SCIENCE CHINA
Mathematics

• ARTICLES •

April 2019 Vol. 62 No. 4: 751–770
https://doi.org/10.1007/s11425-017-9189-9

# Consistent tuning parameter selection in high-dimensional group-penalized regression

Yaguang Li, Yaohua Wu & Baisuo Jin*

*School of Management, University of Science and Technology of China, Hefei* 230026*, China*

*Email: liyg@mail.ustc.edu.cn, wuyh@ustc.edu.cn, jbs@ustc.edu.cn*

Received March 12, 2017; accepted September 26, 2017; published online April 27, 2018

**Abstract** Various forms of penalized estimators with good statistical and computational properties have been proposed for variable selection respecting the grouping structure in the variables. The attractive properties of these shrinkage and selection estimators, however, depend critically on the choice of the tuning parameter. One method for choosing the tuning parameter is via information criteria, such as the Bayesian information criterion (BIC). In this paper, we consider the problem of consistent tuning parameter selection in high dimensional generalized linear regression with grouping structures. We extend the results of the extended regularized information criterion (ERIC) to group selection methods involving concave penalties and then investigate the selection consistency with diverging variables in each group. Moreover, we show that the ERIC-type selector enables consistent identification of the true model and that the resulting estimator possesses the oracle property even when the number of group is much larger than the sample size. Simulations show that the ERIC-type selector can significantly outperform the BIC and cross-validation selectors when choosing true grouped variables, and an empirical example is given to illustrate its use.

**Keywords** Bayesian information criterion, group selection, penalized likelihood, regularization parameter, ultra-high dimensionality

**MSC(2010)** 62H12, 62J12

**Citation:** Li Y G, Wu Y H, Jin B S. Consistent tuning parameter selection in high-dimensional group-penalized regression. Sci China Math, 2019, 62: 751–770, https://doi.org/10.1007/s11425-017-9189-9

## 1 Introduction

Grouping structures arise naturally in many statistical modeling problems. Several methods have been proposed for variable selection that respect grouping structures in variables. Yuan and Lin [20] proposed the group LASSO as a natural extension of the LASSO to take into account the grouping structure of the predictors. Meier et al. [15] further extended the group LASSO to logistic regression. While the group LASSO has many attractive properties, such as sparsity and estimation consistency (see [11,19]), it is not selection-consistent in general. To reduce bias and gain selection consistency, various variable selection methods have been proposed as alternatives of the group LASSO, including the adaptive group LASSO penalty (see [16,23]), the group smoothly clipped absolute deviation penalty (SCAD [18]), and the group minimax concave penalty (MCP, see [10]). These methods can identify the true model consistently, and

* Corresponding author

 

the resulting estimator can be as efficient as an oracle. To employ the group-penalized likelihood in regression analysis, we should first hurdle the computational challenge associated with estimating the group-penalized likelihood, especially for nonconvex penalties. A group least-angle regression (LARS) algorithm can be used for the adaptive group LASSO (see [16, 23]). With the aid of a local linear approximation (LLA) algorithm (see [24]), group LARS can be adopted to solve optimization problems of nonconcave penalized likelihood functions with a group structure. Block coordinate descent algorithms for fitting group-penalized models have recently been shown to be competitive particularly in high-dimensional settings (see [1, 8]). However, the above computational procedures rely on the appropriate selection of the regularization parameter, which is the primary focus of our paper.

A consistent criterion identifies the true model with a probability that approaches 1 in large samples when a set of candidate models contains the true model. Several modifications to the criteria BIC have been proposed to select the tuning parameter consistently in high-dimensional settings (see [2, 7, 21]). Wang et al. [17] considered tuning parameter selection with diverging dimensionality and a modified BIC criterion. However, their analysis was confined to the penalized least-squares method, and the dimensionality of covariates was not allowed to exceed the sample size $n$. Zhang and Shen [22] and Kim et al. [13] proposed different penalty terms to handle settings where the number of true predictors was unbounded under linear regression. Recently, Hui et al. [12] proposed an extended regularized information criterion (ERIC) to select the tuning parameter in adaptive LASSO regression; this selector accounted for the effect of the Laplace prior on coefficients not shrunken to zero. Fan and Tang [7] also accommodated tuning parameter selection for general penalized likelihood methods when the dimensionality grows exponentially with the sample size $n$. Despite their clear merits, however, all of the aforementioned works focus on consistently selecting covariates without a group structure. This gap in the research motivated us to study the issues of regularization parameter selection for penalized likelihood-based models with different group-penalized functions. Very recently, Gao and Carroll [9] established the selection consistency of BIC-type criteria for unbounded true predictors under a broad likelihood setting, including generalized linear models (GLIM, see [14]) as a special case. While this criterion is also used in group penalization, in the present work, we establish theoretical consistency independently by comparing the selection performance of several methods in simulations.

In this paper, we adapt the ERIC selector [12] for choosing regularization parameters in group-penalized likelihood functions. We relate the proposed criterion to the adaptive group LASSO and the group concave penalized likelihood methodology with the GLIM structure. When the true model is among a set of candidate models, we show that the ERIC tuning parameter selector enables us to identify the true group consistently with a diverging true number of groups. Moreover, in the ultra-high-dimensional situation, we also establish group selection consistency for GLIM. Our theoretical investigations, numerical implementations via simulations, and data analysis illustrate that the approach proposed can be significantly superior to BIC in GLIM.

The rest of the paper is organized as follows. Section 2 adapts the ERIC-type selector under a general group-penalized likelihood setting. Section 3 studies the consistency property of the adapted ERIC for generalized linear models with both adaptive group LASSO and group concave penalties, and Section 4 investigates the selection consistency in ultra-high-dimensional GLIM. Monte Carlo simulations are presented in Section 5 to illustrate the use of the extended regularization parameter selectors and an empirical example is given in Section 6. Section 7 provides a discussion of our findings, and technical proofs are given in Appendix A.

## 2 Group-penalized likelihood functions

### 2.1 Group-penalized estimators and penalty functions

We consider group variable selection in GLIM. Let $\{(\boldsymbol{X}_i, y_i); i = 1, \ldots, n\}$ be a sample of independent and identically distributed observations, where $y_i$ is a univariate response, $\boldsymbol{X}_i = (\boldsymbol{x}_{i,1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{i,g_n}^{\mathrm{T}})^{\mathrm{T}}$ is a $p_n$-dimensional vector of covariates with $g_n$ groups of predictors, and $\boldsymbol{x}_{i,j} = (x_{i,j}^1, \ldots, x_{i,j}^{d_j})$ is the $d_j$-

dimensional sub-covariate vector representing the $j$-th group, $j = 1, \ldots, g$. This relation means there are $d_j$ variables in the $j$-th group. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$, $\boldsymbol{X}_j = (\boldsymbol{x}_{1,j}^{\mathrm{T}}, \ldots, \boldsymbol{x}_{n,j}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n \times d_j}$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{g_n})$. The number of covariates is allowed to grow with the sample size. The conditional density of $y_i$ given $\boldsymbol{X}_i$ is assumed to come from the exponential family of distributions taking the canonical form

$$f(y_i \,|\, \boldsymbol{X}_i, \beta, \phi) = \exp\left(\frac{1}{\phi}(y_i\theta_i - b(\theta_i)) + c(y_i, \phi)\right)$$

for suitably chosen functions $b(\cdot)$ and $c(\cdot, \cdot)$, where $\theta = \boldsymbol{X}\boldsymbol{\beta}$ is called the canonical parameter and where $\boldsymbol{\beta} = (\beta_1^{\mathrm{T}}, \ldots, \beta_{g_n}^{\mathrm{T}})^{\mathrm{T}}$ is a $p_n = \sum_{j=1}^{g_n} d_j$-dimensional coefficient vector with $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,d_j})^{\mathrm{T}}$ being the $d_j$-dimensional sub-coefficient vector corresponding to the $j$-th group. In addition, $E(\boldsymbol{y} \,|\, \boldsymbol{X}) = b(\theta) = \mu$. We use the canonical link function, $g(\mu) = \theta$, with the nuisance parameter $\phi$ either known in advance or requiring estimation.

In this paper, we consider group-penalized methods that are usually used to simultaneously accomplish group selection and estimation by maximizing the penalized log-likelihood function given by,

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{y} \,|\, \boldsymbol{\beta}) - \sum_{j=1}^{g_n} \rho(\|\beta_j\|; \sqrt{d_j}\lambda, a), \tag{2.1}$$

where $\ell(\boldsymbol{y} \,|\, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i \,|\, \boldsymbol{X}_i)$, and $\rho(\cdot\,; \lambda, a)$ is a penalty function indexed by the penalty parameters $\lambda$ and $a$ that control the tradeoff between the log-likelihood function and the penalty function. The penalty parameter $\sqrt{d_j}\lambda$ accounts for the group size such that large-sized groups and small-sized groups are fairly penalized.

The penalty function in (2.1) includes many choices. For $\rho(t; \lambda) = \lambda|t|$, the group LASSO penalty (see [15, 20]) can be written as $\rho(\|\beta_j\|; \sqrt{d_j}\lambda) = \lambda\sqrt{d_j}\|\beta_j\|$. Let $\tilde{\boldsymbol{\beta}}$ be the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$. Note that $\tilde{\boldsymbol{\beta}}$ is, under general regularity conditions, well-defined provided $p_n < n$. The adaptive group LASSO estimator is given by

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\max_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{y} \,|\, \boldsymbol{\beta}) - \lambda \sum_{j=1}^{g_n} \sqrt{d_j}\tilde{w}_j \|\beta_j\| \right\}, \tag{2.2}$$

where $\tilde{w}_j = 1/\|\tilde{\beta}_j\|^{\gamma}$ with $\gamma > 0$ as the power parameter. The SCAD penalty is

$$\rho(t; \lambda, a) = \lambda \int_0^{|t|} \min\{1, (a - x/\lambda)_+/(a-1)\}dx, \quad a > 2.$$

The MCP penalty has the form $\rho(t; \lambda, a) = \lambda \int_0^{|t|}(1 - x/(a\lambda))_+ dx$, $a > 1$. These penalties are nearly unbiased and more aggressive than others in enforcing a sparser solution. Using a composite of these penalties and an $\ell_2$ norm of the coefficients in each group, the 2-norm group concave estimator is given by

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\max_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{y} \,|\, \boldsymbol{\beta}) - n \sum_{j=1}^{g_n} \rho(\|\boldsymbol{\beta}_j\|; \sqrt{d_j}\lambda, a) \right\}. \tag{2.3}$$

Moreover, when $a \to \infty$, both the group SCAD and group MCP are simplified to the group LASSO (see [20]), which is defined in (2.2).

### 2.2 High dimensional group-penalized information criteria

Under a general likelihood-based framework, the BIC and CV methods can be used to choose tuning parameters $\lambda$ in group selection. Denote $\mathrm{BIC}(\lambda) = -2\ell(\boldsymbol{y} \,|\, \hat{\boldsymbol{\beta}}_\lambda) + \sum_{j \in \alpha_\lambda} d_j \log(n)$, and $\mathrm{CV}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} -2\ell^{(-i)}(\boldsymbol{y} \,|\, \hat{\boldsymbol{\beta}}_\lambda)$, where $\ell^{(-i)}$ is the log-likelihood estimation discounting the $i$-th part of the data within an interval $[\lambda_1, \lambda_2]$ and $N$ for the $N$-fold cross validation. The CV method has been shown to overfit the true model with a positive probability, and it is asymptotically loss efficient (see [17, 21]). We

review the ERIC (see [12]) motivated from a sound Bayesian perspective; this selector extends the BIC to account for the effect of prior information on the bias-variance tradeoff. If the adaptive LASSO penalty is asymptotically negligible, the ERIC is reduced to the BIC with unpenalized MLEs. The classifier has been shown to be selection-consistent when the number of covariates increases with the sample size, and this consistency holds true in a wider range of cases compared with the BIC (see [12]). However, whether the ERIC is still consistent by selecting the true group covariates in high-dimensional situations, e.g., $g_n = O(n^\kappa)$ for some $0 < \kappa < 1$, remains unknown. Thus, we adapt the ERIC-type criteria for high dimensional group-penalized variable selection as

$$\mathrm{ERIC}_\nu(\lambda) = -2\ell(\boldsymbol{y} \,|\, \hat{\boldsymbol{\beta}}_\lambda) + 2\nu \sum_{j \in \alpha_\lambda} d_j \log(n\phi/\lambda), \tag{2.4}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the penalized estimate obtained by maximizing (2.1). We denote the model associated with $\hat{\boldsymbol{\beta}}_\lambda$ by $\alpha_\lambda$, and consider $\alpha_\lambda = \{j : \|\hat{\beta}_{\lambda,j}\| \neq 0\}$ as the active set excluding the intercept. The number of variables in each group is allowed to diverge with the sample size $n$.

The way ERIC penalizes models is fundamentally different from the way the BIC does, where the latter penalizes a constant value for every new covariate entered into the model. The primary difference lies in the second term, which captures the variance of the nonzero estimates, and the ERIC penalizes models more severely for over-fitting than the BIC. ERIC-type criteria have a dynamic variance penalty, which depends on $\lambda$ itself, meaning it also depends on how complex the model already is. We study (2.4) via an approach similar to that used in [7,12]. More critically, we establish its uniform asymptotic properties by defining a proxy version of the criterion

$$\mathrm{ERIC}^*_\nu(\lambda) = -2\ell(\boldsymbol{y} \,|\, \tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) + 2\nu \sum_{j \in \alpha_\lambda} d_j \log(n\phi/\lambda),$$

where $\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}$ is the unpenalized MLE under model $\alpha_\lambda$ and uniform for all $|\alpha_\lambda| \leqslant K$, where $K$ can be unbounded.

Although, the ERIC-type selector in (2.4) was motivated by the Bayesian framework for the adaptive group LASSO with Laplace priors, our Bayesian friends may also regard the SCAD as the maximum a posteriori using an SCAD prior (the density is the exponential of the negative SCAD penalty function). For the case of a group SCAD, we can apply LLA to the SCAD penalty function in (2.3) (see [24]) and then choose $\rho_{\mathrm{SCAD}}(\|\boldsymbol{\beta}_j\|) = \rho'_{\mathrm{SCAD}}(\|\tilde{\boldsymbol{\beta}}_j\|)\|\boldsymbol{\beta}_j\|$, which provides better unification of adaptive group LASSO and the group SCAD penalty. We can also use the same framework of the ERIC to ensure selection consistency for the group SCAD penalty. The expression in (2.4) can be directly used to the adaptive group LASSO but needs modification of the tuning parameter $\lambda$ by multiplying $n$ when applied to the group nonconvex penalty functions.

## 3 Selection consistency

In this section, we show that ERIC is selection-consistent for group-penalized GLMs, where both $g_n$ and $d_j$ grow at a lower rate than the sample size $n$. Let $\boldsymbol{\beta^0} = (\beta_1^{0\mathrm{T}}, \ldots, \beta_g^{0\mathrm{T}})^{\mathrm{T}}$ denote the true parameter values, with the true model identified by $\alpha_0 = \{j : \|\beta_j^0\| \neq 0\}$ and $p_0 = \sum_{j \in \alpha_0} d_j$. We will develop the concept of selection consistency for both adaptive group LASSO and two concave 2-norm group selection problems.

### 3.1 Consistency for adaptive group LASSO

To obtain the consistency properties for adaptive group LASSO, we assume the following regularity conditions are satisfied.

**Condition 1.** The range of tuning parameters considered in (2.2) lies in the interval $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min} > 0$ and $\lambda_{\max}/n \to 0$, as $n \to \infty$.

**Condition 2.** Let $\lim_{n\to\infty}\frac{\log(g_n)}{\log(n)}=\kappa_1$. Denote $d_{\min}=\min_{1\leqslant j\leqslant g}\{d_j\}$ and $d_{\max}=\max_{1\leqslant j\leqslant g}\{d_j\}$, and let $\lim_{n\to\infty}\frac{\log(d_{\min})}{\log(n)}=\lim_{n\to\infty}\frac{\log(d_{\max})}{\log(n)}=\kappa_2$, where $\kappa_1+\kappa_2\in[0,1)$.

**Condition 3.** For any model $\alpha$ defined in the interval $[\lambda_{\min},\lambda_{\max}]$ such that $|\alpha|\leqslant K$, with $K$ satisfying $K=o(n^{\iota/2})$, where $0<\iota<\nu(1-\kappa_1)-2(\kappa_1+\kappa_2)$ with $2(\kappa_1+\kappa_2)/(1-\kappa_1)<\nu<2(1+\kappa_1+\kappa_2)/(1-\kappa_1)$, we have $\frac{1}{n}\boldsymbol{X}_\alpha^{\mathrm{T}}\boldsymbol{X}_\alpha\to_p\Gamma_\alpha$, and the minimum and maximum eigenvalues satisfy $0<c_1<\zeta_{\min}(\Gamma_\alpha)<\zeta_{\max}(\Gamma_\alpha)<c_2<\infty$. For all $\theta\in\mathbb{R}$, the function $b(\theta)$ has a second-order derivative with $c_0\leqslant b''(\theta)/\phi\leqslant 1/c_0$ and $|b'''(\theta)/\phi|\leqslant 1/c_0$, where $c_0,c_1$, and $c_2$ are positive constants.

**Condition 4.** There exists a positive constant $M$ such that $\min_{j\in\alpha_0}\|\beta_j^0/\sqrt{d_j}\|>M$.

**Condition 5.** (a) $\lambda\sqrt{\frac{p_0}{n}}\to 0$; (b) $\frac{\lambda}{\sqrt{ng_n}}(\frac{n}{p_n})^{\gamma/2}\to\infty$, as $n\to\infty$.

**Remark 3.1.** Condition 1 is identical to that in [12], $\lambda_{\max}$ can be easily chosen such that $\alpha_{\lambda_{\max}}$ is empty, and $\lambda_{\min}$ can be chosen such that $K=|\alpha_{\lambda_{\min}}|$ (see [7]). For group variable selection, Condition 2 permits the numbers of both groups and covariates in each group to diverge with the sample size, as Zou and Zhang [25] did without the group assumption. With Condition 2, we have $\lim_{n\to\infty}\log(p_n)/\log(n)=\kappa_1+\kappa_2<1$. Condition 3 is a mild regularity condition that ensures the Fisher information matrix exists and is non-singular for each $n$. Note that under Conditions 2–3, the MLEs for the full model are well-defined and $\sqrt{n/p_n}$-consistent; thus, weights $\tilde{w}_j=1/\|\tilde{\boldsymbol{\beta}}_j\|^\gamma$ can be calculated for use in the adaptive group LASSO. Condition 4 guarantees the strength of relevant groups. Condition 5 reduces to [12, Condition (A5)] for the adaptive LASSO with diverging $p_n$ and to the conditions in [16] for the adaptive group LASSO, thus achieving consistency in the fixed group $g_n$ and $d_j$ settings.

When the numbers of both groups and variables in each group diverge according to Condition 2, we must first obtain a generalization of the consistency result for the adaptive group LASSO to non-Gaussian responses with diverging numbers of groups and variables in each group. Then, to study the asymptotic behavior of ERIC, we partition the tuning parameter interval $[\lambda_{\min},\lambda_{\max}]$ into the under-fitted, true, and over-fitted subsets, respectively, $\Omega_-=\{\lambda:\alpha_\lambda\not\supset\alpha_0\}$, $\Omega_0=\{\lambda:\alpha_\lambda=\alpha_0\}$, and $\Omega_+=\{\lambda:\alpha_\lambda\supset\alpha_0$ and $\alpha_\lambda=\alpha_0\}$. This partition allows us to assess the performance of regularization parameter selections. Using the proxy ERIC$^*$ and the conditions above, we have the following lemma.

**Lemma 3.2.** *Assume Conditions* 1–4 *are satisfied and that there exists a* $\lambda_0\in\Omega_0$ *satisfying Condition* 5. *Then,*

(a) $\mathrm{P}(\inf_{\lambda\in\Omega_-}\min_{\alpha_\lambda\not\supset\alpha_0}\mathrm{ERIC}_\nu^*(\lambda)>\mathrm{ERIC}_\nu^*(\lambda_0))\to 1$.

(b) $\mathrm{P}(\inf_{\lambda\in\Omega_+}\min_{\alpha_\lambda\not\supseteq\alpha_0}\mathrm{ERIC}_\nu^*(\lambda)>\mathrm{ERIC}_\nu^*(\lambda_0))\to 1$, *if* $\gamma\geqslant\frac{2\kappa_1+(2-\nu)\kappa_2}{\nu(1-\kappa_1-\kappa_2)}-1$.

**Remark 3.3.** Lemma 3.2 is established for unbounded $K$ by Condition 3, which is also studied in [9]. The condition on $\gamma$ reduces to $\gamma\geqslant\frac{2\kappa_1}{\nu(1-\kappa_1)}-1$, which is identical to that in [12] without a group structure assumption. In Lemma 3.2(b), the condition on $\gamma$ reduces to a simple inequality for the case $\nu=1$ ($\gamma\geqslant(2\kappa_1+\kappa_2)/(1-\kappa_1-\kappa_2)$). Then, choosing a $\gamma\geqslant 1$ will be satisfied by Condition 3. In simulations, choosing $\gamma$ is straightforward since $\kappa_1$ and $\kappa_2$ are known. In real applications, we suggest trying several values of $\gamma$, for example, $\gamma=1,2,4,6$, while taking into account the dimensionality of the dataset at hand.

Based on Lemma 3.2 and the estimation consistency given in Appendix A, we obtain the following results.

**Theorem 3.4.** *Suppose Conditions* 1–5 *are satisfied. Then, the tuning parameter* $\hat{\lambda}$ *selected by minimizing* $\mathrm{ERIC}_\nu(\lambda)$ *in* (2.4) *satisfies* $\mathrm{P}\{\alpha_{\hat{\lambda}}=\alpha_0\}\to 1$.

**Corollary 3.5.** *Let* $\hat{\lambda}$ *be the tuning parameter chosen by minimizing BIC. If* $\kappa_1>\frac{1}{2}$ *in Condition* 2, *then,* $\mathrm{P}\{\alpha_{\hat{\lambda}}=\alpha_0\}\not\to 1$.

**Remark 3.6.** Theorem 3.4 guarantees the selection of ERIC if the true model is contained within the set of candidate models. Corollary 3.5 extends the result shown in [2, 12] to the group-penalized likelihood setting. By the above theorems, ERIC is selection-consistent for a wider range of settings compared with BIC.

### 3.2 Consistency for concave 2-norm group selection

In this section, we extend the concept of selection consistency to nonconvex penalty functions, e.g., group SCAD and group MCP. To study the consistent property, we need the following regularity conditions.

**Condition 1′.** The range of tuning parameters considered in (2.3) lies in the interval $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\max} \to 0$, as $n \to \infty$.

**Condition 4′.** Assume $a > 1 + c_*^{-1}$ for the group SCAD and $a > c_*^{-1}$ for the group MCP, where $a$ is the penalty parameter in (2.1), $c_* = c_1 c_0$, and $c_1$ and $c_2$ are the same as those defined in Condition 3. The nonzero group value $\beta_j^0$ satisfies $\min_{j\in\alpha_0} \|\beta_j^0/\sqrt{d_j}\|/\lambda \to \infty$, as $n \to \infty$.

**Condition 5′.** (a) $\lambda = O(n^{-1+\delta})$, where $\frac{1+\kappa_1}{2} < \delta < 1$; (b) $\lambda\sqrt{\frac{nd_{\min}}{p_n}} \to \infty$, as $n \to \infty$.

**Remark 3.7.** Condition $1'$ indicates that a smaller regularization parameter is needed if the sample size is large. Condition $4'$ is similar to the one in [5], which is necessary for obtaining the oracle property, and the condition on $a$ ensures the objective functions in (2.3) are globally convex. Condition $5'$ reduces to the conditions in [5] without a group structure assumption, where $\kappa_1$ is the same as that in Condition 2.

To understand the selection consistency of ERIC, we first have the following consistency results for group SCAD selection problems.

**Lemma 3.8.** *Assume Conditions* 2–3 *and* $4'$–$5'$ *are satisfied. Then, the group SCAD estimates* $\hat{\boldsymbol{\beta}}_\lambda$ *must satisfy*: (a) *estimate consistency*: $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$; (b) *selection consistency*: $\mathrm{P}(\{j : \|\hat{\beta}_{\lambda,j}\| = 0\} = \alpha_0^c) \to 1$.

Similar to the adaptive group LASSO, we show that the group nonconcave penalized likelihood of the generalized linear model with the ERIC selector possesses the oracle property. By the following theorem, for group nonconvex problems, if the true model is contained within the set of candidate models, it can be selected by ERIC.

**Theorem 3.9.** *Suppose Conditions* $1'$, 2–3 *and* $4'$–$5'$ *are satisfied. Then, the tuning parameter* $\hat{\lambda}$ *selected by minimizing* $\mathrm{ERIC}_\nu(\lambda)$ *in* (2.4) *satisfies* $\mathrm{P}\{\alpha_{\hat{\lambda}} = \alpha_0\} \to 1$.

## 4 Ultra-high dimensional data

In this section, we extend our methods separately to the convex and 2-norm concave problems in terms of ultra-high dimensional settings in GLIM (see [6,7]). For the adaptive group LASSO penalty, we apply the maximum marginal likelihood estimators as weights of the adaptive group LASSO inspired by [6,11]. Then, with our ERIC, we achieve selection consistency. For 2-norm concave penalties, penalties such as group SCAD can be used directly, and selection consistency can be guaranteed by using ERIC to select the tuning parameter. We note here that methods such as the LLA algorithm (see [24]) can be used to solve nonconvex penalties, which can be viewed as solving an adaptive group LASSO-type penalty problem, and, consequently, enjoys the same oracle property as the original SCAD-regularized estimator. The detailed implementation for group SCAD is similar to that described by Fan and Tang [7]; thus, we focused on adaptive group LASSO in this section. To obtain selection consistency for the adaptive group LASSO in that case of $\log p_n = o(n^\kappa)$ for some $\kappa > 0$, we first need the following conditions:

**Condition 6.** Conditions 1 and 3 hold true.

**Condition 7.** $\log p_n = o(n^{2\tau})$ and $p_0 = o(n^{2\tau_0})$, for $0 < \tau_0 < \tau < 1/2$.

**Condition 8.** There exists a positive constant $M$ such that $\min_{j\in\alpha_0} \|\beta_j^0/\sqrt{d_j}\| > M$ and $\max_{j\notin\alpha_0} \|\tilde{\beta}_j/\sqrt{d_j}\| \leqslant Mn^{-\tau}$, for $0 < \tau < 1/2$, with a probability approaching 1.

**Condition 9.** Uniformly for all $i = 1, \ldots, n$, $\mathrm{E}|y_i - b'(\boldsymbol{X}_i\boldsymbol{\beta}^0)|^r \leqslant \frac{1}{2}r!L^{r-2}C$, for any $r \geqslant 2$ and some constants $L$ and $C$.

**Condition 10.** (a) $\lambda\sqrt{\frac{p_0}{n}} \to 0$; (b) $\frac{\lambda}{\sqrt{n^{1-2\tau}\log(p_n)}} \to \infty$.

Condition 7 allows the rate of group number $g_n$ to be large as $\exp(o(n^{\kappa_1}))$ for some $0 < \kappa_1 < 1$, which indicates that $g_n$ can be much larger than $n$, while the number of true nonzero variables $p_0$ is permitted to grow with the sample size at a certain rate. Condition 8 can be shown according to Fan and Song [6] without a group structure, where $\tilde{\beta}_j$ are the maximum marginal likelihood estimators. Condition 9 is an assumption on the moments of noise similar to [7, Condition 3]. This condition implies that the tail distribution of noise undergoes exponential decay, which is reasonable for the generalized linear model. Condition 10 can be easily satisfied by choosing $\lambda = o(\sqrt{n})$. Moreover, when the number of covariates is fixed, the choice of $\lambda$ reduces to $\lambda/\sqrt{n} \to 0$ and $\lambda/n^{\frac{1}{2}-\tau} \to \infty$, which is identical to that in [24, Theorem 5], to establish the oracle properties of the one-step estimator for the adaptive LASSO penalty. Similar to the results of adaptive group LASSO in the case $p_n < n$, we first obtain a generalization of the consistency result for adaptive group LASSO in the ultra-high dimensional case, which allows group number $g_n = \exp(o(n^{\kappa_1}))$ for some $0 < \kappa_1 < 1$. Then, we obtain the following result.

**Theorem 4.1.** *Suppose Conditions* 6–10 *hold true. Then, the tuning parameter* $\hat{\lambda}$ *selected by minimizing* $\mathrm{ERIC}_\nu(\lambda)$ *in* (2.4) *satisfies*

$$\mathrm{P}\{\alpha_{\hat{\lambda}} = \alpha_0\} \to 1.$$

## 5 Simulation studies

We conducted a simulation study to compare the performance of $\mathrm{ERIC}_\nu$ against the BIC criteria used to select $\lambda$ in group-penalized GLMs. Three models are considered to compare the performance. To assess the finite sample performance of the proposed methods, we report the following associated features of parameter estimation and variable selection by ERIC, BIC, and CV selectors with adgLASSO, gSCAD, and gMCP: (1) the average of the estimated root mean squared error, $\mathrm{RMSE} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$, (2) the average model size of the identified model $\hat{\alpha}_\lambda = \{j : \|\beta_j\| \neq 0\}$, $\mathrm{MS} = |\hat{\alpha}_\lambda|$, (3) the average false positive rate, $\mathrm{TPR} = |\hat{\alpha}_\lambda^c \cap \alpha_0|/|\alpha_0|$, (4) the average false negative rate, $\mathrm{FNR} = |\hat{\alpha}_\lambda \cap \alpha_0^c|/|\alpha_0|$, and (5) the percentage of correct identified models, denoted by CM. Ideally, we wish to have CM close to 1. To compare model fittings, we further calculate the model error $\mathrm{ME}(\boldsymbol{\beta}) = \mathrm{E}_{\boldsymbol{X}}[\mu(\boldsymbol{X}\boldsymbol{\beta}) - \mu(\boldsymbol{X}\hat{\boldsymbol{\beta}})]^2$, and then report the median of relative model errors (MRME) of the refitted unpenalized estimates for each selected model, which is also used in [21]. For comparison, we evaluate the $\mathrm{ERIC}_\nu$ with $\nu = 0.75$ or $\nu = 1$. Moreover, we also report the results of a BIC-type criterion proposed in [9], which includes GLM as a special case, denoted as

$$\mathrm{BIC}_{gc} = -2\ell(\boldsymbol{y} \,|\, \hat{\boldsymbol{\beta}}) + c\hat{d}_\lambda^* \log(g_n),$$

where $\hat{d}_\lambda^* = \mathrm{tr}(\hat{H}_{\alpha_\lambda}^{-1}\hat{V}_{\alpha_\lambda})$ with $\hat{H}_{\alpha_\lambda}^{-1}$ the observed Hessian matrix and $\hat{V}_{\alpha_\lambda}$ as the sample covariance matrix of the score function. In each of the simulated models, we choose $c = 2$ in the simulation section, and a total of 500 replications are conducted.

**Example 5.1.** We simulate the data from a linear model with $p_n = \lfloor 8n^{1/2} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Covariates $\{\boldsymbol{X}_i; i = 1, \ldots, n\}$ were multivariate normal with the covariance matrix $\Sigma = (\sigma_{ij})$, in which $\sigma_{ij} = 0.2^{|i-j|}$, which are divided into 20 groups, each of size 4. The true regression coefficients shared within each group are $(-2, -1, 1, 2)$, respectively. We consider sample sizes $n = 100$ and error variances $\sigma^2 = 4$ and used $\gamma = 1$ in the adaptive group LASSO weights. From Table 1, all ERIC methods with adaptive group LASSO, group SCAD, and group MCP show better performance than BIC and 10-fold CV in terms of variable selection and model error measures. This result is consistent with the tendency of ERIC to choose larger values of $\lambda$ and, thus, smaller models compared with the BIC criteria (see Figure 1). In model identification, for example, gSCAD-ERIC is more likely than gSCAD-BIC to correctly detect the five true nonzero coefficients, while BIC is slightly more prone than ERIC to overfit when the sample size is small. Both ERIC and $\mathrm{BIC}_{gc}$ are applicable, and using $\nu = 1$ will lead stronger performance compared with $\nu = 0.75$. Finally, 10-fold CV performs very poorly with regard to selecting the true model, although its predictive performance is similar to the BIC-type criteria. As mentioned previously, cross validation is asymptotically less efficient rather than consistent.

**Table 1** Simulation results for the linear regression model

| Method | MRME | RMSE | MS | CM | FPR | FNR |
|---|---|---|---|---|---|---|
| adgLASSO-ERIC$_{0.75}$ | 0.2662 | 0.1097 (0.0251) | 4.070 | 0.938 | 0.0001 | 0.0144 |
| adgLASSO-ERIC$_1$ | 0.2326 | 0.1234 (0.0266) | 4.048 | 0.948 | 0 | 0.0108 |
| adgLASSO-BIC$_{gc}$ | 0.2481 | 0.1101 (0.0241) | 4.062 | 0.944 | 0 | 0.0124 |
| adgLASSO-BIC | 0.2507 | 0.1112 (0.0273) | 4.176 | 0.874 | 0 | 0.0272 |
| adgLASSO-CV | 0.2328 | 0.1365 (0.0385) | 6.514 | 0.282 | 0 | 0.5028 |
| gSCAD-ERIC$_{0.75}$ | 0.2194 | 0.1081 (0.0219) | 4.102 | 0.910 | 0 | 0.0204 |
| gSCAD-ERIC$_1$ | 0.2403 | 0.1088 (0.0254) | 4.042 | 0.957 | 0 | 0.0085 |
| gSCAD-BIC$_{gc}$ | 0.2213 | 0.1111 (0.0240) | 4.054 | 0.954 | 0 | 0.0108 |
| gSCAD-BIC | 0.2397 | 0.1096 (0.0252) | 4.112 | 0.918 | 0 | 0.0224 |
| gSCAD-CV | 0.2314 | 0.1083 (0.0233) | 6.010 | 0.332 | 0 | 0.4020 |
| gMCP-ERIC$_{0.75}$ | 0.2187 | 0.1078 (0.0211) | 4.024 | 0.984 | 0 | 0.0048 |
| gMCP-ERIC$_1$ | 0.2173 | 0.1015 (0.0202) | 4.004 | 0.996 | 0 | 0.0008 |
| gMCP-BIC$_{gc}$ | 0.2182 | 0.1026 (0.0208) | 4.006 | 0.994 | 0 | 0.0012 |
| gMCP-BIC | 0.2508 | 0.1052 (0.0232) | 4.036 | 0.974 | 0 | 0.0072 |
| gMCP-CV | 0.2300 | 0.1097 (0.0264) | 4.714 | 0.662 | 0 | 0.1428 |

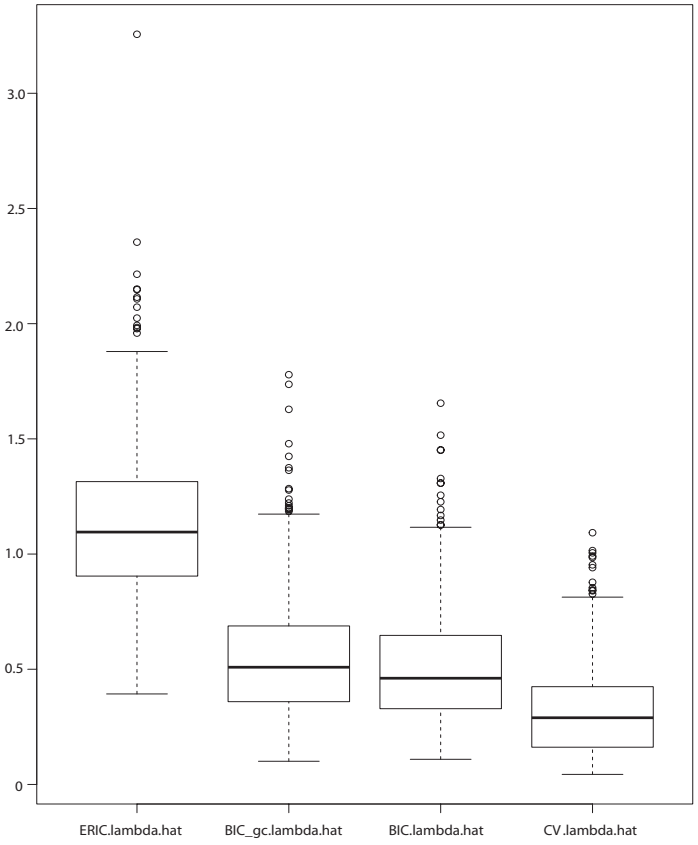


**Figure 1** Boxplots of $\lambda$ values chosen by ERIC, BIC$_{gc}$, BIC and CV based on the adaptive group LASSO from Example 5.1. Both criteria selected the true model

**Table 2** Simulation results for the logistic regression model

| Method | MRME | RMSE | MS | CM | FPR | FNR |
|---|---|---|---|---|---|---|
| adgLASSO-ERIC$_{0.75}$ | 0.3734 | 1.3540 (0.2679) | 3.970 | 0.585 | 0.0126 | 0.0570 |
| adgLASSO-ERIC$_1$ | 0.3051 | 1.3045 (0.2550) | 3.695 | 0.780 | 0.0126 | 0.0020 |
| adgLASSO-BIC$_{gc}$ | 0.2950 | 1.4381 (0.3712) | 3.954 | 0.654 | 0.0121 | 0.0512 |
| adgLASSO-BIC | 0.5125 | 1.5491 (0.4503) | 4.620 | 0.520 | 0.0025 | 0.1364 |
| adgLASSO-CV | 0.5862 | 1.6015 (0.1822) | 7.104 | 0.006 | 0.0016 | 0.6208 |
| gSCAD-ERIC$_{0.75}$ | 0.2907 | 1.4195 (0.3123) | 4.180 | 0.850 | 0.0008 | 0.0400 |
| gSCAD-ERIC$_1$ | 0.2758 | 1.3585 (0.2595) | 3.980 | 0.925 | 0.0020 | 0.0060 |
| gSCAD-BIC$_{gc}$ | 0.2878 | 1.4108 (0.3091) | 4.130 | 0.840 | 0.0010 | 0.0310 |
| gSCAD-BIC | 0.7661 | 1.7654 (0.4534) | 4.600 | 0.514 | 0.0031 | 0.1352 |
| gSCAD-CV | 0.5293 | 1.5083 (0.2714) | 6.665 | 0.064 | 0.0006 | 0.3520 |
| gMCP-ERIC$_{0.75}$ | 0.3564 | 1.3978 (0.2812) | 4.125 | 0.840 | 0.0014 | 0.0310 |
| gMCP-ERIC$_1$ | 0.2541 | 1.3182 (0.2188) | 3.990 | 0.940 | 0.0112 | 0.0180 |
| gMCP-BIC$_{gc}$ | 0.2772 | 1.3586 (0.2557) | 4.085 | 0.855 | 0.0018 | 0.0260 |
| gMCP-BIC | 0.5415 | 1.6174 (0.4012) | 4.481 | 0.570 | 0.0027 | 0.1096 |
| gMCP-CV | 0.5308 | 1.4787 (0.2307) | 6.200 | 0.080 | 0.0012 | 0.2460 |

**Table 3** Simulation results for the ultra-high dimensional regression model

| Method | RMSE | MS | CM | FPR | FNR |
|---|---|---|---|---|---|
| adgLASSO-ERIC$_{0.75}$ | 0.0461 (0.0065) | 4.100 | 0.904 | 0 | 0.0200 |
| adgLASSO-ERIC$_1$ | 0.0463 (0.0061) | 4.054 | 0.946 | 0 | 0.0108 |
| adgLASSO-BIC$_{gc}$ | 0.0461 (0.0062) | 4.070 | 0.934 | 0 | 0.0140 |
| adgLASSO-BIC | 0.0455 (0.0062) | 4.120 | 0.890 | 0 | 0.0240 |
| adgLASSO-CV | 0.0434 (0.0052) | 4.526 | 0.562 | 0 | 0.1052 |
| gSCAD-ERIC$_{0.75}$ | 0.0358 (0.0075) | 4.048 | 0.956 | 0 | 0.0096 |
| gSCAD-ERIC$_1$ | 0.0397 (0.0086) | 4.016 | 0.984 | 0 | 0.0032 |
| gSCAD-BIC$_{gc}$ | 0.0332 (0.0075) | 4.012 | 0.988 | 0 | 0.0024 |
| gSCAD-BIC | 0.0358 (0.0074) | 4.080 | 0.932 | 0 | 0.0160 |
| gSCAD-CV | 0.0329 (0.0027) | 7.040 | 0.022 | 0 | 0.6080 |
| gMCP-ERIC$_{0.75}$ | 0.0338 (0.0074) | 4.050 | 0.964 | 0 | 0.0100 |
| gMCP-ERIC$_1$ | 0.0332 (0.0071) | 4.003 | 0.997 | 0 | 0.0005 |
| gMCP-BIC$_{gc}$ | 0.0333 (0.0076) | 4.002 | 0.998 | 0 | 0.0004 |
| gMCP-BIC | 0.0337 (0.0072) | 4.286 | 0.836 | 0 | 0.0572 |
| gMCP-CV | 0.0332 (0.0018) | 8.044 | 0.010 | 0 | 0.8088 |

**Example 5.2.** We consider a logistic regression model with a rate of divergence of $p_n = \lfloor 6n^{1/2} \rfloor$, and covariates $\boldsymbol{X}$ are generated independently from a multivariate Gaussian distribution; the covariance matrix has a compound symmetric structure with $\rho = 0.5$. Predictors are divided into 30 groups with an equal group size 2. In addition, the true regression coefficients of the first and the last four groups are equal to $(-3, 1.5, 0, 0, 3, -1.5, 0, 0)$. An unpenalized intercept of 1 is also included in the true model. We choose a suitable $\gamma = 6$ for the adaptive group LASSO weights considering computational efficiency. ERIC$_1$ performs best in this setting, with substantial gains in both the proportion of correct models and the model fittings over all BIC-type criteria (see Table 2). The improvement is driven largely by a reduction in false negative rates for ERIC, with little compromise in false positive rates, and results in less over-fitting compared with the BIC criteria. Compared with adaptive group LASSO, the group nonconcave penalties show stronger performance with ERIC. For example, gMCP-ERIC applies greater shrinkage to dramatically reduce the number of false negatives at the risk of missing some truly informative coefficients. Both BIC and 10-fold CV consistently over-fitted for all three group penalties, although this over-fitting did lead to lower false negative rates compared with the ERIC.

**Example 5.3.** For the ultra-high dimensional case, we simulate data from the linear model with $p = 800$ and $\sigma = 1.5$. The covariate vector is normally distributed with mean zero, and the covariance matrix specified is similar to that of Huang et al. [11]. The first 16 covariates $(X_{i1}, \ldots, X_{i16})$ are multivariate normal with the covariance matrix $\Sigma = (\sigma_{ij})$, in which $\sigma_{ij} = 0.6^{|i-j|}$, which are divided into 4 groups, each of size 4. The rest of the covariates are generated in the same way with $\sigma_{ij} = 0.2^{|i-j|}$, and these two parts are independent. The value of $X$ is generated once and then kept fixed. The true regression coefficients are identical to those in Example 5.1. The sample size used in estimation is $n = 100$, and summary statistics are computed based on 500 replications. Results are similar to those obtained from Examples 5.1 and 5.2, $\text{ERIC}_\nu$ and $\text{BIC}_{gc}$ are comparable with $\text{ERIC}_1$ outperforming the BIC criteria and 10-fold CV in model selection and prediction (see Table 3), i.e., ERIC is marginally more successful at detecting true nonzeros than either BIC or CV. ERIC continues to show lower false negative rates than the BIC methods.

**Table 4** The preprocessed student-related variables

| Attribute | Description (domain) |
|---|---|
| Sex | Student's sex (binary: female or male) |
| Age | Student's age (numeric: from 15 to 22) |
| School | Student's school (binary: Gabrie Pereira or Mousinho da Silveira) |
| Address | Student's home address type (binary: urban or rural) |
| Pstatus | Parent's cohabitation status (binary: living together or apart) |
| Medu | Mother's education (numeric: from 0 to 4[a]) |
| Mjob | Mother's job (nominal[b]) |
| Fedu | Father's education (numeric: from 0 to 4[a]) |
| Fjob | Father's job (nominal[b]) |
| Guardian | Student's guardian (nominal: mother, father or other) |
| Famsize | Family size (binary: $\leqslant 3$ or $> 3$) |
| Famrel | Quality of family relationships (numeric: from 1 - very bad to 5 - excellent) |
| Reason | Reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| Traveltime | Home to school travel time (numeric: 1 - $<$ 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour or 4 - $>$ 1 hour) |
| Studytime | Weekly study time (numeric: 1 - $<$ 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours or 4 - $>$ 10 hours) |
| Failures | Number of past class failures (numeric: $n$ if $1 < n < 3$, else 4) |
| Schoolsup | Extra educational school support (binary: yes or no) |
| Famsup | Family educational support (binary: yes or no) |
| Activities | Extra-curricular activities (binary: yes or no) |
| Paidclass | Extra paid classes (binary: yes or no) |
| Internet | Internet access at home (binary: yes or no) |
| Nursery | Attended nursery school (binary: yes or no) |
| Higher | Wants to take higher education (binary: yes or no) |
| romantic | With a romantic relationship (binary: yes or no) |
| Freetime | Free time after school (numeric: from 1 - very low to 5 - very high) |
| Goout | Going out with friends (numeric: from 1 - very low to 5 - very high) |
| Walc | Weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| Dale | Workday alcohol consumption (numeric: from 1 - very low to 5 - very high) |
| Health | Current health status (numeric: from 1 - very bad to 5 - very good) |
| Absences | Number of school absences (numeric: from 0 to 93) |
| G1 | First period grade (numeric: from 0 to 20) |
| G2 | Second-period grade (numeric: from 0 to 20) |
| G3 | Final grade (numeric: from 0 to 20) |

a: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education.

b: teacher, health care related, civil services (e.g., administrative or police), at home or other.

**Table 5** Selection methods comparison for mushroom data

| | gSCAD-ERIC$_1$ | gSCAD-BIC | gSCAD-CV |
|---|---|---|---|
| Mathematics: | | | |
| | Logistic classification model | | |
| No. of factors selected | 1.6300 (1.2198) | 3.0300 (2.6985) | 1.3600 (0.8105) |
| Outsample TPR | 0.8841 (0.0992) | 0.8545 (0.0740) | 0.8886 (0.0970) |
| Outsample FPR | 0.1624 (0.0610) | 0.1522 (0.0569) | 0.1631 (0.0606) |
| Outsample PCC | 0.9078 (0.0105) | 0.9018 (0.0194) | 0.9091 (0.0214) |
| | Linear regression model | | |
| No. of factors selected | 1.7100 (1.3126) | 6.8900 (2.5776) | 3.6600 (3.3851) |
| Outsample RMSE ($\times 10^{-1}$) | 0.5247 (0.0204) | 0.5253 (0.0192) | 0.5253 (0.0198) |
| Portuguese language: | | | |
| | Logistic classification model | | |
| No. of factors selected | 1.4500 (0.9143) | 2.2300 (1.5364) | 1.8300 (1.2477) |
| Outsample TPR | 0.6851 (0.0675) | 0.7005 (0.0739) | 0.6904 (0.0754) |
| Outsample FPR | 0.1033 (0.0708) | 0.1557 (0.1051) | 0.1324 (0.1019) |
| Outsample PCC | 0.9375 (0.0105) | 0.9322 (0.0127) | 0.9341 (0.0122) |
| | Linear regression model | | |
| No. of factors selected | 1.3700 (0.9604) | 7.3800 (2.3689) | 4.5700 (4.6823) |
| Outsample RMSE ($\times 10^{-1}$) | 0.4769 (0.1939) | 0.4775 (0.1895) | 0.4777 (0.1914) |

## 6 An empirical example

In this section, we consider a real data set on student achievement in the secondary level of education in two Portuguese schools; this data set contains 33 attributes, including student grades, demographic, and social- and school-related features, and was collected from school reports and questionnaires (see Table 4). Two data sets with the same attributes and related to performance in two distinct subjects, i.e., Mathematics (mat) and Portuguese language studies (por), were also provided; these data sets were modeled under binary, five-level classification and regression tasks by Cortez and Silva [3]. The target attribute G3 has a strong correlation with attributes G2 and G1 because G3 is the final-year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st- and 2nd-period grades, respectively. To enable complete comparison, we consider both logistic and linear regression to model these two datasets. For fair evaluation, we randomly split the observations equally into two parts. One part of the observations is used to build the model, and the other part is used to evaluate the outsample forecasting accuracy. For reliable comparison, we repeat this procedure 100 times and report key findings in Table 5. We evaluate the prediction performance of the classification model by calculating the outsample percentage of correct classifications (PCC), false positive rate (FPR), and true positive rate (TPR). The results show that, regardless of the selection method used, the optimal model selected by ERIC consistently returns the smallest average model size. When evaluating student grades, a less-sensitive rule, which means a relatively smaller TPR, and a smaller FPR for decreasing rates of Type I errors are preferred. These findings reveal the greater power of the ERIC selector as a logistic classifier compared with other selector types. Using linear regression, we evaluate the outsample forecasting error by calculating the RMSE and find that the optimal model selected by ERIC consistently demonstrates the smallest average model size and the best prediction accuracy.

All of the methods tested suggest that attribute G2 should be included in the final model for both mat and por grades. However, when evaluating the mat grade with the linear model, the proposed ERIC only selected three attributes, i.e., G2, famrel, and absences, while BIC select another five attributes, including age, failures, activities, romantic, and G1. Our ERIC selected fewer attributes than BIC without significantly sacrificing RMSEs, which were 6.1704 and 6.1376, respectively. When evaluating por grades with the classification model, besides G2, BIC selected three other attributes, including school, age, and G1. Moreover, the ERIC selector showed a PCC identical to that of the BIC selector with a relatively

lower FPR, which was 0.0707 and 0.1616, respectively. Therefore, we can conclude that the quality of family relationships, number of school absences, and second-period grades could more directly affect a student's final grade than other factors.

## 7 Discussion

For all penalized likelihood methods, choosing an appropriate tuning parameter is critical to ensure good performance. In this work, we extended the proposed information criterion in Hui et al. [12], to high-dimensional group-penalized regression. We showed that ERIC is selection-consistent when both the true number of groups and the covariates in each group increase with the sample size and that the consistency holds true in a wider range of cases compared with BIC for both convex and nonconvex penalties. Simulations showed that ERIC can outperform the BIC criteria and CV when choosing the optimal tuning parameter. For ultra-high dimensional data, especially with the adaptive group LASSO, we proposed a direct approach to obtain consistency by constructing weights based on fitting GLMs to each group. Simulation results further revealed that all three penalized methods incorporating ERIC showed competitive empirical performance in the ultra-high dimensional case.

**Acknowledgements** This work was supported by National Natural Science Foundation of China (Grant Nos. 11571337 and 71631006) and the Fundamental Research Funds for the Central Universities (Grant No. WK2040160028).

### References

1 Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput, 2015, 25: 173–187
2 Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 2008, 95: 759–771
3 Cortez P, Silva A. Using data mining to predict secondary school student performance. In: Proceedings of the 5th Annual Future Business Technology Conference. Http://www3.dsi.uminho.pt/pcortez/student.pdf, 2008
4 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360
5 Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Ann Statist, 2004, 32: 928–961
6 Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Statist, 2010, 38: 3567–3604
7 Fan Y, Tang C Y. Tuning parameter selection in high dimensional penalized likelihood. J R Stat Soc Ser B Stat Methodol, 2013, 75: 531–552
8 Friedman J, Hastie T, Tibshirani R. A note on the group LASSO and a sparse group LASSO. ArXiv:1001.0736, 2010
9 Gao X, Carroll R J. Data integration with high dimensionality. Biometrika, 2017, 104: 251–272
10 Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Statist Sci, 2012, 27: 481–499
11 Huang J, Ma S, Zhang C H. Adaptive LASSO for sparse high-dimensional regression models. Statist Sinica, 2008, 18: 1603–1618
12 Hui F K C, Warton D I, Foster S D. Tuning parameter selection for the adaptive LASSO using ERIC. J Amer Statist Assoc, 2015, 110: 262–269
13 Kim Y, Kwon S, Choi H. Consistent model selection criteria on high dimensions. J Mach Learn Res, 2012, 13: 1037–1057
14 McCullagh P, Nelder J A. Generalized Linear Models. Boca Raton: CRC Press, 1989
15 Meier L, van de Geer S, Bühlmann P. The group LASSO for logistic regression. J R Stat Soc Ser B Stat Methodol, 2008, 70: 53–71
16 Wang H, Leng C. A note on adaptive group LASSO. Comput Statist Data Anal, 2008, 52: 5277–5286
17 Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B Stat Methodol, 2009, 71: 671–683
18 Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics, 2007, 23: 1486–1494
19 Wei F, Huang J. Consistent group selection in high-dimensional linear regression. Bernoulli, 2010, 16: 1369–1384

20 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol, 2006, 68: 49–67
21 Zhang Y, Li R, Tsai C L. Regularization parameter selections via generalized information criterion. J Amer Statist Assoc, 2010, 105: 312–323
22 Zhang Y, Shen X. Model selection procedure for high-dimensional data. Stat Anal Data Min, 2010, 3: 350–358
23 Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc, 2006, 101: 1418–1429
24 Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Ann Statist, 2008, 36: 1509–1533
25 Zou H, Zhang H H. On the adaptive elastic-net with a diverging number of parameters. Ann Statist, 2009, 37: 1733–1751

## Appendix A

We present the proof of our main theorem and some lemmas that are essential to the proof here as an appendix.

### Appendix A.1

**Lemma A.1.** *Assume Conditions* 1–5 *are satisfied. The adaptive group LASSO estimates* $\hat{\boldsymbol{\beta}}_\lambda$ *in* (2.2) *must satisfy*: (a) *estimate consistency*: $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$ *and* (b) *selection consistency*: $\mathrm{P}(\{j : \|\hat{\beta}_{\lambda,j}\| = 0\} = \alpha_0^c) \to 1$.

*Proof.* The proof below follows an approach similar to that of Fan and Li [4]. Let $\alpha_n = \sqrt{p_n/n}$ and $D(\boldsymbol{u}) = \ell_p(\boldsymbol{\beta}^0 + \alpha_n \boldsymbol{u}) - \ell_p(\boldsymbol{\beta}^0)$, where $\ell_p(\boldsymbol{\beta})$ was defined in (2.1) of the main text. If we can prove that for any given $\varepsilon > 0$, there exists a constant $C$ such that, for a large enough $n$, we have

$$\mathrm{P}\Big(\sup_{\|\boldsymbol{u}\|=C} D(\boldsymbol{u}) < 0\Big) \geqslant 1 - \varepsilon, \tag{A.1}$$

then a local maximizer $\hat{\boldsymbol{\beta}}_\lambda$ of the penalized log-likelihood $\ell_p(\boldsymbol{\beta})$ is guaranteed to exist such that $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$. Let $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{y} \,|\, \boldsymbol{\beta})$ denote the log-likelihood function. To prove the above statement, we first use a Taylor expansion to obtain

$$\begin{aligned} D(\boldsymbol{u}) &\leqslant \ell(\boldsymbol{\beta}^0 + \alpha_n \boldsymbol{u}) - \ell(\boldsymbol{\beta}^0) - \lambda\Big(\sum_{j\in\alpha_0} \sqrt{d_j}\tilde{w}_j(\|\beta_j^0 + \alpha_n u_j\| - \|\beta_j^0\|)\Big) \\ &\leqslant \alpha_n \boldsymbol{u}'\nabla\ell(\boldsymbol{\beta}^0) - \frac{1}{2} n\alpha_n^2 \boldsymbol{u}'\Big(-\frac{1}{n}\nabla^2\ell(\bar{\boldsymbol{\beta}})\Big)\boldsymbol{u} + \lambda\alpha_n \sum_{j\in\alpha_0} \sqrt{d_j}\tilde{w}_j\|u_j\| \\ &\equiv I_1 + I_2 + I_3, \end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ lies on the line segment joining $\boldsymbol{\beta}^0$ and $(\boldsymbol{\beta}^0 + \alpha_n \boldsymbol{u})$. Following an argument similar to [4], we have $\nabla\ell(\boldsymbol{\beta}^0) = O_p(\sqrt{np_n})$ and, thus, $|I_1| \leqslant O_p(\sqrt{np_n}\alpha_n)\|\boldsymbol{u}\| = O_p(n\alpha_n^2)\|\boldsymbol{u}\|$. By Conditions 1–5, we also have $I_2 \leqslant -(1/2)n\alpha_n^2 c_0 c_1 \|\boldsymbol{u}\|^2$. Finally, we consider $I_3$; for $j \in \alpha_0$, we have $\tilde{w}_j \leqslant M^{-\gamma} d_{\max}^{-\gamma/2}$ by Condition 4. Thus

$$I_3 = \lambda\alpha_n \sum_{j\in\alpha_0} \sqrt{d_j}\tilde{w}_j\|u_j\| \leqslant \lambda\alpha_n\Big(\sum_{j\in\alpha_0} d_j\tilde{w}_j^2\Big)^{\frac{1}{2}}\|\boldsymbol{u}\| \leqslant \frac{n\alpha_n^2\|\boldsymbol{u}\|}{M^\gamma d_{\max}^{\gamma/2}}\lambda\sqrt{\frac{p_0}{np}},$$

so we have $|I_3| = o_p(n\alpha_n^2)\|\boldsymbol{u}\|$ by Condition 5(a). Based on the above results, for a large enough $\|\boldsymbol{u}\|$, we know that terms $I_1$ and $I_3$ are asymptotically dominated by $I_2$, which is negative. The probability statement in (A.1) follows immediately from this, and we obtain $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$ as stated.

We now prove selection consistency by showing that, for any $\hat{\boldsymbol{\beta}}$ satisfying $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$, $\mathrm{P}(\{j : \|\hat{\beta}_j\| \neq 0\} = \alpha_0) \to 1$. We first prove that, as the probability approaches 1, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\alpha_0}, \boldsymbol{0})$ is the solution of (2.2). Consider the score equation for $\beta_j$: by the Taylor expansion, we have

$$\frac{\partial\ell_p(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} - \lambda\sqrt{d_j}\tilde{w}_j\frac{\beta_j}{\|\beta_j\|} = \frac{\partial\ell(\boldsymbol{\beta}^0)}{\partial\beta_j} + \sum_{k=1}^{g_n}\frac{\partial^2\ell(\bar{\boldsymbol{\beta}})}{\partial\beta_j\partial\beta_k}\left(\beta_k - \beta_k^0\right) - \lambda\sqrt{d_j}\tilde{w}_j\frac{\beta_j}{\|\beta_j\|}, \tag{A.2}$$

where $\bar{\boldsymbol{\beta}}$ lies on the line segment joining $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}$. According to the definition of $\hat{\boldsymbol{\beta}}$, it is equivalent to show

$$\mathrm{P}\left(\exists\, j \in \alpha_0^c, \left\|\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j}\right\| > \lambda\sqrt{d_j}\tilde{w}_j\right) \to 0. \tag{A.3}$$

For $j \in \alpha_0^c$, $\max_{j\in\alpha_0^c}\{\|\tilde{\beta}_j\|\} = O_p(\sqrt{p/n})$, and by (A.2), we have

$$\begin{aligned}\mathrm{P}\left(\exists\, j \in \alpha_0^c, \left\|\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j}\right\| > \lambda\sqrt{d_j}\tilde{w}_j\right) &\leqslant P\left(\exists\, j \in \alpha_0^c, \left\|\frac{\partial \ell(\boldsymbol{\beta}^0)}{\partial \beta_j}\right\| > \frac{\lambda\sqrt{d_j}\tilde{w}_j}{2}\right)\\ &\quad + P\left(\exists\, j \in \alpha_0^c, \left\|\sum_{k=1}^{g_n}\frac{\partial^2 \ell(\bar{\boldsymbol{\beta}})}{\partial\beta_j\partial\beta_k}\left(\hat{\beta}_k - \beta_k^0\right)\right\| > \frac{\lambda\sqrt{d_j}\tilde{w}_j}{2}\right)\\ &\equiv K_1 + K_2.\end{aligned}$$

For $K_1$, under Condition 5(b) and for a large constant $C$, based on Boole's inequality,

$$\begin{aligned}K_1 &= \mathrm{P}\left(\exists\, j \in \alpha_0^c, \|\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}^0)\| > \frac{\lambda\sqrt{d_j}\tilde{w}_j}{2}\right)\\ &\leqslant \mathrm{P}\left(\exists\, j \in \alpha_0^c, \|\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}^0)\| > \frac{\lambda\sqrt{d_j}}{2}\frac{1}{(\max_{j\in\alpha_0^c}\|\tilde{\beta}_j\|)^\gamma}\right)\\ &\leqslant \mathrm{P}\left(\exists\, j \in \alpha_0^c, \|\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}^0)\| > \frac{\lambda\sqrt{d_j}}{2}\left(C\sqrt{\frac{p}{n}}\right)^{-\gamma}\right)\\ &\leqslant \sum_{j\in\alpha_0^c}\mathrm{P}\left(\|\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}^0)\| > \frac{\lambda\sqrt{d_j}}{2}\left(C\sqrt{\frac{p}{n}}\right)^{-\gamma}\right)\\ &\leqslant \sum_{j\in\alpha_0^c}\frac{\mathrm{E}\|\boldsymbol{X}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}^0)\|^2}{\frac{\lambda^2 d_j}{4}(C\sqrt{\frac{p}{n}})^{-2\gamma}} \leqslant \sum_{j\in\alpha_0^c}\frac{4c_2 n}{\lambda^2(C\sqrt{\frac{p}{n}})^{-2\gamma}} \to 0.\end{aligned}$$

For $K_2$, based on Condition 3 and Markov's inequality,

$$\begin{aligned}K_2 &\leqslant \mathrm{P}\left(\exists\, j \in \alpha_0^c, \left\|\sum_{k=1}^{g}\frac{\partial^2 \ell(\bar{\boldsymbol{\beta}})}{\partial\beta_j\partial\beta_k}(\hat{\beta}_k - \beta_k^0)\right\| > \frac{\lambda\sqrt{d_j}\tilde{w}_j}{2}\right)\\ &\leqslant \frac{\mathrm{E}(\sum_{j\in\alpha_0^c}\|\sum_{k=1}^{g}\frac{\partial^2 \ell(\bar{\boldsymbol{\beta}})}{\partial\beta_j\partial\beta_k}(\hat{\beta}_k - \beta_k^0)\|^2)}{\frac{\lambda^2 d_{\min}}{4}(C\sqrt{\frac{p}{n}})^{-2\gamma}} \leqslant \frac{M\zeta_{\max}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\cdot\zeta_{\max}(\boldsymbol{X}_{\alpha_0^c}^{\mathrm{T}}\boldsymbol{X}_{\alpha_0^c})\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2}{c_0^2\lambda^2 d_{\min}(C\sqrt{\frac{p}{n}})^{-2\gamma}}.\end{aligned}$$

Note that $\zeta_{\max}(\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j) \leqslant \zeta_{\max}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) \leqslant nc_2$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$; thus

$$K_2 \leqslant C\frac{np_n}{\lambda^2 d_{\min}(\sqrt{\frac{p}{n}})^{-2\gamma}} \to 0,$$

under Condition 5(b). We only need to prove that $\mathrm{P}(\min_{j\in\alpha_0}\|\hat{\beta}_j\| > 0) \to 1$. By Condition 4,

$$\min_{j\in\alpha_0}\|\hat{\beta}_j\| \geqslant \min_{j\in\alpha_0}\|\beta_j^0\| - \max\|\hat{\boldsymbol{\beta}}_{\alpha_0} - \boldsymbol{\beta}_{\alpha_0}^0\| \geqslant M\sqrt{d_{\max}} - o_p(1) > 0.$$

Thus, the proof is completed. □

## Appendix A.2 Proof of Lemma 3.2

*Proof of Lemma* 3.2(a). For simplicity, we shall prove this lemma assuming $\phi = 1$ (e.g., Binomial, Poisson response). The proof can be straightforwardly extended to the linear model case with the known $\phi \equiv \sigma^2$, since $\tilde{\sigma}_{\alpha_\lambda}^2 = (1/n)\|\boldsymbol{y} - \boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}\|^2 = O_p(1)$ regardless of $\alpha_\lambda$.

Since $\lambda_0$ satisfies Condition 5, by Lemma A.1, the adaptive group LASSO estimates are $\sqrt{n/p_n}$-consistent and selection-consistent. Define $\tilde{\boldsymbol{\beta}}_\lambda = (\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}, \tilde{\boldsymbol{\beta}}_{\alpha_\lambda^c} = \mathbf{0})$, note that $\alpha_{\lambda_0} = \alpha_0$, and let $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{y}\,|\,\boldsymbol{\beta})$ denote the log-likelihood function. Then, we have

$$\begin{aligned}\frac{\mathrm{ERIC}^*_\nu(\lambda) - \mathrm{ERIC}^*_\nu(\lambda_0)}{n} &= -\frac{2}{n}(\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\tilde{\boldsymbol{\beta}}_{\alpha_0})) + \frac{2\nu}{n}\bigg(\sum_{j\in\alpha_\lambda} d_j \log\left(\frac{n}{\lambda}\right) - \sum_{j\in\alpha_0} d_j \log\left(\frac{n}{\lambda_0}\right)\bigg)\\ &\geqslant -\frac{2}{n}(\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\tilde{\boldsymbol{\beta}}_{\alpha_0})) - \frac{2\nu p_0}{n}\log\left(\frac{n}{\lambda_0}\right) \quad \text{(by Condition 1)}\\ &\equiv J_1 + J_2.\end{aligned}$$

From Conditions 1–2, we have $J_2 = -(2\nu p_0/n)\log(n) + (2\nu p_0/n)\log(\lambda_0) \triangleq L_1 + L_2$. By Condition 5(a), $L_2 \to 0$ can be easily seen. Thus, that $J_1 + L_1$ is asymptotically positive and uniform for all $\lambda \in \Omega^-$ and models $\{\alpha_\lambda \not\supset \alpha_0; |\alpha_\lambda| \leqslant K\}$ is left to be proven.

Without loss of generality, we can rearrange the coefficient vector $\tilde{\boldsymbol{\beta}}_\lambda$ as

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_\lambda &= (\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}, \tilde{\boldsymbol{\beta}}_{\alpha_\lambda^c \cap \alpha_0} = \mathbf{0}, \tilde{\boldsymbol{\beta}}_{(\alpha_0\cup\alpha_\lambda)^c} = \mathbf{0})\\ &= (\tilde{\boldsymbol{\beta}}_\lambda(S), \tilde{\boldsymbol{\beta}}_\lambda(S^c) = \mathbf{0}),\end{aligned}$$

where $\boldsymbol{S} = \alpha_0 \cup \alpha_\lambda$, shift the under-fitted coefficients out of $\alpha_\lambda^c$, and concatenate them with $\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}$ to form $\tilde{\boldsymbol{\beta}}_S$, which will then contain some coefficients that are zero. As this rearrangement of $\tilde{\boldsymbol{\beta}}_\lambda$ does not alter the value of the likelihood, we can write $\ell(\tilde{\boldsymbol{\beta}}_\lambda(S)) = \ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda})$. In the same way, we can rearrange $\tilde{\boldsymbol{\beta}}_{\lambda_0}$ and $\boldsymbol{\beta}^0$ as follows:

$$\begin{aligned}&\tilde{\boldsymbol{\beta}}_{\lambda_0} = (\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{\beta}}_{\alpha_\lambda \cap \alpha_0^c} = \mathbf{0}, \tilde{\boldsymbol{\beta}}_{(\alpha_0\,\cup\,\alpha_\lambda)^c} = \mathbf{0}) = (\tilde{\boldsymbol{\beta}}_{\lambda_0}(S), \tilde{\boldsymbol{\beta}}_{\lambda_0}(S^c) = \mathbf{0}),\\ &\boldsymbol{\beta}^0 = (\boldsymbol{\beta}^0_{\alpha_0}, \boldsymbol{\beta}^0_{\alpha_0^c \cap \alpha_\lambda} = \mathbf{0}, \boldsymbol{\beta}^0_{(\alpha_0\,\cup\,\alpha_\lambda)^c} = \mathbf{0}) = (\boldsymbol{\beta}^0(S), \boldsymbol{\beta}^0(S^c) = \mathbf{0}),\end{aligned}$$

where $S = \alpha_0 \cup \alpha_\lambda$. On the same principle, we have $\ell(\tilde{\boldsymbol{\beta}}_{\lambda_0}(S)) = \ell(\tilde{\boldsymbol{\beta}}_{\alpha_0})$. By the Taylor expansion,

$$-\frac{2}{n}(\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\tilde{\boldsymbol{\beta}}_{\alpha_0})) = \frac{1}{n}(\tilde{\boldsymbol{\beta}}_\lambda(S) - \tilde{\boldsymbol{\beta}}_{\lambda_0}(S))^{\mathrm{T}}\boldsymbol{X}_S^{\mathrm{T}}\bar{\boldsymbol{W}}_S\boldsymbol{X}_S(\tilde{\boldsymbol{\beta}}_\lambda(S) - \tilde{\boldsymbol{\beta}}_{\lambda_0}(S)), \tag{A.4}$$

where $\bar{\boldsymbol{W}}_S$ is a diagonal matrix with elements $b''(\bar{\theta}_i)/\phi$ calculated at $\bar{\boldsymbol{\beta}}_{\alpha_0}$, which lies on the line segment joining $\tilde{\boldsymbol{\beta}}_S$ and $\tilde{\boldsymbol{\beta}}_{\alpha_0}$. From (A.4) and using Conditions 3–4, we have

$$\begin{aligned}\min_{\alpha_\lambda \not\supset \alpha_0}\bigg(-\frac{2}{n}(\ell(\tilde{\boldsymbol{\beta}}_\lambda(S)) - \ell(\tilde{\boldsymbol{\beta}}_{\lambda_0}(S)))\bigg) &\geqslant 2c_0c_1\|\tilde{\boldsymbol{\beta}}_\lambda(S) - \tilde{\boldsymbol{\beta}}_{\lambda_0}(S)\|^2\\ &\geqslant 2c_0c_1(\|\tilde{\boldsymbol{\beta}}_\lambda(S) - \boldsymbol{\beta}^0(S)\|^2 - \|\tilde{\boldsymbol{\beta}}_{\lambda_0}(S) - \boldsymbol{\beta}^0(S)\|^2)\\ &\geqslant 2c_0c_1\bigg(\min_{j\in\alpha_0}\{\|\beta_j^0\|^2\} - O_p\bigg(\frac{p_n}{n}\bigg)\bigg),\end{aligned}$$

where the triangular inequality is used to go from the first to the second line, and the result $\|\tilde{\boldsymbol{\beta}}_{\lambda_0}(S) - \boldsymbol{\beta}^0(S)\|^2 = \|\tilde{\boldsymbol{\beta}}_{\lambda_0} - \boldsymbol{\beta}^0\|^2 = O_p(\sqrt{p_n/n})$ is used to go from the second to the third line. By Condition 4, we can straightforwardly show that the quantity $2c_0c_1(\min_{j\in\alpha_0}\{\|\beta_j^0\|^2\} - O_p(\frac{p_n}{n})) - \frac{2\nu p_0}{n}\log(n)$ is asymptotically positive. Combining the results of $J_1$ and $J_2$, we have

$$\mathrm{P}\Big(\inf_{\lambda\in\Omega_-}\min_{\alpha_\lambda\not\supset\alpha_0;|\alpha_\lambda|\leqslant K}\mathrm{ERIC}^*_\nu(\lambda) > \mathrm{ERIC}^*_\nu(\lambda_0)\Big) \to 1.$$

This relation completes the proof. □

*Proof of Lemma* 3.2(b). As in the proof of Lemma 3.2(a), we shall assume that $\phi = 1$ for simplicity. Since $\lambda_0$ satisfies Condition 5, by Lemma A.1 the adaptive group LASSO estimates are $\sqrt{n/p_n}$-estimation-consistent and selection-consistent. As before, let $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{y}|\boldsymbol{\beta})$ denote the log-likelihood function and

$|\omega_\lambda| = |\alpha_\lambda \cap \alpha_0^c|$ denote the number of over-fitted coefficients. Note it can take values $\omega_\lambda = 1, 2, \ldots, (g_n - g_0)$. We have

$$\begin{aligned}\mathrm{ERIC}^*_\nu(\lambda) - \mathrm{ERIC}^*_\nu(\lambda_0) &= -2(\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\tilde{\boldsymbol{\beta}}_{\alpha_0})) + 2\nu \sum_{j\in\alpha_\lambda} d_j \log\left(\frac{n}{\lambda}\right) - 2\nu \sum_{j\in\alpha_0} d_j \log\left(\frac{n}{\lambda_0}\right)\\ &\geqslant -2(\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\boldsymbol{\beta}^0_{\alpha_0})) + 2\nu \sum_{j\in\alpha_\lambda} d_j \log\left(\frac{n}{\lambda}\right) - 2\nu \sum_{j\in\alpha_0} d_j \log\left(\frac{n}{\lambda_0}\right)\\ &\geqslant -2(\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\boldsymbol{\beta}^0_{\alpha_0})) + 2\nu \sum_{j\in\omega_\lambda} d_j \log\left(\frac{n}{\lambda}\right) + 2\nu \sum_{j\in\alpha_0} d_j \log\left(\frac{\lambda_0}{\lambda}\right).\end{aligned}$$

By the definition of over-fitting, i.e., $\lambda_0/\lambda > 1$ and so $2\nu\sum_{j\in\alpha_0} d_j \log(\lambda_0/\lambda) > 0$. Next, using the Taylor expansion, we have

$$\begin{aligned}\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\boldsymbol{\beta}^0_{\alpha_0}) &= \nabla\ell(\boldsymbol{\beta}^0_{\alpha_\lambda})^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda} - \boldsymbol{\beta}^0_{\alpha_\lambda}) - \frac{n}{2}(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda} - \boldsymbol{\beta}^0_{\alpha_\lambda})^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{X}^{\mathrm{T}}_{\alpha_\lambda}\bar{\boldsymbol{W}}_{\alpha_\lambda}\boldsymbol{X}_{\alpha_\lambda}\right)(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda} - \boldsymbol{\beta}^0_{\alpha_\lambda})\\ &\leqslant \nabla\ell(\boldsymbol{\beta}^0_{\alpha_\lambda})^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda} - \boldsymbol{\beta}^0_{\alpha_\lambda}) - \frac{n}{2}c_0c_1\|\tilde{\boldsymbol{\beta}}_{\alpha_\lambda} - \boldsymbol{\beta}^0_{\alpha_\lambda}\|^2,\end{aligned}$$

where $\bar{\boldsymbol{W}}_{\alpha_\lambda}$ is a diagonal matrix with elements $b''(\bar{\theta}_i)/\phi$ calculated using $\bar{\boldsymbol{\beta}}_{\alpha_\lambda}$, which lies on the line segment joining $\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}$ and $\boldsymbol{\beta}^0_{\alpha_\lambda}$. By using a perfect square trinomial, we can show that

$$\ell(\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) - \ell(\boldsymbol{\beta}^0_{\alpha_0}) \leqslant \frac{1}{2nc_0c_1}\nabla\ell(\boldsymbol{\beta}^0_{\alpha_\lambda})^{\mathrm{T}}\nabla\ell(\boldsymbol{\beta}^0_{\alpha_\lambda}) \leqslant \frac{1}{2nc_0c_1}\|\boldsymbol{X}^{\mathrm{T}}_{\alpha_\lambda}(\boldsymbol{y} - \boldsymbol{\mu}^0)\|^2,$$

where $\boldsymbol{\mu}^0$ is calculated using $\boldsymbol{\beta}^0$. We thus have

$$\mathrm{ERIC}^*_\nu(\lambda) - \mathrm{ERIC}^*_\nu(\lambda_0) \geqslant -\frac{1}{nc_0c_1}\|\boldsymbol{X}^{\mathrm{T}}_{\alpha_\lambda}(\boldsymbol{y} - \boldsymbol{\mu}^0)\|^2 + 2\nu \sum_{j\in\omega_\lambda} d_j \log\left(\frac{n}{\lambda}\right). \tag{A.5}$$

Hence, we can prove uniformly for all $\lambda \in \Omega_+$ that

$$\mathrm{P}\left(\max_{\alpha_\lambda \not\supseteq \alpha_0; |\alpha_\lambda| \leqslant K}\left(\frac{1}{nc_0c_1}\|\boldsymbol{X}^{\mathrm{T}}_{\alpha_\lambda}(\boldsymbol{y} - \boldsymbol{\mu}^0)\|^2\right) > 2\nu \sum_{j\in\omega_\lambda} d_j \log\left(\frac{n}{\lambda}\right)\right) \to 0. \tag{A.6}$$

The required result will then follow immediately. To prove (A.6), we have

$$\begin{aligned}&\mathrm{P}\left(\frac{1}{nc_0c_1}\|\boldsymbol{X}^{\mathrm{T}}_{\alpha_\lambda}(\boldsymbol{y} - \boldsymbol{\mu}^0)\|^2 > 2\nu \sum_{j\in\omega_\lambda} d_j \log\left(\frac{n}{\lambda}\right)\right)\\ &\leqslant \sum_{l\in\alpha_\lambda}\sum_{k=1}^{d_l}\mathrm{P}\left(\frac{1}{\sqrt{nc_0c_1}}|\boldsymbol{x}^{k\mathrm{T}}_l(\boldsymbol{y} - \boldsymbol{\mu}^0)| > \sqrt{\frac{2\nu\sum_{j\in\omega_\lambda} d_j \log(\frac{n}{\lambda})}{d_l}}\right).\end{aligned} \tag{A.7}$$

First, for some $l \in \alpha_\lambda$, denote $\boldsymbol{x}^k_l = (x_{1k}, \ldots, x_{nk})^{\mathrm{T}}$. Consider the tail probability,

$$\mathrm{P}\left(\sum_{i=1}^n \frac{x_{ik}(y_i - \mu^0_i)}{\phi\sqrt{nc_0c_1}} \geqslant M_n\right) = \mathrm{P}\left(\sum_{i=1}^n \frac{a_{ik}(y_i - \mu^0_i)}{\phi} \geqslant M'_n\right)$$

for some $M_n > 0$, where $a_{ik} = x_{ik}/\sqrt{nc_0c_1\sum_{i=1}^n x^2_{ik}b''(\bar{\theta}_i)}$ and $M'_n = M_n/\sqrt{\sum_{i=1}^n x^2_{ik}b''(\bar{\theta}_i)}$, where $\bar{\theta}_i$ is defined shortly. By the exponential Chebychev's inequality, we have for some $t \geqslant 0$,

$$\begin{aligned}\mathrm{P}\left(\sum_{i=1}^n \frac{x_{ik}(y_i - \mu^0_i)}{\sqrt{nc_0c_1}} \geqslant M_n\right) &\leqslant \exp(-tM'_n)\prod_{i=1}^n \mathrm{E}(\exp(ta_{ik}(y_i - \mu^0_i)))\\ &\leqslant \exp(-tM'_n)\prod_{i=1}^n \exp(b(\theta^0_i + a_{ik}t) - b(\theta^0_i) - ta_{ik}\mu^0_i).\end{aligned}$$

$$\leqslant \exp(-tM_n')\prod_{i=1}^{n}\exp\left(\frac{1}{2}a_{ik}^2t^2b''(\bar{\theta}_i)\right)$$
$$\leqslant \exp(-tM_n')\exp\left(\frac{t^2}{2nc_0c_1}\right),$$

where $\bar{\theta}_i$ lies between $\theta_i^0$ and $(\theta_i^0+a_{ik}t)$. If we take $t=M_n'nc_0c_1$ and $M_n=\sqrt{\frac{2\nu\sum_{j\in\omega_\lambda}d_j}{d_l}\log(n/\lambda)}$, then for all $\lambda\in\Omega_+$ and fixed $\omega_\lambda$, we have

$$\mathrm{P}\Big(\sum_{i=1}^{n}\frac{x_{ik}(y_i-\mu_i^0)}{\sqrt{nc_0c_1}}\geqslant\sqrt{\frac{2\nu\sum_{j\in\omega_\lambda}d_j\log(n/\lambda)}{d_l}}\Big)\leqslant\exp\Big(-\frac{2\nu\sum_{j\in\omega_\lambda}d_j\log(n/\lambda)}{2d_l}\Big(\frac{nc_0c_1}{\sum_{i=1}^{n}x_{ik}^2b''(\bar{\theta}_i)}\Big)\Big)$$
$$\leqslant\exp\Big(\frac{-C\nu}{d_l}\sum_{j\in\omega_\lambda}d_j\log(n/\lambda)\Big),$$

for some arbitrary constant $C$ that is assumed to be greater than 1.

Since $\|\boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2/(nc_0c_1)$ is an even function about $(\boldsymbol{y}-\boldsymbol{\mu}^0)$, by (A.7), we have

$$\mathrm{P}\Big(\frac{\|\boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2}{nc_0c_1}\geqslant 2\nu\sum_{j\in\omega_\lambda}d_j\log\Big(\frac{n}{\lambda}\Big)\Big)\leqslant 2\sum_{l\in\alpha_\lambda}\sum_{k=1}^{d_l}\exp\Big(\frac{-C\nu}{d_l}\sum_{j\in\omega_\lambda}d_j\log\Big(\frac{n}{\lambda}\Big)\Big).\tag{A.8}$$

For any $\omega_\lambda$, there are less than $g_n^{|\omega_\lambda|}$ over-fitted models. Thus, by Boole's inequality and (A.8),

$$\mathrm{P}\Big(\max_{\alpha_\lambda\supsetneq\alpha_0;|\alpha_\lambda|\leqslant K}\Big(\frac{1}{nc_0c_1}\|\boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2\Big)>2\nu\sum_{j\in\omega_\lambda}d_j\log\Big(\frac{n}{\lambda}\Big)\Big)$$
$$\leqslant g_n^{|\omega_\lambda|}\times 2\sum_{l\in\alpha_\lambda}\sum_{k=1}^{d_l}\exp\Big(\frac{-C\nu}{d_l}\sum_{j\in\omega_\lambda}d_j\log\Big(\frac{n}{\lambda}\Big)\Big)$$
$$\leqslant O_p\Big(\sum_{l\in\alpha_\lambda}d_l\exp\Big\{|\omega_\lambda|\Big(\log(g_n)-\frac{\nu\sum_{j\in\omega_\lambda}d_j}{|\omega_\lambda|d_l}\log\Big(\frac{n}{\lambda}\Big)\Big)\Big\}\Big).$$

For $\lambda\in\Omega_+$, Condition 5(b) is not satisfied, i.e., $\frac{\lambda}{\sqrt{ng_n}}(\frac{n}{p_n})^{\gamma/2}\leqslant C$ for a constant $C$. Thus, based on Condition 2, with a probability approaching 1, we have

$$\sum_{l\in\alpha_\lambda}d_l\exp\Big\{|\omega_\lambda|\Big(\log(g_n)-\frac{\nu\sum_{j\in\omega_\lambda}d_j}{|\omega_\lambda|d_l}\log\Big(\frac{n}{\lambda}\Big)\Big)\Big\}$$
$$\leqslant Kd_{\max}\exp\Big\{|\omega_\lambda|\Big(\log(g_n)-\frac{\nu}{2}\Big(\log\Big(\frac{n}{g_n}\Big)+\gamma\log\Big(\frac{n}{p_n}\Big)\Big)\Big)\Big\}$$
$$\leqslant Kd_{\max}\exp\Big\{|\omega_\lambda|\log(n)\Big(\kappa_1-\frac{\nu}{2}(1-\kappa_1)-\frac{\nu\gamma}{2}(1-\kappa_1-\kappa_2)\Big)\Big\}$$
$$\leqslant Kd_{\max}\exp\{\zeta|\omega_\lambda|\log(n)\}\quad\Big(\text{where}\quad\zeta=\kappa_1-\frac{\nu}{2}(1-\kappa_1)-\frac{\nu\gamma}{2}(1-\kappa_1-\kappa_2)\Big)$$
$$\leqslant\frac{Kd_{\max}}{n^{-\zeta}}\to 0\quad\text{if}\quad\Big(\gamma\geqslant\frac{2\kappa_1+(2-\nu)\kappa_2}{\nu(1-\kappa_1-\kappa_2)}-1\Big).$$

By the condition $K=o(n^{\iota/2})$, we have a limit that converges to 0.

Thus,

$$\mathrm{P}\Big(\max_{\alpha_\lambda\supsetneq\alpha_0;|\alpha_\lambda|\leqslant K}\Big(\frac{1}{nc_0c_1}\|\boldsymbol{X}^{\mathrm{T}}\mathrm{T}_{\alpha_\lambda}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2\Big)>2\nu\sum_{j\in\omega_\lambda}d_j\log\Big(\frac{n}{\lambda}\Big)\Big)\leqslant o_p(1),$$

for all $\lambda\in\Omega_+$, as required in (A.6). Consequently, we have

$$\mathrm{P}\Big(\inf_{\lambda\in\Omega_+}\min_{\alpha_\lambda\not\supseteq\alpha_0}\mathrm{ERIC}_\nu^*(\lambda)>\mathrm{ERIC}_\nu^*(\lambda_0)\Big)\to 1.$$

This relation completes the proof. □

## Appendix A.3 Proof of Theorem 3.4

*Proof of Theorem* 3.4. As before, let $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{y} \,|\, \boldsymbol{\beta})$ denote the log-likelihood function. Also, we let $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}_{\alpha_\lambda}, \hat{\boldsymbol{\beta}}_{\alpha_\lambda^c} = \mathbf{0})$, and $\tilde{\boldsymbol{\beta}}_\lambda = (\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}, \tilde{\boldsymbol{\beta}}_{\alpha_\lambda^c} = \mathbf{0})$, noting that $\alpha_{\lambda_0} = \alpha_0$.

We first prove the following result linking the log-likelihood functions evaluated at the penalized $\hat{\boldsymbol{\beta}}_{\lambda_0}$ and unpenalized $\tilde{\boldsymbol{\beta}}_{\lambda_0}$ estimators:

$$\frac{1}{n}\ell(\hat{\boldsymbol{\beta}}_{\lambda_0}) = \frac{1}{n}\ell(\tilde{\boldsymbol{\beta}}_{\lambda_0}) + o_p\bigg(\frac{1}{n}\bigg). \tag{A.9}$$

To prove (A.9), consider the following Taylor expansion:

$$\begin{aligned}\frac{1}{n}(\ell(\hat{\boldsymbol{\beta}}_{\lambda_0}) - \ell(\tilde{\boldsymbol{\beta}}_{\lambda_0})) &= \frac{1}{n}(\ell(\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}}) - \ell(\tilde{\boldsymbol{\beta}}_{\alpha_0})) \\ &= -\frac{1}{2}(\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}} - \tilde{\boldsymbol{\beta}}_{\alpha_0})^{\mathrm{T}}\bigg(\frac{1}{n}\boldsymbol{X}_{\alpha_0}^{\mathrm{T}}\bar{\boldsymbol{W}}_{\alpha_0}\boldsymbol{X}_{\alpha_0}\bigg)(\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}} - \tilde{\boldsymbol{\beta}}_{\alpha_0}) \\ &\geqslant -\frac{c_2}{2c_0}\|\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}} - \tilde{\boldsymbol{\beta}}_{\alpha_0}\|^2,\end{aligned} \tag{A.10}$$

where $\bar{\boldsymbol{W}}_{\alpha_0}$ is a diagonal matrix with elements $b''(\bar{\theta}_i)/\phi$ calculated by using $\bar{\boldsymbol{\beta}}_{\alpha_0}$, which lies on the line segment joining $\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}}$ and $\tilde{\boldsymbol{\beta}}_{\alpha_0}$. For GLMs with a canonical link function, we have, by definition with a Taylor expansion,

$$\begin{aligned}&\nabla\ell(\boldsymbol{\beta}_{\alpha_0}^0) + \nabla^2\ell(\bar{\boldsymbol{\beta}}_{\alpha_0})(\tilde{\boldsymbol{\beta}}_{\alpha_0} - \boldsymbol{\beta}_{\alpha_0}^0) = \mathbf{0}, \\ &\nabla\ell(\boldsymbol{\beta}_{\alpha_0}^0) + \nabla^2\ell(\bar{\bar{\boldsymbol{\beta}}}_{\alpha_0})(\hat{\boldsymbol{\beta}}_{\alpha_0} - \boldsymbol{\beta}_{\alpha_0}^0) = \lambda_0 p'(\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}}),\end{aligned}$$

where $p'(\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}})$ is an $|\alpha_0| \times 1$ vector with elements $\{\sqrt{d_j}\tilde{w}_j\frac{\hat{\beta}_j^{\mathrm{T}}}{\|\hat{\beta}_j\|}; j \in \alpha_0\}$. Hence, by Condition 3, we have

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}_{\alpha_0} - \hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}}\|^2 &= \bigg\|\frac{\lambda_0}{nc_0}\bigg(\frac{1}{n}\boldsymbol{X}_{\alpha_0}^{\mathrm{T}}\boldsymbol{X}_{\alpha_0}\bigg)^{-1}p'(\hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}})\bigg\|^2 + o_p(1) \\ &\leqslant O_p\bigg(\bigg(\frac{\lambda_0}{n}\bigg)^2\sum_{j\in\alpha_0} d_j\tilde{\omega}_j^2\bigg) \leqslant O_p\bigg(\bigg(\frac{\lambda_0}{n}\bigg)^2\frac{p_0}{\min_{j\in\alpha_0}\|\tilde{\beta}_j\|^{2\gamma}}\bigg).\end{aligned}$$

It follows that $\|\tilde{\boldsymbol{\beta}}_{\alpha_0} - \hat{\boldsymbol{\beta}}_{\alpha_{\lambda_0}}\|^2 = O_p((\lambda_0/n)^2 p_0)$. Applying this result to (A.10), we obtain

$$-\frac{c_2}{2c_0}O_p\bigg(\bigg(\frac{\lambda_0}{n}\bigg)^2 p_0\bigg) \leqslant \frac{1}{n}(\ell(\hat{\boldsymbol{\beta}}_{\lambda_0}) - \ell(\tilde{\boldsymbol{\beta}}_{\lambda_0})) < 0.$$

By Condition 5(a), we have $(\lambda_0/n)^2 p_0 = o_p(1/n)$ and, thus, the left-hand side of the inequality above approaches 0. The result in (A.9) follows immediately from this.

Given the forms of $\mathrm{ERIC}_\nu(\lambda)$ and $\mathrm{ERIC}_\nu^*(\lambda)$ and using the result in (A.9), we can write

$$\frac{\mathrm{ERIC}_\nu(\lambda) - \mathrm{ERIC}_\nu(\lambda_0)}{n} \geqslant \frac{\mathrm{ERIC}_\nu^*(\lambda) - \mathrm{ERIC}_\nu^*(\lambda_0)}{n} + o_p\bigg(\frac{1}{n}\bigg).$$

From Lemma 3.2, we know that $\mathrm{ERIC}_\nu^*(\lambda) - \mathrm{ERIC}_\nu^*(\lambda_0)$ is guaranteed to be asymptotically positive for both over- and under-fitted models. It follows immediately that

$$\mathrm{P}\Big(\inf_{\lambda\in\Omega_-\cup\Omega_+}\mathrm{ERIC}_\nu(\lambda) > \mathrm{ERIC}_\nu(\lambda_0)\Big) \to 1.$$

By Lemma A.1, $\lambda_0 \in \Omega_0$ asymptotically identifies the true model $\alpha_0$; therefore, $\mathrm{P}(\alpha_{\hat{\lambda}} = \alpha_0) \to 1$ . □

### Appendix A.4 Proof of Corollary 3.5

*Proof of Corollary* 3.5. Define a proxy version of $\mathrm{BIC}(\lambda)$ as $\mathrm{BIC}^*(\lambda) = -2\ell(\boldsymbol{y}\,|\,\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}) + \sum_{j\in\alpha_\lambda} d_j \log(n)$. Following a development similar to the proof of Lemma 3.2(b), we obtain a result analogous to the one in (A.5),

$$\mathrm{BIC}^*(\lambda) - \mathrm{BIC}^*(\lambda_0) \geqslant -\frac{1}{nc_0c_1}\|\boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2 + \sum_{j\in\omega_\lambda} d_j\log(n).$$

Furthermore, similar to (A.8), we obtain the following result on tail probability:

$$\mathrm{P}\bigg(\frac{1}{nc_0c_1}\|\boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2 \geqslant \sum_{j\in\alpha_\lambda} d_j\log(n)\bigg) \leqslant 2\sum_{l\in\alpha_\lambda}\sum_{k=1}^{d_l}\exp\bigg(-\frac{C\sum_{j\in\alpha_\lambda} d_j\log(n)}{2d_l}\bigg).$$

For a fixed $\omega_\lambda$, less than $g_n^{|\omega_\lambda|}$ over-fitted models are available. Thus, by Condition 2,

$$\begin{aligned}&\mathrm{P}\bigg(\max_{\alpha_\lambda\not\supseteq\alpha_0;|\alpha_\lambda|\leqslant K}\bigg(\frac{1}{nc_0c_1}\|\boldsymbol{X}_{\alpha_\lambda}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\mu}^0)\|^2\bigg) > \sum_{j\in\alpha_\lambda} d_j\log(n)\bigg)\\ &\quad\leqslant \sum_{j\in\alpha_\lambda} d_j \times g_n^{|\omega_\lambda|}\times 2\exp\bigg(-\frac{C\sum_{j\in\omega_\lambda} d_j\log(n)}{2d_l}\bigg) \leqslant O_p\bigg(\exp\bigg\{|\omega_\lambda|\log(n)\bigg(\kappa_1-\frac{1}{2}\bigg)\bigg\}\bigg).\end{aligned}$$

Hence, if $\kappa_1 > \frac{1}{2}$, then the above probability does not approach 0. Following a development of proof similar to that of Theorem 3.4, we obtain $\mathrm{P}(\alpha_{\hat{\lambda}} = \alpha_0) \not\to 1$. □

### Appendix A.5 Proof of Lemma 3.8

The proof below follows an approach similar to that in Fan and Peng [5]. Considering the same note in the proof of Lemma A.1, we can easily obtain estimation consistency. We can obtain selection consistency via the same approach applied to Lemma A.1. Thus, the proof is completed.

**Lemma A.2.** *Assume Conditions* 2–3 *and* 4′ *are satisfied and that there exists a* $\lambda_0 \in \Omega_0$ *satisfying Condition* 5′. *Then,*

(a) $\mathrm{P}(\inf_{\lambda\in\Omega_-}\min_{\alpha_\lambda\not\supset\alpha_0}\mathrm{ERIC}_\nu^*(\lambda) > \mathrm{ERIC}_\nu^*(\lambda_0)) \to 1$.

(b) $\mathrm{P}(\inf_{\lambda\in\Omega_+}\min_{\alpha_\lambda\not\supseteq\alpha_0}\mathrm{ERIC}_\nu^*(\lambda) > \mathrm{ERIC}_\nu^*(\lambda_0)) \to 1$, *if* $\nu \geqslant \frac{2(\kappa_1+\kappa_2)}{1-\kappa_1}$.

*Proof.* The proof is similar to that of Lemma 3.2; thus, we omit it here. □

### Appendix A.6 Proof of Theorem 3.9

*Proof of Theorem* 3.9. As before, let $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{y}\,|\,\boldsymbol{\beta})$ denote the log-likelihood function. Also, we let $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}_{\alpha_\lambda}, \hat{\boldsymbol{\beta}}_{\alpha_\lambda^{\mathrm{c}}} = \boldsymbol{0})$, and $\tilde{\boldsymbol{\beta}}_\lambda = (\tilde{\boldsymbol{\beta}}_{\alpha_\lambda}, \tilde{\boldsymbol{\beta}}_{\alpha_\lambda^{\mathrm{c}}} = \boldsymbol{0})$, noting that $\alpha_{\lambda_0} = \alpha_0$. We first prove the following result linking the log-likelihood functions evaluated at the penalized $\hat{\boldsymbol{\beta}}_{\lambda_0}$ and unpenalized $\tilde{\boldsymbol{\beta}}_{\lambda_0}$ estimators:

$$\frac{1}{n}\ell(\hat{\boldsymbol{\beta}}_{\lambda_0}) = \frac{1}{n}\ell(\tilde{\boldsymbol{\beta}}_{\lambda_0}) + o_p(1). \tag{A.11}$$

We can prove (A.11) via the same approach as that applied to Theorem 3.4. Given the forms of $\mathrm{ERIC}_\nu(\lambda)$ and $\mathrm{ERIC}_\nu^*(\lambda)$ and using the result in (A.11), we can write

$$\frac{\mathrm{ERIC}_\nu(\lambda) - \mathrm{ERIC}_\nu(\lambda_0)}{n} \geqslant \frac{\mathrm{ERIC}_\nu^*(\lambda) - \mathrm{ERIC}_\nu^*(\lambda_0)}{n} + o_p(1).$$

From Lemma A.2, it follows immediately that $\mathrm{P}(\inf_{\lambda\in\Omega_-\cup\Omega_+}\mathrm{ERIC}_\nu(\lambda) > \mathrm{ERIC}_\nu(\lambda_0)) \to 1$. By Lemma 3.8, $\lambda_0 \in \Omega_0$ asymptotically identifies the true model $\alpha_0$; therefore, $\mathrm{P}(\alpha_{\hat{\lambda}} = \alpha_0) \to 1$. □

## Appendix A.7

**Lemma A.3.** *Assume Conditions* 6–10 *are satisfied and define*

$$\hat{\boldsymbol{\beta}}_{\alpha_0} = \arg\max_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{y} \,|\, \boldsymbol{\beta}) - \lambda \sum_{j \in \alpha_0} \sqrt{d_j} \tilde{w}_j \|\beta_j\| \right\}.$$

*Then, with a probability approaching* 1, $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}'_{\alpha_0}, \mathbf{0}')'$ *is the adaptive group LASSO estimate to* (2.2) *that satisfies* (a) *estimate consistency*: $\|\hat{\boldsymbol{\beta}}_{\alpha_0} - \boldsymbol{\beta}^0_{\alpha_0}\| = O_p(\sqrt{p_0/n})$ *and* (b) *selection consistency*: $\mathrm{P}(\{j : \|\hat{\beta}_{\lambda,j}\| = 0\} = \alpha_0^c) \to 1$.

*Proof.* The proof is similar to the proof of Lemma A.1; thus, it is omitted here. □

*Proof of Theorem* 4.1. The proof is nearly identical to that for Theorem 3.4 except for the condition $\nu > c \log(g_n)/\log(n)$ for some $1 < c < 2$; thus, the proof is omitted here. □