CrossMark

AIC for the non-concave penalized likelihood method

**Yuta Umezu[1] · Yusuke Shimizu[2] ·
Hiroki Masuda[3] · Yoshiyuki Ninomiya[4]**

Abstract Non-concave penalized maximum likelihood methods are widely used because they are more efficient than the Lasso. They include a tuning parameter which controls a penalty level, and several information criteria have been developed for selecting it. While these criteria assure the model selection consistency, they have a problem in that there are no appropriate rules for choosing one from the class of information criteria satisfying such a preferred asymptotic property. In this paper, we derive an information criterion based on the original definition of the AIC by considering minimization of the prediction error rather than model selection consistency. Concretely speaking, we derive a function of the score statistic that is asymptotically equivalent to the non-concave penalized maximum likelihood estimator and then provide an estimator of the Kullback–Leibler divergence between the true distribution and the estimated distribution based on the function, whose bias converges in mean to zero.

✉ Yoshiyuki Ninomiya
nino@imi.kyushu-u.ac.jp

[1] Department of Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

[2] Department of Mathematics, Josai University, 1-1 Keyakidai, Sakado, Saitama 350-0295, Japan

[3] Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

[4] Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan

Springer

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 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$, where $\lambda \ (> 0)$ is a tuning parameter controlling a penalty level. The Lasso can simultaneously perform estimation and variable selection by exploiting the non-differentiability of the penalty term at the origin. Concretely speaking, if $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, \hat{\beta}_{\lambda,2}, \ldots, \hat{\beta}_{\lambda,p})^{\mathrm{T}}$ is the estimator based on the Lasso, several of its components will shrink to exactly 0 when λ is not close to 0. However, a parameter estimation based on the Lasso is not necessarily efficient, because the Lasso shrinks the estimator to the zero vector too much. To avoid such a problem, it has been proposed to use a penalty term that does not shrink the estimator with a large value. Typical examples of such regularization methods are the Bridge (Frank and Friedman 1993), the smoothly clipped absolute deviation (SCAD; Fan and Li 2001), and the minimax concave penalty (MCP; Zhang 2010), whereas the Bridge uses an ℓ_q penalty term $(0 < q < 1)$, SCAD and MCP use penalty terms that can be approximated by an ℓ_1 penalty term in the neighborhood of the origin, which we call an ℓ_1 type. Although it is difficult to obtain estimates of them as their penalties are non-convex, there are several algorithms, such as the coordinate descent method and the gradient descent method, that assure convergence to a local optimal solution.

On the other hand, in the above regularization methods, we have to choose a proper value for the tuning parameter λ, and this is an important task for appropriate model selection. One of the simplest ways of selecting λ is to use cross-validation (CV; Stone 1974). While the stability selection method (Meinshausen and Bühlmann 2010) based on subsampling in order to avoid problems caused by selecting a model based on only one value of λ would be nice, it carries with it a considerable computational cost as in CV. Recently, information criteria without such a problem have been developed (Yuan and Lin 2007; Wang et al. 2007, 2009; Zhang et al. 2010; Fan and Tang 2013). Here, by letting $\ell(\cdot)$ be the log-likelihood function and $\hat{\boldsymbol{\beta}}_\lambda$ be the estimator of $\boldsymbol{\beta}$ obtained by the above regularization methods, their information criteria take the form $-2\ell(\hat{\boldsymbol{\beta}}_\lambda) + \kappa_n \|\hat{\boldsymbol{\beta}}_\lambda\|_0$. Accordingly, model selection consistency is at least assured for some sequence κ_n that depends on at least the sample size n. For example, the information criterion with $\kappa_n = \log n$ is proposed as the BIC. This approach includes the results for the case in which the dimension of the parameter vector p goes to infinity, and hence, it is considered to be significant. However, there is an arbitrariness in the choice of tuning parameter by this approach. That is, there is a class of κ_n assuring a preferred asymptotic property such as model selection consistency, but there are no appropriate rules for choosing one from the class. For example, since the BIC described above is not derived from the Bayes factor, there is no reason to use $\kappa_n = \log n$ instead of $\kappa_n = 2 \log n$. This is a severe problem because data analysts can choose κ_n arbitrarily and do model selection as they want.

Information criteria without such an arbitrariness problem have been proposed by Efron et al. (2004) or Zou et al. (2007) for Gaussian linear regression and by Ninomiya and Kawano (2016) for generalized linear regression. Concretely speaking, on the basis of the original definition of the C_p or AIC, they derive an unbiased estimator of the mean squared error or an asymptotically unbiased estimator of a Kullback–Leibler divergence. However, these criteria are basically only for the Lasso. In addition, the asymptotic setting used in Ninomiya and Kawano (2016) does not assure even estimation consistency.

Our goal in this paper is to derive an information criterion based on the original definition of AIC in an asymptotic setting that assures estimation consistency for regularization methods using non-concave penalties including the Bridge, SCAD, and MCP. To achieve it, the results presented in Hjort and Pollard (1993) are slightly extended to derive an asymptotic property for the estimator. Then, for the Kullback–Leibler divergence, we construct an asymptotically unbiased estimator by evaluating the asymptotic bias between the divergence and the log-likelihood into which the estimator is plugged. Moreover, we verify that this evaluation is the asymptotic bias in the strict sense; that is, the bias converges in mean to the evaluation. This sort of verification has usually been ignored in the literature (see, e.g., Konishi and Kitagawa 2008).

The rest of the paper is organized as follows. Section 2 introduces the generalized linear model and the regularization method, and it describes some of the assumptions on our asymptotic theory. In Sect. 3, we discuss the asymptotic property of the estimator obtained from the regularization method. In Sect. 4, we use it to evaluate the asymptotic bias, which is needed to derive the AIC. The derived AIC is given in (27). In Sect. 5, we discuss the moment convergence of the estimator to show that the bias converges in mean to our evaluation. Section 6 presents the results of simulation studies showing the validity of the proposed information criterion for several models, and Sect. 7 gives concluding remarks and mentions future work. The proofs are relegated to the appendixes.

2 Setting and assumptions for asymptotics

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

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

with respect to a σ-finite measure. We assume that Θ is the natural parameter space; that is, $\boldsymbol{\theta}$ in Θ satisfies $0 < \int \exp\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\theta} + b(\boldsymbol{y})\}d\boldsymbol{y} < \infty$. Accordingly, 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, $\mathrm{E}[\boldsymbol{y}] = a'(\boldsymbol{\theta})$ and $\mathrm{V}[\boldsymbol{y}] = a''(\boldsymbol{\theta})$. For a function $c(\boldsymbol{\eta})$, we 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 also assume that $\mathrm{V}[\boldsymbol{y}] = a''(\boldsymbol{\theta})$ is positive definite, and hence, $-\log f(\boldsymbol{y}; \boldsymbol{\theta})$ is a strictly convex function with respect to $\boldsymbol{\theta}$.

Let $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ be the ith set of responses and regressors $(i = 1, 2, \ldots, n)$; we assume that $\boldsymbol{y}_i$ are independent r-dimensional random vectors and $\boldsymbol{X}_i$ in $\mathcal{X}$ $(\subset \mathbb{R}^{r \times p})$ are $(r \times p)$-matrices of known constants. We will consider generalized linear models with natural link functions for such data (see McCullagh and Nelder 1989); that is, we will consider a class of density functions $\{f(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta}); \ \boldsymbol{\beta} \in \mathcal{B}\}$ for $\boldsymbol{y}_i$; thus, the log-likelihood function of $\boldsymbol{y}_i$ is given by

$$g_i(\boldsymbol{\beta}) = \boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{\beta} - a(\boldsymbol{X}_i \boldsymbol{\beta}) + b(\boldsymbol{y}_i),$$

where $\boldsymbol{\beta}$ is a p-dimensional coefficient vector and $\mathcal{B}$ $(\subset \mathbb{R}^p)$ is an open convex set. To develop an asymptotic theory for this model, we assume two conditions about the behavior of $\{\boldsymbol{X}_i\}$, as follows:

(C1) $\mathcal{X}$ is a compact set with $\boldsymbol{X}\boldsymbol{\beta} \in \Theta^{\mathrm{int}}$ for all $\boldsymbol{X}$ $(\in \mathcal{X})$ and $\boldsymbol{\beta}$ $(\in \mathcal{B})$.
(C2) The sequence $\{\boldsymbol{X}_i\}$ is ergodic with an invariance measure μ on $\mathcal{X}$. For example, $n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i \boldsymbol{\beta}) \boldsymbol{X}_i$ converges to $\boldsymbol{J}(\boldsymbol{\beta}) \equiv \int_{\mathcal{X}} \boldsymbol{X}^{\mathrm{T}} a''(\boldsymbol{X}\boldsymbol{\beta}) \boldsymbol{X} \mu(\mathrm{d}\boldsymbol{X})$.
(C3) $\boldsymbol{J}(\boldsymbol{\beta})$ is a positive definite matrix.

In the above setting, we can prove the following lemma.

Lemma 1 *Let $\boldsymbol{\beta}^*$ be the true value of $\boldsymbol{\beta}$. Then, under conditions (C1), (C2), and (C3), we obtain the following:*

(R1) *There exists a convex and differentiable function $h(\boldsymbol{\beta})$ such that $n^{-1} \sum_{i=1}^{n} \{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} \xrightarrow{\mathrm{p}} h(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$.*
(R2) *$\boldsymbol{J}_n(\boldsymbol{\beta}) \equiv -n^{-1} \sum_{i=1}^{n} g_i''(\boldsymbol{\beta})$ converges to $\boldsymbol{J}(\boldsymbol{\beta})$.*
(R3) *$\boldsymbol{s}_n \equiv n^{-1/2} \sum_{i=1}^{n} g_i'(\boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} \boldsymbol{s} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\beta}^*))$.*

See Ninomiya and Kawano (2016) for the proof. The condition (C2) in Ninomiya and Kawano (2016) is stronger than the above (C2) and (C3) because the condition requires the rate $\mathrm{o}(n^{-1/2})$ in the convergence. But the proof of the above lemma does not depend on the convergence rate, and we can obtain the proof just by reducing the statements in Ninomiya and Kawano (2016).

Note that we can explicitly write

$$h(\boldsymbol{\beta}) = \int_{\mathcal{X}} [a'(\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - \{a(\boldsymbol{X}\boldsymbol{\beta}^*) - a(\boldsymbol{X}\boldsymbol{\beta})\}] \mu(\mathrm{d}\boldsymbol{X}) \qquad (1)$$

since we assume (C2), and hence, we can prove its convexity and differentiability without using the techniques of convex analysis (Rockafellar 1970).

Let us consider a non-concave penalized maximum likelihood estimator,

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\mathrm{argmin}} \left\{ -\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) + n^{1/2} \sum_{j=1}^{p} p_\lambda(\beta_j) \right\}, \qquad (2)$$

where λ (> 0) is a tuning parameter and $p_\lambda(\beta_j)$ is a penalty term with respect to β_j, which is not necessarily convex. Letting $q \in (0, 1]$, we assume that $p_\lambda(\cdot)$ satisfies the following conditions; hereafter, we call it an ℓ_q type:

(C4) $p_\lambda(\beta)$ is differentiable and $p_\lambda'(\beta)$ is continuous except for at the origin. In addition, $p_\lambda(\beta)$ is symmetric with respect to $\beta = 0$ and monotone non-decreasing with respect to $|\beta|$.

(C5) $\lim_{\beta \to 0} p_\lambda(\beta)/|\beta|^q = \lambda$.

Such penalty terms for the Bridge, the SCAD, and the MCP are

$$p_\lambda^{\text{Bridge}}(\beta) = \lambda|\beta|^q,$$
$$p_\lambda^{\text{SCAD}}(\beta) = \lambda|\beta|1_{\{|\beta|\le(r+1)\lambda\}} - (|\beta| - \lambda)^2/(2r)1_{\{\lambda<|\beta|\le(r+1)\lambda\}}$$
$$+ \lambda^2(1 + r/2)1_{\{|\beta|>(r+1)\lambda\}},$$

and

$$p_\lambda^{\text{MCP}}(\beta) = r\lambda^2/2 - (r\lambda - |\beta|)^2/(2r)1_{\{|\beta|\le r\lambda\}},$$

where $0 < q \le 1$ and $r > 1$. The Bridge penalty is the Lasso penalty itself when $q = 1$, and it has the property that the derivative at the origin diverges when $0 < q < 1$. For the SCAD and MCP penalties, condition (C5) on the behavior in the neighborhood of the origin is satisfied by setting $q = 1$, just like in the Lasso penalty. Thus, it is easy to imagine that a lot of penalties satisfy these conditions. Note that by using such penalties, several components of $\hat{\boldsymbol{\beta}}_\lambda$ tend to exactly 0 because of the non-differentiability at the origin. Also, note that $p_\lambda(\cdot)$ is assumed not to depend on the subscript j of the parameter for simplicity; this is not essential. While Ninomiya and Kawano (2016) put n on the penalty term, we put $n^{1/2}$ on it in this study. From this, we can prove estimation consistency. Moreover, we can prove weak convergence of $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$, although the asymptotic distribution is not normal in general.

3 Asymptotic behavior

3.1 Preparations

Although the objective function in (2) is no longer convex because of the non-convexity of $p_\lambda(\cdot)$, the consistency of $\hat{\boldsymbol{\beta}}_\lambda$ can be derived by using a similar argument to the one in Knight and Fu (2000). First, the following lemma holds.

Lemma 2 $\hat{\boldsymbol{\beta}}_\lambda$ *is a consistent estimator of* $\boldsymbol{\beta}^*$, *that is,* $\hat{\boldsymbol{\beta}}_\lambda \xrightarrow{\text{p}} \boldsymbol{\beta}^*$ *under conditions* (*C1*)–(*C5*).

This lemma is proved through uniform convergence of the random function,

$$\mu_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} - \frac{1}{n^{1/2}}\sum_{j=1}^p\{p_\lambda(\beta_j^*) - p_\lambda(\beta_j)\}. \tag{3}$$

The details are given in Sect. A.1. Hereafter, we will denote $\boldsymbol{J}(\boldsymbol{\beta}^*)$ by $\boldsymbol{J}$ so long as there is no confusion. In addition, we denote $\{j;\ \beta_j^* = 0\}$ and $\{j;\ \beta_j^* \ne 0\}$ by $\mathcal{J}^{(1)}$ and

$\mathcal{J}^{(2)}$, respectively. Moreover, the vector $(u_j)_{j\in\mathcal{J}^{(k)}}$ and the matrix $(J_{ij})_{i\in\mathcal{J}^{(k)},j\in\mathcal{J}^{(l)}}$ will be denoted by $\boldsymbol{u}^{(k)}$ and $\boldsymbol{J}^{(kl)}$, respectively, and we will sometimes express, for example, $\boldsymbol{u}$ as $(\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)})$.

To develop the asymptotic property of the penalized maximum likelihood estimator in (2), which will be used to derive an information criterion, we need to make a small generalization of the result in Hjort and Pollard (1993), as follows:

Lemma 3 *Suppose that $\eta_n(\boldsymbol{u})$ is a strictly convex random function that is approximated by $\tilde{\eta}_n(\boldsymbol{u})$. Let $\boldsymbol{u}^\dagger$ be a subvector of $\boldsymbol{u}$, and let $\phi(\boldsymbol{u})$ and $\psi(\boldsymbol{u}^\dagger)$ be continuous functions such that $\phi_n(\boldsymbol{u})$ and $\psi_n(\boldsymbol{u}^\dagger)$ converge to $\phi(\boldsymbol{u})$ and $\psi(\boldsymbol{u}^\dagger)$ uniformly over $\boldsymbol{u}$ and $\boldsymbol{u}^\dagger$ in any compact set, respectively, and assume that $\phi(\boldsymbol{u})$ is convex and $\psi(\boldsymbol{0}) = 0$. In addition, for*

$$v_n(\boldsymbol{u}) = \eta_n(\boldsymbol{u}) + \phi_n(\boldsymbol{u}) + \psi_n\left(\boldsymbol{u}^\dagger\right) \quad and \quad \tilde{v}_n(\boldsymbol{u}) = \tilde{\eta}_n(\boldsymbol{u}) + \phi(\boldsymbol{u}) + \psi\left(\boldsymbol{u}^\dagger\right),$$

let $\boldsymbol{u}_n$ and $\tilde{\boldsymbol{u}}_n$ be the argmin of $v_n(\boldsymbol{u})$ and $\tilde{v}_n(\boldsymbol{u})$, respectively, and assume that $\tilde{\boldsymbol{u}}_n$ is unique and $\tilde{\boldsymbol{u}}_n^\dagger = \boldsymbol{0}$. Then, for any $\varepsilon\ (>0)$, $\delta\ (>0)$, and $\xi\ (>\delta)$, there exists $\gamma\ (>0)$ such that

$$\begin{aligned}
&P\left(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \delta\right)\\
&\leq P\left(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)\right) + P\left(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \xi\right) + P\left(|\boldsymbol{u}_n^\dagger| > \gamma\right),
\end{aligned} \quad (4)$$

where

$$\Delta_n(\delta) = \sup_{|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n|\leq\delta} |v_n(\boldsymbol{u}) - \tilde{v}_n(\boldsymbol{u})| \quad and \quad \Upsilon_n(\delta) = \inf_{|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n|=\delta} \tilde{v}_n(\boldsymbol{u}) - \tilde{v}_n(\tilde{\boldsymbol{u}}_n). \quad (5)$$

Hjort and Pollard (1993) derived an inequality $P(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \delta) \leq P(2\Delta_n(\delta) \geq \Upsilon_n(\delta))$; they assumed that $v_n(\boldsymbol{u})$ is convex. Although $\phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ is non-convex (hence $v_n(\boldsymbol{u})$ is too), we will use the fact that $\phi_n(\boldsymbol{u})+\psi_n(\boldsymbol{u}^\dagger)$ converge to $\phi(\boldsymbol{u})+\psi(\boldsymbol{u}^\dagger)$ over $\mathcal{U} \equiv \{\boldsymbol{u};\ |\boldsymbol{u}^\dagger| \leq \gamma,\ \delta \leq |\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| \leq \xi\}$. In fact, if n is sufficiently large, the inequality satisfied by the convex function is approximately satisfied for $\phi_n(\boldsymbol{u})$; that is, we have

$$(1 - \delta/l)\,\phi_n(\tilde{\boldsymbol{u}}_n) + (\delta/l)\phi_n(\boldsymbol{u}) - \phi_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) > -\varepsilon/2 \quad (6)$$

in $\mathcal{U}$. Here, $\boldsymbol{w}$ is a unit vector such that $\boldsymbol{u} = \tilde{\boldsymbol{u}}_n + l\boldsymbol{w}$, and l is in $[\delta, \xi]$, since $\delta \leq |\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| \leq \xi$. Moreover, if γ is sufficiently small and n is sufficiently large, since $\psi(\tilde{\boldsymbol{u}}_n^\dagger) = 0$, we have

$$(1 - \delta/l)\,\psi_n\left(\tilde{\boldsymbol{u}}_n^\dagger\right) + (\delta/l)\psi_n\left(\boldsymbol{u}^\dagger\right) - \psi_n\left(\tilde{\boldsymbol{u}}_n^\dagger + \delta\boldsymbol{w}^\dagger\right) > -\varepsilon/2 \quad (7)$$

in $\mathcal{U}$. Hence, we can show that

$$P\left(|\boldsymbol{u}_n^\dagger| \leq \gamma,\ \delta \leq |\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \leq \xi\right) \leq P\left(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)\right) \quad (8)$$

in the same way as in Hjort and Pollard (1993), from which we obtain the above lemma. See Sect. A.2 for the details.

3.2 Limiting distribution

We use Lemma 3 to derive the asymptotic property of the penalized maximum likelihood estimator in (2). Because the asymptotic property depends on the value of q, we will develop our argument by setting $0 < q < 1$. Furthermore, we will use $\tilde{q} = 1/(2q)$ for the sake of simplicity.

Let us define a strictly convex random function,

$$\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \sum_{i=1}^{n} \left\{ g_i(\boldsymbol{\beta}^{*(1)}, \boldsymbol{\beta}^{*(2)}) - g_i\left(\frac{\boldsymbol{u}^{(1)}}{n^{\tilde{q}}}, \frac{\boldsymbol{u}^{(2)}}{n^{1/2}} + \boldsymbol{\beta}^{*(2)}\right) \right\} \tag{9}$$

and

$$\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = -\boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}_n^{(2)} + \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)} \boldsymbol{u}^{(2)}/2, \tag{10}$$

where $\boldsymbol{s}_n^{(2)} = n^{-1/2} \sum_{i=1}^{n} g_i'^{(2)}(\boldsymbol{\beta}^*)$. By making a Taylor expansion around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\boldsymbol{0}, \boldsymbol{0})$, $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ can be expressed as

$$- \sum_{i=1}^{n} \left\{ \frac{1}{n^{\tilde{q}}} \boldsymbol{u}^{(1)\mathrm{T}} g_i'^{(1)}(\boldsymbol{\beta}^*) + \frac{1}{n^{1/2}} \boldsymbol{u}^{(2)\mathrm{T}} g_i'^{(2)}(\boldsymbol{\beta}^*) \right\}$$

$$- \sum_{i=1}^{n} \left\{ \frac{1}{2n^{2\tilde{q}}} \boldsymbol{u}^{(1)\mathrm{T}} g_i''^{(11)}(\boldsymbol{\beta}^*) \boldsymbol{u}^{(1)} + \frac{1}{n^{\tilde{q}+1/2}} \boldsymbol{u}^{(1)\mathrm{T}} g_i''^{(12)}(\boldsymbol{\beta}^*) \boldsymbol{u}^{(2)} \right.$$

$$\left. + \frac{1}{2n} \boldsymbol{u}^{(2)\mathrm{T}} g_i''^{(22)}(\boldsymbol{\beta}^*) \boldsymbol{u}^{(2)} \right\}$$

plus $o_p(1)$. Note that the term $-n^{-1} \sum_{i=1}^{n} \boldsymbol{u}^{(2)\mathrm{T}} g_i''^{(22)}(\boldsymbol{\beta}^*) \boldsymbol{u}^{(2)}$ converges to $\boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J} \boldsymbol{u}^{(2)}$ from (R2), and the terms including $\boldsymbol{u}^{(1)}$ reduce to $o_p(1)$. Accordingly, we see that $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ is asymptotically equivalent to $\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$. Next, letting $\boldsymbol{u}^{\dagger}$ be $\boldsymbol{u}^{(1)}$ and letting

$$\phi_n(\boldsymbol{u}) = n^{1/2} \sum_{j \in \mathcal{J}^{(2)}} \left\{ p_\lambda\left(\frac{u_j}{n^{1/2}} + \beta_j^*\right) - p_\lambda(\beta_j^*) \right\} \tag{11}$$

and

$$\psi_n\left(\boldsymbol{u}^{\dagger}\right) = n^{1/2} \sum_{j \in \mathcal{J}^{(1)}} p_\lambda\left(\frac{u_j}{n^{\tilde{q}}}\right), \tag{12}$$

we can see from (C4) and (C5) that $\phi_n(\boldsymbol{u})$ and $\psi_n(\boldsymbol{u}^\dagger)$ uniformly converge to a function,

$$\phi(\boldsymbol{u}) = \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{p}_\lambda^{\prime(2)} \quad \text{and} \quad \psi\left(\boldsymbol{u}^\dagger\right) = \lambda \|\boldsymbol{u}^{(1)}\|_q^q, \tag{13}$$

over $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ in a compact set, respectively, where $\boldsymbol{p}_\lambda^{\prime(2)} = (p_\lambda'(\beta_j^*))_{j \in \mathcal{J}^{(2)}}$. In addition, letting $v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ and $\tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger)$, we see that the argmins of $v_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ and $\tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ are given by

$$\left(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)}\right) = \left(n^{\tilde{q}} \hat{\boldsymbol{\beta}}_\lambda^{(1)}, n^{1/2} \left(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}\right)\right)$$

and

$$\left(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}\right) = \left(\boldsymbol{0}, \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)}\right)\right).$$

Note that $\psi(\boldsymbol{u}^\dagger)$ is not convex but satisfies that $\psi(\tilde{\boldsymbol{u}}_n^{(1)}) = 0$. Using Lemma 3 together with the above preliminaries, we find that, for any $\varepsilon\ (> 0)$, $\delta\ (> 0)$ and $\xi\ (> \delta)$, there exists $\gamma\ (> 0)$ such that

$$P\left(\left|\left(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)}\right)\right| \geq \delta\right)$$
$$\leq P\left(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)\right) + P\left(\left|\left(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)}\right)\right| \geq \xi\right) + P\left(\left|\boldsymbol{u}_n^{(1)}\right| > \gamma\right), \tag{14}$$

where $\Delta_n(\delta)$ and $\Upsilon_n(\delta)$ are the functions defined in (5). The triangle inequality, the convexity of $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}_n^{(2)}$ and the uniform convergence of $\phi_n(\boldsymbol{u})$ and $\psi_n(\boldsymbol{u}^\dagger)$ imply

$$\Delta_n(\delta)$$
$$\leq \sup_{|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})| \leq \delta} \left|\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{s}_n^{(2)} - \boldsymbol{u}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)} \boldsymbol{u}^{(2)}/2\right|$$
$$+ \sup_{|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})| \leq \delta} |\phi_n(\boldsymbol{u}) - \phi(\boldsymbol{u})| + \sup_{|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})| \leq \delta} \left|\psi_n\left(\boldsymbol{u}^\dagger\right) - \psi\left(\boldsymbol{u}^\dagger\right)\right|$$
$$\xrightarrow{\mathrm{p}} 0. \tag{15}$$

Let $\rho\ (> 0)$ be half the smallest eigenvalue of $\boldsymbol{J}^{(22)}$. Then, a simple calculation gives

$$\Upsilon_n(\delta) = \inf_{|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})| = \delta} \{\lambda \|\boldsymbol{u}^{(1)}\|_q^q + (\boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})^\mathrm{T} \boldsymbol{J}^{(22)} (\boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})/2\}$$
$$\geq \min\{\lambda \delta^q, \rho \delta^2\}. \tag{16}$$

From (15) and (16), by considering a sufficiently small ε and a sufficiently large n, the first term on the right-hand side in (14) can be made arbitrarily small. In addition, we can generalize the result in Radchenko (2005) with respect to the model and the penalty term; thus, for any γ (> 0), we have

$$\mathrm{P}\left(\left|\boldsymbol{u}_n^{(1)}\right| \leq \gamma\right) \to 1 \quad \text{and} \quad |\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| = \mathrm{O}_\mathrm{p}(1). \tag{17}$$

See Sect. A.3 for the proof of (17). From this, by considering a sufficiently large ξ and a sufficiently large n, the second and third terms on the right-hand side in (14) can be made arbitrarily small. Thus, we conclude that

$$\boldsymbol{u}_n^{(1)} = \mathrm{o}_\mathrm{p}(1) \quad \text{and} \quad \boldsymbol{u}_n^{(2)} = \tilde{\boldsymbol{u}}_n^{(2)} + \mathrm{o}_\mathrm{p}(1).$$

Theorem 1 *Let* $\boldsymbol{p}_\lambda'^{(2)} = (p_\lambda'(\beta_j^*))_{j \in \mathcal{J}^{(2)}}$, $\boldsymbol{J}^{(1|2)} = \boldsymbol{J}^{(11)} - \boldsymbol{J}^{(12)} \boldsymbol{J}^{(22)-1} \boldsymbol{J}^{(21)}$, $\boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) = \boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)} \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})$ *and*

$$\hat{\boldsymbol{u}}_n^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\operatorname{argmin}}\{\boldsymbol{u}^{(1)\mathrm{T}} \boldsymbol{J}^{(1|2)} \boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}} \boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda \|\boldsymbol{u}^{(1)}\|_1\}. \tag{18}$$

Under conditions (C1)–(C5), we have

$$n^{1/(2q)}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathrm{o}_\mathrm{p}(1) \quad \text{and} \quad n^{1/2}\left(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}\right) = \boldsymbol{J}^{(22)-1}\left(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right) + \mathrm{o}_\mathrm{p}(1)$$

when $0 < q < 1$, *and we have*

$$n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \hat{\boldsymbol{u}}_n^{(1)} + \mathrm{o}_\mathrm{p}(1) \tag{19}$$

and

$$n^{1/2}\left(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}\right) = -\boldsymbol{J}^{(22)-1} \boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)} + \boldsymbol{J}^{(22)-1}\left(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right) + \mathrm{o}_\mathrm{p}(1) \tag{20}$$

when $q = 1$.

We can obtain the result for the case of $q = 1$ in almost the same way as in the case of $0 < q < 1$ (see Sect. A.4 for details). From Theorem 1, the estimator $\hat{\boldsymbol{\beta}}_\lambda$ in (2) is shown to converge in distribution to some function of a Gaussian distributed random variable. When $0 < q < 1$, we immediately see that it is 0 or the Gaussian distributed random variable itself, and this simple fact is useful for deriving an information criterion explicitly and reducing the computational cost of model selection. On the other hand, when $q = 1$, we can prove weak convergence, since the convex objective function in (18) converges uniformly from the convexity lemma in Hjort and Pollard (1993).

Corollary 1 *Let* $\boldsymbol{s}$ *be a Gaussian distributed random variable with mean* $\boldsymbol{0}$ *and covariance matrix* $\boldsymbol{J}$, *and*

$$\hat{\boldsymbol{u}}^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\arg\min} \left\{ \boldsymbol{u}^{(1)\mathrm{T}} \boldsymbol{J}^{(1|2)} \boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}} \boldsymbol{\tau}_{\lambda}(\boldsymbol{s}) + \lambda \|\boldsymbol{u}^{(1)}\|_1 \right\}. \tag{21}$$

Then, under the same conditions as in Theorem 1, *we have*

$$n^{1/(2q)} \hat{\boldsymbol{\beta}}_{\lambda}^{(1)} \xrightarrow{\mathrm{d}} \boldsymbol{0} \quad and \quad n^{1/2} \left(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)} \right) \xrightarrow{\mathrm{d}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_{\lambda}^{\prime(2)} \right)$$

when $0 < q < 1$, *and we have*

$$n^{1/2} \hat{\boldsymbol{\beta}}_{\lambda}^{(1)} \xrightarrow{\mathrm{d}} \hat{\boldsymbol{u}}^{(1)}$$

and

$$n^{1/2} \left(\hat{\boldsymbol{\beta}}_{\lambda}^{(2)} - \boldsymbol{\beta}^{*(2)} \right) \xrightarrow{\mathrm{d}} -\boldsymbol{J}^{(22)-1} \boldsymbol{J}^{(21)} \hat{\boldsymbol{u}}^{(1)} + \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_{\lambda}^{\prime(2)} \right)$$

when $q = 1$.

In the case of $q = 1$, we still need to solve the minimization problem in (21) for evaluating the AIC, but this is easy because the objective function is convex with respect to $\boldsymbol{u}^{(1)}$, so we can use existing convex optimization techniques. It is known that the proximal gradient method (Rockafellar 1976; Beck and Teboulle 2009) is effective for solving such a minimization problem when the objective function is the sum of a differentiable function and a non-differentiable function. We will use, however, the coordinate descent method (Mazumder et al. 2011) because the objective function can be minimized explicitly for each variable. Actually, when we fix all the elements of $\hat{\boldsymbol{u}}$ except for the jth one, $\hat{u}_j^{(1)}$ is given by

$$\hat{u}_j^{(1)} = \frac{1}{J_{jj}^{(1|2)}} \mathrm{sgn}\left(\tau_j - \sum_{k \neq j} J_{jk}^{(1|2)} \hat{u}_k^{(1)} \right) \max \left\{ \left| \tau_j - \sum_{k \neq j} J_{jk}^{(1|2)} \hat{u}_k^{(1)} \right| - \lambda, 0 \right\}.$$

Then, for the $(t + 1)$th step in the algorithm, we update $u_j^{(1)[t+1]}$ by

$$\underset{u}{\arg\min} \; H\left(u_1^{(1)[t+1]}, u_2^{(1)[t+1]}, \ldots, u_{j-1}^{(1)[t+1]}, u, u_{j+1}^{(1)[t]}, u_{j+2}^{(1)[t]}, \ldots, u_{|\mathcal{J}^{(1)}|}^{(1)[t]} \right)$$

for $j = 1, 2, \ldots, |\mathcal{J}^{(1)}|$, where $H(\boldsymbol{u}^{(1)})$ is the objective function in (21). We repeat this update until $|\boldsymbol{u}^{(1)[t+1]} - \boldsymbol{u}^{(1)[t]}|$ converges.

4 Information criterion

From the perspective of prediction, model selection using the AIC aims to minimize twice the Kullback–Leibler divergence (Kullback and Leibler 1951) between the true distribution and the estimated distribution. Let $(\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_n)$ be a copy of $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$; that is, $(\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_n)$ has the same distribution as

$(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ and is independent of $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$. Then, twice the divergence is written as

$$2\tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\boldsymbol{\beta}^*)\right] - 2\tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right],$$

where $\tilde{g}_i(\boldsymbol{\beta})$ and $\tilde{\mathrm{E}}$ denote a log-likelihood function based on $\tilde{\mathbf{y}}_i$, that is, $\log f(\tilde{\mathbf{y}}_i; X_i\boldsymbol{\beta})$, and the expectation with respect to only $(\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \ldots, \tilde{\mathbf{y}}_n)$, respectively. Because the first term is a constant, i.e., it does not depend on the model selection, we only need to consider the second term, and then the AIC is defined as an asymptotically unbiased estimator for it (Akaike 1973). A simple estimator of the second term in our setting is $-2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda)$, but it underestimates the second term. Consequently, we will minimize the bias correction,

$$-2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{E}\left[\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right]\right], \tag{22}$$

in AIC-type information criteria (see Konishi and Kitagawa 2008). Because the expectation in (22), i.e., the bias term, depends on the true distribution, it cannot be explicitly given in general; thus, we will evaluate it asymptotically in the same way as was done for the AIC.

For the Lasso, Efron et al. (2004) and Zou et al. (2007) developed the C_p-type information criterion as an unbiased estimator of the prediction squared error in a Gaussian linear regression setting, in other words, a finite correction of the AIC (Sugiura 1978) in a Gaussian linear setting with a known variance. For the Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, \ldots, \hat{\beta}_{\lambda,p})$, it can be expressed as

$$\sum_{i=1}^{n} \left\{ \left(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}_\lambda\right)^{\mathrm{T}} \mathrm{V}[\mathbf{y}_i]^{-1} \left(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}_\lambda\right) + \log|2\pi \mathrm{V}[\mathbf{y}_i]| \right\} + 2\left|\left\{j;\ \hat{\beta}_{\lambda,j} \neq 0\right\}\right|,$$

where the index set $\{j;\ \hat{\beta}_{\lambda,j} \neq 0\}$ is called an active set. Unfortunately, since Stein's unbiased risk estimation theory (Stein 1981) was used for deriving this criterion, it was difficult to extend this result to other models. In that situation, Ninomiya and Kawano (2016) relied on statistical asymptotic theory and extended the result to generalized linear models based on the asymptotic distribution of the Lasso estimator. The Lasso estimator in their paper is defined by

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) + n\lambda\|\boldsymbol{\beta}\|_1 \right\},$$

but, as was mentioned in the previous section, estimation consistency is not assured because the order of the penalty term is $\mathrm{O}(n)$. In this study, we derive an information criterion in a setting that estimation consistency holds as in Lemma 2 for not only the Lasso, but also the non-concave penalized likelihood method.

The bias term in (22) can be rewritten as the expectation of

$$\sum_{i=1}^{n} \left\{ g_i(\hat{\boldsymbol{\beta}}_\lambda) - g_i(\boldsymbol{\beta}^*) \right\} - \sum_{i=1}^{n} \left\{ \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{g}_i(\boldsymbol{\beta}^*) \right\}, \tag{23}$$

so we can derive an AIC by evaluating $E[z^{\text{limit}}]$, where z^{limit} is the limit to which (23) converges in distribution. We call $E[z^{\text{limit}}]$ an asymptotic bias. Here, we will develop an argument by setting $0 < q < 1$.

Using Taylor's theorem, the first term in (23) can be expressed as

$$\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \right)^{\mathrm{T}} \sum_{i=1}^{n} g_i'(\boldsymbol{\beta}^*) + \left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \right)^{\mathrm{T}} \sum_{i=1}^{n} g_i'' \left(\boldsymbol{\beta}^\dagger \right) \left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \right) \Big/ 2, \tag{24}$$

where $\boldsymbol{\beta}^\dagger$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. Note that $-n^{-1} \sum_{i=1}^{n} g_i''(\boldsymbol{\beta}^\dagger)$ converges in probability to $\boldsymbol{J}$ from (R2) and Lemma 2. Now we apply Theorem 1. First, the terms including $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ reduce to $o_p(1)$ because $n^{1/(2q)} \hat{\boldsymbol{\beta}}_\lambda^{(1)} = o_p(1)$. Moreover, $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^*)$ is asymptotically equivalent to $\boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})$. Thus, (24) can be expressed as

$$\boldsymbol{s}_n^{(2)\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) - \left(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) \Big/ 2 + o_p(1),$$

and we see that this converges in distribution to

$$\boldsymbol{s}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) - \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) \Big/ 2$$

from (R3). Similarly, the second term in (23) can be expressed as using Taylor's theorem

$$\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \right)^{\mathrm{T}} \sum_{i=1}^{n} \tilde{g}_i'(\boldsymbol{\beta}^*) + \left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \right)^{\mathrm{T}} \sum_{i=1}^{n} \tilde{g}_i'' \left(\boldsymbol{\beta}^\ddagger \right) \left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* \right) \Big/ 2, \tag{25}$$

where $\boldsymbol{\beta}^\ddagger$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$, and by applying Theorem 1 and (R3), we see that this converges in distribution to

$$\tilde{\boldsymbol{s}}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) - \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) \Big/ 2,$$

where $\tilde{\boldsymbol{s}}^{(2)}$ is a copy of $\boldsymbol{s}^{(2)}$. Hence, we have

$$z^{\text{limit}} = \boldsymbol{s}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right) - \tilde{\boldsymbol{s}}^{(2)\mathrm{T}} \boldsymbol{J}^{(22)-1} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right).$$

Because $s^{(2)}$ and $\tilde{s}^{(2)}$ are independently distributed according to $N(\mathbf{0}, J^{(22)})$, the asymptotic bias reduces to

$$\mathrm{E}\left[z^{\mathrm{limit}}\right] = \mathrm{E}\left[s^{(2)\mathrm{T}} J^{(22)-1}\left(s^{(2)} - p_\lambda^{'(2)}\right)\right],$$

and we obtain the following theorem.

Theorem 2 *Under the same conditions as in Theorem 1, we have*

$$\mathrm{E}\left[z^{\mathrm{limit}}\right] = \left|\mathcal{J}^{(2)}\right|$$

when $0 < q < 1$, and we have

$$\mathrm{E}\left[z^{\mathrm{limit}}\right] = \left|\mathcal{J}^{(2)}\right| + K \tag{26}$$

when $q = 1$, where $K = \mathrm{E}[\hat{u}^{(1)\mathrm{T}} s^{(1|2)}]$, $s^{(1|2)} = s^{(1)} - J^{(12)} J^{(22)-1} s^{(2)}$, and $\hat{u}^{(1)}$ is the random vector defined in (21).

We can obtain the result in the case of $q = 1$ in almost the same way as in the case of $0 < q < 1$ (see Sect. A.5 for details). Because the asymptotic bias derived in Theorem 2 depends on an unknown value β^*, we need to evaluate it. Here, we use the fact that $\hat{\beta}_\lambda$ is a consistent estimator of β^* from Lemma 2 and that $J_n(\hat{\beta}_\lambda) = n^{-1} \sum_{i=1}^n X_i^{\mathrm{T}} a''(X_i \hat{\beta}_\lambda) X_i$ converges in probability to J. Concretely speaking, we replace $\mathcal{J}^{(2)}$ by the active set $\hat{\mathcal{J}}^{(2)} = \{j; \ \hat{\beta}_{\lambda,j} \neq 0\}$ and K by its empirical mean $\hat{K}$ obtained by generating samples from $N(\mathbf{0}, J_n(\hat{\beta}_\lambda))$. As a result, we propose the following index as an AIC for the non-concave penalized maximum likelihood method:

$$\mathrm{AIC}_\lambda^{\ell_q\text{-type}} = \begin{cases} -2\sum_{i=1}^n g_i\left(\hat{\beta}_\lambda\right) + 2\left|\hat{\mathcal{J}}^{(2)}\right| & (0 < q < 1) \\ -2\sum_{i=1}^n g_i\left(\hat{\beta}_\lambda\right) + 2\left|\hat{\mathcal{J}}^{(2)}\right| + 2\hat{K} & (q = 1) \end{cases}. \tag{27}$$

When $0 < q < 1$, we can see that the bias term of the information criterion in Efron et al. (2004) or Zou et al. (2007) can be used not only for Gaussian linear regression settings, but also for generalized linear settings. Thus, by minimizing the AIC in (27), we can obtain the optimal value of the tuning parameter λ.

5 Moment convergence

By adding trivial conditions, we can verify that convergence holds in mean for the asymptotic bias in Theorem 2; that is, the second term in (22) converges to $|\mathcal{J}^{(2)}|$ when $0 < q < 1$ and $|\mathcal{J}^{(2)}| + K$ when $q = 1$. Note that this sort of verification is usually ignored in the literature (see, e.g., Konishi and Kitagawa 2008).

To deal with the cases of $0 < q < 1$ and $q = 1$ simultaneously, let us denote $\sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(n^{-1/2}\boldsymbol{u} + \boldsymbol{\beta}^*)\} - n^{1/2}\sum_{j=1}^{p}\{p_\lambda(\beta_j^*) - p_\lambda(n^{-1/2}u_j + \beta_j^*)\}$ by $v_n(\boldsymbol{u})$ also for $0 < q < 1$ in this section and the weak limit of $\boldsymbol{u}_n = \operatorname{argmin}_{\boldsymbol{u}} v_n(\boldsymbol{u})$ by

$$
\begin{aligned}
\tilde{\boldsymbol{u}} &= (\tilde{\boldsymbol{u}}^{(1)}, \tilde{\boldsymbol{u}}^{(2)}) \\
&= \left(\hat{\boldsymbol{u}}^{(1)} 1_{\{q=1\}}, -\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}^{(1)} 1_{\{q=1\}} + \boldsymbol{J}^{(22)-1}\left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)}\right)\right)
\end{aligned}
$$

which is given in Corollary 1.

First, we state the result of applying the theorem in Yoshida (2011) to our problem, which gives sufficient conditions for a polynomial-type large deviation inequality with respect to $\boldsymbol{u}_n$. Note that the theorem in Yoshida (2011) also plays an essential role in Masuda and Shimizu (2017). In this section, we assume that $\mathcal{B}$ is a precompact set. Letting $\alpha \in (0, 1)$, $L > 2$ and $\omega_n(\boldsymbol{u}) = v_n(\boldsymbol{u}) - n^{1/2}\sum_{j\in\mathcal{J}^{(1)}} p_\lambda(n^{-1/2}u_j) + \boldsymbol{u}^\mathrm{T}\boldsymbol{s}_n - \boldsymbol{u}^\mathrm{T}\boldsymbol{J}\boldsymbol{u}/2$, the sufficient conditions can be written as follows:

(A1) $\exists \chi_1 = \chi_1(\boldsymbol{\beta}^*) > 0$, $\exists \chi_2 = \chi_2(\boldsymbol{\beta}^*) > 0$, $\forall \boldsymbol{\beta} \in \mathcal{B}$, $h(\boldsymbol{\beta}) \geq \chi_1 |\boldsymbol{\beta} - \boldsymbol{\beta}^*|^{\chi_2}$.
(A2) $\exists \gamma_1 > 0$, $\exists c_1 > 0$,

$$
\sup_{r>0}\sup_{n>0} r^L \mathrm{P}\left(\sup_{\boldsymbol{u}\in U_n(r)} \frac{|\omega_n(\boldsymbol{u})|}{1+|\boldsymbol{u}|^2} \geq r^{-\gamma_1}\right) \leq c_1,
$$

where $U_n(r) = \{\boldsymbol{u} \in \mathbb{R}^p;\ r \leq |\boldsymbol{u}| \leq n^{(1-\alpha)/2}\}$.
(A3) $\exists \gamma_2 \in [0, 1/2)$, $\exists c_2 \in (\alpha\chi_2, 1 - 2\gamma_2)$,

$$
\sup_{n>0} \mathrm{E}\left[|\boldsymbol{s}_n|^{N_1}\right] < \infty \quad \text{and} \quad \sup_{n>0} \mathrm{E}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\{n^{1/2-\gamma_2}|\mu_n(\boldsymbol{\beta}) - h(\boldsymbol{\beta})|\right\}^{N_2}\right] < \infty
$$

where $N_1 = L(1 - \gamma_1)^{-1}$, $N_2 = L(1 - 2\gamma_2 - c_2)^{-1}$, and $\mu_n(\boldsymbol{\beta})$ is the random function defined in (3).

Theorem 3 (Yoshida 2011) *If there exists α ($\in (0, 1)$) such that (A1)–(A3) hold, we have*

$$
\sup_{r>0}\sup_{n>0} r^L \mathrm{P}\left(\sup_{|\boldsymbol{u}|\geq r} \{-v_n(\boldsymbol{u})\} \geq 0\right) < \infty. \tag{28}
$$

The definition of $\omega_n(\boldsymbol{u})$ may seem somewhat strange, but this can be justified from the non-negativity of $p_\lambda(\cdot)$. In fact, we see that

$$
\mathrm{P}\left(\sup_{|\boldsymbol{u}|\geq r}\{-v_n(\boldsymbol{u})\} \geq 0\right) \leq \mathrm{P}\left(\sup_{|\boldsymbol{u}|\geq r}\left\{-v_n(\boldsymbol{u}) + n^{1/2}\sum_{j\in\mathcal{J}^{(1)}} p_\lambda\left(\frac{u_j}{n^{1/2}}\right)\right\} \geq 0\right)
$$

holds. Therefore, to obtain (28), it suffices to establish a polynomial-type large deviation inequality for a random function $-v_n(\boldsymbol{u}) + n^{1/2}\sum_{j\in\mathcal{J}^{(1)}} p_\lambda(n^{-1/2}u_j)$ instead of $-v_n(\boldsymbol{u})$.

We can easily obtain from (28) that

$$\sup_{r>0} \sup_{n>0} r^L \mathrm{P}\left(|\boldsymbol{u}_n| \geq r\right) < \infty.$$

Moreover, considering the weak convergence of $\boldsymbol{u}_n$ to $\tilde{\boldsymbol{u}}$, we have

$$\mathrm{E}\left[f_L(\boldsymbol{u}_n)\right] \to \mathrm{E}\left[f_L(\tilde{\boldsymbol{u}})\right] \tag{29}$$

for every polynomial growth function $f_L : \mathbb{R}^p \to \mathbb{R}$ whose order is less than L.

The sufficient conditions (A1)–(A3) cannot be derived from only (C1)–(C5); we require additional trivial conditions:

(C6) The eigenvalues of $\boldsymbol{J}(\boldsymbol{\beta})$ are uniformly bounded away from 0 and infinity over $\boldsymbol{\beta} \in \mathcal{B}$.

(C7) There exists δ_1 $(\in (0, 1))$ such that

$$\sup_{n>0}\left\{n^{\delta_1}\left|\frac{1}{n}\sum_{i=1}^{n} g_i''(\boldsymbol{\beta}^*) + \boldsymbol{J}\right|\right\} < \infty.$$

(C8) There exists δ_2 $(\in (0, 1))$ such that

$$\sup_{n>0} \mathrm{E}\left[\left\{n^{\delta_2}\left|\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{X}_i - \int_{\mathcal{X}} a'(\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}\mu(\mathrm{d}\boldsymbol{X})\right|\right\}^k\right] < \infty$$

for all k $(\in \mathbb{N})$ and

$$\sup_{n>0} \sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\{n^{\delta_2}\left|\frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{X}_i\boldsymbol{\beta}) - \int_{\mathcal{X}} a(\boldsymbol{X}\boldsymbol{\beta})\mu(\mathrm{d}\boldsymbol{X})\right|\right\} < \infty.$$

Letting $\alpha \in (0, \min\{2\delta_1, \delta_2, 1/2\})$, we will check the sufficient conditions.

First, it can be easily seen from (C6) that (A1) holds by setting χ_1 to the infimum of the smallest eigenvalue of $\boldsymbol{J}(\boldsymbol{\beta})$ over $\boldsymbol{\beta} \in \mathcal{B}$ and $\chi_2 = 2$, as we obtain

$$h(\boldsymbol{\beta}) = \int_{\mathcal{X}} \left\{(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} a''(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right\} \mu(\mathrm{d}\boldsymbol{X})$$

from using Taylor's theorem for $h(\boldsymbol{\beta})$ in (1), where $\tilde{\boldsymbol{\beta}}$ is a vector between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$.

Next, let us consider (A2). Using Taylor's theorem, $\omega_n(\boldsymbol{u})$ can be written as

$$-\boldsymbol{u}^{\mathrm{T}} \int_0^1 (1-s)\left\{\frac{1}{n}\sum_{i=1}^{n} g_i''\left(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}s}{n^{1/2}}\right) + \boldsymbol{J}\right\}\mathrm{d}s\,\boldsymbol{u}$$

$$+ n^{1/2} \sum_{j\in\mathcal{J}^{(2)}} \left\{p_\lambda\left(\beta_j^* + \frac{u_j}{n^{1/2}}\right) - p_\lambda(\beta_j^*)\right\}.$$

Using Taylor's theorem again for $g_i''(\boldsymbol{\beta}^* + n^{-1/2}\boldsymbol{u}s)$ and (C4), we get

$$
\frac{|\omega_n(\boldsymbol{u})|}{1+|\boldsymbol{u}|^2} \lesssim \frac{|\boldsymbol{u}|^2}{1+|\boldsymbol{u}|^2} \left| \frac{1}{n} \sum_{i=1}^n g_i''(\boldsymbol{\beta}^*) + \boldsymbol{J} \right|
$$
$$
+ \frac{|\boldsymbol{u}|^2}{1+|\boldsymbol{u}|^2} \frac{|\boldsymbol{u}|}{n^{1/2}} \int_0^1 \int_0^1 \left| \frac{1}{n} \sum_{i=1}^n g_i'''\left(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}st}{n^{1/2}}\right) \right| dt\,ds + \frac{|\boldsymbol{u}|}{1+|\boldsymbol{u}|^2},
$$
$$(30)$$

where $A_n \lesssim B_n$ means that $\sup_n(A_n/B_n) < \infty$. Let $0 < \xi < \alpha/(1-\alpha)$. Note that $-\alpha/2 + (1-\alpha)\xi/2 < 0$, and therefore, $-\delta_1 + (1-\alpha)\xi/2 < 0$. Then, for the first term of the right-hand side in (30), it follows from (C7) that

$$
\sup_{\boldsymbol{u}\in U_n(r)} \left\{ \frac{|\boldsymbol{u}|^2}{1+|\boldsymbol{u}|^2} \left| \frac{1}{n} \sum_{i=1}^n g_i''(\boldsymbol{\beta}^*) + \boldsymbol{J} \right| \right\}
$$
$$
= n^{\delta_1} \left| \frac{1}{n} \sum_{i=1}^n g_i''(\boldsymbol{\beta}^*) + \boldsymbol{J} \right| \sup_{\boldsymbol{u}\in U_n(r)} \left(\frac{|\boldsymbol{u}|^2}{1+|\boldsymbol{u}|^2} \frac{|\boldsymbol{u}|^\xi |\boldsymbol{u}|^{-\xi}}{n^{\delta_1}} \right)
$$
$$
\lesssim n^{-\delta_1+(1-\alpha)\xi/2} r^{-\xi} \lesssim r^{-\xi}.
$$
$$(31)$$

In addition, for the second and third terms of the right-hand side in (30), we have

$$
\sup_{\boldsymbol{u}\in U_n(r)} \left\{ \frac{|\boldsymbol{u}|^2}{1+|\boldsymbol{u}|^2} \frac{|\boldsymbol{u}|}{n^{1/2}} \int_0^1 \int_0^1 \left| \frac{1}{n} \sum_{i=1}^n g_i'''\left(\boldsymbol{\beta}^* + \frac{\boldsymbol{u}st}{n^{1/2}}\right) \right| dt\,ds + \frac{|\boldsymbol{u}|}{1+|\boldsymbol{u}|^2} \right\}
$$
$$
\lesssim \sup_{\boldsymbol{u}\in U_n(r)} \left(\frac{|\boldsymbol{u}|^2}{1+|\boldsymbol{u}|^2} \frac{|\boldsymbol{u}|^\xi |\boldsymbol{u}|^{-\xi}}{n^{\alpha/2}} + \frac{|\boldsymbol{u}|}{1+|\boldsymbol{u}|^2} \right) \lesssim n^{-\alpha/2+(1-\alpha)\xi/2} r^{-\xi} + r^{-1} \lesssim r^{-\xi}.
$$
$$(32)$$

Letting $\gamma_1 \in (0,\xi)$, it can be seen that (A2) holds from (30), (31), and (32).

Finally, let us consider (A3). From Burkholder's and Jensen's inequalities (Shiryaev 1996), we have

$$
\sup_{n>0} E[|\boldsymbol{s}_n|^{N_1}] \le \sup_{n>0} E\left[\max_{k\le n} \left| \sum_{i=1}^k \frac{g_i'(\boldsymbol{\beta}^*)}{n^{1/2}} \right|^{N_1} \right]
$$
$$
\lesssim \sup_{n>0} E\left[\left\{ \sum_{i=1}^n \frac{g_i'(\boldsymbol{\beta}^*)^2}{n} \right\}^{N_1/2} \right] \le \sup_{n>0} E\left[\frac{1}{n} \sum_{i=1}^n |g_i'(\boldsymbol{\beta}^*)|^{N_1} \right] < \infty
$$
$$(33)$$

for $N_1 = L(1-\gamma_1)^{-1} \ge 2$. Let us fix γ_2 and c_2 such that $\alpha\chi_2 < c_2 < 1 - 2\gamma_2 < \min\{2\delta_2, 1\}$. Since $(A+B)^{N_2} \lesssim A^{N_2} + B^{N_2}$ when A and B are positive and

$N_2 = L(1 - 2\gamma_2 - c_2)^{-1} \geq 2$, it follows from (C8) that

$$\sup_{n>0} E\left[\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left[n^{1/2-\gamma_2} \left| \frac{1}{n} \sum_{i=1}^{n} \{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} - h(\boldsymbol{\beta}) \right| \right]^{N_2} \right]$$

$$\lesssim \sup_{n>0} E\left[\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left\{ n^{1/2-\gamma_2} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{X}_i (\boldsymbol{\beta}^* - \boldsymbol{\beta}) \right. \right. \right.$$

$$\left. \left. \left. - \int_{\mathcal{X}} a'(\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) \mu(\mathrm{d}\boldsymbol{X}) \right| \right\}^{N_2} \right]$$

$$+ \sup_{n>0} \sup_{\boldsymbol{\beta} \in \mathcal{B}} \left[n^{1/2-\gamma_2} \left| \frac{1}{n} \sum_{i=1}^{n} \{a(\boldsymbol{X}_i \boldsymbol{\beta}^*) - a(\boldsymbol{X}_i \boldsymbol{\beta})\} \right. \right.$$

$$\left. \left. - \int_{\mathcal{X}} \{a(\boldsymbol{X}\boldsymbol{\beta}^*) - a(\boldsymbol{X}\boldsymbol{\beta})\} \mu(\mathrm{d}\boldsymbol{X}) \right| \right]^{N_2}$$

$$< \infty. \tag{34}$$

Further, we obtain from the precompactness of $\mathcal{B}$ that

$$\sup_{n>0} \sup_{\boldsymbol{\beta} \in \mathcal{B}} \left[n^{1/2-\gamma_2} \left| \frac{1}{n^{1/2}} \sum_{j=1}^{p} \left\{ p_\lambda(\beta_j) - p_\lambda(\beta_j^*) \right\} \right| \right]^{N_2} < \infty. \tag{35}$$

Hence, it can be seen that (A3) holds from (33), (34), and (35).

Now, let us summarize the above discussion.

Theorem 4 *Under conditions (C1)–(C8), moment convergence* (29) *holds.*

By looking at the derivation of Theorem 2 carefully, we can see that the second term in (22) can be rewritten as

$$E\left[\boldsymbol{u}_n^{\mathrm{T}} \boldsymbol{s}_n \right] - E\left[\boldsymbol{u}_n^{\mathrm{T}} \{\boldsymbol{J}_n(\boldsymbol{\beta}^\dagger) - \boldsymbol{J}_n(\boldsymbol{\beta}^\ddagger)\} \boldsymbol{u}_n \right] \Big/ 2. \tag{36}$$

Let $\delta \in (0, L/2 - 1)$. For the first term in (36), it follows from the Cauchy–Schwarz inequality, (29), and (33) that

$$\sup_{n>0} E[|\boldsymbol{u}_n^{\mathrm{T}} \boldsymbol{s}_n|^{1+\delta}] \leq \left(\sup_{n>0} E[|\boldsymbol{u}_n|^{2(1+\delta)}] \right)^{1/2} \left(\sup_{n>0} E[|\boldsymbol{s}_n|^{2(1+\delta)}] \right)^{1/2} < \infty.$$

In addition, for the second term in (36), it follows from (29) that

$$\sup_{n>0} E\left[|\boldsymbol{u}_n^{\mathrm{T}} \{\boldsymbol{J}_n(\boldsymbol{\beta}^\dagger) - \boldsymbol{J}_n(\boldsymbol{\beta}^\ddagger)\} \boldsymbol{u}_n|^{1+\delta} \right] / 2 \leq \chi_3 \sup_{n>0} E\left[|\boldsymbol{u}_n|^{2(1+\delta)} \right] < \infty,$$

where χ_3 is the supremum of the largest eigenvalue of $\boldsymbol{J}_n(\boldsymbol{\beta})$ over $\mathcal{B}$. These uniform integrabilities assure the convergence of (36) to $E[\tilde{\boldsymbol{u}}^{\mathrm{T}} \boldsymbol{s}]$.

6 Simulation study

We conducted simulation studies to check the performance of tuning parameter selection based on the AIC in (27). Concretely speaking, we considered a linear regression setting (linear) and a logistic regression setting (logistic) and compared the performances of AIC and fivefold CV. As regularization methods, we used the Bridge ($q = 0.2$), SCAD, and MCP.

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

$$\text{KL} = -\tilde{\text{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})\right],$$

where $\hat{\lambda}$ is the value of the tuning parameter given by each of the criteria, and we evaluated the expectation using an empirical mean of 500 samples. We interpreted that a criterion giving a small KL value is good. Although the original aim of AIC is to minimize KL, as a secondary index for the assessment, we also determined the number of false positives and false negatives:

$$\text{FP} = |\{j; \ \hat{\beta}_j \neq 0 \wedge \beta_j^* = 0\}| \quad \text{and} \quad \text{FN} = |\{j; \ \hat{\beta}_j = 0 \wedge \beta_j^* \neq 0\}|,$$

for each of the criteria.

The AICs we used included the one corresponding to the case $0 < q < 1$ in (27) for the Bridge and the one corresponding to the case $q = 1$ in (27) for SCAD and MCP. Note that the log-likelihood function $g_i(\boldsymbol{\beta})$ for a linear or a logistic regression setting is expressed as

$$y_i X_i \boldsymbol{\beta} - (X_i \boldsymbol{\beta})^2/2 - y_i^2/2 \quad \text{or} \quad y_i X_i \boldsymbol{\beta} - \log\{1 + \exp(X_i \boldsymbol{\beta})\},$$

and $\boldsymbol{J}_n(\boldsymbol{\beta})$ needed for evaluating $\hat{K}$ can be expressed as

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{\text{T}} X_i \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(X_i \boldsymbol{\beta})}{\{1 + \exp(X_i \boldsymbol{\beta})\}^2} X_i^{\text{T}} X_i.$$

The simulation settings were as follows. As p-dimensional regressors X_i, ($i = 1, 2, \ldots, n$), we used vectors obtained from the multivariate Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is ($p \times p$)-covariance matrix whose (i, j)th element was set to $0.5^{|i-j|}$. The true coefficient vector $\boldsymbol{\beta}^*$ was

$$\boldsymbol{\beta}^* = \left(\beta_1^* \boldsymbol{1}_k^{\text{T}}, \beta_2^* \boldsymbol{1}_k^{\text{T}}, \boldsymbol{0}_{p-2k}^{\text{T}}\right)^{\text{T}},$$

where $\boldsymbol{1}_k$ and $\boldsymbol{0}_{p-2k}$, respectively, denote a k-dimensional one vector and a $(p - 2k)$-dimensional zero vector. In addition, (β_1^*, β_2^*) was set to $(0.1, 0.5)$ or $(0.2, 1)$ in the linear regression setting and $(0.5, 1.5)$ or $(1, 2)$ in the logistic regression setting, and

seven cases of the three-tuple (p, k, n) were considered: (8, 2, 50), (8, 2, 100), (8, 2, 150), (8, 1, 100), (8, 3, 100), (12, 3, 100), and (16, 4, 100). We used the local quadratic approximation in Fan and Li (2001) for the parameter estimation and conducted fifty simulations.

Tables 1, 2, and 3 show the results for the Bridge, SCAD, and MCP, respectively. Each table lists the averages and standard deviations of KL, as well as the averages of FP and FN, for the linear and the logistic regression settings. Let us look at the main index in Table 1. While CV gives a smaller KL value than AIC does in about half the cases, the differences between the two values are small. On the other hand, in the cases in which AIC gives a smaller KL value than CV does, the differences tend to be large. Next, let us look at the subindices FP and FN. In the logistic setting, the FP values are almost 0, while those of FN are rather large. That is, we can say that CV causes an imbalance. So long as there is no special reason of given importance on the FP, it will be natural to use the AIC. In Tables 2 and 3, AIC and CV give almost the same values of KL in the linear setting. On the other hand, in the logistic setting, AIC is clearly superior to CV in many cases. On the whole, we can conclude that the AIC in (27) is better than CV.

7 Discussion

Although Ninomiya and Kawano (2016) derived an information criterion for the Lasso in generalized linear models on the basis of the original definition of AIC, which is an asymptotically unbiased estimator of the Kullback–Leibler divergence, they used an asymptotic setting wherein estimation consistency is not assured. In addition, the Lasso itself has a problem in that efficiency is not necessarily high because it shrinks the estimator to the zero vector too much. As a way of dealing with these problems, we derived an information criterion for non-concave penalized maximum likelihood methods including the Bridge, SCAD, and MCP, which are known to be more efficient than the Lasso, on the basis of the original definition of AIC in a setting in which estimation consistency is assured. The AIC in (27) is the only criterion for such non-concave penalized maximum likelihood methods that has the same roots as those of the classic information criteria. Its bias term, including its coefficient, is determined. Therefore, unlike the information criteria that assure model selection consistency, it allows us to perform a model selection without any arbitrariness.

It has been shown through simulation studies that the performance of the AIC in (27) is almost the same as or better than that of the CV. In terms of computational cost, AIC is clearly better than CV in the Bridge-type regularization method because of its simple expression. This fact is a significant advantage when handling large-scale data.

Although the number of tuning parameters to be selected is only one, we can extend our result to regularization methods that have several tuning parameters, such as SELO (Dicker et al. 2012). In addition, although we used the natural link function for our generalized linear models, it is possible to treat different link functions given certain regularity conditions. Similarly to in Ninomiya and Kawano (2016), we can treat the case where there exists a dispersion parameter which will be estimated by a non-sparse method. In this case, our criterion will depend on the dispersion parameter, and

Table 1 Comparison of CV and AIC in (27) for the Bridge penalty. The true parameter vector (β_1^*, β_2^*) is (0.1, 0.5) for Case 1 and (0.2, 1) for Case 2 in the linear regression setting and (0.5, 1.5) and (1,2) in the logistic regression setting

Model	(p, k, n)		Case 1			Case 2		
			KL (sd)	FP	FN	KL (sd)	FP	FN
Linear	(8, 2, 50)	CV	0.676 (0.019)	0.30	1.58	0.645 (0.026)	0.30	1.29
		AIC	0.679 (0.018)	0.09	1.77	0.649 (0.022)	0.11	1.55
	(8, 2, 100)	CV	0.670 (0.016)	0.31	1.31	0.631 (0.018)	0.28	1.05
		AIC	0.672 (0.015)	0.05	1.61	0.634 (0.018)	0.07	1.27
	(8, 2, 150)	CV	0.666 (0.014)	0.32	1.24	0.632 (0.012)	0.40	0.86
		AIC	0.666 (0.013)	0.10	1.45	0.636 (0.014)	0.04	1.17
	(8, 1, 100)	CV	0.687 (0.008)	0.46	0.75	0.658 (0.017)	0.75	0.45
		AIC	0.687 (0.009)	0.12	0.81	0.658 (0.016)	0.13	0.54
	(8, 3, 100)	CV	0.655 (0.014)	0.24	1.86	0.615 (0.020)	0.24	1.40
		AIC	0.659 (0.012)	0.03	2.34	0.626 (0.019)	0.04	2.19
	(12, 3, 100)	CV	0.662 (0.014)	0.47	1.91	0.617 (0.021)	0.46	1.64
		AIC	0.665 (0.014)	0.15	2.38	0.624 (0.018)	0.06	2.17
	(16, 4, 100)	CV	0.652 (0.021)	0.41	3.03	0.610 (0.024)	0.69	2.47
		AIC	0.652 (0.017)	0.12	3.28	0.618 (0.021)	0.12	2.98
Logistic	(8, 2, 50)	CV	0.462 (0.061)	0.01	1.28	0.406 (0.070)	0.04	1.21
		AIC	0.473 (0.153)	0.33	0.69	0.417 (0.129)	0.40	0.40
	(8, 2, 100)	CV	0.419 (0.044)	0.01	1.04	0.348 (0.047)	0.00	0.92
		AIC	0.398 (0.050)	0.31	0.43	0.307 (0.035)	0.50	0.19
	(8, 2, 150)	CV	0.394 (0.024)	0.00	0.94	0.307 (0.033)	0.01	0.67
		AIC	0.376 (0.018)	0.43	0.33	0.271 (0.018)	0.41	0.11
	(8, 1, 100)	CV	0.495 (0.029)	0.00	0.42	0.411 (0.021)	0.00	0.22
		AIC	0.513 (0.033)	0.61	0.21	0.423 (0.035)	0.63	0.02
	(8, 3, 100)	CV	0.408 (0.047)	0.00	1.92	0.348 (0.053)	0.00	1.74
		AIC	0.346 (0.042)	0.22	0.78	0.272 (0.087)	0.35	0.32
	(12, 3, 100)	CV	0.384 (0.031)	0.01	1.82	0.376 (0.056)	0.00	1.68
		AIC	0.397 (0.134)	0.75	0.58	0.346 (0.112)	0.73	0.35
	(16, 4, 100)	CV	0.392 (0.048)	0.01	2.72	0.407 (0.045)	0.00	2.66
		AIC	0.414 (0.122)	1.19	1.05	0.379 (0.137)	1.17	0.60

so we need to replace it by its estimator. In this study, we derived the AIC based on statistical asymptotic theory for which the dimension of the parameter vector is fixed and the sample size diverges. On the other hand, it is becoming important to analyze high-dimensional data wherein the dimension of the parameter vector is comparable to the sample size. Also, for such high-dimensional data, we expect that the AIC-type information criterion will work well from the viewpoint of efficiency. In fact, Zhang et al. (2010) has shown that, when the dimension of the parameter vector increases with the sample size, their criterion close to the proposed information criterion has an asymptotic loss efficiency in a sparse setting under certain conditions. It will be

Table 2 Comparison of CV and AIC in (27) for the SCAD penalty. The true parameter vector (β_1^*, β_2^*) is (0.1, 0.5) for Case 1 and (0.2, 1) for Case 2 in the linear regression setting and (0.5, 1.5) and (1, 2) in the logistic regression setting

Model	(p, k, n)		Case 1			Case 2		
			KL (sd)	FP	FN	KL (sd)	FP	FN
Linear	(8, 2, 50)	CV	0.557 (0.050)	0.69	0.49	0.563 (0.039)	0.87	0.20
		AIC	0.566 (0.055)	0.60	0.59	0.582 (0.056)	0.95	0.20
	(8, 2, 100)	CV	0.521 (0.020)	1.01	0.27	0.518 (0.031)	0.93	0.11
		AIC	0.524 (0.025)	0.92	0.28	0.519 (0.028)	0.91	0.15
	(8, 2, 150)	CV	0.531 (0.013)	0.76	0.24	0.567 (0.012)	1.05	0.03
		AIC	0.534 (0.015)	0.70	0.26	0.569 (0.013)	0.89	0.03
	(8, 1, 100)	CV	0.526 (0.021)	1.24	0.19	0.500 (0.020)	1.26	0.06
		AIC	0.526 (0.025)	1.05	0.24	0.503 (0.023)	1.13	0.06
	(8, 3, 100)	CV	0.491 (0.020)	0.49	0.41	0.555 (0.025)	0.59	0.17
		AIC	0.492 (0.021)	0.43	0.51	0.555 (0.027)	0.48	0.22
	(12, 3, 100)	CV	0.504 (0.020)	1.16	0.37	0.556 (0.023)	1.33	0.15
		AIC	0.509 (0.028)	1.15	0.38	0.561 (0.026)	1.23	0.16
	(16, 4, 100)	CV	0.550 (0.030)	1.54	0.66	0.565 (0.029)	1.80	0.15
		AIC	0.557 (0.035)	1.39	0.66	0.573 (0.031)	1.44	0.24
Logistic	(8, 2, 50)	CV	0.506 (0.032)	0.04	0.82	0.493 (0.023)	0.06	0.59
		AIC	0.477 (0.117)	0.76	0.56	0.511 (0.184)	0.48	0.54
	(8, 2, 100)	CV	0.476 (0.017)	0.07	0.69	0.426 (0.018)	0.04	0.20
		AIC	0.446 (0.059)	0.78	0.41	0.321 (0.037)	0.52	0.25
	(8, 2, 150)	CV	0.451 (0.015)	0.05	0.41	0.394 (0.015)	0.06	0.13
		AIC	0.411 (0.021)	1.09	0.18	0.301 (0.025)	0.95	0.08
	(8, 1, 100)	CV	0.541 (0.017)	0.15	0.14	0.454 (0.024)	0.07	0.06
		AIC	0.542 (0.036)	1.40	0.09	0.406 (0.029)	1.01	0.04
	(8, 3, 100)	CV	0.431 (0.017)	0.05	1.09	0.423 (0.015)	0.05	0.54
		AIC	0.339 (0.043)	0.38	0.66	0.314 (0.056)	0.19	0.55
	(12, 3, 100)	CV	0.449 (0.014)	0.03	0.95	0.420 (0.015)	0.03	0.53
		AIC	0.413 (0.093)	1.44	0.46	0.349 (0.086)	0.86	0.59
	(16, 4, 100)	CV	0.436 (0.013)	0.08	1.50	0.423 (0.018)	0.06	1.19
		AIC	0.438 (0.115)	1.52	0.99	0.356 (0.080)	0.87	1.11

important in terms of both theory and practice to show that the proposed information criterion has a similar asymptotic property.

A Proofs

A.1 Proof of Lemma 2

From (R1), the first term in the right-hand side of (3) converges in probability to $h(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$. In addition, from the convexity of $g_i(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we have

Table 3 Comparison of CV and AIC in (27) for the MCP penalty. The true parameter vector (β_1^*, β_2^*) is (0.1, 0.5) for Case 1 and (0.2, 1) for Case 2 in the linear regression setting and (0.5, 1.5) and (1, 2) in the logistic regression setting

Model	(p, k, n)		Case 1			Case 2		
			KL (sd)	FP	FN	KL (sd)	FP	FN
Linear	(8, 2, 50)	CV	0.545 (0.047)	0.82	0.42	0.556 (0.046)	0.79	0.23
		AIC	0.545 (0.047)	0.67	0.49	0.557 (0.046)	0.71	0.29
	(8, 2, 100)	CV	0.558 (0.020)	0.79	0.38	0.527 (0.023)	0.86	0.13
		AIC	0.560 (0.026)	0.64	0.39	0.530 (0.027)	0.92	0.13
	(8, 2, 150)	CV	0.520 (0.017)	0.91	0.31	0.518 (0.015)	0.94	0.10
		AIC	0.521 (0.018)	0.71	0.38	0.519 (0.015)	0.84	0.11
	(8, 1, 100)	CV	0.502 (0.015)	1.02	0.25	0.539 (0.023)	1.03	0.15
		AIC	0.503 (0.018)	0.88	0.27	0.540 (0.024)	0.99	0.14
	(8, 3, 100)	CV	0.553 (0.021)	0.33	0.53	0.508 (0.028)	0.62	0.10
		AIC	0.556 (0.023)	0.30	0.61	0.510 (0.029)	0.49	0.16
	(12, 3, 100)	CV	0.523 (0.023)	1.24	0.57	0.578 (0.030)	1.45	0.17
		AIC	0.525 (0.024)	1.02	0.57	0.582 (0.028)	1.39	0.19
	(16, 4, 100)	CV	0.530 (0.029)	1.72	0.72	0.563 (0.035)	1.73	0.28
		AIC	0.532 (0.031)	1.45	0.72	0.565 (0.036)	1.53	0.34
Logistic	(8, 2, 50)	CV	0.493 (0.037)	0.04	1.04	0.453 (0.035)	0.06	0.81
		AIC	0.514 (0.159)	0.59	0.59	0.383 (0.090)	0.41	0.52
	(8, 2, 100)	CV	0.447 (0.023)	0.02	0.65	0.397 (0.025)	0.02	0.47
		AIC	0.418 (0.043)	0.79	0.29	0.323 (0.029)	0.54	0.21
	(8, 2, 150)	CV	0.423 (0.017)	0.04	0.54	0.367 (0.019)	0.01	0.17
		AIC	0.390 (0.019)	0.88	0.21	0.308 (0.020)	0.94	0.09
	(8, 1, 100)	CV	0.529 (0.020)	0.10	0.23	0.448 (0.021)	0.13	0.08
		AIC	0.530 (0.036)	0.83	0.17	0.429 (0.027)	1.06	0.06
	(8, 3, 100)	CV	0.429 (0.020)	0.01	1.09	0.409 (0.031)	0.02	0.97
		AIC	0.362 (0.056)	0.33	0.70	0.312 (0.075)	0.16	0.73
	(12, 3, 100)	CV	0.423 (0.027)	0.01	1.10	0.401 (0.017)	0.01	0.99
		AIC	0.389 (0.070)	1.02	0.66	0.352 (0.075)	0.91	0.65
	(16, 4, 100)	CV	0.426 (0.022)	0.02	1.92	0.411 (0.017)	0.02	1.54
		AIC	0.440 (0.136)	1.79	0.94	0.345 (0.107)	1.31	1.02

$$\sup_{\boldsymbol{\beta} \in K} \left| \frac{1}{n} \sum_{i=1}^{n} \{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} - h(\boldsymbol{\beta}) \right| \xrightarrow{\mathrm{p}} 0$$

for any compact set K (Andersen and Gill 1982; Pollard 1991). Accordingly, we have

$$\sup_{\boldsymbol{\beta} \in K} |\mu_n(\boldsymbol{\beta}) - h(\boldsymbol{\beta})| \xrightarrow{\mathrm{p}} 0. \tag{37}$$

Note that in the following inequality,

$$\mu_n(\boldsymbol{\beta}) \geq \frac{1}{n} \sum_{i=1}^{n} \{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} - \frac{1}{n^{1/2}} \sum_{j=1}^{p} p_\lambda(\beta_j^*) \equiv \mu_n^{(0)}(\boldsymbol{\beta}),$$

the argmin of the right-hand side is the maximum likelihood estimator and is $O_p(1)$. Also, note that for some $M \ (> 0)$,

$$P(|\hat{\boldsymbol{\beta}}_\lambda| > M) \leq P\left(\inf_{|\boldsymbol{\beta}|>M} \mu_n(\boldsymbol{\beta}) \leq \mu_n(\boldsymbol{0})\right) \leq P\left(\inf_{|\boldsymbol{\beta}|>M} \mu_n^{(0)}(\boldsymbol{\beta}) \leq \mu_n^{(0)}(\boldsymbol{0})\right)$$

because $p_\lambda(0) = 0$ from (C5). Therefore, we have

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\operatorname{argmin}} \ \mu_n(\boldsymbol{\beta}) = O_p(1). \tag{38}$$

From (37) and (38), we obtain

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\operatorname{argmin}} \ \mu_n(\boldsymbol{\beta}) \overset{\mathrm{p}}{\to} \underset{\boldsymbol{\beta} \in \mathcal{B}}{\operatorname{argmin}} \ h(\boldsymbol{\beta}) = \boldsymbol{\beta}^*.$$

The last equality holds from (1). □

A.2 Proof of (8)

Let $\boldsymbol{u} = \tilde{\boldsymbol{u}}_n + l\boldsymbol{w}$, where $\boldsymbol{w}$ is a unit vector, and let $l \in (\delta, \xi)$. The strong convexity of $\eta_n(\boldsymbol{u})$ implies

$$(1 - \delta/l)\eta_n(\tilde{\boldsymbol{u}}_n) + (\delta/l)\eta_n(\boldsymbol{u}) > \eta_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}),$$

and we thus have

$$\begin{aligned}
(\delta/l)\{v_n(\boldsymbol{u}) - v_n(\tilde{\boldsymbol{u}}_n)\} > \ & v_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - v_n(\tilde{\boldsymbol{u}}_n) \\
& + (1 - \delta/l)\phi_n(\tilde{\boldsymbol{u}}_n) + (\delta/l)\phi_n(\boldsymbol{u}) - \phi_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) \\
& + (1 - \delta/l)\psi_n(\tilde{\boldsymbol{u}}_n^\dagger) + (\delta/l)\psi_n(\boldsymbol{u}^\dagger) - \psi_n(\tilde{\boldsymbol{u}}_n^\dagger + \delta\boldsymbol{w}^\dagger).
\end{aligned}$$

Since it follows that

$$\begin{aligned}
v_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) &- v_n(\tilde{\boldsymbol{u}}_n) \\
&= \left\{v_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - \tilde{v}_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w})\right\} + \left\{\tilde{v}_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - \tilde{v}_n(\tilde{\boldsymbol{u}}_n)\right\} + \left\{\tilde{v}_n(\tilde{\boldsymbol{u}}_n) - v_n(\tilde{\boldsymbol{u}}_n)\right\} \\
&\geq \Upsilon_n(\delta) - 2\Delta_n(\delta),
\end{aligned}$$

we obtain from (6) and (7) that, for any ε (> 0),

$$(\delta/l)\{v_n(\boldsymbol{u}) - v_n(\tilde{\boldsymbol{u}}_n)\} > \Upsilon_n(\delta) - 2\Delta_n(\delta) - \varepsilon$$

for sufficiently large n and sufficiently small γ. If $2\Delta_n(\delta) + \varepsilon < \Upsilon_n(\delta)$, then $v_n(\boldsymbol{u}) \geq v_n(\tilde{\boldsymbol{u}}_n)$ for any $\boldsymbol{u}$ such that $|\boldsymbol{u}^\dagger| \leq \gamma$ and $\delta \leq |\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| \leq \xi$. This means $\boldsymbol{u}_n$ must satisfy $|\boldsymbol{u}_n^\dagger| > \gamma$ or $|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \notin [\delta, \xi]$ in order for $\boldsymbol{u}_n$ to be the argmin of $v_n(\boldsymbol{u})$. Hence, we obtain (8). □

A.3 Proof of (17)

Let us consider a random function $\mu_n(\boldsymbol{\beta})$ in (3). Since $p_\lambda(0) = 0$ from (C5), we have

$$\mu_n(\hat{\boldsymbol{\beta}}_\lambda) = -n^{-1/2}\boldsymbol{s}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2$$
$$+ n^{-1/2}\sum_{j \in \mathcal{J}^{(1)}} p_\lambda(\hat{\beta}_{\lambda,j}) + n^{-1/2}\sum_{j \in \mathcal{J}^{(2)}} p'_\lambda(\beta_j^*)(\hat{\beta}_{\lambda,j} - \beta_j^*)\{1 + o_p(1)\},$$

where $\tilde{\boldsymbol{\beta}}$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. Then, we have

$$0 \geq \mu_n(\hat{\boldsymbol{\beta}}_\lambda) - \mu_n(\boldsymbol{\beta}^*) \neq O_p(n^{-1/2}|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*|) + (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2$$

because $\boldsymbol{s}_n = O_p(1)$. From (C2) and (C3), $\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})$ is positive definite for sufficiently large n, and therefore, it follows that

$$\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* = O_p(n^{-1/2}). \tag{39}$$

Let us express $\mu_n(\boldsymbol{\beta})$ by $\mu_n(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$. Because $0 \geq \mu_n(\hat{\boldsymbol{\beta}}_\lambda^{(1)}, \hat{\boldsymbol{\beta}}_\lambda^{(2)}) - \mu_n(\boldsymbol{0}, \hat{\boldsymbol{\beta}}_\lambda^{(2)})$, we see that

$$-n^{-1/2}\boldsymbol{s}_n^{(1)\mathrm{T}}\hat{\boldsymbol{\beta}}_\lambda^{(1)} + \hat{\boldsymbol{\beta}}_\lambda^{(1)\mathrm{T}}\boldsymbol{J}_n^{(11)}(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}_\lambda^{(1)}/2$$
$$+ \hat{\boldsymbol{\beta}}_\lambda^{(1)\mathrm{T}}\boldsymbol{J}_n^{(11)}(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + n^{-1/2}\sum_{j \in \mathcal{J}^{(1)}} p_\lambda(\hat{\beta}_{\lambda,j})$$

is non-positive. Here, we use the fact that $\sum_{j \in \mathcal{J}^{(1)}} p_\lambda(\hat{\beta}_{\lambda,j})$ reduces to $\lambda\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|_q^q\{1 + o_p(1)\}$ from (C5) and (39) and that $\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})$ is positive definite for sufficiently large n. Accordingly, we have

$$|\hat{\boldsymbol{\beta}}_\lambda^{(1)}|^2 + n^{-1/2}\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|_q^q\{1 + o_p(1)\} = O_p(n^{-1/2}|\hat{\boldsymbol{\beta}}_\lambda^{(1)}|)$$

and thus $\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|_q^q = O_p(|\hat{\boldsymbol{\beta}}_\lambda^{(1)}|)$ since the first term on the left-hand side is nonnegative. Hence, we have

$$P(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1 \tag{40}$$

because $0 < q < 1$ and $\hat{\boldsymbol{\beta}}_\lambda^{(1)} = o_p(1)$. This implies the former in (17). Since $\tilde{\boldsymbol{u}}_n^{(2)}$ is trivially $O_p(1)$, we obtain the latter of (17) from (39) and (40). $\qquad\square$

A.4 Proof of (19) and (20)

Let $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ be the one with $q = 1$ in (9), and let $\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = -\boldsymbol{u}^{\mathrm{T}} \boldsymbol{s}_n + \boldsymbol{u}^{\mathrm{T}} \boldsymbol{J} \boldsymbol{u}/2$ in place of (10). Then, we can obtain $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + o_p(1)$ by taking a Taylor expansion around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\mathbf{0}, \mathbf{0})$. In addition, let $\phi_n(\boldsymbol{u})$ and $\phi(\boldsymbol{u})$ be $\phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ and $\phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger)$ with $q = 1$ in (11), (12), and (13), let $\boldsymbol{u}^\dagger$ be empty vector and $\psi_n(\boldsymbol{u}^\dagger) = \psi(\boldsymbol{u}^\dagger) = 0$, and define $\nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ and $\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger)$ again. Here, note that

$$(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)}) = \underset{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}{\operatorname{argmin}} \nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (n^{1/2} \hat{\boldsymbol{\beta}}_\lambda^{(1)}, n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})).$$

Next, because

$$\begin{aligned}
\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) &= \|\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}\{-\boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)} + (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})\}\|_{\boldsymbol{J}^{(22)}}^2/2 \\
&\quad + \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda\|\boldsymbol{u}^{(1)}\|_1 - \|\boldsymbol{s}_n^{(2)} \\
&\quad - \boldsymbol{p}_\lambda'^{(2)}\|_{\boldsymbol{J}^{(22)-1}}^2/2,
\end{aligned}$$

we see by using $\hat{\boldsymbol{u}}_n^{(1)}$ in (18) that

$$\begin{aligned}
(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}) &= \underset{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}{\operatorname{argmin}} \tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) \\
&= (\hat{\boldsymbol{u}}_n^{(1)}, -\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)} + \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})),
\end{aligned}$$

where we have denoted $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$ by $\|\boldsymbol{x}\|_A^2$ for an appropriate size of matrix A and vector $\boldsymbol{x}$. Now, we apply Lemma 3 and evaluate the right-hand side in (4). In the same way as in (15), it follows that $\Delta_n(\delta)$ converges in probability to 0. Next, the definition of $\tilde{\boldsymbol{u}}_n^{(1)}$ ensures that

$$\boldsymbol{J}^{(1|2)}\tilde{\boldsymbol{u}}_n^{(1)} - \boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda\boldsymbol{\gamma} = \mathbf{0},$$

where $\boldsymbol{\gamma}$ is a $|\mathcal{J}^{(1)}|$-dimensional vector such that $\gamma_j = 1$ when $\hat{u}_{n,j}^{(1)} > 0$, $\gamma_j = -1$ when $\hat{u}_{n,j}^{(1)} < 0$, and $\gamma_j \in [-1, 1]$ when $\hat{u}_{n,j}^{(1)} = 0$. Thus, noting that $\tilde{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{\gamma} = \|\tilde{\boldsymbol{u}}_n^{(1)}\|_1$,

we can write $\tilde{v}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{v}_n(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)})$ as

$$\|\boldsymbol{u}^{(1)} - \tilde{\boldsymbol{u}}_n^{(1)}\|_{\boldsymbol{J}^{(1|2)}}^2/2 + \lambda \sum_{j \in \mathcal{J}^{(1)}} \left(|u_j| - \gamma_j u_j \right)$$

$$+ \|\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}\{-\boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)} + (s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})\}\|_{\boldsymbol{J}^{(22)}}^2/2 \qquad (41)$$

after a simple calculation. Let $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ be unit vectors such that $\boldsymbol{u}^{(1)} = \tilde{\boldsymbol{u}}_n^{(1)} + \zeta \boldsymbol{w}_1$ and $\boldsymbol{u}^{(2)} = \tilde{\boldsymbol{u}}_n^{(2)} + (\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2$, where $0 \leq \zeta \leq \delta$. Then, letting $\rho^{(22)}$ and $\rho^{(1|2)}$ (> 0) be half the smallest eigenvalues of $\boldsymbol{J}^{(22)}$ and $\boldsymbol{J}^{(1|2)}$, respectively, it follows that

$$\Upsilon_n(\delta) \geq \min_{0 \leq \zeta \leq \delta} \left\{ \rho^{(1|2)}\zeta^2 + \rho^{(22)}|(\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2 + \zeta \boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\boldsymbol{w}_1|^2 \right\} > 0$$

because the second term in (41) is nonnegative. Hence, the first term on the right-hand side in (4) converges to 0. In addition, because $(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)})$ is $O_p(1)$ from (39) and $(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)})$ is also $O_p(1)$, the second term on the right-hand side in (4) can be made arbitrarily small by considering a sufficiently large ξ. Thus, we have $|\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| = o_p(1)$, and as a consequence, we obtain (19) and (20).   □

A.5 Proof of (26)

Because $n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \hat{\boldsymbol{u}}_n^{(1)} + o_p(1)$ from Theorem 1, the terms including $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ do not reduce to $o_p(1)$ in this case. Therefore, (24) is expressed as

$$\hat{\boldsymbol{u}}_n^{(1)\mathrm{T}} \left(s_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}s_n^{(2)} \right) + \left(s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1}s_n^{(2)}$$

$$- \hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n/2 - \left(s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)} \left(s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)/2 + o_p(1),$$

and this converges in distribution to

$$\hat{\boldsymbol{u}}^{(1)\mathrm{T}}s^{(1|2)} + \left(s^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1}s^{(2)}$$

$$- \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}/2 - \left(s^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)} \left(s^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)/2.$$

In the same way, (25) is expressed as

$$\hat{\boldsymbol{u}}_n^{(1)\mathrm{T}} \left(\tilde{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\tilde{s}_n^{(2)} \right) + \left(s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1}\tilde{s}_n^{(2)}$$

$$- \hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n/2 - \left(s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)^{\mathrm{T}} \boldsymbol{J}^{(22)} \left(s_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)} \right)/2 + o_p(1),$$

and this converges in distribution to

$$\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\tilde{\boldsymbol{s}}^{(1|2)} + \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)}$$

$$- \hat{\boldsymbol{u}}^{(1)\mathrm{T}} \boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}/2 - \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right)^{\mathrm{T}} \boldsymbol{J}^{(22)} \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right)/2,$$

where $\tilde{\boldsymbol{s}}_n^{(1)}$, $\tilde{\boldsymbol{s}}^{(2)}$, $\tilde{\boldsymbol{s}}^{(1|2)}$, and $\tilde{\boldsymbol{s}}^{(2)}$ are copies of $\boldsymbol{s}_n^{(1)}$, $\boldsymbol{s}^{(2)}$, $\boldsymbol{s}^{(1|2)}$, and $\boldsymbol{s}^{(2)}$, respectively. Thus, we see that

$$z^{\mathrm{limit}} = \hat{\boldsymbol{u}}^{(1)\mathrm{T}} \boldsymbol{s}^{(1|2)} + \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}$$

$$- \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\tilde{\boldsymbol{s}}^{(1|2)} - \left(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}\right)^{\mathrm{T}} \boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)}.$$

Since $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$ are independently distributed according to $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)})$, the asymptotic bias reduces to

$$\mathrm{E}\left[z^{\mathrm{limit}}\right] = \mathrm{E}\left[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}\right] + \mathrm{E}\left[(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}\right].$$

As a result, we obtain (26). □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, F. Csaki (Eds.), *Proceeding of the 2nd international symposium on information theory* (pp. 267–281), Akademiai Kiado.

Andersen, P. K., Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, *10*, 1100–1120.

Beck, A., Teboulle, M. (2009). A fast iterative shrinkage–thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*, 183–202.

Dicker, L., Huang, B., Lin, X. (2012). Variable selection and estimation with the seamless-L_0 penalty. *Statistica Sinica*, *23*, 929–962.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, Y., Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B*, *75*, 531–552.

Frank, L. E., Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*, 109–135.

Hjort, N. L., Pollard, D. (1993). Asymptotics for minimisers of convex processes, arXiv preprint arXiv:1107.3806.

Knight, K., Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*, 1356–1378.

Konishi, S., Kitagawa, G. (2008). *Information criteria and statistical modeling. Springer Series in Statistics*. New York: Springer.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Masuda, H., Shimizu, Y. (2017). Moment convergence in regularized estimation under multiple and mixed-rates asymptotics. *Mathematical Methods of Statistics*, *26*, 81–110.

Mazumder, R., Friedman, J. H., Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, *106*, 1125–1138.

McCullagh, P., Nelder, J. A. (1989). *Generalized linear models, monographs on monographs on statistics and applied probability*. London: Chapman & Hall.

Meinshausen, N., Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society*: *Series B*, *72*, 417–473.

Ninomiya, Y., Kawano, S. (2016). AIC for the LASSO in generalized linear models. *Electronic Journal of Statistics*, *10*, 2537–2560.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, *7*, 186–199.

Radchenko, P. (2005). Reweighting the lasso. In *2005 Proceedings of the American Statistical Association* [*CD-ROM*].

Rockafellar, R. T. (1970). *Convex analysis, Princeton mathematical series*. New Jersey: Princeton University Press.

Rockafellar, R. T. (1976). Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, *1*, 97–116.

Shiryaev, A. N. (1996). *Probability, volume 95 of graduate texts in mathematics* (2nd ed.). New York: Springer.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, *9*, 1135–1151.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*: *Series B, 36*, 111–147.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods*, *7*, 13–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*, 267–288.

Wang, H., Li, R., Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*, 553–568.

Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society*: *Series B*, *71*, 671–683.

Yoshida, N. (2011). Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. *Annals of the Institute of Statistical Mathematics*, *63*, 431–479.

Yuan, M., Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, *94*, 19–35.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.

Zhang, Y., Li, R., Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*, 312–323.

Zou, H., Hastie, T., Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, *35*, 2173–2192.