TR-No. 17-07, Hiroshima Statistical Research Group, 1–24

A Fast Algorithm for Optimizing Ridge Parameters in a Generalized Ridge Regression by Minimizing an Extended GCV Criterion

Mineaki Ohishi, Hirokazu Yanagihara and Yasunori Fujikoshi

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

Abstract

In this paper, we deal with an optimization of ridge parameters in a generalized ridge regression by minimizing a model selection criterion. Optimization methods based on minimizations of a generalized C_p (GC_p) criterion and GCV criterion have several important advantages and several major disadvantages. The most important advantage of two methods is that the speed of an optimization is very fast because minimizers of two criteria are given by closed forms. The serious disadvantage of the two methods is that the optimization methods do not work well when the number of explanatory variables is larger than the sample size. In order to improve the disadvantage while maintaining the advantage, at first we extend the GCV criterion, called the extended GCV (EGCV), by changing the second power of its denominator to the α-power, where α is some positive number larger than 2. Next, we propose a fast optimization algorithm to minimize the EGCV by using specific candidate minimizers of the EGCV of which the number is finite. By conducting numerical examinations, we verify that the optimization method based on the minimization of the EGCV performs well than those based on the minimizations of existing criteria.

(Last Modified: April 14, 2017)

Key words: Generalized ridge regression, Generalized cross-validation criterion, Linear regression model, Model selection criterion, Ridge parameters, Optimization of ridge parameters.

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

1. Introduction

A linear regression model is one of important tools to clarify the relationship between a scalar response variable and one or more explanatory variables. 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 ($\boldsymbol{X}'\mathbf{1}_n = \mathbf{0}_k$) with $\mathrm{rank}(\boldsymbol{X}) = m \leq \min\{n-1, k\}$, where n is the sample size, k is the number of explanatory variables, $\mathbf{1}_n$ is an n-dimensional vector of ones, and $\mathbf{0}_k$ is a k-dimensional vector of zeros. An equation of the linear regression model is

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

where μ is an unknown location parameter, $\boldsymbol{\beta}$ is a k-dimensional vector of unknown regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an n-dimensional vector of independent error variables from a distribution with the mean 0 and variance σ^2.

One of methods to avoid a serious problem of multicollinearity among explanatory variables is the generalized ridge regression (GRR) proposed by Hoerl and Kennard (1970), which gives a shrinkage estimate of $\boldsymbol{\beta}$ towards $\mathbf{0}_k$ via multiple ridge parameters $\theta_1, \ldots, \theta_k$, where $\theta_j \in \mathbb{R}_+ = \{\theta \in \mathbb{R} \mid \theta \geq 0\}$ $(j = 1, \ldots, k)$. Although the GRR was proposed 45 years ago, it is still used in many research fields (e.g., Liu *et al.*, 2011; Shen *et al.*, 2013). Since an estimator of the GRR estimator (GRRE) of $\boldsymbol{\beta}$ changes depending on $\theta_1, \ldots, \theta_k$, it is important how to optimize $\theta_1, \ldots, \theta_k$.

An optimization method based on minimization of a model selection criterion (MSC) with respect to $\theta_1, \ldots, \theta_k$ is one of common methods to optimize $\theta_1, \ldots, \theta_k$ (see e.g., Lawless, 1981; Walker & Page, 2001; Yanagihara, 2013). Most MSCs consist of two terms; one is a discrepancy term that measures a goodness of fit of a model, and the other is a penalty term that applies a penalty to a complexity of a model. The generalized C_p (GC_p) criterion (see Atkinson, 1980; Nagai *et al.*, 2012) and generalized cross-validation (GCV) criterion (see Craven & Wahba, 1979; Ye, 1998; Yanagihara, 2013) measure a goodness of fit of model via the residual sum of squares. An optimization method based on the minimization of GC_p criterion (called the GC_p-minimization method) has the following advantages and disadvantage:

- Advantages of the GC_p-minimization method.
  - (1) The speed of the optimization is very fast because minimizers of GC_p criterion can be derived by the closed forms (see Nagai *et al.*, 2012).
  - (2) The strength of the penalty term can be changed freely.
- A disadvantage of the GC_p-minimization method.
  - (1) This method cannot be used when n is smaller than k because the GC_p requires an asymptotically unbiased estimator of σ^2.

Since we encounter many opportunities to analyze data with huge explanatory variables, i.e.,

high-dimensional explanatory variables, we have to say that the GC_p-minimization method has a serious disadvantage. Since the serious problem is occurred from a requirement of an asymptotically unbiased estimator of σ^2, it is solvable by using a model selection criterion that does not require the asymptotically unbiased estimator. The GCV criterion is one of model selection criteria that do not require an asymptotically unbiased estimator of σ^2. However, an optimization method based on the minimization of GCV criterion (called the GCV-minimization method) also has the following advantages and disadvantage:

- Advantages of the GCV-minimization method.

  (1) The speed of the optimization is very fast because minimizers of the GCV criterion can be derived by the closed forms (see Yanagihara, 2013).

  (2) This method can be used even when n is smaller than k because the GCV does not require an asymptotically unbiased estimator of σ^2.

- A disadvantage of the GCV-minimization method.

  (1) The strength of the penalty term is fixed.

Although the GCV-minimization method can be used when n is smaller than k, this method does not work well then. This is because the GRRE hardly shrink towards $\mathbf{0}_k$ (see Yanagihara, 2013). The problem will be avoided if the strength of the penalty term in the GCV can be changed freely.

On the other hand, a goodness of fit of a model can be also measured by the Kullback-Leibler (KL) discrepancy (Kullback & Leibler, 1951). Hence, we can define a generalized information criterion (GIC; Nishii, 1984) for the GRR under a normal distribution assumption. An optimization method based on the minimization of GIC (called the GIC-minimization method) also has the following advantages and disadvantage:

- Advantages of the GIC-minimization method.

  (1) This method can be used even when n is smaller than k because the GIC does not require an asymptotically unbiased estimator of σ^2.

  (2) The strength of the penalty to a complexity of a model can be changed freely.

- A disadvantage of the GIC-minimization method.

  (1) The speed of the optimization is slow because minimizers of the GIC are derived by an iterative calculation, e.g., the Newton-Raphson method.

In this paper, at the beginning, we propose a fast optimization algorithm to solve a minimization problem of the GIC by using specific candidate minimizers of the GIC of which the number is finite. Unfortunately, we also show that the GIC-minimization method does not work well when n is smaller than k even if the strength of the penalty term is increased. Therefore, we extend the GCV so that the strength of the penalty term can be changed freely. We call this criterion an extended GCV (EGCV), and an optimization method based on the minimization of this criterion the EGCV-minimization method. We also propose a fast optimization algorithm to solve a minimization problem of the EGCV by using specific candidate minimizers of the EGCV of which the number is finite.

This paper is organized as follows: In section 2, we describe several results for deriving main theorems. In section 3, we give a fast optimization algorithm of the GIC-minimization method, and show the reason why the GIC-minimization method does not work well when k is larger than n. In section 4, we propose the EGCV criterion and fast optimization algorithm to solve a minimization problem of EGCV. Numerical examinations are conducted in Section 5. Technical details are provided in the Appendix.

2. Preliminaries

It follows from the results in Yanagihara (2013) that the GRRE of $\boldsymbol{\beta}$ and a hat matrix of the GRR are invariant to any changes in $\theta_{m+1}, \ldots, \theta_k$ when $m < k$. From this fact, we set $\theta_{m+1} = \cdots = \theta_k = \infty$ to reduce a variance of the GRRE of $\boldsymbol{\beta}$. Therefore, it is enough just to minimize MSC with respect to $\theta_1, \ldots, \theta_m$ for optimizing ridge parameters in the GRR.

Let $\boldsymbol{Q}$ be a kth orthogonal matrix that diagonalizes $\boldsymbol{X}'\boldsymbol{X}$, i.e.,

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

where $\boldsymbol{D}$ is an mth diagonal matrix defined by

$$\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_m), \quad d_1 \geq \cdots \geq d_m \text{ are nonzero eigenvalues of } \boldsymbol{X}'\boldsymbol{X}, \tag{2.1}$$

and $\boldsymbol{O}_{k,m}$ is a $k \times m$ matrix of zeros, and let $\boldsymbol{\theta}$ be an m-dimensional vector given by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \in \mathbb{R}_+^m$, and $\boldsymbol{\Theta}$ be a kth diagonal matrix given by $\boldsymbol{\Theta} = \mathrm{diag}(\theta_1, \ldots, \theta_m, 0, \ldots, 0)$, where $\mathbb{R}_+^m$ is the mth Cartesian power set of $\mathbb{R}_+$. Notice that $d_1, \ldots, d_m$ are positive because $\boldsymbol{X}'\boldsymbol{X}$ is a positive semidefinite matrix. Moreover, let $\boldsymbol{M}_{\boldsymbol{\theta}}$ be a kth square matrix defined by $\boldsymbol{M}_{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}'$. In particular, we write $\boldsymbol{M}_{\boldsymbol{\theta}} = \boldsymbol{M}$ when $\boldsymbol{\theta} = \boldsymbol{0}_m$. Then, the GRRE of $\boldsymbol{\beta}$ is expressed by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} = \boldsymbol{M}_{\boldsymbol{\theta}}^{+}\boldsymbol{X}'\boldsymbol{y}, \tag{2.2}$$

where $\boldsymbol{A}^+$ is the Moore-Penrose inverse matrix of a matrix $\boldsymbol{A}$ (for details of the Moore-Penrose inverse matrix, see e.g., Harville, 1997, chap. 20). It is well known fact that the least square estimator (LSE) of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \boldsymbol{M}^+ \boldsymbol{X}' \boldsymbol{y}. \tag{2.3}$$

It is clear that (2.2) coincides with (2.3) when $\boldsymbol{\theta} = \boldsymbol{0}_m$. Let $\hat{\mu}$ be the LSE of μ given by $\hat{\mu} = \bar{y}$, where $\bar{y}$ is the sample mean of $y_1, \ldots, y_n$, i.e., $\bar{y} = \sum_{j=1}^n y_i/n$. The equation (2.2) and $\hat{\mu}$ imply a predictive value of $\boldsymbol{y}$ as

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

where $\boldsymbol{J}_n$ is an $n \times n$ projection matrix defined by $\boldsymbol{J}_n = \boldsymbol{1}_n\boldsymbol{1}_n'/n$. By using $\boldsymbol{M}_{\boldsymbol{\theta}}^+$ and $\hat{\boldsymbol{y}}_{\boldsymbol{\theta}}$, we define an estimator of σ^2 and a generalized degree of freedom (df) of the GRR that are main bodies of the discrepancy and penalty terms of MSC, respectively, as

$$\hat{\sigma}^2(\boldsymbol{\theta}) = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\boldsymbol{\theta}}) = \frac{1}{n}\boldsymbol{y}'\left(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{X}\boldsymbol{M}_{\boldsymbol{\theta}}^+\boldsymbol{X}'\right)^2\boldsymbol{y}, \tag{2.4}$$

$$\mathrm{df}(\boldsymbol{\theta}) = \mathrm{tr}(\boldsymbol{J}_n + \boldsymbol{X}\boldsymbol{M}_{\boldsymbol{\theta}}^+\boldsymbol{X}') = 1 + \mathrm{tr}(\boldsymbol{M}_{\boldsymbol{\theta}}^+\boldsymbol{M}). \tag{2.5}$$

Most MSCs for the GRR are defined by a bivariate function of $\hat{\sigma}^2(\boldsymbol{\theta})$ and $\mathrm{df}(\boldsymbol{\theta})$.

Let $\boldsymbol{z}$ be an m-dimensional vector defined by

$$\boldsymbol{z} = (z_1, \ldots, z_m)' = (\boldsymbol{D}^{-1/2}, \boldsymbol{O}_{m,k-m})\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{D}^{-1/2}\boldsymbol{Q}_1'\boldsymbol{X}'\boldsymbol{y}, \tag{2.6}$$

where $\boldsymbol{Q}_1$ is a $k \times m$ matrix defined by $\boldsymbol{Q} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2)$. Here, we assume that all $z_1, \ldots, z_m$ are not 0. If z_j is accidentally 0, then we set $d_j = 0$ and use $(z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_m)'$ as $\boldsymbol{z}$. By using the result in Yanagihara (2013), $\hat{\sigma}^2(\boldsymbol{\theta})$ in (2.4) and $\mathrm{df}(\boldsymbol{\theta})$ in (2.5) can be rewritten as

$$\hat{\sigma}^2(\boldsymbol{\theta}) = \frac{1}{n}\left\{n\hat{\sigma}_0^2 + \sum_{j=1}^m \left(\frac{\theta_j}{d_j + \theta_j}\right)^2 z_j^2\right\}, \quad \mathrm{df}(\boldsymbol{\theta}) = 1 + m - \sum_{j=1}^m \frac{\theta_j}{d_j + \theta_j}, \tag{2.7}$$

where $\hat{\sigma}_0^2$ is the maximum likelihood estimator of σ^2 under normality, which is given by

$$\hat{\sigma}_0^2 = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{X}\boldsymbol{M}^+\boldsymbol{X}')\boldsymbol{y}. \tag{2.8}$$

It should be kept in mind that $\hat{\sigma}_0^2 = 0$ holds in most cases of the linear regression with high-dimensional explanatory variables. Notice that $\theta_j/(d_j + \theta_j)$ monotonically increases θ_j, and $\hat{\sigma}^2(\boldsymbol{0}_m) = \hat{\sigma}_0^2$, $\mathrm{df}(\boldsymbol{0}_m) = m + 1$, and

$$\lim_{\theta_1 \to \infty, \ldots, \theta_m \to \infty} \hat{\sigma}^2(\boldsymbol{\theta}) = \hat{\sigma}_\infty^2 = \frac{1}{n}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{y}, \quad \lim_{\theta_1 \to \infty, \ldots, \theta_m \to \infty} \mathrm{df}(\boldsymbol{\theta}) = 1. \tag{2.9}$$

Hence, the ranges of $\hat{\sigma}^2(\boldsymbol{\theta})$ and $\mathrm{df}(\boldsymbol{\theta})$ are given by

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

Let $f(r,u)$ be a bivariate function which satisfies the following assumptions:

(A1) $f(r,u)$ is a continuous function at any $(r,u) \in (0, \hat{\sigma}^2_\infty] \times [1,n)$.

(A2) $f(r,u) > 0$ for any $(r,u) \in (0, \hat{\sigma}^2_\infty] \times [1,n)$.

(A3) $f(r,u)$ is the first-order partially differentiable at any $(r,u) \in (0, \hat{\sigma}^2_\infty] \times [1,n)$, and

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

Let $\mathbb{D}$ be the domain of f. It is easy to see that $m = n-1 \Rightarrow \hat{\sigma}^2_0 = 0$. Recall that $m \leq n-1$. Hence, we have

$$\hat{\sigma}^2_0 \neq 0 \Rightarrow \mathbb{D} \subseteq (0, \hat{\sigma}^2_\infty] \times [1,n).$$

By using the bivariate function f, MSC for the GRR can be expressed as

$$\mathrm{MSC}(\boldsymbol{\theta}) = f(\hat{\sigma}^2(\boldsymbol{\theta}), \mathrm{df}(\boldsymbol{\theta})). \tag{2.10}$$

The relationship between an existing criterion and the bivariate function f is as follows:

$$f(r,u) = \begin{cases} f_{GC_p}(r,u) = nr/s^2_0 + \alpha u & (GC_p : \text{when } \hat{\sigma}^2_0 \neq 0) \\ f_{\mathrm{GCV}}(r,u) = r(1-u/n)^{-2} & (\mathrm{GCV} : \text{when } u \neq n) \\ f_{\mathrm{GIC}}(r,u) = r\exp(\alpha u/n) & (\mathrm{GIC}) \end{cases}, \tag{2.11}$$

where α is some positive value expressing the strength of the penalty term, and s^2_0 is an ordinary unbiased estimator of σ^2 when $m < n-1$, which is given by $s^2_0 = n\hat{\sigma}^2_0/(n-m-1)$. The GC_p includes several famous MSCs, e.g., the GC_p coincides with the C_p proposed by Mallows (1973) when $\alpha = 2$, and the modified C_p (MC_p) proposed by Fujikoshi and Satoh (1997) and Yanagihara *et al*. (2009) when $\alpha = 1 + 2/(n-m-1)$. Moreover, the original GIC under normality is expressed as $n\log r + \alpha u$. The GIC in this paper is defined by an exponential transformation of the original GIC divided by n. The GIC includes several famous MSCs, e.g., the GIC coincides the AIC proposed by Akaike (1973) when $\alpha = 2$, the HQC proposed by Hannan and Quinn (1978) when $\alpha = 2\log\log n$, and the BIC proposed by Schwarz (1978) when $\alpha = \log n$.

In order to express an optimized GRRE of $\boldsymbol{\beta}$ systematically, we prepare the following class of $\boldsymbol{\theta}$ for any $h \in \mathbb{R}_+$:

$$\hat{\boldsymbol{\theta}}(h) = (\hat{\theta}_1(h), \ldots, \hat{\theta}_m(h))', \quad \hat{\theta}_j(h) = \begin{cases} \dfrac{d_j h}{z^2_j - h} & (h \leq z^2_j) \\ \infty & (h > z^2_j) \end{cases}. \tag{2.12}$$

The GRRE of $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\theta}}(\hat{h})} = \hat{\boldsymbol{\beta}}(\hat{h}) = \boldsymbol{Q}_1 \boldsymbol{V}(\hat{h}) \boldsymbol{Q}_1' \hat{\boldsymbol{\beta}}, \tag{2.13}$$

where $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$ given by (2.3), and $\boldsymbol{V}(h)$ is a kth diagonal matrix of which the jth diagonal element is given by

$$v_j(h) = \frac{d_j}{d_j + \hat{\theta}_j(h)} = \begin{cases} 1 - \dfrac{h}{z_j^2} & (h \leq z_j^2) \\ 0 & (h > z_j^2) \end{cases}.$$

An optimized $\boldsymbol{\theta}$ by minimizing $\mathrm{MSC}(\boldsymbol{\theta})$ when the domain of f is a subset of $(0, \hat{\sigma}_\infty^{-2}] \times [1, n)$ is expressed as in the following theorem (the proof is given in Appendix A.1):

Lemma 1. *Let*

$$\phi(h) = \mathrm{MSC}(\hat{\boldsymbol{\theta}}(h)) \quad (h \in \mathbb{R}_+ \backslash \{0\}). \tag{2.14}$$

Suppose that $\mathbb{D} \subseteq (0, \hat{\sigma}_\infty^{-2}] \times [1, n)$ *and* $\exists \xi > 0$ *s.t.* $\phi(\xi) < \lim_{h \to 0} \phi(h)$*. Then, optimized ridge parameters by minimizing the* $\mathrm{MSC}(\boldsymbol{\theta})$ *are given by* $\hat{\boldsymbol{\theta}}(\hat{h})$*, where* $\hat{h}$ *is given by*

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

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

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

and let R_j $(j = 0, \ldots, m)$ be a range defined by

$$R_j = \begin{cases} (0, t_1] & (j = 0) \\ (t_j, t_{j+1}] & (j = 1, \ldots, m-1) \\ (t_m, \infty) & (j = m) \end{cases}. \tag{2.15}$$

By using t_j and R_j, we can clarify properties of $\phi(h)$ given in (2.14) as in the following lemma (the proof is given in Appendix A.2):

Lemma 2. *The* $\phi(h)$ *satisfies the following properties:*

(P1) $\phi(h)$ *is a continuous at any* $h \in \mathbb{R}_+ \backslash \{0\}$.

(P2) $\phi(h) = f(\hat{\sigma}_\infty^{-2}, 1)$ *for every* $h \geq t_m$.

(P3) *$\phi(h)$ is a piecewise function as*

$$\phi(h) = \phi_a(h) \ (h \in R_a, a = 0, 1, \ldots, m).$$

The specific form of $\phi_a(h)$ can be expressed as

$$\phi_a(h) = f\left(\hat{\sigma}_0^2 + (c_{1,a} + h^2 c_{2,a})/n, 1 + m - a - hc_{2,a}\right),$$

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

$$c_{1,a} = \sum_{j=1}^{a} t_j, \quad c_{2,a} = \sum_{j=a+1}^{m} \frac{1}{t_j}. \tag{2.16}$$

Let s_a^2 $(a = 0, 1, \ldots, m)$ be a statistic defined by

$$s_a^2 = \frac{n\hat{\sigma}_0^2 + c_{1,a}}{n - m - 1 + a}, \tag{2.17}$$

and, let a^* be a positive integer defined by

$$a^* \in \{0, 1, \ldots, m-1\} \ s.t. \ s_{a^*}^2 \in R_{a^*}, \tag{2.18}$$

From Lemma 2.1 in Yanagihara (2013), we can see that a^* exists unique when $\hat{\sigma}_0^2 \neq 0$. Let

$$GC_p(\boldsymbol{\theta}) = f_{GC_p}(\hat{\sigma}^2(\boldsymbol{\theta}), \mathrm{df}(\boldsymbol{\theta})), \quad \mathrm{GCV}(\boldsymbol{\theta}) = f_{\mathrm{GCV}}(\hat{\sigma}^2(\boldsymbol{\theta}), \mathrm{df}(\boldsymbol{\theta})).$$

From the results in Nagai *et al*. (2012) and Yanagihara (2013), the minimizers of the GC_p and GCV can be expressed as in the following theorem (the proof is omitted because it is easy to obtain from the results in Nagai *et al*., 2012; Yanagihara, 2013):

Theorem 1. *Optimized ridge parameters minimizing $GC_p(\boldsymbol{\theta})$ and $\mathrm{GCV}(\boldsymbol{\theta})$ are $\hat{\boldsymbol{\theta}}(\hat{h}_{GC_p})$ and $\hat{\boldsymbol{\theta}}(\hat{h}_{\mathrm{GCV}})$, respectively, where $\hat{h}_{GC_p}$ and $\hat{h}_{\mathrm{GCV}}$ are defined as*

$$\hat{h}_{GC_p} = \frac{1}{2}\alpha s_0^2 \ (\hat{\sigma}_0^2 \neq 0), \quad \hat{h}_{\mathrm{GCV}} = \begin{cases} s_{a^*}^2 & (\textit{when } \hat{\sigma}_0^2 \neq 0) \\ t_1 & (\textit{when } \hat{\sigma}_0^2 = 0 \textit{ and } m = n-1) \\ 0 & (\textit{when } \hat{\sigma}_0^2 = 0 \textit{ and } m < n-1) \end{cases},$$

where s_a^2 is given by (2.17) *and a_* is the integer given by* (2.18).

By using the results in Theorem 1, optimized GRREs of $\boldsymbol{\beta}$ by minimizing the GC_p and GCV can be derived as $\hat{\boldsymbol{\beta}}(\hat{h}_{GC_p})$ and $\hat{\boldsymbol{\beta}}(\hat{h}_{\mathrm{GCV}})$, respectively.

3. Algorithm to Solve the Minimization Problem of the GIC

Theorem 1 in the previous section indicates that the GC_p-minimization method cannot be used when $\hat{\sigma}_0^2$ in 2.8 is 0, i.e., most cases of the linear regression with high dimensional explanatory variables, the GCV-minimization method hardly shrink the GRRE of $\boldsymbol{\beta}$ towards $\mathbf{0}_k$ when $\hat{\sigma}_0^2 = 0$. On the other hand, the GIC is one of model selection criteria that can be defined even if $\hat{\sigma}_0^2 = 0$ and of which the size of the penalty term can be changed freely. However, at the moment, there is no efficient algorithm to solve the minimization problem of the GIC. Hence, we propose the fast algorithm for the GIC-minimization method.

Let

$$\mathrm{GIC}(\boldsymbol{\theta}) = f_{\mathrm{GIC}}(\hat{\sigma}^2(\boldsymbol{\theta}), \mathrm{df}(\boldsymbol{\theta})), \tag{3.1}$$

where $f_{\mathrm{GIC}}(r, u)$ is given by (2.11), and $\hat{\sigma}^2(\boldsymbol{\theta})$ and $\mathrm{df}(\boldsymbol{\theta})$ are given by (2.7), and let

$$\phi_{\mathrm{GIC}}(h) = f_{\mathrm{GIC}}(\hat{\sigma}^2(\hat{\boldsymbol{\theta}}(h)), \mathrm{df}(\hat{\boldsymbol{\theta}}(h))) \quad (h \in \mathbb{R}_+ \backslash \{0\}),$$

where $\hat{\boldsymbol{\theta}}(h)$ is given by (2.12). From Lemma 2, we can see that $\phi_{\mathrm{GIC}}(h)$ is a piecewise function as

$$\phi_{\mathrm{GIC}}(h) = \phi_{\mathrm{GIC},a}(h) \quad (h \in R_a, a = 0, 1, \ldots, m), \tag{3.2}$$

where R_a is the range given by (2.15) and $\phi_{\mathrm{GIC},a}(h)$ is defined by

$$\phi_{\mathrm{GIC},a}(h) = \left\{\hat{\sigma}_0^2 + \frac{1}{n}(c_{1,a} + h^2 c_{2,a})\right\} \exp\left\{\frac{1}{n}\alpha(1 + m - a - hc_{2,a})\right\}.$$

Here $c_{1,a}$ and $c_{2,a}$ $(a = 0, 1, \ldots, m)$ are given by (2.16). Furthermore, we define the function $\psi_{\mathrm{GIC}}(h)$ $(h \in (0, t_m])$ which is a piecewise function as

$$\psi_{\mathrm{GIC}}(h) = \psi_{\mathrm{GIC},a}(h) \quad (h \in R_a, a = 0, 1, \ldots, m - 1), \tag{3.3}$$

where $\psi_{\mathrm{GIC},a}(h)$ is defined by

$$\begin{aligned}\psi_{\mathrm{GIC},a}(h) &= \frac{n^2}{c_{2,a} \exp\{\alpha(1 + m - a - hc_{2,a})/n\}} \frac{\partial}{\partial h} \phi_{\mathrm{GIC},a}(h) \\ &= -\alpha c_{2,a} h^2 + 2nh - \alpha(n\hat{\sigma}_0^2 + c_{1,a}).\end{aligned} \tag{3.4}$$

The minimizer of GIC can be expressed as in the following theorem (the proof is given in Appendix A.3):

Theorem 2. *Let $\xi_{\mathrm{GIC},a}$ be one of the two distinct roots or the double root of the quadratic equation $\psi_{\mathrm{GIC},a}(h) = 0$ as*

$$\xi_{\mathrm{GIC},a} = \frac{n - \sqrt{n^2 - \alpha^2 c_{2,a}(n\hat{\sigma}_0^2 + c_{1,a})}}{\alpha c_{2,a}}, \tag{3.5}$$

and let $\mathcal{A}_{\mathrm{GIC}}$ *and* $\mathcal{T}_{\mathrm{GIC}}$ *be sets defined by*

$$\mathcal{A}_{\mathrm{GIC}} = \{a \in \{0, 1, \ldots, m-1\} \mid \xi_{\mathrm{GIC},a} \in R_a\}, \quad \mathcal{T}_{\mathrm{GIC}} = \begin{cases} \{t_m\} & (\textit{when } \hat{\sigma}_\infty^2 > 2t_m/\alpha) \\ \{\emptyset\} & (\textit{when } \hat{\sigma}_\infty^2 \leq 2t_m/\alpha) \end{cases}, \tag{3.6}$$

where $\hat{\sigma}_\infty^2$ *is given by* (2.9)*. We define a set of candidate minimizers of* $\phi_{\mathrm{GIC}}(h)$ *as*

$$\mathcal{S}_{\mathrm{GIC}} = \left\{ \bigcup_{a \in \mathcal{A}_{\mathrm{GIC}}} \{\xi_{\mathrm{GIC},a}\} \right\} \bigcup \mathcal{T}_{\mathrm{GIC}}. \tag{3.7}$$

Then, an optimized ridge parameters by minimizing $\mathrm{GIC}(\boldsymbol{\theta})$ *are given by* $\hat{\boldsymbol{\theta}}(\hat{h}_{\mathrm{GIC}})$*, where* $\hat{h}_{\mathrm{GIC}}$ *is given by*

$$\hat{h}_{\mathrm{GIC}} = \begin{cases} \arg\min_{h \in \mathcal{S}_{\mathrm{GIC}}} \phi_{\mathrm{GIC}}(h) & (\textit{when } \hat{\sigma}_0^2 \neq 0) \\ 0 & (\textit{when } \hat{\sigma}_0^2 = 0) \end{cases}. \tag{3.8}$$

It should be emphasized that elements of $\mathcal{S}_{\mathrm{GIC}}$ can be written by closed forms, and the number of elements of $\mathcal{S}_{\mathrm{GIC}}$ is equal to or smaller than $m + 1$, i.e., $\#(\mathcal{S}_{\mathrm{GIC}}) \leq m + 1$. Hence, when $\hat{\sigma}_0^2 \neq 0$, by using the results in Theorem 2, we can optimize $\boldsymbol{\theta}$ quickly as follows:

A Fast Algorithm to Derive the Optimized GRRE of $\boldsymbol{\beta}$ by Minimizing the GIC

(1) Calculate elements of $\mathcal{S}_{\mathrm{GIC}}$ by using $\xi_{\mathrm{GIC},a}$, $\mathcal{A}_{\mathrm{GIC}}$ and $\mathcal{T}_{\mathrm{GIC}}$.

(2) Determine $\hat{h}_{\mathrm{GIC}}$ by comparing function values of $\phi_{\mathrm{GIC}}(h)$ at the points in $\mathcal{S}_{\mathrm{GIC}}$.

(3) Calculate the optimized GRRE of $\boldsymbol{\beta}$ by minimizing the GIC as $\hat{\boldsymbol{\beta}}(\hat{h}_{\mathrm{GIC}})$, where $\hat{\boldsymbol{\beta}}(h)$ is given by (2.13).

Unfortunately, Theorem 2 also indicates that the optimized GRRE of $\boldsymbol{\beta}$ by minimizing GIC is always equal to the LSE of $\boldsymbol{\beta}$ when $\hat{\sigma}_0^2 = 0$. Thus, we cannot help but say that the GIC-minimization method does not work well in the case of high-dimensional explanatory variables.

When $\hat{\sigma}_0^2 \neq 0$, we cannot derive $\#(\mathcal{S}_{\mathrm{GIC}})$ without comparing function values of $\phi_{\mathrm{GIC}}(h)$. However, $\#(\mathcal{S}_{\mathrm{GIC}})$ becomes 1 under the specific α when $\hat{\sigma}_0^2 \neq 0$. The result is summarized as in the following corollary (the proof is given as Appendix A.4):

Corollary 1. *Suppose that* $\alpha \leq n/m$ *and* $\hat{\sigma}_0^2 \neq 0$*. Let* $a^\dagger$ *be a positive integer defined by*

$$a^\dagger \in \{0, 1, \ldots, m-1\} \textit{ s.t. } \xi_{\mathrm{GIC},a^\dagger} \in R_{a^\dagger}.$$

The integer $a^\dagger$ exists unique when $\hat{\sigma}^2_\infty \leq 2t_m/\alpha$. Then $\hat{h}_{\rm GIC}$ in (3.8) *can be rewritten as*

$$\hat{h}_{\rm GIC} = \begin{cases} \xi_{{\rm GIC},a^\dagger} & (when\ \hat{\sigma}^2_\infty \leq 2t_m/\alpha) \\ t_m & (when\ \hat{\sigma}^2_\infty > 2t_m/\alpha) \end{cases}.$$

4. Algorithm to Solve the Minimization Problem of the EGCV

Theorem 2 in the previous section indicates that the GIC-minimization method does not shrink the GRRE of $\boldsymbol{\beta}$ towards $\mathbf{0}_k$ at all when $\hat{\sigma}^2_0$ in (2.8) is 0, i.e., most cases of the linear regression with high dimensional explanatory variables. From Theorems 1 and 2, we would have to say regrettably that there are no existing MSCs which work well when $\hat{\sigma}^2_0 = 0$. Hence, we have to propose new MSC which works well even when $\hat{\sigma}^2_0 = 0$. On the other hand, the cause of the problem in GCV is that the strength of the penalty to a complexity of a model is small. Hence, we extend the GCV so that the size of the penalty to a complexity of a model can be changed freely. It is called the extended GCV (EGCV) in this paper. Let

$$f_{\rm EGCV}(r,u) = \frac{r}{(1-u/n)^\alpha},\ \alpha > 2 \text{ and } u < n. \tag{4.1}$$

Then, the EGCV for the GRR is defined as

$$\mathrm{EGCV}(\boldsymbol{\theta}) = f_{\rm EGCV}(\hat{\sigma}^2(\boldsymbol{\theta}), \mathrm{df}(\boldsymbol{\theta})), \tag{4.2}$$

where $\hat{\sigma}^2(\boldsymbol{\theta})$ and $\mathrm{df}(\boldsymbol{\theta})$ are given by (2.7). It is easy to see that the EGCV coincides with the GCV if $\alpha = 2$.

Let

$$\phi_{\rm EGCV}(h) = f_{\rm EGCV}(\hat{\sigma}^2(\hat{\boldsymbol{\theta}}(h)), \mathrm{df}(\hat{\boldsymbol{\theta}}(h))) \quad (h \in \mathbb{R}_+),$$

where $\hat{\boldsymbol{\theta}}(h)$ is given by (2.12). From Lemma 2, we can see that $\phi_{\rm EGCV}(h)$ is a piecewise function as

$$\phi_{\rm EGCV}(h) = \phi_{{\rm EGCV},a}(h) \quad (h \in R_a, a = 0, 1, \ldots, m), \tag{4.3}$$

where R_a is the range given by (2.15) and $\phi_{{\rm EGCV},a}(h)$ is defined by

$$\phi_{{\rm EGCV},a}(h) = \frac{\hat{\sigma}^2_0 + c_{1,a}/n + h^2 c_{2,a}/n}{(b + a/n + h c_{2,a}/n)^\alpha}. \tag{4.4}$$

Here $c_{1,a}$ and $c_{2,a}$ $(a = 0, 1, \ldots, m)$ are given by (2.16), and b is the constant defined by

$$b = 1 - \frac{1}{n}(m+1). \tag{4.5}$$

Furthermore, we define the function $\psi_{\rm EGCV}(h)$ $(h \in (0, t_m])$ which is a piecewise function as

$$\psi_{\mathrm{EGCV}}(h) = \psi_{\mathrm{EGCV},a}(h) \quad (h \in R_a, a = 0, 1, \ldots, m-1), \tag{4.6}$$

where $\psi_{\mathrm{EGCV},a}(h)$ is defined by

$$\begin{aligned}\psi_{\mathrm{EGCV},a}(h) &= \frac{n^2(b + a/n + c_{2,a}h/n)^{\alpha+1}}{c_{2,a}} \frac{\partial}{\partial h}\phi_{\mathrm{EGCV},a}(h) \\ &= -(\alpha-2)c_{2,a}h^2 + 2(a+nb)h - \alpha(n\hat{\sigma}_0^2 + c_{1,a}).\end{aligned} \tag{4.7}$$

The minimizer of EGCV can be expressed as in the following theorem (the proof is given in Appendix A.5):

Theorem 3. *Let $\xi_{\mathrm{EGCV},a}$ be one of the two distinct roots or the double root of the quadratic equation $\psi_{\mathrm{EGCV},a} = 0$ as*

$$\xi_{\mathrm{EGCV},a} = \frac{(a+nb) - \sqrt{(a+nb)^2 - \alpha(\alpha-2)c_{2,a}(n\hat{\sigma}_0^2 + c_{1,a})}}{(\alpha-2)c_{2,a}},$$

and let $\mathcal{A}_{\mathrm{EGCV}}$ and $\mathcal{T}_{\mathrm{EGCV}}$ be sets defined by

$$\mathcal{A}_{\mathrm{EGCV}} = \{a \in \{0, 1, \ldots, m-1\} \mid \xi_{\mathrm{EGCV},a} \in R_a\}, \ \mathcal{T}_{\mathrm{EGCV}} = \begin{cases} \{t_m\} & (\textit{when } \hat{\sigma}_\infty^2 > 2(1-n^{-1})t_m/\alpha) \\ \{\emptyset\} & (\textit{when } \hat{\sigma}_\infty^2 \le 2(1-n^{-1})t_m/\alpha) \end{cases},$$

where $\hat{\sigma}_\infty^2$ is given by (2.9)*. We define a set of candidate minimizers of $\phi_{\mathrm{EGCV}}(h)$ as*

$$\mathcal{S}_{\mathrm{EGCV}} = \left\{ \bigcup_{a \in \mathcal{A}_{\mathrm{EGCV}}} \{\xi_{\mathrm{EGCV},a}\} \right\} \bigcup \mathcal{T}_{\mathrm{EGCV}}. \tag{4.8}$$

Then, an optimized ridge parameters by minimizing $\mathrm{EGCV}(\boldsymbol{\theta})$ *are given by $\hat{\boldsymbol{\theta}}(\hat{h}_{\mathrm{EGCV}})$, where $\hat{h}_{\mathrm{EGCV}}$ is given by*

$$\hat{h}_{\mathrm{EGCV}} = \begin{cases} \arg\min_{h \in \mathcal{S}_{\mathrm{EGCV}}} \phi_{\mathrm{EGCV}}(h) & (\textit{when } \hat{\sigma}_0^2 \neq 0 \textit{ or } b = 0) \\ 0 & (\textit{when } \hat{\sigma}_0^2 = 0 \textit{ and } b \neq 0) \end{cases}. \tag{4.9}$$

It should be emphasized that elements of $\mathcal{S}_{\mathrm{EGCV}}$ can be written by closed forms, and the number of elements of $\mathcal{S}_{\mathrm{EGCV}}$ is equal to or smaller than $m+1$, i.e., $\#(\mathcal{S}_{\mathrm{EGCV}}) \le m+1$. Hence, when $\hat{\sigma}_0^2 \neq 0$ or $b = 0$, by using the results in Theorem 3, we can optimize $\boldsymbol{\beta}$ quickly as follows:

A Fast Algorithm to Derive the Optimized GRRE of $\boldsymbol{\beta}$ by Minimizing the EGCV

(1) Calculate elements of $\mathcal{S}_{\mathrm{EGCV}}$ by using $\xi_{\mathrm{EGCV},a}$, $\mathcal{A}_{\mathrm{EGCV}}$ and $\mathcal{T}_{\mathrm{EGCV}}$.

(2) Determine $\hat{h}_{\mathrm{EGCV}}$ by comparing function values of $\phi_{\mathrm{EGCV}}(h)$ at the points in $\mathcal{S}_{\mathrm{EGCV}}$.

(3) Calculate the optimized GRRE of $\boldsymbol{\beta}$ by minimizing the EGCV as $\hat{\boldsymbol{\beta}}(\hat{h}_{\mathrm{EGCV}})$, where $\hat{\boldsymbol{\beta}}(h)$ is given by (2.13).

When $\hat{\sigma}_0^2 \neq 0$ or $m = n-1$, we cannot derive $\#(\mathcal{S}_{\mathrm{EGCV}})$ without comparing function values of $\phi_{\mathrm{EGCV}}(h)$. However, $\#(\mathcal{S}_{\mathrm{EGCV}})$ becomes 1 under the specific α when $\hat{\sigma}_0^2 \neq 0$ or $m = n-1$. The result is summarized as in the following corollary (the proof is given as Appendix A.6):

Corollary 2. *Suppose that* $\alpha \leq (n+m-1)/m$, *and* $\{\hat{\sigma}_0^2 \neq 0\} \vee \{b = 0\}$. *Let* $a^{\ddagger}$ *be a positive integer defined by*

$$a^{\ddagger} \in \{0, 1, \ldots, m-1\} \; s.t. \; \xi_{\mathrm{EGCV},a^{\ddagger}} \in R_{a^{\ddagger}}.$$

The integer $a^{\ddagger}$ *exists unique when* $\hat{\sigma}_{\infty}^2 \leq 2(1-n^{-1})t_m/\alpha$. *Then* $\hat{h}_{\mathrm{EGCV}}$ *in* (4.9) *can be rewritten as*

$$\hat{h}_{\mathrm{EGCV}} = \begin{cases} \xi_{\mathrm{EGCV},a^{\ddagger}} & (when\ \hat{\sigma}_{\infty}^2 \leq 2(1-n^{-1})t_m/\alpha) \\ t_m & (when\ \hat{\sigma}_{\infty}^2 > 2(1-n^{-1})t_m/\alpha) \end{cases}.$$

5. Numerical Study

In this section, we compared performances of the GC_p-, GIC- and EGCV-minimization methods with $\alpha = 2$, $2\log\log n$ and $\log n$ by conducting numerical examinations. We generated data from the simulation model $N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n)$, 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 of which elements are identically and independently distributed according to $U(-1, 1)$, $\boldsymbol{\Phi}(\rho)$ is a $k \times k$ symmetric matrix of which the (i, j)th element is $\rho^{|i-j|}$ and $\boldsymbol{\eta}$ is an n-dimensional vector of which the jth element is given by

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

The simulation model is the same as that in Yanagihara (2013).

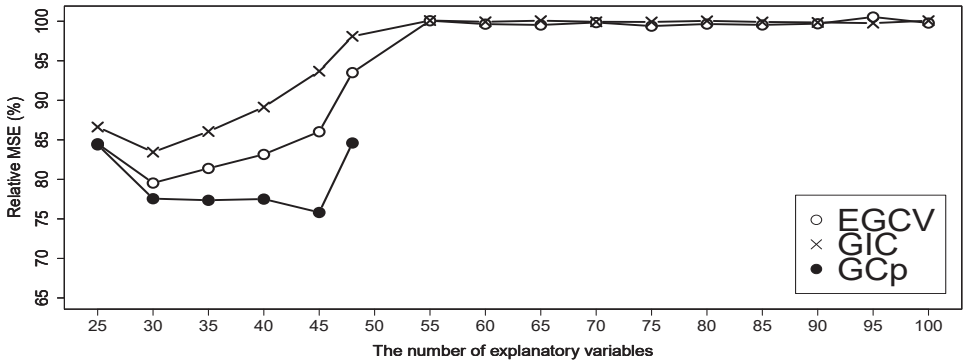
Let $\hat{\boldsymbol{\theta}}$ be an optimized ridge parameters by minimizing MSC, $\hat{\boldsymbol{y}}$ be the predictive value of $\boldsymbol{y}$ based on the LSE of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{\theta}}}$ be the predictive value of $\boldsymbol{y}$ in the GRR based on $\boldsymbol{\theta}$, i.e.,

$$\hat{\boldsymbol{y}} = (\boldsymbol{J}_n + \boldsymbol{X}\boldsymbol{M}^{+}\boldsymbol{X}')\boldsymbol{y}, \quad \hat{\boldsymbol{y}}_{\hat{\boldsymbol{\theta}}} = (\boldsymbol{J}_n + \boldsymbol{X}\boldsymbol{M}_{\hat{\boldsymbol{\theta}}}^{+}\boldsymbol{X}')\boldsymbol{y}.$$

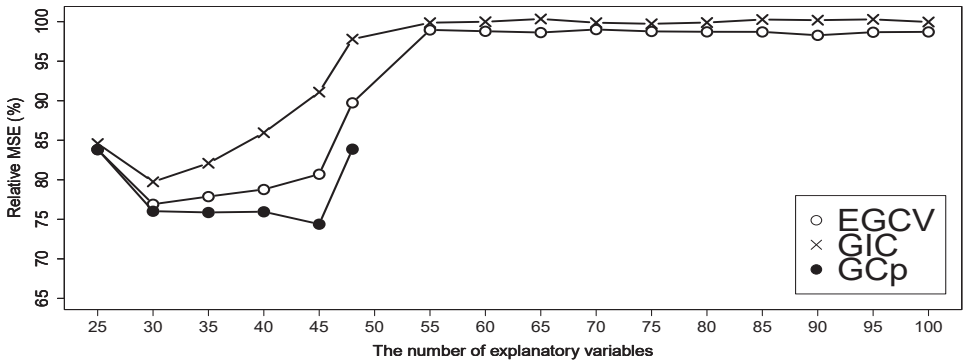
We used the following relative mean square error (MSE) as a measurement of performance of each method.

$$\frac{\mathrm{E}\left[(\hat{\boldsymbol{y}}_{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\beta})'(\hat{\boldsymbol{y}}_{\boldsymbol{\theta}} - \boldsymbol{X}\boldsymbol{\beta})\right]}{\mathrm{E}\left[(\hat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\hat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})\right]} \times 100(\%)$$

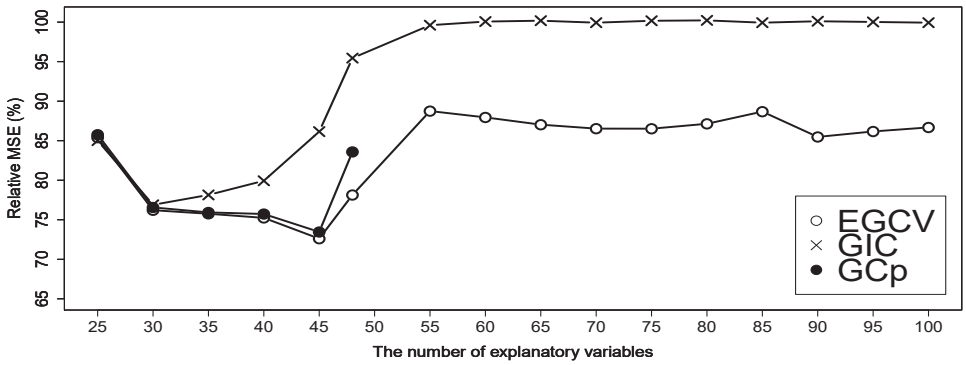
Notice that $\mathrm{E}\left[(\hat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\hat{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})\right] = m+1$ in this case. The above expectation was evaluated by Monte Carlo simulation with 10,000 iterations.



(a) Case of $\alpha = 2$

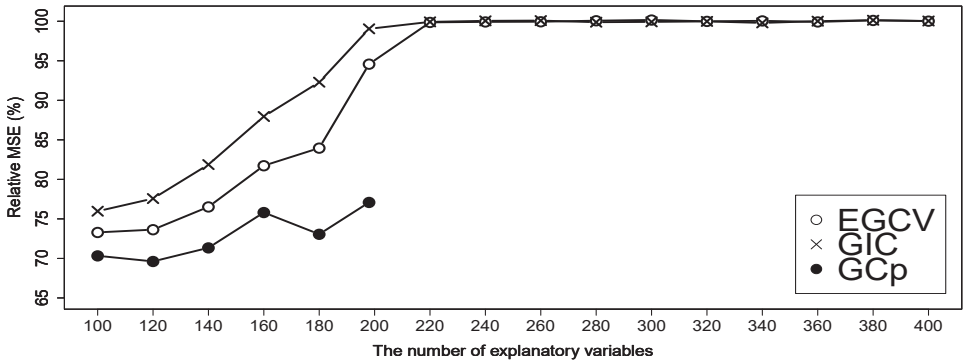


(b) Case of $\alpha = 2 \log \log n$

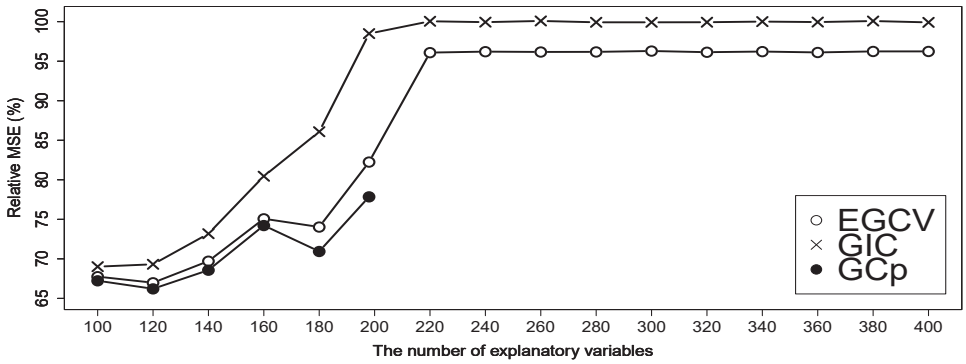


(c) Case of $\alpha = \log n$

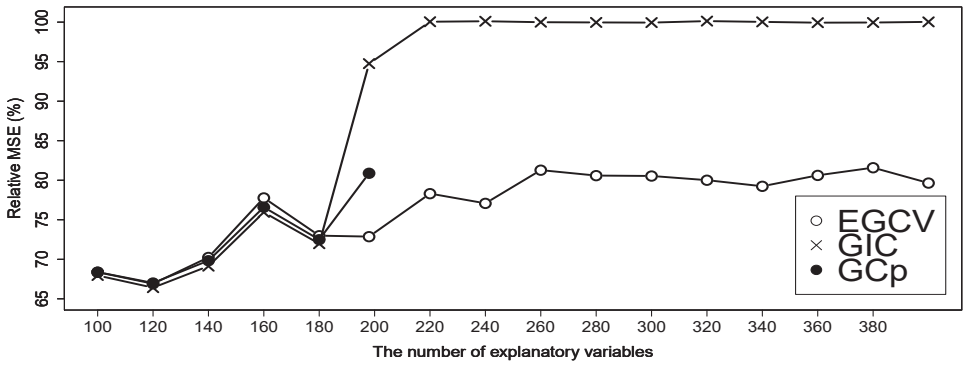
Figure 1. Case of $n = 50$



(a) Case of $\alpha = 2$

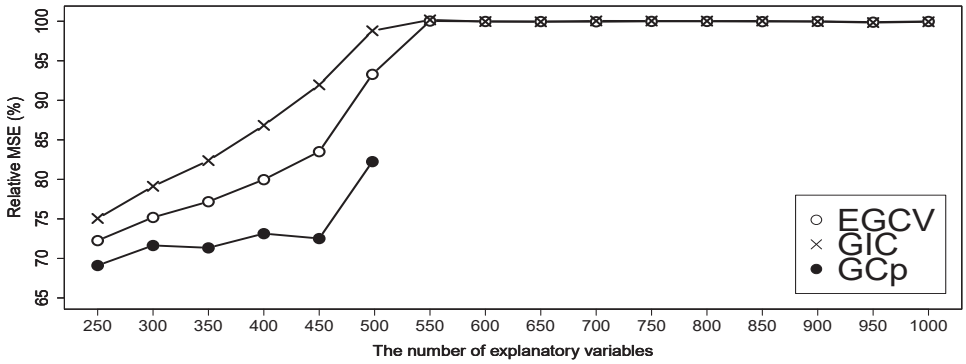


(b) Case of $\alpha = 2 \log \log n$

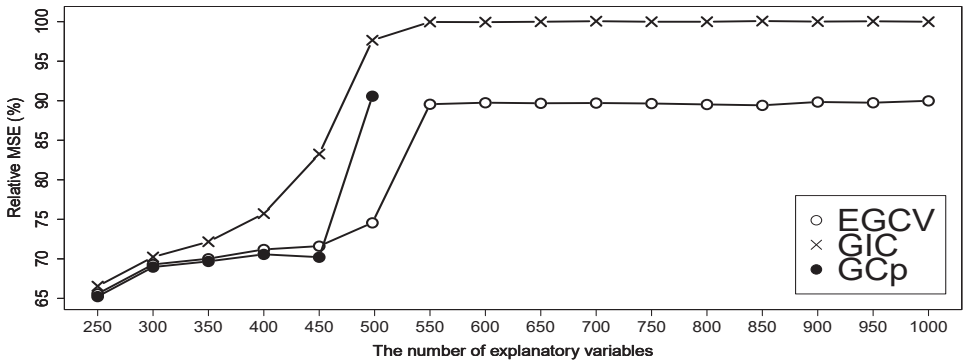


(c) Case of $\alpha = \log n$

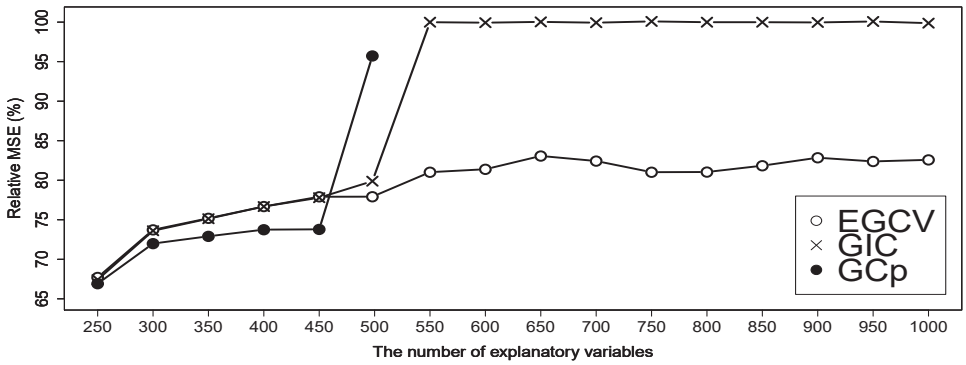
Figure 2. Case of $n = 200$



(a) Case of $\alpha = 2$



(b) Case of $\alpha = 2 \log \log n$



(c) Case of $\alpha = \log n$

Figure 3. Case of $n = 500$

Figures 1, 2 and 3 shows the results when $n = 50$, 200 and 500, respectively, with $\rho = 0.99$ and $k = k_1, \ldots, k_{16}$, where

$$k_j = \begin{cases} n(4+j)/10 & (j = 1, \ldots, 5) \\ n-2 & (j = 6) \\ n\{1 + (j-6)/10\} & (j = 7, \ldots, 16) \end{cases}.$$

In the figures, there were no results of GC_p-minimization methods when $k = k_7, \ldots, k_{16}$ because the GC_p was not definable when $k \geq k_7$. From figures, we can see that the performances of the EGCV-minimization method with $\alpha = \log n$ were well in most cases. On the other hand, the performances of the GIC-minimization method were very bad when $k \geq k_7$ because the GIC-minimization method dose not shrink the GRRE of $\boldsymbol{\beta}$ towards $\mathbf{0}_k$ at all when $k \geq k_7$. Also, the performances of the GCV-minimization method were also bad when $k \geq k_7$ because the GCV-minimization method hardly shrink the GRRE of $\boldsymbol{\beta}$ towards $\mathbf{0}_k$ when $k \geq k_7$. Moreover, as for the number of candidate minimizers, the following results were derived:

$$\mathrm{mode}\,\{\#(\mathcal{S}_{\mathrm{GIC}})\} = 2, \quad \min\{\#(\mathcal{S}_{\mathrm{GIC}})\} = 2, \quad \max\{\#(\mathcal{S}_{\mathrm{GIC}})\} = 7,$$

and

$$\mathrm{mode}\,\{\#(\mathcal{S}_{\mathrm{EGCV}})\} = \begin{cases} 3 & (k \geq n-1 \ \& \ \alpha = 2\log\log n) \\ 2 & (\text{otherwise}) \end{cases},$$

$$\min\{\#(\mathcal{S}_{\mathrm{GIC}})\} = 2, \quad \max\{\#(\mathcal{S}_{\mathrm{EGCV}})\} = \begin{cases} 3 & (k < n-1) \\ 12 & (k \geq n-1) \end{cases}.$$

Hence our algorithm will be efficient because the number of candidate minimizers was small.

Acknowledgment The authors thank Dr. Shintaro Hashimoto, Mr. Tomoyuki Nakagawa, Mr. Ryoya Oda and Mr. Shota Ochiai, of Hiroshima University, for helpful comments.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), pp. 267–281. Akadémiai Kiadó, Budapest.

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

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.

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

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

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.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics*, **22**, 79–86.

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

Liu, Z., Shen, Y. & Ott, J. (2011). Multilocus association mapping using generalized ridge logistic regression. *BMC Bioinformatics*, **12**, 384–391.

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

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.

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

Shen, X., Alam, M., Fikse, F. & Rönnegård, L. (2013). A novel generalized ridge regression method for quantitative genetics. *Genetics*, **193**, 1255–1268.

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

Yanagihara, H. (2013). Explicit solution to the minimization problem of generalized cross-validation criterion for selecting ridge parameters in generalized ridge regression. *TR-No*. 13-07, *Hiroshima Statistical Research Group*, Hiroshima University.

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.1. Proof of Lemma 1

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

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

where $\theta_1, \ldots, \theta_m$ are ridge parameters, and d_j is given by (2.1). Since $\boldsymbol{\theta} \to \boldsymbol{\delta}$ is one to one, we minimize MSC via $\boldsymbol{\delta}$ instead of $\boldsymbol{\theta}$. By using $\boldsymbol{\delta}$, we rewrite $\hat{\sigma}^2(\boldsymbol{\theta})$ and $\mathrm{df}(\boldsymbol{\theta})$ in (2.7) as a function with respect to $\boldsymbol{\delta}$ as

$$\hat{\sigma}^2(\boldsymbol{\theta}) = r(\boldsymbol{\delta}) = \frac{1}{n}\left(n\hat{\sigma}_0^2 + \sum_{j=1}^{m} \delta_j^2 z_j^2\right), \quad \mathrm{df}(\boldsymbol{\theta}) = u(\boldsymbol{\delta}) = 1 + m - \sum_{j=1}^{m} \delta_j, \tag{A.1}$$

where z_j^2 and $\hat{\sigma}_0^2$ are given by (2.6) and (2.8), respectively. By using $r(\boldsymbol{\delta})$ and $u(\boldsymbol{\delta})$, we rewrite $\mathrm{MSC}(\boldsymbol{\theta})$ in (2.10) as a function with respect to $\boldsymbol{\delta}$ as

$$\mathrm{MSC}(\boldsymbol{\theta}) = g(\boldsymbol{\delta}) = f(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})). \tag{A.2}$$

Since $\mathbb{D} \subseteq (0, \hat{\sigma}^2_\infty] \times [1, n)$, we know that $\boldsymbol{\delta} \neq \mathbf{0}_m$. From the assumption of f, we have $\dot{f}_r(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) > 0$ and $\dot{f}_u(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) > 0$, where $\dot{f}_r(r, u)$ and $\dot{f}_u(r, u)$ are the first partial derivatives of f with respect to r and u, respectively. Let

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

It is clear that $\tau(\boldsymbol{\delta}) > 0$ when $\mathbb{D} \subseteq (0, \hat{\sigma}^2_\infty] \times [1, n)$. Notice that

$$\begin{aligned}\frac{\partial}{\partial \delta_j} g(\boldsymbol{\delta}) &= \frac{\partial r(\boldsymbol{\delta})}{\partial \delta_j} \cdot \frac{\partial f(r, u)}{\partial r}\bigg|_{(r,u)=(r(\boldsymbol{\delta}),u(\boldsymbol{\delta}))} + \frac{\partial u(\boldsymbol{\delta})}{\partial \delta_j} \cdot \frac{\partial f(r, u)}{\partial u}\bigg|_{(r,u)=(r(\boldsymbol{\delta}),u(\boldsymbol{\delta}))} \\ &= \frac{1}{n} 2z_j^2 \delta_j \dot{f}_r(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) - \dot{f}_u(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) = \frac{2}{n} z_j^2 \dot{f}_r(r(\boldsymbol{\delta}), u(\boldsymbol{\delta})) \left\{ \delta_j - \frac{\tau(\boldsymbol{\delta})}{z_j^2} \right\}. \quad \text{(A.3)}\end{aligned}$$

Let $\boldsymbol{\delta}^\star = (\delta_1^\star, \ldots, \delta_m^\star)'$ be the minimizer of $g(\boldsymbol{\delta})$, i.e.,

$$\boldsymbol{\delta}^\star = \arg \min_{\boldsymbol{\delta} \in [0,1]^m \setminus \{\mathbf{0}_m\}} g(\boldsymbol{\delta}),$$

where $[0, 1]^m$ is the mth Cartesian power of the set $[0, 1]$. From (A.3), necessary condition of $\boldsymbol{\delta}^\star$ is given by

$$\delta_j^\star = \begin{cases} \dfrac{\tau(\boldsymbol{\delta}^\star)}{z_j^2} & (\tau(\boldsymbol{\delta}^\star) \leq z_j^2) \\ 1 & (\tau(\boldsymbol{\delta}^\star) > z_j^2) \end{cases} \quad (j = 1, \ldots, m).$$

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

$$\mathcal{G} = \left\{ \boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)' \in [0, 1]^m \mid \boldsymbol{\delta} = \hat{\boldsymbol{\delta}}(h), {}^\forall h \in \mathbb{R}_+ \setminus \{0\} \right\},$$

where $\hat{\boldsymbol{\delta}}(h)$ is the m-dimensional vector of which jth element is defined by

$$\hat{\delta}_j(h) = \begin{cases} \dfrac{h}{z_j^2} & (h \leq z_j^2) \\ 1 & (h > z_j^2) \end{cases} \quad (j = 1, \ldots, m). \quad \text{(A.4)}$$

It follows from the fact $\mathcal{G} \subseteq [0, 1]^m \setminus \{\mathbf{0}_m\}$ that

$$g(\boldsymbol{\delta}^\star) = \min_{\boldsymbol{\delta} \in [0,1]^m \setminus \{\mathbf{0}_m\}} g(\boldsymbol{\delta}) \leq \min_{\boldsymbol{\delta} \in \mathcal{G}} g(\boldsymbol{\delta}) = \min_{h \in \mathbb{R}_+ \setminus \{0\}} g(\hat{\boldsymbol{\delta}}(h)).$$

Moreover, it is clear that $\boldsymbol{\delta}^\star \in \mathcal{G}$. This implies that

$$g(\boldsymbol{\delta}^\star) \geq \min_{h \in \mathbb{R}_+ \setminus \{0\}} g(\hat{\boldsymbol{\delta}}(h)).$$

Therefore, we have $\boldsymbol{\delta}^\star = \hat{\boldsymbol{\delta}}(\hat{h})$, where

$$\hat{h} = \arg\min_{h\in\mathbb{R}_+\setminus\{0\}} g(\hat{\boldsymbol{\delta}}(h)).$$

Recall that $g(\boldsymbol{\delta}) = \mathrm{MSC}(\boldsymbol{\theta})$ and

$$\theta_j = \begin{cases} \dfrac{d_j\delta_j}{1-\delta_j} & (\text{if } \delta_j \neq 1) \\ \infty & (\text{if } \delta_j = 1) \end{cases}.$$

Moreover, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}}$ is rewitten as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{\theta}} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{Q}\boldsymbol{\Theta}\boldsymbol{Q}')^+\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{Q}(\boldsymbol{D}+\boldsymbol{\Theta})^+\boldsymbol{Q}'\boldsymbol{X}\boldsymbol{y} \\ &= \boldsymbol{Q}(\boldsymbol{D}+\boldsymbol{\Theta})^+\boldsymbol{D}\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{D}^+\boldsymbol{Q}'\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{V}\boldsymbol{Q}'\hat{\boldsymbol{\beta}},\end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{D}+\boldsymbol{\Theta})^+\boldsymbol{D}$ and $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$ given by (2.3). When $\boldsymbol{\theta} \neq \boldsymbol{0}_m$, the jth diagonal element of $\boldsymbol{V}$ is given by

$$v_j = \begin{cases} \dfrac{d_j}{d_j+\theta_j} & (\text{when } \theta_j < \infty) \\ 0 & (\text{when } \theta_j = \infty) \end{cases}.$$

It follows from the simple calculation that $d_j/\{d_j+\hat{\theta}_j(h)\} = 1-h/z_j^2$ when $h \le z_j^2$. Consequently, Lemma 1 is proved.

A.2. Proof of Lemma 2

From (A.2), we can see that

$$\phi(h) = \mathrm{MSC}(\hat{\boldsymbol{\theta}}(h)) = f(r(\hat{\boldsymbol{\delta}}(h)), u(\hat{\boldsymbol{\delta}}(h))),$$

where $\hat{\boldsymbol{\theta}}(h)$ is given by (2.12), $\hat{\boldsymbol{\delta}}(h)$ are given by (A.4), and $r(\boldsymbol{\delta})$ and $u(\boldsymbol{\delta})$ are given by (A.1), respectively. Since $t_1,\ldots,t_m$ are order statistics of $z_1^2,\ldots,z_m^2$, we have

$$\begin{aligned} r(\hat{\boldsymbol{\delta}}(h)) &= \hat{\sigma}_0^2 + \frac{1}{n}\sum_{j=1}^m \left\{ \mathrm{I}(z_j^2 \ge h)\left(\frac{h}{z_j^2}-1\right)+1 \right\}^2 z_j^2 = \hat{\sigma}_0^2 + \frac{1}{n}\sum_{j=1}^m \left\{ \mathrm{I}(t_j \ge h)\left(\frac{h}{t_j}-1\right)+1 \right\}^2 t_j, \\ u(\hat{\boldsymbol{\delta}}(h)) &= 1+m-\sum_{j=1}^m \left\{ \mathrm{I}(z_j^2 \ge h)\left(\frac{h}{z_j^2}-1\right)+1 \right\} = 1+m-\sum_{j=1}^m \left\{ \mathrm{I}(t_j \ge h)\left(\frac{h}{t_j}-1\right)+1 \right\}, \end{aligned}$$

where $\mathrm{I}(x \ge h)$ is an indicator function, i.e., $\mathrm{I}(x \ge h) = 1$ if $x \ge h$ and $\mathrm{I}(x \ge h) = 0$ if $x < h$. It is clear that $r(\hat{\boldsymbol{\delta}}(h))$ and $u(\hat{\boldsymbol{\delta}}(h))$ are continuous functions on $h \in \mathbb{R}_+\setminus\{0\}$. Therefore, the property (P1) is proved.

Notice that when $h \in R_a$, we have

$$r(\hat{\boldsymbol{\delta}}(h)) = r_a(\hat{\boldsymbol{\delta}}(h)) = \hat{\sigma}_0^2 + \frac{1}{n}\left(\sum_{j=1}^{a} t_j + \sum_{j=a+1}^{m} \frac{h^2}{t_j}\right) = \hat{\sigma}_0^2 + \frac{1}{n}(c_{1,a} + h^2 c_{2,a}),$$

$$u(\hat{\boldsymbol{\delta}}(h)) = u_a(\hat{\boldsymbol{\delta}}(h)) = 1 + m - \left(\sum_{j=1}^{a} 1 + \sum_{j=a+1}^{m} \frac{1}{t_j}\right) = 1 + m - a - hc_{2,a},$$

where $c_{1,a}$ and $c_{2,a}$ are given by (2.16). The property (P3) is proved by this equation. Notice that $c_{2,m} = 0$ and

$$n\hat{\sigma}_0^2 + c_{1,m} = n\hat{\sigma}_0^2 + \boldsymbol{z}'\boldsymbol{z} = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{J}_n - \boldsymbol{X}\boldsymbol{M}^+\boldsymbol{X}')\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{X}\boldsymbol{M}^+\boldsymbol{X}'\boldsymbol{y} = n\hat{\sigma}_\infty^2.$$

These imply that $r_m(\hat{\boldsymbol{\delta}}(h)) = \hat{\sigma}_\infty^2$ and $u_m(\hat{\boldsymbol{\delta}}(h)) = 1$ whem $h \in R_m$. Hence, the property (P2) is proved.

A.3. Proof of Theorem 2

At first, we derive the minimizer of GIC($\boldsymbol{\theta}$) in (3.1) when $\hat{\sigma}_0^2$ in (2.8) is 0. It is easy to see that $\hat{\sigma}^2(\boldsymbol{0}_m) = 0$ when $\hat{\sigma}_0^2 = 0$, and $f_{\mathrm{GIC}}(0, u) = 0$ holds for any $u \in [1, n]$, where $\hat{\sigma}(\boldsymbol{\theta})$ is given by (2.4) and $f_{\mathrm{GIC}}(r, u)$ is given by (2.11). Moreover, $f_{\mathrm{GIC}}(r, u) > 0$ holds for any $(r, u) \in (0, \hat{\sigma}_\infty^2] \times [1, n]$, where $\hat{\sigma}_\infty^2$ is given by (2.9). Notice that $\hat{\sigma}^2(\boldsymbol{\theta}) = 0$ holds if and only if $\boldsymbol{\theta} = \boldsymbol{0}_m$ and $\hat{\sigma}_0^2 = 0$, and $\hat{\boldsymbol{\theta}}(0) = \boldsymbol{0}_m$ holds, where $\hat{\boldsymbol{\theta}}(h)$ is given by (2.12). These results imply that the minimizer of GIC($\boldsymbol{\theta}$) is $\hat{\boldsymbol{\theta}}(0)$ when $\hat{\sigma}_0^2 = 0$.

Next, we derive the minimizer of GIC($\boldsymbol{\theta}$) when $\hat{\sigma}_0^2 \neq 0$. Then the domain of f_{GIC} is included in $(0, \hat{\sigma}_\infty^2] \times [1, n)$. Moreover,

$$\lim_{h \to 0} \psi_{\mathrm{GIC},0}(h) = -\alpha\hat{\sigma}_0^2 < 0,$$

where $\psi_{\mathrm{GIC},a}(h)$ is given by (3.4). This indicates that some positive number ξ exists such that $\phi_{\mathrm{GIC}}(\xi) < \lim_{h \to 0} \phi_{\mathrm{GIC}}(h)$, where $\phi_{\mathrm{GIC}}(h)$ is given by (3.2). Therefore, by using Lemma 1, the minimizer of GIC($\boldsymbol{\theta}$) is given by $\hat{\boldsymbol{\theta}}(\hat{h}_{\mathrm{GIC}})$, where $\hat{h}_{\mathrm{GIC}}$ is the minimizer of $\phi_{\mathrm{GIC}}(h)$ as

$$\hat{h}_{\mathrm{GIC}} = \arg\min_{h \in \mathbb{R}_+ \setminus \{0\}} \phi_{\mathrm{GIC}}(h).$$

Hence, in order to prove Theorem 2 when $\hat{\sigma}_0^2 \neq 0$, it is enough just to show $\hat{h}_{\mathrm{GIC}} \in \mathcal{S}_{\mathrm{GIC}}$, where $\mathcal{S}_{\mathrm{GIC}}$ is a set of candidate minimizers of $\phi_{\mathrm{GIC}}(h)$ given by (3.7). From (P1) and (P2) in Lemma 2, we can see that $\phi_{\mathrm{GIC}}(h)$ is continuous at any $h \in \mathbb{R}_+ \setminus \{0\}$ and $\phi_{\mathrm{GIC}}(h) = \hat{\sigma}_\infty^2 \exp(\alpha/n)$ when $h \in R_m$. The $\psi_{\mathrm{GIC}}(h)$ in (3.3) is also continuous at any $h \in (0, t_m]$ because of

$$\begin{aligned}
\psi_{\mathrm{GIC},a}(t_{a+1}) &= -\alpha c_{2,a} t_{a+1}^2 + 2nt_{a+1} - \alpha(n\hat{\sigma}_0^2 + c_{1,a}) \\
&= -\alpha\left(c_{2,a+1} + \frac{1}{t_{a+1}}\right) t_{a+1}^2 + 2nt_{a+1} - \alpha\left(n\hat{\sigma}_0^2 + c_{1,a+1} - t_{a+1}\right) \\
&= -\alpha c_{2,a+1} t_{a+1}^2 + 2nt_{a+1} - \alpha(n\hat{\sigma}_0^2 + c_{1,a+1}) = \psi_{\mathrm{GIC},a+1}(t_{a+1}) \quad (a = 0, \ldots, m-2),
\end{aligned}$$

where $c_{1,a}$ and $c_{2,a}$ are given by (2.16). It should be emphasized that the sign of $\partial\phi_{\mathrm{GIC},a}(h)/\partial h$ is equal to that of $\psi_{\mathrm{GIC},a}(h)$. Hence, we have the necessary condition of $\hat{h}_{\mathrm{GIC}}$ as

$$\begin{aligned}&\Big\{\Big\{\psi_{\mathrm{GIC}}(\hat{h}_{\mathrm{GIC}})=0\Big\}\bigwedge\Big\{\exists\epsilon_0>0\ s.t.\ \forall\epsilon\in(0,\epsilon_0),\ \psi_{\mathrm{GIC}}(\hat{h}_{\mathrm{GIC}}-\epsilon)<0\Big\}\Big\}\\ &\quad\bigvee\Big\{\hat{h}_{\mathrm{GIC}}=t_m\ s.t.\ \psi_{\mathrm{GIC},m-1}(t_m)<0\Big\}.\end{aligned}\tag{A.5}$$

Notice that

$$\psi_{\mathrm{GIC},m-1}(t_m)=-\frac{\alpha}{t_m}t_m^2+2nt_m-\alpha(n\hat{\sigma}_\infty^2-t_m)=2nt_m-\alpha n\hat{\sigma}_\infty^2.$$

Hence, we derive

$$\psi_{\mathrm{GIC},m-1}(t_m)<0\Longleftrightarrow\hat{\sigma}_\infty^2<\frac{2}{\alpha}t_m.\tag{A.6}$$

Recall that $\psi_{\mathrm{GIC},a}(h)$ is the concave quadratic function. Hence, we have

$$\begin{aligned}&\Big\{\psi_{\mathrm{GIC}}(\hat{h}_{\mathrm{GIC}})=0\Big\}\bigwedge\Big\{\exists\epsilon_0>0\ s.t.\ \forall\epsilon\in(0,\epsilon_0),\ \psi_{\mathrm{GIC}}(\hat{h}_{\mathrm{GIC}}-\epsilon)<0\Big\}\\ \Longleftrightarrow&\Big\{\hat{h}_{\mathrm{GIC}}\text{ is the smaller of the two real distinct roots or the double}\\ &\quad\text{root of }\psi_{\mathrm{GIC},a}(h)=0\text{ which is included in }R_a\ (a=0,1,\ldots,m-1)\Big\}.\end{aligned}\tag{A.7}$$

Notice that $\xi_{\mathrm{GIC},a}$ in (3.5) is the smaller of the two real distinct roots or the double root of $\psi_{\mathrm{GIC},a}(h)=0$ when $\alpha^2c_{2,a}(n\hat{\sigma}_0^2+c_{1,a})\le n^2$. Therefore, from (A.5), (A.6) and (A.7), $\hat{h}_{\mathrm{GIC}}\in\mathcal{S}_{\mathrm{GIC}}$ can be shown. Consequently, Theorem 2 is proved.

A.4. Proof of Corollary 1

It follows from (3.4) that

$$\psi_{\mathrm{GIC},a}(h)=-\alpha c_{2,a}\left(h-\frac{n}{\alpha c_{2,a}}\right)^2+\frac{n^2}{\alpha c_{2,a}}-\alpha(n\hat{\sigma}_0^2+c_{1,a}),$$

where $\hat{\sigma}_0^2$ is given by (2.8), and $c_{1,a}$ and $c_{2,a}$ are given by (2.16). This indicates that the h-coordinate of the vertex of $\psi_{\mathrm{GIC},a}(h)$ is $n/(\alpha c_{2,a})$. It follows from the equation $t_1\le\cdots\le t_m$ that for any $a\in\{0,1,\ldots,m-1\}$,

$$c_{a,2}=\sum_{j=a+1}^{m}\frac{1}{t_j}\le\sum_{j=a+1}^{m}\frac{1}{t_{a+1}}=\frac{m-a}{t_{a+1}}\le\frac{m}{t_{a+1}}.\tag{A.8}$$

Hence, we have

$$\frac{n}{\alpha c_{2,a}}\ge\frac{n}{\alpha m}t_{a+1}\quad(a=0,1,\ldots,m-1).$$

From this result, we can see that $n/(\alpha c_{2,a})\ge t_{a+1}$ if $\alpha\le n/m$. This indicates that $\psi_{\mathrm{GIC},a}(h)$ is a strictly increasing function at $h\in R_a$ in (2.15) if $\alpha\le n/m$ because the h-coordinate of the vertex of $\psi_{\mathrm{GIC},a}(h)$ is larger or equal to t_{a+1}, which is the right end point of R_a. Hence, it should

be emphasized that $\psi_{\mathrm{GIC}}(h)$ is a strictly increasing function at $h \in (0, t_m]$ when $\alpha \leq n/m$. Recall that $\psi_{\mathrm{GIC}}(h)$ is continuous at any $h \in (0, t_m]$ and $\lim_{h\to 0} \psi_{\mathrm{GIC},0}(h) < 0$ when $\hat{\sigma}_0^2 \neq 0$, where $\hat{\sigma}_0^2$ is given by (2.8). Therefore, when $\hat{\sigma}_0^2 \neq 0$ and $\alpha \leq n/m$, we have

$$\begin{aligned}
\psi_{\mathrm{GIC},m-1}(t_m) \geq 0 &\Longrightarrow \text{the unique solution of } \psi_{\mathrm{GIC}}(h) = 0 \text{ exisits in } h \in (0, t_m] \\
&\Longleftrightarrow \{\#(\mathcal{A}_{\mathrm{GIC}}) = 1\} \wedge \{\mathcal{T}_{\mathrm{GIC}} = \emptyset\}, \\
\psi_{\mathrm{GIC},m-1}(t_m) < 0 &\Longrightarrow \text{there is no solution of } \psi_{\mathrm{GIC}}(h) = 0 \text{ in } h \in (0, t_m] \\
&\Longleftrightarrow \{\mathcal{A}_{\mathrm{GIC}} = \emptyset\} \wedge \{\mathcal{T}_{\mathrm{GIC}} = \{t_m\}\},
\end{aligned}$$

where $\mathcal{A}_{\mathrm{GIC}}$ and $\mathcal{T}_{\mathrm{GIC}}$ are given by (3.6). From the above results and (A.6), Corollary 1 can be proved.

A.5. Proof of Theorem 3

At first, we derive the minimizer of $\mathrm{EGCV}(\boldsymbol{\theta})$ in (4.2) when $\hat{\sigma}_0^2 = 0$ and $b \neq 0$, where $\hat{\sigma}_0^2$ and b are given by (2.8) and (4.5), respectively. It is easy to see that $\hat{\sigma}^2(\mathbf{0}_m) = 0$ when $\hat{\sigma}_0^2 = 0$, $\mathrm{df}(\boldsymbol{\theta}) < n$ when $b \neq 0$, and $f_{\mathrm{EGCV}}(0, u) = 0$ holds for any $u \in [1, n)$, where $\hat{\sigma}^2(\boldsymbol{\theta})$ is given by (2.4) and $f_{\mathrm{EGCV}}(r, u)$ is given by (4.1). Moreover, $f_{\mathrm{EGCV}}(r, u) > 0$ holds for any $(r, u) \in (0, \hat{\sigma}_\infty^2] \times [1, n)$, where $\hat{\sigma}_\infty^2$ is given by (2.9). Notice that $\hat{\sigma}^2(\boldsymbol{\theta}) = 0$ holds if and only if $\boldsymbol{\theta} = \mathbf{0}_m$ and $\hat{\sigma}_0^2 = 0$, and $\hat{\boldsymbol{\theta}}(0) = \mathbf{0}_m$ holds, where $\hat{\boldsymbol{\theta}}(h)$ is given by (2.12). These results imply that the minimizer of $\mathrm{EGCV}(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\theta}}(0)$ when $\hat{\sigma}_0^2 = 0$ and $b \neq 0$.

Next, we consider when $\hat{\sigma}_0^2 \neq 0$ or $b = 0$. Since $r < n$ in $f_{\mathrm{EGCV}}(r, u)$, it is necessary to satisfy the inequality $\mathrm{df}(\boldsymbol{\theta}) < n$, where $\mathrm{df}(\boldsymbol{\theta})$ is given by (2.5). From (2.7), we can see that $\mathrm{df}(\boldsymbol{\theta}) < n \Leftrightarrow \boldsymbol{\theta} \neq \mathbf{0}_m \Rightarrow \hat{\sigma}^2(\boldsymbol{\theta}) \neq 0$, where $\hat{\sigma}^2(\boldsymbol{\theta})$ is given by (2.4). This indicates that the domain of f_{EGCV} is included in $(0, \hat{\sigma}_\infty^2] \times [1, n)$. Notice that

$$\phi_{\mathrm{EGCV},0}(h) = \frac{\hat{\sigma}_0^2 + h^2 c_{2,0}/n}{(b + h c_{2,0}/n)^\alpha}, \quad \psi_{\mathrm{EGCV},0}(h) = -(\alpha - 2) c_{2,0} h^2 + 2nbh - \alpha n \hat{\sigma}_0^2,$$

where $\phi_{\mathrm{EGCV},a}(h)$, $\psi_{\mathrm{EGCV},a}(h)$ and $c_{2,a}$ are given by (4.4), (4.7) and (2.16), respectively. Recall that $b = 0 \Rightarrow \hat{\sigma}_0^2 = 0$. It follows from the above equations and the assumption $\alpha > 2$ that

$$\lim_{h\to 0} \phi_{\mathrm{EGCV},0}(h) = \begin{cases} \hat{\sigma}_0^2 / b^\alpha & (\hat{\sigma}_0^2 \neq 0) \\ \infty & (b = 0) \end{cases}, \quad \lim_{h\to 0} \psi_{\mathrm{EGCV},0}(h) = \begin{cases} -\alpha n \hat{\sigma}_0^2 < 0 & (\hat{\sigma}_0^2 \neq 0) \\ 0 & (b = 0) \end{cases}.$$

These results imply that some positive number ξ exists such that $\phi_{\mathrm{EGCV}}(\xi) < \lim_{h\to 0} \phi_{\mathrm{EGCV}}(h)$ when $\hat{\sigma}_0^2 \neq 0$ or $b = 0$, where $\phi_{\mathrm{EGCV}}(h)$ is given by (4.3). Therefore, by using Lemma 1, the minimizer of $\mathrm{EGCV}(\boldsymbol{\theta})$ is given by $\hat{\boldsymbol{\theta}}(\hat{h}_{\mathrm{EGCV}})$, where $\hat{h}_{\mathrm{EGCV}}$ is the minimizer of $\phi_{\mathrm{EGCV}}(h)$ as

$$\hat{h}_{\mathrm{EGCV}} = \arg \min_{h \in \mathbb{R}_+ \setminus \{0\}} \phi_{\mathrm{EGCV}}(h).$$

From (P1) and (P2) in Lemma 2, we can see that $\phi_{\mathrm{EGCV}}(h)$ is continuous at any $h \in \mathbb{R}_+ \backslash \{0\}$ and $\phi_{\mathrm{EGCV}}(h) = \hat{\sigma}^2_\infty/(1-n^{-1})^\alpha$ when $h \in R_m$, where $\hat{\sigma}^2_\infty$ is given by (2.9). The $\psi_{\mathrm{EGCV}}(h)$ which is given by (4.6) is also continuous at any $h \in (0, t_m]$ because of

$$\begin{aligned}
\psi_{\mathrm{EGCV},a}(t_{a+1}) &= -(\alpha-2)c_{2,a}t^2_{a+1} + 2(a+nb)t_{a+1} - \alpha(n\hat{\sigma}^2_0 + c_{1,a}) \\
&= -(\alpha-2)\left(c_{2,a+1} + \frac{1}{t_{a+1}}\right)t^2_{a+1} + 2(a+nb)t_{a+1} - \alpha\left(n\hat{\sigma}^2_0 + c_{1,a+1} - t_{a+1}\right) \\
&= -(\alpha-2)c_{2,a+1}t^2_{a+1} + 2(a+1+nb)t_{a+1} - \alpha(n\hat{\sigma}^2_0 + c_{1,a+1}) \\
&= \psi_{\mathrm{EGCV},a+1}(t_{a+1}) \quad (a = 0, \ldots, m-2),
\end{aligned}$$

where $c_{1,a}$ is given by (2.16). Notice that

$$\psi_{\mathrm{EGCV},m-1}(t_m) = -\frac{\alpha-2}{t_m}t^2_m + 2(n-2)t_m - \alpha(n\hat{\sigma}^2_\infty - t_m) = 2(n-1)t_m - \alpha n\hat{\sigma}^2_\infty.$$

Hence, by the same way as in the proof of Theorem 2, we can show that a set of candidate minimizers of $\phi_{\mathrm{GIC}}(h)$ is given by $\mathcal{S}_{\mathrm{EGCV}}$ in (4.8). Consequently, Theorem 3 is proved.

A.6. Proof of Corollary 2

It follows from (4.7) that

$$\psi_{\mathrm{EGCV},a}(h) = -(\alpha-2)c_{2,a}\left\{h - \frac{a+nb}{(\alpha-2)c_{2,a}}\right\}^2 + \frac{(a+nb)^2}{(\alpha-2)c_{a,2}} - \alpha(n\hat{\sigma}^2_0 + c_{1,a}),$$

where $\hat{\sigma}^2_0$ and b are given by (2.8) and (4.5), respectively, and $c_{a,1}$ and $c_{a,2}$ are given by (2.16). This indicates that the h-coordinate of the vertex of $\psi_{\mathrm{EGCV},a}(h)$ is $(a+nb)/\{(\alpha-2)c_{a,2}\}$. From (A.8), we derive

$$\frac{a+nb}{(\alpha-2)c_{2,a}} \geq \frac{a+nb}{(\alpha-2)m}t_{a+1} \geq \frac{n-m-1}{(\alpha-2)m}t_{a+1} \quad (a = 0, 1, \ldots, m-1).$$

From this result, we can see that $(a+nb)/\{(\alpha-2)c_{2,a}\} \geq t_{a+1}$ if $\alpha \leq (n+m-1)/m$. This indicates that $\psi_{\mathrm{EGCV},a}(h)$ is a strictly increasing function at $h \in R_a$ in (2.15) if $\alpha \leq (n+m-1)/m$ because the h-coordinate of the vertex of $\psi_{\mathrm{EGCV},a}(h)$ is larger or equal to t_{a+1}, which is the right end point of R_a. It should be emphasized that $\psi_{\mathrm{GIC}}(h)$ is a strictly increasing function at $h \in (0, t_m]$ when $\alpha \leq (n+m-1)/m$. Hence, by the same way as in the proof of Corollary 1, we can prove Corollary 2.