TR-No. 19-05, Hiroshima Statistical Research Group, 1–15

Equivalence between Adaptive-Lasso and Generalized Ridge Estimators in Linear Regression with Orthogonal Explanatory Variables after Optimizing Regularization Parameters

Mineaki Ohishi[1*], **Hirokazu Yanagihara**[1] **and Shuichi Kawano**[2]

[1]Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan
[2]Department of Computer and Network Engineering, Graduate School of Informatics and Engineering
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

Abstract

In this paper, we deal with a penalized least-squares (PLS) method for a linear regression model with orthogonal explanatory variables. The used penalties are an adaptive-Lasso (AL)-type ℓ_1 penalty (AL-penalty) and a generalized ridge (GR)-type ℓ_2 penalty (GR-penalty). Since the estimators obtained by minimizing the PLS methods strongly depend on the regularization parameters, we optimize them by a model selection criterion (MSC)-minimization method. The estimators based on the AL-penalty and the GR-penalty have different properties, and it is universally recognized that these are completely different estimators. However, in this paper, we show an interesting result that the two estimators are exactly equal when the explanatory variables are orthogonal after optimizing the regularization parameters by the MSC-minimization method.

(Last Modified: February 25, 2019)

Key words: Adaptive-Lasso, C_p criterion, GCV criterion, Generalized ridge regression, GIC, Linear regression, Model selection criterion, Optimization problem, Regularization parameters, Sparsity.

*Corresponding author
E-mail address: mineaki-ohishi@hiroshima-u.ac.jp (Mineaki Ohishi)

1. Introduction

We deal with a linear regression model with an n-dimensional vector of response variables $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and an $n \times k$ matrix of nonstochastic explanatory variables $\boldsymbol{X}$, where n is the sample size and k is the number of explanatory variables. Here, without loss of generality, we

assume that $\boldsymbol{y}$ and $\boldsymbol{X}$ are centralized, i.e., $\boldsymbol{y}'\mathbf{1}_n = 0$ and $\boldsymbol{X}'\mathbf{1}_n = \mathbf{0}_k$, where $\mathbf{1}_n$ is an n-dimensional vector of ones and $\mathbf{0}_k$ is a k-dimensional vector of zeros. Moreover, in this paper, we particularly assume that the following equations hold:

$$\mathrm{rank}(\boldsymbol{X}) = k < n - 1, \quad \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_k), \quad d_1 \geq \cdots \geq d_k > 0.$$

The relation $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{D}$ indicates that the explanatory variables are orthogonal. Examples of models with orthogonal explanatory variables include those of principal component analysis (Massy, 1965; Jolliffe, 1982; Yanagihara, 2018), generalized ridge (GR) regression (Hoerl & Kennard, 1970), and smoothing using orthogonal basis functions (Yanagihara, 2012; Hagiwara, 2017).

The least-squares (LS) method is widely used for estimating the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ of a linear regression model. The LS estimator (LSE) of $\boldsymbol{\beta}$ is obtained by minimizing the residual sum of squares (RSS) defined by

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{1.1}$$

There also exist penalized LS (PLS) methods for estimating $\boldsymbol{\beta}$. In a PLS method, an estimator of $\boldsymbol{\beta}$ is obtained from a minimization of a penalized RSS (PRSS) defined by adding the RSS to a penalty term. There are many kinds of PLS methods. One such method is the GR regression proposed by Hoerl & Kennard (1970), which is designed to avoid multicollinearity among explanatory variables. The GR estimator (GRE) of $\boldsymbol{\beta}$ is obtained by minimizing the $\mathrm{PRSS}^{\mathrm{GR}}$ defined by adding the RSS to the GR-type ℓ_2 penalty (GR-penalty) as

$$\mathrm{PRSS}^{\mathrm{GR}}(\boldsymbol{\beta}) = \mathrm{RSS}(\boldsymbol{\beta}) + \sum_{j=1}^{k} \theta_j \beta_j^2, \tag{1.2}$$

where $\theta_j \in \mathbb{R}_+ = \{\theta \in \mathbb{R} \mid \theta \geq 0\}$ $(j = 1, \ldots, k)$ are regularization parameters called ridge parameters. When $\theta_1 = \cdots = \theta_k = 0$, the $\mathrm{PRSS}^{\mathrm{GR}}$ coincides with the usual RSS. Most researchers consider it commonsense that the GRE does not have sparsity. Since the value of the GRE of $\boldsymbol{\beta}$ depends on ridge parameters, the optimization of these parameters is very important. Methods for optimizing ridge parameters include model selection criterion (MSC)-minimization methods, for example, the generalized C_p (GC_p; Atkinson, 1980)- and GCV (Craven & Wahba, 1979)-minimization methods (Nagai *et al.*, 2012; Yanagihara, 2018), and a fast algorithm for minimizing MSC (Ohishi *et al.*, 2018).

Moreover, Lasso, proposed by Tibshirani (1996), and adaptive-Lasso (AL), proposed by Zou (2006) as an extension of the Lasso, give sparse estimates of unknown parameters. The AL estimator (ALE) of $\boldsymbol{\beta}$ is obtained by minimizing $\mathrm{PRSS}^{\mathrm{AL}}$, which is defined by changing the GR-penalty in (1.2) to the AL-type ℓ_1 penalty (AL-penalty) as

$$\mathrm{PRSS}^{\mathrm{AL}}(\boldsymbol{\beta}) = \mathrm{RSS}(\boldsymbol{\beta}) + 2\lambda \sum_{j=1}^{k} w_j |\beta_j|, \tag{1.3}$$

where $\lambda \in \mathbb{R}_+$ is a regularization parameter called a tuning parameter and w_j ($j = 1, \ldots, k$) is a weight. The $\mathrm{PRSS}^{\mathrm{AL}}$ with $\lambda = 0$ coincides with RSS, and the AL with $w_j = 1$ coincides with the ordinary Lasso. The ALE of $\boldsymbol{\beta}$ usually cannot be obtained without a numerical search algorithm, e.g., LARS (Efron *et al.*, 2004), Coordinate Descent (Friedman *et al.*, 2010), or ADMM (Boyd *et al.*, 2011). However, in the case of using orthogonal explanatory variables as in this paper, the ALE can be obtained in closed form. For the weights w_j ($j = 1, \ldots, k$) in the AL, Zou (2006) proposed $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|^{\delta}$ ($\delta \in \mathbb{R}_+ \backslash \{0\}$), where $\hat{\beta}_j^{\mathrm{LS}}$ is the LSE of β_j. Using these weights, it is known that the ALE satisfies the oracle property (Fan & Li, 2001). Since the value of the ALE of $\boldsymbol{\beta}$ depends on a tuning parameter, the optimization of this parameter is very important. Methods for optimizing the tuning parameter include MSC-minimization methods as in the case of GR. As examples, there are the CV- and an ERIC (Francis *et al.*, 2015)-minimization methods (Zou, 2006; Francis *et al.*, 2015) and selection stability (Sun *et al.*, 2013).

In this paper, we give the closed form of the tuning parameter optimized by the GCV-minimization method when $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$. Moreover, although it is widely recognized that the GRE and the ALE are different estimators because the GRE does not have sparsity and the ALE has sparsity, we show an interesting result that the GRE and the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ are exactly equal after optimizing the regularization parameters by the GCV-minimization method.

This paper is organized as follows: In Section 2, we show that the tuning parameter of the AL optimized by the GCV-minimization method can be obtained in closed form. Moreover, we show the equivalence between the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and the GRE after optimizing the regularization parameters. In Section 3, we show the equivalence between the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and the GRE after optimizing the regularization parameters by the MSC-minimization method. Technical details are provided in the Appendix.

2. Equivalence between Two Estimators Optimized by the GCV-Minimization Method

In the beginning of this section, we consider the ALE of $\boldsymbol{\beta}$ with the tuning parameter optimized by the GCV-minimization method. Since the explanatory variables are orthogonal, it follows from the singular-value decomposition that

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

where $\boldsymbol{O}_{n,k}$ is an $n \times k$ matrix of zeros, $\boldsymbol{P}$ is an orthogonal matrix of order n, and $\boldsymbol{P}_1$ satisfying $\boldsymbol{P}_1' \boldsymbol{P}_1 = \boldsymbol{I}_k$ and $\boldsymbol{P}_1' \boldsymbol{1}_n = \boldsymbol{0}_k$ is the $n \times k$ matrix that consists of the first k columns of $\boldsymbol{P}$. Using $\boldsymbol{P}_1$,

we define the k-dimensional vector $\boldsymbol{z}_1$ as

$$\boldsymbol{z}_1 = (z_1, \ldots, z_k)' = \boldsymbol{P}_1'\boldsymbol{y}. \tag{2.2}$$

Then the LSE of $\boldsymbol{\beta}$ that minimizes the RSS in (1.1) is given as

$$\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = \left(\hat{\beta}_1^{\mathrm{LS}}, \ldots, \hat{\beta}_k^{\mathrm{LS}}\right)' = \boldsymbol{D}^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{D}^{-1/2}\boldsymbol{z}_1 = \left(z_1/\sqrt{d_1}, \ldots, z_k/\sqrt{d_k}\right)'. \tag{2.3}$$

When the explanatory variables are orthogonal as in this paper, the ALE of $\boldsymbol{\beta}$ that minimizes (1.3) can be obtained in closed form, given as in the following theorem (the proof is given in Appendix A.1).

Theorem 1. *Let $\boldsymbol{L}_\lambda$ be the diagonal matrix of order k of which the jth diagonal element is defined by*

$$\ell_{\lambda,j} = \frac{1}{d_j} S\left(1, \lambda w_j / \left(|z_j|\sqrt{d_j}\right)\right),$$

where $S(x, a)$ is a soft-thresholding operator, i.e., $S(x, a) = \mathrm{sign}(x)(|x| - a)_+$. Then the ALE of $\boldsymbol{\beta}$ that minimizes the $\mathrm{PRSS}^{\mathrm{AL}}$ *in* (1.3) *is given by*

$$\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{AL}} = (\hat{\beta}_{\lambda,1}^{\mathrm{AL}}, \ldots, \hat{\beta}_{\lambda,k}^{\mathrm{AL}})' = \boldsymbol{L}_\lambda \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{L}_\lambda \boldsymbol{D}^{1/2}\boldsymbol{z}_1, \tag{2.4}$$

that is, $\hat{\beta}_{\lambda,j}^{\mathrm{AL}}$ is expressed as

$$\hat{\beta}_{\lambda,j}^{\mathrm{AL}} = \frac{1}{\sqrt{d_j}} S\left(z_j, \lambda w_j / \sqrt{d_j}\right). \tag{2.5}$$

From Theorem 1 and the result in Ohishi & Yanagihara (2017), we can see that the ALE coincides with the ordinary Lasso estimator when $w_j = 1$ and the LSE in (2.3) when $\lambda = 0$.

Let $\boldsymbol{H}_\lambda^{\mathrm{AL}}$ be a hat matrix of the AL, i.e., $\boldsymbol{H}_\lambda^{\mathrm{AL}} = \boldsymbol{X}\boldsymbol{L}_\lambda\boldsymbol{X}'$. Then, the predictive value of $\boldsymbol{y}$ from the AL is given by

$$\hat{\boldsymbol{y}}_\lambda^{\mathrm{AL}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{AL}} = \boldsymbol{H}_\lambda^{\mathrm{AL}}\boldsymbol{y}.$$

The GCV criterion for optimizing a tuning parameter consists of the following estimator of variance $\hat{\sigma}_{\mathrm{AL}}^2$ and generalized degrees of freedom $\mathrm{df}_{\mathrm{AL}}$:

$$\hat{\sigma}_{\mathrm{AL}}^2(\lambda) = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}}_\lambda^{\mathrm{AL}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_\lambda^{\mathrm{AL}}) = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{L}_\lambda\boldsymbol{X}')^2\boldsymbol{y}, \tag{2.6}$$

$$\mathrm{df}_{\mathrm{AL}}(\lambda) = 1 + \mathrm{tr}(\boldsymbol{H}_\lambda^{\mathrm{AL}}) = 1 + \mathrm{tr}(\boldsymbol{L}_\lambda\boldsymbol{D}). \tag{2.7}$$

The generalized degrees of freedom in (2.7) with $w_j = 1$ coincides with that proposed by Tibshirani (1996). Using the above equations, the GCV criterion for optimizing a tuning parameter is

given by

$$\mathrm{GCV}_{\mathrm{AL}}(\lambda) = \frac{\hat{\sigma}^2_{\mathrm{AL}}(\lambda)}{\{1 - \mathrm{df}_{\mathrm{AL}}(\lambda)/n\}^2}.$$

In this paper, a common weight proposed by Zou (2006) with $\delta = 1$ is used as w_j:

$$w_j = \frac{1}{|\hat{\beta}^{\mathrm{LS}}_j|} = \frac{\sqrt{d_j}}{|z_j|} \quad (j = 1, \ldots, k).$$

Then, from Theorem 1, the jth element of the ALE with $w_j = 1/|\hat{\beta}^{\mathrm{LS}}_j|$ is calculated as

$$\hat{\beta}^{\mathrm{AL}}_{\lambda,j} = \frac{1}{\sqrt{d_j}} S\left(z_j, \lambda/|z_j|\right). \tag{2.8}$$

The $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ are rewritten as in the following lemma (the proof is given in Appendix A.2).

Lemma 1. *The $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ with $w_j = 1/|\hat{\beta}^{\mathrm{LS}}_j|$ can be expressed as*

$$\hat{\sigma}^2_{\mathrm{AL}}(\lambda) = \frac{1}{n}\left[n\hat{\sigma}^2_0 + \sum_{j=1}^{k}\left\{S\left(\lambda/z_j^2, 1\right) + 1\right\}^2 z_j^2\right], \quad \mathrm{df}_{\mathrm{AL}}(\lambda) = 1 - \sum_{j=1}^{k} S\left(\lambda/z_j^2, 1\right), \tag{2.9}$$

where $\hat{\sigma}^2_0$ is given by

$$\hat{\sigma}^2_0 = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{D}^{-1}\boldsymbol{X}')\boldsymbol{y}. \tag{2.10}$$

Since when $k < n-1$, $\hat{\sigma}^2_0 \neq 0$ in most cases, we assume $\hat{\sigma}^2_0 \neq 0$ in this paper. Moreover, let $t_0 = 0$, t_j $(j = 1, \ldots, k)$ be the jth-order statistic of $z_1^2, \ldots, z_k^2$, i.e.,

$$t_j = \begin{cases} \min\{z_1^2, \ldots, z_k^2\} & (j = 1) \\ \min\{\{z_1^2, \ldots, z_k^2\} \backslash \{t_1, \ldots, t_{j-1}\}\} & (j = 2, \ldots, k) \end{cases}, \tag{2.11}$$

R_j $(j = 0, 1, \ldots, k)$ be the range defined by

$$R_j = \begin{cases} (t_j, t_{j+1}] & (j = 0, 1, \ldots, k-1) \\ (t_k, \infty) & (j = k) \end{cases}, \tag{2.12}$$

and s^2_a $(a = 0, 1, \ldots, k)$ be the estimators of variance defined by

$$s^2_a = \frac{n\hat{\sigma}^2_0 + \sum_{j=0}^{a} t_j}{n - k - 1 + a} \quad (a = 0, 1, \ldots, k). \tag{2.13}$$

As the relation between R_a and s^2_a, Yanagihara (2018) showed that the following statement is true:

$$\exists! a_* \in \{0, \ldots, k-1\} \; s.t. \; s^2_{a_*} \in R_{a_*}. \tag{2.14}$$

Then, the tuning parameter optimized by the GCV-minimization method is as in the following theorem (the proof is given in Appendix A.3).

Theorem 2. *Let $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and let $\hat{\lambda}$ be the tuning parameter optimized by the GCV-minimization method, i.e.,*

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}_+} \mathrm{GCV}_{\mathrm{AL}}(\lambda).$$

Then, the closed form of $\hat{\lambda}$ is given by $\hat{\lambda} = s_{a_*}^2$.

By using Theorem 2, from (2.8), we can see that the ALE of $\boldsymbol{\beta}$ with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ after optimizing the tuning parameter by the GCV-minimization method is obtained as the following closed form:

$$\hat{\beta}_{\hat{\lambda},j}^{\mathrm{AL}} = \frac{1}{\sqrt{d_j}} S\left(z_j, s_{a_*}^2/|z_j|\right). \tag{2.15}$$

Next, in order to show the equivalence between the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and the GRE, we consider the GRE of $\boldsymbol{\beta}$ with the ridge parameters optimized by the GCV-minimization method. The GRE that minimizes the $\mathrm{PRSS}^{\mathrm{GR}}$ in (1.2) is given by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{\mathrm{GR}} = \boldsymbol{D}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{D}_{\boldsymbol{\theta}}^{-1}\boldsymbol{D}^{1/2}\boldsymbol{z}_1,$$

where $\boldsymbol{D}_{\boldsymbol{\theta}} = \boldsymbol{D} + \mathrm{diag}(\theta_1, \ldots, \theta_k)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Since it is easy to see that the jth element of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{\mathrm{GR}}$ depends on only θ_j, we write it as

$$\hat{\beta}_{\theta_j,j}^{\mathrm{GR}} = \frac{\sqrt{d_j} z_j}{d_j + \theta_j}. \tag{2.16}$$

Let $\boldsymbol{H}_{\boldsymbol{\theta}}^{\mathrm{GR}}$ be a hat matrix of the GR, i.e., $\boldsymbol{H}_{\boldsymbol{\theta}}^{\mathrm{GR}} = \boldsymbol{X}\boldsymbol{D}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}'$. Then, the predictive value of $\boldsymbol{y}$ from the GR is given by

$$\hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{\mathrm{GR}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}^{\mathrm{GR}} = \boldsymbol{H}_{\boldsymbol{\theta}}^{\mathrm{GR}}\boldsymbol{y}.$$

The GCV criterion for optimizing the ridge parameters consists of the following estimator of variance $\hat{\sigma}_{\mathrm{GR}}^2$ and generalized degrees of freedom $\mathrm{df}_{\mathrm{GR}}$:

$$\hat{\sigma}_{\mathrm{GR}}^2(\boldsymbol{\theta}) = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{\mathrm{GR}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}^{\mathrm{GR}}) = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{X}\boldsymbol{D}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}')^2\boldsymbol{y}, \tag{2.17}$$

$$\mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta}) = 1 + \mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\theta}}^{\mathrm{GR}}) = 1 + \mathrm{tr}(\boldsymbol{D}_{\boldsymbol{\theta}}^{-1}\boldsymbol{D}). \tag{2.18}$$

Using the above equations, the GCV criterion for optimizing ridge parameters is given by

$$\mathrm{GCV}_{\mathrm{GR}}(\boldsymbol{\theta}) = \frac{\hat{\sigma}_{\mathrm{GR}}^2(\boldsymbol{\theta})}{\{1 - \mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta})/n\}^2}.$$

Yanagihara (2018) showed that the ridge parameters optimized by the GCV-minimization method are obtained as the following closed forms:

$$\hat{\theta}_j = \begin{cases} \dfrac{d_j s_{a_*}^2}{z_j^2 - s_{a_*}^2} & (s_{a_*}^2 < z_j^2) \\ \infty & (s_{a_*}^2 \geq z_j^2) \end{cases} \quad (j = 1, \ldots, k).$$

As the result, we have

$$\frac{\sqrt{d_j}}{d_j + \hat{\theta}_j} = \frac{1}{\sqrt{d_j}} S\left(1, s_{a_*}^2 / z_j^2\right).$$

Consequently, from the above equation and (2.16), the GRE after optimizing the ridge parameters by the GCV-minimization method is given by

$$\hat{\beta}_{\hat{\theta}_j, j}^{\mathrm{GR}} = \frac{1}{\sqrt{d_j}} S\left(z_j, s_{a_*}^2 / |z_j|\right). \tag{2.19}$$

From the result that (2.19) includes 0, we can see that the GRE after optimizing the ridge parameters has sparsity. By comparing (2.15) and (2.19), we can obtain the following theorem.

Theorem 3. *When the explanatory variables are orthogonal, the ALE with* $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ *is exactly equal to the GRE after optimizing the regularization parameters by the GCV-minimization method, i.e.,* $\hat{\beta}_{\lambda, j}^{\mathrm{AL}} = \hat{\beta}_{\hat{\theta}_j, j}^{\mathrm{GR}}$ $(j = 1, \ldots, k)$.

3. Equivalence between Two Estimators Optimized by the MSC-Minimization Method

In the previous section, we showed the equivalence between the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and the GRE after optimizing the regularization parameters by the GCV-minimization method. In this section, we show that the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ is equal to the GRE optimized not only by the GCV-minimization method but also by a general MSC-minimization method. First, we consider the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ after optimizing the tuning parameter by the MSC-minimization method. The MSC for optimizing a tuning parameter can be expressed by a bivariate function with respect to the $\hat{\sigma}_{\mathrm{AL}}^2(\lambda)$ in (2.6) and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ in (2.7). From Lemma 1, we obtain the following lemma about the ranges of $\hat{\sigma}_{\mathrm{AL}}^2(\lambda)$ and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ (the proof is given in Appendix A.4).

Lemma 2. *Ranges of* $\hat{\sigma}_{\mathrm{AL}}^2(\lambda)$ *and* $\mathrm{df}_{\mathrm{AL}}(\lambda)$ *with* $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ *are given by*

$$\hat{\sigma}_{\mathrm{AL}}^2(\lambda) \in [\hat{\sigma}_0^2,\ \hat{\sigma}_\infty^2], \quad \mathrm{df}_{\mathrm{AL}}(\lambda) \in [1,\ k+1],$$

where $\hat{\sigma}_0^2$ *is given by* (2.10) *and* $\hat{\sigma}_\infty^2 = \lim_{\lambda \to \infty} \hat{\sigma}_{\mathrm{AL}}^2(\lambda)$*, i.e.,*

$$\hat{\sigma}_\infty^2 = \frac{1}{n} \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{y}.$$

A general expression of the MSC comes from using the following bivariate function given by Ohishi *et al.* (2018).

Definition 1. *The $f(r,u)$ is a bivariate function that satisfies the following conditions:*

(C1) *$f(r,u)$ is a continuous function at any $(r,u) \in (0, \hat{\sigma}^2_\infty] \times [1, u_0)$,*

(C2) *$f(r,u) > 0$ for any $(r,u) \in (0, \hat{\sigma}^2_\infty] \times [1, u_0)$,*

(C3) *$f(r,u)$ is first-order partially differentiable at any $(r,u) \in (0, \hat{\sigma}^2_\infty] \times [1, u_0)$ and*

$$\dot{f}_r(r,u) = \frac{\partial}{\partial r} f(r,u) > 0, \quad \dot{f}_u(r,u) = \frac{\partial}{\partial u} f(r,u) > 0, \quad \forall (r,u) \in (0, \hat{\sigma}^2_\infty] \times [1, u_0),$$

where $u_0 \leq n$.

By using the bivariate function $f(r,u)$, the MSC for optimizing a tuning parameter can be expressed as

$$\mathrm{MSC}_{\mathrm{AL}}(\lambda) = f(\hat{\sigma}^2_{\mathrm{AL}}(\lambda),\ \mathrm{df}_{\mathrm{AL}}(\lambda)). \tag{3.1}$$

Specific forms of the functions f of existing criteria, for example, a generalized C_p (GC_p; Atkinson, 1980), a generalized information criterion (GIC; Nishii, 1984) under normality, and an extended GCV (EGCV; Ohishi *et al.*, 2018), are expressed as follows:

$$f(r,u) = \begin{cases} nr/s_0^2 + \alpha u & (GC_p) \\ r \exp(\alpha u/n) & (\mathrm{GIC}) \\ r/(1-u/n)^\alpha & (\mathrm{EGCV} : u < n) \end{cases},$$

where s_0^2 is given by (2.13) and α is some positive value expressing the strength of a penalty for model complexity. We can see that $s_0^2 \neq 0$ because we assume that $\hat{\sigma}_0^2 \neq 0$. Moreover, the original GIC under normality is expressed as $n \log r + \alpha u$. The GIC in this paper is defined as an exponential transformation of the original GIC divided by n. Using (3.1), the tuning parameter optimized by the MSC-minimization method is given by

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}_+} \mathrm{MSC}_{\mathrm{AL}}(\lambda).$$

Hence, it follows from $\hat{\lambda}$ and (2.8) that the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ after optimizing the tuning parameter by the MSC-minimization method is given by

$$\hat{\beta}^{\mathrm{AL}}_{\hat{\lambda},j} = \frac{1}{\sqrt{d_j}} S\left(z_j, \hat{\lambda}/|z_j|\right). \tag{3.2}$$

Next, in order to show the equivalence between the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and the GRE, we give the GRE after optimizing the ridge parameters by the MSC-minimization method. From

Ohishi *et al.* (2018), the ridge parameters optimized by the MSC-minimization method are

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)' = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{\theta}), \quad \mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{\theta}) = f(\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta}),\ \mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta})), \tag{3.3}$$

where f is given by Definition 1 and $\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta})$ and $\mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta})$ are given by (2.17) and (2.18), respectively. The following lemma describes the ranges of $\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta})$ and $\mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta})$ (the proof is given in Ohishi *et al*., 2018).

Lemma 3. *Ranges of* $\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta})$ *and* $\mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta})$ *are given by*

$$\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta}) \in [\hat{\sigma}^2_0,\ \hat{\sigma}^2_\infty], \quad \mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta}) \in [1,\ k+1].$$

Here, we consider the following class of ridge parameters defined by Ohishi *et al.* (2018):

$$\forall h \in \mathbb{R}_+,\ \boldsymbol{g}(h) = (g_1(h), \ldots, g_k(h))',\ g_j(h) = \begin{cases} \dfrac{d_j h}{z_j^2 - h} & (h < z_j^2) \\ \infty & (h \geq z_j^2) \end{cases}. \tag{3.4}$$

In the class, k ridge parameters are written in terms of one parameter h, and hence the codomain of the class becomes smaller than that of $\boldsymbol{\theta}$. Nevertheless, Ohishi *et al.* (2018) showed that the optimal ridge parameters are included in the class. Hence, it follows from (3.3) and (3.4) that the ridge parameters optimized by the MSC-minimization method are given by

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)' = \boldsymbol{g}(\hat{h}) = (g_1(\hat{h}), \ldots, g_k(\hat{h}))',\ g_j(\hat{h}) = \begin{cases} \dfrac{d_j \hat{h}}{z_j^2 - \hat{h}} & (\hat{h} < z_j^2) \\ \infty & (\hat{h} \geq z_j^2) \end{cases},$$

$$\hat{h} = \arg\min_{h \in \mathbb{R}_+} \mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{g}(h)).$$

Equation (2.16) and $\hat{\boldsymbol{\theta}} = \boldsymbol{g}(\hat{h})$ imply that the GRE after optimizing the ridge parameters by the MSC-minimization method is given as

$$\hat{\beta}^{\mathrm{GR}}_{\hat{\theta}_j, j} = \frac{1}{\sqrt{d_j}} S\left(z_j, \hat{h}/|z_j|\right). \tag{3.5}$$

From the result that (3.5) includes 0, we can see that the GRE after optimizing the ridge parameters has sparsity as in (2.19).

Equations (3.2) and (3.5) imply that if $\hat{\lambda} = \hat{h}$, the ALE with $w_j = 1/|\hat{\beta}^{\mathrm{LS}}_j|$ is exactly equal to the GRE after optimizing the regularization parameters. The equality is shown if the function for optimizing λ is the same as that for optimizing h. The equivalence of the two functions can be derived by the following lemma (the proof is given in Appendix A.5).

Lemma 4. *When* $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$*, we have*

$$\forall x \in \mathbb{R}_+,\ \mathrm{MSC}_{\mathrm{AL}}(x) = \mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{g}(x)).$$

Lemma 4 implies that $\hat{\lambda} = \hat{h}$. Hence, we can obtain the following theorem.

Theorem 4. *Suppose that* $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$*. Let* $\hat{\lambda}$ *and* $\hat{h}$ *be the minimizers of* $\mathrm{MSC}_{\mathrm{AL}}(\lambda)$ *and* $\mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{g}(h))$*, respectively, i.e.,*

$$\hat{\lambda} = \arg\min_{\lambda \in \mathbb{R}_+} \mathrm{MSC}_{\mathrm{AL}}(\lambda), \quad \hat{h} = \arg\min_{h \in \mathbb{R}_+} \mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{g}(h)).$$

When the explanatory variables are orthogonal, $\hat{\lambda}$ *is exactly equal to* $\hat{h}$*, and hence the ALE is exactly equal to the GRE, i.e.,* $\hat{\beta}_{\hat{\lambda},j}^{\mathrm{AL}} = \hat{\beta}_{\hat{\theta}_j,j}^{\mathrm{GR}}$ $(j = 1, \ldots, k)$*.*

Theorem 4 shows the equivalence between the ALE with $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$ and the GRE after optimizing the regularization parameters by the MSC-minimization method when the explanatory variables are orthogonal. When we use the PLS method based on the AL-penalty or the GR-penalty, although we have to calculate $\hat{\lambda}$ or $\hat{h}$, the values can be obtained in a calculation of order $O(k)$ by using a fast algorithm proposed by Ohishi *et al.* (2018).

4. Conclusion

In this paper, we dealt with the PLS methods based on the AL-penalty and the GR-penalty when the explanatory variables are orthogonal. Although the estimators obtained from these penalties are different, we showed the interesting result that the two estimators with the regularization parameters optimized by the MSC-minimization method are exactly equal. The equivalence of the two estimators was derived from the result that the function for optimizing the tuning parameter in the AL is equal to that for optimizing the ridge parameters in the GR. Therefore, the two PLS methods are completely equivalent when the explanatory variables are orthogonal. For the case of general explanatory variables, although the ALE cannot be obtained without iterative calculation, the GRE can be obtained in closed form. If the equivalence or some relationship between the ALE and the GRE can be obtained for general explanatory variables, we may easily obtain the ALE through the GRE. The results in this paper suggest that possibility.

Acknowledgment The authors thank Prof. Emer. Yasunori Fujikoshi, Dr. Shintaro Hashimoto, and Mr. Ryoya Oda, of Hiroshima University and Dr. Tomoyuki Nakagawa of Tokyo University of Science, for helpful comments.

References

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

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122.

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.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32**, 407–499.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

Francis, K. C. H., David, I. W. & Scott, D. F. (2015). Tuning parameter selection for the adaptive lasso using ERIC. *J. Amer. Statist. Assoc*., **110**, 262–269.

Friedman, J. H., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*., **33**, 1–22.

Hagiwara, K. (2017). A scaling and non-negative garrote in soft-thresholding. *IEICE Trans. Inf. & Syst.*, **E100.D**, 2702–2710.

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

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *J. Appl. Statist.*, **31**, 300–303.

Massy, W. F. (1965). Principal components regression in explanatory statistical research. *J. Amer. Statist. Assoc.*, **60**, 234–256.

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

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

Ohishi, M. & Yanagihara, H. (2017). Minimization algorithm of model selection criterion for optimizing tuning parameter in Lasso estimator when explanatory variables are orthogonal. *RIMS Kôkyûroku*, **2047**, 124–140 (*in Japanese*).

Ohishi, M., Yanagihara, H. & Fujikoshi, Y. (2018). A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. (submit for publication).

Sun, W., Wang, J. & Fang, Y. (2013). Consistent selection of tuning parameters via variable selection stability. *J. Mach. Learn. Res.*, **14**, 3419–3440.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267–288.

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

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

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

Appendix

A.1. Proof of Theorem 1

Using the orthogonal matrix $\boldsymbol{P}$ in (2.1), we define the n-dimensional vector $\boldsymbol{z} = \boldsymbol{P}'\boldsymbol{y}$. Then, since $\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2)$, by using $\boldsymbol{z}_1$ in (2.2), $\boldsymbol{z}$ can be partitioned as

$$\boldsymbol{z} = (z_1, \ldots, z_n)' = \boldsymbol{P}'\boldsymbol{y} = \begin{pmatrix} \boldsymbol{P}_1'\boldsymbol{y} \\ \boldsymbol{P}_2'\boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{pmatrix}. \tag{A.1}$$

These equations imply that

$$\begin{aligned}\mathrm{PRSS}^{\mathrm{AL}}(\boldsymbol{\beta}) &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{P}\boldsymbol{P}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\sum_{j=1}^{k} w_j|\beta_j| \\ &= \left\{\boldsymbol{z} - \begin{pmatrix}\boldsymbol{D}^{1/2}\\ \boldsymbol{O}_{n-k,k}\end{pmatrix}\boldsymbol{\beta}\right\}' \left\{\boldsymbol{z} - \begin{pmatrix}\boldsymbol{D}^{1/2}\\ \boldsymbol{O}_{n-k,k}\end{pmatrix}\boldsymbol{\beta}\right\} + 2\lambda\sum_{j=1}^{k} w_j|\beta_j| \\ &= \sum_{j=1}^{k}\left\{\left(z_j - \sqrt{d_j}\beta_j\right)^2 + 2\lambda w_j|\beta_j|\right\} + \sum_{j=k+1}^{n} z_j^2.\end{aligned}$$

Hence, the minimization of the $\mathrm{PRSS}^{\mathrm{AL}}$ is equivalent to that of the following function:

$$\begin{aligned}\zeta(\beta_j \mid d_j) &= \left(z_j - \sqrt{d_j}\beta_j\right)^2 + 2\lambda w_j|\beta_j| \\ &= d_j\beta_j^2 - 2\left\{z_j\sqrt{d_j} - \lambda w_j\,\mathrm{sign}(\beta_j)\right\}\beta_j + z_j^2 \quad (j=1,\ldots,k).\end{aligned}$$

Since $d_j > 0$, $\zeta(\beta_j \mid d_j)$ is a piecewise quadratic function. If the sign of $z_j d_j^{1/2} - \lambda w_j\,\mathrm{sign}(\beta_j)$ that is the β_j-coordinate of the vertex of $\zeta(\beta_j \mid d_j)$ is equal to the sign of β_j, then the minimizer of $\zeta(\beta_j \mid d_j)$ is the β_j-coordinate of the vertex, and otherwise it is 0. This result implies (2.5). Moreover, by using $\boldsymbol{X} = \boldsymbol{P}_1\boldsymbol{D}^{1/2}$ and $\boldsymbol{z}_1 = \boldsymbol{P}_1'\boldsymbol{y}$, we have

$$\hat{\boldsymbol{\beta}}_\lambda^{\mathrm{AL}} = \begin{pmatrix}\hat{\beta}_{\lambda,1}^{\mathrm{AL}}\\ \vdots \\ \hat{\beta}_{\lambda,k}^{\mathrm{AL}}\end{pmatrix} = \begin{pmatrix}\ell_{\lambda,1}\sqrt{d_1}z_1\\ \vdots \\ \ell_{\lambda,k}\sqrt{d_k}z_k\end{pmatrix} = \boldsymbol{L}_\lambda\boldsymbol{D}^{1/2}\boldsymbol{z}_1 = \boldsymbol{L}_\lambda\boldsymbol{D}^{1/2}\boldsymbol{P}_1'\boldsymbol{y} = \boldsymbol{L}_\lambda\boldsymbol{X}'\boldsymbol{y}.$$

Consequently, Theorem 1 is proved.

A.2. Proof of Lemma 1

First, we show the result about $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$. The $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ can be calculated as

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{AL}}(\lambda) &= \frac{1}{n}\boldsymbol{y}'\left\{\boldsymbol{P}\boldsymbol{P}' - \boldsymbol{P}\begin{pmatrix}\boldsymbol{I}_k\\ \boldsymbol{O}_{n-k,k}\end{pmatrix}\boldsymbol{D}^{1/2}\boldsymbol{L}_\lambda\boldsymbol{D}^{1/2}(\boldsymbol{I}_k, \boldsymbol{O}_{k,n-k})\boldsymbol{P}'\right\}^2\boldsymbol{y} \\ &= \frac{1}{n}\boldsymbol{z}'\left\{\boldsymbol{I}_n - \begin{pmatrix}\boldsymbol{D}^{1/2}\boldsymbol{L}_\lambda\boldsymbol{D}^{1/2} & \boldsymbol{O}_{k,n-k}\\ \boldsymbol{O}_{n-k,k} & \boldsymbol{O}_{n-k,n-k}\end{pmatrix}\right\}^2\boldsymbol{z} \\ &= \frac{1}{n}(\boldsymbol{z}_1', \boldsymbol{z}_2')\begin{pmatrix}(\boldsymbol{I}_k - \boldsymbol{D}^{1/2}\boldsymbol{L}_\lambda\boldsymbol{D}^{1/2})^2 & \boldsymbol{O}_{k,n-k}\\ \boldsymbol{O}_{n-k,k} & \boldsymbol{I}_{n-k,n-k}\end{pmatrix}\begin{pmatrix}\boldsymbol{z}_1\\ \boldsymbol{z}_2\end{pmatrix} \\ &= \frac{1}{n}\left\{\boldsymbol{z}_1'(\boldsymbol{I}_k - \boldsymbol{D}^{1/2}\boldsymbol{L}_\lambda\boldsymbol{D}^{1/2})^2\boldsymbol{z}_1 + \boldsymbol{z}_2'\boldsymbol{z}_2\right\},\end{aligned}$$

where $\boldsymbol{P}$ and $\boldsymbol{z}$ are the matrix of order n and the n-dimensional vector given by (2.1) and (A.1),

respectively. Notice that $\boldsymbol{I}_k - \boldsymbol{D}^{1/2}\boldsymbol{L}_\lambda\boldsymbol{D}^{1/2}$ is a diagonal matrix and that $\hat{\beta}_j^{\mathrm{LS}} = z_j/d_j^{1/2}$. When $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$, the jth diagonal element is expressed as

$$1 - \ell_{\lambda,j}d_j = 1 - S\left(1, \lambda/z_j^2\right) = S\left(\lambda/z_j^2, 1\right) + 1.$$

Moreover, $\boldsymbol{z}_2'\boldsymbol{z}_2$ can be expressed as

$$\boldsymbol{z}_2'\boldsymbol{z}_2 = \boldsymbol{y}'\boldsymbol{P}_2\boldsymbol{P}_2'\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_1\boldsymbol{P}_1')\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{D}^{-1}\boldsymbol{X}')\boldsymbol{y} = n\hat{\sigma}_0^2.$$

Hence, $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ is given by (2.9).

Next, we show the result about $\mathrm{df}_{\mathrm{AL}}(\lambda)$. When $w_j = 1/|\hat{\beta}_j^{\mathrm{LS}}|$, the jth element of $\boldsymbol{L}_\lambda\boldsymbol{D}$ is expressed as

$$\ell_{\lambda,j}d_j = 1 - \left\{S\left(\lambda/z_j^2, 1\right) + 1\right\} = -S\left(\lambda/z_j^2, 1\right).$$

Hence, $\mathrm{df}_{\mathrm{AL}}(\lambda)$ is given by (2.9). Consequently, Lemma 1 is proved.

A.3. Proof of Theorem 2

The $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ in Lemma 1 are rewritten as the following piecewise functions:

$$\hat{\sigma}^2_{\mathrm{AL}}(\lambda) = \hat{\sigma}^2_{\mathrm{AL},a}(\lambda) = \hat{\sigma}_0^2 + \frac{1}{n}(c_{1,a} + c_{2,a}\lambda^2) \quad (\lambda \in R_a), \tag{A.2}$$

$$\mathrm{df}_{\mathrm{AL}}(\lambda) = \mathrm{df}_{\mathrm{AL},a}(\lambda) = 1 + k - a - c_{2,a}\lambda \quad (\lambda \in R_a), \tag{A.3}$$

where R_a is the range given by (2.12), $\hat{\sigma}_0^2$ and t_j are given by (2.10) and (2.11), respectively, and $c_{1,a}$ and $c_{2,a}$ are nonnegative constants defined by

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

Hence, the GCV criterion for optimizing the tuning parameter is also expressed as a piecewise function, as follows

$$\mathrm{GCV}_{\mathrm{AL}}(\lambda) = \phi_a(\lambda) = \frac{\hat{\sigma}^2_{\mathrm{AL},a}(\lambda)}{\{1 - \mathrm{df}_{\mathrm{AL},a}(\lambda)/n\}^2} \quad (\lambda \in R_a).$$

In order to obtain $\hat{\lambda}$ that is the minimizer of the GCV, we have to solve the minimization problem of $\phi_a(\lambda)$. Since $c_{2,k} = 0$ when $a = k$, $\phi_k(\lambda)$ is the constant $\hat{\sigma}^2_\infty/(1 - n^{-1})^2$ at any $\lambda \in R_k$. When $a < k$, the derivative of $\phi_a(\lambda)$ is given by

$$\frac{d}{d\lambda}\phi_a(\lambda) = \frac{c_{2,a}}{n^2\{b + (a + c_{2,a}\lambda)\}^3} \cdot \psi_a(\lambda), \tag{A.4}$$

where

$$b = 1 - \frac{1}{n}(k+1), \quad \psi_a(\lambda) = (a+nb)\lambda - (n\hat{\sigma}_0^2 + c_{1,a}).$$

Here, by using $\psi_a(\lambda)$, we define the function $\psi(\lambda)$ ($\lambda \in (0, t_k]$) as

$$\psi(\lambda) = \psi_a(\lambda) \quad (\lambda \in R_a).$$

Since $c_{2,a}/n^2\{b + (a + c_{2,a}\lambda)\}^3$ that is the coefficient of $\psi_a(\lambda)$ in (A.4) is positive, it is enough to examine the sign of $\psi_a(\lambda)$ in order to search for the local minimum of $\phi_a(\lambda)$. Hence, we should find the point such that the sign of the linear function $\psi_a(\lambda)$ changes negative to positive. The $\psi_a(\lambda)$ is monotonic increasing function at any $\lambda \in R_a$ because $a + nb$ that is the gradient of the linear function is positive. It follows from the simple calculation that $\psi(\lambda)$ is continuous at any $\lambda \in (0, t_k]$, i.e.,

$$\psi_a(t_{a+1}) = \psi_{a+1}(t_{a+1}) \quad (a = 0, 1, \ldots, k-2).$$

Notice that $\psi_0(0) = -n\hat{\sigma}_0^2 < 0$ and $\psi_{k-1}(t_k) = (n-2)t_k - (n\hat{\sigma}_0^2 + c_{1,k-1}) > 0$. Hence, since $\psi(\lambda)$ is a piecewise increasing linear function with $\psi(0) < 0$ and $\psi(t_k) > 0$, λ satisfying $\psi(\lambda) = 0$ uniquely exists, i.e., the following statement is true:

$$\exists! a_* \in \{0, \ldots, k-1\} \; s.t. \; \psi_{a_*}(\lambda) = 0, \; \lambda \in R_{a_*}.$$

Notice that $n\hat{\sigma}_0^2 + c_{1,a} = (n-k-1+a)s_a^2$. Consequently, Theorem 2 is proved by solving $\psi_{a_*}(\lambda) = 0$.

A.4. Proof of Lemma 2

It follows from (A.2) and (A.3) that $\hat{\sigma}^2_{\mathrm{AL},a}(\lambda)$ is a monotonic increasing function and $\mathrm{df}_{\mathrm{AL},a}(\lambda)$ is a monotonic decreasing function. Notice that

$$\hat{\sigma}^2_{\mathrm{AL},a}(t_{a+1}) = \hat{\sigma}^2_{\mathrm{AL},a+1}(t_{a+1}), \quad \mathrm{df}_{\mathrm{AL},a}(t_{a+1}) = \mathrm{df}_{\mathrm{AL},a+1}(t_{a+1}) \quad (a = 0, 1, \ldots, k-2),$$

where t_j is the jth-order statistic given by (2.11). Hence, $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ is a continuous monotonic increasing function and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ is a continuous monotonic decreasing function. Moreover, Lemma 1 implies

$$\begin{aligned}
\hat{\sigma}^2_{\mathrm{AL}}(0) &= \hat{\sigma}_0^2, \quad \mathrm{df}_{\mathrm{AL}}(0) = k+1, \\
\lim_{\lambda\to\infty} \hat{\sigma}^2_{\mathrm{AL}}(\lambda) &= \lim_{\lambda\to\infty} \hat{\sigma}^2_{\mathrm{AL},k}(\lambda) = \frac{1}{n}(n\hat{\sigma}_0^2 + \boldsymbol{z}_1'\boldsymbol{z}_1) = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{y}, \\
\lim_{\lambda\to\infty} \mathrm{df}_{\mathrm{AL}}(\lambda) &= \lim_{\lambda\to\infty} \mathrm{df}_{\mathrm{AL},k}(\lambda) = 1.
\end{aligned}$$

Consequently, Lemma 2 is proved.

A.5. Proof of Lemma 4

From results in Yanagihara (2018), $\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta})$ in (2.17) and $\mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta})$ in (2.18) are expressed as

$$\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{\theta}) = \frac{1}{n}\left\{n\hat{\sigma}^2_0 + \sum_{j=1}^{k}\left(\frac{\theta_j}{d_j+\theta_j}\right)^2 z_j^2\right\}, \quad \mathrm{df}_{\mathrm{GR}}(\boldsymbol{\theta}) = 1 + k - \sum_{j=1}^{k}\frac{\theta_j}{d_j+\theta_j}.$$

These equations imply

$$\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{g}(h)) = \frac{1}{n}\left\{n\hat{\sigma}^2_0 + \sum_{j=1}^{k}\left(\frac{g_j(h)}{d_j+g_j(h)}\right)^2 z_j^2\right\}, \quad \mathrm{df}_{\mathrm{GR}}(\boldsymbol{g}(h)) = 1 + k - \sum_{j=1}^{k}\frac{g_j(h)}{d_j+g_j(h)}, \quad \text{(A.5)}$$

where z_j and $\hat{\sigma}^2_0$ are given by (2.2) and (2.10), respectively. Moreover, by using (3.4) and the soft-thresholding operator, we have

$$\frac{g_j(h)}{d_j+g_j(h)} = S(h/z_j^2, 1) + 1.$$

Hence, (A.5) can be expressed as

$$\hat{\sigma}^2_{\mathrm{GR}}(\boldsymbol{g}(h)) = \frac{1}{n}\left[n\hat{\sigma}^2_0 + \sum_{j=1}^{k}\left\{S(h/z_j^2, 1) + 1\right\}^2 z_j^2\right] = \hat{\sigma}^2_{\mathrm{AL}}(h),$$

$$\mathrm{df}_{\mathrm{GR}}(\boldsymbol{g}(h)) = 1 - \sum_{j=1}^{k} S(h/z_j^2, 1) = \mathrm{df}_{\mathrm{AL}}(h),$$

where $\hat{\sigma}^2_{\mathrm{AL}}(\lambda)$ and $\mathrm{df}_{\mathrm{AL}}(\lambda)$ are given by (2.9). Recall that $\mathrm{MSC}_{\mathrm{AL}}(h) = f(\hat{\sigma}^2_{\mathrm{AL}}(h), \mathrm{df}_{\mathrm{AL}}(h))$ from (3.1). Hence we have

$$\mathrm{MSC}_{\mathrm{AL}}(h) = \mathrm{MSC}_{\mathrm{GR}}(\boldsymbol{g}(h)).$$

Consequently, Lemma 4 is proved.