Modified C_p Criterion for Optimizing Ridge and Smooth Parameters in the MGR Estimator for the Nonparametric GMANOVA model

(Last Modified: March 18, 2011)

Isamu Nagai[1]

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

Abstract

Longitudinal trends of observations can be estimated using the generalized multivariate analysis of variance (GMANOVA) model proposed by Potthoff and Roy (1964). In the present paper, we consider estimating the trends nonparametrically using known basis functions. Then, as in non-parametric regression, an overfitting problem occurs. Satoh and Yanagihara (2010) showed that the GMANOVA model is equivalent to the varying coefficient model with non-longitudinal covariates. Hence, as in the case of the ordinary linear regression model, when the number of covariates becomes large, the estimator of the varying coefficient becomes unstable. In the present paper, we avoid the overfitting problem and the instability problem by applying the concept behind penalized smoothing spline regression and multivariate generalized ridge regression. In addition, we propose two criteria to optimize hyperparameters, namely, a smoothing parameter and ridge parameters. Finally, we compare the ordinary least square estimator and the new estimator.

Key words: Generalized ridge regression; Mallows' C_p statistic; GMANOVA; Non-iterative estimator; Shrinkage estimator; Varying coefficient model.

1. Introduction

We consider the generalized multivariate analysis of variance (GMANOVA) model with n observations of p-dimensional vectors of response variables. This model was proposed by Potthoff and Roy (1964). Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$, $\boldsymbol{A}$, $\boldsymbol{X}$ and $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ be an $n \times p$ matrix of response variables, an $n \times k$ matrix of non-stochastic centerized between-individual explanatory variables (i.e., $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}_k$) of $\mathrm{rank}(\boldsymbol{A}) = k$ $(k < n)$, a $p \times q$ matrix of non-stochastic within-individual explanatory variables of $\mathrm{rank}(\boldsymbol{X}) = q$ $(q \le p)$, and an $n \times p$ matrix of error variables, respectively, where n is the

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

sample size, $\mathbf{1}_n$ is an n-dimensional vector of ones and $\mathbf{0}_k$ is a k-dimensional vector of zeros. Then, the GMANOVA model is expressed as

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

where $\boldsymbol{\mu}$ is a q-dimensional unknown vector and $\Xi = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)'$ is a $k \times q$ unknown regression coefficient matrix. We assume that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n \overset{\text{i.i.d.}}{\sim} N_p(\mathbf{0}_p, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a $p \times p$ unknown covariance matrix of $\text{rank}(\boldsymbol{\Sigma}) = p$. Then, we can express the GMANOVA model as

$$\boldsymbol{Y} \sim N_{n \times p}(\mathbf{1}_n \boldsymbol{\mu}' \boldsymbol{X}' + \boldsymbol{A} \Xi \boldsymbol{X}', \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n).$$

Let $\boldsymbol{S}$ be an unbiased estimator of the unknown covariance matrix $\boldsymbol{\Sigma}$ that is given by

$$\boldsymbol{S} = \frac{1}{n-k-1} \boldsymbol{Y}' \{\boldsymbol{I}_n - n^{-1} \mathbf{1}_n \mathbf{1}_n' - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1} \boldsymbol{A}'\} \boldsymbol{Y}. \tag{1.2}$$

Then, the maximum likelihood (ML) estimators of $\boldsymbol{\mu}$ and Ξ are given by $n^{-1}(\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{Y}'\mathbf{1}_n$ and $(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{S}^{-1}\boldsymbol{X})^{-1}$, respectively. The ML estimators are the unbiased and asymptotically efficiency estimators of $\boldsymbol{\mu}$ and Ξ.

In the GMANOVA model, $\boldsymbol{x}(t) = (1, t, \ldots, t^{q-1})'$, $(t = t_1, \ldots, t_p)$ is often used as the jth row vector of $\boldsymbol{X}$. Then, we estimate the longitudinal trends of $\boldsymbol{Y}$ using $(q-1)$-polynomial curves. However, occasionally, the polynomial curve cannot thoroughly express flexible longitudinal trends. Hence, we consider estimating the longitudinal trends nonparametrically in the same manner as Riedel and Imre (1993) and Kshirsagar and Smith (1995), i.e., we use the known basis function as $\boldsymbol{x}(t)$ and assume that p is large. In the present paper, we refer to the GMANOVA model with $\boldsymbol{X}$ obtained from the basis function as the nonparametric GMANOVA model. In the nonparametric GMANOVA model, it is well known that the ML estimators become unstable because $\boldsymbol{S}^{-1}$ becomes unstable when p is large. Thus, we deal with the least square (LS) estimators of $\boldsymbol{\mu}$ and Ξ, which are obtained by minimizing $\text{tr}\{(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\Xi\boldsymbol{X}')(\boldsymbol{Y} - \mathbf{1}_n\boldsymbol{\mu}'\boldsymbol{X}' - \boldsymbol{A}\Xi\boldsymbol{X}')'\}$. Then, the LS estimators of $\boldsymbol{\mu}$ and Ξ are obtained by $\hat{\boldsymbol{\mu}} = n^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}'\mathbf{1}_n$ and

$$\hat{\Xi} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{1.3}$$

respectively. Note that $\hat{\boldsymbol{\mu}}$ does not depend on $\boldsymbol{A}$. The LS estimators are simple and unbiased estimators of $\boldsymbol{\mu}$ and Ξ. However, as well as ordinary nonparametric regression model, the LS estimators cause an overfitting problem when we use basis functions to estimate the longitudinal trends nonparametrically. In order to avoid the overfitting problem, we use $\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K}$ instead of $\boldsymbol{X}'\boldsymbol{X}$ as the penalized smoothing spline regression (see e.g., Green and Silverman (1994)), where $\lambda\ (\geq 0)$ is a smoothing parameter and $\boldsymbol{K}$ is a $q \times q$ known penalty matrix.

Let $\boldsymbol{y}_i = (y_i(t_1), \ldots, y_i(t_p))'$, and let $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_1), \ldots, \varepsilon_i(t_p))'$. Then, the GMANOVA model can be expressed as

$$y_i(t) = \boldsymbol{x}(t)'\boldsymbol{\mu} + \sum_{j=1}^{k} a_{ij}\boldsymbol{x}(t)'\boldsymbol{\xi}_j + \varepsilon_i(t), \ (i = 1, \ldots, n; \ t = t_1, \ldots, t_p),$$

where a_{ij} is the (i, j)th element of $\boldsymbol{A}$. This expression indicates that the GMANOVA model is equivalent to the varying coefficient model with non-longitudinal covariates (Satoh & Yanagihara, 2010), i.e.,

$$y_i(t) = \phi_0(t) + \sum_{j=1}^{k} a_{ij}\phi_j(t) + \varepsilon_i(t), \ (i = 1, \ldots, n; \ t = 1, \ldots, t_p), \tag{1.4}$$

where $\phi_0(t) = \boldsymbol{x}(t)'\boldsymbol{\mu}$ and $\phi_j(t) = \boldsymbol{x}(t)'\boldsymbol{\xi}_j$, $(j = 1, \ldots, k)$. Hence, estimating the longitudinal trends in the GMANOVA model nonparametrically is equivalent to estimating the varying coefficients $\phi_j(t)$, $(j = 0, \ldots, k)$ nonparametrically. However, when multicollinearity occurs in $\boldsymbol{A}$, the estimate of $\phi_j(t)$, $(j = 1, \ldots, k)$ becomes unstable, as does the ordinary LS estimator of regression coefficient, because the variance of an estimator of $\phi_j(t)$ becomes large. Hence, we avoid the multicollinearity problem in $\boldsymbol{A}$ by the ridge regression.

When $\boldsymbol{X} = \boldsymbol{I}_p$ and $p = 1$ in the model (1.1), Hoerl and Kennard (1970) proposed a ridge regression. This estimator is generally defined by adding $\theta\boldsymbol{I}_k$ to $\boldsymbol{A}'\boldsymbol{A}$ in (1.3), where $\theta \geq 0$ is referred to as a ridge parameter. Since the ridge estimator changes with θ, optimization of θ is very important. One method for optimizing θ is minimizing the C_p criterion proposed by Mallows (1973; 1995) in the univariate linear regression model (for multivariate case, see e.g., Sparks, Coutsourides and Troskie (1983)). For the case in which $\boldsymbol{X} = \boldsymbol{I}_p$ and $p \geq 1$, Yanagihara and Satoh (2010) proposed the C_p and its bias-corrected C_p (modified C_p; MC_p) criteria for optimizing the ridge parameter. However, an optimal θ cannot be obtained without an iterative computational algorithm because an optimal θ cannot be obtained in closed form.

On the other hand, Hoerl and Kennard (1970) also proposed a generalized ridge (GR) regression in the univariate linear regression model, i.e., the model (1.1) with $\boldsymbol{X} = \boldsymbol{I}_p$ and $p = 1$, simultaneously with the ridge regression. The GR estimator is defined not by a single ridge parameter, but rather by multiple ridge parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, $(\theta_i \geq 0, \ i = 1, \ldots, k)$. Then, several authors proposed a non-iterative GR estimator (see, e.g., Lawless (1981)). Yanagihara, Nagai and Satoh (2009) proposed a GR regression in the multivariate linear regression model, i.e., the model (1.1) with $\boldsymbol{X} = \boldsymbol{I}_p$ and $p \geq 1$. We call this generalized ridge regression the multivariate GR (MGR) regression. They also proposed the C_p and MC_p criteria for optimizing ridge parameters $\boldsymbol{\theta}$ in the MGR regression. They showed that the optimized $\boldsymbol{\theta}$ by minimizing two criteria are obtained in closed form. Nagai, Yanagihara and Satoh (2010) proposed non-iterative MGR estimators by extending non-iterative GR estimators. Several computational tasks are required in estimating $\phi_j(t)$ nonparametrically because we determine the optimal λ and the number of basis functions simultaneously. Fortunately, Yanagihara, Nagai and Satoh (2009) reported that the performance of the MGR regression is the almost same as that of the multivariate ridge regression. Hence, we use the MGR regression in order to avoid the multicollinearity problem that occurs in $\boldsymbol{A}$ in order to reduce the number of computational tasks.

The remainder of the present paper is organized as follows: In Section 2, we propose new estimators using the concept of the penalized smoothing spline regression and the MGR regression. In Section 3, we show the target mean squared error (MSE) of a predicted value of $\boldsymbol{Y}$. We then propose the C_p and MC_p criteria to optimize ridge parameters and smoothing parameters in the new estimator. Using these criteria, we show that the optimized ridge parameters are obtained in closed form under the fixed λ. We also show the magnitude relationship between the optimized ridge parameters. In Section 4, we compare the LS estimator in (1.3) with the proposed estimator through numerical studies. In Section 5, we present our conclusions.

2. The New Estimators

In the model (1.1), we consider estimating the longitudinal trends nonparametrically by using basis functions $\boldsymbol{X}$. Then, we consider the following estimators in order to avoid the overfitting problem in the nonparametric GMANOVA model, $\hat{\boldsymbol{\mu}}_\lambda = n^{-1}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}\boldsymbol{X}'\boldsymbol{Y}'\boldsymbol{1}_n$ and

$$\hat{\Xi}_\lambda = (\hat{\boldsymbol{\xi}}_{\lambda,1}, \ldots, \hat{\boldsymbol{\xi}}_{\lambda,k})' = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}, \tag{2.1}$$

where λ (≥ 0) is a smoothing parameter and $\boldsymbol{K}$ is a $q \times q$ known penalty matrix. In this estimator, we must determine $\boldsymbol{K}$ before using this estimator. Since $\boldsymbol{K}$ is usually set as some nonnegative definite matrix, we assume that $\boldsymbol{K}$ is a nonnegative definite matrix. If $\lambda\boldsymbol{K} = \boldsymbol{O}_{q,q}$, where $\boldsymbol{O}_{q,q}$ is a $q \times q$ matrix of zeros, then this estimator corresponds to the LS estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\Xi}$ in (1.3). Note that this estimator controls the smoothness of each estimated curve $\hat{\phi}_0(t) = \boldsymbol{x}(t)'\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\phi}_j(t) = \boldsymbol{x}(t)'\hat{\boldsymbol{\xi}}_{\lambda,j}$, $(j = 1, \ldots, k)$ through only one parameter λ. When we use this estimator, we need to optimize the parameter λ because this estimator changes with λ.

If multicollinearity occurs in $\boldsymbol{A}$, then the LS estimator $\hat{\Xi}$ in (1.3) and the proposed estimator $\hat{\Xi}_\lambda$ in (2.1) are not good estimators in the sense of having large variance. Note that neither the LS estimator $\hat{\boldsymbol{\mu}}$ nor the proposed estimator $\hat{\boldsymbol{\mu}}_\lambda$ depend on $\boldsymbol{A}$. Hence, we avoid the multicollinearity problem for estimating Ξ. Multicollinearity often occurs when k becomes large. Using the following estimator, the multicollinearity problem in $\boldsymbol{A}$ can be avoided,

$$\hat{\Xi}_{\theta,\lambda} = (\hat{\boldsymbol{\xi}}_{\theta\lambda,1}, \ldots, \hat{\boldsymbol{\xi}}_{\theta\lambda,k})' = (\boldsymbol{A}'\boldsymbol{A} + \theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}, \tag{2.2}$$

where $\theta \geq 0$ is a ridge parameter. This estimator with $\boldsymbol{K} = \boldsymbol{I}_q$ corresponds to the estimator of Takane, Jung and Hwang (2011). If $\theta = 0$, then this estimator corresponds to the estimator in

(2.1). Note that $(\boldsymbol{A}'\boldsymbol{A}+\theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'\boldsymbol{Y}$ in this estimator corresponds to the ridge estimator for a multivariate linear model (Yanagihara & Satoh, 2010). In this estimator, we need to optimize θ and λ because this estimator changes with these parameters. However, we cannot obtain the optimized θ and λ in closed form. Thus, we need to use an iterative computational algorithm to optimize two parameters. From another point of view, this estimator controls the smoothness of each estimated curve $\hat{\phi}_j(t) = \boldsymbol{x}(t)'\hat{\boldsymbol{\xi}}_{\theta\lambda,j}$, $(j = 1, \ldots, k)$ through only one parameter λ. Hence, this estimator is not a well fitting curve when the smoothnesses of the true curves differ.

Hence, we apply the concept of the MGR estimator (Yanagihara, Nagai & Satoh, 2009) to $(\boldsymbol{A}'\boldsymbol{A}+\theta\boldsymbol{I}_k)^{-1}\boldsymbol{A}'\boldsymbol{Y}$ in order to obtain the optimized ridge parameter in closed form. Here, we derive the MGR estimator for the nonparametric GMANOVA model as follows:

$$\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda} = (\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}, \tag{2.3}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$, $(\theta_i \geq 0,\ i = 1, \ldots, k)$ is also a ridge parameter, $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta})$, and $\boldsymbol{Q}$ is the $k \times k$ orthogonal matrix that diagonalizes $\boldsymbol{A}'\boldsymbol{A}$, i.e., $\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{D}$ where $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_k)$ and $d_1, \ldots, d_k$ are eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$. It is clearly that $d_i > 0$, $(i = 1, \ldots, k)$. In this estimator, since $\boldsymbol{\theta}$ shrinks the estimators of $\phi_j(t)$, $(j = 1, \ldots, k)$ to 0, we can regard $\boldsymbol{\theta}$ as controlling the smoothness of $\phi_1(t), \ldots, \phi_k(t)$. Therefore, in this estimator, rough smoothness of the estimated curves is controlled by λ, and each smoothness of $\phi_1(t), \ldots, \phi_k(t)$ is controlled by $\boldsymbol{\theta}$.

Clearly, $\hat{\boldsymbol{\Xi}}_{\boldsymbol{0}_k,0} = \hat{\boldsymbol{\Xi}}$ and $\hat{\boldsymbol{\Xi}}_{\boldsymbol{0}_k,\lambda} = \hat{\boldsymbol{\Xi}}_\lambda$. The $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ with $\boldsymbol{\theta} = \boldsymbol{1}_k\theta$ for some θ (≥ 0) corresponds to $\hat{\boldsymbol{\Xi}}_{\theta,\lambda}$ in (2.2). Thus, the estimator $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ includes these estimators. The estimator $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ is more flexible than these estimators $\hat{\boldsymbol{\Xi}}_\lambda$ and $\hat{\boldsymbol{\Xi}}_{\theta,\lambda}$ because $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ has $k+1$ parameters and $\hat{\boldsymbol{\Xi}}_\lambda$ or $\hat{\boldsymbol{\Xi}}_{\theta,\lambda}$ has only one or two parameters. Hence, we consider $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ in estimating the longitudinal trends or the varying coefficient curve, while avoiding the overfitting and multicollinearity problems in the nonparametric GMANOVA model. When $\boldsymbol{X} = \boldsymbol{I}_p$ and $\lambda\boldsymbol{K} = \boldsymbol{O}_{q,q}$, $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ corresponds to the MGR estimator in Yanagihara, Nagai and Satoh (2009).

3. Main Results

3.1. Target MSE

In order to define the MSE of the predicted value of $\boldsymbol{Y}$, we prepare the following discrepancy function for measuring the distance between $n \times p$ matrices $\boldsymbol{E}$ and $\boldsymbol{F}$:

$$r(\boldsymbol{E}, \boldsymbol{F}) = \mathrm{tr}\{(\boldsymbol{E} - \boldsymbol{F})\boldsymbol{\Sigma}^{-1}(\boldsymbol{E} - \boldsymbol{F})'\}.$$

Since $\boldsymbol{\Sigma}$ is an unknown covariance matrix, we use the unbiased estimator $\boldsymbol{S}$ in (1.2) instead of $\boldsymbol{\Sigma}$ to estimate $r(\boldsymbol{E}, \boldsymbol{F})$. Hence, we estimate $r(\boldsymbol{E}, \boldsymbol{F})$ using the following sample discrepancy function:

$$\hat{r}(\boldsymbol{E}, \boldsymbol{F}) = \mathrm{tr}\{(\boldsymbol{E} - \boldsymbol{F})\boldsymbol{S}^{-1}(\boldsymbol{E} - \boldsymbol{F})'\}. \tag{3.1}$$

These two functions, $r(\boldsymbol{E}, \boldsymbol{F})$ and $\hat{r}(\boldsymbol{E}, \boldsymbol{F})$, correspond to the summation of the Mahalanobis distance and the sample Mahalanobis distance between the rows of $\boldsymbol{E}$ and $\boldsymbol{F}$, respectively. Clearly, $r(\boldsymbol{E}, \boldsymbol{F}) = r(\boldsymbol{F}, \boldsymbol{E})$ and $\hat{r}(\boldsymbol{E}, \boldsymbol{F}) = \hat{r}(\boldsymbol{F}, \boldsymbol{E})$. Through simple calculation, we obtain the following properties:

$$r(\boldsymbol{E}_1 + \boldsymbol{E}_2, \boldsymbol{E}_3) = r(\boldsymbol{E}_1, \boldsymbol{E}_3) + 2\mathrm{tr}\{(\boldsymbol{E}_1 - \boldsymbol{E}_3)\boldsymbol{\Sigma}^{-1}\boldsymbol{E}_2\} + r(\boldsymbol{E}_2, \boldsymbol{O}_{n,p}),$$
$$\hat{r}(\boldsymbol{E}_1 + \boldsymbol{E}_2, \boldsymbol{E}_3) = \hat{r}(\boldsymbol{E}_1, \boldsymbol{E}_3) + 2\mathrm{tr}\{(\boldsymbol{E}_1 - \boldsymbol{E}_3)\boldsymbol{S}^{-1}\boldsymbol{E}_2\} + \hat{r}(\boldsymbol{E}_2, \boldsymbol{O}_{n,p}),$$

for any $n \times p$ matrices $\boldsymbol{E}_1$, $\boldsymbol{E}_2$ and $\boldsymbol{E}_3$.

Using the discrepancy function r, the MSE of the predicted value of $\boldsymbol{Y}$ is defined as

$$\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}] = E[r(E[\boldsymbol{Y}], \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})], \tag{3.2}$$

where $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}_\lambda'\boldsymbol{X}' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}\boldsymbol{X}'$, which is the predicted value of $\boldsymbol{Y}$ when we use $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ in (2.3). In the present paper, we regard $\boldsymbol{\theta}$ and λ making the MSE the smallest as the principle optimum. However, we cannot use the MSE in (3.2) in actual application because this MSE includes unknown parameters. Hence, we must estimate (3.2) in order to estimate the optimum $\boldsymbol{\theta}$ and λ.

3.2. The C_p and MC_p criteria

Let $\boldsymbol{H}_{\boldsymbol{\theta}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{Q}\Theta\boldsymbol{Q}')^{-1}\boldsymbol{A}'$ and $\boldsymbol{G}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}\boldsymbol{X}'$. Note that $\boldsymbol{Y} = E[\boldsymbol{Y}] + \boldsymbol{\mathcal{E}}$. Hence, we obtain

$$\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}] = E[r(\boldsymbol{Y} - \boldsymbol{\mathcal{E}}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]$$

From the properties of the function r and using $E[\mathrm{tr}(\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] = E[\mathrm{tr}(\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')]$, since $E[\boldsymbol{Y}]$ is a non-stochastic variable and $E[\boldsymbol{\mathcal{E}}] = \boldsymbol{O}_{n,p}$ and $E[\mathrm{tr}(\boldsymbol{E}_4)] = \mathrm{tr}(E[\boldsymbol{E}_4])$ for any square matrix $\boldsymbol{E}_4$, we obtain

$$\begin{aligned}
\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}] &= E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})] - 2E[\mathrm{tr}\{(\boldsymbol{Y} - \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}'\}] + E[r(\boldsymbol{\mathcal{E}}, \boldsymbol{O}_{n,p})] \\
&= E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})] - 2E[\mathrm{tr}(\boldsymbol{Y}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] + 2E[\mathrm{tr}(\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] + E[\mathrm{tr}(\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] \\
&= E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})] - E[\mathrm{tr}(\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] + 2E[\mathrm{tr}(\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')].
\end{aligned}$$

Note that $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}_\lambda'\boldsymbol{X}' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}\boldsymbol{X}' = (n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{H}_{\boldsymbol{\theta}})\boldsymbol{Y}\boldsymbol{G}_\lambda = (n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{H}_{\boldsymbol{\theta}})(E[\boldsymbol{Y}] + \boldsymbol{\mathcal{E}})\boldsymbol{G}_\lambda$. Thus, we can calculate $E[\mathrm{tr}(\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')]$ as follows:

$$\begin{aligned}
E[\mathrm{tr}(\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] &= E[\mathrm{tr}\{(n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{H}_{\boldsymbol{\theta}})(E[\boldsymbol{Y}] + \boldsymbol{\mathcal{E}})\boldsymbol{G}_\lambda\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}'\}] \\
&= \mathrm{tr}\{(n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{H}_{\boldsymbol{\theta}})E[\boldsymbol{\mathcal{E}}\boldsymbol{G}_\lambda\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}']\},
\end{aligned}$$

because $E[\boldsymbol{Y}]$, $\boldsymbol{G}_\lambda$ and $\boldsymbol{\Sigma}^{-1}$ are non-stochastic variables. For calculating the expectations in the MSE, we prove the following lemma.

Lemma 3.1. For any $p \times p$ non-stochastic matrix $\boldsymbol{J}$, we obtain $E[\boldsymbol{\mathcal{E}}\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}'] = \mathrm{tr}(\boldsymbol{J})\boldsymbol{I}_n$.

Proof. Since $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$, we obtain the (i,j)th element of $E[\boldsymbol{\mathcal{E}}\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}']$ as $E[\boldsymbol{\varepsilon}_i'\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_j]$, $(i = 1, \ldots, n;\ j = 1, \ldots, n)$. We obtain $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'] = \delta_{i,j}\boldsymbol{\Sigma}$ because $\boldsymbol{\varepsilon}_i \perp\!\!\!\perp \boldsymbol{\varepsilon}_j$ for any $i \neq j$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}$ for any i, where $\delta_{i,j}$ is defined as $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ if $i \neq j$. Hence, we obtain $E[\boldsymbol{\varepsilon}_i'\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_j] = E[\mathrm{tr}(\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_i')] = \mathrm{tr}(\delta_{j,i}\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) = \delta_{j,i}\mathrm{tr}(\boldsymbol{J})$. This result means that $E[\boldsymbol{\varepsilon}_i'\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_j] = \mathrm{tr}(\boldsymbol{J})$ if $i = j$ and $E[\boldsymbol{\varepsilon}_i'\boldsymbol{J}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_j] = 0$ if $i \neq j$. Thus, the lemma is proven. □

Using this lemma, we obtain $E[\mathrm{tr}(\boldsymbol{\mathcal{E}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}')] = \mathrm{tr}(\mathrm{tr}(\boldsymbol{I}_p)\boldsymbol{I}_n) = np$ and $E[\boldsymbol{\mathcal{E}}\boldsymbol{G}_\lambda\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}'] = \mathrm{tr}(\boldsymbol{G}_\lambda)\boldsymbol{I}_n$. Hence, we obtain

$$\begin{aligned}\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}] &= E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})] - np + 2\mathrm{tr}\{(n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{H}_{\boldsymbol{\theta}})\mathrm{tr}(\boldsymbol{G}_\lambda)\boldsymbol{I}_n\} \\ &= E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})] - np + 2\mathrm{tr}(\boldsymbol{G}_\lambda)\mathrm{tr}(n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{H}_{\boldsymbol{\theta}}) \\ &= E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})] - np + 2\mathrm{tr}(\boldsymbol{G}_\lambda)\{1 + \mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}})\}.\end{aligned}$$

By replacing $E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]$ with $\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})$, we can propose the instinctive estimator of MSE, referred to as the C_p criterion, as follows:

$$C_p(\boldsymbol{\theta}, \lambda) = \hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}) - np + 2\mathrm{tr}(\boldsymbol{G}_\lambda)\{\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}}) + 1\}. \tag{3.3}$$

When we use this criterion, we optimize the ridge parameter $\boldsymbol{\theta}$ and the smoothing parameter λ by the following algorithm:

(i) We obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda) = \arg\min_{\boldsymbol{\theta}} C_p(\boldsymbol{\theta}, \lambda)$, where $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda) = (\hat{\theta}_1^{(\mathrm{C})}(\lambda), \ldots, \hat{\theta}_k^{(\mathrm{C})}(\lambda))'$, $(\hat{\theta}_i^{(\mathrm{C})}(\lambda) \geq 0,\ i = 1, \ldots, k)$ if λ is given.

(ii) We obtain $\hat{\lambda}^{(\mathrm{C})} = \arg\min_{\lambda \geq 0} C_p(\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda), \lambda)$, where $\hat{\lambda}^{(\mathrm{C})} \geq 0$.

(iii) We can obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) = \arg\min_{\boldsymbol{\theta}} C_p(\boldsymbol{\theta}, \hat{\lambda}^{(\mathrm{C})})$, where $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) = (\hat{\theta}_1^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}), \ldots, \hat{\theta}_k^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}))'$, $(\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) \geq 0,\ i = 1, \ldots, k)$ under fixed $\hat{\lambda}^{(\mathrm{C})}$.

(iv) We optimize the ridge parameter and the smoothing parameter as $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\lambda}^{(\mathrm{C})}$, respectively.

Note that this C_p criterion corresponds to that in Yanagihara, Nagai and Satoh (2009) when $\boldsymbol{X} = \boldsymbol{I}_p$ and $\lambda\boldsymbol{K} = \boldsymbol{O}_{q,q}$.

There is some bias between the MSE in (3.2) and the C_p criterion in (3.3) because the C_p criterion is obtained by replacing $E[r(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]$ in the MSE with $\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})$. Generally, when the sample size n is small or the number of explanatory variables k is large, this bias becomes large. Then, we cannot obtain the higher-accuracy estimation of the optimum parameters because we cannot obtain the higher-accuracy estimation of MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}$ in (3.2). Hence, we correct the bias between $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}]$

and the C_p criterion. To correct the bias, we assume $n-k-p-2>0$. Let $\boldsymbol{W}_{\boldsymbol{\theta},\lambda}=(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})'(\boldsymbol{Y}-\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})$ and $\boldsymbol{W}=(n-k-1)\boldsymbol{S}$. Note that $\boldsymbol{W}\sim W_p(n-k-1,\boldsymbol{\Sigma})$ and $\boldsymbol{W}_{\boldsymbol{\theta},\lambda}-\boldsymbol{W}\perp\!\!\!\perp \boldsymbol{W}^{-1}$ because $\boldsymbol{A}'\boldsymbol{1}_n=\boldsymbol{0}_k$ and $\boldsymbol{A}'\{\boldsymbol{I}_n-\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\}=\boldsymbol{O}_{k,n}$. Then, we obtain

$$E[\hat{r}(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]=(n-k-1)E[\mathrm{tr}\{(\boldsymbol{W}_{\boldsymbol{\theta},\lambda}-\boldsymbol{W})\boldsymbol{W}^{-1}+\boldsymbol{I}_p\}].$$

Since $E[\boldsymbol{W}^{-1}]=\boldsymbol{\Sigma}^{-1}/(n-k-p-2)$, $E[\boldsymbol{W}]=(n-k-1)\boldsymbol{\Sigma}$ (see, e.g., Siotani, Hayakawa and Fujikoshi (1985)) and $\mathrm{tr}\{E[\boldsymbol{W}_{\boldsymbol{\theta},\lambda}\boldsymbol{\Sigma}^{-1}]\}=E[r(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]$, we obtain

$$\begin{aligned}E[\hat{r}(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]&=\frac{n-k-1}{n-k-p-2}\mathrm{tr}\{E[\boldsymbol{W}_{\boldsymbol{\theta},\lambda}-\boldsymbol{W}]\boldsymbol{\Sigma}^{-1}+(n-k-p-2)\boldsymbol{I}_p\}\\&=\frac{n-k-1}{n-k-p-2}\{E[r(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]-(p+1)p\}.\end{aligned}$$

Therefore, we obtain the unbiased estimator for $E[r(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})]$ as $c_{\mathrm{M}}\hat{r}(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})+p(p+1)$, where $c_{\mathrm{M}}=1-(p+1)/(n-k-1)$. This implies that the bias corrected C_p criterion, denoted as MC_p (modified C_p) criterion, is obtained by

$$MC_p(\boldsymbol{\theta},\lambda)=c_{\mathrm{M}}\hat{r}(\boldsymbol{Y},\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})+p(p+1-n)+2\mathrm{tr}(\boldsymbol{G}_\lambda)\{\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}})+1\}. \tag{3.4}$$

As in the case of using the C_p, we optimize $\boldsymbol{\theta}$ and λ using this criterion as follows:

(i) We obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)=\arg\min_{\boldsymbol{\theta}} MC_p(\boldsymbol{\theta},\lambda)$, where $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)=(\hat{\theta}_1^{(\mathrm{M})}(\lambda),\ldots,\hat{\theta}_k^{(\mathrm{M})}(\lambda))'$, $(\hat{\theta}_i^{(\mathrm{M})}(\lambda)\geq 0,\ i=1,\ldots,k)$ if λ is given.

(ii) We obtain $\hat{\lambda}^{(\mathrm{M})}=\arg\min_{\lambda\geq 0} MC_p(\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda),\lambda)$.

(iii) We obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})=\arg\min_{\boldsymbol{\theta}} MC_p(\boldsymbol{\theta},\hat{\lambda}^{(\mathrm{M})})$, where $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})=(\hat{\theta}_1^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}),\ldots,\hat{\theta}_k^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}))'$, $(\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})\geq 0,\ i=1,\ldots,k)$ under fixed $\hat{\lambda}^{(\mathrm{M})}$.

(iv) We optimize the ridge parameter and the smoothing parameter as $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ and $\hat{\lambda}^{(\mathrm{M})}$, respectively.

Note that the MC_p criterion corresponds to that in Yanagihara, Nagai and Satoh (2009) when $\boldsymbol{X}=\boldsymbol{I}_p$ and $\lambda\boldsymbol{K}=\boldsymbol{O}_{q,q}$. The MC_p criterion completely omits the bias between the MSE of $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}$ in (3.2) and the C_p criterion in (3.3) by using a number of constant terms c_{M} and $p(p+1)$. If $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda)$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)$ can be expressed in closed form for any fixed $\lambda\geq 0$, we do not need the above iterative computational algorithm.

3.3. Optimizations using the C_p and MC_p criteria

Using the generalized C_p (GC_p) criterion, which is given in (A.1), we can express the C_p and MC_p criteria as follows:

$$\begin{aligned}C_p(\boldsymbol{\theta},\lambda)&=GC_p(\boldsymbol{\theta},\lambda|1)-np+2\mathrm{tr}(\boldsymbol{G}_\lambda),\\MC_p(\boldsymbol{\theta},\lambda)&=GC_p(\boldsymbol{\theta},\lambda|c_{\mathrm{M}})+p(p+1-n)+2\mathrm{tr}(\boldsymbol{G}_\lambda).\end{aligned}$$

Note that the terms with respect to $\boldsymbol{\theta}$ in the C_p and MC_p criteria correspond to $GC_p(\boldsymbol{\theta}, \lambda|1)$ and $GC_p(\boldsymbol{\theta}, \lambda|c_{\mathrm{M}})$, respectively. Hence, we consider obtaining the optimum $\boldsymbol{\theta}$ by minimizing the GC_p criterion. From Theorem A.1, the optimum $\boldsymbol{\theta}$ is obtained in closed form as (A.2). Using the closed form in (A.2), we obtain $\hat{\theta}_i^{(\mathrm{C})}(\lambda)$ and $\hat{\theta}_i^{(\mathrm{M})}(\lambda)$ for each $i = 1, \ldots, k$ and any fixed $\lambda \geq 0$ as follows:

$$\hat{\theta}_i^{(\mathrm{C})}(\lambda) = \hat{\theta}_i^{(\mathrm{G})}(\lambda|1) = \begin{cases} 0 & (t_i^{(\mathrm{C})}(\lambda) \geq 0) \\ \dfrac{-d_i t_i^{(\mathrm{C})}(\lambda)}{t_i^{(\mathrm{C})}(\lambda) + u_{ii}} & (t_i^{(\mathrm{C})}(\lambda) < 0 \; and \; 0 < t_i^{(\mathrm{C})}(\lambda) + u_{ii}) \\ \infty & (otherwise) \end{cases}, \tag{3.5}$$

$$\hat{\theta}_i^{(\mathrm{M})}(\lambda) = \hat{\theta}_i^{(\mathrm{G})}(\lambda|c_{\mathrm{M}}) = \begin{cases} 0 & (t_i^{(\mathrm{M})}(\lambda) \geq 0) \\ \dfrac{-d_i t_i^{(\mathrm{M})}(\lambda)}{t_i^{(\mathrm{M})}(\lambda) + u_{ii}} & (t_i^{(\mathrm{M})}(\lambda) < 0 \; and \; 0 < t_i^{(\mathrm{M})}(\lambda) + c_{\mathrm{M}} u_{ii}) \\ \infty & (otherwise) \end{cases}, \tag{3.6}$$

where u_{ii} and v_{ii} are the (i, i)th elements of $\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{G}_\lambda \boldsymbol{S}^{-1}\boldsymbol{G}_\lambda \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Q}$ and $\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{G}_\lambda \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Q}$, respectively, $t_i^{(\mathrm{C})}(\lambda) = t_i(\lambda|1) = u_{ii} - v_{ii} - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)$ and $t_i^{(\mathrm{M})}(\lambda) = t_i(\lambda|c_{\mathrm{M}}) = c_{\mathrm{M}}(u_{ii} - v_{ii}) - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)$. Note that u_{ii} and v_{ii} vary with λ. Since $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda)$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)$ are regarded as a function of λ, we can regard the C_p and MC_p criteria for optimizing $\boldsymbol{\theta}$ and λ in (3.3) and (3.4) as a function of λ. This means that we can use these criteria to optimize λ.

Then, we can rewrite the optimization algorithms to optimize the ridge parameter $\boldsymbol{\theta}$ and the smoothing parameter λ by minimizing the C_p and MC_p criteria in (3.3) and (3.4) as follows:

(i) We obtain $\hat{\lambda}^{(\mathrm{C})} = \underset{\lambda \geq 0}{\mathrm{argmin}} C_p(\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda), \lambda)$ and $\hat{\lambda}^{(\mathrm{M})} = \underset{\lambda \geq 0}{\mathrm{argmin}} MC_p(\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda), \lambda)$.

(ii) We optimize the ridge parameter and the smoothing parameter as $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$, respectively, by using $\hat{\lambda}^{(\mathrm{C})}$, $\hat{\lambda}^{(\mathrm{M})}$ and the closed forms in (3.5) and (3.6).

This means that we can reduce the processing time to optimize the parameters, and we need to use the optimization algorithm for only one parameter, λ, for any k.

3.4. Magnitude relationships between optimized ridge parameters

In this subsection, we prove the magnitude relationships between $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$, $(i = 1, \ldots, k)$.

Lemma 3.2. For any $\lambda \geq 0$, we obtain $\mathrm{tr}(\boldsymbol{G}_\lambda) \geq 0$.

Proof. Since we assume $\boldsymbol{K}$ as a nonnegative definite matrix, there exists $\boldsymbol{L}$ that satisfies $\boldsymbol{K} = \boldsymbol{L}'\boldsymbol{L}$ (see, e.g., Harville (1997)). Then, since $\lambda \geq 0$, we have $\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K} = (\boldsymbol{X}', \sqrt{\lambda}\boldsymbol{L}')(\boldsymbol{X}', \sqrt{\lambda}\boldsymbol{L}')'$. Hence, $\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K}$ is a nonnegative definite matrix. This means that all of the eigenvalues of $\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K}$ are nonnegative. Hence, all of the eigenvalues of $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}$ are nonnegative. Thus, $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}$ is also a nonnegative definite matrix for any $\lambda \geq 0$. Since $\boldsymbol{G}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{K})^{-1}\boldsymbol{X}'$, we obtain $\boldsymbol{G}_\lambda$ as a nonnegative definite matrix for any $\lambda \geq 0$. Thus, the lemma is proven. □

Using the same idea, we have $\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}}) \geq 0$ for any $\boldsymbol{\theta}$ ($\theta_i \geq 0,\ i = 1, \ldots, k$). Therefore, the final terms of the C_p and MC_p criteria in (3.3) and (3.4) are always greater than $\mathrm{tr}(\boldsymbol{G}_\lambda) \geq 0$. In order to prove the magnitude relationship between $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$, we consider two situations in which $\hat{\lambda}^{(\mathrm{C})} = \hat{\lambda}^{(\mathrm{M})}$ is satisfied and $\hat{\lambda}^{(\mathrm{C})} \neq \hat{\lambda}^{(\mathrm{M})}$ is satisfied.

First, we consider $\hat{\lambda}^{(\mathrm{C})} = \hat{\lambda}^{(\mathrm{M})}$ to be satisfied. Let $\hat{\lambda} = \hat{\lambda}^{(\mathrm{C})} = \hat{\lambda}^{(\mathrm{M})}$ ($\hat{\lambda} \geq 0$). Using $\hat{\lambda}$, we obtain the following corollary.

Corollary 3.1. For any $\hat{\lambda} \geq 0$, we obtain $c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}) \geq t_i^{(\mathrm{M})}(\hat{\lambda})$.

Proof. Through simple calculation, we obtain

$$c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}) - t_i^{(\mathrm{M})}(\hat{\lambda}) = d_i \mathrm{tr}(\boldsymbol{G}_{\hat{\lambda}})(1 - c_{\mathrm{M}}).$$

Since $d_i > 0$, $0 < c_{\mathrm{M}} < 1$ and $\mathrm{tr}(\boldsymbol{G}_{\hat{\lambda}}) \geq 0$ from lemma 3.2, the corollary is proven. □

This corollary indicates that $t_i^{(\mathrm{C})}(\hat{\lambda}) \geq 0$ is satisfied when $t_i^{(\mathrm{M})}(\hat{\lambda}) \geq 0$ is satisfied because $c_{\mathrm{M}} > 0$ and $t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii} > 0$ is satisfied when $t_i^{(\mathrm{M})}(\hat{\lambda}) + c_{\mathrm{M}} u_{ii} > 0$ is satisfied because $c_{\mathrm{M}}\{t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii}\} > t_i^{(\mathrm{M})}(\hat{\lambda}) + c_{\mathrm{M}} u_{ii} > 0$ and $c_{\mathrm{M}} > 0$. Using these relationships, we obtain the following theorem.

Theorem 3.1. For any $\hat{\lambda} \geq 0$, we obtain $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}) \geq \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda})$.

Proof. We consider the following situations:

(1) $t_i^{(\mathrm{M})}(\hat{\lambda}) \geq 0$ is satisfied,

(2) $t_i^{(\mathrm{M})}(\hat{\lambda}) < 0 < t_i^{(\mathrm{M})}(\hat{\lambda}) + c_{\mathrm{M}} u_{ii}$ is satisfied,

(3) $t_i^{(\mathrm{M})}(\hat{\lambda}) + c_{\mathrm{M}} u_{ii} < 0$ is satisfied.

In (1), $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}) = \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}) = 0$, because $t_i^{(\mathrm{C})}(\hat{\lambda}) \geq 0$. In (3), $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}) \geq \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda})$, because $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda})$ becomes ∞. Hence, we only consider situation (2). Note that $t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii} > 0$, because $c_{\mathrm{M}}\{t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii}\} > 0$ and $c_{\mathrm{M}} > 0$. This means that $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda})$ does not become ∞. This theorem holds when $t_i^{(\mathrm{C})}(\hat{\lambda}) \geq 0$, because, in this case, $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}) = 0$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}) \geq 0$. We also consider $t_i^{(\mathrm{C})}(\hat{\lambda}) < 0 < t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii}$ to be satisfied. Then, we obtain

$$\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}) - \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}) = \frac{d_i u_{ii}\{c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}) - t_i^{(\mathrm{M})}(\hat{\lambda})\}}{\{t_i^{(\mathrm{M})}(\hat{\lambda}) + c_{\mathrm{M}} u_{ii}\}\{t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii}\}}.$$

Since $\boldsymbol{S}^{-1}$ is a positive definite matrix, $u_{ii} \geq 0$ for any $\hat{\lambda} \geq 0$. From corollary 3.1, we have $c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}) \geq t_i^{(\mathrm{M})}(\hat{\lambda})$ for any $\hat{\lambda} \geq 0$. Hence we obtain $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}) \geq \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda})$ for any $\hat{\lambda} \geq 0$ since $d_i > 0$, $t_i^{(\mathrm{M})}(\hat{\lambda}) + c_{\mathrm{M}} u_{ii} > 0$ and $t_i^{(\mathrm{C})}(\hat{\lambda}) + u_{ii} > 0$. Thus, the theorem is proven. □

This theorem corresponds to that in Nagai, Yanagihara and Satoh (2010) when $\boldsymbol{X} = \boldsymbol{I}_p$ and $\lambda \boldsymbol{K} = \boldsymbol{O}_{q,q}$.

From Theorem 3.1, we obtained the relationships between $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ for the case in which the optimized smoothing parameters $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ are the same. However, $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$

are optimized by minimizing the C_p and MC_p criteria in (3.3) and (3.4). Hence, $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ are generally different. Thus, we consider the relationship between $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ when $\hat{\lambda}^{(\mathrm{C})} \neq \hat{\lambda}^{(\mathrm{M})}$. Since u_{ii} is regarded as a function of λ, we write u_{ii} as $u_{ii}(\hat{\lambda}^{(\mathrm{C})})$ and $u_{ii}(\hat{\lambda}^{(\mathrm{M})})$ for each optimized smoothing parameter.

Theorem 3.2. We consider the following situations:

(1) $t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) + u_{ii}(\hat{\lambda}^{(\mathrm{C})}) \leq 0$ or $t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) \geq 0$ is satisfied,

(2) $t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) < 0 < t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) + u_{ii}(\hat{\lambda}^{(\mathrm{C})})$ and $t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) < 0 < t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) + c_{\mathrm{M}} u_{ii}(\hat{\lambda}^{(\mathrm{M})})$ are satisfied,

(3) $c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) u_{ii}(\hat{\lambda}^{(\mathrm{M})}) \leq t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) u_{ii}(\hat{\lambda}^{(\mathrm{C})})$ is satisfied,

(4) $t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) u_{ii}(\hat{\lambda}^{(\mathrm{C})}) \leq c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) u_{ii}(\hat{\lambda}^{(\mathrm{M})})$ is satisfied,

(5) $t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) \geq 0$ or $t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) + c_{\mathrm{M}} u_{ii}(\hat{\lambda}^{(\mathrm{M})}) \leq 0$ is satisfied.

For any $\hat{\lambda}^{(\mathrm{C})} \geq 0$ and $\hat{\lambda}^{(\mathrm{M})} \geq 0$, we obtain the following relationships based on the above situations:

(i) If (1), then $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) \leq \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$,

(ii) If (2) and (3), then we obtain $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) \leq \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$,

(iii) If (2) and (4), then we obtain $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) \leq \hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$,

(iv) If (5), then $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) \leq \hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$.

Proof. In (1) and (5), the relationships (i) and (iv) are true. Hence, we need only prove relationships (ii) and (iii). Then, we obtain $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ using the closed forms of (3.5) and (3.6) and each optimized smoothing parameter. Through simple calculation, we obtain

$$\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) - \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) = \frac{d_i\{c_{\mathrm{M}} t_i^{\mathrm{C}}(\hat{\lambda}^{(\mathrm{C})}) u_{ii}(\hat{\lambda}^{(\mathrm{M})}) - t_i^{\mathrm{M}}(\hat{\lambda}^{(\mathrm{M})}) u_{ii}(\hat{\lambda}^{(\mathrm{C})})\}}{\{t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) + u_{ii}(\hat{\lambda}^{(\mathrm{C})})\}\{t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) + c_{\mathrm{M}} u_{ii}(\hat{\lambda}^{(\mathrm{M})})\}}.$$

Since $d_i > 0$, $t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) + u_{ii}(\hat{\lambda}^{(\mathrm{C})}) > 0$ and $t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) + c_{\mathrm{M}} u_{ii}(\hat{\lambda}^{(\mathrm{M})}) > 0$, the sign of $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) - \hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ is the same as the sign of $c_{\mathrm{M}} t_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}) u_{ii}(\hat{\lambda}^{(\mathrm{M})}) - t_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}) u_{ii}(\hat{\lambda}^{(\mathrm{C})})$. Hence, we obtain relationships (ii) and (iii). Thus, the theorem is proven. □

4. Numerical studies

In this section, we compare the LS estimator $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Xi}}$ in (1.3) with the proposed estimator $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$ in (2.3) through a numerical study. Let $\boldsymbol{R}_r = \mathrm{diag}(1, \ldots, r)$, and let $\boldsymbol{\Delta}_r(\rho)$ be an $r \times r$ matrix as follows:

$$\boldsymbol{\Delta}_r(\rho) = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{r-1} \\ \rho & 1 & \rho & \cdots & \rho^{r-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{r-3} \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ \rho^{r-1} & \rho^{r-2} & \rho^{r-3} & \cdots & 1 \end{pmatrix}.$$

The explanatory matrix $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{N}\boldsymbol{\Psi}^{1/2}$ where $\boldsymbol{\Psi} = \boldsymbol{R}_k^{1/2}\boldsymbol{\Delta}_k(\rho_a)\boldsymbol{R}_k^{1/2}$, $\boldsymbol{N}$ is an $n \times k$ matrix and each row vector of $\boldsymbol{N}$ is generated from the independent k-dimensional normal distribution with mean $\boldsymbol{0}_k$ and covariance matrix $\boldsymbol{I}_k$. Let $\boldsymbol{m}_i$, $(i = 1, \ldots, 12)$ be a p-dimensional vector. We set each $\boldsymbol{m}_i$ as follows:

$$\begin{aligned}
&\boldsymbol{m}_1 = \boldsymbol{h}(\boldsymbol{t}; e^2, e^{-1.5}, e^1),\ \boldsymbol{m}_2 = \boldsymbol{h}(\boldsymbol{t}; e^2, e^{-1.5}, e^2),\ \boldsymbol{m}_3 = \boldsymbol{h}(\boldsymbol{t}; e^2, e^{-2.0}, e^1),\ \boldsymbol{m}_4 = \boldsymbol{h}(\boldsymbol{t}; e^2, e^{-2.0}, e^2),\\
&\boldsymbol{m}_5 = \boldsymbol{h}(\boldsymbol{t}; e^2, e^{-2.5}, e^1),\ \boldsymbol{m}_6 = \boldsymbol{h}(\boldsymbol{t}; e^2, e^{-2.5}, e^2),\ \boldsymbol{m}_7 = \boldsymbol{h}(\boldsymbol{t}; e^3, e^{-1.5}, e^1),\ \boldsymbol{m}_8 = \boldsymbol{h}(\boldsymbol{t}; e^3, e^{-1.5}, e^2),\\
&\boldsymbol{m}_9 = \boldsymbol{h}(\boldsymbol{t}; e^3, e^{-2.0}, e^1),\ \boldsymbol{m}_{10} = \boldsymbol{h}(\boldsymbol{t}; e^3, e^{-2.0}, e^2),\ \boldsymbol{m}_{11} = \boldsymbol{h}(\boldsymbol{t}; e^3, e^{-2.5}, e^1),\ \boldsymbol{m}_{12} = \boldsymbol{h}(\boldsymbol{t}; e^3, e^{-2.5}, e^2),
\end{aligned}$$

where $\boldsymbol{t} = (1, \ldots, p)'$ and the ith element of $\boldsymbol{h}(\boldsymbol{t}; z_1, z_2, z_3)$ is $z_1\{1 - \exp(-z_2 t_i)\}^{z_3}$. Each element of $\boldsymbol{h}(\boldsymbol{t}; z_1, z_2, z_3)$ is Richard's growth curve model (Richard, 1959). We set the longitudinal trends using these $\boldsymbol{m}_i$ as $\boldsymbol{M}_k(\boldsymbol{t}) = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k)'$. Note that $\boldsymbol{m}_{i+6} = e\boldsymbol{m}_i$, $(i = 1, \ldots, 6)$, which indicates that the last six rows of $\boldsymbol{M}_{12}(\boldsymbol{t})$ are obtained by changing the scale of $\boldsymbol{M}_6(\boldsymbol{t})$. The response matrix $\boldsymbol{Y}$ is generated by $N_{n\times p}(\boldsymbol{A}\boldsymbol{M}(\boldsymbol{t}), \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n)$ where $\boldsymbol{\Sigma} = \boldsymbol{R}_p^{1/2}\boldsymbol{\Delta}_p(\rho_y)\boldsymbol{R}_p^{1/2}$. Then, we standardized $\boldsymbol{A}$. Let $\boldsymbol{k}_i = (\boldsymbol{0}'_{i-1}, 1, -2, 1, \boldsymbol{0}'_{q-2-i})'$, $(i = 1, \ldots, q-2)$ which is a q-dimensional vector, and $\boldsymbol{K} = (\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{q-2})(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_{q-2})'$. We set each element of $\boldsymbol{X}$ as a cubic B-spline basis function. Since $\boldsymbol{X}$ is set using the cubic B-spline, we note that $3 \le q \le p$. Additional details concerning $\boldsymbol{K}$ and $\boldsymbol{X}$ are reported in Green and Silverman (1994). We simulate $10,000$ repetitions for each n, p, k, ρ_a, and ρ_y. In each repetition, we fixed $\boldsymbol{A}$, but $\boldsymbol{Y}$ varies. We search $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ using fminsearch, which is a program in the software Matlab used to search for a minimum value, because $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ cannot be obtained in closed form. In searching $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$, we transform $\lambda' = \exp(\lambda)$ and search optimized λ' by each criterion because $\hat{\lambda}^{(\mathrm{C})} \ge 0$ and $\hat{\lambda}^{(\mathrm{M})} \ge 0$. In the search algorithm, the starting point for the search is set as $\lambda = 0$. Then, we obtain the optimized ridge parameters $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ using the closed forms of (3.5) and (3.6) in each repetition. In each repetition, we need to optimize q because $\boldsymbol{X}$ and $\boldsymbol{K}$ vary with q. We calculate $C_p(\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})}), \hat{\lambda}^{(\mathrm{C})})$ and $MC_p(\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})}), \hat{\lambda}^{(\mathrm{M})})$ for each $q = 3, \ldots, p$ in each repetition. Then, we adopt the optimized q by minimizing each criterion in each repetition. After that, we calculate $r(E[\boldsymbol{Y}], \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}})/(np)$ for each criterion, where $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}'_{\hat{\lambda}}\boldsymbol{X}' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}\boldsymbol{X}' = n^{-1}\boldsymbol{1}_n\boldsymbol{1}'_n\boldsymbol{Y}\boldsymbol{G}_{\hat{\lambda}} + \boldsymbol{H}_{\hat{\boldsymbol{\theta}}(\hat{\lambda})}\boldsymbol{Y}\boldsymbol{G}_{\hat{\lambda}}$, which is obtained using $\hat{\lambda}$ and $\hat{\boldsymbol{\theta}}(\hat{\lambda})$ for each criterion and the optimized q in each repetition. The average of $r(E[\boldsymbol{Y}], \hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}})$ over $10,000$ repetitions is regarded as the MSE of $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}$. We compare the values predicted using the estimators $\hat{\boldsymbol{\mu}}_{\hat{\lambda}}$ and $\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}$ with those using the LS estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Xi}}$ and the estimators $\hat{\boldsymbol{\mu}}_{\hat{\lambda}}$ and $\hat{\boldsymbol{\Xi}}_{\hat{\lambda}}$ in (2.1). When we use $\hat{\boldsymbol{\Xi}}_{\lambda}$, we obtain $\hat{\lambda}$ by minimizing $C_p(\boldsymbol{0}_k, \lambda)$ and $MC_p(\boldsymbol{0}_k, \lambda)$. As in the case of using $\hat{\boldsymbol{\Xi}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}$, we adopt q by using each criterion in each repetition for $\hat{\boldsymbol{\Xi}}$ and $\hat{\boldsymbol{\Xi}}_{\lambda}$. Some of the results are shown in Tables 1 and 2. The values in the tables are obtained by $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}]/(np)$, $\mathrm{MSE}[\hat{\boldsymbol{Y}}_{\hat{\lambda}}]/(np)$ where $\hat{\boldsymbol{Y}}_{\hat{\lambda}} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}'_{\hat{\lambda}}\boldsymbol{X}' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}_{\hat{\lambda}}\boldsymbol{X}'$, and $\mathrm{MSE}[\hat{\boldsymbol{Y}}]/(np)$ where $\hat{\boldsymbol{Y}} = \boldsymbol{1}_n\hat{\boldsymbol{\mu}}'\boldsymbol{X}' + \boldsymbol{A}\hat{\boldsymbol{\Xi}}\boldsymbol{X}'$.

Please insert Tables 1 and 2 around here

Each estimator optimized by using the MC_p criterion for λ, $\boldsymbol{\theta}$, and q is more improve than that by using the C_p criterion for each estimator in almost all situations. This indicates that the MC_p

criterion is a better estimator of the MSE of each predicted value of $\boldsymbol{Y}$ than the C_p criterion. The reasons for this are that the MC_p criterion is an unbiased estimator of MSE and each of the parameters in each estimator is optimized by minimizing the MC_p criterion. When $k = 6$, $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ provides a greater improvement than either $\hat{\Xi}_\lambda$ or $\hat{\Xi}$ in all situations. The estimator $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$, which is optimized using the MC_p criterion, has the smallest MSE among these estimators for almost situations when $k = 6$. Here, $\hat{\Xi}_\lambda$ provides a greater improvement than $\hat{\Xi}$ when $k = 6$ in all situations. When ρ_a is large, the estimator $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ provides a greater improvement than $\hat{\Xi}_\lambda$ in most situations when $k = 12$. On the other hand, $\hat{\Xi}_\lambda$ provides a greater improvement than $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ in most situations when ρ_a is small, $k = 12$ and $p = 10$. If $k = 12$, then $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ and $\hat{\Xi}_\lambda$ improve the LS estimator. Comparing the results for $k = 6$ with the results for $k = 12$ reveals that these estimators become poor estimators when k becomes large. The reasons for this are thought to be that $\boldsymbol{S}^{-1}$ and $\boldsymbol{A}$ become unstable and the $\boldsymbol{M}_{12}(\boldsymbol{t})$ has some curves that are in a different scale. Each MSE using each method and the C_p criterion is similar to that using the MC_p criterion if n becomes large because c_{M} is close to 1. When ρ_a becomes large, $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ improves the LS estimator more than when ρ_a is small. Since ρ_a controls the correlation in $\boldsymbol{A}$, the multicollinearity in $\boldsymbol{A}$ becomes large when ρ_a becomes large. Then, $\hat{\Xi}_\lambda$ is not a good estimator because $(\boldsymbol{A}'\boldsymbol{A})^{-1}$ is unstable. Hence, we can avoid the multicollinearity problem in $\boldsymbol{A}$ by using $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$, which is one of the purposes of the present study. In all situations, the new estimators improve the LS estimator $\hat{\Xi}$. In addition, $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ is better than $\hat{\Xi}_\lambda$ in most situations, especially when k is small or ρ_a is large. In general, $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ optimized using MC_p is the best method.

5. Conclusions

In the present paper, we estimate the longitudinal trends nonparametrically by using the nonparametric GMANOVA model in (1.1), which is defined using basis functions as $\boldsymbol{X}$ in the GMANOVA model. When we use basis functions as $\boldsymbol{X}$, the LS estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\Xi}$ incur overfitting. In order to avoid this problem, we proposed $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\Xi}_\lambda$ in (2.1) using the smoothing parameter λ (≥ 0) and the $q \times q$ known penalty non-negative definite matrix $\boldsymbol{K}$. However, if multicollinearity occurs in $\boldsymbol{A}$, $\hat{\Xi}$ and $\hat{\Xi}_\lambda$ are not good estimators due to large variance. In the present paper, we also proposed $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ in (2.3) in order to avoid the multicollinearity problem that occurs in $\boldsymbol{A}$ and the overfitting problem by using basis functions as $\boldsymbol{X}$. The estimator $\hat{\Xi}_\lambda$ controls the smoothness of each estimated longitudinal curve using only one parameter λ. On the other hand, in the estimator $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$, the rough smoothness of estimated longitudinal curves is controlled using λ, and each smoothness of $\phi_1(t), \ldots, \phi_k(t)$ in the varying coefficient model (1.4) is controlled by $\boldsymbol{\theta}$.

We also proposed the C_p and MC_p criteria in (3.3) and (3.4) for optimizing the ridge parameter $\boldsymbol{\theta}$ and the smoothing parameter λ. Then, using the GC_p criterion in (A.1) and minimizing this criterion in Theorem A.1, we obtain the optimized $\boldsymbol{\theta}$ using the C_p and MC_p criteria in closed form as (3.5) and (3.6) for any λ. Thus, we can regard the C_p and MC_p criteria as a function of λ. Hence, we need to optimize only one parameter λ in order to optimize $k+1$ parameters in $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ using these criteria. On the other hand, we must optimize two parameters when we use $\hat{\Xi}_{\theta,\lambda}$ in (2.2). This optimization is difficult and requires a complicated program and a long processing time for simulation or analysis of real data because the optimized θ cannot be obtained in closed form even if λ is fixed. This is the advantage of using $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$. This advantage does not appear to be important because of the high calculation power of CPUs. However, this advantage is made clear when we use $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ together with variable selection. Even if k becomes large, then this advantage remains when $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ is used because the optimized $\boldsymbol{\theta}$ obtained using each criterion is always obtained as (3.5) and (3.6) for any k. Furthermore, we must optimize q if we use model (1.1) to estimate the longitudinal trends. This means that we optimize the parameters in the estimators and calculate the valuation of the estimator for each q, and then we compare these values in order to optimize q. Since this optimization requires an iterative computational algorithm, we must reduce the processing time for estimating the parameters in the estimator. Hence, the advantage of using $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ is very important. This optimized ridge parameter in (3.5) and (3.6) corresponds to that in Yanagihara, Nagai and Satoh (2009) when $\boldsymbol{X} = \boldsymbol{I}_p$ and $\lambda\boldsymbol{K} = \boldsymbol{O}_{q,q}$.

Using some matrix properties, we showed that $\mathrm{tr}(\boldsymbol{G}_\lambda)$ and $\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}})$ in the C_p and MC_p criteria are always nonnegative. From $\mathrm{tr}(\boldsymbol{G}_\lambda) \geq 0$ for any $\lambda \geq 0$ in lemma 3.2, we also established the relationship between $t_i^{(\mathrm{C})}(\lambda)$ and $t_i^{(\mathrm{M})}(\lambda)$ for any $\lambda \geq 0$ in corollary 3.1. Then, in Theorem 3.1, we established the relationship between $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ if $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ are the same, where $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ are obtained by minimizing the C_p and MC_p criteria. Note that this relationship corresponds to that in Nagai, Yanagihara and Satoh (2010) when $\boldsymbol{X} = \boldsymbol{I}_p$ and $\lambda\boldsymbol{K} = \boldsymbol{O}_{q,q}$. In Theorem 3.2, we also established the relationships between $\hat{\theta}_i^{(\mathrm{C})}(\hat{\lambda}^{(\mathrm{C})})$ and $\hat{\theta}_i^{(\mathrm{M})}(\hat{\lambda}^{(\mathrm{M})})$ for the more general case, in which $\hat{\lambda}^{(\mathrm{C})}$ and $\hat{\lambda}^{(\mathrm{M})}$ are different. The reason of the relationship in Theorem 3.2 is occurred is that $\hat{\theta}_i^{(\mathrm{C})}(\lambda)$ and $\hat{\theta}_i^{(\mathrm{M})}(\lambda)$ for each $i = 1, \ldots, k$ can be regarded as a function of λ.

The numerical results reveal that $\hat{\Xi}_\lambda$ and $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ have some following properties. These estimation methods $\hat{\Xi}_\lambda$ and $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ improve the LS estimator in all situations, especially when ρ_a is large. This indicates that the proposed estimators are better than the LS estimator. Even if ρ_a becomes large, we note that $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ is stable because we add the ridge parameter to $\boldsymbol{A}'\boldsymbol{A}$ in the LS estimator. This result indicates that the multicollinearity problem in $\boldsymbol{A}$ can be avoided by using the estimator in (2.3). These estimators can be used to estimate the true longitudinal trends nonparametrically using basis functions as $\boldsymbol{X}$ without overfitting. The LS estimator and the proposed estimators $\hat{\Xi}_\lambda$ and $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ optimized using the MC_p criterion provide a greater improvement than the estimators optimized using the C_p criterion in most situations. The reason for this is that the MC_p criterion is the unbias estimator of MSE of the predicted value of $\boldsymbol{Y}$. Based on the present numerical study, $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$

can be used to estimate the longitudinal trends in most situations. In addition, the MC_p can be used to optimize the smoothing parameter λ and the number of basis functions q. Hence, we can use $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\Xi}}_{\boldsymbol{\theta},\lambda}$, the parameters $\boldsymbol{\theta}$, λ, and q of which are optimized by the MC_p criterion for estimating the longitudinal trends.

Acknowledgments

I would like to express my deepest gratitude to Dr. Hirokazu Yanagihara of Hiroshima University for his valuable ideas and useful discussions. In addition, I would like to thank Prof. Yasunori Fujikoshi, Prof. Hirofumi Wakaki, Dr. Kenichi Satoh and Dr. Kengo Kato of Hiroshima University for their useful suggestions and comments. Finally, I would like to thank Dr. Tomoyuki Akita of Hiroshima University for his advice with regard to programming.

Appendix

A.1. Minimization of the GC_p criterion

In this appendix, we show that the optimizations using the C_p and MC_p criteria in (3.3) and (3.4) are obtained in closed form as (3.5) and (3.6) for any λ (≥ 0). Nagai, Yanagihara and Satoh (2010) proposed the generalized C_p (GC_p) criterion for the MGR regression (originally the GC_p criterion for selection variables in the univariate regression was proposed by Atkinson (1980)). Similar to their idea, we proposed the GC_p criterion for the nonparametric GMANOVA model.

By omitting constant terms and some terms with respect to λ in the C_p and MC_p criteria in (3.3) and (3.4), these criteria are included in a class of criteria specified by α (> 0). This class is expressed by the GC_p criterion as

$$GC_p(\boldsymbol{\theta}, \lambda|\alpha) = \alpha\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}) + 2\mathrm{tr}(\boldsymbol{G}_\lambda)\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}}), \tag{A.1}$$

where the function $\hat{r}$ is given by (3.1). Note that $GC_p(\boldsymbol{\theta}, \lambda|1)$ and $GC_p(\boldsymbol{\theta}, \lambda|c_{\mathrm{M}})$ correspond to the terms with respect to $\boldsymbol{\theta}$ in the C_p and MC_p criteria. Using this GC_p criterion, we can deal systematically with the C_p and MC_p criteria for optimizing $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|\alpha) = (\hat{\theta}_1^{(\mathrm{G})}(\lambda|\alpha), \dots, \hat{\theta}_k^{(\mathrm{G})}(\lambda|\alpha))'$, and let $(\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha) \geq 0,\ i = 1, \dots, k)$ minimize the GC_p criterion for any λ (≥ 0). Then, $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda)$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)$ are obtained as $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda) = \hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|1)$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda) = \hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|c_{\mathrm{M}})$, respectively. Thus, we can deal systematically with the optimizations of $\boldsymbol{\theta}$ when we use the C_p and MC_p criteria. This means that we need only obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|\alpha)$ in order to obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda)$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)$ for any λ and some α. If $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|\alpha)$ is obtained in closed form for any fixed λ, we do not need to use the iterative computational algorithm for optimizing the ridge parameter $\boldsymbol{\theta}$. In order to obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|\alpha)$, we obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha)$, $(i = 1, \dots, k)$ in closed form, as shown in the following theorem.

Theorem A.1. For each i and any λ (≥ 0), $\hat{\theta}_i^{(\mathrm{G})}(\lambda)$ is obtained as

$$\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha) = \begin{cases} 0 & (t_i(\lambda|\alpha) \geq 0) \\ \dfrac{-d_i t_i(\lambda|\alpha)}{t_i(\lambda|\alpha) + \alpha u_{ii}} & (t_i(\lambda|\alpha) < 0 \text{ and } 0 < t_i(\lambda|\alpha) + \alpha u_{ii}) \\ \infty & (\text{otherwise}) \end{cases}, \tag{A.2}$$

where $t_i(\lambda|\alpha) = \alpha(v_{ii} - u_{ii}) - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)$.

Proof. Since $\hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda} = \boldsymbol{1}_n \hat{\boldsymbol{\mu}}_\lambda' \boldsymbol{X}' + \boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{Y} \boldsymbol{G}_\lambda$ and we use the properties of the function $\hat{r}$ in Section 3.1, we can calculate $\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})$ in the GC_p criterion in (A.1) as follows:

$$\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}) = \hat{r}(\boldsymbol{Y}, \boldsymbol{1}_n \hat{\boldsymbol{\mu}}_\lambda' \boldsymbol{X}') + 2\mathrm{tr}\{(\boldsymbol{1}_n \hat{\boldsymbol{\mu}}_\lambda \boldsymbol{X}' - \boldsymbol{Y}) \boldsymbol{S}^{-1} (\boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{Y} \boldsymbol{G}_\lambda)'\} + \hat{r}(\boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{Y} \boldsymbol{G}_\lambda, \boldsymbol{O}_{n,p}).$$

Since $\boldsymbol{G}_\lambda = \boldsymbol{G}_\lambda'$ for any λ, $\boldsymbol{H}_{\boldsymbol{\theta}} = \boldsymbol{H}_{\boldsymbol{\theta}}'$ and $\boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{1}_n = \boldsymbol{0}_n$ for any $\boldsymbol{\theta}$, the second term in the right-hand side of the above equation can be calculated as

$$\mathrm{tr}\{(\boldsymbol{1}_n \hat{\boldsymbol{\mu}}_\lambda \boldsymbol{X}' - \boldsymbol{Y}) \boldsymbol{S}^{-1} (\boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{Y} \boldsymbol{G}_\lambda)'\} = -\mathrm{tr}(\boldsymbol{Y} \boldsymbol{S}^{-1} \boldsymbol{G}_\lambda \boldsymbol{Y}' \boldsymbol{H}_{\boldsymbol{\theta}}).$$

Note that $\boldsymbol{H}_{\boldsymbol{\theta}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^{-1}\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{Q}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{Q}'\boldsymbol{A}'$ because $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{D}$. Hence, we obtain the following results:

$$\begin{aligned} \mathrm{tr}(\boldsymbol{Y} \boldsymbol{S}^{-1} \boldsymbol{G}_\lambda \boldsymbol{Y}' \boldsymbol{H}_{\boldsymbol{\theta}}) &= \mathrm{tr}\{\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{G}_\lambda \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Q}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\}, \\ \hat{r}(\boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{Y} \boldsymbol{G}_\lambda, \boldsymbol{O}_{n,p}) &= \mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}} \boldsymbol{Y} \boldsymbol{G}_\lambda \boldsymbol{S}^{-1} \boldsymbol{G}_\lambda \boldsymbol{Y}' \boldsymbol{H}_{\boldsymbol{\theta}}) \\ &= \mathrm{tr}\{\boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{G}_\lambda \boldsymbol{S}^{-1}\boldsymbol{G}_\lambda \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Q}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\}. \end{aligned}$$

Since $\boldsymbol{D}$ and $(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}$ are diagonal matrices, we obtain $(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1} = \boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-2}$. Hence, $\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda})$ is calculated as

$$\hat{r}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}_{\boldsymbol{\theta},\lambda}) = \hat{r}(\boldsymbol{Y}, \boldsymbol{1}_n \hat{\boldsymbol{\mu}}_\lambda' \boldsymbol{X}') - 2\mathrm{tr}\{\boldsymbol{V}(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\} + \mathrm{tr}\{\boldsymbol{U}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-2}\},$$

where $\boldsymbol{V} = \boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{S}^{-1}\boldsymbol{G}_\lambda \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Q}$ and $\boldsymbol{U} = \boldsymbol{Q}'\boldsymbol{A}'\boldsymbol{Y}\boldsymbol{G}_\lambda \boldsymbol{S}^{-1}\boldsymbol{G}_\lambda \boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Q}$. Clearly, $\boldsymbol{V}$ and $\boldsymbol{U}$ change with λ. Based on this result and $\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}}) = \mathrm{tr}\{\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})\}$, we can calculate the GC_p criterion in (A.1) as follows:

$$GC_p(\boldsymbol{\theta}|\lambda, \alpha) = \alpha \hat{r}(\boldsymbol{Y}, \boldsymbol{1}_n \hat{\boldsymbol{\mu}}_\lambda' \boldsymbol{X}') + \alpha \mathrm{tr}\{\boldsymbol{U}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-2}\} - 2\mathrm{tr}\{(\alpha \boldsymbol{V} - \mathrm{tr}(\boldsymbol{G}_\lambda)\boldsymbol{D})(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\},$$

Then, we calculate the second and third terms in the right-hand side of the above equation as follows:

$$\alpha \mathrm{tr}\{\boldsymbol{U}\boldsymbol{D}(\boldsymbol{D} + \boldsymbol{\Theta})^{-2}\} - 2\mathrm{tr}\{(\alpha \boldsymbol{V} - \mathrm{tr}(\boldsymbol{G}_\lambda)\boldsymbol{D})(\boldsymbol{D} + \boldsymbol{\Theta})^{-1}\} = \sum_{i=1}^{k} \left\{ \frac{\alpha d_i u_{ii}}{(d_i + \theta_i)^2} - 2\frac{\alpha v_{ii} - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)}{d_i + \theta_i} \right\},$$

where u_{ij} and v_{ij} are the (i, j)th elements of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Clearly, u_{ij} and v_{ij} also vary with λ. Note that $u_{ii} \geq 0$, $(i = 1, \ldots, k)$ for any $\lambda \geq 0$ because $\boldsymbol{S}^{-1}$ is a positive definite matrix (see, e.g., Harville (1997)). Let $\varphi_i(\theta_i)$, $(i = 1, \ldots, k)$ be as follows:

$$\varphi_i(\theta_i, \lambda|\alpha) = \frac{\alpha d_i u_{ii}}{(d_i + \theta_i)^2} - 2\frac{\alpha v_{ii} - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)}{d_i + \theta_i}. \tag{A.3}$$

Using $\varphi_i(\theta_i, \lambda|\alpha)$, we can express $GC_p(\boldsymbol{\theta}, \lambda|\alpha)$ as follows:

$$GC_p(\boldsymbol{\theta}, \lambda|\alpha) = \alpha\hat{r}(\boldsymbol{Y}, \mathbf{1}_n\hat{\boldsymbol{\mu}}'_\lambda \boldsymbol{X}') + \sum_{i=1}^{k} \varphi_i(\theta_i, \lambda|\alpha).$$

Since $\alpha\hat{r}(\boldsymbol{Y}, \mathbf{1}_n\hat{\boldsymbol{\mu}}'_\lambda \boldsymbol{X}')$ does not depend on $\boldsymbol{\theta}$, we can obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha)$ by minimizing $\varphi_i(\theta_i, \lambda|\alpha)$ for each $i = 1, \ldots, k$ and any λ (≥ 0). In order to obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha)$, we consider the following function for $w \in \mathbb{R}$:

$$\varphi_i(w) = \frac{\alpha d_i u_{ii}}{(d_i + w)^2} - 2\frac{\alpha v_{ii} - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)}{d_i + w}, \tag{A.4}$$

If we restrict w to be greater than or equal to 0, then this function is equivalent to the function $\varphi_i(\theta_i, \lambda|\alpha)$ in (A.3), which must be minimized. Note that $\lim_{w \to \pm\infty} \varphi_i(w) = 0$ and $\lim_{w \to -d_i \pm 0} \varphi_i(w) = +\infty$. Letting $\dot{\varphi}_i(w) = \partial\varphi_i(w)/\partial w$, we obtain

$$\dot{\varphi}_i(w) = -\frac{2}{(d_i + w)^3} \left\{\alpha d_i u_{ii} - (\alpha v_{ii} - d_i \mathrm{tr}(\boldsymbol{G}_\lambda))(d_i + w)\right\}.$$

Let $\hat{w}$ satisfy $\dot{\varphi}_i(w)|_{w=\hat{w}} = 0$ and $\hat{w} \neq \pm\infty$, then $\hat{w}$ is obtained by

$$\hat{w} = \frac{-d_i t_i(\lambda|\alpha)}{t_i(\lambda|\alpha) + \alpha u_{ii}}, \ \ (\text{if } t_i(\lambda|\alpha) + \alpha u_{ii} \neq 0),$$

where $t_i(\lambda|\alpha) = \alpha(v_{ii} - u_{ii}) - d_i \mathrm{tr}(\boldsymbol{G}_\lambda)$. Note that the function $\varphi_i(w)$ in (A.4) has a minimum value at $\hat{w}$, which is $\dot{\varphi}_i(w)|_{w<\hat{w}} < 0$ and $\dot{\varphi}_i(w)|_{w>\hat{w}} > 0$. Note that the sign of $t_i(\lambda|\alpha)$ is the same as the sign of $\dot{\varphi}_i(w)|_{w=0}$. In order to obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha)$ (≥ 0), we consider the following situations:

(1) $t_i(\lambda|\alpha) \geq 0$ is satisfied,
(2) $t_i(\lambda|\alpha) < 0$ and $t_i(\lambda|\alpha) + \alpha u_{ii} > 0$ are satisfied,
(3) $t_i(\lambda|\alpha) < 0$ and $t_i(\lambda|\alpha) + \alpha u_{ii} < 0$ are satisfied,

In (1), $-d_i < \hat{w} < 0$, because $u_{ii} \geq 0$ and $\alpha > 0$. In addition, $\varphi(w) \geq \varphi(0)$ for any $w \geq 0$, because $\hat{w} < 0$, and $t_i(\lambda|\alpha) \geq 0$ indicates that the sign of $\dot{\varphi}_i(w)|_{w=0}$ is nonnegative. This means that the minimum value of $\varphi(w)$ in $w \geq 0$ is obtained when $w = 0$ in situation (1). In (2), $\hat{w} > 0$, and then the minimum value of $\varphi(w)$ in $w \geq 0$ is obtained when $w = \hat{w}$. In (3), since $\hat{w} < -d_i$ and $\dot{\varphi}(w)|_{w=0} < 0$, we obtain $\varphi(0) > \varphi(w_1) > \varphi(w_2)$ for any $w_2 > w_1 > 0$. Hence, $\varphi(w)$ is minimized when $w = \infty$ in $w \geq 0$. From the above results, we obtain $\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha)$ (≥ 0) as follows:

$$\hat{\theta}_i^{(\mathrm{G})}(\lambda|\alpha) = \begin{cases} 0 & (t_i(\lambda|\alpha) \geq 0) \\ \dfrac{-d_i t_i(\lambda|\alpha)}{t_i(\lambda|\alpha) + \alpha u_{ii}} & (t_i(\lambda|\alpha) < 0 \ \mathit{and} \ 0 < t_i(\lambda|\alpha) + \alpha u_{ii}) \\ \infty & (\mathit{otherwise}) \end{cases}, \ \ (i = 1, \ldots, k).$$

Thus, the theorem is proven. □

Note that $\hat{\boldsymbol{\theta}}^{(\mathrm{G})}(\lambda|\alpha)$ corresponds to that in Nagai, Yanagihara and Satoh (2010) when $\boldsymbol{X} = \boldsymbol{I}_p$ and $\lambda\boldsymbol{K} = \boldsymbol{O}_{q,q}$. Since we obtain $\hat{\boldsymbol{\theta}}^{(\mathrm{C})}(\lambda)$ and $\hat{\boldsymbol{\theta}}^{(\mathrm{M})}(\lambda)$ in closed form as (A.2) for any λ, we must optimize only one parameter λ in order to optimize $k+1$ parameters. The use of $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ is advantageous because only an iterative computational algorithm is required for optimizing only one parameter λ for any k. This means that we can reduce the processing time required to optimize the parameters in the estimator $\hat{\Xi}_{\boldsymbol{\theta},\lambda}$ which is defined by (2.3). When we use $\hat{\Xi}_{\lambda}$ in (2.1), we also need the same iterative computational algorithm to optimize only one parameter λ.

On the other hand, when we use $\hat{\Xi}_{\theta,\lambda}$ in (2.2), the GC_p criterion for optimizing θ for any fixed λ is obtained as

$$GC_p(\theta|\lambda,\alpha) = \alpha\hat{r}(\boldsymbol{Y}, \mathbf{1}_n\hat{\boldsymbol{\mu}}'_\lambda\boldsymbol{X}') + \sum_{i=1}^{k}\varphi_i(\theta,\lambda|\alpha).$$

Since we need to minimize $\sum_{i=1}^{k}\varphi_i(\theta,\lambda|\alpha)$ in order to optimize θ, we cannot obtain $\hat{\theta}^{(\mathrm{G})}(\lambda|\alpha)$ that minimizes this GC_p criterion for $\hat{\Xi}_{\theta,\lambda}$ in closed form, even if λ is fixed. Thus, we use an iterative computational algorithm to optimize the parameters λ and θ simultaneously. This iterative computational algorithm for optimizing two parameters is difficult and requires a longer processing time than the optimization of a single parameter.

References

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

[2] Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman & Hall/CRC.

[3] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective.* New York Springer.

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

[5] Kshirsagar, A. M. & Smith, W. B. (1995). *Growth Curves.* Marcel Dekker.

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

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

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

[9] Nagai, I., Yangihara, H. & Satoh, K. (2010). Optimization of Ridge Parameters in Multivariate Generalized Ridge Regression by Plug-in Methods. *TR 10-03, Statistical Research Group, Hiroshima University.*

[10] Potthoff, R. F. & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.

[11] Riedel, K. S. & Imre, K. (1993). Smoothing spline growth curves with covariates. *Comm. Statist. A – Theory methods*, **22**, 1795–1818

[12] Richard, F. J. (1959). A flexible growth function for empirical use. *J. Exp. Bot.*, **10**, 290–301.

[13] Satoh, K. & Yanagihara, H. (2010). Estimation of varying coefficients for a growth curve model. *Amer. J. Math. Management Sci.* (in press).

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

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

[16] Takane, Y., Jung, K. & Hwang, H. (2011). Regularized reduced rank growth curve models. *Comput. Statist. Data Anal.*, **55**, 1041–1052.

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

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

Table 1. MSE when q is selected using each criterion for each method in each repetition ($k = 6$)

				Using $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}$		Using $\hat{\boldsymbol{Y}}_{\hat{\lambda}}$		Using $\hat{\boldsymbol{Y}}$	
ρ_y	ρ_a	p	n	C_p	MC_p	C_p	MC_p	C_p	MC_p
0.2	0.2	5	30	0.127	**0.123**	0.133	0.125	0.206	0.199
			50	0.080	**0.079**	0.082	0.080	0.121	0.119
		10	30	0.119	0.098	0.121	**0.090**	0.168	0.119
			50	0.063	0.058	0.062	**0.056**	0.080	0.070
	0.8	5	30	0.110	**0.101**	0.143	0.135	0.206	0.199
			50	0.067	**0.065**	0.088	0.086	0.122	0.119
		10	30	0.111	**0.080**	0.128	0.093	0.170	0.119
			50	0.056	**0.049**	0.063	0.057	0.080	0.070
	0.99	5	30	0.090	**0.078**	0.147	0.140	0.207	0.199
			50	0.054	**0.050**	0.090	0.088	0.122	0.120
		10	30	0.095	**0.060**	0.129	0.094	0.169	0.118
			50	0.045	**0.036**	0.064	0.058	0.079	0.069
0.8	0.2	5	30	0.133	**0.131**	0.154	0.147	0.208	0.201
			50	0.087	**0.086**	0.093	0.092	0.122	0.120
		10	30	0.123	**0.101**	0.136	0.106	0.179	0.133
			50	0.069	**0.065**	0.070	0.065	0.089	0.080
	0.8	5	30	0.113	**0.103**	0.159	0.153	0.207	0.200
			50	0.066	**0.063**	0.094	0.092	0.122	0.120
		10	30	0.108	**0.074**	0.140	0.107	0.178	0.131
			50	0.055	**0.047**	0.072	0.065	0.088	0.078
	0.99	5	30	0.092	**0.078**	0.162	0.156	0.208	0.201
			50	0.053	**0.049**	0.095	0.094	0.122	0.120
		10	30	0.096	**0.059**	0.142	0.108	0.178	0.131
			50	0.046	**0.037**	0.073	0.066	0.087	0.078
Average				0.086	**0.074**	0.110	0.098	0.147	0.130

Table 2. MSE when q is selected using each criterion for each method in each repetition ($k = 12$)

				Using $\hat{\boldsymbol{Y}}_{\hat{\boldsymbol{\theta}}(\hat{\lambda}),\hat{\lambda}}$		Using $\hat{\boldsymbol{Y}}_{\hat{\lambda}}$		Using $\hat{\boldsymbol{Y}}$	
ρ_y	ρ_a	p	n	C_p	MC_p	C_p	MC_p	C_p	MC_p
0.2	0.2	5	30	0.299	**0.292**	0.312	0.296	0.383	0.364
			50	0.184	0.183	0.183	**0.180**	0.222	0.217
		10	30	0.317	0.247	0.326	**0.226**	0.382	0.248
			50	0.146	0.137	0.146	**0.134**	0.165	0.150
	0.8	5	30	0.285	**0.279**	0.313	0.295	0.384	0.365
			50	0.175	**0.173**	0.182	0.179	0.223	0.218
		10	30	0.305	0.223	0.329	**0.216**	0.378	0.226
			50	0.145	0.132	0.145	**0.129**	0.155	0.135
	0.99	5	30	0.224	**0.204**	0.314	0.296	0.383	0.364
			50	0.142	**0.138**	0.183	0.180	0.222	0.218
		10	30	0.270	**0.173**	0.330	0.211	0.377	0.221
			50	0.134	**0.119**	0.143	0.123	0.148	0.127
0.8	0.2	5	30	0.323	**0.321**	0.342	0.331	0.387	0.368
			50	0.204	0.204	0.205	**0.203**	0.224	0.219
		10	30	0.330	0.277	0.344	**0.256**	0.389	0.282
			50	0.165	0.153	0.167	**0.152**	0.178	0.159
	0.8	5	30	0.298	**0.294**	0.337	0.321	0.385	0.367
			50	0.191	**0.191**	0.200	0.197	0.224	0.220
		10	30	0.309	**0.244**	0.346	0.251	0.386	0.265
			50	0.161	**0.150**	0.166	0.151	0.175	0.159
	0.99	5	30	0.228	**0.208**	0.338	0.322	0.386	0.368
			50	0.142	**0.137**	0.199	0.196	0.223	0.219
		10	30	0.263	**0.170**	0.347	0.236	0.384	0.247
			50	0.126	**0.106**	0.161	0.139	0.166	0.145
Average				0.224	**0.198**	0.252	0.217	0.289	0.245