k^{1/2}$. Furthermore, $\rho = 0.99$ and thus this simulation is a highly correlated setting. Finally, $\boldsymbol{\Xi}$ and $\boldsymbol{X}_0$ are fixed throughout the simulation iterations.

Let $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}$ be the predictive matrix of $\boldsymbol{Y}$ obtained from the optimal MGR estimator based on minimizing the MSC and $\hat{\boldsymbol{Y}}$ be the predictive matrix of $\boldsymbol{Y}$ obtained from the LS estimator, i.e., $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{\boldsymbol{0}_k}$. Then, we evaluate the prediction accuracy of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}$ by the following relative mean square error (RMSE):

$$\mathrm{RMSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}] = \frac{\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}]}{p(k+1)} \times 100(\%), \quad \mathrm{MSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}] = \mathrm{E}\left[\mathrm{tr}\left\{(\boldsymbol{X}\boldsymbol{\Xi} - \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}})'(\boldsymbol{X}\boldsymbol{\Xi} - \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}})\boldsymbol{\Sigma}^{-1}\right\}\right].$$

In this setting, $\mathrm{MSE}[\hat{\boldsymbol{Y}}] = p(k + 1)$. This means that the prediction accuracies are evaluated in terms of the amount of improvement of the prediction accuracy of $\hat{\boldsymbol{Y}}$. Specifically, RMSE $< 100$ means the prediction accuracy of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}$ is superior to that of $\hat{\boldsymbol{Y}}$ and RMSE $> 100$ means the prediction accuracy of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}$ is inferior to that of $\hat{\boldsymbol{Y}}$. The smaller the RMSE value, the better the prediction accuracy. The expectation of the MSE is evaluated by Monte Carlo simulation with 10,000 iterations. Furthermore, it can be considered that the MSE value strongly relates to the amount of shrinkage of the MGR estimator, in particular, more shrinkage is required when there are highly correlated variables in $\boldsymbol{X}$. When $\hat{\delta}_j = 1$, the amount of shrinkage of the MGR estimator is maximized and this means that the $j$th eigenvalue (and corresponding eigenvector) is removed from the model. From this, we measure the amount of shrinkage of the MGR estimator by calculating the following relative number of removed eigenvalues (RNRE):

$$\mathrm{RNRE}(\hat{\boldsymbol{\delta}}) = \frac{\#(\{j \in \{1, \ldots, k\} \mid \hat{\delta}_j = 1\})}{k} \times 100\ (\%).$$

The RNRE expresses the ratio of the number of removed eigenvalues. If the RNRE value is small (large), then the amount of shrinkage is also small (large).

In this simulation, we estimate the mean structure of model. Thus, we use the LH-, LR-, and BNP-distances as the distance in the MSC. RMSE comparison 1 explores the prediction accuracies of predictive matrices where ridge parameters are optimized by the following methods:

- $GC_p$: $GC_p$ criterion minimization method.
- EGCV: EGCV criterion minimization method.
- GIC: GIC minimization method via the iterative method with the initial vector $\mathbf{0}_k$.
- BNPC: BNPC minimization method via the iterative method with the initial vector $\mathbf{0}_k$.
- PIM1: PIM with EGCV criterion and GIC.
- PIM2: PIM with EGCV criterion and BNPC.

For all MSCs, we use $\alpha = 2, 2\log\log n, \log n$, and they are labeled as 1, 2, and 3, respectively. Furthermore, the quartic equation in the BNPC minimization method is solved by the R function "polyroot".

Table 1 summarizes the RMSE and RNRE values for $\rho_y = 0.2, 0.5, 0.9$ and $k = 0.1n, 0.3n, 0.5n$ when $p = 5$ and $n = 50$. From this table, it can be discerned that the prediction accuracy of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\delta}}}$ is greater than that of $\hat{\boldsymbol{Y}}$ in most cases. We also found that although the RNRE values increase as $\alpha$ increases, i.e., as the amount of shrinkage increases, the prediction accuracies deteriorate because the amount of shrinkage is too large. Although the RMSE values tend to increase with increasing $\rho_y$ or $k$, this is caused by decreasing shrinkage. Table 2 summarizes the results when $p = 5$ and $n = 200$. Overall, tends are similar to those in Table 1. However, when $n = 50$ the amount of shrinkage substantially decreases. Table 3 summarizes the results when $p = 5$ and $n = 500$. In this case, the optimal ridge parameters often do not lead to improvements in prediction accuracies. This is because the amount of shrinkage is too large for the BNPC and too small for the methods. Tables 4 – 6 show the results when $p = 10$,

Table 1. RMSE comparison 1 when $p = 5$ and $n = 50$

| | | $\rho_y$ | 0.2 | | | 0.5 | | | 0.9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k$ | 5 | 15 | 25 | 5 | 15 | 25 | 5 | 15 | 25 |
| $GC_p$ | 1 | RMSE | 49.40 | 47.52 | 48.94 | 51.43 | 50.44 | 52.18 | 65.17 | 66.74 | **71.24** |
| | | RNRE | 36.22 | 31.99 | 28.52 | 34.80 | 29.95 | 26.40 | 24.02 | 19.03 | 14.69 |
| | 2 | RMSE | 45.30 | 44.06 | 45.24 | 47.60 | 47.54 | 49.20 | 63.96 | **66.23** | 71.53 |
| | | RNRE | 50.10 | 45.39 | 41.63 | 48.26 | 42.82 | 38.77 | 34.74 | 28.39 | 22.35 |
| | 3 | RMSE | 43.19 | 43.69 | **44.27** | 45.65 | 48.03 | 49.31 | 66.18 | 69.91 | 76.91 |
| | | RNRE | 64.97 | 60.97 | 58.01 | 63.32 | 57.91 | 54.72 | 47.83 | 40.62 | 33.31 |
| EGCV | 1 | RMSE | 49.59 | 48.21 | 50.37 | 51.60 | 51.01 | 53.33 | 65.21 | 66.88 | 71.40 |
| | | RNRE | 35.72 | 30.11 | 25.18 | 34.27 | 28.31 | 23.49 | 23.80 | 18.35 | 13.68 |
| | 2 | RMSE | 45.29 | 44.00 | 44.99 | 47.57 | 47.47 | 48.97 | **63.95** | 66.28 | 72.02 |
| | | RNRE | 50.34 | 45.71 | 43.00 | 48.50 | 43.29 | 40.47 | 35.08 | 29.25 | 24.39 |
| | 3 | RMSE | 43.12 | 44.08 | 45.02 | 45.60 | 48.70 | 51.01 | 66.52 | 71.98 | 86.35 |
| | | RNRE | 66.00 | 64.46 | 66.79 | 64.36 | 61.71 | 64.07 | 49.11 | 44.60 | 43.72 |
| GIC | 1 | RMSE | 50.02 | 50.80 | 57.71 | 52.08 | 53.59 | 60.48 | 65.78 | 69.00 | 76.44 |
| | | RNRE | 37.11 | 27.27 | 16.50 | 35.57 | 25.32 | 14.83 | 24.39 | 14.91 | 6.82 |
| | 2 | RMSE | 45.21 | 44.74 | 47.42 | 47.60 | 48.28 | 51.31 | 64.38 | 66.66 | 72.30 |
| | | RNRE | 53.65 | 46.23 | 38.21 | 51.77 | 43.42 | 34.76 | 37.31 | 27.15 | 15.71 |
| | 3 | RMSE | 42.88 | 44.54 | 45.42 | 45.51 | 49.67 | 51.73 | 67.94 | 72.38 | 83.41 |
| | | RNRE | 70.30 | 68.78 | 70.97 | 69.00 | 66.07 | 68.09 | 53.81 | 46.64 | 42.07 |
| BNPC | 1 | RMSE | 48.17 | 45.60 | 50.84 | 50.42 | 49.82 | 61.45 | 65.45 | 68.53 | 145.94 |
| | | RNRE | 46.53 | 54.76 | 85.12 | 44.61 | 51.70 | 84.47 | 31.15 | 29.32 | 57.38 |
| | 2 | RMSE | 43.73 | 48.99 | 59.09 | 46.39 | 57.16 | 75.94 | 66.25 | 102.30 | $***$ |
| | | RNRE | 66.23 | 79.62 | 91.35 | 64.70 | 78.68 | 91.40 | 49.52 | 62.36 | 91.06 |
| | 3 | RMSE | **42.80** | 66.04 | 89.59 | 45.60 | 79.82 | 118.06 | 76.70 | $***$ | $***$ |
| | | RNRE | 78.76 | 88.69 | 92.94 | 78.34 | 88.57 | 93.34 | 69.75 | 85.94 | 94.79 |
| PIM1 | 1 | RMSE | 48.84 | 47.34 | 48.91 | 50.94 | 50.30 | 52.11 | 65.06 | 66.78 | 71.46 |
| | | RNRE | 39.93 | 34.80 | 30.91 | 38.20 | 32.79 | 29.01 | 26.84 | 21.69 | 17.51 |
| | 2 | RMSE | 44.59 | **43.59** | 44.28 | 47.01 | **47.43** | **48.84** | 64.38 | 67.36 | 75.72 |
| | | RNRE | 56.00 | 52.84 | 52.60 | 54.31 | 50.29 | 50.00 | 39.97 | 35.09 | 32.80 |
| | 3 | RMSE | **42.80** | 45.39 | 46.58 | **45.43** | 51.05 | 53.90 | 68.89 | 77.46 | 103.47 |
| | | RNRE | 71.71 | 72.21 | 76.51 | 70.49 | 69.92 | 74.35 | 55.97 | 52.99 | 56.55 |
| PIM2 | 1 | RMSE | 48.02 | 46.50 | 47.93 | 50.22 | 49.78 | 51.97 | 65.02 | 67.28 | 78.35 |
| | | RNRE | 45.41 | 42.84 | 44.79 | 43.57 | 40.82 | 42.86 | 31.25 | 27.98 | 30.58 |
| | 2 | RMSE | 43.90 | 44.34 | 46.32 | 46.49 | 49.30 | 52.71 | 65.52 | 72.54 | 103.73 |
| | | RNRE | 63.43 | 63.54 | 67.10 | 61.89 | 61.48 | 65.29 | 47.13 | 46.20 | 52.24 |
| | 3 | RMSE | 42.64 | 49.52 | 50.09 | 45.46 | 57.84 | 59.65 | 73.92 | 113.98 | $***$ |
| | | RNRE | 77.31 | 80.65 | 84.19 | 76.70 | 79.66 | 82.99 | 65.95 | 67.35 | 73.37 |

Note: Emboldened entries represent the minimum of the RMSE values in each column; $***$ denotes values greater than 150.

and we can see that tends are similar compared to the case where $p = 5$.

In RMSE comparison 1, the iteration method was used to optimize the ridge parameters using the GIC and BNPC minimization methods. However, these optimal ridge parameters can

Table 2. RMSE comparison 1 when $p = 5$ and $n = 200$

| | | $\rho_y$ | 0.2 | | | 0.5 | | | 0.9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k$ | 20 | 60 | 100 | 20 | 60 | 100 | 20 | 60 | 100 |
| $GC_p$ | 1 | RMSE | 68.57 | 75.92 | 81.39 | 73.54 | 80.32 | 84.51 | 89.94 | 92.30 | 94.24 |
| | | RNRE | 19.37 | 13.63 | 9.84 | 15.86 | 10.71 | 7.94 | 5.17 | 3.77 | 2.56 |
| | 2 | RMSE | 73.88 | 82.70 | 89.17 | 80.85 | 88.41 | 92.48 | 97.35 | 98.59 | 99.68 |
| | | RNRE | 37.35 | 26.81 | 20.12 | 31.19 | 21.43 | 16.44 | 12.30 | 7.62 | 5.63 |
| | 3 | RMSE | 89.84 | 101.56 | 110.81 | 100.53 | 110.59 | 114.55 | 117.17 | 116.58 | 115.30 |
| | | RNRE | 55.56 | 41.74 | 32.45 | 48.38 | 34.31 | 27.06 | 21.37 | 12.67 | 9.73 |
| EGCV | 1 | RMSE | **68.56** | **75.88** | **81.27** | **73.52** | **80.25** | **84.39** | 89.93 | 92.27 | 94.20 |
| | | RNRE | 19.16 | 13.23 | 9.34 | 15.73 | 10.45 | 7.62 | 5.16 | 3.74 | 2.52 |
| | 2 | RMSE | 74.32 | 84.71 | 95.15 | 81.45 | 90.81 | 98.34 | 97.81 | 99.70 | 101.84 |
| | | RNRE | 38.18 | 29.04 | 24.51 | 31.99 | 23.36 | 20.07 | 12.62 | 8.08 | 6.43 |
| | 3 | RMSE | 92.62 | 115.19 | $***$ | 104.26 | 129.24 | $***$ | 120.70 | 128.37 | 148.20 |
| | | RNRE | 57.77 | 48.88 | 48.61 | 50.87 | 42.05 | 42.90 | 22.52 | 15.02 | 14.72 |
| GIC | 1 | RMSE | 68.76 | 76.45 | 83.20 | 73.60 | 80.41 | 85.79 | **89.66** | **91.89** | 94.34 |
| | | RNRE | 18.08 | 9.19 | 3.57 | 14.79 | 7.07 | 2.79 | 4.57 | 2.29 | 0.75 |
| | 2 | RMSE | 74.21 | 80.60 | 83.03 | 81.29 | 85.67 | 85.72 | 96.60 | 94.62 | **94.11** |
| | | RNRE | 38.07 | 24.25 | 12.93 | 31.78 | 18.89 | 9.73 | 11.91 | 5.69 | 2.35 |
| | 3 | RMSE | 94.07 | 110.50 | 136.68 | 106.84 | 123.88 | 141.34 | 120.95 | 114.66 | 103.06 |
| | | RNRE | 59.24 | 47.18 | 42.03 | 52.74 | 40.57 | 35.23 | 22.57 | 12.38 | 6.67 |
| BNPC | 1 | RMSE | 68.88 | 76.24 | 82.13 | 73.89 | 80.43 | 85.06 | 89.88 | 91.92 | 94.25 |
| | | RNRE | 20.41 | 12.45 | 5.71 | 16.51 | 9.24 | 4.03 | 5.01 | 2.61 | 0.87 |
| | 2 | RMSE | 78.89 | 119.39 | $***$ | 87.55 | 147.83 | $***$ | 99.54 | 99.52 | $***$ |
| | | RNRE | 45.23 | 51.28 | 94.39 | 38.44 | 49.56 | 95.01 | 14.02 | 8.05 | 5.59 |
| | 3 | RMSE | 110.61 | $***$ | $***$ | 130.66 | $***$ | $***$ | $***$ | $***$ | $***$ |
| | | RNRE | 68.76 | 83.54 | 97.03 | 65.39 | 85.89 | 97.02 | 30.66 | 94.16 | 97.15 |
| PIM1 | 1 | RMSE | 68.67 | 76.04 | 81.62 | 73.68 | 80.49 | 84.80 | 90.06 | 92.44 | 94.41 |
| | | RNRE | 19.83 | 13.84 | 10.06 | 16.28 | 10.99 | 8.23 | 5.39 | 3.93 | 2.72 |
| | 2 | RMSE | 75.29 | 86.59 | 101.64 | 82.87 | 93.71 | 106.56 | 98.83 | 101.75 | 105.66 |
| | | RNRE | 39.87 | 30.98 | 27.91 | 33.62 | 25.36 | 23.46 | 13.40 | 8.83 | 7.37 |
| | 3 | RMSE | 95.74 | 121.86 | $***$ | 109.29 | 141.32 | $***$ | 126.93 | 147.81 | $***$ |
| | | RNRE | 60.36 | 52.26 | 53.79 | 54.14 | 47.10 | 50.12 | 24.11 | 18.30 | 25.30 |
| PIM2 | 1 | RMSE | 68.80 | 76.32 | 82.79 | 73.88 | 80.89 | 86.11 | 90.21 | 92.67 | 94.78 |
| | | RNRE | 20.62 | 14.67 | 11.50 | 16.91 | 11.76 | 9.49 | 5.64 | 4.17 | 3.01 |
| | 2 | RMSE | 76.58 | 90.23 | 122.20 | 84.85 | 100.37 | 140.97 | 100.18 | 106.38 | $***$ |
| | | RNRE | 41.95 | 34.07 | 35.28 | 35.68 | 29.12 | 32.68 | 14.36 | 10.24 | 15.98 |
| | 3 | RMSE | 100.12 | 133.69 | $***$ | 116.54 | $***$ | $***$ | 139.98 | $***$ | $***$ |
| | | RNRE | 63.44 | 56.30 | 59.23 | 58.54 | 53.67 | 58.34 | 26.59 | 34.39 | 67.16 |

Note: Emboldened entries represent the minimum of the RMSE values in each column; $***$ denotes values greater than 150.

also be calculated by using the coordinate descent algorithm or the PIM with the $GC_p$ criterion. RMSE comparison 2 confirms whether the three algorithms minimize the MSC or not by comparing the results obtained from these algorithms. Although the initial vector used in

Table 3. RMSE comparison 1 when $p = 5$ and $n = 500$

| | | $\rho_y$ | 0.2 | | | 0.5 | | | 0.9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k$ | 50 | 150 | 250 | 50 | 150 | 250 | 50 | 150 | 250 |
| $GC_p$ | 1 | RMSE | 87.42 | 94.49 | 94.98 | 89.93 | 95.84 | 96.11 | 97.21 | 99.19 | 98.84 |
| | | RNRE | 6.75 | 2.27 | 2.14 | 5.23 | 1.58 | 1.56 | 1.11 | 0.11 | 0.40 |
| | 2 | RMSE | 98.40 | 104.50 | 103.78 | 100.49 | 105.05 | 104.03 | 103.37 | 103.50 | 102.07 |
| | | RNRE | 15.90 | 6.95 | 5.71 | 12.45 | 5.07 | 4.13 | 3.62 | 0.60 | 1.07 |
| | 3 | RMSE | 130.30 | 131.79 | 129.36 | 130.42 | 130.22 | 127.53 | 120.89 | 116.88 | 112.61 |
| | | RNRE | 26.47 | 14.69 | 11.48 | 21.95 | 11.59 | 8.63 | 7.77 | 1.97 | 2.09 |
| EGCV | 1 | RMSE | 87.40 | 94.47 | 94.95 | 89.92 | 95.83 | 96.08 | 97.20 | 99.19 | 98.84 |
| | | RNRE | 6.72 | 2.26 | 2.11 | 5.21 | 1.57 | 1.55 | 1.11 | 0.11 | 0.40 |
| | 2 | RMSE | 99.18 | 106.79 | 108.56 | 101.17 | 106.90 | 107.68 | 103.57 | 103.84 | 102.57 |
| | | RNRE | 16.33 | 7.79 | 7.05 | 12.76 | 5.66 | 4.97 | 3.70 | 0.63 | 1.14 |
| | 3 | RMSE | 137.37 | ∗ ∗ ∗ | ∗ ∗ ∗ | 136.36 | 149.81 | ∗ ∗ ∗ | 122.82 | 121.41 | 121.10 |
| | | RNRE | 27.93 | 19.22 | 23.61 | 23.31 | 15.38 | 18.15 | 8.11 | 2.38 | 2.66 |
| GIC | 1 | RMSE | **87.10** | **93.58** | 94.63 | **89.59** | **94.93** | 95.70 | **96.92** | **98.70** | **98.62** |
| | | RNRE | 5.94 | 1.17 | 0.60 | 4.59 | 0.78 | 0.44 | 0.92 | 0.04 | 0.10 |
| | 2 | RMSE | 97.08 | 98.71 | 95.24 | 99.03 | 99.46 | 96.23 | 102.06 | 100.40 | 98.76 |
| | | RNRE | 15.27 | 4.64 | 2.32 | 11.84 | 3.19 | 1.62 | 3.24 | 0.25 | 0.37 |
| | 3 | RMSE | 134.11 | 129.57 | 115.52 | 132.87 | 126.22 | 111.41 | 119.31 | 108.82 | 101.60 |
| | | RNRE | 27.37 | 14.24 | 8.70 | 22.65 | 10.65 | 5.69 | 7.62 | 1.20 | 1.00 |
| BNPC | 1 | RMSE | 87.22 | 93.62 | **94.55** | 89.70 | 94.98 | **95.65** | 96.96 | 98.71 | **98.62** |
| | | RNRE | 6.28 | 1.30 | 0.67 | 4.82 | 0.86 | 0.48 | 0.95 | 0.04 | 0.11 |
| | 2 | RMSE | 100.13 | 103.66 | ∗ ∗ ∗ | 101.56 | 102.85 | 97.88 | 102.71 | 100.70 | 98.84 |
| | | RNRE | 17.00 | 6.77 | 4.32 | 13.10 | 4.41 | 2.34 | 3.51 | 0.29 | 0.41 |
| | 3 | RMSE | ∗ ∗ ∗ | ∗ ∗ ∗ | ∗ ∗ ∗ | ∗ ∗ ∗ | ∗ ∗ ∗ | ∗ ∗ ∗ | 124.39 | 114.88 | ∗ ∗ ∗ |
| | | RNRE | 32.58 | 98.69 | $100^*$ | 27.14 | 99.36 | $100^*$ | 8.64 | 1.85 | 1.43 |
| PIM1 | 1 | RMSE | 87.44 | 94.54 | 95.05 | 89.97 | 95.89 | 96.19 | 97.23 | 99.21 | 98.86 |
| | | RNRE | 6.83 | 2.31 | 2.19 | 5.29 | 1.61 | 1.60 | 1.13 | 0.12 | 0.41 |
| | 2 | RMSE | 99.66 | 107.65 | 111.08 | 101.67 | 107.89 | 110.47 | 103.83 | 104.11 | 103.06 |
| | | RNRE | 16.66 | 8.14 | 7.67 | 13.06 | 5.96 | 5.49 | 3.82 | 0.68 | 1.21 |
| | 3 | RMSE | 140.71 | ∗ ∗ ∗ | ∗ ∗ ∗ | 139.49 | ∗ ∗ ∗ | ∗ ∗ ∗ | 124.29 | 124.55 | 132.50 |
| | | RNRE | 28.64 | 20.44 | 28.29 | 24.08 | 16.84 | 24.37 | 8.46 | 2.66 | 3.24 |
| PIM2 | 1 | RMSE | 87.49 | 94.61 | 95.20 | 90.01 | 95.97 | 96.34 | 97.26 | 99.23 | 98.89 |
| | | RNRE | 6.94 | 2.38 | 2.28 | 5.38 | 1.66 | 1.67 | 1.16 | 0.12 | 0.43 |
| | 2 | RMSE | 100.23 | 108.98 | 119.90 | 102.25 | 109.44 | 121.35 | 104.11 | 104.46 | 103.86 |
| | | RNRE | 17.03 | 8.64 | 9.25 | 13.37 | 6.39 | 7.14 | 3.97 | 0.73 | 1.31 |
| | 3 | RMSE | 145.66 | ∗ ∗ ∗ | ∗ ∗ ∗ | 143.99 | ∗ ∗ ∗ | ∗ ∗ ∗ | 126.11 | 131.62 | ∗ ∗ ∗ |
| | | RNRE | 29.54 | 23.09 | 35.64 | 25.03 | 20.81 | 38.87 | 8.88 | 3.19 | 59.05 |

Note: Emboldened entries represent the minimum of the RMSE values in each column; ∗ ∗ ∗ denotes values greater than 150; $^*$ denotes an exact value.

the iterative method is $\mathbf{0}_k$, the PIM with the $GC_p$ criterion is the iterative method by changing the initial vector from $\mathbf{0}_k$ to the ridge parameters optimized by the $GC_p$ criterion minimization method. Hence, by comparing the results obtained from the two methods, we can confirm

Table 4. RMSE comparison 1 when $p = 10$ and $n = 50$

|  |  | $\rho_y$ | 0.2 |  |  | 0.5 |  |  | 0.9 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | $k$ | 5 | 15 | 25 | 5 | 15 | 25 | 5 | 15 | 25 |
| $GC_p$ | 1 | RMSE | 47.25 | 43.85 | 48.37 | 49.21 | 46.54 | 51.65 | 61.62 | 62.52 | 68.72 |
|  |  | RNRE | 24.78 | 19.56 | 12.83 | 23.24 | 18.20 | 11.48 | 14.52 | 10.18 | 5.72 |
|  | 2 | RMSE | 42.71 | 38.81 | 42.17 | 44.95 | 41.82 | 46.06 | 59.32 | 60.16 | 65.84 |
|  |  | RNRE | 42.72 | 35.07 | 24.76 | 40.42 | 32.90 | 22.35 | 27.17 | 19.76 | 11.70 |
|  | 3 | RMSE | 40.61 | 37.16 | 38.43 | 43.07 | 40.62 | 43.24 | 60.65 | 61.73 | 66.31 |
|  |  | RNRE | 63.05 | 55.28 | 43.57 | 60.99 | 52.36 | 39.81 | 43.92 | 34.58 | 22.26 |
| EGCV | 1 | RMSE | 47.36 | 44.27 | 49.05 | 49.30 | 46.90 | 52.21 | 61.65 | 62.63 | 68.87 |
|  |  | RNRE | 24.49 | 18.68 | 11.86 | 22.94 | 17.40 | 10.71 | 14.44 | 9.94 | 5.53 |
|  | 2 | RMSE | 42.60 | 38.34 | 40.39 | 44.85 | 41.40 | 44.55 | **59.30** | **60.09** | **65.50** |
|  |  | RNRE | 43.50 | 37.75 | 31.19 | 41.25 | 35.54 | 28.33 | 27.92 | 21.71 | 14.89 |
|  | 3 | RMSE | 40.55 | 37.68 | 38.50 | 43.03 | 41.41 | 44.97 | 61.08 | 63.98 | 73.76 |
|  |  | RNRE | 64.88 | 62.77 | 63.23 | 62.91 | 59.69 | 59.41 | 46.10 | 41.77 | 37.62 |
| GIC | 1 | RMSE | 46.61 | 46.01 | 58.58 | 48.71 | 48.85 | 61.80 | 61.73 | 65.21 | 76.91 |
|  |  | RNRE | 32.68 | 21.04 | 6.86 | 30.68 | 19.21 | 5.69 | 19.44 | 9.33 | 1.96 |
|  | 2 | RMSE | 41.50 | 37.85 | 40.88 | 44.02 | 41.08 | 45.91 | 59.74 | 61.02 | 69.04 |
|  |  | RNRE | 58.42 | 50.97 | 40.64 | 56.12 | 47.47 | 34.49 | 39.19 | 27.44 | 10.41 |
|  | 3 | RMSE | 40.14 | 40.11 | 40.84 | 42.93 | 44.33 | 49.30 | 65.66 | 68.82 | 81.02 |
|  |  | RNRE | 77.07 | 78.59 | 82.57 | 76.24 | 75.06 | 78.72 | 60.97 | 57.74 | 51.56 |
| BNPC | 1 | RMSE | 42.71 | 41.12 | 51.02 | 45.39 | 45.33 | 68.20 | 61.06 | 68.18 | $***$ |
|  |  | RNRE | 64.79 | 79.62 | 91.59 | 62.73 | 75.73 | 91.38 | 44.03 | 54.31 | 78.60 |
|  | 2 | RMSE | **39.83** | 52.83 | 72.67 | **42.90** | 62.50 | 97.41 | 68.37 | 91.06 | $***$ |
|  |  | RNRE | 79.04 | 89.89 | 93.29 | 78.69 | 87.98 | 93.56 | 67.08 | 73.92 | 92.37 |
|  | 3 | RMSE | 41.21 | 61.96 | 111.45 | 43.32 | 79.90 | 138.42 | 82.95 | $***$ | $***$ |
|  |  | RNRE | 80.03 | 93.12 | 95.21 | 79.99 | 92.84 | 95.38 | 78.97 | 83.33 | 94.81 |
| PIM1 | 1 | RMSE | 45.28 | 41.15 | 43.12 | 47.42 | 44.00 | 46.92 | 60.75 | 61.49 | 66.64 |
|  |  | RNRE | 36.96 | 32.41 | 27.88 | 34.84 | 30.58 | 25.62 | 23.15 | 18.80 | 14.55 |
|  | 2 | RMSE | 41.01 | **37.02** | **37.71** | 43.57 | **40.32** | **43.17** | 59.85 | 62.00 | 70.16 |
|  |  | RNRE | 61.72 | 59.62 | 61.03 | 59.60 | 56.79 | 57.30 | 43.06 | 39.97 | 37.79 |
|  | 3 | RMSE | 40.20 | 41.42 | 42.05 | 42.99 | 46.22 | 52.06 | 67.20 | 73.52 | 102.37 |
|  |  | RNRE | 77.91 | 81.02 | 85.84 | 77.25 | 77.75 | 83.09 | 63.07 | 64.21 | 65.64 |
| PIM2 | 1 | RMSE | 42.94 | 39.22 | 41.39 | 45.49 | 42.37 | 47.38 | 60.56 | 64.27 | 80.77 |
|  |  | RNRE | 61.28 | 61.61 | 64.39 | 59.20 | 59.00 | 61.61 | 43.11 | 44.50 | 46.22 |
|  | 2 | RMSE | 39.90 | 41.28 | 42.22 | 42.97 | 46.11 | 52.02 | 66.13 | 73.27 | 111.18 |
|  |  | RNRE | 78.32 | 81.28 | 83.70 | 77.78 | 78.42 | 81.46 | 64.66 | 65.39 | 67.05 |
|  | 3 | RMSE | 40.45 | 53.45 | 48.40 | 43.14 | 63.21 | 63.07 | 81.84 | 95.55 | $***$ |
|  |  | RNRE | 79.99 | 90.19 | 91.41 | 79.97 | 88.32 | 90.86 | 78.23 | 75.29 | 80.91 |

Note: Emboldened entries represent the minimum of the RMSE values in each column; $***$ denotes values greater than 150.

whether the iterative method depends on the initial vector or not.

Table 7 compares the three algorithms for solving the GIC minimization method in terms of the RMSE, i.e., from the iterative method (GIC_IM), the coordinate descent algorithm

Table 5. RMSE comparison 1 when $p = 10$ and $n = 200$

| | | $\rho_y$ | 0.2 | | | 0.5 | | | 0.9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $k$ | 20 | 60 | 100 | 20 | 60 | 100 | 20 | 60 | 100 |
| $GC_p$ | 1 | RMSE | **57.38** | **67.35** | **72.47** | **62.60** | **72.10** | **76.68** | **83.29** | **88.44** | **90.86** |
| | | RNRE | 16.40 | 10.25 | 7.31 | 13.21 | 8.01 | 5.61 | 3.39 | 2.11 | 1.34 |
| | 2 | RMSE | 61.78 | 73.26 | 78.22 | 68.67 | 78.80 | 82.61 | 90.30 | 93.85 | 95.20 |
| | | RNRE | 40.86 | 27.18 | 20.49 | 33.81 | 21.51 | 16.27 | 11.54 | 6.27 | 4.25 |
| | 3 | RMSE | 76.54 | 92.15 | 98.54 | 87.31 | 100.35 | 103.24 | 111.52 | 111.78 | 110.88 |
| | | RNRE | 63.02 | 45.94 | 36.23 | 55.35 | 38.00 | 29.93 | 23.44 | 13.10 | 8.85 |
| EGCV | 1 | RMSE | 57.39 | 67.37 | 72.49 | 62.61 | **72.10** | **76.68** | **83.29** | **88.44** | **90.86** |
| | | RNRE | 16.24 | 10.00 | 6.98 | 13.11 | 7.86 | 5.43 | 3.38 | 2.10 | 1.33 |
| | 2 | RMSE | 62.35 | 76.04 | 85.74 | 69.43 | 81.88 | 89.90 | 90.91 | 95.27 | 97.88 |
| | | RNRE | 42.32 | 31.11 | 27.98 | 35.16 | 24.74 | 22.49 | 12.08 | 6.98 | 5.29 |
| | 3 | RMSE | 79.09 | 107.06 | 149.90 | 90.91 | 119.41 | $***$ | 115.99 | 126.01 | $***$ |
| | | RNRE | 65.46 | 54.49 | 55.01 | 58.39 | 47.89 | 48.46 | 25.20 | 17.08 | 17.74 |
| GIC | 1 | RMSE | 57.70 | 68.76 | 77.17 | 62.88 | 73.26 | 80.71 | **83.29** | 88.83 | 92.49 |
| | | RNRE | 16.01 | 6.06 | 1.54 | 12.75 | 4.55 | 1.09 | 3.03 | 0.98 | 0.18 |
| | 2 | RMSE | 63.40 | 73.48 | 74.93 | 70.68 | 78.53 | 78.36 | 90.51 | 90.67 | 90.93 |
| | | RNRE | 45.18 | 27.99 | 15.35 | 37.62 | 21.40 | 10.36 | 12.30 | 4.58 | 1.37 |
| | 3 | RMSE | 83.94 | 110.14 | 144.62 | 97.64 | 122.96 | $***$ | 121.44 | 118.06 | 105.65 |
| | | RNRE | 69.91 | 56.63 | 54.61 | 64.36 | 51.24 | 47.69 | 27.74 | 15.57 | 7.72 |
| BNPC | 1 | RMSE | 57.66 | 68.03 | 87.36 | 63.06 | 72.70 | 78.78 | 83.50 | 88.68 | 92.23 |
| | | RNRE | 21.14 | 12.84 | 13.01 | 16.82 | 8.71 | 3.12 | 3.88 | 1.36 | 0.26 |
| | 2 | RMSE | 71.91 | $***$ | $***$ | 82.22 | $***$ | $***$ | 96.74 | $***$ | $***$ |
| | | RNRE | 59.59 | 71.87 | 92.54 | 52.43 | 69.76 | 93.04 | 17.72 | 15.24 | 77.56 |
| | 3 | RMSE | 105.05 | $***$ | $***$ | 126.17 | $***$ | $***$ | $***$ | $***$ | $***$ |
| | | RNRE | 82.08 | 87.21 | 95.67 | 80.44 | 87.80 | 96.30 | 50.65 | 91.06 | 99.96 |
| PIM1 | 1 | RMSE | 57.43 | 67.50 | 72.73 | 62.73 | 72.31 | 77.03 | 83.47 | 88.61 | 91.07 |
| | | RNRE | 18.23 | 11.71 | 8.94 | 14.77 | 9.26 | 7.03 | 3.91 | 2.48 | 1.72 |
| | 2 | RMSE | 64.42 | 80.28 | 96.26 | 72.19 | 87.80 | 103.49 | 93.06 | 99.03 | 107.22 |
| | | RNRE | 47.27 | 36.70 | 35.58 | 39.83 | 30.17 | 30.40 | 14.29 | 8.93 | 8.09 |
| | 3 | RMSE | 85.21 | 120.90 | $***$ | 99.44 | 136.43 | $***$ | 128.16 | $***$ | $***$ |
| | | RNRE | 70.72 | 60.15 | 63.32 | 65.58 | 56.05 | 59.38 | 29.66 | 23.98 | 33.96 |
| PIM2 | 1 | RMSE | 57.59 | 68.00 | 74.63 | 62.99 | 72.97 | 79.46 | 83.71 | 88.93 | 91.81 |
| | | RNRE | 20.71 | 14.32 | 13.19 | 16.93 | 11.42 | 10.84 | 4.60 | 3.02 | 2.49 |
| | 2 | RMSE | 67.66 | 88.23 | 118.06 | 76.73 | 100.31 | 140.80 | 96.47 | 110.29 | $***$ |
| | | RNRE | 53.40 | 44.46 | 45.76 | 46.19 | 39.71 | 42.74 | 17.45 | 13.81 | 23.59 |
| | 3 | RMSE | 95.20 | $***$ | $***$ | 112.90 | $***$ | $***$ | $***$ | $***$ | $***$ |
| | | RNRE | 77.46 | 68.98 | 71.41 | 74.39 | 65.19 | 70.63 | 38.43 | 49.73 | 67.58 |

Note: Emboldened entries represent the minimum of the RMSE values in each column; $***$ denotes values greater than 150.

(GIC_CD), and the PIM with the $GC_p$ criterion (PIM_$GC_p$). Settings are as per RMSE comparison 1, where $\alpha$ is only $\alpha = 2$. From these results, it can be discerned that there is equivalent performance among the three algorithms. Although there is a bit of error, it can be considered

Table 6. RMSE comparison 1 when $p = 10$ and $n = 500$

|  |  | $\rho_y$ | 0.2 |  |  | 0.5 |  |  | 0.9 |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | $k$ | 50 | 150 | 250 | 50 | 150 | 250 | 50 | 150 | 250 |
| $GC_p$ | 1 | RMSE | 81.19 | 88.94 | 90.00 | 85.14 | 91.51 | **92.14** | 95.37 | 97.82 | 97.67 |
|  |  | RNRE | 4.32 | 1.56 | 1.66 | 2.93 | 0.89 | 1.12 | 0.38 | 0.03 | 0.20 |
|  | 2 | RMSE | 93.28 | 98.48 | 97.82 | 96.38 | 100.05 | 99.03 | 100.55 | 101.06 | 100.33 |
|  |  | RNRE | 15.09 | 7.32 | 6.46 | 10.64 | 4.69 | 4.47 | 2.03 | 0.27 | 0.77 |
|  | 3 | RMSE | 131.47 | 128.21 | 124.25 | 132.40 | 127.22 | 122.64 | 119.74 | 113.93 | 111.01 |
|  |  | RNRE | 28.73 | 17.16 | 13.68 | 22.07 | 12.58 | 10.26 | 4.66 | 1.06 | 1.63 |
| EGCV | 1 | RMSE | 81.19 | 88.93 | **89.99** | 85.14 | 91.51 | **92.14** | 95.36 | 97.82 | 97.67 |
|  |  | RNRE | 4.31 | 1.55 | 1.64 | 2.93 | 0.89 | 1.11 | 0.38 | 0.03 | 0.20 |
|  | 2 | RMSE | 94.53 | 101.57 | 103.97 | 97.40 | 102.38 | 103.49 | 100.75 | 101.35 | 100.85 |
|  |  | RNRE | 15.80 | 8.72 | 8.72 | 11.11 | 5.54 | 5.89 | 2.07 | 0.29 | 0.83 |
|  | 3 | RMSE | 142.15 | $***$ | $***$ | 141.78 | $***$ | $***$ | 122.24 | 118.84 | 122.10 |
|  |  | RNRE | 31.45 | 23.46 | 30.84 | 24.22 | 17.55 | 22.35 | 4.93 | 1.35 | 2.43 |
| GIC | 1 | RMSE | **81.06** | 88.96 | 91.45 | **84.98** | 91.45 | 93.26 | **95.24** | **97.75** | 97.99 |
|  |  | RNRE | 3.66 | 0.59 | 0.22 | 2.44 | 0.32 | 0.14 | 0.28 | 0.01 | 0.02 |
|  | 2 | RMSE | 92.92 | 93.59 | 90.43 | 95.60 | 95.22 | 92.38 | 99.36 | 98.62 | **97.64** |
|  |  | RNRE | 15.10 | 4.93 | 2.18 | 10.48 | 2.82 | 1.33 | 1.83 | 0.10 | 0.19 |
|  | 3 | RMSE | 145.39 | 144.66 | 131.69 | 143.38 | 137.46 | 119.42 | 118.85 | 106.79 | 100.25 |
|  |  | RNRE | 32.53 | 20.22 | 14.90 | 24.88 | 14.36 | 9.51 | 4.75 | 0.68 | 0.77 |
| BNPC | 1 | RMSE | 81.18 | **88.85** | 91.17 | 85.07 | **91.39** | 93.08 | 95.26 | **97.75** | 97.98 |
|  |  | RNRE | 4.18 | 0.76 | 0.29 | 2.75 | 0.39 | 0.18 | 0.31 | 0.01 | 0.02 |
|  | 2 | RMSE | 99.64 | $***$ | $***$ | 100.71 | 106.53 | $***$ | 100.06 | 98.92 | 97.71 |
|  |  | RNRE | 18.76 | 16.06 | 99.59 | 13.00 | 6.64 | 69.70 | 2.09 | 0.14 | 0.23 |
|  | 3 | RMSE | $***$ | $***$ | $***$ | $***$ | $***$ | $***$ | 127.85 | $***$ | $***$ |
|  |  | RNRE | 50.52 | 96.67 | 99.90 | 40.25 | 96.76 | 100.00 | 5.99 | 37.40 | 90.00 |
| PIM1 | 1 | RMSE | 81.27 | 89.05 | 90.15 | 85.21 | 91.62 | 92.29 | 95.39 | 97.84 | 97.70 |
|  |  | RNRE | 4.55 | 1.70 | 1.85 | 3.09 | 0.98 | 1.25 | 0.41 | 0.04 | 0.22 |
|  | 2 | RMSE | 96.06 | 104.80 | 111.93 | 98.81 | 105.66 | 111.48 | 101.12 | 101.89 | 101.90 |
|  |  | RNRE | 16.79 | 10.01 | 10.84 | 11.94 | 6.62 | 7.83 | 2.24 | 0.35 | 0.95 |
|  | 3 | RMSE | $***$ | $***$ | $***$ | $***$ | $***$ | $***$ | 125.33 | 127.66 | $***$ |
|  |  | RNRE | 34.26 | 28.55 | 38.11 | 26.62 | 22.41 | 34.55 | 5.45 | 1.94 | 5.85 |
| PIM2 | 1 | RMSE | 81.36 | 89.21 | 90.51 | 85.30 | 91.78 | 92.62 | 95.41 | 97.86 | 97.73 |
|  |  | RNRE | 4.82 | 1.90 | 2.18 | 3.27 | 1.09 | 1.46 | 0.44 | 0.04 | 0.25 |
|  | 2 | RMSE | 98.10 | 111.58 | 137.23 | 100.68 | 113.05 | 147.94 | 101.54 | 102.68 | 104.57 |
|  |  | RNRE | 18.01 | 12.06 | 15.23 | 12.95 | 8.61 | 12.78 | 2.43 | 0.45 | 1.18 |
|  | 3 | RMSE | $***$ | $***$ | $***$ | $***$ | $***$ | $***$ | 129.75 | $***$ | $***$ |
|  |  | RNRE | 38.48 | 38.87 | 46.20 | 30.39 | 36.78 | 48.39 | 6.17 | 7.24 | 65.25 |

Note: Emboldened entries represent the minimum of the RMSE values in each column; $***$ denotes values greater than 150.

that the error is made when convergence judgment. Thus, the three algorithms all converge and achieve minimization of the GIC. Furthermore, we found that the iterative method does not depend on the initial vector.

Table 7. RMSE comparison 2 (GIC; $\alpha = 2$)

| | | | $p = 5$ | | | $p = 10$ | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\rho_y$ | $k$ | GIC_IM | GIC_CD | PIM_$GC_p$ | GIC_IM | GIC_CD | PIM_$GC_p$ |
| 50 | 0.2 | 5 | 50.02 | 50.02 | 50.01 | 46.61 | 46.60 | 46.61 |
| | | 15 | 50.80 | 50.79 | 50.78 | 46.01 | 45.99 | 45.97 |
| | | 25 | 57.71 | 57.69 | 57.66 | 58.58 | 58.54 | 58.47 |
| | 0.5 | 5 | 52.08 | 52.08 | 52.07 | 48.71 | 48.70 | 48.70 |
| | | 15 | 53.59 | 53.58 | 53.57 | 48.85 | 48.83 | 48.81 |
| | | 25 | 60.48 | 60.46 | 60.43 | 61.80 | 61.77 | 61.71 |
| | 0.9 | 5 | 65.78 | 65.78 | 65.77 | 61.73 | 61.73 | 61.73 |
| | | 15 | 69.00 | 68.99 | 68.99 | 65.21 | 65.20 | 65.18 |
| | | 25 | 76.44 | 76.44 | 76.43 | 76.91 | 76.90 | 76.89 |
| 200 | 0.2 | 20 | 68.76 | 68.76 | 68.76 | 57.70 | 57.70 | 57.70 |
| | | 60 | 76.45 | 76.45 | 76.45 | 68.76 | 68.76 | 68.76 |
| | | 100 | 83.20 | 83.19 | 83.19 | 77.17 | 77.16 | 77.16 |
| | 0.5 | 20 | 73.60 | 73.60 | 73.60 | 62.88 | 62.88 | 62.88 |
| | | 60 | 80.41 | 80.41 | 80.41 | 73.26 | 73.26 | 73.25 |
| | | 100 | 85.79 | 85.79 | 85.79 | 80.71 | 80.71 | 80.70 |
| | 0.9 | 20 | 89.66 | 89.66 | 89.67 | 83.29 | 83.29 | 83.29 |
| | | 60 | 91.89 | 91.89 | 91.89 | 88.83 | 88.83 | 88.83 |
| | | 100 | 94.34 | 94.34 | 94.34 | 92.49 | 92.49 | 92.49 |
| 500 | 0.2 | 50 | 87.10 | 87.10 | 87.10 | 81.06 | 81.06 | 81.06 |
| | | 150 | 93.58 | 93.58 | 93.58 | 88.96 | 88.96 | 88.96 |
| | | 250 | 94.63 | 94.63 | 94.63 | 91.45 | 91.45 | 91.45 |
| | 0.5 | 50 | 89.59 | 89.59 | 89.59 | 84.98 | 84.98 | 84.98 |
| | | 150 | 94.93 | 94.93 | 94.93 | 91.45 | 91.45 | 91.45 |
| | | 250 | 95.70 | 95.70 | 95.70 | 93.26 | 93.26 | 93.26 |
| | 0.9 | 50 | 96.92 | 96.92 | 96.92 | 95.24 | 95.24 | 95.24 |
| | | 150 | 98.70 | 98.70 | 98.70 | 97.75 | 97.75 | 97.75 |
| | | 250 | 98.62 | 98.62 | 98.62 | 97.99 | 97.99 | 97.99 |

Table 8 shows a runtime comparison of the three algorithms for the GIC minimization method in terms of time (s) per repeat, where the reported values are 10,000 times the actual values. The PIM is the fastest algorithm in most cases. Although sometimes the iterative method is faster than the PIM, this is related the initial vector and the amount of shrinkage. The difference between the PIM and the iterative method is the initial vector, and the iterative method is faster when the amount of shrinkage is small, i.e., the optimal ridge parameters are close to the initial vector $\mathbf{0}_k$. On the other hand, the coordinate descent algorithm is overwhelmingly slowest of all. Hence, the best option for solving the GIC minimization method is to use the PIM with the $GC_p$ criterion.

Table 9 compares the three algorithms for solving the BNPC minimization method, in terms of RMSE as similar to Table 7. It can be discerned that the three algorithms converge and achieve minimization of the BNPC, and the iterative method does not depend on the initial

Table 8. Runtime comparison (GIC; ×1/10,000 (s))

|  |  |  | $p = 5$ |  |  | $p = 10$ |  |  |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\rho_y$ | $k$ | GIC_IM | GIC_CD | PIM_$GC_p$ | GIC_IM | GIC_CD | PIM_$GC_p$ |
| 50 | 0.2 | 5 | 3.04 | 10.81 | **2.30** | 4.57 | 12.46 | **3.48** |
|  |  | 15 | 5.15 | 43.41 | **3.76** | 8.78 | 52.13 | **6.10** |
|  |  | 25 | 8.06 | 95.43 | **5.87** | 13.65 | 122.30 | **10.74** |
|  | 0.5 | 5 | 3.40 | 10.97 | **2.09** | 4.65 | 12.41 | **3.53** |
|  |  | 15 | 6.04 | 46.88 | **3.66** | 8.86 | 52.57 | **6.45** |
|  |  | 25 | 8.48 | 104.61 | **5.87** | 13.57 | 124.52 | **11.11** |
|  | 0.9 | 5 | 3.52 | 12.28 | **2.23** | 5.70 | 12.36 | **4.05** |
|  |  | 15 | 6.04 | 46.51 | **3.80** | 8.97 | 52.90 | **6.70** |
|  |  | 25 | 8.74 | 104.99 | **6.15** | 13.61 | 121.29 | **11.77** |
| 200 | 0.2 | 20 | 4.00 | 52.16 | **2.76** | 4.99 | 57.31 | **3.62** |
|  |  | 60 | 7.51 | 205.03 | **5.26** | 9.66 | 228.49 | **6.95** |
|  |  | 100 | 15.20 | 443.87 | **12.82** | 20.92 | 489.25 | **18.33** |
|  | 0.5 | 20 | 3.96 | 52.96 | **2.75** | 4.92 | 58.35 | **3.43** |
|  |  | 60 | 7.64 | 207.20 | **5.77** | 10.24 | 227.12 | **7.16** |
|  |  | 100 | 15.74 | 450.99 | **13.95** | 22.31 | 504.68 | **20.00** |
|  | 0.9 | 20 | 3.74 | 49.59 | **2.53** | 4.69 | 54.66 | **3.40** |
|  |  | 60 | 6.12 | 174.76 | **4.48** | 9.25 | 197.91 | **7.01** |
|  |  | 100 | **9.30** | 303.43 | 9.91 | 16.13 | 379.86 | **13.15** |
| 500 | 0.2 | 50 | 4.70 | 128.49 | **3.16** | 5.86 | 137.78 | **4.05** |
|  |  | 150 | 13.66 | 456.75 | **10.85** | 23.53 | 528.58 | **20.47** |
|  |  | 250 | 41.80 | 851.61 | **38.38** | 81.79 | 1051.05 | **54.42** |
|  | 0.5 | 50 | 4.50 | 126.82 | **3.24** | 5.87 | 134.42 | **3.99** |
|  |  | 150 | 13.52 | 440.54 | **11.35** | 22.19 | 521.07 | **18.68** |
|  |  | 250 | **34.72** | 798.15 | 37.24 | 67.56 | 986.76 | **47.96** |
|  | 0.9 | 50 | 3.91 | 109.08 | **2.66** | 5.13 | 112.50 | **3.37** |
|  |  | 150 | 10.16 | 348.91 | **10.19** | 15.00 | 360.90 | **14.80** |
|  |  | 250 | **21.28** | 559.92 | 26.01 | **32.38** | 607.38 | 37.82 |

Note: Emboldened entries represent the fastest time in each column.

vector.

Table 10 shows a runtime comparison of the three algorithms for the BNPC minimization method in terms of time (s) as per Table 8. Similar to what was noted above regarding the GIC minimization method, to solve the BNPC minimization method, using the PIM with the $GC_p$ criterion is the best option.

**Acknowledgment** I would like to express my deepest gratitude to my supervisor, Prof. Hirokazu Yanagihara of Hiroshima University, for stimulating my interest in statistics and for his invaluable support and guidance. I also thank Prof. Yasunori Fujikoshi, Prof. Hirofumi Wakaki, Dr. Keisuke Fukui, Dr. Shinpei Imori, Dr. Shintaro Hashimoto, and Dr. Ryoya Oda

Table 9. RMSE comparison 2 (BNPC; $\alpha = 2$)

| | | | $p = 5$ | | | $p = 10$ | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\rho_y$ | $k$ | BNPC_IM | BNPC_CD | PIM_$GC_p$ | BNPC_IM | BNPC_CD | PIM_$GC_p$ |
| 50 | 0.2 | 5 | 48.17 | 48.15 | 48.16 | 42.71 | 42.67 | 42.71 |
| | | 15 | 45.60 | 45.58 | 45.60 | 41.12 | 47.36 | 41.12 |
| | | 25 | 50.84 | 60.14 | 50.83 | 51.02 | 602.41 | 51.02 |
| | 0.5 | 5 | 50.42 | 50.41 | 50.42 | 45.39 | 45.36 | 45.39 |
| | | 15 | 49.82 | 49.83 | 49.82 | 45.33 | 55.91 | 45.33 |
| | | 25 | 61.45 | 87.35 | 61.45 | 68.20 | 1127.10 | 68.20 |
| | 0.9 | 5 | 65.45 | 65.45 | 65.45 | 61.06 | 61.12 | 61.06 |
| | | 15 | 68.53 | 68.56 | 68.52 | 68.18 | 138.36 | 68.18 |
| | | 25 | 145.94 | 510.55 | 146.63 | 193.23 | 9745.56 | 193.62 |
| 200 | 0.2 | 20 | 68.88 | 68.88 | 68.88 | 57.66 | 57.67 | 57.66 |
| | | 60 | 76.24 | 76.24 | 76.24 | 68.03 | 68.04 | 68.03 |
| | | 100 | 82.13 | 82.13 | 82.12 | 87.36 | 89.21 | 87.41 |
| | 0.5 | 20 | 73.89 | 73.89 | 73.89 | 63.06 | 63.06 | 63.06 |
| | | 60 | 80.43 | 80.44 | 80.44 | 72.70 | 72.70 | 72.69 |
| | | 100 | 85.06 | 85.06 | 85.05 | 78.78 | 78.77 | 78.74 |
| | 0.9 | 20 | 89.88 | 89.88 | 89.88 | 83.50 | 83.51 | 83.50 |
| | | 60 | 91.92 | 91.92 | 91.92 | 88.68 | 88.68 | 88.68 |
| | | 100 | 94.25 | 94.24 | 94.24 | 92.23 | 92.23 | 92.22 |
| 500 | 0.2 | 50 | 87.22 | 87.22 | 87.22 | 81.18 | 81.18 | 81.18 |
| | | 150 | 93.62 | 93.62 | 93.62 | 88.85 | 88.85 | 88.85 |
| | | 250 | 94.55 | 94.55 | 94.55 | 91.17 | 91.17 | 91.17 |
| | 0.5 | 50 | 89.70 | 89.70 | 89.70 | 85.07 | 85.08 | 85.08 |
| | | 150 | 94.98 | 94.98 | 94.98 | 91.39 | 91.39 | 91.39 |
| | | 250 | 95.65 | 95.65 | 95.65 | 93.08 | 93.08 | 93.08 |
| | 0.9 | 50 | 96.96 | 96.96 | 96.96 | 95.26 | 95.26 | 95.27 |
| | | 150 | 98.71 | 98.71 | 98.71 | 97.75 | 97.75 | 97.75 |
| | | 250 | 98.62 | 98.62 | 98.62 | 97.98 | 97.98 | 97.98 |

of Hiroshima University and Dr. Mariko Yamamura of Radiation Effects Research Foundation for their helpful comments and encouragement.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory. Akadémiai Kiadó. Budapest. 267–281.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Inc. New Jersey.

Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model.

Table 10. Runtime comparison (BNPC; ×1/10,000 (s))

| | | | $p = 5$ | | | $p = 10$ | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $\rho_y$ | $k$ | BNPC_IM | BNPC_CD | PIM_$GC_p$ | BNPC_IM | BNPC_CD | PIM_$GC_p$ |
| 50 | 0.2 | 5 | 4.39 | 68.80 | **3.43** | 5.92 | 76.06 | **4.95** |
| | | 15 | 9.73 | 337.69 | **8.10** | 11.66 | 281.44 | **10.57** |
| | | 25 | 11.36 | 621.02 | **9.32** | 12.21 | 472.33 | **9.96** |
| | 0.5 | 5 | 4.34 | 71.33 | **3.45** | 6.07 | 73.57 | **5.23** |
| | | 15 | 10.17 | 364.55 | **8.64** | 12.77 | 283.79 | **11.63** |
| | | 25 | 12.20 | 650.78 | **10.31** | 13.28 | 482.18 | **10.76** |
| | 0.9 | 5 | 5.29 | 72.58 | **4.26** | 7.73 | 70.38 | **6.87** |
| | | 15 | 12.20 | 418.82 | **9.86** | 16.47 | 320.11 | **14.59** |
| | | 25 | 18.36 | 886.39 | **15.46** | 23.12 | 604.47 | **20.86** |
| 200 | 0.2 | 20 | 5.27 | 376.38 | **4.04** | 7.73 | 390.91 | **6.68** |
| | | 60 | 17.64 | 2295.43 | **16.25** | 26.98 | 2116.54 | **24.41** |
| | | 100 | 27.54 | 3635.11 | **26.46** | 53.83 | 4494.30 | **48.62** |
| | 0.5 | 20 | 5.47 | 388.16 | **4.37** | 8.05 | 414.10 | **7.12** |
| | | 60 | 20.60 | 2500.19 | **19.32** | 28.91 | 2341.24 | **26.35** |
| | | 100 | 28.81 | 3581.45 | **26.60** | 41.19 | 3368.07 | **39.14** |
| | 0.9 | 20 | 6.28 | 418.43 | **4.93** | 11.00 | 492.37 | **9.63** |
| | | 60 | 15.07 | 1676.25 | **12.75** | 25.66 | 2106.34 | **25.16** |
| | | 100 | 19.40 | 2759.43 | **17.84** | 45.79 | 3869.82 | **39.98** |
| 500 | 0.2 | 50 | 6.28 | 958.66 | **5.03** | 10.85 | 1175.67 | **9.16** |
| | | 150 | 35.73 | 4509.76 | **29.75** | 68.50 | 6669.25 | **64.67** |
| | | 250 | 59.00 | 6650.66 | **53.98** | 142.62 | 8408.83 | **127.37** |
| | 0.5 | 50 | 6.19 | 928.01 | **4.70** | 11.59 | 1274.18 | **10.46** |
| | | 150 | 40.44 | 4488.88 | **34.88** | 53.09 | 4748.88 | **48.16** |
| | | 250 | 52.95 | 5933.89 | **47.70** | 176.77 | 12455.95 | **148.31** |
| | 0.9 | 50 | 4.80 | 774.13 | **3.45** | 6.37 | 788.19 | **4.89** |
| | | 150 | 14.96 | 2655.99 | **11.09** | 57.38 | 5380.94 | **55.70** |
| | | 250 | **34.27** | 4518.84 | 39.54 | 118.36 | 8531.93 | **97.23** |

Note: Emboldened entries represent the fastest time in each column.

*Biometrika*, **67**, 413–418.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Inc. New York.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

David, A. C. (2004). *Galois Theory*. John Wiley & Sons, Inc. New Jersey.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

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

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B. Stat. Methodl.*, **41**, 190–195.

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

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

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

Mori, Y. & Suzuki, T. (2018). Generalized ridge estimator and model selection criteria in multivariate linear regression. *J. Multivariate Anal.*, **165**, 243–261.

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

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

Ohishi, M., Yanagihara, H. & Fujikoshi, Y. (2020a). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. *J. Statist. Plann. Inference*, **204**, 187–205.

Ohishi, M., Yanagihara, H. & Kawano, S. (2020b). Equivalence between adaptive-lasso and generalized ridge estimators in linear regression with orthogonal explanatory variables after optimizing regularization parameters. *Ann. Inst. Statist. Math.*, (in press).

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **58**, 267–288.

Tignol, J.-P. (2001). *Galois' Theory of Algebraic Equations*. World Scientific Publishing. Singapore.

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

Xin, X., Hu, J. & Liu, L. (2017). On the oracle property of a generalized adaptive elastic-net for multivariate linear regression with a diverging number of parameters. *J. Multivariate Anal.*, **162**, 16–31.

Yanagihara, H. (2018). Explicit solution to the minimization problem of generalized cross-validation criterion for selecting ridge parameters in generalized ridge regression. *Hiroshima Math. J.*, **48**, 203–222.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418–1429.

## Appendix

### A.1. Proof of Theorem 1

Let $r(\boldsymbol{\delta}) = \mathrm{tr}(\boldsymbol{B}^*_{\boldsymbol{\delta}})$ and $u(\boldsymbol{\delta}) = \mathrm{df}(\boldsymbol{\delta})$. From Lemma 2, the domain of $f$ is included in $[0, r_+] \times [p, np)$. We define $\tau(\boldsymbol{\delta})$ as

$$\tau(\boldsymbol{\delta}) = \frac{nbp\dot{f}_u(r(\boldsymbol{\delta}), u(\boldsymbol{\delta}))}{2\dot{f}_r(r(\boldsymbol{\delta}), u(\boldsymbol{\delta}))}.$$

It is straightforward that $\tau(\boldsymbol{\delta}) > 0$ from $f$ satisfies Definition 4. Then, we have

$$\begin{aligned}\frac{\partial}{\partial \delta_j}\mathrm{MSC}(\boldsymbol{\delta}) &= \frac{\partial}{\partial \delta_j} r(\boldsymbol{\delta}) \cdot \frac{\partial}{\partial r} f(r,u)\bigg|_{(r,u)=(r(\boldsymbol{\delta}),u(\boldsymbol{\delta}))} + \frac{\partial}{\partial \delta_j} u(\boldsymbol{\delta}) \cdot \frac{\partial}{\partial u} f(r,u)\bigg|_{(r,u)=(r(\boldsymbol{\delta}),u(\boldsymbol{\delta}))} \\ &= \frac{2}{nb}\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j\delta_j\dot{f}_r(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) - p\dot{f}_u(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) \\ &= \frac{2}{nb}\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j\dot{f}_r(r(\boldsymbol{\delta}), u(\boldsymbol{\delta}))\left(\delta_j - \frac{\tau(\boldsymbol{\delta})}{\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j}\right),\end{aligned}$$

$$\frac{\partial}{\partial \delta_j}\mathrm{MSC}(\boldsymbol{\delta})\bigg|_{\boldsymbol{\delta}=\boldsymbol{0}_k} < 0.$$

Let $\boldsymbol{\delta}^\star = (\delta_1^\star, \dots, \delta_k^\star)'$ be the minimizer of $\mathrm{MSC}(\boldsymbol{\delta})$. Then, $\delta_j^\star \neq 0$ $(j = 1, \dots, k)$, and the necessary condition of $\delta_j^\star$ is given by

$$\delta_j^\star = \begin{cases} \dfrac{\tau(\boldsymbol{\delta}^\star)}{\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j} & (\tau(\boldsymbol{\delta}^\star) < \boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j) \\ 1 & (\tau(\boldsymbol{\delta}^\star) \geq \boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j) \end{cases} \quad (j = 1, \dots, k).$$

Let $\mathcal{G}$ be a set defined by

$$\mathcal{G} = \left\{ \boldsymbol{\delta} \in [0,1]^k \mid \boldsymbol{\delta} = \hat{\boldsymbol{\delta}}(h),\ \forall h \in \mathbb{R}_+ \backslash \{0\} \right\},$$

where $\hat{\boldsymbol{\delta}}(h)$ is a $k$-dimensional vector of which the $j$th element is given by

$$\hat{\delta}_j(h) = \begin{cases} \dfrac{h}{\boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j} & (h < \boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j) \\ 1 & (h \geq \boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j) \end{cases} \quad (j = 1, \dots, k).$$

Then, from $\boldsymbol{\delta}^\star$ is the minimizer of MSC($\boldsymbol{\delta}$), the following equation holds:

$$\mathrm{MSC}(\boldsymbol{\delta}^\star) = \min_{\boldsymbol{\delta} \in [0,1]^k \backslash \{\mathbf{0}_k\}} \mathrm{MSC}(\boldsymbol{\delta}) \leq \min_{\boldsymbol{\delta} \in \mathcal{G}} \mathrm{MSC}(\boldsymbol{\delta}) = \min_{h \in \mathbb{R}_+ \backslash \{0\}} \mathrm{MSC}(\hat{\boldsymbol{\delta}}(h)).$$

Whereas, because $\boldsymbol{\delta}^\star \in \mathcal{G}$ the following equation holds:

$$\mathrm{MSC}(\boldsymbol{\delta}^\star) \geq \min_{\boldsymbol{\delta} \in \mathcal{G}} \mathrm{MSC}(\boldsymbol{\delta}) = \min_{h \in \mathbb{R}_+ \backslash \{0\}} \mathrm{MSC}(\hat{\boldsymbol{\delta}}(h)).$$

These results lead to

$$\mathrm{MSC}(\boldsymbol{\delta}^\star) = \min_{h \in \mathbb{R}_+ \backslash \{0\}} \mathrm{MSC}(\hat{\boldsymbol{\delta}}(h)),$$

and hence, we have

$$\boldsymbol{\delta}^\star = \hat{\boldsymbol{\delta}}(\hat{h}), \quad \hat{h} = \arg \min_{h \in \mathbb{R}_+ \backslash \{0\}} \mathrm{MSC}(\hat{\boldsymbol{\delta}}(h)).$$

Consequently, Theorem 1 is proved.

### A.2. Proof of Lemma 4

To prove Lemma 4, it is sufficient to prove $\mathrm{df}(h_1) \geq \mathrm{df}(h_2)$. From Lemma 1, df($h$) is expressed as

$$\mathrm{df}(h) = p + p \sum_{j=1}^{k} \mathrm{soft}(1, h / \boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j).$$

Therefore, we have

$$\mathrm{df}(h_1) - \mathrm{df}(h_2) = p \sum_{j=1}^{k} \left\{ \mathrm{soft}(1, h_1 / \boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j) - \mathrm{soft}(1, h_2 / \boldsymbol{z}_j' \boldsymbol{S}^{-1} \boldsymbol{z}_j) \right\},$$

and regarding the RHS of the above equation, the following equation holds:

$$\mathrm{soft}(1, h_1/\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j) - \mathrm{soft}(1, h_2/\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j) = \begin{cases} 0 & (\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j \leq h_1) \\ 1 - \dfrac{h_1}{\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j} > 0 & (h_1 < \boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j \leq h_2) \\ \dfrac{h_2 - h_1}{\boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j} > 0 & (h_2 < \boldsymbol{z}_j'\boldsymbol{S}^{-1}\boldsymbol{z}_j) \end{cases}.$$

Hence, $\mathrm{df}(h_1) \geq \mathrm{df}(h_2)$ holds with quality only when $t_k \leq h_1$. Consequently, Lemma 4 is proved.

## A.3. Proof of Proposition 4

First, we prove (1) by reductio ad absurdum. Let $\hat{h}_{\alpha_1} = t_k$ and suppose that $\hat{h}_{\alpha_2} \neq t_k$. Then, the definition $\hat{h}_\alpha$ gives

$$\phi(\hat{h}_{\alpha_2} \mid \alpha_1) \geq \phi(t_k \mid \alpha_1), \quad \phi(t_k \mid \alpha_2) \geq \phi(\hat{h}_{\alpha_2} \mid \alpha_2),$$

and we have $\hat{h}_{\alpha_2} \neq t_k \Rightarrow \hat{h}_{\alpha_2} < t_k$ from (P2) in Proposition 1. Furthermore, $\phi(h \mid \alpha) = \eta(h \mid \alpha - \alpha_0)\phi(h \mid \alpha_0)$ holds from the definition of $\phi(h \mid \alpha)$. Therefore, from Lemma 4, we have

$$\phi(t_k \mid \alpha_2) = \eta(t_k \mid \alpha_2 - \alpha_1)\phi(t_k \mid \alpha_1) < \eta(\hat{h}_{\alpha_2} \mid \alpha_2 - \alpha_1)\phi(\hat{h}_{\alpha_2} \mid \alpha_1) = \phi(\hat{h}_{\alpha_2} \mid \alpha_2).$$

However, this contradicts $\phi(t_k \mid \alpha_2) \geq \phi(\hat{h}_{\alpha_2} \mid \alpha_2)$. Hence, (1) is proved.

Next, regarding (2), it is sufficient to prove $\hat{h}_{\alpha_1} < \hat{h}_{\alpha_2}$. We approach this via reductio ad absurdum again. Let $\alpha_1 < \alpha_2$ and suppose that $\hat{h}_{\alpha_2} \leq \hat{h}_{\alpha_1}$. Now, we have $\hat{h}_{\alpha_2} < t_k$ from $\hat{h}_{\alpha_2} \neq t_k$. Therefore,

$$\phi(\hat{h}_{\alpha_1} \mid \alpha_2) = \eta(\hat{h}_{\alpha_1} \mid \alpha_2 - \alpha_1)\phi(\hat{h}_{\alpha_1} \mid \alpha_1) < \eta(\hat{h}_{\alpha_2} \mid \alpha_2 - \alpha_1)\phi(\hat{h}_{\alpha_2} \mid \alpha_1) = \phi(\hat{h}_{\alpha_2} \mid \alpha_2).$$

However, this contradicts the definition of $\hat{h}_{\alpha_2}$. Hence, (2) is proved.

Consequently, Proposition 4 is proved.

## A.4. Proof of Proposition 5

First, we prove that the sequence $\{\delta_j^{(i)}\}$ $(i = 0, 1, \ldots)$ is a monotonically increasing sequence when $\delta_j^{(1)} \geq \delta_j^{(0)}$ $(j = 1, \ldots, k)$. Suppose that $\delta_j^{(i)} \geq \delta_j^{(i-1)}$ $(j = 1, \ldots, k)$. Then, $\delta_j^{(i)}$ is updated as

$$\delta_j^{(i+1)} = \zeta_j(\boldsymbol{\delta}^{(i)}) = 1 - \mathrm{soft}\left(1, \tau(\boldsymbol{\delta}^{(i)})/\boldsymbol{z}_j^{*\prime}\dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}^{(i)}}^*)\boldsymbol{z}_j^*\right),$$

and we have

$$\tau(\boldsymbol{\delta}^{(i)}) \geq \tau(\boldsymbol{\delta}^{(i-1)}), \quad \boldsymbol{z}_j^{*\prime}\dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}^{(i)}}^*)\boldsymbol{z}_j^* \leq \boldsymbol{z}_j^{*\prime}\dot{\boldsymbol{G}}(\boldsymbol{B}_{\boldsymbol{\delta}^{(i-1)}}^*)\boldsymbol{z}_j^*.$$

This gives $\delta_j^{(i+1)} \geq \delta_j^{(i)}$ for all $j = 1, \ldots, k$, and hence the sequence $\{\delta_j^{(i)}\}$ is a monotonically increasing sequence. Moreover, the sequence is bounded. Hence, the iterative method converges. In contrast, the sequence is bounded and monotonically decreasing when $\delta_j^{(1)} \leq \delta_j^{(0)}$ $(j = 1, \ldots, k)$, and hence, the iterative method converges. Consequently, Proposition 5 is proved.

### A.5. Proof of Theorem 3

Now, we have

$$\dot{f}_j(0) = -\frac{c_{j,0}}{\dot{f}_{j,1}(0)} < 0, \quad \dot{f}_j(\delta) = 0 \Longleftrightarrow \delta = \frac{1 \pm \sqrt{1 - c_{j,2}c_{j,0}/c_{j,1}^2}}{c_{j,2}/c_{j,1}}.$$

Therefore, $\hat{\delta}_j \neq 0$ and the smaller of the two real distinct roots or the double root of the quadratic equation $\dot{f}_{j,2}(\delta) = 0$ is the local minimizer. Notice that $\delta \in [0, 1]$. Then, to obtain the minimizer of $f_j(\delta)$, it is sufficient to confirm whether the local minimizer is included in $[0, 1]$ or not.

When $1 - c_{j,2}c_{j,0}/c_{j,1}^2 \geq 0$, there is one local minimizer, and let this be $\tilde{\delta}_j$, i.e.,

$$\tilde{\delta}_j = \frac{1 - \sqrt{1 - c_{j,2}c_{j,0}/c_{j,1}^2}}{c_{j,2}/c_{j,1}}.$$

This is positive and the following equation holds when $c_{j,2} > c_{j,1}$:

$$\tilde{\delta}_j < 1 - \sqrt{1 - c_{j,2}c_{j,0}/c_{j,1}^2} < 1.$$

Hence, we can obtain (1) in Theorem 3.

When $1 - c_{j,2}c_{j,0}/c_{j,1}^2 < 0$, there are no stationary points, and therefore $f_j(\delta)$ is a monotonically decreasing function. Hence, we can obtain (2) in Theorem 3.

Consequently, Theorem 3 is proved.

### A.6. Proof of Theorem 4

Now, we have

$$\dot{f}_j(0) = -\frac{c_{j,0}}{\dot{f}_{j,1}(0)} < 0, \quad \dot{f}_j(\delta) = 0 \Longleftrightarrow \dot{f}_{j,2}(\delta) = 0.$$

Thus $\hat{\delta}_j \neq 0$. Moreover, from $\delta \in [0, 1]$, minimizer candidates are local minimizers of $\dot{f}_{j,2}(\delta)$ included in $(0, 1)$ and the right end point of the range. Hence, we can obtain the set of minimizer candidates $\mathcal{S}_j$ by calculating stationary points of the cubic function $\dot{f}_{j,2}(\delta)$ and by confirming whether each stationary point is included in $(0, 1)$ or not. Consequently, Theorem 4 is proved.

### A.7. Proof of Theorem 6

To prove the equivalence between the two estimators, it is sufficient to prove $\hat{h}_{\boldsymbol{A}} = \hat{\lambda}_{\boldsymbol{A}}$. The two terms which constitute the MSC for optimizing ridge parameters are

$$\mathrm{tr}\left\{\hat{\Sigma}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(h \mid \boldsymbol{A}))\boldsymbol{A}^{-1}\right\} = b\,\mathrm{tr}\left(\hat{\Sigma}_0\boldsymbol{A}^{-1}\right) + \frac{1}{n}\sum_{j=1}^{k}\left\{1 - \mathrm{soft}\left(1, h/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right)\right\}^2 \boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j,$$

$$\mathrm{df}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(h \mid \boldsymbol{A})) = p + p\sum_{j=1}^{k}\mathrm{soft}\left(1, h/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right).$$

On the other hand, when $w_j = 1/\|\hat{\boldsymbol{\gamma}}_j\|$, from Lemma 5, the two terms which constitute the MSC for optimizing the tuning parameter are given by

$$\mathrm{tr}\left\{\hat{\Sigma}_{\mathrm{L}}(\lambda)\boldsymbol{A}^{-1}\right\} = b\,\mathrm{tr}\left(\hat{\Sigma}_0\boldsymbol{A}^{-1}\right) + \frac{1}{n}\sum_{j=1}^{k}\left\{1 - \mathrm{soft}\left(1, \lambda/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right)\right\}^2 \boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j,$$

$$\mathrm{df}_{\mathrm{L}}(\lambda) = p + p\sum_{j=1}^{k}\mathrm{soft}\left(1, \lambda/\boldsymbol{z}_j'\boldsymbol{A}^{-1}\boldsymbol{z}_j\right).$$

Hence, for all $x \in \mathbb{R}_+$, the following equation holds:

$$\mathrm{MSC}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(x \mid \boldsymbol{A}) \mid \boldsymbol{A}) = \mathrm{MSC}_{\mathrm{L}}(x \mid \boldsymbol{A}).$$

Thus $\hat{h}_{\boldsymbol{A}} = \hat{\lambda}_{\boldsymbol{A}}$ and consequently, Theorem 6 is proved.

### A.8. Proof of Theorem 7

From (6.3), the MGR estimator under the ridge parameters optimized by minimizing $\mathrm{MSC}_{\mathrm{R}}(\boldsymbol{\theta} \mid \boldsymbol{I}_p)$ is given by

$$\hat{\boldsymbol{\xi}}^{\mathrm{R}}_{\hat{\theta}_j, j} = \frac{1}{\sqrt{d_j}}\,\mathrm{soft}\left(1, \hat{h}_{\boldsymbol{I}_p}/\|\boldsymbol{z}_j\|^2\right)\boldsymbol{z}_j,$$

$$\hat{h}_{\boldsymbol{I}_p} = \arg\min_{h \in \mathbb{R}_+ \backslash \{0\}} \phi(h \mid \boldsymbol{I}_p), \quad \phi(h \mid \boldsymbol{I}_p) = \mathrm{MSC}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(h \mid \boldsymbol{I}_p) \mid \boldsymbol{I}_p).$$

Therefore, it is sufficient to prove $\hat{h}_{\boldsymbol{I}_p} = \hat{\lambda}_{\boldsymbol{I}_p}$. Similar to Appendix A.7, for all $x \in \mathbb{R}_+$, the following equations hold:

$$\mathrm{tr}\left\{\hat{\Sigma}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(x))\right\} = \mathrm{tr}\left\{\hat{\Sigma}_{\mathrm{L}}(x)\right\} = \mathrm{tr}(\hat{\Sigma}_0) + \frac{1}{n}\sum_{j=1}^{k}\left\{1 - \mathrm{soft}\left(1, x/\|\boldsymbol{z}_j\|^2\right)\right\}^2 \|\boldsymbol{z}_j\|^2,$$

$$\mathrm{df}_{\mathrm{R}}(\hat{\boldsymbol{\theta}}(x)) = \mathrm{df}_{\mathrm{L}}(x) = p + p\sum_{j=1}^{k}\mathrm{soft}\left(1, x/\|\boldsymbol{z}_j\|^2\right).$$

Hence, $\hat{h}_{\boldsymbol{I}_p} = \hat{\lambda}_{\boldsymbol{I}_p}$ holds and consequently, Theorem 7 is proved.