Supplementary Material for "Regularization Parameter Selections via Generalized Information Criterion" by Yiyun Zhang, Runze Li and Chih-Ling Tsai

In this supplementary note, we study the asymptotic loss efficiency for the generalized linear model (GLIM) under the Kullback-Leibler (KL) loss. To the best of our knowledge, there is no literature studying the asymptotic loss efficiency for GLIM. After a series of attempts, we found that the extension of the asymptotic loss efficiency from the linear regression model to GLIM is a challenging task. For example, the Taylor expansion commonly used in the asymptotic analysis cannot be utilized when a candidate model is not in the neighborhood of the true model. As a result, a general framework for theoretical developments is required. To this end, (a) we first present the asymptotic theory for the GLIM estimator with the given candidate model, which also allows the candidate model to be misspecified. This theory leads us to study the asymptotic bias and variance of the parameter estimator. (b) From the result of asymptotic bias, we next find that the parameter estimator is not consistent when the candidate model is misspecified. Thus, we restrict ourselves to candidate models in the neighborhood of the true model so that the Taylor expansion is applicable. (c) We further demonstrate that the KL loss is asymptotically equivalent to a squared loss for the linear regression model. (d) We finally show the KL asymptotic loss efficiency for the classical AIC-type variable selection criterion and for the AIC-type tuning parameter selector.

1. The Asymptotic Theory of the GLIM Estimator

It is known that when the candidate model is a correct model (i.e., the true model or the overfitted model, see Shao, 1997), under certain regularity conditions, the resulting parameter estimator of GLIM is consistent and follows an asymptotic multivariate normal distribution. In this section, we present the asymptotic theory of the GLIM estimator without assuming that the candidate model is a correct model.

Consider the generalized linear model, whose density function is

$$f(y;\theta,\phi) = \exp\left\{ \left[y\theta - b(\theta)\right]/a(\phi) + c(y,\phi)\right\},$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are suitably chosen functions, θ is the canonical parameter, and ϕ is a scale parameter that is also called the dispersion parameter. Additionally, $a(\cdot)$ is assumed to be positive and $b(\cdot)$ is a second order smooth function with $b''(\theta) > 0$ and $b'''(\theta)$ bounded for every $\theta \in \Theta$. Given θ, denote the mean and variance of y by μ and σ^2. It can be easily shown that $E(y) = \mu = b'(\theta)$ and $Var(y) = \sigma^2 = a(\phi)b''(\theta)$. In general, the canonical parameter θ is related to the systematic parameter η though the pre-specified link function $g(\cdot)$, so that $\eta = g(\mu) = g \circ b'(\theta)$. In this note, we restrict the discussion to canonical link functions. That is, we assume $g^{-1}(\cdot) = b'(\cdot)$ so that $\theta = \eta$. Furthermore, we denote the true canonical parameter by θ_0, where θ_0 lies in a set Θ with bounded support. Moreover, we assume that the canonical parameter θ is a function of covariates $\mathbf{x}$ (i.e., $\theta = \theta(\mathbf{x})$), where $\mathbf{x}$ is fixed. Analogous to classical variable selection criteria (Shao, 1997), our results are still valid for the random $\mathbf{x}$ if we impose additional conditions.

The goal of variable selection is to find a best submodel $\alpha \subset \bar{\alpha}$ that is parsimonious and in which $\theta(\mathbf{x})$ is well approximated by $\mathbf{x}_\alpha \boldsymbol{\beta}_\alpha$ for some parameter $\boldsymbol{\beta}_\alpha$. For the sake of simplicity, we only consider the case where the dispersion parameter ϕ is known, which includes the normal linear regression with known variance, the logistic regression model, and the Poisson log-linear model. It is noteworthy that our results can be carried out when the dispersion

parameter is replaced by its consistent estimator.

For the given candidate model α, the conditional density function of $y|\mathbf{x}$ is

$$f_\alpha(y;\boldsymbol{\beta}_\alpha,\phi)=\exp\left\{\left[y\mathbf{x}_\alpha^T\boldsymbol{\beta}_\alpha-b(\mathbf{x}_\alpha^T\boldsymbol{\beta}_\alpha)\right]/a(\phi)+c(y,\phi)\right\}.$$

Let f_0 be the true (conditional) density function of $y|\mathbf{x}$. Then, the Kullback-Leibler (KL) discrepancy between the true model and the candidate model is (omitting irrelevant terms)

$$\begin{aligned}\rho\left(\theta_0(\mathbf{x}),\mathbf{x}_\alpha^T\boldsymbol{\beta}_\alpha\right) &= E_0\left\{\log\left(\frac{f_0}{f_\alpha}\right)\right\}\\ &= \frac{1}{a(\phi)}\left[b'(\theta_0(\mathbf{x}))(\theta_0(\mathbf{x})-\mathbf{x}_\alpha^T\boldsymbol{\beta}_\alpha)-b(\theta_0(\mathbf{x}))+b(\mathbf{x}_\alpha^T\boldsymbol{\beta}_\alpha)\right], \qquad (1)\end{aligned}$$

where E_0 denotes expectation under the true model.

Based on a sample of n observations, the objective function, $n^{-1}\sum_{i=1}^n\rho\left(\theta_0(\mathbf{x}_i),\mathbf{x}_{\alpha_i}^T\boldsymbol{\beta}_\alpha\right)$, is convex, because its second derivative with respect to $\boldsymbol{\beta}_\alpha$ is $n^{-1}\sum_{i=1}^n b''(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha)\mathbf{x}_{\alpha i}\mathbf{x}_{\alpha i}^T/a(\phi)$, which is a positive definite matrix. This allows us to adopt the approach of Hjort and Pollard (1993) by assuming that there exists a unique optimal parameter $\boldsymbol{\beta}_\alpha^*$ that is the minimizer of the limit of $n^{-1}\sum_{i=1}^n\rho\left(\theta_0(\mathbf{x}_i),\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha\right)$ over $\boldsymbol{\beta}_\alpha$ for each candidate model α. We next introduce some notation from Hjort and Pollard (1993) so that we are able to employ their Theorem 2.3 to establish the asymptotic distribution of the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}_\alpha$, $\hat{\boldsymbol{\beta}}_\alpha^*=\text{argmax}_{\boldsymbol{\beta}_\alpha}\ell(\boldsymbol{\beta}_\alpha)$, where $\ell(\boldsymbol{\beta}_\alpha)=\sum_{i=1}^n\left[\frac{y_i\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha-b(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha)}{a(\phi)}+c(y_i,\phi)\right]$.

Let $g_i(y_i,\boldsymbol{\beta}_\alpha|\mathbf{x}_i)=a(\phi)^{-1}[-y_i\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha+b(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha)]$. Then, the MLE $\hat{\boldsymbol{\beta}}_\alpha^*$ is the minimizer of $\sum_{i=1}^n g_i(y_i,\boldsymbol{\beta}_\alpha|\mathbf{x}_{\alpha i})$, and

$$\begin{aligned}g_i(y_i,\boldsymbol{\beta}_\alpha^*+\mathbf{t}|\mathbf{x}_{\alpha i})-g_i(y_i,\boldsymbol{\beta}_\alpha^*|\mathbf{x}_{\alpha i}) &= \frac{1}{a(\phi)}\left[-y_i\mathbf{x}_{\alpha i}^T\mathbf{t}+b(\mathbf{x}_{\alpha i}^T(\boldsymbol{\beta}_\alpha^*+\mathbf{t}))-b(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)\right]\\ &= \{\delta(\mathbf{x}_{\alpha i})+D_i(y_i|\mathbf{x}_{\alpha i})\}^T\mathbf{t}+R_i(y_i,\mathbf{t}|\mathbf{x}_{\alpha i}),\end{aligned}$$

where $\mathbf{t} = (t_1, \cdots, t_{d_\alpha})^T$,

$$\begin{aligned}
\delta(\mathbf{x}_{\alpha i}) &= a(\phi)^{-1}\left[b'(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*) - b'(\theta(\mathbf{x}_i))\right]\mathbf{x}_{\alpha i}, \\
D_i(y_i|\mathbf{x}_{\alpha i}) &= -a(\phi)^{-1}\left[y_i - b'(\theta(\mathbf{x}_i))\right]\mathbf{x}_{\alpha i}, \\
\text{and } R_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}) &= a(\phi)^{-1}\left[b(\mathbf{x}_{\alpha i}^T(\boldsymbol{\beta}_\alpha^* + \mathbf{t})) - b(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*) - b'(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)\mathbf{x}_{\alpha i}^T\mathbf{t}\right].
\end{aligned}$$

Because $Ey_i = b'(\theta(\mathbf{x}_i))$, it is easy to verify that $E(D_i(y_i|\mathbf{x}_{\alpha i})) = 0$. In addition,

$$Var(D_i(y_i|\mathbf{x}_{\alpha i})) = \mathbf{B}_i(\mathbf{x}_{\alpha i}) = b''(\theta(\mathbf{x}_i))\mathbf{x}_{\alpha i}\mathbf{x}_{\alpha i}^T,$$

$$ER_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}) = \frac{1}{2}\mathbf{t}^T\mathbf{A}_i(\mathbf{x}_{\alpha i})\mathbf{t} + v_{i,0}(\mathbf{t}|\mathbf{x}_{\alpha i}) \text{ and } VarR_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}) = v_i(\mathbf{t}|\mathbf{x}_i),$$

where $\mathbf{A}_i(\mathbf{x}_{\alpha i}) = b''(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)\mathbf{x}_{\alpha i}\mathbf{x}_{\alpha i}^T/a(\phi)$,

$$v_{i,0}(\mathbf{t}|\mathbf{x}_{\alpha i}) = \frac{1}{6a(\phi)}\sum_{j=1}^{d_\alpha}\sum_{k=1}^{d_\alpha}\sum_{l=1}^{d_\alpha} b'''(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)x_{\alpha ij}x_{\alpha ik}x_{\alpha il}\xi_j\xi_k\xi_l,$$

and $0 \le \xi_j \le t_j$ for $j = 1, \cdots, d_\alpha$. For the sake of simplification, we further introduce the following notation:

$$\begin{aligned}
\mathbf{J}_n &= \sum_{i=1}^n \mathbf{A}_i(\mathbf{x}_{\alpha i}) = \frac{1}{a(\phi)}\mathbf{X}_\alpha^T\text{diag}\{b''(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)\}_{i=1}^n\mathbf{X}_\alpha \\
\text{and } \mathbf{K}_n &= \sum_{i=1}^n \mathbf{B}_i(\mathbf{x}_{\alpha i}) = \frac{1}{a(\phi)}\mathbf{X}_\alpha^T\text{diag}\{b''(\theta(\mathbf{x}_i))\}_{i=1}^n\mathbf{X}_\alpha.
\end{aligned}$$

Theorem 1. *Assume that* $\max_{i=1,\ldots,n}\frac{|\mathbf{x}_i|}{\sqrt{n}} \to 0$ *as* $n \to \infty$, $\mathbf{J}_n/n$ *is bounded away from 0, and* $\mathbf{K}_n/n$ *is bounded. Then,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\alpha^* - \boldsymbol{\beta}_\alpha^*) = -(\mathbf{J}_n/n)^{-1}\left\{n^{-1/2}\sum_{i=1}^n \delta(\mathbf{x}_{\alpha i}) + n^{-1/2}\sum_{i=1}^n D(y_i|\mathbf{x}_{\alpha i})\right\} + o_P(1). \quad (2)$$

Furthermore, if $\mathbf{J}_n/n \to \mathbf{J}$ *and* $\mathbf{K}_n/n \to \mathbf{K}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\alpha^* - \boldsymbol{\beta}_\alpha^*) = N\left(-\mathbf{J}^{-1}n^{-1/2}\sum_{i=1}^n \delta(\mathbf{x}_{\alpha i}), \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1}\right) + o_P(1). \quad (3)$$

The proof of Theorem 1 is given in the Appendix.

Let

$$\nabla \ell(\boldsymbol{\beta}_\alpha^*) = \frac{\partial \ell(\boldsymbol{\beta}_\alpha^*)}{\partial \boldsymbol{\beta}_\alpha} = a(\phi)^{-1}\mathbf{X}_\alpha^T\{\mathbf{y} - b'(\mathbf{X}_\alpha\boldsymbol{\beta}_\alpha^*)\},$$

$$\nabla^2 \ell(\boldsymbol{\beta}_\alpha^*) = \frac{\partial^2 \ell(\boldsymbol{\beta}_\alpha^*)}{\partial \boldsymbol{\beta}_\alpha^T \partial \boldsymbol{\beta}_\alpha} = -a(\phi)^{-1}\mathbf{X}_\alpha^T\mathbf{V}_\alpha^T\mathbf{V}_\alpha\mathbf{X}_\alpha,$$

where $\mathbf{V}_\alpha = \text{diag}\{b''(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)^{1/2}, ..., b''(\mathbf{x}_{\alpha n}^T\boldsymbol{\beta}_\alpha^*)^{1/2}\}$. Theorem 1 indicates that, although the asymptotic expansion of the MLE still has the classical form

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\alpha^* - \boldsymbol{\beta}_\alpha^* &= -\left[\nabla^2 \ell(\boldsymbol{\beta}_\alpha^*)\right]^{-1} \nabla \ell(\boldsymbol{\beta}_\alpha^*) + o_P(1/\sqrt{n}) \\ &= \left(\mathbf{X}_\alpha^T\mathbf{V}_\alpha^T\mathbf{V}_\alpha\mathbf{X}_\alpha\right)^{-1}\mathbf{X}_\alpha^T\{\mathbf{y} - b'(\mathbf{X}_\alpha\boldsymbol{\beta}_\alpha^*)\} + o_P(1/\sqrt{n}), \end{aligned} \tag{4}$$

$E\nabla \ell(\boldsymbol{\beta}_\alpha^*) \neq 0$ leads to the bias induced by $\delta(\mathbf{x}_{\alpha i})$.

2. The Set of Candidate Models

To study the asymptotic efficiency of variable (or tuning parameter) selection, we need to employ the Taylor expansion for $b(\hat{\theta}_\alpha^*(\mathbf{x}_i))$ at $\theta = \theta_0(\mathbf{x}_i)$, where $\hat{\theta}_\alpha^*(\mathbf{x}_i) = \mathbf{x}_{\alpha i}^T\hat{\boldsymbol{\beta}}_\alpha^*$ and $\theta_0(\cdot)$ is the true canonical parameter. However, if α is not a correct model, then Theorem 1 shows that $\hat{\boldsymbol{\beta}}_\alpha^*$ is not an asymptotically unbiased estimator of $\boldsymbol{\beta}_\alpha^*$. As a result, the difference between $\hat{\theta}_\alpha^*(\mathbf{x}_i)$ and $\theta_0(\mathbf{x}_i)$ is not vanishing and the Taylor expansion cannot be applied. Therefore, in the rest of this note, we only focus on the set of candidate models given below.

$$\mathcal{C} = \{\alpha : \sup_{1 \leq i \leq n} |\hat{\theta}_\alpha^*(\mathbf{x}_i) - \theta_0(\mathbf{x}_i)| \to 0 \text{ in probability, as } n \to \infty\}. \tag{5}$$

For the model in $\mathcal{C}$, the approximate bias is small so that $\hat{\theta}_\alpha^*(\mathbf{x}_i)$ is within a neighborhood of $\theta_0(\mathbf{x}_i)$ and the Taylor expansion is applicable.

3. Asymptotic Representation of KL Loss

Based on a random sample of size n, the log-likelihood function can be expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_i(y_i; \boldsymbol{\theta}, \phi) = \sum_{i=1}^n \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right],$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)^T$. Let $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\theta}_0 = (\theta_{01}, ..., \theta_{0n})^T$. It can be either $\hat{\boldsymbol{\theta}}^*_\alpha = \mathbf{X}\hat{\boldsymbol{\beta}}^*_\alpha$, the maximum likelihood estimator under the model α, or $\hat{\boldsymbol{\theta}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$, the penalized parameter estimator under the penalized model α_λ. Then, twice the average Kullback-Leibler loss can be written as

$$L_{KL}(\hat{\boldsymbol{\beta}}) = \frac{2}{n}\sum_{i=1}^{n} \rho(\theta_0(\mathbf{x}_i), \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) = \frac{2}{n}E_0\left\{\ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}})\right\}. \tag{6}$$

Note that $\hat{\boldsymbol{\beta}}$ and hence $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are treated as nonrandom in calculating the above expectation. Furthermore, the KL risk is defined as $R_{KL}(\hat{\boldsymbol{\beta}}) = E_0[L_{KL}(\hat{\boldsymbol{\beta}})]$.

The following Lemma shows that, for $\alpha \in \mathcal{C}$, both $\ell(\hat{\boldsymbol{\theta}}^*_\alpha)$ and $L_{KL}(\hat{\boldsymbol{\beta}})$ have asymptotic approximations.

Lemma 1. *For $\alpha \in \mathcal{C}$, $\ell(\hat{\boldsymbol{\theta}}^*_\alpha)$ can be asymptotically expanded as*

$$\tilde{\ell}(\hat{\boldsymbol{\theta}}^*_\alpha) = \frac{1}{a(\phi)}\sum_{i=1}^{n}\left\{y_i(\hat{\theta}^*_{\alpha i} - \theta_{0i}) - \left[b'(\theta_{0i})(\hat{\theta}^*_{\alpha i} - \theta_{0i}) + \frac{1}{2}b''(\theta_{0i})(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2\right]\right\},$$

in the sense that for any bounded weights w_i satisfying $0 < c_1 < w_i < c_2$,

$$\frac{\left|\ell(\hat{\boldsymbol{\theta}}^*_\alpha) - \ell(\boldsymbol{\theta}_0) - \tilde{\ell}(\hat{\boldsymbol{\theta}}^*_\alpha)\right|}{\sum_{i=1}^{n} w_i(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2} \to 0$$

in probability as $n \to \infty$, where c_1 and c_2 are some positive constants, $b'(\theta_{0i}) = E_0(y_i)$, and $b''(\theta_{0i}) = Var_0(y_i)/a(\phi)$ for $i = 1 \cdots, n$. In addition, twice the average KL loss has the expansion

$$L_{KL}(\hat{\boldsymbol{\beta}}^*_{\alpha i}) = \frac{1}{na(\phi)}\left\{\sum_{i=1}^{n} b''(\theta_{0i})^{-1}(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2 + o_P\left(\sum_{i=1}^{n} b''(\theta_{0i})^{-1}(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2\right)\right\},$$

if we consider $w_i = b''(\theta_{0i})^{-1}$.

The proof of Lemma 1 is outlined in the Appendix. Note that although Lemma 1 focuses on the MLE $\hat{\boldsymbol{\theta}}^*_\alpha$, it can be shown that the lemma also holds for any estimator $\hat{\boldsymbol{\theta}}$ that satisfies the condition in (5).

To get further insights from the expansions of $\ell(\hat{\boldsymbol{\theta}})$ and $L_{KL}(\hat{\boldsymbol{\beta}})$, we introduce the following notation. Denote

$$y_i^\dagger = b''(\theta_{0i})^{-1/2}(y_i - b'(\theta_{0i})) + b''(\theta_{0i})^{1/2}\theta_{0i} \quad \text{and} \quad \hat{\theta}_i^\dagger = b''(\theta_{0i})^{1/2}\hat{\theta}_i,$$

(or in vector notation,

$$\mathbf{y}^\dagger = \mathbf{V}_0^{-1}(\mathbf{y} - b'(\boldsymbol{\theta}_0)) + \mathbf{V}_0\boldsymbol{\theta}_0 \quad \text{and} \quad \hat{\boldsymbol{\theta}}^\dagger = \mathbf{V}_0\hat{\boldsymbol{\theta}},$$

where $b'(\boldsymbol{\theta}_0) = (b'(\theta_{01}, \cdots, b'(\theta_{0n})^T$ and $\mathbf{V}_0 = \text{diag}\{b''(\theta_{01})^{1/2}, ..., b''(\theta_{0n})^{1/2}\}$). After algebraic simplification by adding the constant term with respect to $\boldsymbol{\theta}$ (i.e., $\sum_{i=1}^n b''(\theta_{0i})^{-1}(y_i - b'(\theta_{0i}))^2$), we have

$$\ell(\hat{\boldsymbol{\theta}}) = -\frac{1}{2a(\phi)} \parallel \mathbf{y}^\dagger - \hat{\boldsymbol{\theta}}^\dagger \parallel^2 (1 + o_P(1)), \tag{7}$$

which is asymptotically a quadratic function. In addition, using the fact that $\sum_{i=1}^n b''(\theta_{0i})^{-1}(\hat{\theta}_i - \theta_{0i})^2 = \sum_{i=1}^n (\hat{\theta}_i^\dagger - \theta_{0i}^\dagger)^2$, we obtain

$$L_{KL}(\hat{\boldsymbol{\beta}}) = \frac{1}{na(\phi)} \parallel \hat{\boldsymbol{\theta}}^\dagger - \boldsymbol{\theta}_0^\dagger \parallel^2 (1 + o_P(1)), \tag{8}$$

where $\boldsymbol{\theta}_0^\dagger = (\theta_{01}^\dagger, \cdots, \theta_{0n}^\dagger)^T = \mathbf{V}_0\boldsymbol{\theta}_0$. This indicates that the KL loss can be treated asymptotically as a squared loss.

Remark 1 (Normal linear regression). In normal linear regression with known variance, $\mathbf{V}_0$ is an identity matrix and $\boldsymbol{\theta}_0 = b'(\boldsymbol{\theta}_0)$. Hence, $\mathbf{y}^\dagger$ becomes $\mathbf{y}$. Accordingly, the maximization of the log-likelihood function is the same as the minimization of the error sum of squares $\parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2$. Furthermore, twice the average KL loss reduces to $(n\sigma^2)^{-1} \parallel X\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0 \parallel^2$.

4. Asymptotic Loss Efficiency of GLIM

Following Lemma 1, we can express the likelihood function and the KL loss as their corresponding quadratic functions asymptotically. This enables us to adopt the strategies used to develop asymptotic loss efficiency for the least squares setting to GLIM. For the

sake of simplicity, we only consider the leading terms of the right-hand side of equations (7) and (8) in the rest of this section. Because we study the efficiency of selection, our focus is choosing the model α, from all the candidate models, that minimizes the following generalized information criterion,

$$\mathrm{GIC}^*_{\kappa_n}(\alpha) = \frac{1}{n} D(\mathbf{y}; \hat{\boldsymbol{\theta}}^*_\alpha) + \frac{1}{n}\kappa_n d_\alpha, \tag{9}$$

where $D(\mathbf{y}; \hat{\boldsymbol{\theta}}^*_\alpha)$ is the scaled deviance function.

To establish the asymptotic loss efficiency, we need to further restrict the candidate model lying in the set of $\mathcal{D} \subset \mathcal{C}$ such that

$$\sup_{\alpha \in \mathcal{D}} \sup_{1 \le i \le n} |\hat{\theta}^*_{\alpha i} - \theta_{0i}| \to 0,$$

in probability. In other words, we consider the models with MLEs being uniformly close to the truth. Then, we adopt the loss efficiency from Li (1987) to define the asymptotic KL loss efficiency for a classical selection criterion as follows:

$$\frac{L_{KL}(\hat{\boldsymbol{\beta}}^*_{\hat{\alpha}})}{\inf_{\alpha \in \mathcal{D}} L_{KL}(\hat{\boldsymbol{\beta}}^*_\alpha)} \to 1 \text{ } in \text{ } probability, \text{ } as \text{ } n \to \infty. \tag{10}$$

To study efficiency, we next identify the systematic bias of the candidate model α via the KL loss. Let $\mathbf{X}^\dagger_\alpha = \mathbf{V}_0 \mathbf{X}_\alpha$. Then, minimizing $\| \mathbf{y}^\dagger - \mathbf{X}^\dagger_\alpha \boldsymbol{\beta} \|^2$, we obtain the MLE of $\boldsymbol{\beta}$ (in an asymptotical sense), $\hat{\boldsymbol{\beta}}^*_\alpha = (\mathbf{X}^{\dagger T}_\alpha \mathbf{X}^\dagger_\alpha)^{-1} \mathbf{X}^{\dagger T}_\alpha \mathbf{y}^\dagger$. Subsequently, the scaled KL loss is

$$\begin{aligned} a(\phi) L_{KL}(\hat{\boldsymbol{\beta}}^*_\alpha) = \frac{1}{n} \| \hat{\boldsymbol{\theta}}^\dagger_\alpha - \boldsymbol{\theta}^\dagger_0 \|^2 &= \frac{1}{n} \| \mathbf{H}^\dagger_\alpha \mathbf{y}^\dagger - \mathbf{H}^\dagger_\alpha \boldsymbol{\theta}^\dagger_0 \|^2 + \frac{1}{n} \| \mathbf{H}^\dagger_\alpha \boldsymbol{\theta}^\dagger_0 - \boldsymbol{\theta}^\dagger_0 \|^2 \\ &= \frac{\boldsymbol{\epsilon}^{\dagger T} \mathbf{H}^\dagger_\alpha \boldsymbol{\epsilon}^\dagger}{n} + \Delta^\dagger_\alpha, \end{aligned} \tag{11}$$

where $\boldsymbol{\epsilon}^\dagger = \mathbf{y}^\dagger - \boldsymbol{\theta}^\dagger_0 = \mathbf{V}_0^{-1}(\mathbf{y} - b'(\boldsymbol{\theta}_0))$ is the "standardized" error term whose components have mean 0 and variance $a(\phi)$, $\mathbf{H}^\dagger_\alpha = \mathbf{X}^\dagger_\alpha (\mathbf{X}^{\dagger T}_\alpha \mathbf{X}^\dagger_\alpha)^{-1} \mathbf{X}^{\dagger T}_\alpha$, and $\Delta^\dagger_\alpha = \frac{1}{n} \| \mathbf{H}^\dagger_\alpha \boldsymbol{\theta}^\dagger_0 - \boldsymbol{\theta}^\dagger_0 \|^2$. The quantity $\Delta^\dagger_\alpha$ is the distance between the true $\boldsymbol{\theta}^\dagger_0$ and its projection on the space spanned by $\mathbf{X}^\dagger_\alpha$, which corresponds to the systematic bias of model α. Moreover, taking the expectation

from both sides of (11) and using the fact that $E\left(\boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger\right) = tr\mathbf{H}_\alpha^\dagger E\left(\boldsymbol{\epsilon}^\dagger\boldsymbol{\epsilon}^{\dagger T}\right) = d_\alpha a(\phi)$, we have

$$a(\phi)R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*) = \Delta_\alpha^\dagger + \frac{a(\phi)d_\alpha}{n}. \tag{12}$$

Finally, we adopt strategies from Li (1987) to show the asymptotic loss efficiency of the classical AIC-type variable selection (i.e., the $\mathrm{GIC}_{\kappa_n}^*$ with $\kappa_n \to 2$ in (9)), which includes AIC. To this end, we also consider the following technical conditions from Li (1987).

(A1) For any candidate model $\alpha \in \mathcal{A}$, the largest eigenvalue of $\frac{1}{n}\mathbf{X}_\alpha^{\dagger T}\mathbf{X}_\alpha^\dagger$ is bounded uniformly by some finite number.

(A2) For $i = 1, \cdots, n$, $E|y_i|^{4q} < \infty$ for some integer q.

(A3) For $\alpha \in \mathcal{D}$, the risk of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_\alpha^*$ satisfies

$$\sum_{\alpha\in\mathcal{D}}[nR(\hat{\boldsymbol{\beta}}_\alpha^*)]^{-q} \to 0. \tag{13}$$

Theorem 2. *Assume conditions (A1)–(A3) hold. Then $\hat{\alpha}$, the model selected from $\mathcal{D}$ by $GIC_{\kappa_n}^*$ with $\kappa_n \to 2$, is asymptotically loss efficient in the sense of (10).*

The proof is given in the Appendix.

After obtaining the asymptotic efficiency of the classical AIC variable selection criterion, we further study the asymptotic loss efficiency of the nonconcave penalized likelihood estimator with the AIC tuning parameter selector. Specifically, we choose the tuning parameter λ from the range $[0, \lambda_{\max}]$ by minimizing

$$\mathrm{GIC}_{\kappa_n}(\lambda) = \frac{1}{n}D(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda) + \frac{1}{n}\kappa_n d_{\alpha_\lambda}, \tag{14}$$

where $\hat{\boldsymbol{\mu}}_\lambda = (g^{-1}(\mathbf{x}_1^T\hat{\boldsymbol{\beta}}_\lambda), \cdots, g^{-1}(\mathbf{x}_n^T\hat{\boldsymbol{\beta}}_\lambda))^T$. Analogous to the classical variable selection, we define the asymptotic KL loss efficiency for the tuning parameter selector as follows:

$$\frac{L_{KL}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_{\lambda\in\Lambda} L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)} \to 1 \text{ in } \textit{probability}, \text{ as } n \to \infty, \tag{15}$$

where $\hat{\lambda}$ is the minimizer of (14) and $\Lambda = \{\lambda \in [0, \lambda_{\max}] : \alpha_\lambda \in \mathcal{D}\}$.

To show efficiency, we need an additional condition given below to regularize the penalized estimator.

(A4) Let $\flat = (\flat_1, ..., \flat_d)^T$, where $\flat_j = p'_\lambda(|\hat{\beta}_{\lambda j}|)\text{sgn}(\hat{\beta}_{\lambda j})$ for all j such that $|\hat{\beta}_{\lambda j}| > 0$, and $\flat_j = 0$ otherwise, and $\hat{\beta}_{\lambda j}$ is the j-th component of the penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$. In addition, we assume that, in probability,

$$\sup_{\lambda \in \Lambda} \frac{\| \flat \|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \to 0,$$

where $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ obtained from the model α_λ.

In the following theorem, we show the asymptotic loss efficiency of the AIC-type tuning parameter selector in GLIM.

Theorem 3. *Assume conditions (A1)-(A4) hold. Then the penalized estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ with $\hat{\lambda}$ selected by minimizing the GIC_{κ_n} criterion with $\kappa_n \to 2$ is asymptotically loss efficient in the sense of (15).*

The proof is given in the Appendix.

Appendix: Technical Proofs

Proof of Theorem 1. Because $b'''(\cdot)$ is assumed to be bounded and $\max_i |\mathbf{x}_i|/\sqrt{n} \to 0$, we have that

$$\sum_{i=1}^{n} v_{i,0}(\mathbf{s}/\sqrt{n}|\mathbf{x}_i) \xrightarrow{P} 0, \quad \text{for each } \mathbf{s}.$$

In addition, $v_i(\mathbf{s}/\sqrt{n}|\mathbf{x}_i) = 0$. These together with the conditions that $\mathbf{J}_n/n$ is bounded away from 0 and both $\mathbf{K}_n/n$ and $\mathbf{L}_n/n$ are bounded, allow us to directly apply Theorem 2.3 from Hjort and Pollard (1993) and conclude (2). Then, Applying the Central Limit Theorem, (3) follows immediately from (2). □

Proof of Lemma 1. At each θ_{0i} $(i = 1, \cdots, n)$, we have the expansion

$$b(\hat{\theta}^*_{\alpha i}) = b(\theta_{0i}) + b'(\theta_{0i})(\hat{\theta}^*_{\alpha i} - \theta_{0i}) + \frac{1}{2}b''(\theta_{0i})(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2 + \frac{1}{6}b'''(\zeta_i)(\hat{\theta}^*_{\alpha i} - \theta_{0i})^3,$$

for some ζ_i such that $|\zeta_i - \theta_{0i}| < |\hat{\theta}^*_{\alpha i} - \theta_{0i}|$. Note that ζ_i is in a neighborhood of θ_{0i} which is an interior point of Θ. Therefore $|b'''(\zeta_i)| < K$ for some constant $K > 0$ by assumption. Hence,

$$\frac{\left|\ell(\hat{\boldsymbol{\theta}}^*_\alpha) - \ell(\boldsymbol{\theta}_0) - \tilde{\ell}(\hat{\boldsymbol{\theta}}^*_\alpha)\right|}{\sum_{i=1}^n w_i(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2} = \frac{\left|\sum_{i=1}^n \frac{1}{6} b'''(\zeta_i)(\hat{\theta}^*_{\alpha i} - \theta_{0i})^3\right|}{\sum_{i=1}^n w_i(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2} \leq \frac{K}{6c_1} \sup_{1 \leq i \leq n} |\hat{\theta}^*_{\alpha i} - \theta_{0i}|.$$

Because $\sup_{1 \leq i \leq n} |\hat{\theta}^*_{\alpha i} - \theta_{0i}| \to 0$ in probability as $n \to \infty$, the right-hand side of the above equation goes to zero in probability as $n \to \infty$. This completes the proof. □

Before proving Theorems 2 and 3, we introduce the following lemma so that we only need to show a variable selection criterion is asymptotically (uniformly) equivalent to the KL loss to establish the KL loss efficiency. It is applicable to both classical variable and tuning parameter selections. For the sake of simplicity, we state the lemma given below via λ and Λ, while it still holds if they are replaced by α and $\mathcal{D}$, respectively. In addition, we denote C as a generic constant number in the proofs of Theorems 2 and 3.

Lemma 2. *Suppose $L(\lambda) > 0$ for all $\lambda \in \Lambda$. Assume $C(\lambda) = L(\lambda) + r(\lambda)$, where*

$$\sup_{\lambda \in \Lambda} \left|\frac{r(\lambda)}{L(\lambda)}\right| \to 0, \quad \textit{in probability.}$$

In addition, let $\hat{\lambda} = arginf_{\lambda \in \Lambda} C(\lambda)$, and assume $L(\hat{\lambda})$ is bounded. Then, we have

$$\frac{L(\hat{\lambda})}{\inf_{\lambda \in \Lambda} L(\lambda)} \to 1, \quad \textit{in probability.}$$

Proof. With probability tending to 1,

$$L(\hat{\lambda}) = C(\hat{\lambda}) - \frac{r(\hat{\lambda})}{L(\hat{\lambda})} L(\hat{\lambda}) \leq C(\lambda) + \left|\frac{r(\hat{\lambda})}{L(\hat{\lambda})}\right| L(\hat{\lambda}), \text{ for any } \lambda \in \Lambda.$$

Taking $\inf_{\lambda \in \Lambda}$ of the right-hand side of the above equation, we have

$$\begin{aligned} \inf_{\lambda \in \Lambda} L(\lambda) \leq L(\hat{\lambda}) &\leq \inf_{\lambda \in \Lambda} \left\{ L(\lambda) \left[1 + \frac{r(\lambda)}{L(\lambda)}\right] \right\} + \frac{r(\hat{\lambda})}{L(\hat{\lambda})} L(\hat{\lambda}) \\ &\leq \inf_{\lambda \in \Lambda} L(\lambda) \left[1 + \sup_{\lambda \in \Lambda} \left|\frac{r(\lambda)}{L(\lambda)}\right|\right] + \left|\frac{r(\hat{\lambda})}{L(\hat{\lambda})}\right| L(\hat{\lambda}). \end{aligned}$$

From the assumption in the lemma, the right-hand side of the above equation goes to $\inf_{\lambda \in \Lambda} L(\lambda)$ as n goes to ∞. This completes the proof. □

Proof of Theorem 2. To prove this theorem, we adapt the techniques used in Theorem 2.1 of Li (1987) by noting that the components of $\boldsymbol{\epsilon}^\dagger = \mathbf{V}_0^{-1}(\mathbf{y} - b'(\boldsymbol{\theta}_0))$ are independent with mean 0 and equal variances $a(\phi)$, although they are not identically distributed. Applying Lemma 1, we obtain the following approximation of $\mathrm{GIC}_2^*(\alpha)$:

$$\mathrm{GIC}_2^*(\alpha) = -\frac{2}{n}\ell(\hat{\boldsymbol{\beta}}_\alpha^*) + \frac{2d_\alpha}{n} \approx \frac{\parallel \mathbf{y}^\dagger - \hat{\boldsymbol{\theta}}_\alpha^\dagger \parallel^2}{na(\phi)} + \frac{2d_\alpha}{n}$$

$$= \quad L_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*) + \frac{1}{a(\phi)}\left\{ \frac{\parallel \boldsymbol{\epsilon}^\dagger \parallel^2}{n} + \frac{2\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger}{n} + \frac{2(a(\phi)d_\alpha - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger)}{n} + \frac{(\kappa_n - 2)a(\phi)d_\alpha}{n} \right\} \quad (16)$$

Considering $\mathrm{GIC}_2^*(\alpha)$, $L_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)$, and the second term in (16) as $C(\lambda)$, $L(\lambda)$, and $r(\lambda)$ in Lemma 2 respectively, we can then apply Lemma 2 to show that Theorem 2 holds if we demonstrate that

$$\sup_{\alpha\in\mathcal{D}} \frac{\left|\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger\right|}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} \to 0, \quad (17)$$

$$\sup_{\alpha\in\mathcal{D}} \frac{\left|a(\phi)d_\alpha - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger\right|}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} \to 0, \quad (18)$$

$$\sup_{\alpha\in\mathcal{D}} \left| \frac{(\kappa_n - 2)d_\alpha}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} \right| \to 0, \quad (19)$$

and

$$\sup_{\alpha\in\mathcal{D}} \left| \frac{L_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)}{R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} - 1 \right| \to 0. \quad (20)$$

To prove (17), we employ Chebyshev's inequality to obtain

$$P\left\{ \sup_{\alpha\in\mathcal{D}} \frac{\left|\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger\right|}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} > \delta \right\} \leq \sum_{\alpha\in\mathcal{D}} \frac{E(\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger)^{2q}}{n^{2q}\delta^{2q}R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)^{2q}}.$$

Subsequently, applying Theorem 2 of Whittle (1960), the right-hand side of the above equation is no greater than

$$C\delta^{-2q}\sum_{\alpha\in\mathcal{D}} \frac{\parallel (\mathbf{I} - \mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger \parallel^{2q}}{n^{2q}R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)^{2q}} \leq C\delta^{-2q}\sum_{\alpha\in\mathcal{D}} \left[nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)\right]^{-q},$$

which goes to zero by condition (A3). Analogously, using Theorem 2 of Whittle (1960) and expansion (12), equation (18) can be shown by noting that

$$E(a(\phi)d_\alpha - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger)^{2q} \leq Cd_\alpha^q \leq \frac{CR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)^q}{n^q}.$$

Next, equation (12) and the assumption that $\kappa_n \to 2$ together lead to equation (19). Finally, from the expansion of (11), equation (20) can be shown in the same manner as (18). □

Before proving Theorem 3, we establish the following two lemmas. Lemma 3 evaluates the difference between a penalized estimator and its corresponding least squares estimator, while Lemma 4 demonstrates that their resulting losses are asymptotically equivalent.

Lemma 3. *Let* $\hat{\boldsymbol{\theta}}_\lambda^\dagger = \mathbf{X}^\dagger \hat{\boldsymbol{\beta}}_\lambda$ *and* $\hat{\boldsymbol{\theta}}_{\alpha_\lambda}^{\dagger *} = \mathbf{X}^\dagger \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*$. *Under condition (A1),*

$$\| \hat{\boldsymbol{\theta}}_\lambda^\dagger - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}^{\dagger *} \|^2 \leq nC \| \natural \|^2,$$

where C *is a constant number and* $\natural$ *is defined in condition (A4).*

Proof. Without loss of generality, we assume that the first d_{α_λ} components of $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*$ are nonzero, and denote them by $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ and $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}$, respectively. Thus, $\hat{\boldsymbol{\theta}}_\lambda^\dagger = \mathbf{X}^\dagger \hat{\boldsymbol{\beta}}_\lambda = \mathbf{X}_{\alpha_\lambda}^\dagger \hat{\boldsymbol{\beta}}_\lambda^{(1)}$ and $\hat{\boldsymbol{\theta}}_{\alpha_\lambda}^\dagger = \mathbf{X}^\dagger \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^* = \mathbf{X}_{\alpha_\lambda}^\dagger \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}$. From the expansion of the log-likelihood function in (7), the resulting penalized log-likelihood function $Q(\boldsymbol{\beta}) \propto -\frac{1}{2n} \| \mathbf{y}^\dagger - \mathbf{X}^\dagger \boldsymbol{\beta} \|^2 - a(\phi) \sum_{j=1}^n p_\lambda(|\beta_j|)$ by ignoring a constant with respect to $\boldsymbol{\beta}$. From the proofs of Theorems 1 and 2 in Fan and Li (2001), with probability tending to 1, we have that $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ is the solution of the following equation,

$$\frac{1}{n} \mathbf{X}_{\alpha_\lambda}^{\dagger T} \left(\mathbf{y}^\dagger - \mathbf{X}_{\alpha_\lambda}^\dagger \boldsymbol{\beta}_\lambda^{(1)} \right) + a(\phi) \natural^{(1)} = \mathbf{0},$$

where $\natural^{(1)}$ is the subvector of $\natural$ that corresponds to $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$. Accordingly,

$$\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \left(\mathbf{X}_{\alpha_\lambda}^{\dagger T} \mathbf{X}_{\alpha_\lambda}^\dagger \right)^{-1} \mathbf{X}_{\alpha_\lambda}^{\dagger T} \mathbf{y}^\dagger + \left(\frac{1}{n} \mathbf{X}_{\alpha_\lambda}^{\dagger T} \mathbf{X}_{\alpha_\lambda}^\dagger \right)^{-1} a(\phi) \natural^{(1)} = \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)} + \left(\frac{1}{n} \mathbf{X}_{\alpha_\lambda}^{\dagger T} \mathbf{X}_{\alpha_\lambda}^\dagger \right)^{-1} a(\phi) \natural^{(1)}.$$

In addition, the eigenvalues of $\left(\frac{1}{n} \mathbf{X}_{\alpha_\lambda}^{\dagger T} \mathbf{X}_{\alpha_\lambda}^\dagger \right)^{-1}$ are bounded under condition (A1). Hence,

$$\| \hat{\boldsymbol{\theta}}_\lambda^\dagger - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}^{\dagger *} \|^2 = \| \mathbf{X}_{\alpha_\lambda}^\dagger (\hat{\boldsymbol{\beta}}_\lambda^{(1)} - \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}) \|^2 = n a(\phi)^2 \natural^{(1)T} \left(\frac{1}{n} \mathbf{X}_{\alpha_\lambda}^{\dagger T} \mathbf{X}_{\alpha_\lambda}^\dagger \right)^{-1} \natural^{(1)} \leq nC \| \natural \|^2 .$$

This completes the proof.

Lemma 4. *If conditions (A1)—(A4) hold, then*

$$\sup_{\lambda \in \Lambda} \left| \frac{L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*)} - 1 \right| \to 0$$

in probability, as $n \to \infty$.

Proof. After algebraic simplification, we have

$$L(\hat{\boldsymbol{\beta}}_\lambda) - L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}) = \frac{\| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2}{na(\phi)} + \frac{2(\boldsymbol{\theta}^{\dagger}_0 - \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda})^T(\hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda})}{na(\phi)} = I_1 + I_2.$$

Under conditions (A1)-(A3), we know that (20) holds. This, together with condition (A4) and Lemma 3, implies

$$\sup_{\lambda\in\Lambda} \left| \frac{I_1}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \right| = \sup_{\lambda\in\Lambda} \left\{ \frac{\| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2}{na(\phi)R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} - \frac{\| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2}{na(\phi)L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \left[\frac{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} - 1 \right] \right\} \to 0.$$

Applying the Cauchy-Schwarz inequality, we then obtain

$$I_2 \le \frac{2 \| \boldsymbol{\theta}^{\dagger}_0 - \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} \| \cdot \| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|}{n} = 2\sqrt{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \cdot \frac{1}{\sqrt{n}} \| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \| .$$

As a result, $\sup_{\lambda\in\Lambda} \left| \frac{I_2}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \right| \to 0$, and Lemma 4 follows immediately.

Proof of Theorem 3. To show the asymptotic efficiency of the AIC-type selector, we first examine the relative difference between $\mathrm{GIC}_{\kappa_n}(\lambda)$ and $L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$ asymptotically. It is noteworthy that, ignoring a constant,

$$\begin{aligned}
\mathrm{GIC}_{\kappa_n}(\lambda) &\approx \frac{\| \mathbf{y}^{\dagger} - \mathbf{X}^{\dagger}\hat{\boldsymbol{\beta}}_\lambda \|^2}{na(\phi)} + \frac{\kappa_n d_{\alpha_\lambda}}{n} \\
&= \frac{\| \mathbf{y}^{\dagger} - \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} \|^2}{na(\phi)} + \frac{\| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2}{na(\phi)} + \frac{\kappa_n d_{\alpha_\lambda}}{n} \\
&= L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) + \left[L_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) - L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) \right] + \frac{1}{a(\phi)} \left\{ \frac{\| \boldsymbol{\epsilon}^{\dagger} \|^2}{n} + \frac{1}{n} \| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2 \right. \\
&\quad \left. + \frac{2}{n} \boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}^{\dagger}_{\alpha_\lambda})\boldsymbol{\theta}^{\dagger}_0 + \frac{2}{n} \left(a(\phi)d_{\alpha_\lambda} - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}^{\dagger}_{\alpha_\lambda}\boldsymbol{\epsilon}^{\dagger} \right) + \frac{1}{n}(\kappa_n - 2)a(\phi)d_{\alpha_\lambda} \right\}.
\end{aligned}$$

Let $J_1 = L_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) - L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$, $J_2 = \| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2 / n$, $J_3 = 2\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}^{\dagger}_{\alpha_\lambda})\boldsymbol{\theta}^{\dagger}_0/n$, $J_4 = 2(a(\phi)d_{\alpha_\lambda} - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}^{\dagger}_{\alpha_\lambda}\boldsymbol{\epsilon}^{\dagger})/n$, and $J_5 = (\kappa_n - 2)a(\phi)d_{\alpha_\lambda}/n$. By similar arguments used in the proof of Theorem 2, we obtain that, in probability,

$$\sup_{\lambda\in\Lambda} \left| \frac{J_j}{L_{KL}(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \right| \to 0, \quad \text{for } j = 1, ..., 5.$$

Then by Lemma 4, the above equations also hold if $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$ is replaced by $\hat{\boldsymbol{\beta}}_\lambda$. These imply that, ignoring a constant with respect to $\hat{\boldsymbol{\beta}}_\lambda$, the difference between $\mathrm{GIC}_{\kappa_n}(\lambda)$ and $L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$ is negligible in comparison to $L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$ when $\kappa_n \to 2$. We next apply Lemma 2, which completes the proof. □

References

Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, *96*, 1348-1360.

Hjort, N., and Pollard, D. (1993). Asymptotics for Minimisers of Convex Processes. *http://www.stat.yale.edu/~pollard/Papers/convex.pdf*.

Li, K.-C. (1987). Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, *15*, 958-975.

Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, *7*, 221-264.

Whittle, P. (1960). Bounds For the Moments of Linear and Quadratic Forms in Independent Variables. *Theory of Probability and Its Applications*, *5*, 302-305.