TR-No. 13-07, Hiroshima Statistical Research Group, 1–17

Explicit Solution to the Minimization Problem of Generalized Cross-Validation Criterion for Selecting Ridge Parameters in Generalized Ridge Regression

Hirokazu Yanagihara

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

Abstract

This paper considers optimization of the ridge parameters in generalized ridge regression (GRR) by minimizing a model selection criterion. GRR has a major advantage over ridge regression (RR) in that a solution to the minimization problem for one model selection criterion, i.e., Mallows' C_p criterion, can be obtained explicitly with GRR, but such a solution for any model selection criteria, e.g., C_p criterion, cross-validation (CV) criterion, or generalized CV (GCV) criterion, cannot be obtained explicitly with RR. On the other hand, C_p criterion is at a disadvantage compared to CV and GCV criteria because a good estimate of the error variance is required in order for C_p criterion to work well. In this paper, we show that ridge parameters optimized by minimizing GCV criterion can also be obtained by closed forms in GRR. We can overcome one disadvantage of GRR by using GCV criterion for the optimization of ridge parameters.

(Last Modified: May 17, 2013)

Key words: Explicit optimal solution, Generalized ridge regression, Generalized cross-validation criterion, Linear regression model, High-dimensional explanatory variables, Multiple ridge parameters, Principal component regression, Selection of ridge parameters.

MSC 2010 subject classifications: Primary 62J07, Secondary 62F07.

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

1. Introduction

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be an n-dimensional vector of response variables and $\boldsymbol{X}$ be an $n \times k$ matrix of nonstochastic centralized explanatory variables ($\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$) with $\mathrm{rank}(\boldsymbol{X}) = m$ $(\leq \min\{k, n-1\})$, where n is the sample size, $\boldsymbol{1}_n$ is an n-dimensional vector of ones, and $\boldsymbol{0}_k$ is a k-dimensional vector of zeros. We assume a linear relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$, expressed by the liner regression model:

$$\boldsymbol{y} = \mu\boldsymbol{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where μ is an unknown location parameter, $\boldsymbol{\beta}$ is a k-dimensional vector of unknown regression coefficients, and ε is an n-dimensional vector of independent error variables from a distribution with mean 0 and error variance σ^2.

The ordinary least square (OLS) method is widely used for estimating the unknown parameters in (1.1). This is because although the OLS estimators of μ and $\boldsymbol{\beta}$ are given by simple forms, they have several desirable theoretical properties. The OLS estimators of μ and $\boldsymbol{\beta}$ are given by $\hat{\mu} = \bar{y}$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{+}\boldsymbol{X}'\boldsymbol{y}$, respectively, where $\bar{y}$ is a sample mean of the elements of $\boldsymbol{y}$, i.e., $\bar{y} = \mathbf{1}_n'\boldsymbol{y}/n$, and $\boldsymbol{A}^{+}$ is the Moore-Penrose inverse matrix of $\boldsymbol{A}$ (for details of the Moore-Penrose inverse matrix, see, e.g., Harville, 1997, chap. 20). However, when multicollinearity occurs in $\boldsymbol{X}$, the OLS estimator of $\boldsymbol{\beta}$ is not a good estimator in the sense that it has a large variance. The ridge regression (RR) estimation proposed by Hoerl and Kennard (1970) is one of the methods that avoids the problems from multicollinearity. The RR estimator is defined by adding $\theta\boldsymbol{I}_k$ to $\boldsymbol{X}'\boldsymbol{X}$ in $\hat{\boldsymbol{\beta}}$, where $\theta \in \mathbb{R}_{+} = \{\theta \in \mathbb{R} | \, \theta \geq 0\}$ is called a ridge parameter. Since the estimates provided by the RR estimator depend heavily on the value of θ, the optimization of θ is a very important problem. One of the optimization methods is to choose a ridge parameter that minimizes a model selection criterion, e.g., Mallows' C_p (Mallows, 1973, 1995), cross-validation (CV; Stone, 1974) and generalized CV (GCV; Craven & Wahba, 1979) criteria (see, e.g., Golub *et al*., 1979; Yanagihara & Satoh, 2010). However, an optimal value of θ cannot be obtained without an iterative computational algorithm.

Hoerl and Kennard (1970) proposed not only the RR but also a generalized ridge regression (GRR) in their paper. Although GRR estimation was proposed over 40 years ago, even today, many researchers study the theoretical properties of the GRR estimator (e.g., Jimichi, 2008), and use GRR for real data analysis (e.g., Smyth *et al*., 2011), and for developing new statistical procedures based on GRR (e.g., Batah *et al*., 2008; Jensen & Ramirez, 2010; Yanagihara, 2012). The GRR estimator is defined not by a single ridge parameter but by multiple ridge parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)' \in \mathbb{R}_{+}^{k}$, i.e., the GRR estimator of $\boldsymbol{\beta}$ is defined by replacing $\theta\boldsymbol{I}_k$ in the RR estimator of $\boldsymbol{\beta}$ with $\boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'$, where $\mathbb{R}_{+}^{k}$ is the kth Cartesian power of $\mathbb{R}_{+}$, $\boldsymbol{\Theta}$ is a kth diagonal matrix whose jth diagonal element is θ_j, and $\boldsymbol{Q}$ is the kth orthogonal matrix that diagonalizes $\boldsymbol{X}'\boldsymbol{X}$. Even though the number of ridge parameters has increased, we can obtain $\boldsymbol{\theta}$ minimizing C_p criterion by closed form (see, e.g., Lawless, 1981; Walker & Page, 2001; Yanagihara *et al*., 2009; Nagai *et al*., 2012). However, C_p criterion is at a disadvantage compared to the CV or GCV criteria because a good estimate of the error variance σ^2 is required in order for C_p criterion to work well. In an extended GRR, several authors have tried solving the minimization problem for a model selection criterion other than C_p criterion by using the Newton-Raphson method (e.g., Gu & Wahba, 1991; Wood, 2000). In this paper, we show that ridge parameters optimized by minimizing the GCV criterion can also be obtained by closed forms in the original GRR. We can overcome one of the disadvantages of GRR by using GCV criterion for the optimization of the ridge parameters.

This paper is organized as follows: In Section 2, we describe the use of GCV criterion for selecting the ridge parameters for GRR, and we present some lemmas to express explicitly the optimal solution of GCV criterion. In Section 3, we show an explicit solution to the minimization problem of GCV criterion for GRR, and present additional theorems on GRR after optimizing the ridge parameters. In Section 4, we apply GRR to a linear regression model with high-dimensional explanatory variables. A numerical examination is conducted at the end of Section 4. Technical details are provided in the Appendix.

2. Preliminaries

Let $\boldsymbol{Q}$ be the kth orthogonal matrix that diagonalizes $\boldsymbol{X}'\boldsymbol{X}$ as

$$\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{O}_{m,k-m} \\ \boldsymbol{O}_{k-m,m} & \boldsymbol{O}_{k-m,k-m} \end{pmatrix}, \tag{2.1}$$

where $\boldsymbol{O}_{k,m}$ is a $k \times m$ matrix of zeros, and

$$\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_m) \text{ and } d_1, \ldots, d_m \text{ are nonzero eigenvalues of } \boldsymbol{X}'\boldsymbol{X}. \tag{2.2}$$

We note that $d_1, \ldots, d_m$ are positive, because we assume that $\boldsymbol{X}'\boldsymbol{X}$ is a positive semidefinite matrix. Without loss of generality, it assumes that $d_1 \geq \cdots \geq d_m$. Moreover, let $\boldsymbol{M}_{\boldsymbol{\theta}}$ be a $k \times k$ matrix defined by

$$\boldsymbol{M}_{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}',$$

where $\boldsymbol{\Theta}$ is the kth diagonal matrix given by $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_k)$. In particular, we write $\boldsymbol{M}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} = \boldsymbol{0}_k$ as $\boldsymbol{M}$. Then, a GRR estimator of $\boldsymbol{\beta}$ is defined by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = \boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.3}$$

It is clear that the GRR estimator in (2.3) with $\boldsymbol{\theta} = \boldsymbol{0}_k$ coincides with the ordinary least square (OLS) estimator defined by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{M}^{+}\boldsymbol{X}'\boldsymbol{y}, \tag{2.4}$$

where $\boldsymbol{M}^{+}$ is the Moore-Penrose inverse matrix of $\boldsymbol{M}$, i.e.,

$$\boldsymbol{M}^{+} = \begin{pmatrix} \boldsymbol{D}^{-1} & \boldsymbol{O}_{m,k-m} \\ \boldsymbol{O}_{k-m,m} & \boldsymbol{O}_{k-m,k-m} \end{pmatrix}.$$

Equation (2.3) leads to a predictor of $\boldsymbol{y}$ derived from GRR as

$$\hat{\boldsymbol{y}}_{\boldsymbol{\theta}} = \bar{y}\boldsymbol{1}_n + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = (\boldsymbol{J}_n + \boldsymbol{X}\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}')\boldsymbol{y}, \tag{2.5}$$

where $\boldsymbol{J}_n$ is an $n \times n$ projection matrix defined by $\boldsymbol{J}_n = \boldsymbol{1}_n\boldsymbol{1}_n'/n$.

Notice that $\mathrm{tr}(\boldsymbol{J}_n + \boldsymbol{X}\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}') = 1 + \mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M})$. Thus, according to a general formula of the GCV criterion provide by Craven and Wahba (1979), the GCV criterion for selecting $\boldsymbol{\theta}$ can be defined by

$$\mathrm{GCV}(\boldsymbol{\theta}) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}})}{n[1 - \{1 + \mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M})\}/n]^2}. \tag{2.6}$$

A main aim of this paper is to obtain the closed form of the minimizers of $\mathrm{GCV}(\boldsymbol{\theta})$. Let $z_1, \ldots, z_m$ be elements of an m-dimensional vector defined by

$$(z_1, \ldots, z_m)' = (\boldsymbol{D}^{-1/2}, \boldsymbol{O}_{m,k-m})\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{y}, \tag{2.7}$$

and let t_j $(j = 1, \ldots, m)$ be the jth-order statistic of $z_1^2, \ldots, z_m^2$, i.e.,

$$t_j = \begin{cases} \min\{z_1^2, \ldots, z_m^2\} & (j = 1) \\ \min\{\{z_1^2, \ldots, z_m^2\} \backslash \{t_1, \ldots, t_{j-1}\}\} & (j = 2, \ldots, m-1) \end{cases}. \tag{2.8}$$

The following statistic based on $t_1, \ldots, t_m$ plays a big role in expressing the closed form of the minimizers of GCV criterion:

$$s_0^2 = \frac{\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{X}\boldsymbol{M}^{+}\boldsymbol{X}')\boldsymbol{y}}{n-m-1}, \quad s_\alpha^2 = \frac{(n-m-1)s_0^2 + \sum_{j=1}^{\alpha} t_j}{n-m-1+\alpha} \ (\alpha = 1, \ldots, m). \tag{2.9}$$

When the sample size is smaller than the number of explanatory variables, $m \leq n-1$ holds because $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$ is satisfied. It is also easy to see that $s_0^2 = 0$ holds when $m = n-1$. It should be kept in mind that $s_0^2 = 0$ holds in most cases of high-dimensional explanatory variables. The term s_α^2 has the following property (the proof is given in Appendix A):

Lemma 1. *Let a_* be an integer defined by*

$$a_* \in \{0, 1, \ldots, m\} \ s.t. \ s_{a_*}^2 \in R_{a_*}, \tag{2.10}$$

where R_α is a range given by

$$R_\alpha = \begin{cases} (0, t_1] & (\alpha = 0) \\ (t_\alpha, t_{\alpha+1}] & (\alpha = 1, \ldots, m-1) \\ (t_m, \infty) & (\alpha = m) \end{cases}. \tag{2.11}$$

Then following properties are satisfied:

(1) *Case of $s_0^2 \neq 0$: ${}^{\exists !}a_* \in \{0, 1, \ldots, m\}$ s.t. $s_{a_*}^2 \in R_{a_*}$. Then $s_{a_*}^2 \leq s_0^2$ is satisfied.*

(2) *Case of $s_0^2 = 0$: $\neg({}^{\exists}a_* \in \{0, 1, \ldots, m\}$ s.t. $s_{a_*}^2 \in R_{a_*})$.*

On the other hand, the GRR estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ in (2.3) and GCV($\boldsymbol{\theta}$) in (2.6) satisfy the following property (the proof is given in Appendix B):

Lemma 2. *The GRR estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ and GCV($\boldsymbol{\theta}$) are invariant to any changes in $\theta_{m+1}, \ldots, \theta_k$.*

From Lemma 2, we set $\theta_{m+1} = \cdots = \theta_k = \infty$ for simplicity. Moreover, Lemma 2 indicates that GCV($\boldsymbol{\theta}$) can be regarded as a function with respect to $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_m)'$. In particular, the GCV criterion can be expressed as the following lemma (the proof is given in Appendix C):

Lemma 3. *The GCV($\boldsymbol{\theta}$) can be written as*

$$\mathrm{GCV}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}_1) = \frac{\{(n-m-1)s_0^2 + \sum_{j=1}^{m}\{\theta_j/(d_j+\theta_j)\}^2 z_j^2\}/n}{\{1 - (m+1-\sum_{j=1}^{m}\theta_j/(d_j+\theta_j))/n\}^2}. \tag{2.12}$$

Notice that when $s_0^2 \neq 0$,

$$\frac{\partial}{\partial \theta_\alpha} g(\boldsymbol{\theta}_1)\bigg|_{\boldsymbol{\theta}_1=\mathbf{0}_m} = -\frac{2s_0^2}{d_\alpha(n-m-1)} < 0.$$

This implies that $g(\boldsymbol{\theta}_1)$ does not reach a minimum at $\mathbf{0}_m$ when $s_0^2 \neq 0$. On the other hand, $g(\boldsymbol{\theta}_1)$ is not determinate when $s_0^2 = 0$ and $\theta_1 = \cdots = \theta_m = 0$. Hence, we search for optimal solutions of $g(\boldsymbol{\theta}_1)$ in $\boldsymbol{\theta}_1 \in \mathbb{R}_+^m \backslash \{\mathbf{0}_m\}$.

3. Main Results

3.1. Optimal Solutions of GCV Criterion

The ridge parameters $\theta_1, \ldots, \theta_m$ that minimize $g(\boldsymbol{\theta}_1)$ in (2.12) are derived as in the following theorem (the proof is given in Appendix D):

Theorem 1. *Let $\hat{\theta}_1, \ldots, \hat{\theta}_m$ be optimal solutions of $g(\boldsymbol{\theta}_1)$, i.e.,*

$$\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_1, \ldots, \hat{\theta}_m)' = \arg \min_{\boldsymbol{\theta}_1 \in \mathbb{R}_+^m \backslash \{\mathbf{0}_m\}} g(\boldsymbol{\theta}_1).$$

Then, an explicit form of $\hat{\theta}_j$ $(j = 1, \ldots, m)$ is given as follows:

(1) *Case of $s_0^2 \neq 0$:*

$$\hat{\theta}_j = \begin{cases} \infty & (s_{a_*}^2 > z_j^2), \\ d_j/(z_j^2/s_{a_*}^2 - 1) & (s_{a_*}^2 \leq z_j^2), \end{cases} \tag{3.1}$$

where d_j, z_j, and s_α^2 are given by (2.2), (2.7), and (2.9), respectively, and the integer a_* is given by (2.10).

(2) *Case of $s_0^2 = 0$:* $^\forall h \in (0, t_1]$,

$$\hat{\theta}_j = d_j/(z_j^2/h - 1), \tag{3.2}$$

where t_j is given by (2.8). To minimize the covariance matrix of the GRR estimator, we define $h = t_1$. Hence

$$\hat{\theta}_j = \begin{cases} \infty & (z_j^2 = t_1), \\ d_j/(z_j^2/t_1 - 1) & (z_j^2 \neq t_1). \end{cases} \tag{3.3}$$

By using equation (3.1) or (3.3), we can obtain a closed form of the GRR estimator of $\boldsymbol{\beta}$ after optimizing $\boldsymbol{\theta}$ by GCV criterion. However, the expression is somewhat difficult to use in actual data analysis because equations (3.1) and (3.3) involve ∞. Hence, we give another expression of the GRR estimator after optimizing $\boldsymbol{\theta}$ by GCV criterion. Let $\boldsymbol{V}$ be an mth diagonal matrix defined by $\boldsymbol{V} = \mathrm{diag}(v_1, \ldots, v_m)$, where

$$v_j = \begin{cases} 0 \quad (s_{a_*}^2 > z_j^2); \quad 1 - s_{a_*}^2/z_j^2 \quad (s_{a_*}^2 \leq z_j^2), & (\text{when } s_0^2 \neq 0), \\ 0 \quad (t_1 = z_j^2); \quad 1 - t_1/z_j^2 \quad (t_1 \neq z_j^2), & (\text{when } s_0^2 = 0). \end{cases} \tag{3.4}$$

Then, the GRR estimator after optimizing $\boldsymbol{\theta}$ by GCV criterion is given by

$$\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{Q}_1 \boldsymbol{V} \boldsymbol{Q}_1' \hat{\boldsymbol{\beta}}, \tag{3.5}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator of $\boldsymbol{\beta}$ given by (2.4), and $\boldsymbol{Q}_1$ is a $k \times m$ matrix that consists of the first m columns of $\boldsymbol{Q}$, which is given by (2.1).

3.2. Relationships between the Optimal Solutions of GCV and the Generalized C_p Criteria

When $s_0^2 \neq 0$, C_p and the modified C_p (MC_p; Yanagihara *et al.*, 2009) criteria can be defined. Their optimal solutions are also given by closed forms, and they are unified as solutions of the minimization problem of the following generalized C_p (GC_p) criterion:

$$GC_p(\boldsymbol{\theta}|\lambda) = (\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}) + 2\lambda \mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M}),$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{\theta}}$ is the predictor of $\boldsymbol{y}$ given by (2.5) (originally, the GC_p criterion for the model (1.1) was proposed by Atkinson, 1980). Solutions of $GC_p(\boldsymbol{\theta}|\lambda)$ with $\lambda = s_0^2$ and $c_{\mathrm{M}} s_0^2$ correspond to those of C_p and MC_p criteria, respectively, where $c_{\mathrm{M}} = 1 + 2/(n - m - 3)$. Since it follows from Lemma 2 that $GC_p(\boldsymbol{\theta}|\lambda)$ is invariant to any changes in $\theta_{m+1}, \ldots, \theta_k$, we take $\theta_{m+1} = \cdots = \theta_k = \infty$ for simplicity as well as the minimization of the GCV criterion. By extending the result in Nagai *et al*. (2012), the optimal solutions of $GC_p(\boldsymbol{\theta}|\lambda)$ are given by

$$\hat{\theta}_j(\lambda) = \begin{cases} \infty & (\lambda > z_j^2), \\ d_j/(z_j^2/\lambda - 1) & (\lambda \leq z_j^2). \end{cases} \tag{3.6}$$

By comparing (3.1) with (3.6), it is clear that the optimal solutions of GCV criterion are a special case of those of GC_p criterion with $\lambda = s_{a_*}^2$. Suppose that $\lambda_1 \leq \lambda_2$. Then it is easy to see that $\hat{\theta}_j(\lambda_1) \leq \hat{\theta}_j(\lambda_2)$. Notice that $c_{\mathrm{M}} > 1$ holds. Moreover, from Lemma 1 (1), $s_{a_*}^2 \leq s_0^2$ holds. Consequently, the following theorem is derived:

Theorem 2. *The optimal solutions of GCV criterion can be regarded as the special case of those of GC_p criterion with $\lambda = s_{a_*}^2$, where a_* is the integer defined by* (2.10). *Let $\hat{\theta}_j^{(\mathrm{C})}$ and $\hat{\theta}_j^{(\mathrm{M})}$* ($j = 1, \ldots, m$) *be optimal solutions of C_p and MC_p criteria, respectively, when $s_0^2 \neq 0$. Then, the following inequality always holds:*

$$\hat{\theta}_j \leq \hat{\theta}_j^{(\mathrm{C})} \leq \hat{\theta}_j^{(\mathrm{M})}.$$

Theorem 2 indicates that even though GCV criterion does not require an estimator of σ^2, it estimates σ^2 automatically by $s_{a_*}^2$. Furthermore, $s_{a_*}^2$ always underestimates σ^2. This results in less shrinkage of the OLS estimator with the GRR optimized by GCV criterion than it does by C_p criterion or MC_p criterion.

Additionally, we consider choosing a threshold value λ in (3.6) by minimizing the $\mathrm{GCV}(\hat{\boldsymbol{\theta}}(\lambda))$, where $\hat{\boldsymbol{\theta}}(\lambda) = (\hat{\theta}_1(\lambda), \ldots, \hat{\theta}_m(\lambda), \infty, \ldots, \infty)'$, and $\hat{\theta}_j(\lambda)$ is given by (3.6). It is obviously that $\min_{\boldsymbol{\theta} \in \mathbb{R}_+^k} \mathrm{GCV}(\boldsymbol{\theta}) \leq \min_{\lambda \in \mathbb{R}_+} \mathrm{GCV}(\hat{\boldsymbol{\theta}}(\lambda))$. From Theorem 1, the ridge parameters that minimize $\mathrm{GCV}(\boldsymbol{\theta})$ can be expressed as $\hat{\boldsymbol{\theta}}(a_*)$. Hence, we derive the following theorem:

Theorem 3. *An explicit solution to the minimization problem of* $\mathrm{GCV}(\hat{\boldsymbol{\theta}}(\lambda))$ *can be obtained as*

$s^2_{a_*}$, *i.e.*,

$$s^2_{a_*} = \arg\min_{\lambda \in \mathbb{R}_+} \mathrm{GCV}(\hat{\boldsymbol{\theta}}(\lambda)).$$

Theorem 3 indicates that the GRR with $\boldsymbol{\theta}$ optimized by the GCV criterion is equivalent to the GRR with $\boldsymbol{\theta}$ optimized by GC_p criterion after choosing the threshold value λ by GCV criterion.

3.3. Generalized Degrees of Freedom in the Optimized GRR

In this subsection, we derive an estimate for the generalized degrees of freedom (GDF), as proposed by Ye (1998), for the GRR after optimizing $\boldsymbol{\theta}$ by GCV criterion under the normal distributed assumption. Suppose that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$. From Efron (2004), the GDF of the GRR after optimizing $\boldsymbol{\theta}$ is given by

$$\gamma = E\left[\sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial y_i}\right],$$

where $\hat{\mu}_i$ $(i = 1, \dots, n)$ is the ith element of $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{\theta}}} = \bar{y}\mathbf{1}_n + \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}}$, and $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}}$ is the GRR estimator of $\boldsymbol{\beta}$ after optimizing GCV, which is given by (3.5). Hence, we can see that the GDF is estimated by $\hat{\gamma} = \sum_{i=1}^n \partial\hat{\mu}_i/\partial y_i$. After a simple calculation, we obtain the explicit form of $\hat{\gamma}$ as in the following theorem (the proof is given in Appendix E):

Theorem 4. *Suppose that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n)$. Let $w_j = I(v_j \neq 0)$ $(j = 1, \dots,)$, where $I(x \neq 0)$ is the indicator function, i.e., $I(x \neq 0) = 1$ if $x \neq 0$ and $I(x \neq 0) = 0$ if $x = 0$, $\boldsymbol{V} = \mathrm{diag}(v_1, \dots, v_m)$ is given by (3.4), and let $\boldsymbol{W}$ be an mth diagonal matrix whose jth diagonal element is w_j. Then, an estimator of the GDF is derived as*

$$\hat{\gamma} = 1 + 2\mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{V}). \tag{3.7}$$

In particular $\mathrm{tr}(\boldsymbol{W}) = m - a_*$ holds when $s_0^2 \neq 0$ and $\mathrm{tr}(\boldsymbol{W}) = m - 1$ holds when $s_0^2 = 0$, where the integer a_* is given by (2.10).

4. Application to the Case of High-Dimensional Explanatory Variables

4.1. Principle Component Regression Hybridized with the GRR

In this section, we consider the case of high-dimensional explanatory variables, i.e., the case of $n \leq k$, which has been studied by, e.g., Srivastava and Kubokawa (2007), and Fan and Lv (2010). In this paper, the case of $m = n - 1$ is considered. Even when $m = n - 1$, GRR can work, and the optimal solutions of GCV criterion can be obtained by the closed forms, as in Theorem 1. However, it seems from Theorem 1 that the optimal $\boldsymbol{\theta}_1$ will become very small. Thus, there is a possibility that GRR cannot work effectively. In order to avoid such a risk, we apply GRR to a regression model in which the various small singular values of $\boldsymbol{X}$ are eliminated, i.e., the GRR is applied to a principal component regression (PCR; see, e.g., Draper & Smith, 1981, chap. 6.9; Liu *et al*., 2003). Let $\boldsymbol{D}_r = (d_1, \dots, d_r)$ $(r < m)$ be a rth diagonal matrix, where d_j is the jth largest eigenvalue of $\boldsymbol{X}'\boldsymbol{X}$ defined by (2.2), and let $\boldsymbol{X}_r$ be an $n \times k$ matrix defined by

$$\boldsymbol{X}_r = \boldsymbol{P} \begin{pmatrix} \boldsymbol{D}_r^{1/2} & \boldsymbol{O}_{r,k-r} \\ \boldsymbol{O}_{n-r,r} & \boldsymbol{O}_{n-r,k-r} \end{pmatrix} \boldsymbol{Q}'.$$

After eliminating $m - r$ principal components and replacing $\boldsymbol{X}$ with $\boldsymbol{X}_r$, the reduced model, called the r-PCR model, can be expressed. It is equivalent to the following liner regression model:

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{X}_r \boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{4.1}$$

We know that a predictor of $\boldsymbol{y}$ derived from the model (4.1) with $r = m$ corresponds to $\boldsymbol{y}$. Thus, we do not consider the case of $r = m$. Let $\mathrm{GCV}(\boldsymbol{\theta}|r)$ be the GCV criterion for selecting $\boldsymbol{\theta}_r$ in the r-PCR model (4.1) to which the GRR is applied, and let $\hat{\boldsymbol{\theta}}_r$ be the minimizer of $\mathrm{GCV}(\boldsymbol{\theta}|r)$. Then, $\hat{\boldsymbol{\theta}}_r$ can be also obtained in closed form from Theorem 1.

The most important choice in PCR is to determine how many singular values are eliminated, i.e., it is important to choose the optimal r. We can use the estimate of the GDF calculated in Theorem 4 with the new GCV criterion for selecting r for the PCR hybridized with the GRR. For the r-PCR model (4.1) derived from the GRR after optimizing $\boldsymbol{\theta}_r$, let $\hat{\boldsymbol{y}}_{r,\hat{\boldsymbol{\theta}}_r}$ be a predictor of $\boldsymbol{y}$ and let $\hat{\gamma}_r$ be the estimator of GDF. As in Ye (1998), we propose a new GCV criterion for selecting r as

$$\mathrm{GCV}^{\#}(r) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}}_{r,\hat{\boldsymbol{\theta}}_r})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{r,\hat{\boldsymbol{\theta}}_r})}{n(1 - \hat{\gamma}_r/n)^2}. \tag{4.2}$$

Unfortunately, there is a possibility that $1 - \hat{\gamma}_r/n \le 0$, in which case, we reject r. Let $\mathcal{S}$ be a set of integers defined by $\mathcal{S} = \{r \in \{0, 1, \ldots, m-1\}|\ 1 - \hat{\gamma}_r/n > 0\}$. Then, an optimal r is found by minimizing the GCV criterion in (4.2) is as follows:

$$\hat{r} = \arg\min_{r \in \mathcal{S}} \mathrm{GCV}^{\#}(r).$$

4.2. Numerical Study

We evaluated the proposed method by applying it to data from $N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n)$, where $\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{X}_0 \boldsymbol{\Phi}(\rho)^{1/2}$ and $\boldsymbol{\beta} = \boldsymbol{M}^+ \boldsymbol{X}' \boldsymbol{\eta}$. Here, $\boldsymbol{X}_0$ is an $n \times k$ matrix whose elements are identically and independently distributed according to $U(-1, 1)$, $\boldsymbol{\Phi}(\rho)$ is a $k \times k$ symmetric matrix whose (a, b)th element is $\rho^{|a-b|}$, and $\boldsymbol{\eta}$ is an n-dimensional vector whose jth element is given by

$$\sqrt{\frac{12n(n-1)}{4n^2 + 6n - 1}} \left\{ (-1)^{j-1} \left(1 - \frac{j-1}{n} \right) - \frac{1}{2n} \right\}.$$

In this setting, it should be emphasized that $||\boldsymbol{\beta}||$ does not become large even when k is increased. If $||\boldsymbol{\beta}||$ becomes large as k is increased, a value close to $\mathbf{0}_m$ is frequently chosen as the optimal $\boldsymbol{\theta}$. Needless to say, such a situation is meaningless in applications of GRR. Therefore, we avoid such a situation by controlling the elements of $\boldsymbol{\beta}$.

The following three methods were applied to simulated data:

Method 1: ordinary GRR (GRR with all of the principle components).

Method 2: PCR hybridized with GRR (i.e., the proposed method).

Table 1. MSEs of coefficients and a predictor in each method

			MSE of Coefficients (%)			MSE of Predictor (%)		
n	k	ρ	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
20	20	0.80	95.69	29.28	29.30	98.72	96.39	101.82
		0.90	98.10	29.20	29.22	98.77	95.19	101.80
		0.99	100.63	29.14	29.17	99.10	94.71	101.85
	40	0.80	98.06	63.29	63.87	98.93	94.63	100.27
		0.90	99.25	63.35	63.37	99.73	95.22	100.90
		0.99	97.84	63.32	63.08	98.60	95.44	100.86
	100	0.80	99.15	93.76	94.13	99.04	95.41	99.94
		0.90	99.01	96.60	97.15	99.12	97.20	102.77
		0.99	99.03	99.33	100.02	98.80	96.79	103.60
	200	0.80	98.55	87.53	90.82	98.32	92.20	98.91
		0.90	98.92	85.91	91.02	98.55	90.30	101.48
		0.99	98.88	86.48	92.37	98.51	87.80	101.35
50	50	0.80	100.24	77.08	77.06	99.69	96.51	99.33
		0.90	100.91	77.42	77.29	99.97	97.01	99.38
		0.99	100.15	77.32	77.51	99.81	97.36	99.08
	100	0.80	100.30	75.76	74.62	99.57	90.22	89.28
		0.90	99.94	75.61	73.96	99.77	91.39	92.40
		0.99	100.08	75.15	73.79	99.84	86.65	92.24
	250	0.80	99.72	78.76	77.62	99.69	86.64	88.70
		0.90	99.90	78.83	77.80	99.74	91.84	95.19
		0.99	100.26	78.52	77.90	100.02	89.30	98.42
	500	0.80	99.46	86.05	87.77	99.59	92.65	97.63
		0.90	99.56	87.51	89.43	99.47	91.99	98.90
		0.99	99.68	90.53	92.37	99.88	95.59	100.85

Method 3: ordinary PCR (PCR without GRR) with an optimal r ($r = 0, 1, \ldots, m-1$) chosen by minimizing GCV criterion as

$$\mathrm{GCV}^{\#}_{\mathrm{p}}(r) = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}}_r)'(\boldsymbol{y} - \hat{\boldsymbol{y}}_r)}{n\{1-(1+r)/n\}^2},$$

where $\hat{\boldsymbol{y}}_r = \{\boldsymbol{J}_n + \boldsymbol{X}_r\boldsymbol{M}_r^+\boldsymbol{X}_r'\}\boldsymbol{y}$.

Let $\hat{\boldsymbol{\beta}}_j$ be an estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{y}}_j$ be a predictor of $\boldsymbol{y}$, as derived from Method j ($j = 1, 2, 3$). We compared the following two characteristics of each method, based on 10,000 iterations:

- MSE of coefficients (%): $100 \times E[(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta})]/\mathrm{tr}(\boldsymbol{M}^+)$, $\mathrm{tr}(\boldsymbol{M}^+)$ is the MSE of the OLS estimator of $\boldsymbol{\beta}$.
- MSE of predictor (%): $100 \times E[(\hat{\boldsymbol{y}}_j - \boldsymbol{X}\boldsymbol{\beta})'(\hat{\boldsymbol{y}}_j - \boldsymbol{X}\boldsymbol{\beta})]/(n-1)$, where $(n-1)$ is the MSE of a predictor of $\boldsymbol{y}$ derived from the OLS estimation.

Table 1 shows the two characteristics for $n = 20, 100$, $k = n, 2n, 5n, 10n$ and $\rho = 0.8, 0.9, 0.99$. When the characteristic is less than 100, it means that the method used improved the performance of the OLS estimation, as measured by the MSE. From the table, we can see that in most cases and for both MSEs Method 2 resulted in the smallest (best) values. Those of Method 1 were the worst. These results indicate that GRR does not work effectively when k is larger than n. If PCR is used instead of GRR, although the result is improved, it is still insufficient. Using GRR and PCR simultaneously is expected to improve the results more than using either one alone.

Acknowledgment The author thanks Prof. Hirofumi Wakaki, Hiroshima University, for helpful comments on the proof of the uniqueness of the solution. This research was supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Challenging Exploratory Research, #22650058, 2010-2012.

References

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

Batah, F. S. M., Ramanathan, T. V. & Gore, S. D. (2008). The efficiency of modified jackknife and ridge type regression estimators: a comparison. *Surv. Math. Appl.*, **3**, 111–122.

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.

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis* (2nd. ed.). John Wiley & Sons, Inc., New York.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, **99**, 619–632.

Fan, J. & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Stat. Sinica*, **20**, 101–148.

Golub, G. H., Heath, M. & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Gu, C. & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, **12**, 383–398.

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

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

Jensen, D. R. & Ramirez, D. E. (2010). Surrogate models in ill-conditioned systems. *J. Statist. Plann. Inference*, **140**, 2069–2077.

Jimichi, M. (2008). Exact moments of feasible generalized ridge regression estimator and numerical evaluations. *J. Japanese Soc. Comput. Statist.*, **21**, 1–20.

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

Liu, R. X., Kuang, J., Gong, Q. & Hou, X. L. (2003). Principal component regression analysis with SPSS. *Comput. Meth. Prog. Bio.*, **71**, 141–147.

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

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

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.

Smyth, R., Narayan, P. K. & Shi, H. (2011). Substitution between energy and classical factor inputs in the Chinese steel sector. *Appl. Energ.*, *88*, 361–367.

Srivastava, M. S. & Kubokawa, T. (2007). Empirical Bayes regression analysis with many regressors but fewer observations. *J. Statist. Plann. Inference*, **137**, 3778–3792.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc. Ser*. **B**, **36**, 111–147.

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

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. Ser*. **B**, **62**, 413–428.

Yanagihara, H. (2012). A non-iterative optimization method for smoothness in penalized spline regression. *Stat. Comput.*, **22**, 527–544.

Yanagihara, H. & Satoh, K. (2010). An unbiased C_p criterion for multivariate ridge regression. *J. Multivariate Anal.*, **101**,

1226–1238.
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).
Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, **93**, 120–131.

Appendix

A. Proof of Lemma 1

In order to prove Lemma 1 (1), we show that if the integer a_* in (2.10) exists, it is unique. Later, we will use reductio ad absurdum to prove the existence of the integer a_*. Notice that the following equation is satisfied for any integers $\alpha \in \{0, 1, \dots, m-1\}$:

$$s_{\alpha+1}^2 = \frac{(n-m-1+\alpha)s_\alpha^2 + t_{\alpha+1}}{n-m+\alpha} = \frac{n-m-1+\alpha}{n-m+\alpha}(s_\alpha^2 - t_{\alpha+1}) + t_{\alpha+1},$$

where t_j and s_α^2 are given by (2.8) and (2.9), respectively. This implies that

$$s_{\alpha+1}^2 - t_{\alpha+1} = \frac{n-m-1+\alpha}{n-m+\alpha}(s_\alpha^2 - t_{\alpha+1}) \;\; ({}^\forall\alpha \in \{0, 1, \dots, m-1\}).$$

From the above equation, we can see that the following statements are true:

$$s_\alpha^2 - t_{\alpha+1} \le 0 \Rightarrow s_{\alpha+1}^2 - t_{\alpha+1} \le 0, \quad s_\alpha^2 - t_\alpha > 0 \Rightarrow s_{\alpha-1}^2 - t_\alpha > 0. \tag{A.1}$$

Moreover, the following statements are also true because $t_1 \le \cdots \le t_m$ holds:

$$s_\alpha^2 - t_\alpha \le 0 \Rightarrow s_\alpha^2 - t_{\alpha+1} \le 0, \quad s_\alpha^2 - t_{\alpha+1} > 0 \Rightarrow s_\alpha^2 - t_\alpha > 0. \tag{A.2}$$

Suppose that an integer a_* exists. Combining (A.1) and (A.2) yields

$$s_{a_*}^2 - t_{a_*+1} \le 0 \Rightarrow s_{a_*+1}^2 - t_{a_*+1} \le 0 \Rightarrow s_{a_*+1}^2 - t_{a_*+2} \le 0 \Rightarrow \cdots \Rightarrow s_m^2 - t_m \le 0,$$

and

$$s_{a_*}^2 - t_{a_*} > 0 \Rightarrow s_{a_*-1}^2 - t_{a_*} > 0 \Rightarrow s_{a_*-1}^2 - t_{a_*-1} > 0 \Rightarrow \cdots \Rightarrow s_0^2 - t_1 > 0.$$

Hence, we find

$$s_\alpha^2 \le t_\alpha \;\; ({}^\forall\alpha \in \{a_*+1, \dots, m\}), \quad s_\alpha^2 > t_{\alpha+1} \;\; ({}^\forall\alpha \in \{0, 1, \dots, a_*-1\}).$$

These equations indicate that $s_\alpha^2 \notin R_\alpha$ when $\alpha \ne a_*$, where R_α is given by (2.11). Consequently, the integer a_* is uniquely determined if a_* exists. Next we show the existence of the integer a_*. Since $R_\alpha^c = (0, t_\alpha] \bigcup (t_{\alpha+1}, \infty)$, we can see that the following statement is true:

$$\{s_\alpha^2 - t_\alpha > 0\} \bigcap \{s_\alpha^2 \notin R_\alpha\} \Rightarrow s_\alpha^2 - t_{\alpha+1} > 0. \tag{A.3}$$

Suppose that the integer a_* does not exist, i.e., $s_\alpha^2 \notin R_\alpha$ holds ${}^\forall\alpha = \{0, 1, \dots, m\}$. This implies that

$s_0^2 > t_1$. Combining (A.1) and (A.3) yields

$$s_0^2 - t_1 > 0 \Rightarrow s_1^2 - t_1 > 0 \Rightarrow s_1^2 - t_2 > 0 \Rightarrow \cdots \Rightarrow s_m^2 - t_m > 0.$$

However, $s_m^2 - t_m > 0$ contradicts the assumption $s_m^2 \notin R_m$. Consequently, by reductio ad absurdum, the integer a_* exists.

Next, we derive an upper bound for $s_{a_*}^2$. Let $x_1 = \cdots = x_{n-m-1} = s_0^2$ and $x_{n-m-1+j} = t_j$ ($j = 1, \dots, m$). Then s_α^2 is regarded as the sample mean of $x_1, \dots, x_{n-m-1+\alpha}$. It follows from a property of the sample mean that

$$s_\alpha^2 \le \max_{j \in \{1,\dots,n-m-1+\alpha\}} x_j = \max\{s_0^2, t_\alpha\} \ \ ({}^\forall \alpha \in \{1, \dots, m\}). \tag{A.4}$$

Since $s_0^2 > 0$ and $\bigcup_{j=0}^m R_j = (0, \infty]$ hold, an integer $b \in \{0, 1, \dots, m\}$ exists such that $s_0 \in R_b$. When $b = m$, it follows from the inequality $s_0^2 > t_m$ and (A.4) that

$$s_\alpha^2 \le \max\{s_0^2, t_\alpha\} \le \max\{s_0^2, t_m\} = s_0^2 \ \ ({}^\forall \alpha \in \{1, \dots, m\}).$$

When $b \le m - 1$, inequalities $s_0^2 \le t_\alpha \ {}^\forall \alpha \in \{b + 1, \dots, m\}$ and $s_0^2 > t_\alpha \ {}^\forall \alpha \in \{1, \dots, b\}$ are satisfied, because $s_0^2 \in R_b$. It follows from these results and (A.4) that

$$s_\alpha^2 \le \max\{s_0^2, t_\alpha\} = \begin{cases} t_\alpha & ({}^\forall \alpha \in \{b + 1, \dots, m\}), \\ s_0^2 & ({}^\forall \alpha \in \{1, \dots, b\}). \end{cases} \tag{A.5}$$

The upper equation on the right side of (A.5) indicates that $s_\alpha^2 \notin R_\alpha$ holds ${}^\forall \alpha \in \{b+1, \dots, m\}$. Hence it seems that the integer a_* is less than or equal to b. This result and the lower equation on the right side of (A.5) lead us to the conclusion that $s_{a_*}^2 \le s_0^2$.

Finally, we give the proof of Lemma 1 (2). When $s_0^2 = 0$, s_α^2 is expressed as the sample mean of $t_1, \dots, t_m$, i.e., $s_\alpha^2 = \alpha^{-1} \sum_{j=1}^\alpha t_j$ ($\alpha = 1, \dots, m$). It is clear that $s_0^2 \notin R_0$. Moreover, from a property of the sample mean and the inequality $t_1 \le \cdots \le t_m$, we derive

$$s_\alpha^2 \le \max_{j \in \{1,\dots,m\}} t_j = t_\alpha \ \ ({}^\forall \alpha \in \{1, \dots, m\}).$$

The above equation indicates that $s_\alpha^2 \notin R_\alpha$ holds ${}^\forall \alpha \in \{1, \dots, m\}$. Therefore, Lemma 1 (2) is proved.

B. Proof of Lemma 2

Let $\boldsymbol{P}$ be an nth orthogonal matrix that diagonalizes $\boldsymbol{X}\boldsymbol{X}'$ as

$$\boldsymbol{P}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{P} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{n-m,m} & \boldsymbol{O}_{n-m,n-m} \end{pmatrix}, \tag{B.1}$$

where $\boldsymbol{D}$ is an mth diagonal matrix given by (2.2). The singular value decomposition of $\boldsymbol{X}$ is expressed as

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

where $\boldsymbol{Q}$ is given by (2.1). Let $\boldsymbol{\Theta}_1 = \mathrm{diag}(\theta_1, \ldots, \theta_m)$ and $\boldsymbol{\Theta}_2 = \mathrm{diag}(\theta_{m+1}, \ldots, \theta_k)$. It follows from (B.2) that

$$\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{Q}\begin{pmatrix} (\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}^{1/2} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{k-m,m} & \boldsymbol{O}_{k-m,n-m} \end{pmatrix}\boldsymbol{P}'\boldsymbol{y}. \tag{B.3}$$

Moreover, the equations (B.2) and (B.3) imply that

$$\boldsymbol{X}\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}' = \boldsymbol{P}\begin{pmatrix} \boldsymbol{D}^{1/2}(\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}^{1/2} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{n-m,m} & \boldsymbol{O}_{n-m,n-m} \end{pmatrix}\boldsymbol{P}'. \tag{B.4}$$

The results in (B.3) and (B.4) indicate that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ in (2.3) and $\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M})$ in (2.6) are independent of $\boldsymbol{\Theta}_2$. Consequently, Lemma 2 is proved.

C. Proof of Lemma 3

Let $\boldsymbol{u}$ be an n-dimensional vector derived by centralizing $\boldsymbol{y}$, i.e., $\boldsymbol{u} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{y}$. Moreover, let us decompose $\boldsymbol{P}$ in (B.1) to

$$\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2), \tag{C.1}$$

where $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are $n \times m$ and $n \times (n-m)$ matrices, respectively. It follows from the equation $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$ and (B.2) that

$$\boldsymbol{P}_1'\boldsymbol{u} = (\boldsymbol{D}^{-1/2}, \boldsymbol{O}_{m,k-m})\boldsymbol{Q}'\boldsymbol{Q}\begin{pmatrix} \boldsymbol{D}^{1/2} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{k-m,m} & \boldsymbol{O}_{k-m,n-m} \end{pmatrix}\boldsymbol{P}'\boldsymbol{u} = (\boldsymbol{D}^{-1/2}, \boldsymbol{O}_{m,k-m})\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{y}.$$

Since $\boldsymbol{P}_1'\boldsymbol{u}$ is equal to $(z_1, \ldots, z_m)'$ in (2.7), we write the following n-dimensional vector as $\boldsymbol{z}$:

$$\boldsymbol{z} = (z_1, \ldots, z_n)' = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{P}_1'\boldsymbol{u} \\ \boldsymbol{P}_2'\boldsymbol{u} \end{pmatrix}. \tag{C.2}$$

Notice that $\boldsymbol{P}_2\boldsymbol{P}_2' = \boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}^{+}\boldsymbol{X}'$ and $\boldsymbol{X}'\boldsymbol{J}_n = \boldsymbol{O}_{k,n}$. Thus, we have

$$\boldsymbol{z}_2'\boldsymbol{z}_2 = \boldsymbol{u}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}^{+}\boldsymbol{X})\boldsymbol{u} = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}^{+}\boldsymbol{X})(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{y} = (n-m-1)s_0^2, \tag{C.3}$$

where s_0^2 is given by (2.9). By using the equation $\boldsymbol{X}'\boldsymbol{1}_n = \boldsymbol{0}_k$, and (B.4) and (C.3), the residual sum of squares in (2.6) can be rewritten as

$$\begin{aligned}
(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}) &= \boldsymbol{u}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}')^2\boldsymbol{u} \\
&= \boldsymbol{u}'\boldsymbol{P}\left\{\boldsymbol{I}_n - \begin{pmatrix} \boldsymbol{D}^{1/2}(\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}^{1/2} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{n-m,m} & \boldsymbol{O}_{n-m,n-m} \end{pmatrix}\right\}^2 \boldsymbol{P}'\boldsymbol{u} \\
&= \boldsymbol{z}_1'\{\boldsymbol{I}_m - \boldsymbol{D}^{1/2}(\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}^{1/2}\}^2\boldsymbol{z}_1 + \boldsymbol{z}_2'\boldsymbol{z}_2 \\
&= (n-m-1)s_0^2 + \sum_{j=1}^{m}\left(\frac{\theta_j}{d_j+\theta_j}\right)^2 z_j^2.
\end{aligned} \tag{C.4}$$

Moreover, from (B.4), $\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M})$ can be rewritten as

$$\mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^{-1}\boldsymbol{M}) = \mathrm{tr}\left\{\begin{pmatrix} \boldsymbol{D}^{1/2}(\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}^{1/2} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{n-m,m} & \boldsymbol{O}_{n-m,n-m} \end{pmatrix}\right\}$$
$$= \mathrm{tr}\left\{(\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}\right\} = m - \sum_{j=1}^{m}\left(\frac{\theta_j}{d_j+\theta_j}\right). \tag{C.5}$$

By substituting (C.4) and (C.5) into (2.6), GCV($\boldsymbol{\theta}$) is expressed as (2.12).

D. Proof of Theorem 1

Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)'$ be an m-dimensional vector whose jth element $\delta_j \in [0,1]$ ($j = 1, \ldots, m$) is defined by

$$\delta_j = \frac{\theta_j}{d_j+\theta_j}.$$

From Lemma 3, $g(\boldsymbol{\theta}_1)$ in (2.12) is expressed as the following function with respect to $\boldsymbol{\delta}$:

$$g(\boldsymbol{\theta}_1) = f(\boldsymbol{\delta}) = \frac{r(\boldsymbol{\delta})}{c(\boldsymbol{\delta})^2}, \tag{D.1}$$

where

$$r(\boldsymbol{\delta}) = \frac{1}{n}\left\{(n-m-1)s_0^2 + \sum_{j=1}^{m}\delta_j^2 z_j^2\right\}, \quad c(\boldsymbol{\delta}) = 1 - \frac{1}{n}\left\{m+1-\sum_{j=1}^{m}\delta_j\right\},$$

and z_j and s_0^2 are given by (2.7) and (2.9), respectively. Let $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_m)'$ be a minimizer of $f(\boldsymbol{\delta})$ in (D.1), i.e.,

$$\hat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}\in[0,1]^m} f(\boldsymbol{\delta}),$$

where $[0,1]^m$ is the mth Cartesian power of the set $[0,1]$. Notice that

$$\frac{\partial}{\partial\delta_\alpha} f(\boldsymbol{\delta}) = \frac{2}{nc(\boldsymbol{\delta})^3}\left\{c(\boldsymbol{\delta})\delta_\alpha z_\alpha^2 - r(\boldsymbol{\delta})\right\}.$$

Hence, we find that a necessary condition of $\hat{\boldsymbol{\delta}}$ is

$$\hat{\delta}_j = \begin{cases} 1 & (\text{if } h(\hat{\boldsymbol{\delta}}) > z_j^2), \\ h(\hat{\boldsymbol{\delta}})/z_j^2 & (\text{if } h(\hat{\boldsymbol{\delta}}) \le z_j^2), \end{cases} \tag{D.2}$$

where $h(\hat{\boldsymbol{\delta}}) = r(\hat{\boldsymbol{\delta}})/c(\hat{\boldsymbol{\delta}})$.

Suppose that $h(\hat{\boldsymbol{\delta}}) \in R_a$, where $a \in \{0, 1, \ldots m\}$, and R_α is a range defined by (2.11). Then the equation (D.2) leads us to the result that $\hat{\delta}_j = 1$ when $j \in \mathcal{J}_a = \{j \in \{1, \ldots, m\}|\ z_j^2 \le t_a\}$ and $\hat{\delta}_j = h(\hat{\boldsymbol{\delta}})/z_j^2$ when $j \in \mathcal{J}_a^c = \{j \in \{1, \ldots, m\}|\ z_j^2 \ge t_{a+1}\}$, where t_j is given by (2.8). Notice that

$$\sum_{j=1}^{m}\hat{\delta}_j = \sum_{j\in\mathcal{J}_a} 1 + \sum_{j\in\mathcal{J}_a^c}\frac{h(\hat{\boldsymbol{\delta}})}{z_j^2} = a + h(\hat{\boldsymbol{\delta}})\sum_{j=a+1}^{m}\frac{1}{t_j},$$
$$\sum_{j=1}^{m}\hat{\delta}_j^2 z_j^2 = \sum_{j\in\mathcal{J}_a} z_j^2 + \sum_{j\in\mathcal{J}_a^c}\frac{h(\hat{\boldsymbol{\delta}})^2}{z_j^4}z_j^2 = \sum_{j=1}^{a} t_j + h(\hat{\boldsymbol{\delta}})^2\sum_{j=a+1}^{m}\frac{1}{t_j}.$$

These imply

$$r(\hat{\boldsymbol{\delta}}) = \frac{1}{n}\left\{(n-m-1+a)s_a^2 + h(\hat{\boldsymbol{\delta}})^2 \sum_{j=a+1}^{m} \frac{1}{t_j}\right\}, \quad c(\hat{\boldsymbol{\delta}}) = \frac{1}{n}\left\{n-m-1+a+h(\hat{\boldsymbol{\delta}}) \sum_{j=a+1}^{m} \frac{1}{t_j}\right\},$$

where s_α^2 is given by (2.9). It follows from the above equation and the definition of $h(\boldsymbol{\delta})$ that

$$h(\hat{\boldsymbol{\delta}}) = \frac{(n-m-1+a)s_a^2 + h(\hat{\boldsymbol{\delta}})^2 \sum_{j=a+1}^{m} 1/t_j}{n-m-1+a+h(\hat{\boldsymbol{\delta}}) \sum_{j=a+1}^{m} 1/t_j}.$$

By solving the above equation, an explicit form of $h(\hat{\boldsymbol{\delta}})$ is given as

$$h(\hat{\boldsymbol{\delta}}) = \begin{cases} s_a^2; & \text{(when } s_0^2 \neq 0) \\ {}^\forall h \in (0, t_1] \quad (a=0); \quad s_a^2 \quad (a=1,\ldots,m); & \text{(when } s_0^2 = 0). \end{cases}$$

From Lemma 1, we find that the integer $a \in \{0, 1, \ldots m\}$ such that $s_a^2 \in R_a$ is uniquely determined as a_* when $s_0^2 \neq 0$, where a_* is defined by (2.10), and the integer $a \in \{1, \ldots m\}$ such that $s_a^2 \in R_a$ does not exist when $s_0^2 = 0$. Therefore, we derive

$$h(\hat{\boldsymbol{\delta}}) = \begin{cases} s_{a_*}^2 & (s_0^2 \neq 0), \\ {}^\forall h \in (0, t_1] & (s_0^2 = 0). \end{cases} \tag{D.3}$$

Recall that $\hat{\delta}_j = \hat{\theta}_j/(d_j + \hat{\theta}_j)$. By using (D.2) and (D.3), the equations (3.1) and (3.2) are obtained.

Finally, from the same calculation as in (B.4), the covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ is derived as

$$Cov[\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}] = \sigma^2 \boldsymbol{M}_{\boldsymbol{\theta}}^{-1} \boldsymbol{M} \boldsymbol{M}_{\boldsymbol{\theta}}^{-1} = \boldsymbol{Q} \begin{pmatrix} (\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1}\boldsymbol{D}(\boldsymbol{D}+\boldsymbol{\Theta}_1)^{-1} & \boldsymbol{O}_{m,k-m} \\ \boldsymbol{O}_{k-m,m} & \boldsymbol{O}_{k-m,k-m} \end{pmatrix} \boldsymbol{Q}'.$$

The equation indicates that a larger θ_j reduces the covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$. Since the largest h is t_1, equation (3.3) is obtained.

E. Proof of Theorem 4

Since $\boldsymbol{V}$ given in (3.4) and $\boldsymbol{D}$ given in (2.2) are diagonal matrices, $\boldsymbol{D}^{1/2}\boldsymbol{V}\boldsymbol{D}^{-1/2} = \boldsymbol{V}$ holds. By using this result, the definition of $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}}$ in (3.5), and the singular value decomposition of $\boldsymbol{X}$ in (B.2), we derive

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}} = \boldsymbol{P} \begin{pmatrix} \boldsymbol{D}^{1/2} & \boldsymbol{O}_{m,k-m} \\ \boldsymbol{O}_{n-m,m} & \boldsymbol{O}_{n-m,k-m} \end{pmatrix} \boldsymbol{Q}'\boldsymbol{Q}_1\boldsymbol{V}\boldsymbol{Q}_1'\boldsymbol{Q} \begin{pmatrix} \boldsymbol{D}^{-1/2} & \boldsymbol{O}_{m,n-m} \\ \boldsymbol{O}_{k-m,m} & \boldsymbol{O}_{k-m,n-m} \end{pmatrix} \boldsymbol{P}'\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_1'\boldsymbol{y},$$

where $\boldsymbol{P}_1$ is given by (C.1). This equation leads to another expression of the predictor of $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{\theta}}}$ as

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{\theta}}} = (\boldsymbol{J}_n + \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_1')\boldsymbol{y}.$$

It follows from the above equation and the result $\boldsymbol{P}_1'\boldsymbol{P}_1 = \boldsymbol{I}_m$ that

$$\begin{aligned}\hat{\gamma} &= \frac{\partial}{\partial \boldsymbol{y}'}(\boldsymbol{J}_n + \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_1')\boldsymbol{y} = \mathrm{tr}(\boldsymbol{J}_n + \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_1') + \sum_{i=1}^{n} \boldsymbol{e}_i'\boldsymbol{P}_1\left(\frac{\partial}{\partial y_i}\boldsymbol{V}\right)\boldsymbol{P}_1'\boldsymbol{y} \\ &= 1 + \mathrm{tr}(\boldsymbol{V}) + \sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{e}_i'\boldsymbol{p}_j\left(\frac{\partial v_j}{\partial y_i}\right)\boldsymbol{p}_j'\boldsymbol{y} = 1 + \mathrm{tr}(\boldsymbol{V}) + \sum_{j=1}^{m}\left(\frac{\partial v_j}{\partial \boldsymbol{y}'}\right)\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y}, \end{aligned} \tag{E.1}$$

where $\boldsymbol{e}_i$ is an n-dimensional vector such that the ith element is 1 and the others are 0, and $\boldsymbol{p}_j$ is the jth column vector of $\boldsymbol{P}_1$, i.e., $\boldsymbol{P}_1 = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m)$.

At first, we consider the case of $s_0^2 \neq 0$. Recall that the number of v_js that are zero is a_*, where a_* is given by (2.10). Thus, $\mathrm{tr}(\boldsymbol{W}) = m - a_*$ is satisfied, where $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_m)$ is given in Theorem 4. Let $\boldsymbol{L}$ be an mth diagonal matrix defined by $\boldsymbol{L} = \mathrm{diag}(z_1^2, \ldots, z_m^2)$, where z_j is given by (2.7). Then, we have

$$\sum_{j=1}^{m}\frac{w_j}{z_j^2}\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{P}_1'\boldsymbol{y}, \quad \frac{w_j s_{a_*}^2}{z_j^2} = w_j - v_j, \tag{E.2}$$

where s_α^2 is given by (2.9). Notice that

$$\frac{\partial v_j}{\partial \boldsymbol{y}} = -\frac{w_j}{z_j^4}\left\{\left(\frac{\partial s_{a_*}^2}{\partial \boldsymbol{y}}\right)z_j^2 - s_{a_*}^2\left(\frac{\partial z_j^2}{\partial \boldsymbol{y}}\right)\right\}, \tag{E.3}$$

and $\partial s_{a_*}^2/\partial \boldsymbol{y}$ does not depend on j. From the above results and (E.2), the last part of (E.1) is expressed as

$$\begin{aligned}\sum_{j=1}^{m}\left(\frac{\partial v_j}{\partial \boldsymbol{y}'}\right)\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y} &= -\sum_{j=1}^{m}\frac{w_j}{z_j^4}\left\{\left(\frac{\partial s_{a_*}^2}{\partial \boldsymbol{y}'}\right)z_j^2 - s_{a_*}^2\left(\frac{\partial z_j^2}{\partial \boldsymbol{y}'}\right)\right\}\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y} \\ &= -\left(\frac{\partial s_{a_*}^2}{\partial \boldsymbol{y}'}\right)\boldsymbol{P}_1\boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{P}_1'\boldsymbol{y} + \sum_{j=1}^{m}\frac{w_j - v_j}{z_j^2}\left(\frac{\partial z_j^2}{\partial \boldsymbol{y}'}\right)\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y}. \end{aligned} \tag{E.4}$$

On the other hand, by using the same method as in Appendix C, $s_{a_*}^2$ and z_j^2 are rewritten as

$$s_{a_*}^2 = \frac{1}{n-m-1+a_*}\boldsymbol{y}'\{\boldsymbol{P}_2\boldsymbol{P}_2' + \boldsymbol{P}_1(\boldsymbol{I}_m - \boldsymbol{W})\boldsymbol{P}_1'\}\boldsymbol{y}, \quad z_j^2 = (\boldsymbol{p}_j'\boldsymbol{y})^2,$$

where $\boldsymbol{P}_2$ is given by (C.1). These equations imply that

$$\frac{\partial s_{a_*}^2}{\partial \boldsymbol{y}} = \frac{2}{n-m-1+a_*}\{\boldsymbol{P}_2\boldsymbol{P}_2' + \boldsymbol{P}_1(\boldsymbol{I}_m - \boldsymbol{W})\boldsymbol{P}_1'\}\boldsymbol{y}, \quad \frac{\partial z_j^2}{\partial \boldsymbol{y}} = 2\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y}. \tag{E.5}$$

It follows from $\boldsymbol{P}_1'\boldsymbol{P}_1 = \boldsymbol{I}_m$, $\boldsymbol{P}_2'\boldsymbol{P}_1 = \boldsymbol{O}_{n-m,m}$, $\boldsymbol{W}^2 = \boldsymbol{W}$, and $z_j = \boldsymbol{p}_j'\boldsymbol{y}$ that

$$\boldsymbol{y}'\{\boldsymbol{P}_2\boldsymbol{P}_2' + \boldsymbol{P}_1(\boldsymbol{I}_m - \boldsymbol{W})\boldsymbol{P}_1'\}\boldsymbol{P}_1\boldsymbol{W}\boldsymbol{L}^{-1}\boldsymbol{P}_1'\boldsymbol{y} = 0, \quad \frac{1}{z_j^2}\boldsymbol{y}'\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y} = 1. \tag{E.6}$$

By using (E.6) after substituting (E.5) into (E.4), we derive

$$\sum_{j=1}^{m}\left(\frac{\partial v_j}{\partial \boldsymbol{y}'}\right)\boldsymbol{p}_j\boldsymbol{p}_j'\boldsymbol{y} = 2\sum_{j=1}^{m}(w_j - v_j) = 2\{\mathrm{tr}(\boldsymbol{W}) - \mathrm{tr}(\boldsymbol{V})\}. \tag{E.7}$$

Next, we consider the case of $s_0^2 = 0$. In order to give the proof, it is only necessary to replace $\partial s_{a_*}^2/\partial \boldsymbol{y}$ in (E.3) with $\partial t_1/\partial \boldsymbol{y}$, where t_j is given by (2.8). Notice that $t_j = \boldsymbol{y}'\boldsymbol{P}_1(\boldsymbol{I}_m - \boldsymbol{W})\boldsymbol{P}_1'\boldsymbol{y}$. Thus, by using the same method that was used in the proof of the case $s_0^2 \neq 0$, we can see that the equation (E.7) is satisfied even when $s_0^2 = 0$. Consequently, equation (3.7) is derived from (E.1) and (E.7).