Electronic Journal of Statistics
Vol. 10 (2016) 2537–2560
ISSN: 1935-7524
DOI: 10.1214/16-EJS1179

AIC for the Lasso in generalized linear models

Yoshiyuki Ninomiya

Institute of Mathematics for Industry
Kyushu University
Nishi-ku, Fukuoka 819-0395, Japan
e-mail: nino@imi.kyushu-u.ac.jp

and

Shuichi Kawano

Graduate School of Informatics and Engineering
The University of Electro-Communications
Chofu, Tokyo 182-8585, Japan
e-mail: skawano@ai.is.uec.ac.jp

Abstract: The Lasso is a popular regularization method that can simultaneously do estimation and model selection. It contains a regularization parameter, and several information criteria have been proposed for selecting its proper value. While any of them would assure consistency in model selection, we have no appropriate rule to choose between the criteria. Meanwhile, a finite correction to the AIC has been provided in a Gaussian regression setting. The finite correction is theoretically assured from the viewpoint not of the consistency but of minimizing the prediction error and does not have the above-mentioned difficulty. Our aim is to derive such a criterion for the Lasso in generalized linear models. Towards this aim, we derive a criterion from the original definition of the AIC, that is, an asymptotically unbiased estimator of the Kullback-Leibler divergence. This becomes the finite correction in the Gaussian regression setting, and so our criterion can be regarded as its generalization. Our criterion can be easily obtained and requires fewer computational tasks than does cross-validation, but simulation studies and real data analyses indicate that its performance is almost the same as or superior to that of cross-validation. Moreover, our criterion is extended for a class of other regularization methods.

MSC 2010 subject classifications: Primary 62J07, 62J12; secondary 62E20, 62F12.
Keywords and phrases: Convexity lemma, information criterion, Kullback-Leibler divergence, statistical asymptotic theory, tuning parameter, variable selection.

Received March 2016.

Contents

1. Introduction

The Lasso (Tibshirani 1996) is a regularization method that imposes an ℓ_1 penalty term $\lambda||\boldsymbol{\beta}||_1$ on an estimating function with respect to an unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, where λ (> 0) is the regularization parameter. If $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, \ldots, \hat{\beta}_{\lambda,p})^{\mathrm{T}}$ is the estimator of $\boldsymbol{\beta}$ by the Lasso, several of its components will be shrunk to exactly 0 when λ is not close to 0, which means that the Lasso can simultaneously do estimation and model selection. In addition, the Lasso is computationally feasible in general, and so it is one of key methods in current and future research directions in the area of statistics and machine learning. On the other hand, as mentioned in Meinshausen and Bühlmann (2010), a remaining challenge is to select the proper value for the regularization parameter λ. The estimated parameters continuously shrink toward 0 as λ increases, and so the selection of λ is important for the selection of an appropriate model.

One of the simplest methods for selecting λ is to use cross-validation (CV; Stone 1974). On the other hand, Meinshausen and Bühlmann (2010) proposed a stability selection method based on subsampling in order to avoid problems caused by selecting a model based on only one value of λ. Their method of stability selection avoids the problem they discussed, and it is a new and attractive model selection scheme; however, it requires a considerable number of computational tasks, comparable to the number required for CV.

As analytical methods for selecting λ, in general, there are two approaches. The first obtains a class of λ that has desirable properties for model selection or estimation accuracy, under some regularity conditions. For example, Zhao and Yu (2006) and Meinshausen and Yu (2009) obtained the expression $\lambda = \lambda_n$, which depends on at least data size, n; this requires that the model selection be consistent. In addition, for example, Bunea, Tsybakov and Wegkamp (2007), van de Geer (2008), Wainwright (2009), Sun and Zhang (2012), and Chételat, Lederer and Salmon (2014) obtained a more rigorous evaluation that is essentially of the form $\mathrm{P}(\text{estimation error} \leq \delta_\lambda) \geq 1 - \epsilon_\lambda$ for a class of λ, where

δ_λ and ϵ_λ are constants depending on at least λ. The second approach uses an information criterion that takes the form of $-2l(\hat{\boldsymbol{\beta}}_\lambda) + \eta_\lambda$, where $l(\cdot)$ is the log-likelihood function and η_λ is a penalty term that depends on at least λ; they showed that the model selection based on the λ that minimizes the information criterion is consistent (e.g., Yuan and Lin 2007; Wang, Li and Leng 2009; Zhang, Li and Tsai 2010; Fan and Tang 2013). Both approaches to selecting λ include the results for the case in which the dimension of the parameter vector p goes to infinity, and these are valuable because the Lasso with this value of λ has been shown to have desirable properties. However, the choice of λ remains somewhat arbitrary. When λ_n and η_λ, as defined above, satisfy the consistency requirement, $c \times \lambda_n$ and $c \times \eta_\lambda$ also satisfy it for a fixed coefficient c. For the rigorous evaluation in the first approach, no value of λ minimizes both δ_λ and ϵ_λ, and so we have no appropriate rule for choosing the optimal value of λ. It will be a severe problem because this arbitrariness for the choice of λ leads to the arbitrariness of model selection.

In a Gaussian linear regression setting, an appropriate selection of λ theoretically assured from the viewpoint of classic statistics can be achieved through an information criterion obtained by Efron et al. (2004) and Zou, Hastie and Tibshirani (2007). They derived an unbiased estimator of the true prediction error as a C_P-type criterion, through an elegant use of Stein's unbiased estimation theory (Stein 1981). In other words, we can say that they derived a finite correction to Akaike's information criterion (AIC; Akaike 1973) (Sugiura 1978; Hurvich and Tsai 1989) for the Lasso in Gaussian settings with known variance, because in these settings the true prediction error becomes essentially the same as the Kullback-Leibler divergence (Kullback and Leibler 1951). The corrected AIC can be expressed as $-2l(\hat{\boldsymbol{\beta}}_\lambda) + 2|\{j : \hat{\beta}_{\lambda,j} \neq 0\}|$, and so it is easy to use for model selection. Our aim in this paper is to derive such an information criterion for the Lasso in more general settings that is assured theoretically and can be computed without heavy computational tasks. Towards this aim, we obtain an asymptotically unbiased estimator of the Kullback-Leibler divergence, that is, we obtain the AIC for the Lasso, based on its original definition under the framework of generalized linear models (see McCullagh and Nelder 1983).

The remainder of this paper is organized as follows. After introducing in Section 2 generalized linear models and the assumptions for our asymptotic theory, some limiting distributions for the Lasso estimator are given in Section 3. The purpose of our asymptotic theory is to approximate some statistics for the given finite samples, and the limiting distributions obtained based on the asymptotic theory are different from those in Knight and Fu (2000). In Section 4, we use the limiting distributions to achieve our main goal which is the derivation of the AIC for the Lasso, and in Sections 5 and 6, its validity is demonstrated for several models thorough the performance of simulations and real data analyses. In Section 7, the AIC for the Lasso is extended for several cases including the cases of using other penalty terms, and a discussion is provided in Section 8. The program code used in Sections 5 and 6 is available from `https://sites.google.com/site/shuichikawanoen/research/aic.r`.

2. Setting and assumptions

Let us consider a natural exponential family with a natural parameter $\boldsymbol{\theta}$ ($\in \Theta \subset \mathbb{R}^r$) for an r-dimensional random variable $\boldsymbol{y}$, whose density is

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \exp\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\theta} - a(\boldsymbol{\theta}) + b(\boldsymbol{y})\}$$

with respect to a σ-finite measure μ. We assume that $\boldsymbol{\theta}$ in Θ satisfies $0 < \int \exp\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\theta} + b(\boldsymbol{y})\}\mathrm{d}\mu(\boldsymbol{y}) < \infty$, that is, Θ is the natural parameter space. Then all the derivatives of $a(\boldsymbol{\theta})$ and all the moments of $\boldsymbol{y}$ exist in the interior Θ^{int} of Θ, and, in particular, $\mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{y}) = a'(\boldsymbol{\theta})$ and $\mathrm{V}_{\boldsymbol{\theta}}(\boldsymbol{y}) = a''(\boldsymbol{\theta})$. For a function $c(\boldsymbol{\eta})$, we will denote $\partial c(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ and $\partial^2 c(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}$ by $c'(\boldsymbol{\eta})$ and $c''(\boldsymbol{\eta})$, respectively. We assume that $\mathrm{V}_{\boldsymbol{\theta}}(\boldsymbol{y}) = a''(\boldsymbol{\theta})$ is positive definite and so $-\log f(\boldsymbol{y};\boldsymbol{\theta})$ is a convex function with respect to $\boldsymbol{\theta}$.

Let $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ be the i-th set of responses and regressors ($1 \leq i \leq n$); we assume that the $\boldsymbol{y}_i$ are independent r-dimensional random vectors and $\boldsymbol{X}_i$ are $(r \times p)$-matrices of known constants. We will consider generalized linear models with natural link functions for such data, that is, we consider a class of density functions $\{f(\boldsymbol{y};\boldsymbol{X}_i\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathcal{B}\}$ for $\boldsymbol{y}_i$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a coefficient vector and $\mathcal{B}$ is an open convex set. We denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^*$.

Before developing the asymptotic theory for this model, let us explain about two types of aims for using asymptotic theory. The first aim is to approximate something well for the case where the data size goes to infinity in the future. The second aim is to approximate something well for the given real data whose size is large but of course finite. Assumptions which suit the first one sometimes do not suit the second one. For example, for a regression model with regressors $\{\boldsymbol{x}_i\}$, let us consider the case where the limiting frequency distribution of $\{\boldsymbol{x}_i : 1 \leq i < \infty\}$, $p_\infty(\cdot)$, is quite different from the frequency distribution of the given real data $\{\boldsymbol{x}_i : 1 \leq i \leq n_0\}$, $p_{n_0}(\cdot)$. In this situation, the asymptotic variance of regression coefficients' estimator should be evaluated based on $p_\infty(\cdot)$ if we want to evaluate the variance in the future. If we want to evaluate the variance for the given real data, however, we should evaluate it based on not $p_\infty(\cdot)$ but $p_{n_0}(\cdot)$. That is, as a limiting frequency distribution of $\{\boldsymbol{x}_i\}$, although to assume $p_\infty(\cdot)$ will suit the first aim, it is better for the second aim to assume $p_{n_0}(\cdot)$. Roughly speaking, our asymptotic theory is to approximate some statistics for the given real data $\{(\boldsymbol{y}_i, \boldsymbol{X}_i) : 1 \leq i \leq n_0\}$. In light of this, we assume

(C1) $\{\boldsymbol{X}_i\}$ lies in a compact set $\mathcal{X}$ with $\boldsymbol{X}\boldsymbol{\beta} \in \Theta^{\mathrm{int}}$ for all $\boldsymbol{X} \in \mathcal{X}$ and $\boldsymbol{\beta} \in \mathcal{B}$,

and

(C2) $\sum_{i=1}^n a(\boldsymbol{X}_i\boldsymbol{\beta})/n$, $\sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a'(\boldsymbol{X}_i\boldsymbol{\beta})/n$ and $\sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i/n$ converge with a rate $\mathrm{o}(1/\sqrt{n})$ for each $\boldsymbol{\beta}$, and the limit of $\sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i/n$ is positive definite,

even if we know that $\{\boldsymbol{X}_i\}$ diverges in the future.

Let $g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})$ be the log-likelihood for $\boldsymbol{y}_i$, i.e., $g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) = \log f(\boldsymbol{y}_i;\boldsymbol{X}_i\boldsymbol{\beta})$. Under the above-mentioned model with the conditions (C1) and (C2), we obtain several expressions that will be used for our asymptotic theory, as follows:

(R1) There exists a convex and differentiable function $h(\boldsymbol{\beta})$ such that $\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n \xrightarrow{\mathrm{p}} h(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$;

(R2) $\sum_{i=1}^{n} \mathrm{E}\{-g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n - \partial h(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathrm{o}(1/\sqrt{n})$.

(R3) There exists a positive definite matrix $\boldsymbol{J}(\boldsymbol{\beta})$ such that $\boldsymbol{J}_n(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} \mathrm{E}\{-g''_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n \to \boldsymbol{J}(\boldsymbol{\beta})$;

(R4) $\sum_{i=1}^{n}[g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) - \mathrm{E}\{g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}]/\sqrt{n} \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\beta}^*))$.

The expression (R3) is a direct consequence of (C2) because $-g''_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) = \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i$. The other expressions will be proved in the appendix.

3. Limiting distribution

Let $||\cdot||_1$ be the ℓ_1 norm, i.e., $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p}|\beta_j|$. For the above-mentioned model, the Lasso estimator of $\boldsymbol{\beta}^*$ is

$$\hat{\boldsymbol{\beta}}_\lambda \equiv \underset{\boldsymbol{\beta}\in\mathcal{B}}{\operatorname{argmin}} \Big\{ -\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) + n\lambda||\boldsymbol{\beta}||_1 \Big\}, \tag{1}$$

where λ is a regularization parameter. Here we put n in the penalty term for our asymptotic theory; this corresponds to the setting in Theorem 1 of Knight and Fu (2000) which provided the limiting value but did not discuss the limiting distribution. If we assume that the penalty term is $\mathrm{o}(n)$, $\hat{\boldsymbol{\beta}}_\lambda$ converges in probability to $\boldsymbol{\beta}^*$. We think, however, this closeness between $\hat{\boldsymbol{\beta}}_\lambda$ and $\boldsymbol{\beta}^*$ does not reflect the characteristic of $\hat{\boldsymbol{\beta}}_\lambda$ for the given real data, because it is moved to $\boldsymbol{0}$ from $\boldsymbol{\beta}^*$. As mentioned in the previous section, the purpose of our asymptotic theory is to approximate some statistics for the given real data. The penalty term is assumed to be $\mathrm{O}(n)$, because in this case, $\hat{\boldsymbol{\beta}}_\lambda$ converges to a vector made by moving $\boldsymbol{\beta}^*$ close to $\boldsymbol{0}$ (this will be shown below). Actually, for more tractable regularization methods, an information criterion is already derived by assuming $\mathrm{O}(n)$ for the penalty term (see, e.g., Konishi and Kitagawa 2008).

To consider the limiting value of $\hat{\boldsymbol{\beta}}_\lambda$, we define a random function, as follows:

$$u_n(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\} + \lambda||\boldsymbol{\beta}||_1.$$

The function $u_n(\boldsymbol{\beta})$ is convex with respect to $\boldsymbol{\beta}$, and $\operatorname{argmin}_{\boldsymbol{\beta}\in\mathcal{B}} u_n(\boldsymbol{\beta})$ is equal to $\hat{\boldsymbol{\beta}}_\lambda$. In Knight and Fu (2000), the same type of random function was defined for the Gaussian case, but their function did not sum over the $g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*)$. We did this, however, so we would not need to be concerned with $b(\boldsymbol{y})$ in $g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})$. From (R1), we see that $u_n(\boldsymbol{\beta})$ converges in probability to $h(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1$ for each $\boldsymbol{\beta}$. Because $u_n(\boldsymbol{\beta})$ is a convex function with respect to $\boldsymbol{\beta}$, similarly to in Knight and Fu (2000), we can apply the convexity lemma from Andersen and Gill (1982) or Pollard (1991).

We assume the condition

(C3) $h(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1$ has a unique minimum in $\mathcal{B}$,

and we denote $\mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{h(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1\}$ by $\boldsymbol{\beta}^{**} = (\beta_1^{**}, \ldots, \beta_p^{**})^{\mathrm{T}}$. Because $h(\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_1$ is a convex function, (C3) will hold if $\mathcal{B}$ is sufficiently large. Using the convexity lemma, we obtain the following result.

Lemma. *The Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda$ in (1) converges in probability to $\boldsymbol{\beta}^{**}$ under the conditions (C1), (C2), and (C3).*

Because $h(\boldsymbol{\beta})$ is convex and differentiable, we can easily check that

$$\beta_j^{**} = 0 \quad \Leftrightarrow \quad -\lambda < \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} < \lambda \tag{2}$$

and

$$\beta_j^{**} \neq 0 \quad \Leftrightarrow \quad \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} = -\lambda \times \mathrm{sgn}(\beta_j^{**}). \tag{3}$$

Let us denote $\{j : \beta_j^{**} = 0\}$ and $\{j : \beta_j^{**} \neq 0\}$ by $\mathcal{J}^{(1)}$ and $\mathcal{J}^{(2)}$, respectively. In addition, for a p-dimensional vector $\boldsymbol{\beta}$, an $(r \times p)$-matrix $\boldsymbol{X}$, and a $(p \times p)$-matrix $\boldsymbol{J}$, the vector $(\beta_j)_{j\in\mathcal{J}^{(k)}}$ is denoted by $\boldsymbol{\beta}^{(k)}$, the matrix $(\boldsymbol{X}_{ij})_{1\leq i\leq r, j\in\mathcal{J}^{(k)}}$ is denoted by $\boldsymbol{X}^{(k)}$, and the matrix $(\boldsymbol{J}_{ij})_{i\in\mathcal{J}^{(k)}, j\in\mathcal{J}^{(l)}}$ is denoted by $\boldsymbol{J}^{(kl)}$ $(k, l \in \{1, 2\})$. Note that $\boldsymbol{\beta}^{**(1)} = \boldsymbol{0}$. Moreover, we sometimes express, for example, $\boldsymbol{\beta}$ by $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$. We need this notation in order to investigate the asymptotic behavior of $\hat{\boldsymbol{\beta}}_\lambda$ in more detail, since the asymptotic behaviors of $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ and $\hat{\boldsymbol{\beta}}_\lambda^{(2)}$ are different.

Let us define another random function, as follows:

$$\begin{aligned} v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \equiv \sum_{i=1}^{n} \Big\{ & g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**}) - g_{\boldsymbol{y}_i, \boldsymbol{X}_i}\Big(\frac{\boldsymbol{u}^{(1)}}{n}, \frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} + \boldsymbol{\beta}^{**(2)}\Big)\Big\} \\ & + n\lambda\Big|\Big|\frac{\boldsymbol{u}^{(1)}}{n}\Big|\Big|_1 + n\lambda\Big|\Big|\frac{\boldsymbol{u}^{(2)}}{\sqrt{n}} + \boldsymbol{\beta}^{**(2)}\Big|\Big|_1 - n\lambda||\boldsymbol{\beta}^{**}||_1. \end{aligned}$$

Note that $\mathrm{argmin}_{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}\, v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (n\hat{\boldsymbol{\beta}}_\lambda^{(1)}, \sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)}))$. Using the Taylor expansion around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\boldsymbol{0}, \boldsymbol{0})$, this random function can be expressed as

$$\begin{aligned} & -\sum_{i=1}^{n}\Big\{g'^{(1)}_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})^{\mathrm{T}}\frac{\boldsymbol{u}^{(1)}}{n} + g'^{(2)}_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})^{\mathrm{T}}\frac{\boldsymbol{u}^{(2)}}{\sqrt{n}}\Big\} \\ & -\sum_{i=1}^{n}\Big[\frac{\boldsymbol{u}^{(1)\mathrm{T}}}{n}\Big\{g''^{(11)}_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\frac{\boldsymbol{u}^{(1)}}{2n} + g''^{(12)}_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\frac{\boldsymbol{u}^{(2)}}{\sqrt{n}}\Big\} + \frac{1}{2}\frac{\boldsymbol{u}^{(2)\mathrm{T}}}{\sqrt{n}}g''^{(22)}_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\frac{\boldsymbol{u}^{(2)}}{\sqrt{n}}\Big] \\ & + \lambda||\boldsymbol{u}^{(1)}||_1 + \sqrt{n}\lambda\boldsymbol{u}^{(2)\mathrm{T}}\mathrm{sgn}(\boldsymbol{\beta}^{**(2)}) + \mathrm{o_P}(1), \end{aligned}$$

where $\mathrm{sgn}(\boldsymbol{\beta}^{**(2)})$ is the vector whose components are $\mathrm{sgn}(\beta_j^{**})$ $(j \in \mathcal{J}^{(2)})$. In the quadratic form, the terms including $\boldsymbol{u}^{(1)}$ reduce to $o_P(1)$, and from (R3), the remainder term $-\sum_{i=1}^n (\boldsymbol{u}^{(2)}/\sqrt{n})^{\mathrm{T}} g''^{(22)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})(\boldsymbol{u}^{(2)}/\sqrt{n})/2$ converges to $\boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{J}(\boldsymbol{\beta}^{**})\boldsymbol{u}^{(2)}/2$. In addition, from (R2), (R4), and (3), we have that there exists a $|\mathcal{J}^{(2)}|$-dimensional random vector $\boldsymbol{s}^{(2)}$ having a $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^*))$ distribution, such that $\sum_{i=1}^n g'^{(2)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})/\sqrt{n} - \sqrt{n}\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)})$ converges in distribution to $\boldsymbol{s}^{(2)}$. We can also easily obtain that $-\sum_{i=1}^n g'^{(1)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})/n$ converges in probability to $\partial h(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^{(1)}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}$. Thus, for each $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$, it follows that $v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ converges in distribution to

$$
\begin{aligned}
&v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\
&\quad \equiv \sum_{j \in \mathcal{J}^{(1)}} \left\{ \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} u_j + \lambda |u_j| \right\} - \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}^{(2)} + \frac{1}{2} \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**}) \boldsymbol{u}^{(2)}.
\end{aligned}
$$

From (2), the first term is a non-negative function of $\boldsymbol{u}^{(1)}$. That is, $v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ has a unique minimum at $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)})$, and $v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ is convex. Similarly to in Knight and Fu (2000), we can apply the convexity lemma from Hjort and Pollard (1993) or Geyer (1996), and then we have $\mathrm{argmin}_{(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})} v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \xrightarrow{\mathrm{d}} \mathrm{argmin}_{(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})} v(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$.

Theorem 1. *For the Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda$ in (1), we have*

$$
n(\hat{\boldsymbol{\beta}}_\lambda^{(1)} - \boldsymbol{\beta}^{**(1)}) \xrightarrow{\mathrm{P}} \boldsymbol{0} \tag{4}
$$

and

$$
\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)}) \\
&\quad = \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n g'^{(2)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**}) - \sqrt{n}\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)}) \right\} + o_P(1) \\
&\quad \xrightarrow{\mathrm{d}} \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1} \boldsymbol{s}^{(2)} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1} \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^*) \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}),
\end{aligned} \tag{5}
$$

under the conditions (C1), (C2), and (C3).

A small generalization of the above-mentioned convexity lemma is required to prove (5), and so the proof will be given in the appendix.

4. Bias evaluation

Model selection can be approached by trying to reduce twice the Kullback-Leibler divergence (Kullback and Leibler 1951) between the true distribution and the estimated distribution,

$$
2\tilde{\mathrm{E}}\Big\{ \sum_{i=1}^n g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) \Big\} - 2\tilde{\mathrm{E}}\Big\{ \sum_{i=1}^n g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) \Big\},
$$

where $(\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n)$ is a copy of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, in other words, $(\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n)$ is distributed according to the distribution of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ and is independent of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, and $\tilde{\mathrm{E}}$ denotes the expectation with respect to only $(\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_n)$. Because the first term on the right-hand side does not depend on the model selection, we need to consider only the second term. A simple estimator of the second term is $-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda)$, but this underestimates it. We then consider minimizing the bias correction,

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{E}\Big[\sum_{i=1}^{n} g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}\Big\{\sum_{i=1}^{n} g_{\tilde{\boldsymbol{y}}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda)\Big\}\Big], \tag{6}$$

in AIC-type information criteria (see, e.g., Chapter 3 in Konishi and Kitagawa 2008). Because the second term depends on the true distribution, it cannot be given explicitly. In a Gaussian linear regression setting with a known common variance, that is, when $g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta})$ can be written as $-(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})/2 - (r/2)\log(2\pi)$ by standardizing the data, it can be shown by an elegant use of Stein's unbiased estimation theory (Stein 1981) that the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_\lambda$, $|\{j : \hat{\beta}_{\lambda,j} \neq 0\}|$ is an unbiased estimator of the second term (Efron et al. 2004; Zou, Hastie and Tibshirani 2007). This means that

$$\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_\lambda)^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_\lambda) + nr\log(2\pi) + 2|\{j : \hat{\beta}_{\lambda,j} \neq 0\}| \tag{7}$$

can be regarded as the AICc, a finite correction of the AIC (Sugiura 1978; Hurvich and Tsai 1989). However, this criterion cannot be extended for the general case, and so we evaluate the second term asymptotically in the same way as was done for the AIC. That is, considering that $\mathrm{E}[\sum_{i=1}^{n} g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}\{\sum_{i=1}^{n} g_{\tilde{\boldsymbol{y}}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda)\}]$ can be rewritten as the expectation of

$$\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} - \sum_{i=1}^{n}\{g_{\tilde{\boldsymbol{y}}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\tilde{\boldsymbol{y}}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\}, \tag{8}$$

we use

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{E}(z^{\mathrm{limit}}) \tag{9}$$

in place of (6), where z^{limit} is the limit to which (8) converges in distribution; we say that $\mathrm{E}(z^{\mathrm{limit}})$ is an asymptotic bias.

Now we evaluate (8). Using a Taylor expansion around $\boldsymbol{\beta}^{**}$, the first term can be expressed as

$$(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^{**})^{\mathrm{T}}\sum_{i=1}^{n} g'_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{**}) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^{**})^{\mathrm{T}}\Big\{\sum_{i=1}^{n} g''_{\boldsymbol{y}_i, \boldsymbol{X}_i}(\boldsymbol{\beta}^{\dagger})\Big\}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^{**}),$$

where $\boldsymbol{\beta}^{\dagger}$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^{**}$. Using (4) in Theorem 1, the terms including $\hat{\boldsymbol{\beta}}_\lambda^{(1)} - \boldsymbol{\beta}^{**(1)}$ reduce to $o_{\mathrm{P}}(1)$. For the remainder terms,

as shown in the previous section, $\sum_{i=1}^n g'^{(2)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})/\sqrt{n} - \sqrt{n}\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)})$ converges in distribution to $\boldsymbol{s}^{(2)}$. In addition, we can easily check from (R3) and Lemma 3 that $\sum_{i=1}^n g''^{(22)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{\dagger})/n = -\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**}) + \mathrm{o_P}(1)$. Thus, by using also (5) in Theorem 1, we have

$$\sum_{i=1}^n \{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} - (\hat{\boldsymbol{\beta}}^{(2)}_\lambda - \boldsymbol{\beta}^{**(2)})^{\mathrm{T}} n\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)}) \xrightarrow{\mathrm{d}} \frac{1}{2}\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}. \tag{10}$$

Using the same type of Taylor expansion, the second term on the right-hand side of (8) can be expressed as

$$(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^{**})^{\mathrm{T}} \sum_{i=1}^n g'_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**}) + \frac{1}{2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^{**})^{\mathrm{T}} \Big\{ \sum_{i=1}^n g''_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{\ddagger}) \Big\} (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^{**}),$$

where $\boldsymbol{\beta}^{\ddagger}$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^{**}$. Similarly to in the above analysis, we see that the terms including $\hat{\boldsymbol{\beta}}^{(1)}_\lambda - \boldsymbol{\beta}^{**(1)}$ reduce to $\mathrm{o_P}(1)$, and $\sum_{i=1}^n g''^{(2)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{\ddagger})/n$ converges to $-\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})$. In addition, there exists a $|\mathcal{J}^{(2)}|$-dimensional random vector $\tilde{\boldsymbol{s}}^{(2)}$ having a $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*}))$ distribution, such that $\sum_{i=1}^n g'^{(2)}_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})/\sqrt{n} - \sqrt{n}\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)})$ converges in distribution to $\tilde{\boldsymbol{s}}^{(2)}$. Then, we have

$$\sum_{i=1}^n \{g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} - (\hat{\boldsymbol{\beta}}^{(2)}_\lambda - \boldsymbol{\beta}^{**(2)})^{\mathrm{T}} n\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)}) \xrightarrow{\mathrm{d}} \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\tilde{\boldsymbol{s}}^{(2)} - \frac{1}{2}\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}. \tag{11}$$

Thus, it follows from (10) and (11) that

$$z^{\mathrm{limit}} = -\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\tilde{\boldsymbol{s}}^{(2)} + \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}$$

Because $\boldsymbol{s}^{(2)}$ and $\tilde{\boldsymbol{s}}^{(2)}$ are independently distributed according to $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*}))$, we obtain the following theorem.

Theorem 2. *The asymptotic bias in (9) is given by*

$$\mathrm{E}(z^{\mathrm{limit}}) = \mathrm{tr}\{\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*})\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\}$$

under the conditions (C1), (C2), and (C3).

We cannot know the values of $\boldsymbol{\beta}^{*}$ or $\boldsymbol{\beta}^{**}$, and so we replace $\mathrm{tr}\{\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*})\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\}$ by its consistent estimator. Let $\hat{\mathcal{J}}^{(2)} = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$ for $\hat{\boldsymbol{\beta}}_\lambda = (\beta_{\lambda,1}, \ldots, \beta_{\lambda,p})^{\mathrm{T}}$, which is called an active set, and let $\boldsymbol{J}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i/n$. Defining $|\hat{\mathcal{J}}^{(2)}| \times |\hat{\mathcal{J}}^{(2)}|$ matrices $\hat{\boldsymbol{J}}^{*(22)}_n$ and $\hat{\boldsymbol{J}}^{**(22)}_n$ as $(\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_0)_{jk})_{j\in\hat{\mathcal{J}}^{(2)},k\in\hat{\mathcal{J}}^{(2)}}$ and $(\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_\lambda)_{jk})_{j\in\hat{\mathcal{J}}^{(2)},k\in\hat{\mathcal{J}}^{(2)}}$, respectively, we have

$$\mathrm{tr}(\hat{\boldsymbol{J}}^{*(22)}_n\hat{\boldsymbol{J}}^{**(22)-1}_n) = \mathrm{tr}\{\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*})\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\} + \mathrm{o_P}(1). \tag{12}$$

See the appendix for the proof. Thus, we propose the following index as an AIC for the Lasso:

$$\mathrm{AIC}_{\lambda}^{\mathrm{Lasso}} = -2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_{\lambda}) + 2\mathrm{tr}(\hat{\boldsymbol{J}}_n^{*(22)}\hat{\boldsymbol{J}}_n^{**(22)-1}). \tag{13}$$

Here we use $\hat{\boldsymbol{\beta}}_0$ as a consistent estimator of $\boldsymbol{\beta}^*$ as done in the adaptive Lasso (Zou 2006). When $\hat{\boldsymbol{\beta}}_0$ is expected to be unstable, for example, when p is large, we propose the use of a more stable but consistent estimator in place of $\hat{\boldsymbol{\beta}}_0$. We thus have only to select the λ that minimizes this $\mathrm{AIC}_{\lambda}^{\mathrm{Lasso}}$.

When $g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) = -(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})/2 - (r/2)\log(2\pi)$, we have $a''(\boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{I}_r$, where $\boldsymbol{I}_r$ is the $r \times r$ identity matrix. That is, $\boldsymbol{J}_n(\boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$ and so $\hat{\boldsymbol{J}}_n^{*(22)} = \hat{\boldsymbol{J}}_n^{**(22)}$, which means that (13) reduces to (7). Hence, the AIC in (13) can be regarded as a generalization of the AICc for the Gaussian linear regression when the variance is known.

5. Simulation study

In this section, to check the performance of the AIC in (13), we perform some simulation studies using logistic regression, Poisson regression, and Gaussian graphical models, and we compare it with CV and the criterion with the penalty term derived by Efron et al. (2004) and Zou, Hastie and Tibshirani (2007). This last criterion can be written as

$$\mathrm{AICc}_{\lambda}^{\mathrm{Lasso}} = -2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_{\lambda}) + 2|\hat{\mathcal{J}}^{(2)}|. \tag{14}$$

It is not theoretically assured, that is, it is not a finite correction of the AIC for the cases of the above models, but for simplicity, we will call it the AICc.

We assessed the performance in terms of the second term of the Kullback-Leibler divergence:

$$\mathrm{KL} = -\tilde{\mathrm{E}}\Big\{\sum_{i=1}^{n} g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})\Big\},$$

where $\hat{\lambda}$ is the value of λ selected by each criterion, $(\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_n)$ is a copy of $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n)$, and $\tilde{\mathrm{E}}$ denotes the expectation with respect to only $(\tilde{\boldsymbol{y}}_1,\ldots,\tilde{\boldsymbol{y}}_n)$. We evaluated the expectation using test datasets of size 1000. As a secondary index for the assessment, we also determined the rates of false positives and false negatives:

$$\mathrm{FP} = |\{j : \hat{\beta}_j \neq 0 \ \wedge \ \beta_j^* = 0\}|/|\{j : \beta_j^* = 0\}|$$

and

$$\mathrm{FN} = |\{j : \hat{\beta}_j = 0 \ \wedge \ \beta_j^* \neq 0\}|/|\{j : \beta_j^* \neq 0\}|,$$

for each of the three criteria. When a criterion has a larger FP but a smaller FN, compared to the other criteria, no conclusion can be made from this secondary index.

5.1. Logistic and Poisson regression models

As simple examples of the generalized linear model, here we consider a logistic regression model

$$g_{y_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) = y_i\boldsymbol{X}_i\boldsymbol{\beta} - \log\{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})\} \qquad (y_i \in \{0,1\})$$

and a Poisson regression model

$$g_{y_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) = y_i\boldsymbol{X}_i\boldsymbol{\beta} - \exp(\boldsymbol{X}_i\boldsymbol{\beta}) - \log y_i! \qquad (y_i \in \{0,1,2,\ldots\}),$$

where $\boldsymbol{X}_i$ is a $(1\times p)$-matrix of known constants, and $\boldsymbol{\beta}$ is a p-dimensional coefficient vector. For these models, $\boldsymbol{J}_n(\boldsymbol{\beta})$ can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})\}^2}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\exp(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i,$$

respectively, and so we can easily obtain the AIC in (13).

The simulation settings were as follows. For the p-dimensional regressors, we used vectors obtained from the multivariate Gaussian distribution $\mathrm{N}_p(\boldsymbol{0},\Sigma)$, where the (i,j)-element of Σ was set to $0.5^{|i-j|}$. Here we do not regard the regressors as random vectors. The true coefficient vector $\boldsymbol{\beta}^*$ was

$$\boldsymbol{\beta}^* = (\underbrace{\beta_1^*,\ldots,\beta_1^*}_{k},\underbrace{\beta_2^*,\ldots,\beta_2^*}_{k},\underbrace{0,\ldots,0}_{p-2k})^{\mathrm{T}},$$

and seven cases were considered for the pairs of p and k, as follows: $(p,k) = (8,1),(8,2),(8,3),(16,2),(32,2),(500,5),(1000,5)$. We generated a dataset of size $n=100$ or $n=200$, and used the Lasso to estimate the coefficient vector (we used the package `glmnet` in `R`). One hundred simulations were conducted.

Tables 1 and 2 show the averages and standard deviations of the KL, and the averages of the FP and FN for the logistic and Poisson regression models, respectively. In all cases, the average of the KL for the AIC is almost equal to or smaller than those for CV and the AICc. In the logistic regression, the average KL for CV tends to be clearly larger than that for the AIC when $n=100$ and p is large, and the average KL for the AICc is sometimes considerably larger than those for the AIC and CV especially when p is large. In the Poisson regression, because the average KL is almost the same for all criteria, we check the secondary index. Then, we see that the sum of FP and FN values for CV is sometimes clearly larger than those for the AIC and AICc especially when p is small. Thus, we can say that the AICc and CV perform poorly in comparison with the other criteria, at least in the case of the logistic and Poisson regressions, respectively. We can thus conclude that, overall in these simple examples, the AIC is superior to CV and the AICc.

TABLE 1
Comparison among the cross-validation (CV), the AIC in (13) and the AICc in (14) for the logistic regression models.

				$n = 100$			$n = 200$		
(β_1^*, β_2^*)	p	k		KL (SD)	FP	FN	KL (SD)	FP	FN
			CV	0.230 (0.018)	0.38	0.15	0.215 (0.006)	0.41	0.09
(6.0,0.5)	8	1	AIC	0.234 (——)	0.41	0.14	0.215 (——)	0.42	0.09
			AICc	0.259 (0.047)	0.48	0.13	0.222 (0.014)	0.47	0.09
			CV	0.161 (0.010)	0.41	0.18	0.139 (0.003)	0.43	0.09
(6.0,0.5)	8	2	AIC	0.158 (——)	0.38	0.19	0.138 (——)	0.40	0.10
			AICc	0.166 (0.028)	0.48	0.16	0.140 (0.007)	0.49	0.09
			CV	0.134 (0.008)	0.29	0.18	0.112 (0.004)	0.35	0.10
(6.0,0.5)	8	3	AIC	0.132 (——)	0.25	0.22	0.112 (——)	0.27	0.12
			AICc	0.130 (0.016)	0.34	0.16	0.110 (0.005)	0.37	0.09
			CV	0.173 (0.017)	0.39	0.21	0.147 (0.007)	0.45	0.12
(6.0,0.5)	16	2	AIC	0.171 (——)	0.38	0.22	0.147 (——)	0.41	0.13
			AICc	0.194 (0.041)	0.46	0.19	0.154 (0.015)	0.53	0.11
			CV	0.182 (0.017)	0.28	0.21	0.156 (0.009)	0.33	0.15
(6.0,0.5)	32	2	AIC	0.179 (——)	0.32	0.20	0.155 (——)	0.35	0.15
			AICc	0.201 (0.040)	0.28	0.23	0.167 (0.019)	0.40	0.15
			CV	0.263 (0.014)	0.05	0.43	0.190 (0.009)	0.08	0.32
(6.0,0.5)	500	5	AIC	0.254 (——)	0.05	0.43	0.187 (——)	0.07	0.34
			AICc	0.293 (0.026)	0.01	0.47	0.216 (0.018)	0.02	0.40
			CV	0.292 (0.015)	0.03	0.44	0.195 (0.008)	0.05	0.40
(6.0,0.5)	1000	5	AIC	0.283 (——)	0.03	0.43	0.195 (——)	0.04	0.40
			AICc	0.316 (0.027)	0.01	0.48	0.232 (0.015)	0.01	0.45
			CV	0.210 (0.014)	0.41	0.03	0.194 (0.005)	0.42	0.01
(6.5,1.0)	8	1	AIC	0.212 (——)	0.41	0.05	0.194 (——)	0.41	0.01
			AICc	0.233 (0.032)	0.51	0.03	0.199 (0.013)	0.47	0.01
			CV	0.149 (0.009)	0.41	0.08	0.127 (0.004)	0.47	0.01
(6.5,1.0)	8	2	AIC	0.147 (——)	0.36	0.10	0.127 (——)	0.38	0.02
			AICc	0.150 (0.016)	0.47	0.08	0.126 (0.009)	0.49	0.01
			CV	0.133 (0.010)	0.33	0.12	0.107 (0.003)	0.11	0.11
(6.5,1.0)	8	3	AIC	0.131 (——)	0.28	0.14	0.107 (——)	0.08	0.14
			AICc	0.129 (0.018)	0.35	0.11	0.104 (0.004)	0.12	0.11
			CV	0.165 (0.016)	0.39	0.09	0.138 (0.006)	0.45	0.03
(6.5,1.0)	16	2	AIC	0.160 (——)	0.38	0.10	0.137 (——)	0.39	0.03
			AICc	0.174 (0.029)	0.47	0.10	0.141 (0.012)	0.52	0.03
			CV	0.178 (0.018)	0.28	0.13	0.152 (0.007)	0.36	0.05
(6.5,1.0)	32	2	AIC	0.173 (——)	0.30	0.12	0.149 (——)	0.35	0.06
			AICc	0.190 (0.033)	0.30	0.14	0.159 (0.018)	0.44	0.05
			CV	0.277 (0.016)	0.05	0.36	0.195 (0.009)	0.08	0.24
(6.5,1.0)	500	5	AIC	0.267 (——)	0.05	0.36	0.192 (——)	0.07	0.24
			AICc	0.297 (0.026)	0.02	0.41	0.218 (0.016)	0.02	0.30
			CV	0.304 (0.018)	0.03	0.39	0.211 (0.009)	0.05	0.28
(6.5,1.0)	1000	5	AIC	0.293 (——)	0.03	0.38	0.208 (——)	0.04	0.29
			AICc	0.326 (0.026)	0.01	0.44	0.243 (0.018)	0.01	0.36

KL, the averages of the Kullback-Leibler divergences; SD, the standard deviations of the difference between the Kullback-Leibler divergences for CV or the AICc and for the AIC; FP, the averages of the false positive rates; FN, the averages of the false negative rates.

TABLE 2
Comparison among the cross-validation (CV), the AIC in (13) and the AICc in (14) for the Poisson regression models.

				$n = 100$			$n = 200$		
(β_1^*, β_2^*)	p	k		KL (SD)	FP	FN	KL (SD)	FP	FN
			CV	1.343 (0.012)	0.24	0.22	1.333 (0.006)	0.24	0.23
(0.5,0.1)	8	1	AIC	1.343 (——)	0.21	0.23	1.333 (——)	0.20	0.23
			AICc	1.343 (0.002)	0.20	0.23	1.333 (0.000)	0.20	0.23
			CV	1.366 (0.011)	0.29	0.11	1.350 (0.004)	0.31	0.10
(0.5,0.1)	8	2	AIC	1.366 (——)	0.22	0.13	1.349 (——)	0.26	0.12
			AICc	1.366 (0.001)	0.22	0.13	1.349 (0.002)	0.26	0.12
			CV	1.405 (0.017)	0.27	0.10	1.373 (0.009)	0.22	0.05
(0.5,0.1)	8	3	AIC	1.409 (——)	0.18	0.13	1.373 (——)	0.17	0.06
			AICc	1.410 (0.006)	0.19	0.13	1.372 (0.001)	0.17	0.06
			CV	1.380 (0.034)	0.26	0.17	1.348 (0.008)	0.23	0.09
(0.5,0.1)	16	2	AIC	1.381 (——)	0.24	0.19	1.349 (——)	0.20	0.11
			AICc	1.381 (0.002)	0.23	0.19	1.349 (0.000)	0.20	0.11
			CV	1.379 (0.038)	0.17	0.19	1.356 (0.015)	0.18	0.09
(0.5,0.1)	32	2	AIC	1.380 (——)	0.16	0.21	1.356 (——)	0.16	0.09
			AICc	1.379 (0.010)	0.14	0.22	1.356 (0.004)	0.16	0.10
			CV	2.399 (0.174)	0.05	0.26	1.755 (0.057)	0.05	0.10
(0.5,0.1)	500	5	AIC	2.377 (——)	0.04	0.27	1.741 (——)	0.05	0.10
			AICc	2.380 (0.027)	0.04	0.27	1.741 (0.003)	0.05	0.10
			CV	2.579 (0.150)	0.03	0.30	1.826 (0.074)	0.03	0.14
(0.5,0.1)	1000	5	AIC	2.574 (——)	0.02	0.31	1.820 (——)	0.03	0.14
			AICc	2.579 (0.055)	0.02	0.31	1.821 (0.020)	0.03	0.15
			CV	1.356 (0.013)	0.26	0.09	1.366 (0.004)	0.30	0.03
(0.6,0.2)	8	1	AIC	1.356 (——)	0.24	0.08	1.367 (——)	0.28	0.03
			AICc	1.356 (0.002)	0.24	0.08	1.367 (0.003)	0.27	0.03
			CV	1.404 (0.016)	0.35	0.04	1.370 (0.007)	0.36	0.01
(0.6,0.2)	8	2	AIC	1.404 (——)	0.28	0.04	1.369 (——)	0.27	0.01
			AICc	1.404 (0.004)	0.28	0.04	1.369 (0.000)	0.27	0.01
			CV	1.438 (0.057)	0.24	0.01	1.406 (0.010)	0.32	0.00
(0.6,0.2)	8	3	AIC	1.444 (——)	0.23	0.02	1.406 (——)	0.27	0.00
			AICc	1.443 (0.022)	0.23	0.02	1.405 (0.001)	0.29	0.00
			CV	1.379 (0.034)	0.25	0.06	1.374 (0.010)	0.28	0.01
(0.6,0.2)	16	2	AIC	1.380 (——)	0.21	0.07	1.374 (——)	0.26	0.01
			AICc	1.379 (0.002)	0.21	0.07	1.374 (0.003)	0.26	0.01
			CV	1.426 (0.033)	0.19	0.05	1.385 (0.015)	0.21	0.01
(0.6,0.2)	32	2	AIC	1.428 (——)	0.17	0.05	1.384 (——)	0.19	0.01
			AICc	1.423 (0.007)	0.15	0.06	1.384 (0.001)	0.19	0.01
			CV	5.870 (0.412)	0.05	0.09	2.583 (0.095)	0.05	0.00
(0.6,0.2)	500	5	AIC	5.823 (——)	0.04	0.10	2.572 (——)	0.05	0.01
			AICc	5.838 (0.197)	0.05	0.10	2.573 (0.035)	0.05	0.01
			CV	6.926 (0.703)	0.03	0.12	3.314 (0.309)	0.03	0.01
(0.6,0.2)	1000	5	AIC	6.748 (——)	0.03	0.12	3.277 (——)	0.03	0.01
			AICc	6.756 (0.239)	0.03	0.12	3.270 (0.056)	0.03	0.01

KL, the averages of the Kullback-Leibler divergences; SD, the standard deviations of the difference between the Kullback-Leibler divergences for CV or the AICc and for the AIC; FP, the averages of the false positive rates; FN, the averages of the false negative rates.

5.2. Gaussian graphical model

Suppose that a q-dimensional random vector $\boldsymbol{z}_i$ is distributed according to a multivariate Gaussian distribution $\mathrm{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Without loss of generality, we can assume that the mean vector is the zero vector. The graphical Lasso estimates the covariance matrix $\boldsymbol{\Sigma}$, under the assumption that the precision matrix $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$ is sparse (Yuan and Lin 2007; Friedman, Hastie and Tibshirani 2008). Let us denote the p-dimensional vectors $\mathrm{vech}(\boldsymbol{z}_i\boldsymbol{z}_i^{\mathrm{T}})$ and $-\mathrm{vech}\{\boldsymbol{C} - \mathrm{diag}(\boldsymbol{C})/2\}$ by $\boldsymbol{y}_i$ and $\boldsymbol{\beta}$, respectively, where $p = q(q+1)/2$ and $\mathrm{vech}(\cdot)$ is the half-vectorization of a symmetric matrix. Then, the log-likelihood of $\boldsymbol{y}_i$ is

$$g_{\boldsymbol{y}_i}(\boldsymbol{\beta}) = \boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{\beta} + \log|\boldsymbol{C}|^{1/2} - \log(2\pi)^{p/2}.$$

By simple calculations, we have $\partial(-\log|\boldsymbol{C}|^{1/2})/\partial(-\boldsymbol{C}_{ii}/2) = \boldsymbol{\Sigma}_{ii}$ and $\partial(-\log|\boldsymbol{C}|^{1/2})/\partial(-\boldsymbol{C}_{ij}) = \boldsymbol{\Sigma}_{ij}$, for $i \neq j$. Moreover, we obtain $\partial\boldsymbol{\Sigma}_{kl}/\partial(-\boldsymbol{C}_{ii}/2) = 2\boldsymbol{\Sigma}_{ki}\boldsymbol{\Sigma}_{il}$ and $\partial\boldsymbol{\Sigma}_{kl}/\partial(-\boldsymbol{C}_{ij}) = \boldsymbol{\Sigma}_{ki}\boldsymbol{\Sigma}_{jl} + \boldsymbol{\Sigma}_{kj}\boldsymbol{\Sigma}_{il}$, for $i \neq j$. Therefore, for this model, the elements of $\boldsymbol{J}_n(\boldsymbol{\beta})$ can be written as

$$\boldsymbol{J}_n(\boldsymbol{\beta})_{(i-1)q+j,(k-1)q+l} = \boldsymbol{\Sigma}_{ki}\boldsymbol{\Sigma}_{jl} + \boldsymbol{\Sigma}_{kj}\boldsymbol{\Sigma}_{il},$$

and so we can easily obtain the AIC in (13).

The simulation settings were those used by Yuan and Lin (2007). The models for the precision matrix are as follows:

(M1) AR(1) model with $c_{ii} = 1$ $(1 \leq i \leq q)$ and $c_{i,i-1} = c_{i-1,i} = 0.5$ $(2 \leq i \leq q)$.
(M2) AR(2) model with $c_{ii} = 1$ $(1 \leq i \leq q)$, $c_{i,i-1} = c_{i-1,i} = 0.5$ $(2 \leq i \leq q)$, and $c_{i,i-2} = c_{i-2,i} = 0.25$ $(3 \leq i \leq q)$.
(M3) AR(3) model with $c_{ii} = 1$ $(1 \leq i \leq q)$, $c_{i,i-1} = c_{i-1,i} = 0.4$ $(2 \leq i \leq q)$, $c_{i,i-2} = c_{i-2,i} = 0.3$ $(3 \leq i \leq q)$ and $c_{i,i-3} = c_{i-3,i} = 0.2$ $(4 \leq i \leq q)$.

We considered four cases for the pair of n and q, as follows: $(n, q) = (25, 5)$, $(50, 5)$, $(50, 10)$, $(100, 10)$. The parameter vector $\boldsymbol{\beta}$ was estimated by the graphical Lasso using the package `glasso` in `R`. The simulations were conducted 100 times.

Table 3 shows the averages and standard deviations of the KL, along with the averages of the FP and FN. Also in this model, the average of the KL for the AIC is almost equal to or smaller than those for CV and the AICc, but the differences are small. For the cases of $(n, q) = (25, 5), (50, 10)$, the difference between the KL averages for the AIC and CV becomes slightly large.

6. Real data analyses

We investigated the effectiveness of our criterion through real data analyses. We used eight benchmark datasets, which are depicted in Table 4. The "pima" and "biodegradation" datasets were available from the UCI database (http://archive.ics.uci.edu/ml/index.html). The "colon", "leukemia", "takeover.

TABLE 3
Comparison among the cross-validation (CV), the AIC in (13) and the AICc in (14) for the Gaussian graphical models.

		Model (M1)			Model (M2)			Model (M3)		
(n, q)		KL (SD)	FP	FN	KL (SD)	FP	FN	KL (SD)	FP	FN
(25,5)	CV	3.537 (0.101)	0.47	0.01	5.012 (0.194)	0.51	0.33	5.403 (0.164)	0.24	0.58
	AIC	3.493 (——)	0.46	0.00	4.946 (——)	0.57	0.25	5.326 (——)	0.40	0.41
	AICc	3.501 (0.050)	0.45	0.01	4.964 (0.084)	0.54	0.27	5.338 (0.058)	0.37	0.45
(50,5)	CV	3.030 (0.005)	0.50	0.00	4.494 (0.032)	0.55	0.16	4.864 (0.080)	0.58	0.34
	AIC	3.031 (——)	0.49	0.00	4.498 (——)	0.52	0.19	4.843 (——)	0.54	0.34
	AICc	3.031 (0.000)	0.49	0.00	4.500 (0.012)	0.52	0.20	4.846 (0.020)	0.53	0.35
(50,10)	CV	2.620 (0.000)	0.37	0.00	8.846 (0.042)	0.42	0.16	9.504 (0.075)	0.37	0.38
	AIC	2.620 (——)	0.37	0.00	8.841 (——)	0.42	0.16	9.477 (——)	0.40	0.34
	AICc	2.620 (0.000)	0.37	0.00	8.852 (0.035)	0.41	0.17	9.483 (0.040)	0.38	0.36
(100,10)	CV	2.299 (0.000)	0.36	0.00	8.358 (0.006)	0.37	0.10	9.005 (0.014)	0.35	0.31
	AIC	2.299 (——)	0.36	0.00	8.359 (——)	0.37	0.10	9.010 (——)	0.34	0.32
	AICc	2.299 (0.000)	0.36	0.00	8.360 (0.009)	0.37	0.11	9.011 (0.006)	0.33	0.32

KL, the averages of the Kullback-Leibler divergences; SD, the standard deviations of the difference between the Kullback-Leibler divergences for CV or the AICc and for the AIC; FP, the averages of the false positive rates; FN, the averages of the false negative rates.

bids.case", "doctor.visits", and "mathmark" datasets were, respectively, obtained from the packages `HiDimDA`, `plsgenomics`, `CountsEPPM`, `AER`, and `gRbase` in `R`. The "flow.cytometry" dataset was obtained from Sachs et al. (2005). Logistic regression models were applied into the "pima", "biodegradation", "colon", and "leukemia" datasets, Poisson regression models were applied into the "takeover.bids.case" and "doctor.visits" datasets, and Gaussian graphical models were applied into the "flow.cytometry" and "mathmark" datasets. The covariates were standardized for each dataset.

We randomly divided the observed data into training samples for constructing the models and test samples for computing the KL. The numbers of training samples are shown in Table 4, while the remaining observations were regarded as test samples. Note that we randomly selected 1000 observations in advance if the number of the observed data was larger than 1000. We repeated these procedures 10 times.

Table 5 shows the results. In almost all cases, the AIC is superior to the other criteria. In particular, the results for the "colon" and "leukemia" datasets suggests that the AIC is sometimes clearly superior to CV.

7. Extensions

The AIC in (13) can be extended for more general cases. In this section, we will indicate the broad possibilities of this by providing an actual AIC for some particular cases.

TABLE 4
Sample size and the number of covariates in real datasets, and the number of training samples used in each analysis.

	sample size	# of covariates	# of training samples	model
pima	200	7	100	logistic
biodegradation	1055	41	100	logistic
colon	62	2000	40	logistic
leukemia	38	3051	20	logistic
takeover.bids.case	126	14	50	Poisson
doctor.visits	5190	11	100	Poisson
flow.cytometry	7466	11	100	graphical
mathmark	88	5	50	graphical

TABLE 5
Averages (standard deviations) of the KL in real data analyses.

	CV	AIC	AICc
pima	0.4935 (0.0304)	0.4808 (———)	0.4939 (0.0365)
biodegradation	0.4481 (0.0359)	0.4436 (———)	0.4800 (0.0449)
colon	0.5303 (0.1506)	0.4640 (———)	0.4939 (0.0671)
leukemia	0.3733 (0.1007)	0.2863 (———)	0.3354 (0.1077)
takeover.bids.case	1.626 (0.028)	1.614 (——)	1.618 (0.011)
doctor.visits	0.7168 (0.0280)	0.7131 (———)	0.7193 (0.0196)
flow.cytometry	12.71 (0.24)	12.63 (——)	12.63 (0.00)
mathmark	3.395 (0.007)	3.392 (——)	3.392 (0.000)

7.1. Model with a nuisance parameter

For the Gaussian linear regression model in which each random vector $\boldsymbol{z}_i$ is independently distributed according to the r-dimensional Gaussian distribution $\mathrm{N}_r(\boldsymbol{X}_i\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}_r)$ with unknown variance parameter σ^2, the estimation is usually performed without any penalization, even if $\boldsymbol{\gamma}$ is estimated by the Lasso. We will begin by considering such an example, that is, a case in which there are several parameters with no penalty terms. For simplicity, we will denote all parameters by $\boldsymbol{\beta}$ as before, and let $\mathcal{J}$ be an index set of β_j, which is estimated without penalization. In this setting, the estimator of $\boldsymbol{\beta}$ can be written as $\hat{\boldsymbol{\beta}}_\lambda \equiv \operatorname{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{-\sum_{i=1}^n g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) + n\lambda\sum_{j\notin\mathcal{J}}|\beta_j|\}$.

Let us define $\boldsymbol{\beta}^{**}$ by $\operatorname{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{h(\boldsymbol{\beta}) + \lambda\sum_{j\notin\mathcal{J}}|\beta_j|\}$, $\mathcal{J}^{(1)}$ by $\{j : \beta_j^{**} = 0\} \cap \overline{\mathcal{J}}$ and $\mathcal{J}^{(2)}$ by $\{j : \beta_j^{**} \neq 0\} \cup \mathcal{J}$. As a result, Lemma 3, Theorem 1, and Theorem 2 hold. Their derivations are a little more complicated than those in Sections 3 and 4, because (2) and (3) hold only when $j \in \overline{\mathcal{J}}$ and $\partial h(\boldsymbol{\beta})/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} = 0$ always holds when $j \in \mathcal{J}$. Noting that $\hat{\mathcal{J}}^{(2)}$ includes $\mathcal{J}$ with probability one, the AIC in this case is also given by (13).

For the above-mentioned Gaussian linear regression model with unknown variance, if $\boldsymbol{\gamma}$ and σ^2 are estimated by the Lasso and without penalization, respectively, one might think that the penalty term in the AIC will be

$2(|\hat{\mathcal{J}}^{(2)}|+1)$. However, $\boldsymbol{J}_n(\boldsymbol{\beta})$ becomes

$$\boldsymbol{J}_n(\sigma^2,\boldsymbol{\gamma})=\sum_{i=1}^{n}\begin{pmatrix} 4\sigma^2\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\gamma}+2r\sigma^4 & 2\sigma^2\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i \\ 2\sigma^2\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i\boldsymbol{\gamma} & \sigma^2\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i \end{pmatrix}$$

in this setting, and so the penalty term in the AIC cannot be easily expressed by the number of elements in the active set.

7.2. Other convex penalties

Here, in place of $n\lambda||\boldsymbol{\beta}||_1$, we consider a general convex penalty term $n\lambda\eta(\boldsymbol{\beta})$. Since the Lasso can simultaneously do estimation and model selection by shrinking the estimators, we will assume that $\eta(\boldsymbol{\beta})$ is symmetric with respect to $\beta_j=0$ and nondifferentiable with respect to β_j at $\beta_j=0$. We assume that $\eta(\boldsymbol{\beta})$ is differentiable with respect to β_j at $\beta_j\neq 0$ for the sake of simplicity, and, also for simplicity, we will denote $\partial\eta(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\beta_j\to+0}$ by $\partial\eta(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\beta_j=0}$. The asymptotic properties of $\hat{\boldsymbol{\beta}}_\lambda\equiv\mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{-\sum_{i=1}^{n}g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})+n\lambda\eta(\boldsymbol{\beta})\}$ can be derived similarly to those of the Lasso estimators.

First, it follows from a convexity lemma that $\hat{\boldsymbol{\beta}}_\lambda$ converges in probability to $\boldsymbol{\beta}^{**}\equiv\mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{h(\boldsymbol{\beta})+\lambda\eta(\boldsymbol{\beta})\}$. In addition, (2) and (3) can be rewritten as

$$\beta_j^{**}=0 \quad\Leftrightarrow\quad -\lambda\frac{\partial\eta(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}<\frac{\partial h(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}<\lambda\frac{\partial\eta(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}$$

and

$$\beta_j^{**}\neq 0 \quad\Leftrightarrow\quad \frac{\partial h(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}=-\lambda\frac{\partial\eta(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}, \tag{15}$$

and we can show the pointwise convergence of

$$\begin{aligned} v_n(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})\equiv\sum_{i=1}^{n}&\Big\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})-g_{\boldsymbol{y}_i,\boldsymbol{X}_i}\Big(\frac{\boldsymbol{u}^{(1)}}{n},\frac{\boldsymbol{u}^{(2)}}{\sqrt{n}}+\boldsymbol{\beta}^{**(2)}\Big)\Big\} \\ &-n\lambda\Big\{\eta(\boldsymbol{\beta}^{**})-\eta\Big(\frac{\boldsymbol{u}^{(1)}}{n},\frac{\boldsymbol{u}^{(2)}}{\sqrt{n}}+\boldsymbol{\beta}^{**(2)}\Big)\Big\} \end{aligned} \tag{16}$$

to

$$\begin{aligned} &\sum_{j\in\mathcal{J}^{(1)}}\Big\{\frac{\partial h(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}u_j+\lambda\frac{\partial\eta(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}|u_j|\Big\} \\ &\quad-\boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{s}^{(2)}+\boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})\boldsymbol{u}^{(2)}/2 \end{aligned} \tag{17}$$

from the equality in (15). Here, $\boldsymbol{K}_\lambda(\boldsymbol{\beta})$ is $\boldsymbol{J}(\boldsymbol{\beta})+\lambda\partial^2\eta(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$, and $\lambda\partial^2\eta(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}$ is the term that does not exist for the Lasso. Then, from a convexity lemma and the inequality in (15), it holds that

$$n(\hat{\boldsymbol{\beta}}_\lambda^{(1)}-\boldsymbol{\beta}^{**(1)})\xrightarrow{\mathrm{P}}\boldsymbol{0} \tag{18}$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)}) \xrightarrow{\mathrm{d}} \boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}. \tag{19}$$

Let us consider the bias evaluation, in a way similar to what we did for the Lasso. By using

$$\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} - (\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)})^{\mathrm{T}} n\lambda \frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}$$
$$\xrightarrow{\mathrm{d}} \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)} - \frac{1}{2}\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}$$

in place of (10) and using

$$\sum_{i=1}^{n}\{g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) - g_{\tilde{\boldsymbol{y}}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})\} - (\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)})^{\mathrm{T}} n\lambda \frac{\partial \eta(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}$$
$$\xrightarrow{\mathrm{d}} \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\tilde{\boldsymbol{s}}^{(2)} - \frac{1}{2}\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}$$

in place of (11), we have $z^{\mathrm{limit}} = -\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\tilde{\boldsymbol{s}}^{(2)} + \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}$. Then we have $\mathrm{E}(z^{\mathrm{limit}}) = \mathrm{tr}\{\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{*})\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\}$ in place of Theorem 2. Thus, the AIC for this case is

$$-2\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{tr}\Big\{\hat{\boldsymbol{J}}_n^{*(22)}\Big(\hat{\boldsymbol{J}}_n^{**(22)} + \lambda\frac{\partial^2\eta(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(2)}\partial\boldsymbol{\beta}^{(2)\mathrm{T}}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda}\Big)^{-1}\Big\}. \tag{20}$$

7.3. Nonconvex penalties

Here we consider the case with a nonconvex penalty term $n\lambda\eta(\boldsymbol{\beta})$. We assume that $\eta(\boldsymbol{\beta})$ has the same properties as in Section 7.2, except for convexity, and we assume that $\eta(\boldsymbol{\beta})$ is a nondecreasing function with respect to each $|\beta_j|$. For simplicity, we will also assume that $\partial\eta(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\beta_j=0}$, which denotes $\partial\eta(\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\beta_j\to+0}$, is a finite value. This setting does not allow use of the convexity lemmas for showing the asymptotic properties of $\hat{\boldsymbol{\beta}}_\lambda \equiv \mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{-\sum_{i=1}^{n} g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) + n\lambda\eta(\boldsymbol{\beta})\}$, but we can use the same approach as was used for the nonconvex case treated in Knight and Fu (2000).

First, we can easily show that $u_n(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{*}) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n + \lambda\eta(\boldsymbol{\beta})$ uniformly converges in probability to $h(\boldsymbol{\beta}) + \lambda\eta(\boldsymbol{\beta})$ on any compact set of $\boldsymbol{\beta}$ and that $\mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\, u_n(\boldsymbol{\beta})$ is $\mathrm{O_P}(1)$. Thus, $\hat{\boldsymbol{\beta}}_\lambda$ converges in probability to $\boldsymbol{\beta}^{**} \equiv \mathrm{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\{h(\boldsymbol{\beta}) + \lambda\eta(\boldsymbol{\beta})\}$. We can also easily show that (16) converges uniformly to (17) on any compact set of $\boldsymbol{\beta}$, and so if $\mathrm{argmin}_{(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})} v_n(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})$ is $\mathrm{O_P}(1)$, (18) and (19) hold. It is necessary to place some conditions in order to assure that this is $\mathrm{O_P}(1)$. For example, such conditions can be that $v_n(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})$ is convex when n and $|\boldsymbol{u}^{(2)}|$ are large enough, and $\boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})$ is positive definite. In that case, we have $n(\hat{\boldsymbol{\beta}}_\lambda^{(1)} - \boldsymbol{\beta}^{**(1)}) \xrightarrow{\mathrm{p}} \boldsymbol{0}$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)}) \xrightarrow{\mathrm{d}} \boldsymbol{K}_\lambda^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}^{(2)}$.

The remaining evaluation of the bias is the same as in Section 7.2, and the AIC in this case can be given by (20).

8. Discussion

By generalizing and modifying the original definition of the AIC, various criteria have been proposed, e.g., the GIC (Konishi and Kitagawa 1996), the DIC (Spiegelhalter et al. 2002), the FIC (Claeskens and Hjort 2003), and the GAIC (Lv and Liu 2014). For the Lasso, a popular regularization method, there was not even a naive AIC, except for in the case of Gaussian linear regression. In this study, we used the original definition of the AIC to obtain an AIC for the Lasso for a generalized linear model. For several settings, BICs for the Lasso have been proposed, such as those by Yuan and Lin (2007) and Wang, Li and Leng (2009). But such BICs have not been derived from Bayes factors, and so the AIC in (13) can be regarded as the only criterion for the Lasso that has the same roots as those of the classic information criteria.

The penalty term in (13) is written by using an information matrix with respect to the active set, and simulation studies indicated that its value is close to twice the number of members in the active set. We can interpret this to mean that the active set contributes to the penalty by approximately the usual amount, and the other parameters do not. While the active set consists of parameters in the model that are selected by the Lasso, it is adaptively selected from the full set, and so the above interpretation is not necessarily obvious. As was remarked for the Gaussian case in Lockhart et al. (2014), the adaptive selection costs extra bias, and shrinking the nonzero coefficients decreases the bias by approximately the same amount. This is an interesting phenomenon, but it is not clear why they should be almost the same amount.

The selection of the regularization parameter for the Lasso by the AIC in (13) requires few computations, compared to those required by CV. Nevertheless, its performance is almost the same as or better than that of CV. It is particularly worth noting that the AIC is not inferior to CV even if the dimension of the coefficient vector p is particularly large, although the AIC is based on the asymptotic theory with a fixed p. It would be interesting to determine why the AIC works well for the case of large p, and the theoretical clarification of this will be an area for our future work.

As mentioned above, previously there was not even a naive AIC for the Lasso. Thus, in this paper, we have considered only the most basic setting for our theorems, and they have been extended for a few settings. Using these extensions, we will be able to obtain, for example, the AIC for the generalized Lasso (Tibshirani and Taylor 2011) and the AIC for more general penalty terms. To check their performance in specific problems will be an important area for our future work. Beyond the selection of the regularization parameter for the Lasso, "inference after selection" problem is a current challenging topic and recently several methods are proposed (e.g., Lee et al. 2013 and Javanmard and Montanari 2014). To check the compatibility between the AIC in (13) and such methods will be also an important area for our future work.

Appendix

In this section, we give short proofs of the expressions used for our asymptotic theory.

Proof of (R1)

From (C2), the expectation of $\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n$ converges, and we denote this limit by $h(\boldsymbol{\beta})$. From (C1), the variance of $\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/\sqrt{n}$ is bounded, and so we can easily check that it converges in probability to $h(\boldsymbol{\beta})$. Because $\sum_{i=1}^{n}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n$ is convex and differentiable, $h(\boldsymbol{\beta})$ can also be shown to be convex and differentiable. See Theorems 10.8 and 25.7 in Rockafellar (1970).

Proof of (R2)

It holds from Theorem 25.7 in Rockafellar (1970) that

$$\lim_{n\to\infty}\frac{\partial}{\partial\boldsymbol{\beta}}\Big[\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\{g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^*) - g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}\Big] = \frac{\partial h(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}.$$

With direct calculation, the left-hand side reduces to $\lim_{n\to\infty}\sum_{i=1}^{n}\boldsymbol{X}_i^{\mathrm{T}}a'(\boldsymbol{X}_i\boldsymbol{\beta})/n$, and so $\partial h(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is shown to be the limit of $\sum_{i=1}^{n}\mathrm{E}\{-g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}/n$. From (C2), its convergence rate is $\mathrm{o}(1/\sqrt{n})$, and thus we obtain (R2).

Proof of (R4)

The asymptotic normality for the score function $\sum_{i=1}^{n}g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})$ can be obtained by applying the approach in Xie and Yang (2003), while their asymptotic normality is shown for standard generalized estimating equations estimators. For any given p-dimensional vector $\boldsymbol{\alpha}$ with $||\boldsymbol{\alpha}|| = 1$, let $\boldsymbol{y}_i^* = a''(\boldsymbol{X}_i\boldsymbol{\beta}^*)^{-1/2}\{\boldsymbol{y}_i - a'(\boldsymbol{X}_i\boldsymbol{\beta}^*)\}$ and $\omega_{ni} = \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2}[g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) - \mathrm{E}\{g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}] = \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2}\boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{y}_i - a'(\boldsymbol{X}_i\boldsymbol{\beta}^*)\}$. By the Cauchy-Schwarz inequality, it follows that

$$\omega_{ni}^2 \le \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2}\boldsymbol{X}_i^{\mathrm{T}}a''(\boldsymbol{X}_i\boldsymbol{\beta}^*)\boldsymbol{X}_i\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2}\boldsymbol{\alpha}\times\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^*.$$

Let $\gamma_{ni} = \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2}\boldsymbol{X}_i^{\mathrm{T}}a''(\boldsymbol{X}_i\boldsymbol{\beta}^*)\boldsymbol{X}_i\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2}\boldsymbol{\alpha}$. The minimum eigenvalue of $n\boldsymbol{J}_n(\boldsymbol{\beta}^*)$ goes to infinity because of (R3), and so $\max_i\gamma_{ni}\to 0$. We also note that $\sum_{i=1}^{n}\gamma_{ni} = 1$. Letting $\epsilon > 0$, we have

$$\mathrm{E}\{\omega_{ni}^2\mathrm{I}(|\omega_{ni}| > \epsilon)\} \le \mathrm{E}\Big\{\gamma_{ni}\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^*\mathrm{I}\Big(\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^* > \frac{\epsilon^2}{\gamma_{ni}}\Big)\Big\} \le \gamma_{ni}\mathrm{E}\Big\{(\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^*)^2\frac{\gamma_{ni}}{\epsilon^2}\Big\}.$$

In the first and second inequalities, we just use $\omega_{ni}^2 \le \gamma_{ni}\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^*$ and $\mathrm{I}(\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^* > \epsilon^2/\gamma_{ni}) \le \boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^*\gamma_{ni}/\epsilon^2$, respectively. For this model, the moments of $\boldsymbol{y}_i$ exist,

and so we have $\mathrm{E}\{(\boldsymbol{y}_i^{*\mathrm{T}}\boldsymbol{y}_i^*)^2\} < k$ for a constant $k > 0$. Thus, if we also use $\sum_{i=1}^n \gamma_{ni} = 1$, it holds that

$$\sum_{i=1}^{n} \mathrm{E}\{\omega_{ni}^2 \mathrm{I}(|\omega_{ni}| > \epsilon)\} \leq \frac{k}{\epsilon^2} \max_i \gamma_{ni} \to 0 \qquad (n \to \infty).$$

By the Lindeberg central limit theorem and the Cramér-Wold device, we have

$$\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \sum_{i=1}^{n} [g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}) - \mathrm{E}\{g'_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})\}] \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_p),$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. From this and (R3), we obtain (R4).

Proof of (5)

Let $\boldsymbol{s}_n^{(2)} = \sum_{i=1}^n g'^{(2)}_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta}^{**})/\sqrt{n} - \sqrt{n}\lambda \times \mathrm{sgn}(\boldsymbol{\beta}^{**(2)})$, and let $\tilde{\boldsymbol{s}}_n^{(2)} = \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})^{-1}\boldsymbol{s}_n^{(2)}$ and $\tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \sum_{j \in \mathcal{J}^{(1)}} \{\partial h(\boldsymbol{\beta})/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} u_j + \lambda|u_j|\} - \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{s}_n^{(2)} + \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})\boldsymbol{u}^{(2)}/2$. Note that $\tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ has a unique argmin $(\boldsymbol{0}, \tilde{\boldsymbol{s}}_n^{(2)})$. Defining $\Delta_n(\delta)$ as the supremum of $|v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})|$ over $\{|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{s}}_n^{(2)})| \leq \delta\}$, it follows from Lemma 2 in Hjort and Pollard (1993) that

$$\begin{aligned}
&\mathrm{P}\{|(n\hat{\boldsymbol{\beta}}_\lambda^{(1)}, \sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)}) - \tilde{\boldsymbol{s}}_n^{(2)})| \geq \delta\} \\
&\quad \leq \mathrm{P}\Big[\Delta_n(\delta) \geq \frac{1}{2}\Big\{ \inf_{|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{s}}_n^{(2)})| = \delta} \tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{v}_n(\boldsymbol{0}, \tilde{\boldsymbol{s}}_n^{(2)}) \Big\}\Big].
\end{aligned}$$

By a simple calculation, we have

$$\begin{aligned}
&\tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{v}_n(\boldsymbol{0}, \tilde{\boldsymbol{s}}_n^{(2)}) \\
&\quad = \sum_{j \in \mathcal{J}^{(1)}} \Big\{ \frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} u_j + \lambda|u_j| \Big\} + \frac{1}{2}(\boldsymbol{u}^{(2)} - \tilde{\boldsymbol{s}}_n^{(2)})^{\mathrm{T}} \boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})(\boldsymbol{u}^{(2)} - \tilde{\boldsymbol{s}}_n^{(2)}).
\end{aligned}$$

Let ρ_1 be the minimum value of $\lambda + h(\boldsymbol{\beta})/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}$ for $j \in \mathcal{J}^{(1)}$ and $\lambda - h(\boldsymbol{\beta})/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}}$ for $j \in \mathcal{J}^{(1)}$, and let ρ_2 be half the smallest eigenvalue of $\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})$. Note that ρ_1 and ρ_2 are positive because of (2) and the positive definiteness of $\boldsymbol{J}^{(22)}(\boldsymbol{\beta}^{**})$. Thus, we can obtain that

$$\mathrm{P}\{|\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{**(2)}) - \tilde{\boldsymbol{s}}_n^{(2)}| \geq \delta\} \leq \mathrm{P}\Big\{\Delta_n(\delta) \geq \frac{1}{2}\min(\rho_1\delta, \rho_2\delta^2)\Big\} \to 0.$$

Note that it holds $\Delta_n(\delta) \xrightarrow{\mathrm{p}} 0$ from the convexity lemma in Andersen and Gill (1982) or Pollard (1991).

Proof of (12)

Because $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_\lambda$ are consistent estimators of $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{**}$, respectively, it is sufficient to show that $\hat{\mathcal{J}}^{(2)}$ is a consistent estimator of $\mathcal{J}^{(2)}$.

Let us consider $\boldsymbol{\beta}$ satisfying $|\boldsymbol{\beta} - \boldsymbol{\beta}^{**}| \leq C/\sqrt{n}$ for some constant $C > 0$. Noting that $\partial\{\sum_{i=1}^n -g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})/n\}/\partial\beta_j = \partial h(\boldsymbol{\beta})/\partial\beta_j + \mathrm{o_P}(1)$ and $\partial^2\{\sum_{i=1}^n -g_{\boldsymbol{y}_i,\boldsymbol{X}_i}(\boldsymbol{\beta})/n\}/\partial\beta_j\partial\beta_k = \mathrm{O}(1)$, we have

$$\frac{\partial u_n(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial h_n(\boldsymbol{\beta})}{\partial\beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{**}} + \lambda \times \mathrm{sgn}(\beta_j) + \mathrm{o_P}(1)$$

uniformly in $\boldsymbol{\beta}$ by Taylor's expansion. Therefore, it follows from (2) that $\partial u_n(\boldsymbol{\beta})/\partial\beta_j < 0$ for $\beta_j < 0$ and $\partial u_n(\boldsymbol{\beta})/\partial\beta_j > 0$ for $\beta_j > 0$ with probability tending to 1 as $n \to \infty$ when $j \in \mathcal{J}^{(1)}$. Because $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{**} = \mathrm{O_P}(1/\sqrt{n})$, $\beta_j^* = 0$ for $j \in \mathcal{J}^{(1)}$, and $\beta_j^* \neq 0$ for $j \in \mathcal{J}^{(2)}$, we can conclude that $\mathrm{P}(\hat{\beta}_j \neq 0) = \mathrm{o}(1)$ for $j \in \mathcal{J}^{(1)}$ and $\mathrm{P}(\hat{\beta}_j = 0) = \mathrm{o}(1)$ for $j \in \mathcal{J}^{(2)}$.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive and helpful comments. This research was partially supported by a Grant-in-Aid for Scientific Research (23500353, 24700280, 15K15947, 16H06429, 16H06430) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

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 and F. Csaki, Budapest: Akademiai Kiado* 716–723. MR0483125

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. DOI: 10.1214/009053606000001587. MR2351101

Chételat, D., Lederer, J. and Salmon, J. (2014). Optimal two-step prediction in regression. *arXiv preprint arXiv:1410.5014*.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. With discussions and a rejoinder by the authors. DOI: 10.1198/016214503000000819. MR2041482

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. With discussion, and a rejoinder by the authors. DOI: 10.1214/009053604000000067. MR2060166

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. DOI: 10.1111/rssb.12001. MR3065478

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GEYER, C. J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript*.

HJORT, N. L. and POLLARD, D. (1993). Asymptotics for Minimisers of Convex Processes. *Unpublished manuscript*.

HURVICH, C. M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307. DOI: 10.1093/biomet/76.2.297. MR1016020

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. DOI: 10.1214/aos/1015957397. MR1805787

KONISHI, S. and KITAGAWA, G. (1996). Generalised information criteria in model selection. *Biometrika* **83** 875–890. DOI: 10.1093/biomet/83.4.875. MR1440051

KONISHI, S. and KITAGAWA, G. (2008). *Information criteria and statistical modeling. Springer Series in Statistics*. Springer, New York. DOI: 10.1007/978-0-387-71887-3. MR2367855

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics* **22** 79–86. MR0039968

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2013). Exact post-selection inference, with application to the lasso. *arXiv preprint arXiv:1311.6238*. MR3485948

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. DOI: 10.1214/13-AOS1175. MR3210970

LV, J. and LIU, J. S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 141–167. DOI: 10.1111/rssb.12023. MR3153937

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized linear models. Monographs on Statistics and Applied Probability*. Chapman & Hall, London. MR0727836

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. DOI: 10.1111/j.1467-9868.2010.00740.x. MR2758523

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. DOI: 10.1214/07-AOS582. MR2488351

POLLARD, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199. DOI: 10.1017/S0266466600004394. MR1128411

ROCKAFELLAR, R. T. (1970). *Convex analysis. Princeton Mathematical Series, No. 28*. Princeton University Press, Princeton, N.J. MR0274683

SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. and NOLAN, G. P.

(2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. DOI: 10.1111/1467-9868.00353. MR1979380

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–515. MR0415896

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods* **7** 13–26.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. DOI: 10.1093/biomet/ass043. MR2999166

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. DOI: 10.1214/11-AOS878. MR2850205

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. DOI: 10.1214/009053607000000929. MR2396809

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 671–683. DOI: 10.1111/j.1467-9868.2008.00693.x. MR2749913

Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.* **31** 310–347. DOI: 10.1214/aos/1046294467. MR1962509

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. DOI: 10.1093/biomet/asm018. MR2367824

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. With supplementary material available online. DOI: 10.1198/jasa.2009.tm08013. MR2656055

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. DOI: 10.1198/016214506000000735. MR2279469

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. DOI: 10.1214/009053607000000127. MR2363967