TR-No. 22-05, Hiroshima Statistical Research Group, 1–19

Stable Estimation of the Slant Parameter in Skew Normal Regression via an MM Algorithm and Ridge Shrinkage

Mineaki Ohishi[1]*, Hirokazu Yanagihara[2], Hirofumi Wakaki[2] and Masahiko Ono[2]

[1]Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University
1-1-89 Higashi-Senda-machi, Naka-ku, Hiroshima 730-0053, Japan
[2]Graduate School of Advanced Science and Engineering, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima 739-8526, Japan

Abstract

This paper deals with a skew normal linear regression model in which the error is distributed according to a skew normal distribution. The skew normal distribution has three parameters: a location parameter, a scale parameter, and a slant parameter. Their maximum likelihood estimators can be obtained with an R package `sn`, an EM algorithm, and so on. However, estimation via likelihood maximization causes the estimate of the slant parameter to be particularly unstable. To improve the stability of the slant parameter estimation, we derive a new algorithm based on the MM principle and propose a stable estimation method for the slant parameter using ridge shrinkage.

(Last Modified: October 26, 2022)

Key words: MM algorithm, Ridge regression, Skew normal distribution,

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

1. Introduction

A skew normal distribution is a generalization of a normal distribution that allows "skewness" (for details of a skew normal distribution, see, e.g., Azzalini & Capitanio, 2014). If a random variable y has the probability density function

$$\frac{2}{\omega}\phi\left(\frac{y-\eta}{\omega}\right)\Phi\left(\frac{\nu(y-\eta)}{\omega}\right), \tag{1.1}$$

it is said that y is distributed according to a skew normal distribution with location parameter η, scale parameter ω (> 0), and slant parameter ν, which can be represented as $y \sim SN(\eta, \omega^2, \nu)$. The slant parameter ν is a measure of the asymmetry of the skew normal distribution. When $\nu = 0$, $SN(\eta, \omega^2, \nu)$ coincides with $N(\eta, \omega^2)$. In the framework of regression, the skew normal distribution is adopted in order to accommodate asymmetry in the error distribution (e.g., Azzalini & Capitanio, 1999; Cancho *et al.*, 2010). For example, Aigner *et al.* (1977) expressed the error distribution of a stochastic frontier model whose error distribution consists of a normal distribution and another certain distribution in terms of a normal distribution and a half-normal distribution. The skew normal distribution can be used for such an error distribution.

We consider the following skew normal linear regression model for a response variable y_i and k explanatory variables $x_{i1}, \ldots, x_{ik}$.

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \quad (i = 1, \ldots, n), \\ \varepsilon_i &\sim SN\left(0, \psi^{-1}, \gamma\psi^{-1/2}\right) \quad (\psi > 0), \end{aligned} \tag{1.2}$$

where $\beta_0, \beta_1, \ldots, \beta_k$, ψ, and γ are unknown parameters. When $\gamma = 0$, (1.2) reduces to a normal linear regression model. Moreover, if the error term of the stochastic frontier model treated by Aigner *et al.* (1977) is defined by

$$\varepsilon_i = \varepsilon_{1,i} + \varepsilon_{2,i}, \quad \varepsilon_{1,1}, \ldots, \varepsilon_{1,n} \sim i.i.d.\ N(0, \sigma_1^2), \quad \varepsilon_{2,1}, \ldots, \varepsilon_{2,n} \sim i.i.d.\ N^+(0, \sigma_2^2),$$

there is the following relationship between the above error term and (1.2).

$$\psi = \frac{1}{\sigma_1^2 + \sigma_2^2}, \quad \gamma = \frac{\sigma_2/\sigma_1}{\sqrt{\sigma_1^2 + \sigma_2^2}},$$

where $N^+(0, \sigma^2)$ represents the half-normal distribution with scale parameter σ. In this paper, we identify γ rather than $\gamma\psi^{-1/2}$ as the slant parameter and focus on the estimation of γ. From (1.1), a negative log-likelihood function for (1.2) is given by

$$\begin{aligned} \ell(\boldsymbol{\xi}) &= -\frac{n}{2}\log\frac{2\psi}{\pi} + \frac{\psi}{2}\sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 - \sum_{i=1}^{n}\log\Phi(\gamma r_i(\boldsymbol{\beta})), \quad \boldsymbol{\xi} = (\boldsymbol{\beta}', \gamma, \psi)', \\ r_i(\boldsymbol{\beta}) &= y_i - \boldsymbol{x}_i'\boldsymbol{\beta}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)', \quad \boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})'. \end{aligned} \tag{1.3}$$

By minimizing $\ell(\boldsymbol{\xi})$, a maximum likelihood estimator (MLE) for (1.2) can be obtained. Although the minimizer of $\ell(\boldsymbol{\xi})$ cannot be obtained in closed form, there are various algorithms for solving the minimization problem. For example, we can use a function `selm` of a package `sn` (e.g., Azzalini, 2022) in R (e.g., R Core Team, 2021) that minimizes $\ell(\boldsymbol{\xi})$ by a function `optim`. As another option, the expectation-minimization (EM) algorithm (Dempster *et al.*,

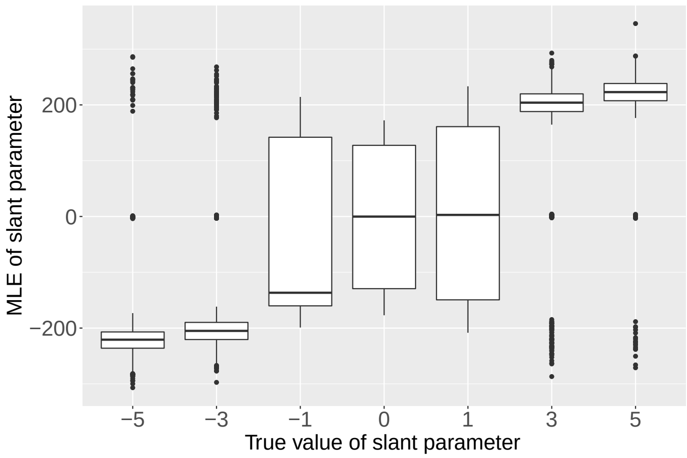

Figure 1. Unstableness of MLE of the slant parameter using `selm`

1977) can be applied (Azzalini & Capitanio, 2014). Although these algorithms give the MLE for (1.2), the MLE of γ is particularly unstable. Figure 1 displays boxplots of the MLE of γ obtained by `selm` when $n = 100$, $k = 30$, and the number of iterations is 1,000. The vertical axis shows the MLE; the horizontal axis shows the true value of γ (for details of the setting, see section 4). The figure confirms that the MLE of γ is far from the true value and that its variance is large.

To reduce the unstableness of the maximum likelihood estimation procedure, this paper proposes the use of ridge regression (Hoerl & Kennard, 1970). Ridge regression was first proposed as a way to avoid problems associated with multicollinearity among the explanatory variables by shrinking the estimator towards zero using penalized estimation with the ℓ_2-norm. Although the main purpose of ridge regression is to address multicollinearity, it can be applied for many purposes, e.g., smoothing in nonparametric regression (Yanagihara, 2012) and the regularization of the covariance matrix in multivariate regression (Yamamura *et al.*, 2010; Kubokawa & Srivastava, 2012). In this paper, we seek to reduce the unstableness of the MLE of slant parameter γ through ridge shrinkage. To incorporate ridge shrinkage into the estimation of γ, we first devise a new algorithm to determine MLE based on the majorization-minimization (MM) algorithm (Hunter & Lange, 2004). Since the MM algorithm minimizes a surrogate function which gives an upper bound of an objective function, we essentially derive the surrogate function. In particular, we evaluate the objective function more tightly. Note that although, in a broad sense, an EM algorithm is a type of MM algorithm, we distinguish between the two. In our MM algorithm for MLE, each parameter ($\boldsymbol{\beta}$, γ, and ψ) can be updated in closed form. Moreover, the update equation of γ is obtained from a minimization of a quadratic function.

Hence, ridge shrinkage can be easily introduced, and an update equation for ridge shrinkage can be obtained in closed form as well.

The remainder of the paper is organized as follows: In section 2, we describe the MM algorithm for calculating the MLE for model (1.2) by deriving surrogate functions for $\boldsymbol{\beta}$ and γ, and discuss the initial values for the algorithm. In section 3, we introduce ridge shrinkage into our MM algorithm to stabilize the estimation of slant parameter γ. In section 4, we evaluate the performance of our proposed method. Technical details are provided in the Appendix.

2. MM Algorithm for Maximum Likelihood Estimation

In this section, we propose a new algorithm to minimize the objective function $\ell(\boldsymbol{\xi})$ in (1.3). To minimize the function, we first apply a block-wise coordinate descent algorithm. Specifically, we search the solution by repeating the following minimizations for each parameter:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \ell(\boldsymbol{\xi}), \quad \min_{\gamma\in\mathbb{R}} \ell(\boldsymbol{\xi}), \quad \min_{\psi\in\mathbb{R}} \ell(\boldsymbol{\xi}).$$

By ignoring the constant terms, the objective functions for these sub-problems are given by

$$\ell_1(\boldsymbol{\beta}) = \frac{\psi}{2}\sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 - \sum_{i=1}^{n} \log \Phi(\gamma r_i(\boldsymbol{\beta})), \tag{2.1}$$

$$\ell_2(\gamma) = -\sum_{i=1}^{n} \log \Phi(\gamma r_i(\boldsymbol{\beta})), \tag{2.2}$$

$$\ell_3(\psi) = -\frac{n}{2}\log\psi + \frac{\psi}{2}\sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2. \tag{2.3}$$

At the minimization of $\ell_1(\boldsymbol{\beta})$, γ and ψ are regarded as constants. Similarly, $\boldsymbol{\beta}$ and ψ are regarded as constants at the minimization of $\ell_2(\gamma)$, and $\boldsymbol{\beta}$ and γ are regarded as constants at the minimization of $\ell_3(\psi)$. Since $\ell_3(\psi)$ is a strictly convex function of ψ, a stationary point uniquely exists and is the minimizer. The derivative of $\ell_3(\psi)$ is given by

$$\dot{\ell}_3(\psi) = \frac{d}{d\psi}\ell_3(\psi) = -\frac{n}{2\psi} + \frac{1}{2}\sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2,$$

and hence, the minimizer of $\ell_3(\psi)$ is given in closed form as

$$\hat{\psi} = \frac{n}{\sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2}. \tag{2.4}$$

Since $\ell_1(\boldsymbol{\beta})$ and $\ell_2(\gamma)$ are strictly convex functions of $\boldsymbol{\beta}$ and γ, respectively, their minimizers are stationary points. However, it is difficult to obtain these stationary points directly. Hence, we adopt an MM algorithm to minimize $\ell_1(\boldsymbol{\beta})$ and $\ell_2(\gamma)$. The MM algorithm minimizes an objective function by repeating a minimization of its surrogate function giving the upper bound of the objective function. The following property guarantees that a solution obtained by the MM algorithm is the minimizer (Hunter & Lange, 2004):

Proposition 1. *Let $f(\boldsymbol{\xi})$ be a convex function and for given $\boldsymbol{\xi}_0$, let $f^+(\boldsymbol{\xi} \mid \boldsymbol{\xi}_0)$ be a surrogate function of f satisfying*

$$f(\boldsymbol{\xi}) \leq f^+(\boldsymbol{\xi} \mid \boldsymbol{\xi}_0), \quad f^+(\boldsymbol{\xi}_0 \mid \boldsymbol{\xi}_0) = f(\boldsymbol{\xi}_0).$$

Then, we have

$$f(\boldsymbol{\xi}^\dagger) \leq f(\boldsymbol{\xi}_0), \quad \boldsymbol{\xi}^\dagger = \arg\min_{\boldsymbol{\xi}} f^+(\boldsymbol{\xi} \mid \boldsymbol{\xi}_0).$$

The following lemma is the key to obtaining surrogate functions of $\ell_1(\boldsymbol{\beta})$ and $\ell_2(\gamma)$ (the proof is given in Appendix A.1).

Lemma 1. *For all $x \in \mathbb{R}$, we have*

$$\frac{x\phi(x)}{\Phi(x)} + \left\{\frac{\phi(x)}{\Phi(x)}\right\}^2 \leq 1.$$

We describe the MM algorithms for minimizing $\ell_1(\boldsymbol{\beta})$ and $\ell_2(\gamma)$ in subsections 2.1 and 2.2. From the results, the new algorithm to minimize the objective function $\ell(\boldsymbol{\xi})$ is summarized in Algorithm 1, where Algorithm 2 with $d_{\max}$ and Algorithm 3 are given in the following subsections.

Algorithm 1 Main algorithm to minimize (1.3)

Require: initial vector for $\boldsymbol{\xi} = (\boldsymbol{\beta}', \gamma, \psi)'$
repeat
Update $\boldsymbol{\beta}$ via Algorithm 2 with $L_1 = d_{\max}(\psi + 3\gamma^2)$
Update γ via Algorithm 3 with $L_2 = 3\sum_{i=1}^n r_i(\boldsymbol{\beta})^2$
Update ψ by (2.4)
until solution converges

2.1. Update $\boldsymbol{\beta}$

We now consider minimizing $\ell_1(\boldsymbol{\beta})$ in (2.1) under given γ and ψ to obtain an update equation for $\boldsymbol{\beta}$. Although $\ell_1(\boldsymbol{\beta})$ is a strictly convex function of $\boldsymbol{\beta}$, directly minimizing it is difficult. Hence, we minimize a surrogate function of $\ell_1(\boldsymbol{\beta})$ based on the MM algorithm.

We first derive the surrogate function. The following theorem gives the upper bound of $\ell_1(\boldsymbol{\beta})$ (the proof is given in Appendix A.2).

Theorem 1. *Let $d_{\max}$ be the maximum eigenvalue of $\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$ and $L_1 = d_{\max}(\psi + \gamma^2)$. Moreover, for $\boldsymbol{b} \in \mathbb{R}^{k+1}$, we define $\ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b})$ as*

$$\ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b}) = \ell_1(\boldsymbol{b}) + \boldsymbol{g}(\boldsymbol{b})'(\boldsymbol{\beta} - \boldsymbol{b}) + \frac{L_1}{2}\|\boldsymbol{\beta} - \boldsymbol{b}\|^2,$$

$$\boldsymbol{g}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}\ell_1(\boldsymbol{\beta}) = -\psi \sum_{i=1}^{n} r_i(\boldsymbol{\beta})\boldsymbol{x}_i + \gamma \sum_{i=1}^{n} \frac{\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))}\boldsymbol{x}_i.$$

Then, we have

$$\ell_1(\boldsymbol{\beta}) \le \ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b}), \quad \ell_1^+(\boldsymbol{b} \mid \boldsymbol{b}) = \ell_1(\boldsymbol{b}).$$

From the above theorem, we obtain the surrogate function $\ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b})$ of $\ell_1(\boldsymbol{\beta})$. Hence, the minimizer of $\ell_1(\boldsymbol{\beta})$ can be obtained by repeating the following update:

$$\boldsymbol{\beta}^{(m+1)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(m)}) \quad (m = 0, 1, \ldots),$$

where m is an iteration number and $\boldsymbol{\beta}^{(0)}$ is an initial vector for $\boldsymbol{\beta}$. The $\ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b})$ can be rewritten as

$$\ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b}) = \ell_1(\boldsymbol{b}) + \frac{L_1}{2}\|\boldsymbol{\beta} - \boldsymbol{z}(\boldsymbol{b})\|^2 + \frac{L_1}{2}\|\boldsymbol{b}\|^2 - \boldsymbol{g}(\boldsymbol{b})'\boldsymbol{b},$$

where $\boldsymbol{z}(\boldsymbol{b}) = \boldsymbol{b} - \boldsymbol{g}(\boldsymbol{b})/L_1$. Hence, $\ell_1^+(\boldsymbol{\beta} \mid \boldsymbol{b})$ is minimized at

$$\hat{\boldsymbol{\beta}} = \boldsymbol{z}(\boldsymbol{b}).$$

As a result, the MM algorithm for minimizing $\ell_1(\boldsymbol{\beta})$ is given in Algorithm 2.

Algorithm 2 Update $\boldsymbol{\beta}$

Require: γ, ψ, L_1, initial vector $\boldsymbol{\beta}^{(0)}$ for $\boldsymbol{\beta}$
  $m \leftarrow 0$
  repeat
    $\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \boldsymbol{g}(\boldsymbol{\beta}^{(m)})/L_1$
    $m \leftarrow m + 1$
  until solution converges

2.2. Update γ

We next consider minimizing $\ell_2(\gamma)$ in (2.2) under given $\boldsymbol{\beta}$ and ψ to obtain an update equation for γ. Similar to the minimization of $\ell_1(\boldsymbol{\beta})$, we apply an MM algorithm to this minimization problem.

The following theorem gives the upper bound of $\ell_2(\gamma)$ (the proof is given in Appendix A.3).

Theorem 2. *Let* $L_2 = \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2$ *and for* $c \in \mathbb{R}$*, we define* $\ell_2^+(\gamma \mid c)$ *as*

$$\ell_2^+(\gamma \mid c) = \ell_2(c) + \dot{\ell}_2(c)(\gamma - c) + \frac{L_2}{2}(\gamma - c)^2,$$
$$\dot{\ell}_2(\gamma) = \frac{d}{d\gamma}\ell_2(\gamma) = -\sum_{i=1}^{n} \frac{\phi(\gamma r_i(\boldsymbol{\beta}))r_i(\boldsymbol{\beta})}{\Phi(\gamma r_i(\boldsymbol{\beta}))}.$$

Then, we have

$$\ell_2(\gamma) \leq \ell_2^+(\gamma \mid c), \quad \ell_2^+(c \mid c) = \ell_2(c).$$

From the above theorem, we obtain the surrogate function $\ell_2^+(\gamma)$ of $\ell_2(\gamma)$. Hence, the minimizer of $\ell_2(\gamma)$ can be obtained by repeating the following update:

$$\gamma^{(m+1)} = \arg\min_{\gamma \in \mathbb{R}} \ell_2^+(\gamma \mid \gamma^{(m)}) \quad (m = 0, 1, \ldots),$$

where $\gamma^{(0)}$ is an initial value for γ. Since $\ell_2^+(\gamma \mid c)$ is a quadratic function of γ, $\ell_2^+(\gamma \mid c)$ is minimized at

$$\hat{\gamma} = c - \frac{\dot{\ell}_2(c)}{L_2}. \tag{2.5}$$

As a result, the MM algorithm for minimizing $\ell_2(\gamma)$ is given in Algorithm 3.

Algorithm 3 Update γ

Require: $\boldsymbol{\beta}$, ψ, L_2, initial value $\gamma^{(0)}$ for γ
  $m \leftarrow 0$
  repeat
    $\gamma^{(m+1)} = \gamma^{(m)} - \dot{\ell}_2(\gamma^{(m)})/L_2$
    $m \leftarrow m + 1$
  until solution converges

2.3. Initial values

Although the objective function $\ell(\boldsymbol{\xi})$ in (1.3) is convex for each parameter ($\boldsymbol{\beta}$, γ, and ψ), i.e., $\ell_1(\boldsymbol{\beta})$, $\ell_2(\gamma)$, and $\ell_3(\psi)$ are all convex, it is not convex for $\boldsymbol{\xi}$. Hence, estimation results may depend on the initial values and the algorithm. In fact, although the following point is a stationary point, it is not a local minimum; rather, it is a saddle point.

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \\ \psi \end{pmatrix} = \begin{pmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ 0 \\ n/\boldsymbol{y}'\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right\}\boldsymbol{y} \end{pmatrix},$$

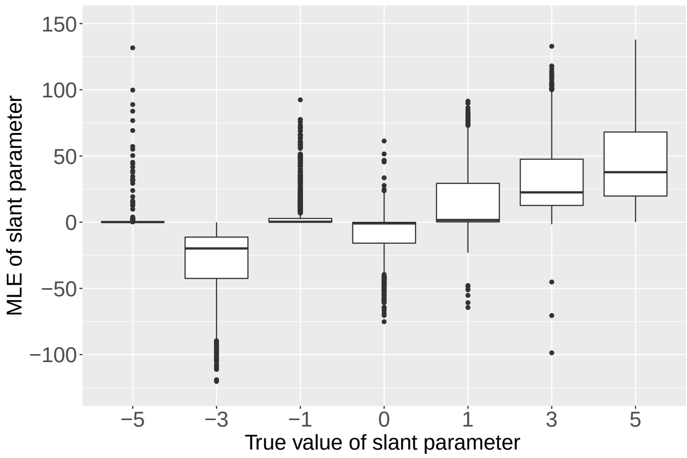

Figure 2. Unstableness of the MLE of the slant parameter using the EM algorithm

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Figure 2 is the EM algorithm version of figure 1 and shows a large difference between the two algorithms. Regarding the initial values for the algorithms, `selm` and the EM algorithm adopt the following moment estimators:

$$\boldsymbol{\beta}^\dagger = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \left(\delta^\dagger\sqrt{2/\pi\psi^\dagger}, \mathbf{0}_k'\right)', \quad \gamma^\dagger = \sqrt{\psi^\dagger}\nu^\dagger, \quad \psi^\dagger = \frac{1-(2/\pi)(\delta^\dagger)^2}{s^2},$$

where

$$\nu^\dagger = \frac{R}{\sqrt{2/\pi - (1-2/\pi)R^2}}, \quad R = \left(\frac{2u}{4-\pi}\right)^{1/3}, \quad \delta^\dagger = \frac{\nu^\dagger}{\sqrt{1+(\nu^\dagger)^2}},$$
$$s^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i, \quad u = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i-\bar{y}}{s}\right)^3.$$

Problems with the sign of $\gamma^\dagger$ then occur in the EM algorithm. Table 1 summarizes the sign of

Table 1. Summary of the sign results using the EM algorithm

	True value of γ					
	-5	-3	-1	1	3	5
A	1000	0	999	54	46	0
B	0	–	0	0	0	–

the results in figure 2. Let s^* be the sign of the true value of γ, $\hat{s}$ be the sign of the MLE of γ

obtained with the EM algorithm, and $s^\dagger = \mathrm{sign}(\gamma^\dagger)$. In the table, "A" denotes the number satisfying $s^\dagger \neq s^*$, and "B" denotes the number satisfying $\hat{s} = s^*$ when $s^\dagger \neq s^*$. Table 1 suggests that if the sign of the initial value differs from the sign of the true value, then the sign of the MLE will also differ from the sign of the true value. The cause of this may be that $\gamma = 0$ is a saddle point, and, in the algorithm, a current solution cannot move across a saddle point. From the above, it appears that using the moment estimator as the initial value for γ is risky, and the same problem may occur in our MM algorithm.

To mitigate this problem, we devised a simple approach. The procedure can be described as follows:

1. Calculate the two MLEs $\hat{\boldsymbol{\xi}}^-$ and $\hat{\boldsymbol{\xi}}^+$, where $\hat{\boldsymbol{\xi}}^-$ is the MLE of $\boldsymbol{\xi}$ under the initial value $\gamma^{(0)} = -1$ and $\hat{\boldsymbol{\xi}}^+$ is the MLE of $\boldsymbol{\xi}$ under the initial value $\gamma^{(0)} = 1$.
2. Select $\gamma^{(0)} = -1$ if $\ell(\hat{\boldsymbol{\xi}}^-) < \ell(\hat{\boldsymbol{\xi}}^+)$ and $\gamma^{(0)} = 1$ if $\ell(\hat{\boldsymbol{\xi}}^-) > \ell(\hat{\boldsymbol{\xi}}^+)$.

By selecting the sign of the initial value, we can expect that the problem with respect to the sign of the initial value is solved. For simplicity, we refer to using the moment estimator as the initial value for γ as initial value-type I (IV-I) and refer to selecting the sign of the initial value for γ using the above approach as initial value-type II (IV-II). For IV-II, the initial vector and value for $\boldsymbol{\beta}$ and ψ are their saddle points. The performances of IV-I and IV-II are numerically compared in section 4.

3. Stable Estimation for the Slant Parameter via Ridge Shrinkage

In the previous section, we derived an MM algorithm to minimize the objective function $\ell(\boldsymbol{\xi})$ in (1.3). In this section, we incorporate ridge shrinkage into our MM algorithm to reduce the unstableness of the MLE of slant parameter γ. Specifically, γ is updated based on minimizing the following penalized function, which adds a penalty term to the objective function for γ in (2.2).

$$\tilde{\ell}_2(\gamma \mid \theta) = \ell_2(\gamma) + \frac{\theta}{2}\gamma^2,$$

where $\theta \geq 0$ is a regularization parameter called the ridge parameter, which adjusts the strength of the penalty. This means that $\boldsymbol{\xi}$ is estimated based on minimizing the following penalized negative log-likelihood function:

$$\tilde{\ell}(\boldsymbol{\xi} \mid \theta) = \ell(\boldsymbol{\xi}) + \frac{\theta}{2}\gamma^2.$$

Then, $\boldsymbol{\beta}$ and ψ are updated, similar to the previous section. On the other hand, the update of γ

is accomplished by replacing the surrogate function in our MM algorithm with

$$\tilde{\ell}_2^+(\gamma \mid \theta, c) = \ell_2^+(\gamma \mid c) + \frac{\theta}{2}\gamma^2,$$

and its minimizer is given by

$$\hat{\gamma}_\theta = \frac{L_2 c - \dot{\ell}_2(c)}{L_2 + \theta}.$$

When $\theta = 0$, $\hat{\gamma}_\theta$ coincides with $\hat{\gamma}$ in (2.5). Since θ is non-negative, the estimator of γ is shrunk towards zero via ridge shrinkage. Hence, the MM algorithm with ridge shrinkage that provides a stable estimator of γ is given by replacing the update equation of γ in Algorithm 3 in the following equation:

$$\gamma^{(m+1)} = \frac{L_2 \gamma^{(m)} - \dot{\ell}_2(\gamma^{(m)})}{L_2 + \theta}.$$

As above, ridge shrinkage for γ can be easily implemented. However, the ridge estimator of γ depends on θ, which is an unknown parameter, and the amount of shrinkage of the estimator varies according th the value of θ. Hence, the stableness of the ridge estimator is entrusted to θ, meaning that the optimization of θ is extremely important. In general, as an optimization method for regularization parameters in penalized estimation methods such as ridge regression and Lasso (Tibshirani, 1996), cross-validation or a model selection criterion minimization method using, for example, the C_p criterion (Mallows, 1973) or the GCV criterion (Craven & Wahba, 1979), is employed (e.g., Zou, 2006; Ohishi *et al.*, 2020). In terms of the calculation cost, the model selection criterion minimization method is more reasonable. In this paper, for optimizing θ, the generalized information criterion (GIC; Konishi & Kitagawa, 1996) is applied. Let $\hat{\boldsymbol{\xi}}_\theta$ be the estimator for $\boldsymbol{\xi}$ under ridge parameter θ, i.e.,

$$\hat{\boldsymbol{\xi}}_\theta = \arg \min_{\boldsymbol{\xi} \in \mathbb{R}^{k+3}} \tilde{\ell}(\boldsymbol{\xi} \mid \theta),$$

and we rewrite $\ell(\boldsymbol{\xi})$ and $\tilde{\ell}(\boldsymbol{\xi} \mid \theta)$ as

$$\ell(\boldsymbol{\xi}) = \sum_{i=1}^n \ell_{(i)}(\boldsymbol{\xi}), \quad \ell_{(i)}(\boldsymbol{\xi}) = -\frac{1}{2}\log\frac{2\psi}{\pi} + \frac{\psi}{2} r_i(\boldsymbol{\beta})^2 - \log \Phi(\gamma r_i(\boldsymbol{\beta})),$$

$$\tilde{\ell}(\boldsymbol{\xi} \mid \theta) = \sum_{i=1}^n \tilde{\ell}_{(i)}(\boldsymbol{\xi} \mid \theta), \quad \tilde{\ell}_{(i)}(\boldsymbol{\xi} \mid \theta) = \ell_{(i)}(\boldsymbol{\xi}) + \frac{\theta}{2n}\gamma^2.$$

Then, the GIC for optimizing θ is given by

$$\mathrm{GIC}(\theta) = 2\ell(\hat{\boldsymbol{\xi}}_\theta) + 2\,\mathrm{tr}\left\{\boldsymbol{B}(\hat{\boldsymbol{\xi}}_\theta \mid \theta)\right\}, \ \boldsymbol{B}(\boldsymbol{\xi} \mid \theta) = \left\{\frac{\partial^2}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}'}\tilde{\ell}(\boldsymbol{\xi} \mid \theta)\right\}^{-1} \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\xi}}\tilde{\ell}_{(i)}(\boldsymbol{\xi} \mid \theta)\frac{\partial}{\partial \boldsymbol{\xi}'}\ell_{(i)}(\boldsymbol{\xi}),$$

and the value of θ optimized by the GIC minimization method is given by

$$\hat{\theta} = \arg \min_{\theta \in [0,\infty)} \mathrm{GIC}(\theta).$$

4. Numerical Study

In this section, we use simulation to evaluate the performance of our MM algorithm with ridge shrinkage for the stable estimation of slant parameter γ. Performance is measured in terms of the following mean square error (MSE):

$$\mathrm{MSE}(\theta) = \mathrm{E}\left[(\hat{\gamma}_\theta - \gamma)^2\right],$$

where $\hat{\gamma}_\theta$ is the ridge estimator of γ and the expectation in the MSE is evaluated via a Monte Carlo simulation with 1,000 iterations. When $\theta = 0$, $\hat{\gamma}_\theta$ is the MLE, and we can write $\mathrm{MSE}_0 = \mathrm{MSE}(0)$. The simulation data are generated by

$$\boldsymbol{y} \sim SN(\boldsymbol{X}\mathbf{1}_k, 1, \gamma), \quad \boldsymbol{X} = \boldsymbol{X}_0\boldsymbol{\Psi}(0.5)^{1/2},$$

where $\boldsymbol{X}_0$ is an $n \times k$ matrix in which the elements are identically and independently distributed according to $U(-1, 1)$, and $\boldsymbol{\Psi}(\rho)$ is a $k \times k$ matrix in which the (i, j)th element is given by $\rho^{|i-j|}$. The ridge parameter is compared for six values: $\hat{\theta}$, which is optimized by the GIC minimization method, $\theta_1 = |\gamma^\dagger|$, $\theta_2 = 1/n$, $\theta_3 = 1/\sqrt{n}$, $\theta_4 = 1/\log n$, and $\theta_5 = 1/2 \log\log n$.

Tables 2, 3, and 4 show the MSE results when $k = 10, 30, 50$, respectively. The minimum value in each row is in bold font; the "mean" and "s. d." values in the bottom two rows are the mean and standard deviation of the values in each column. In terms of the MLE values, it can be seen that the estimation results are relatively stable when n is large or $|\gamma|$ is small. In other cases, MSE_0 takes a very large value. Moreover, MSE_0 becomes larger as k increases. From the results, it appears that a substantial amount of shrinkage of the estimator is needed when n is small or $|\gamma|$ is large. On the other hand, under $\hat{\theta}$, which is optimized by the GIC minimization method, we can see the relationship $\mathrm{MSE}(\hat{\theta}) < \mathrm{MSE}_0$ in numerous cases and can thus say that ridge shrinkage tends to stabilize the estimator of γ. However, despite the fact that $\mathrm{MSE}(\hat{\theta})$ is less than MSE_0 in many cases, it is difficult to conclude that $\mathrm{MSE}(\hat{\theta})$ is sufficiently small. Hence, we compared five fixed values of θ: $\theta_1, \ldots, \theta_5$. If $|\gamma^\dagger|$ corresponds to $|\gamma|$, we can expect that θ_1 adjusts the amount of shrinkage of the estimator according to the data's skewness. Unfortunately, such a relationship does not occur in this simulation. Nevertheless, when $k = 10$, we can see the relationship $\mathrm{MSE}(\theta_1) < \mathrm{MSE}(\hat{\theta})$ in many cases. On the other hand, when $k = 30, 50$, $\mathrm{MSE}(\theta_1)$ decreases and the inequality is reversed. The reason for this is that $|\gamma^\dagger|$ decreases as k increases. The $\theta_2, \ldots, \theta_5$ values decrease as n increases. Since these values shrink more when n is small, we can expect an improvement in MSE when n is small. In fact, θ_4 and θ_5 are shown to greatly improve MSE and perform well even when n is small. Regarding the mean and standard deviation values, the tables show that $\mathrm{MSE}(\theta_4)$ with IV-II is

Table 2. MSE of the ridge estimate of the slant parameter when $k = 10$

		IV-I							IV-II						
γ	n	MLE	$\hat{\theta}$	θ_1	θ_2	θ_3	θ_4	θ_5	MLE	$\hat{\theta}$	θ_1	θ_2	θ_3	θ_4	θ_5
-5	100	3079.633	1755.592	4.642	64.639	4.329	4.218	8.832	3220.495	1883.991	2.411	69.157	2.012	**1.989**	8.932
	300	29.426	28.635	21.185	24.486	21.492	21.168	21.305	53.171	49.971	0.697	18.343	1.827	**0.637**	2.264
	500	12.043	11.323	10.996	11.957	11.347	10.982	11.399	2.327	1.112	0.539	2.191	1.133	**0.494**	1.321
-3	100	72.616	30.861	8.950	12.166	9.132	8.965	8.857	1352.255	659.376	1.420	49.390	2.759	**1.390**	3.093
	300	1.362	1.308	0.445	1.164	0.738	0.517	0.415	1.278	1.175	0.317	1.070	0.627	0.389	**0.284**
	500	7.841	7.667	7.773	7.840	7.822	7.780	7.667	0.329	0.194	0.213	0.327	0.291	0.223	**0.151**
-1	100	147.501	60.569	0.981	11.971	1.477	1.047	**0.834**	105.681	65.189	1.880	12.402	3.024	2.126	1.218
	300	0.823	0.737	0.784	0.822	0.809	0.784	0.718	0.701	0.638	0.662	0.700	0.689	0.664	**0.585**
	500	**0.150**	0.181	0.167	0.150	0.152	0.159	0.240	0.374	0.352	0.362	0.374	0.373	0.366	0.358
0	100	63.394	21.904	0.775	8.756	1.297	0.848	**0.426**	39.967	26.093	1.556	8.181	2.317	1.695	0.843
	300	0.373	0.327	0.333	0.373	0.365	0.348	**0.255**	0.669	0.606	0.607	0.668	0.656	0.629	0.471
	500	0.289	0.258	0.271	0.289	0.285	0.275	**0.212**	0.512	0.483	0.486	0.512	0.506	0.491	0.388
1	100	60.285	29.476	1.883	7.097	2.337	1.880	1.258	78.815	47.004	2.133	9.950	2.739	2.100	**1.217**
	300	**0.230**	0.275	0.263	0.230	0.236	0.251	0.397	0.699	0.642	0.644	0.698	0.686	0.661	0.584
	500	**0.147**	0.189	0.166	0.147	0.150	0.158	0.244	0.405	0.386	0.387	0.405	0.402	0.393	0.363
3	100	1373.003	611.626	1.053	45.785	2.127	**1.005**	2.084	1178.254	565.713	1.373	43.154	2.277	1.107	2.863
	300	2.763	3.394	2.036	2.561	2.224	2.030	1.930	2.988	3.337	0.388	1.675	0.643	0.371	**0.258**
	500	7.791	7.618	7.726	7.790	7.772	7.730	7.618	0.280	0.172	0.192	0.279	0.250	0.194	**0.145**
5	100	77.545	31.879	24.963	27.298	25.063	24.975	24.965	3193.562	1888.662	3.937	71.499	**1.826**	1.914	9.216
	300	27.392	26.235	22.869	24.679	23.101	22.881	22.897	34.745	31.200	0.682	14.136	1.747	**0.599**	2.218
	500	25.079	24.828	24.955	25.076	25.038	24.980	24.834	1.846	0.799	**0.427**	1.768	1.007	0.464	1.342
	mean	237.604	126.423	6.820	13.585	7.014	6.809	7.018	441.398	248.909	1.015	14.613	1.323	**0.900**	1.815
	s. d.	715.085	395.596	8.931	16.844	8.917	8.935	8.940	992.947	573.621	0.938	23.002	0.922	**0.670**	2.570

Table 3. MSE of the ridge estimate of the slant parameter when $k = 30$

		IV-I							IV-II						
γ	n	MLE	$\hat{\theta}$	θ_1	θ_2	θ_3	θ_4	θ_5	MLE	$\hat{\theta}$	θ_1	θ_2	θ_3	θ_4	θ_5
-5	100	3340.218	135.726	241.501	35.852	27.607	25.055	33.343	7740.139	383.102	239.430	10.417	**4.769**	16.932	199.353
	300	2533.607	1257.167	615.286	16.652	1.921	**1.733**	565.294	2470.163	1247.638	612.927	16.838	1.990	1.740	577.862
	500	178.721	270.903	103.445	7.723	2.372	2.011	181.459	187.718	278.012	108.746	7.165	1.420	**1.011**	188.720
-3	100	7915.044	512.400	298.282	18.743	4.696	**3.327**	376.412	6497.484	508.018	267.550	19.496	6.496	6.301	308.859
	300	399.183	242.753	141.542	6.689	1.739	**0.250**	84.494	413.449	283.664	146.274	7.417	1.859	0.261	124.659
	500	4.021	6.516	3.718	2.725	2.386	2.013	5.415	3.634	6.717	2.973	1.053	0.584	**0.143**	5.872
-1	100	2736.400	108.361	240.752	25.587	9.946	2.555	20.277	2673.413	171.049	228.397	28.202	11.553	**1.953**	88.468
	300	1.914	1.552	1.911	1.837	1.747	1.351	1.601	1.374	3.463	1.314	1.089	0.997	**0.771**	2.764
	500	**0.165**	0.195	0.166	0.168	0.175	0.257	0.227	0.479	0.430	0.479	0.473	0.461	0.418	0.471
0	100	2271.708	109.716	231.158	26.587	9.258	**1.758**	23.211	1869.591	138.810	217.745	30.876	13.431	2.447	64.039
	300	0.449	0.355	0.448	0.437	0.415	**0.302**	0.426	0.888	0.730	0.887	0.869	0.831	0.610	0.877
	500	0.304	0.257	0.304	0.300	0.289	**0.221**	0.277	0.592	0.533	0.592	0.586	0.569	0.449	0.581
1	100	3386.430	116.721	239.150	22.588	7.337	**1.429**	21.037	2798.681	181.635	233.451	29.631	11.947	2.155	91.135
	300	0.435	0.416	0.434	0.418	**0.401**	0.486	0.499	1.052	2.823	1.047	0.995	0.927	0.729	2.405
	500	0.623	0.554	0.623	0.618	0.607	0.579	0.567	0.471	0.438	0.471	0.467	0.459	**0.423**	0.475
3	100	7889.808	493.031	292.849	18.842	5.063	**3.883**	392.612	6464.111	491.159	264.161	19.524	6.241	6.185	313.556
	300	9.256	9.018	9.253	9.212	9.138	8.993	8.985	417.582	279.631	150.591	7.919	1.834	**0.238**	127.439
	500	9.225	9.033	9.225	9.208	9.168	9.033	9.001	4.141	5.230	3.166	1.052	0.549	**0.131**	4.414
5	100	8417.202	289.513	243.637	7.480	**1.586**	9.587	171.270	7113.067	296.824	223.972	9.613	4.698	17.615	127.497
	300	446.758	227.894	123.793	24.066	21.879	21.729	126.054	2663.678	1317.416	653.165	16.894	1.931	**1.718**	649.095
	500	175.403	250.940	101.697	7.225	1.381	**1.009**	158.937	178.716	259.486	104.498	7.446	1.429	1.016	165.827
	mean	1891.280	192.525	138.056	11.569	5.672	4.646	103.876	1976.211	278.896	164.849	10.382	3.570	**3.012**	144.970
	s. d.	2852.587	291.101	157.468	10.731	7.229	6.905	158.436	2688.666	372.682	187.270	10.307	**4.107**	5.046	184.230

Table 4. MSE of the ridge estimate of the slant parameter when $k = 50$

		IV-I							IV-II						
γ	n	MLE	$\hat{\theta}$	θ_1	θ_2	θ_3	θ_4	θ_5	MLE	$\hat{\theta}$	θ_1	θ_2	θ_3	θ_4	θ_5
-5	100	9743.796	1452.303	666.777	115.292	53.707	25.019	2101.914	10635.342	151.341	331.347	24.694	**14.759**	23.951	67.484
	300	5749.722	953.460	1294.735	57.142	6.744	**1.279**	586.022	5433.928	903.467	1265.201	55.601	6.944	1.362	540.524
	500	35.637	30.425	31.004	25.447	25.168	24.920	26.937	2415.377	1311.278	1255.374	48.227	5.387	**0.704**	627.718
-3	100	11884.151	277.741	396.865	35.793	10.004	**4.853**	205.160	9309.417	175.960	374.619	34.298	12.206	8.614	104.639
	300	1529.782	231.142	448.550	28.400	8.818	4.983	103.383	3145.230	512.400	938.296	52.029	8.982	**0.473**	271.802
	500	194.896	179.197	128.450	9.295	2.193	0.222	74.526	221.349	207.845	144.406	11.544	2.512	**0.221**	103.560
-1	100	5868.098	345.562	434.582	71.824	30.577	5.262	367.858	5951.734	185.704	378.801	51.852	20.610	**2.049**	123.726
	300	111.912	21.212	52.147	4.474	1.358	**0.722**	5.983	58.135	29.767	34.802	5.170	2.249	1.059	22.471
	500	1.534	1.248	1.533	1.514	1.464	1.217	1.250	0.585	0.486	0.585	0.578	0.558	**0.481**	0.579
0	100	4152.802	156.127	376.218	64.088	29.705	5.594	118.404	4309.235	127.267	369.093	58.601	25.962	**3.369**	55.291
	300	61.373	13.530	36.078	3.777	1.258	**0.462**	2.572	26.980	16.654	19.413	4.576	2.245	0.945	11.327
	500	0.418	0.335	0.418	0.413	0.398	**0.310**	0.388	0.721	0.613	0.721	0.713	0.692	0.546	0.710
1	100	6335.383	166.210	412.667	61.267	24.937	3.818	122.312	5607.683	160.266	368.736	49.260	19.301	**2.040**	90.422
	300	74.549	14.641	34.422	3.770	1.874	1.135	4.383	57.487	28.230	34.282	5.300	2.356	**1.036**	18.338
	500	0.237	0.239	0.237	**0.219**	0.224	0.301	0.293	0.611	0.518	0.610	0.602	0.584	0.502	0.597
3	100	7859.552	1576.954	537.355	88.945	36.213	9.574	3456.729	8599.791	185.280	355.797	32.496	11.171	**8.677**	89.105
	300	9.443	9.044	9.418	9.293	9.197	9.012	9.000	3447.274	549.139	1023.605	54.833	8.999	**0.444**	298.168
	500	9.231	9.027	9.230	9.211	9.165	9.015	8.982	154.715	184.893	103.369	8.453	2.066	**0.196**	92.881
5	100	11786.820	137.250	327.419	16.793	**2.481**	13.922	65.453	9494.834	117.185	306.663	23.216	15.430	23.833	45.073
	300	5687.602	906.837	1280.751	56.431	6.715	**1.284**	583.921	5356.298	829.148	1249.881	55.002	6.848	1.325	479.613
	500	25.339	25.063	25.338	25.311	25.244	25.038	25.000	2418.222	1351.036	1261.193	47.816	5.328	**0.682**	614.979
	mean	3386.775	309.883	309.724	32.795	13.688	7.045	374.784	3649.759	334.689	467.466	29.755	8.342	**3.929**	174.238
	s. d.	4216.805	483.028	388.946	33.206	14.880	8.392	846.468	3614.851	419.397	479.031	22.430	7.352	**7.056**	211.187

best, except for the standard deviation when $k = 30$ (in this case, the standard deviation of MSE(θ_4) with IV-II is second best). From these results, it would appear that the combination of $\theta = \theta_4$ and IV-II is the best choice.

Finally, we can consider the initial values for the algorithm. Table 5 is our MM algorithm version of Table 1, where "B1" and "B2" are same as "B" in Table 1 and they indicate IV-I and IV-II, respectively. The figure indicates that the same problem that occurs in the case of the

Table 5. Summary of the sign results using the MM algorithm

	True value of γ					
	-5	-3	-1	1	3	5
A	1000	0	999	54	46	0
B1	0	–	0	0	0	–
B2	968	–	638	30	42	–

EM algorithm also occurs with IV-I, but the problem is less serious with IV-II. However, tables 2, 3, and 4 show that IV-II does not necessarily improve the MSE relative to IV-I. Nevertheless, IV-II is superior to IV-I if the estimate is appropriately shrunk. In MSE(θ_3) and MSE(θ_4), IV-II, on average, improves the MSE when compared to IV-I.

Acknowledgment This work was supported by JSPS KAKENHI Grant number JP18K03415, JP20H04151, JP21K13834.

References

Aigner, D., Lovell, C. A. K. & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *J. Econometrics*, **6**, 21–37.

Azzalini, A. (2022). *The R package* `sn`*: The skew-normal and related distributions such as the skew-t and the SUN (version 2.0.2)*. Università degli Studi di Padova, Italia.
URL: *https://cran.r-project.org/package=sn*

Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. R. Stat. Soc. Ser. B. Stat. Methodl.*, **61**, 579–602.

Azzalini, A. & Capitanio, A. (2014). *The Skew-Normal and Related Families*. Cambridge University Press. Cambridge.

Cancho, V. G., Lachos, V. H. & Ortega, E. M. M. (2010). A nonlinear regression model with skew-normal errors. *Statist. Papers*, **51**, 547–558.

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.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodl.*, **39**, 1–38.

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

Hunter, D. R. & Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist.*, **58**, 30–37.

Konishi, S. & Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika*, **83**, 875–890.

Kubokawa, T. & Srivastava, M. S. (2012). Selection of variables in multivariate regression models for large dimensions. *Comm. Statist. Theory Methods*, **41**, 2465–2489.

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

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
URL: *https://www.R-project.org/*

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

Yamamura, M., Yanagihara, H. & Srivastava, M. S. (2010). Variable selection in multivariate linear regression models with fewer observations than the dimension. *Japan. J. Appl. Stat.*, **39**, 1–19.

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

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

Appendix

A.1. Proof of Lemma 1

Let $f(x)$ be a function on $\mathbb{R}$ defined by

$$f(x) = \frac{x\phi(x)}{\Phi(x)} + \left\{ \frac{\phi(x)}{\Phi(x)} \right\}^2 = \frac{x\phi(x)\Phi(x) + \phi(x)^2}{\Phi(x)^2}.$$

To prove $f(x) \leq 1$ for all $x \in \mathbb{R}$, we show (1) $\lim_{x\to-\infty} f(x) = 1$ and (2) $f(x)$ is strictly decreasing on $\mathbb{R}$. The (1) is proved by repeatedly applying l'Hôpitals's rule as

$$\begin{aligned}\lim_{x\to-\infty} f(x) &= \lim_{x\to-\infty} \frac{\Phi(x)(1-x^2) - x\phi(x)}{2\Phi(x)} = \lim_{x\to-\infty} \frac{-x\Phi(x)}{\phi(x)} = \lim_{x\to-\infty} \frac{-\Phi(x) - x\phi(x)}{-x\phi(x)} \\ &= \lim_{x\to-\infty} \frac{x^2-2}{x^2-1} = \lim_{x\to-\infty} \left\{ 1 - \frac{1}{x^2-1} \right\} = 1.\end{aligned}$$

Regarding (2), the derivative of $f(x)$ is given by

$$\frac{d}{dx} f(x) = \frac{\phi(x)}{\Phi(x)^3} g(x), \quad g(x) = (1-x^2)\Phi(x)^2 - 3x\phi(x)\Phi(x) - 2\phi(x)^2.$$

To show that $f(x)$ is strictly decreasing, it is sufficient to show $g(x) < 0$. Since $\lim_{x\to-\infty} g(x) = 0$, it is sufficient to show $\dot{g}(x) = dg(x)/dx < 0$. The first-order and second-order derivatives of $g(x)$ are given by

$$\begin{aligned}\dot{g}(x) &= -2x\Phi(x)^2 + (x^2-1)\phi(x)\Phi(x) + x\phi(x)^2, \\ \ddot{g}(x) &= \frac{d^2}{dx^2} g(x) = -2\Phi(x)^2 - (x^3+x)\phi(x)\Phi(x) - x^2\phi(x)^2.\end{aligned}$$

Since $\dot{g}(0) = -\phi(0)/2 < 0$, and $\ddot{g}(x) < 0$ for $x \geq 0$, we have $\dot{g}(x) < 0$ for $x \geq 0$. For the case $x < 0$, we consider a higher derivative. The third-order derivative of $g(x)$ is given by

$$\dddot{g}(x) = \frac{d^3}{dx^3} g(x) = \phi(x)h(x), \quad h(x) = (x^4 - 2x^2 - 5)\Phi(x) + (x^3 - 3x)\phi(x),$$

and the first-order and second-order derivatives of $h(x)$ are given by

$$\begin{aligned}\dot{h}(x) &= 4(x^3 - x)\Phi(x) + 4(x^2-2)\phi(x), \\ \ddot{h}(x) &= 4(3x^2-1)\Phi(x) + 12x\phi(x) = -4\Phi(x) + 12x\{x\Phi(x) + \phi(x)\}.\end{aligned}$$

Now, the following equations hold:

$$\lim_{x\to-\infty} \{x\Phi(x) + \phi(x)\} = 0, \quad \frac{d}{dx}\{x\Phi(x) + \phi(x)\} = \Phi(x) > 0.$$

These results give $x\Phi(x) + \phi(x) > 0$, and hence, $\ddot{h}(x) < 0$ holds for $x < 0$, which gives, with the fact that $\lim_{x\to-\infty} \dot{h}(x) = 0$, that $\dot{h}(x) < 0$ for $x < 0$ and $\lim_{x\to-\infty} h(x) = 0$. Hence, $h(x) < 0$ holds for $x < 0$. Moreover, we have $\dddot{g}(x) < 0$ for $x < 0$. This result and $\lim_{x\to-\infty} \ddot{g}(x) = 0$ lead $\ddot{g}(x) < 0$ for $x < 0$. Thus, $\ddot{g}(x)$ is always negative and $\lim_{x\to-\infty} \dot{g}(x) = 0$ holds. From the above, $\dot{g}(x) < 0$ holds and $f(x)$ is strictly decreasing.

Consequently, Lemma 1 is proved.

A.2. Proof of Theorem 1

The second-order Taylor expansion near $\boldsymbol{\beta} = \boldsymbol{b}$ gives

$$\ell_1(\boldsymbol{\beta} \mid 0) = \ell_1(\boldsymbol{b} \mid 0) + \boldsymbol{g}(\boldsymbol{b})'(\boldsymbol{\beta} - \boldsymbol{b}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{b})'\boldsymbol{H}\left(\tau\boldsymbol{\beta} + (1-\tau)\boldsymbol{b}\right)(\boldsymbol{\beta} - \boldsymbol{b}) \quad (\tau \in (0,1)),$$

$$\boldsymbol{H}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\ell_1(\boldsymbol{\beta} \mid 0) = \psi\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' + \gamma^2\sum_{i=1}^n \frac{\phi(\gamma r_i(\boldsymbol{\beta}))\{\gamma r_i(\boldsymbol{\beta})\Phi(\gamma r_i(\boldsymbol{\beta})) + \phi(\gamma r_i(\boldsymbol{\beta}))\}}{\Phi(\gamma r_i(\boldsymbol{\beta}))^2}\boldsymbol{x}_i\boldsymbol{x}_i'.$$

Moreover, $\boldsymbol{H}(\boldsymbol{\beta})$ can be rewritten as

$$\boldsymbol{H}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\left[\psi + \gamma^2\left\{\frac{\gamma r_i(\boldsymbol{\beta})\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))} + \left(\frac{\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))}\right)^2\right\}\right].$$

For all $\boldsymbol{a} \in \mathbb{R}^{k+1}$, it holds from Lemma 1 that

$$\begin{aligned}\boldsymbol{a}'\boldsymbol{H}(\boldsymbol{\beta})\boldsymbol{a} &= \sum_{i=1}^n \boldsymbol{a}'\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{a}\left[\psi + \gamma^2\left\{\frac{\gamma r_i(\boldsymbol{\beta})\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))} + \left(\frac{\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))}\right)^2\right\}\right] \\ &\le (\psi + \gamma^2)\boldsymbol{a}'\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{a} \le d_{\max}(\psi + \gamma^2)\|\boldsymbol{a}\|^2,\end{aligned}$$

where $d_{\max}$ is the maximum eigenvalue of $\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$. Consequently, we have

$$\ell_1(\boldsymbol{\beta} \mid 0) \le \ell_1(\boldsymbol{b} \mid 0) + \boldsymbol{g}(\boldsymbol{b})'(\boldsymbol{\beta} - \boldsymbol{b}) + \frac{d_{\max}(\psi + \gamma^2)}{2}\|\boldsymbol{\beta} - \boldsymbol{b}\|^2,$$

and Theorem 1 is proved.

A.3. Proof of Theorem 2

The second-order Taylor expansion near $\gamma = c$ gives

$$\ell_2(\gamma) = \ell_2(c) + \dot{\ell}_2(c)(\gamma - c) + \frac{1}{2}\ddot{\ell}_2\left(\tau\gamma + (1-\tau)c\right)(\gamma - c)^2 \quad (\tau \in (0,1)),$$

$$\ddot{\ell}_2(\gamma) = \frac{d^2}{d\gamma^2}\ell_2(\gamma) = \sum_{i=1}^n \frac{r_i(\boldsymbol{\beta})^2\phi(\gamma r_i(\boldsymbol{\beta}))\{\gamma r_i(\boldsymbol{\beta})\Phi(\gamma r_i(\boldsymbol{\beta})) + \phi(\gamma r_i(\boldsymbol{\beta}))\}}{\Phi(\gamma r_i(\boldsymbol{\beta}))^2}.$$

Moreover, Lemma 1 leads to

$$\ddot{\ell}_2(\gamma) = \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 \left[\frac{\gamma r_i(\boldsymbol{\beta})\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))} + \left\{ \frac{\phi(\gamma r_i(\boldsymbol{\beta}))}{\Phi(\gamma r_i(\boldsymbol{\beta}))} \right\}^2 \right] \leq \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2.$$

Consequently, Theorem 2 is proved.