Variable selection and estimation in generalized linear models with the seamless L_0 penalty

Zilin Li[1], Sijian Wang[2] and Xihong Lin[3]
[1]Department of Mathematics, Tsinghua University, Beijing China
[2]Department of Biostatistics & Medical Informatics and Statistics, University of Wisconsin, Madison
[3]Department of Biostatistics, Harvard University
xlin@hsph.harvard.edu

August 6, 2012

Abstract

In this paper, we propose variable selection and estimation in generalized linear models using the seamless L_0 (SELO) penalized likelihood approach. The SELO penalty is a smooth function that very closely resembles the discontinuous L_0 penalty. We develop an efficient algorithm to fit the model, and show that the SELO-GLM procedure has the oracle property in the presence of a diverging number of variables. We propose a Bayesian Information Criterion (BIC) to select the tuning parameter. We show that under some regularity conditions, the proposed SELO-GLM/BIC procedure consistently selects the true model. We perform simulation studies to evaluate the finite sample performance of the proposed methods. Our simulation studies show that the proposed SELO-GLM procedure has a better finite sample performance than several existing methods, especially when the number of variables is large and the signals are weak. We apply the SELO-GLM to analyze a breast cancer genetic dataset to identify the SNPs that are associated with breast cancer risk.

Key words: BIC; Consistency; Coordinate descent algorithm; Model selection; Oracle property; Penalized likelihood methods; SELO penalty; Tuning parameter selection.

1 Introduction

Generalized linear models (GLMs) (McCullagh and Nelder 1989) provide a flexible framework to study the association between a family of continuous and discrete outcomes and a set of independent variables. In modern health science studies, such as genomic studies, a large number of predictors are often collected. For example, in genome-wide association studies, tens of thousands to millions of Single Nucleotide Polymorphisms (SNPs) are measured across the genome (Hunter et al. 2007). One is interested in identifying a subset of SNPs that are associated with disease outcomes. Effective variable selection can also lead to parsimonious models with better prediction accuracy and easier interpretation.

Penalized likelihood methods provide an attractive approach to perform variable selection and regression coefficient estimation by simultaneously identifying a subset of variables that are associated with a response. An important class of penalized likelihood approaches is based on the L_0 penalty function. For example, C_p (Mallows 1973), AIC (Akaike 1974), BIC (Schwarz 1978) and RIC (Foster and George 1994) are all motivated from L_0 penalized likelihood regression. The L_0 penalty directly penalizes the number of non-zero coefficients in the model, and is intuitively suitable for the purpose of variable selection. However, there are two major drawbacks of the L_0 penalized likelihood procedure. First, because the L_0 penalty is not continuous at the origin point, the resulting estimators are likely to be unstable (Breiman 1996). Second, implementation of the L_0 penalized likelihood procedure is NP-hard and involves an exhaustive search over all possible models. As a consequence, the L_0 penalized estimation procedure is computationally infeasible when the number of potential predictors is moderate or large.

As alternatives to L_0 penalty, penalized likelihood estimation based on continuous penalty functions has attracted lots of attention in the recent literature. Tibshirani (1996) proposed a LASSO method, which uses the L_1 penalty for variable selection. The L_1 penalty approximates the L_0-norm of regression coefficients (i.e., the number of nonzero elements) by its L_1-norm, and contiguously shrinks the estimated coefficients toward zero to identify important variables. However, LASSO may not be consistent for model selection (Zhao and Yu 2006; Zou 2006) and the estimated regression coefficients are often not asymptotically normally distributed (Knight and Fu 2000). Fan and Li (2001) proposed a non-concave SCAD penalty which penalizes large coefficients less and hence reduces estimation bias. Under generalized linear models, these authors also proved that, when the tuning parameter is properly selected, the SCAD procedure consistently identifies the true model, and the estimators for the non-zero coefficients in the true model have the same asymptotic normal

distribution as if the true model were known, i.e., it has the "oracle" property. The results in Fan and Li (2001) require the number of variables p is fixed as sample size n goes to infinity. Fan and Peng (2004) extended the results to allow both n and p tend to infinity with $p^5/n \to 0$. Zou (2006) proposed an adaptive LASSO penalty which is a weighted version of LASSO penalty. When the weights are properly constructed, estimation bias can be reduced. Under linear models, Zou (2006) showed that the adaptive LASSO method has the "oracle" property with fixed p, and Zou and Zhang (2009) extended the results to allow both n and p tend to infinity with $\log p/\log n \to \nu \in [0, 1)$. Zhang (2010) proposed a non-concave MCP penalty and developed a MC+ penalized likelihood procedure ("+" refers to the algorithm used for implementing MCP). Under linear models, Zhang (2010) proved that the MC+ procedure may select the correct model with probability tending to 1 and the corresponding estimators have good properties in terms of L^p-loss.

Different from these penalized likelihood methods, Dicker et al. (2012) proposed a seamless L_0 (SELO) penalty to explicitly mimic the L_0 penalty while addressing the two issues associated with the L_0 penalized likelihood procedure discussed above. The SELO penalty very closely approximates the L_0 penalty function and hence is able to effectively perform variable selection. The continuity of the SELO penalty implies that the SELO estimators are likely to be more stable than those obtained through the L_0 procedure, and allows the efficient coordinate descent algorithm (Friedman et al. 2007; Wu and Lange 2008) to be implemented to calculate regression coefficient estimators. Under linear models, Dicker et al. (2012) showed that SELO estimators enjoy the "oracle" property when both p and n tend to infinity with $p/n \to 0$. They also proposed a SELO/BIC procedure to select the tuning parameters and showed it is consistent for model selection. Their simulation studies showed that the SELO method outperforms the existing penalized likelihood methods especially for weaker signals.

In this paper, we develop a SELO penalized likelihood method for variable selection in generalized linear models. We denote the corresponding procedure to be SELO-GLM. We show that under GLMs, the SELO-GLM procedure enjoys the "oracle" property when both p and n tend to infinity with $p^5/n \to 0$. We propose a BIC for selecting the tuning parameters. We show that, when the tuning parameters are selected using the BIC, with probability tending to 1, the SELO-GLM procedure identifies the correct model. We propose an efficient algorithm using the coordinate descent algorithm to calculate the SELO-GLM estimator. We conduct simulation studies to evaluate the finite sample performance of the proposed method and compare it with the existing methods. We also apply the SELO-GLM method to analysis of a breast cancer genetic dataset to identify SNPs which are

associated with breast cancer risk.

The remaining of the paper is organized as follows. Section 2 describes our proposed SELO-GLM method and the corresponding coordinate descent algorithm to calculate the estimator. Section 3 presents the theoretical properties of the SELO-GLM procedure, as well as proposing the BIC for estimating the tuning parameter and studying its theoretical properties. Section 4 and Section 5 demonstrate the finite sample performance of the SELO-GLM method using simulation studies and real data analysis, respectively. We conclude the paper with Section 6.

2 Methods

2.1 Variable selection via penalized likelihood

Suppose that a random sample of n subjects is observed. Let Y_i be a response and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ be a vector of p predictors for the ith subject. Assume Y_i follows a distribution in the exponential family with mean $\mu_i = E(Y_i)$ and variance $V_i = var(Y_i) = a(\phi)V(\mu_i)$, where ϕ is a scale parameter, $a(\cdot)$ is a known function, and $v(\cdot)$ is a variance function. GLMs model the mean μ_i of Y_i as a function of covariates through a known monotone link function g:

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{ip}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p+1) \times 1$ vector of unknown regression coefficients.

The density function of Y_i in the exponential family is

$$L(Y_i; \theta_i, \phi) = \exp\left[\{Y_i\theta_i - b(\theta_i)\}/\{a(\phi)\} + c(Y_i, \phi)\right], \tag{2}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot)$ are known functions and θ_i is a canonical parameter and satisfies $\mu_i = b'(\theta_i)$. We hence write θ_i below as $\theta(\boldsymbol{X}_i, \boldsymbol{\beta})$ to denote it is a function of $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$. For binary data, the dispersion parameter $\phi = 1$. For normal data, $a(\phi)$ is the residual variance σ^2.

The Penalized likelihood estimation solves the following optimization problem:

$$\min_{\boldsymbol{\beta}} -\frac{1}{n}\ell_n(\boldsymbol{\beta}) + p(\boldsymbol{\beta}), \tag{3}$$

where

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Big[Y_i\theta(\boldsymbol{X}_i, \boldsymbol{\beta}) - b\{\theta(\boldsymbol{X}_i, \boldsymbol{\beta})\}\Big] \tag{4}$$

and $p(\boldsymbol{\beta})$ is a certain penalty function that depends on a tuning parameter λ.

Tibshirani (1996) proposed the L_1-penalty (LASSO penalty):

$$p_{LASSO}(\boldsymbol{\beta}) = \lambda\sum_{j=1}^{p}|\beta_j|. \tag{5}$$

Fan and Li (2001) proposed the SCAD penalty that satisfies

$$p'_{SCAD}(\beta_j) = \lambda \operatorname{sgn}(\beta_j)\Big[I\{|\beta_j| \le \lambda\} + \frac{\max\{a\lambda - \beta_j, 0\}}{(a-1)\lambda} I\{|\beta_j| > \lambda\}\Big], \tag{6}$$

where $a > 2$ is another tuning parameter. Fan and Li (2001) recommended setting $a = 3.7$. Essentially, the SCAD penalty is a quadratic spline with knots at $\pm\lambda$ and $\pm a\lambda$.

Zou (2006) proposed the adaptive LASSO penalty:

$$p_{ALASSO}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} w_j |\beta_j|, \tag{7}$$

where w_j's are data-dependent weights. If the coefficient of a variable is non-zero, its weight is expected to be small and the corresponding coefficient is penalized less. If the coefficient of a variable is zero, its weight is expected to be large and the corresponding coefficient is penalized more. A typical weight, if $p < n$, is $w_j = 1/\hat{\beta}_{MLE,j}$, where $\hat{\beta}_{MLE,j}$ is the unpenalized maximum likelihood estimator.

Zhang (2010) proposed the minimax concave penalty (MCP) :

$$p_{MCP}(\beta_j) = \lambda\Big[|\beta_j| - \frac{\beta_j^2}{2\gamma\lambda} I\{0 \le |\beta_j| < \gamma\lambda\} + \frac{\lambda^2\gamma}{2} I\{|\beta_j| \ge \gamma\lambda\}\Big], \tag{8}$$

where $\gamma > 0$ is another tuning parameter. Similar to the SCAD penalty, the MCP penalty is also a quadratic spline. The parameter γ determines the concavity of the penalty function. When $\gamma \to \infty$, the MCP penalty becomes the LASSO penalty, and when $\gamma \to 0^+$, the MCP penalty becomes the L_0 penalty.

2.2 SELO penalty and SELO-GLM method

The L_0 penalty has the form

$$p_{L_0}(\beta_j) = \lambda I\{\beta_j \ne 0\}, \tag{9}$$

and directly penalizes the number of non-zero parameters. Although the L_0 penalty is intuitively suitable for variable selection and has desirable theoretical properties, the associated variable selection and estimation procedures tend to be unstable, especially when the data contains only weak signals, and are computationally infeasible for even moderately large p, as implementation generally requires a combinatorial search. This is largely due to the fact that the L_0 penalty is discontinuous.

To overcome these limitations of the L_0 penalty, Dicker et al. (2012) introduced a continuous approximation to the L_0 penalty: the SELO penalty, which is defined by

$$p_{SELO}(\beta_j) = \frac{\lambda}{\log(2)} \log\Big(\frac{|\beta_j|}{|\beta_j| + \tau} + 1\Big), \tag{10}$$

where $\lambda \geq 0$ and $\tau > 0$ are two tuning parameters. We can see that when τ is small, $p_{SELO}(\beta_j) \approx \lambda \boldsymbol{I}\{\beta_j \neq 0\}$. Figure 1 of Dicker et al. (2012) compares the SELO penalty function with the SCAD, L_1 (LASSO) and L_0 penalties (left panel) and the MCP penalty, for various values of γ (right panel). It shows that the SELO penalty mimics the L_0 penalty much more closely than the L_1, SCAD and MC+ penalties. Since the SELO penalty is continuous at the origin point, the corresponding estimation procedure is more stable than the L_0 procedure. Furthermore, in the next section, we will see that the corresponding optimization problem can be efficiently solved by a combination of Newton-Raphson and coordinate descent algorithms.

In this paper, we apply the SELO penalty to GLMs and propose a SELO-GLM method by considering the following optimization problem

$$\min_{\boldsymbol{\beta}} -\frac{1}{n}\ell_n(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} p_{SELO}(\beta_j). \tag{11}$$

2.3 The Coordinate Descent Algorithm for Fitting SELO-GLM

When there is no penalty, the negative log-likelihood in the objective function $-\ell_n(\boldsymbol{\beta})$ is usually minimized by the Newton-Raphson method, which amounts to iteratively reweighted least squares (IWLS). Specifically, if the current estimators of the regression parameters are $\tilde{\boldsymbol{\beta}}$, we form a quadratic approximation to the negative log-likelihood (the second order Taylor expansion at the current estimates), which is

$$\ell_Q(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n} W_i(Z_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + R(\tilde{\boldsymbol{\beta}}), \tag{12}$$

where Z_i is the GLM working vector, W_i is the GLM weight,

$$Z_i = \eta_i + (Y_i - \mu_i)\frac{d\eta_i}{d\mu_i}, \quad W_i = \Big[\Big(\frac{d\eta_i}{d\mu_i}\Big)^2 V_i\Big]^{-1},$$

$$\eta_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j, \mu_i = g^{-1}(\eta_i), V_i = V(\mu_i) = a(\phi_i)b''(\boldsymbol{X}_i, \boldsymbol{\beta}),$$

and Z_i, μ_i, W_i, η_i are all evaluated at $\tilde{\boldsymbol{\beta}}$, and $R(\tilde{\boldsymbol{\beta}})$ is a term that does not depend on $\boldsymbol{\beta}$. The update in Newton-Raphson algorithm is obtained by minimizing ℓ_Q iteratively until convergence.

We use this iterative approach to minimize the SELO-GLM penalized negative log-likelihood. For each value of λ, we compute the quadratic approximation ℓ_Q about the SELO-GLM penalized negative log-likelihood at the current estimates $\tilde{\boldsymbol{\beta}}$. Then We minimize ℓ_Q by solving the penalized weighted least-squares problem:

$$\min_{\boldsymbol{\beta}} \ell_Q(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n} W_i(Z_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} p_{SELO}(\beta_j). \tag{13}$$

Following Dicker et al. (2012), we apply the coordinate descent algorithm to solve the above optimization problem. The idea of the coordinate descent algorithm is to find a local optima of a multivariate optimization problem by solving a sequence of univariate optimization problems. This may be very effective if the univariate optimization problem is simple.

For fixed $\tilde{\boldsymbol{\beta}}_{-k} = (\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1}, \tilde{\beta}_{k+1}, \ldots, \tilde{\beta}_p)$, define the univariate function by $\ell_Q^{(k)} = \ell_Q(\beta_k; \tilde{\boldsymbol{\beta}}_{-k})$. Given $\boldsymbol{\beta}_{-k}$, it is straightforward to minimize $\ell_Q(\beta_k; \tilde{\boldsymbol{\beta}}_{-k})$ with respect to β_k. Indeed, finding critical points of $\ell_Q(\beta_k; \tilde{\boldsymbol{\beta}}_{-k})$ $(k > 0)$ amounts to find the roots of cubic equations (when $k = 0$, the update of β_0 is trivial). We apply the Cardano's formula to calculate the explicit solution. The coordinate descent algorithm is implemented by minimizing $\ell_Q(\beta_k; \tilde{\boldsymbol{\beta}}_{-k})$ with respect to β_k for each k a time, and using the solution to update $\boldsymbol{\beta}$; at the next step, $\ell_Q^{(k+1)} = \ell_Q(\beta_{k+1}; \tilde{\boldsymbol{\beta}}_{-(k+1)})$ is minimized and the minimizer is again used to update $\boldsymbol{\beta}$. In this way, we cycle through the indices $k = 0, \ldots, p$; this may be performed multiple times until some convergence criterion is reached. In general, determining the theoretical convergence properties of algorithms for non-convex minimization problems is difficult. However, it is clear that $\ell_Q^{(k+1)} \leq \ell_Q^{(k)}$. In practice, we use the estimator from unpenalized GLMs when $p < n$ or tuned ridge penalized GLMs when $p > n$ as the initial value, and have found that numerical convergence of the algorithm generally occurs rapidly.

3 Theoretical Properties

3.1 Problem setup and notation

We assume that the data $\boldsymbol{D}_i = (\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$ are independent and identically distributed. Conditional on $\boldsymbol{X}_i$, Y_i has a density $f(\boldsymbol{X}_i^T \boldsymbol{\beta}, Y_i)$. Let $\boldsymbol{\beta}_n^* = (\beta_{n0}^*, \beta_{n1}^*, \ldots, \beta_{np}^*)$ be the underlying true parameters. Let $A_n = \{j : \beta_{nj}^* \neq 0\}$ be the set of indices of signal variables and denote its cardinality by $|A_n|$. In $\boldsymbol{\beta}_n^*$, A_n and other similar notation introduced later, the subscript n means that the true values of these parameters may change as n changes as we allow p go to infinity. For given λ and τ, the SELO-GLM estimator is obtained by minimizing (11).

We mainly present two theoretical properties of the SELO-GLM procedure. The first result implies that, under certain regularity conditions, the SELO-GLM estimator has the "oracle" property when λ_n and τ_n are properly tuned. That is, SELO-GLM consistently selects the correct model, and, by restricting the parameters to the set of the correct model, the SELO-GLM estimator is asymptotically normal and has the same asymptotic variance as the unpenalized generalized linear model estimator based on the correct model (unknown in practice). The second result is about the tuning parameter selection. Under certain regularity conditions, we show that when the tuning

parameters are selected based on BIC, the SELO-GLM procedure consistently identifies the correct model. For both results, we allow the number of variable diverges, i.e., both n and p tend to infinity.

3.2 The "Oracle" property of SELO-GLM

We first introduce the following regularity conditions:

(A) The observations $\boldsymbol{D}_i = (\boldsymbol{X}_i, \boldsymbol{Y}_i), i = 1, \cdots, n$, are independent and identically distributed with the probability density $f(\boldsymbol{D}, \boldsymbol{\beta}_n)$, which has a common support, and the model is identifiable. Furthermore, the first and second derivatives of the likelihood function satisfy the following equations:

$$E_{\boldsymbol{\beta}_n}\left\{\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n)}{\partial \beta_{nj}}\right\} = 0, \quad j = 0, 1, \cdots, p$$
$$E_{\boldsymbol{\beta}_n}\left\{\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n)}{\partial \beta_{nj}}\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n)}{\partial \beta_{nk}}\right\} = -E_{\boldsymbol{\beta}_n}\left\{\frac{\partial^2 \log f(\boldsymbol{D}, \boldsymbol{\beta}_n)}{\partial \beta_{nj}\partial \beta_{nk}}\right\}, \quad j, k = 0, 1, \ldots, p.$$

(B) The Fisher information matrix

$$I_n(\boldsymbol{\beta}_n^*) = E\left\{\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n^*)}{\partial \boldsymbol{\beta}})(\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n^*)}{\partial \boldsymbol{\beta}})^T\right\}$$

Then there exists $C_1, C_2, C_3, C_4 > 0$ such that:

$$0 < C_1 < \lambda_{\min}\{I_n(\boldsymbol{\beta}_n^*)\} < \lambda_{\max}\{I_n(\boldsymbol{\beta}_n^*)\} < C_2 < \infty$$
$$E_{\boldsymbol{\beta}_n}\left\{\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n^*)}{\partial \beta_{nj}}\frac{\partial \log f(\boldsymbol{D}, \boldsymbol{\beta}_n^*)}{\partial \beta_{nk}}\right\}^2 < C_3 < \infty, \quad j, k = 0, 1, \ldots, p$$
$$E_{\boldsymbol{\beta}_n}\left\{\frac{\partial^2 \log f(\boldsymbol{D}, \boldsymbol{\beta}_n^*)}{\partial \beta_{nj}\partial \beta_{nk}}\right\}^2 < C_4 < \infty, \quad j, k = 0, 1, \ldots, p,$$

where $\lambda_{\min}\{I_n(\boldsymbol{\beta}_n^*)\}$ and $\lambda_{\max}\{I_n(\boldsymbol{\beta}_n^*)\}$ are the smallest and largest eigenvalues of the Fisher information matrix $I_n(\boldsymbol{\beta}_n^*)$.

(C) There is a large enough open subset ω_n of $\Omega_n \in \mathbb{R}^p$ which contains the true parameter point $\boldsymbol{\beta}_n^*$, such that for almost all $\boldsymbol{D}$ the density admits all third derivatives $\partial^3 f(\boldsymbol{D}, \boldsymbol{\beta}_n)/\partial \beta_{nj}\partial \beta_{nk}\partial \beta_{nl}$ for all $\boldsymbol{\beta}_n \in \omega_n$. Furthermore, there are functions $M_{njkl}(\boldsymbol{D})$ and $C_5 > 0$ such that $|\partial^3 f(\boldsymbol{D}, \boldsymbol{\beta}_n)/\partial \beta_{nj}\partial \beta_{nk}\partial \beta_{nl}| \leq M_{njkl}(\boldsymbol{D})$ for all $\boldsymbol{\beta}_n \in \omega_n$ and $E_{\boldsymbol{\beta}_n} M_{njkl}(\boldsymbol{D})^2 < C_5 < \infty$ for all p, n and j, k, l.

(D) $\lambda_n/\rho_n^2 \to 0$, where $\rho_n = \min_{j:\beta_{nj}^* \neq 0} |\beta_{nj}^*|$.

These regularity conditions are part of conditions in Fan and Peng (2004). Basically, conditions (B) and (C) impose on the second and fourth moments of the likelihood function. The information

matrix of the likelihood function is assumed to be positive definite, and its eigenvalues are uniformly bounded. Condition (D) is necessary for obtaining the oracle property, which is implicitly assumed when p is fixed and $\lambda_n \to 0$.

The following theorem states our main results.

Theorem 1 *Let $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n$ be independent and identically distributed, each with a density $f(\boldsymbol{D}, \boldsymbol{\beta})$ that satisfies regularity conditions (A)-(D). When $n, p \to \infty$ with $p^5/n \to 0$, $\tau_n = O(\sqrt{\frac{1}{pn}})$ and $\lambda_n \tau_n (\frac{n}{p})^{\frac{3}{2}} \to \infty$, there exists a sequence of local minima of SELO-GLM, $\hat{\boldsymbol{\beta}}_n$, such that*

(i) [Estimation consistency]

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\|_2 = O_p(\sqrt{p/n})$$

(ii) [Model selection consistency]

$$\lim_{n\to\infty} P\Big(\{j : \hat{\beta}_{nj} \neq 0\} = A_n\Big) = 1$$

(iii) [Asymptotic normality and efficiency]

$$\sqrt{n}\boldsymbol{B}_n \boldsymbol{I}_{n,A_n}(\hat{\boldsymbol{\beta}}_{A_n} - \boldsymbol{\beta}_{A_n}^*) \to N(0, G),$$

where $\boldsymbol{B}_n$ is a $q \times |A_n|$ matrix such that $\boldsymbol{B}_n \boldsymbol{B}_n^T \to \boldsymbol{G}$, $\boldsymbol{G}$ is a $q \times q$ nonnegative symmetric matrix, and $I_{n,A_n} = I_{n,A_n}(\boldsymbol{\beta}_{n,A_n}^*)$ is the Fisher information matrix given $\boldsymbol{\beta}_{n,A_n^c}^* = \mathbf{0}$.

The proofs are provided in the Appendix. The first part of the theorem shows the existence of a $\sqrt{n/p}$-consistent penalized likelihood estimator. The second part of the theorem shows that SELO-GLM can consistently remove the null variables with a probability tending to 1. The third part of the theorem shows that the estimates for the non-zero coefficients in the true model have the same asymptotic distribution as they would have if the non-zero coefficients were known in advance. Therefore, overall, we can say that asymptotically, SELO-GLM performs as well as that if the true underlying model were given in advance, i.e., it has the oracle property. In addition, the rates of λ_n and τ_n can be satisfied when taking $\tau_n = \sqrt{1/(pn)}$ and $\lambda_n = \sqrt{p/n}$. The condition $n, p \to \infty$ with $p^5/n \to 0$ is the same as the condition in Fan and Peng (2004), which is stronger than the condition $p/n \to 0$ for SELO-linear model in Dicker et al. (2012). This is partially because generalized linear models are more complicated than linear models. If we apply our proof to linear models, we can get the same rate as in Dicker et al. (2012).

3.3 Tuning parameter selection for SELO-GLM

In practice, the actual values of the tuning parameters λ_n and τ_n need to be estimated in order to implement the SELO-GLM procedure. Tuning parameter selection is an important issue in penalized likelihood procedures. One often proceeds by finding estimators which correspond to a range of tuning parameter values. The preferred estimator is then identified as the one for the tuning parameter value to optimize some criteria, such as GCV (Golub et al. 1979; Tibshirani 1996; Fan and Li 2001), AIC (Zou et al. 2007), or BIC (Zou et al. 2007; Wang et al. 2007; Zhang et al. 2010; Chen and Chen 2012). It is well known that GCV and AIC-based methods are not consistent for model selection in the sense that, as $n \to \infty$, they may select irrelevant predictors with non-vanishing probability (Shao 1993; Wang et al. 2007; Zhang et al. 2010). On the other hand, BIC-based tuning parameter selection roughly corresponds to maximize the posterior probability of selecting the true model in an appropriate Bayesian formulation (Schwarz 1978) and has been shown to be consistent for model selection in several settings (Zou et al. 2007; Wang et al. 2007; Zhang et al. 2010; Chen and Chen 2012). Thus, if variable selection and identification of the true model is the primary goal, then BIC tuning parameter selection may be preferred over GCV and AIC.

BIC is defined as follows

$$BIC(\hat{\boldsymbol{\beta}}_n) = -2\ell_n(\hat{\boldsymbol{\beta}}_n) + \log n \times df, \tag{14}$$

where $\ell_n(\hat{\boldsymbol{\beta}}_n)$ is the log-likelihood of the data evaluated at $\hat{\boldsymbol{\beta}}_n$ and df is the degrees of freedom of the fitted model. We propose estimating df by the number of non-zero coefficients for SELO-GLM estimators. Under generalized linear models, Zhang et al. (2010) considered the BIC-type tuning parameter selector for the SCAD procedure. They proved that the BIC-type selector enables identification of the true model consistently, but their results require p to be fixed. Chen and Chen (2012) proposed an extended BIC which was shown to be consistent for variable selection under generalized linear models. Their results allow a diverging p, but requires the true model is fixed. In the following, we prove that when BIC is used, the SELO-GLM/BIC procedure is consistent for model selection with diverging p, which also allows the true model to be changed as n changes.

Before stating our main results, we introduce some notation. Let α_n be a subset of $\{X_1, \ldots, X_p\}$, $\alpha_n^* \in \{X_1, \ldots, X_p\}$ be the subset that contains all the signal variables affecting Y (the underlying true model), $m(\alpha_n)$ be the number of variables in α_n (model size), and $\boldsymbol{\beta}(\alpha_n)$ be a vector of the components in $\boldsymbol{\beta}$ that correspond to the features in α_n. Denote $A_{1n} = \{\alpha_n : \alpha_n^* \subset \alpha_n, \alpha_n^* \neq \alpha_n\}$ be the set of over-fitted models, i.e., any model in A_{1n} includes all of signal variables and at least one

null variables. Denote $A_{2n} = \{\alpha_n : \alpha_n^* \nsubseteq \alpha_n\}$ be the set of under-fitted models, i.e., any model in A_{2n} does not include at least one signal variables. We also denote $\alpha_{\lambda_n,\tau_n}$ be the model identified by SELO-GLM when the tuning parameters are (λ_n, τ_n).

To prove the model selection consistency of SELO-GLM/BIC procedure, besides regularity conditions (A)-(D), we need the following conditions as well. Let $\boldsymbol{H}_n(\boldsymbol{\beta})$ be the negative second derivative of the log-likelihood function of data. Under generalized linear models with canonical links, it can be written as

$$\boldsymbol{H}_n(\boldsymbol{\beta}) = -\frac{\partial^2 \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} W_i \boldsymbol{X}_i \boldsymbol{X}_i^T. \tag{15}$$

where W_i is the GLM working weight.

(E) There exist positive constants C_1, C_2 such that for all sufficiently large n,

$$C_1 < \lambda_{\min}\Big\{\frac{1}{n} H_n(\boldsymbol{\beta}_n^*)\Big\} < \lambda_{\max}\Big\{\frac{1}{n} H_n(\boldsymbol{\beta}_n^*)\Big\} < C_2,$$

where $\lambda_{\min}\Big\{\frac{1}{n} H_n(\boldsymbol{\beta}_n^*)\Big\}$ and $\lambda_{\max}\Big\{\frac{1}{n} H_n(\boldsymbol{\beta}_n^*)\Big\}$ are the smallest and largest eigenvalues of $\frac{1}{n} H_n(\boldsymbol{\beta}_n^*)$.

(F) For any given $\epsilon > 0$, there exists a constant $\delta > 0$ such that, when n is sufficiently large,

$$(1-\epsilon) H_n\{\boldsymbol{\beta}_n^*(\alpha_n)\} \leq H_n\{\boldsymbol{\beta}(\alpha_n)\} \leq (1+\epsilon) H_n\{\boldsymbol{\beta}_n^*(\alpha_n)\}$$

for all α_n and $\boldsymbol{\beta}(\alpha_n) \in A_{1n} \bigcup \{\alpha_n^*\}$ satisfies $\|\boldsymbol{\beta}(\alpha_n) - \boldsymbol{\beta}_n^*(\alpha_n)\| \leq \delta$.

(G) Denote by X_{ij} the jth component of $\boldsymbol{X}_i$.

$$\max_{1\leq j\leq d} \max_{1\leq i\leq n} \{\frac{X_{ij}^2}{\sum_{i=1}^n X_{ij}^2 W_i}\} = o((\log n)^{-1}).$$

(H) $\rho_n > C n^{-\frac{1}{5}}$ for some constant C, where $\rho_n = \min_{j\in\alpha_n^*} |\beta_{nj}^*|$.

Conditions (E)-(H) are part of conditions assumed in Chen and Chen (2012). Condition (E) is similar to the UUP condition in Candes and Tao (2007). Condition (F) extends (E) to a small neighborhood of $\boldsymbol{\beta}_n^*$. These two require the true model to stay at some distance from wrong models as n increases. Condition (G) can be violated only if the square of a feature has a severely skewed distribution, however, such variables would have readily been screened out before a variable selection procedure is applied. Condition (H) is similar to condition (D), which is necessary for the proof and is implicitly assumed when p is fixed.

The following theorem states our main results.

Theorem 2 *Let $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_n$ be independent and identically distributed, each with a density $f(\boldsymbol{D}, \boldsymbol{\beta})$ that satisfies regularity conditions (A)-(H). When $n, p \to \infty$ with $p^5/n \to 0$, $\tau_n = O(\sqrt{\frac{1}{pn}})$, $\lambda_n \tau_n (\frac{n}{p})^{\frac{3}{2}} \to \infty$ and $\frac{\lambda_n p n^{\frac{2}{5}}}{\log n} \to 0$ Then*

(a) [Consistency of important variable selection]

$$P\left\{\inf_{\alpha_{\lambda,\tau} \in A_{2n}} BIC(\lambda, \tau) > BIC(\lambda_n, \tau_n)\right\} \to 1$$

(b) [Consistency of unimportant variable removal] If we further assume $m(\alpha_n^*) = o(\log n)$ and $\log p = o(\log n)$, then

$$P\left\{\inf_{\alpha_{\lambda,\tau} \in A_{1n}} BIC(\lambda, \tau) > BIC(\lambda_n, \tau_n)\right\} \to 1$$

The proofs are provided in Appendix. The first part of the theorem shows that the SELO-GLM/BIC procedure consistently selects all signal variables. The second part of theorem shows that, when the size of the true model is not too large (for example, if the true model is fixed), the SELO-GLM/BIC procedures also consistently remove all null variables. The condition $\log p = o(\log n)$ is comparable to the condition $p^5 = o(\log n)$ which is assumed to prove the "oracle" property of SELO-GLM. The $m(\alpha_n^*) = o(\log n)$ constraint on the size of the true model is restrictive, but to our knowledge, our results are already stronger than all the existing results, which either assumes the dimension p is fixed or the true model size $m(\alpha_n^*)$ is fixed. In addition, the required rates of τ_n and λ_n in Theorem 2 can be satisfied when taking $\tau_n = \sqrt{1/(pn)}$ and $\lambda_n = \log n / n^{3/5}$.

4 Simulations

In this section, we perform simulation studies to evaluate the finite sample performance of the SELO-GLM method, and compare the results with several existing methods, including LASSO, Adaptive LASSO (ALASSO), SCAD and MCP. We considered two examples based on logistic regression. The dimension is moderate in the first example and high in the second example. The details of the settings are described as follows.

Example 1: There are $p = 8$ predictors. Each predictor X_j is generated from a standard normal distribution and the correlation between predictors are $Cov(X_i, X_j) = 0.5^{|i-j|}$. The true model is

$$logit(Pr(Y_i = 1 | \boldsymbol{X}_i)) = -0.3 + 0.5X_{i1} + 0.5X_{i5},\ i = 1, \ldots, n.$$

The Bayes error is about 0.356.

Example 2: There are $p = 20$ predictors. The setup is almost the same as in Example 1. We add 12 null variables $(X_9, \ldots, X_{20})$ to make the problem more difficult. Each predictor X_j is generated from

a standard normal distribution and the correlation between predictors are $Cov(X_i, X_j) = 0.5^{|i-j|}$. The true model is the same as in example 1.

For each setup, we considered two sample sizes: $n = 200$ and $n = 300$. We repeated the simulation for $1,000$ times. The SELO-GLM was fitted using the algorithm described in Section 2.3. LASSO and ALASSO were fitted using the R package "glmnet". SCAD and MCP were fitted using the R package "ncvreg".

Regarding to tuning parameter selection, for SELO-GLM, following Dicker et al. (2012), we fixed $\tau = 0.01$. For SCAD, following Fan and Li (2001), we fixed $a = 3.7$. For MCP, Zhang (2010) suggested using $\gamma = 2/(1 - \max_{i \neq j} |\boldsymbol{X}_i^T \boldsymbol{X}_j|/n)$, which was about 4 in our simulations. Zhang et al. (2010) also observed that the performance of MCP tends to be better when γ becomes smaller. We tried several values of γ which were less than 4. When $\gamma < 2$, the R package "ncvreg" reported that the algorithm failed to converge quite often. When γ took values between 2 and 4, we observed that the performances of MCP with $\gamma = 2$ were uniformly better than the performance of MCP with other γ values. Therefore, we only reported the performance of MCP with $\gamma = 2$.

To select the remaining tuning parameter in each of the four methods, we considered two approaches. The first approach was based on BIC (14). The model with the smallest BIC value was selected. In BIC tuning parameter selection, the *df* (degrees of freedom) needs to be specified. For SELO-GLM, by Theorem 2, we used the number of nonzero coefficients (including intercept). For LASSO, following Zou et al. (2007), we also used the number of number coefficients. For SCAD, following Fan and Li (2001); Zhang et al. (2010), we estimated the degrees of freedom be the trace of the approximate linear projection matrix:

$$\hat{df} = tr\left\{ \Big(\frac{\partial^2 Q_n(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big)^{-1} \frac{\partial^2 \ell_n(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\}, \tag{16}$$

where $\ell_n(\boldsymbol{\beta})$ is the log-likelihood function and $Q_n = -\frac{1}{n}\ell_n + p_{SCAD}(\boldsymbol{\beta})$. To our knowledge, BIC tuning parameter selection for MCP has not been discussed in the literature. We used the number of nonzero coefficients to estimate its degrees of freedom.

Our second approach for tuning parameter selection was based on data validation (DV). To be specific, in each simulation, besides the training data, we also independently generated a set of validation data with the same distribution and in a same sample size as training data. Then for each model, we calculated its misclassification error on the validation set. The model with the smallest validation error was selected.

To evaluate the variable selection performance of methods, we considered the false positive rate

(FPR) and false negative rate (FNR), which are defined as follows.

$$\text{FPR} = \frac{\#\text{ of selected important variables}}{\#\text{ of selected variables}}, \quad \text{FNR} = \frac{\#\text{ of removed unimportant variables}}{\#\text{ of removed variables}}$$

Our calculations of FPR and FNR did not include the intercept term. Also, when no variable was selected, FPR was set to be 0, and when all variables were selected, FNR was set to be 0. We also calculated the model size, which is the number of variables in the model (the intercept term is not included, since it is always selected), and the percentage that the correct model is identified within $1,000$ repetitions. To evaluate the prediction performance of methods, in each simulation, we generated an independent test set with sample size $n = 10,000$ to calculate the misclassification rates of each method. We also calculated the square error for estimated coefficients, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$.

Table 1 summarizes the results for Example 1. We can see that, both of the SELO-GLM and MCP selected a smaller model and the correct model more frequently than other methods, regardless of the tuning methods using BIC or data validation. SELO-GLM tended to select a smaller model than MCP. When $n = 200$ and BIC was used to select the tuning parameter, SELO-GLM and MCP had a same percentage of identifying the correct model, but SELO-GLM had slightly higher percentages than MCP in other settings. All of methods had comparable performance in terms of misclassification rates and square errors. When data validation (DV) was used as the tuning procedure, we can see that all of methods had slightly better prediction performances (lower misclassification rates) but much worse variable selection performances than they had when BIC was used as the tuning procedure. It is not surprising that the DV approach yielded better prediction performance than the BIC approach, because the DV approach is specifically designed to minimize the misclassification rate. The worse variable selection performance of the DV approach is partially because the best predictive model tends to include some null variables especially when the signal in the data is weak and the sample size is relatively small (Hastie et al. 2009). In addition, when the sample size increased from $n = 200$ to $n = 300$, all of methods had better performances in all metrics as expected.

Table 2 summarizes the results for Example 2. When the number of parameter was high, the SELO-GLM has a clear advantage in model selection over all other methods (including MCP). To be specific, SELO-GLM had significantly higher percentage of identifying the correct model than all other methods, regardless of the choice of the tuning methods. This advantage was even bigger when the sample size increased from $n = 200$ to $n = 300$. For other metrics, the SELO-GLM had comparable performances with all other methods. Also, similar to Example 1, we observe that when data validation (DV) was used as the tuning procedure, all methods had a slightly better prediction

performance but a worse variable selection performance than they had when BIC was used as the tuning procedure. When the sample size increased from $n = 200$ to $n = 300$, all of methods had a better performances in all metrics as expected.

We also conducted additional simulations with the compound symmetry correlation structure and additional simulations with larger p, e.g., $p = 50$. The simulation results were similar to what we reported here. SELO-GLM has better performances than other methods in model selection. We omit the detailed results due to the limit of the space.

In summary, the simulation studies illustrate that the SELO-GLM has advantage in identifying correct models over other methods, especially when the dimension of the parameter space is high and the signal is weak. In our simulation studies, we fixed $a = 3.7$ for SCAD, $\gamma = 2$ for MCP and $\tau = 0.01$ for SELO-GLM. If these parameters are also tuned carefully, all of three methods may have better performance; nevertheless, the simulations succeed in demonstrating the differences between the methods in terms of correct model identification. We also would like to point out that, based on our numerical studies, the performance of SELO-GLM was quite robust to the choice of τ, while the performance of MCP was sensitive to the selection of γ.

5 Real data analysis

Hunter et al. (2007) reported a study to assess the association between the FGFR2 gene and the risk of sporadic postmenopausal breast cancer using the data from a genome-wide association study (GWAS) of the Nurses' Health Study (NHS) cohort, which was part of the National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMS) Study. Using the Illumina HumanHap500 array, this study initially genotyped 1,183 women with postmenopausal invasive breast cancer and 1,185 individually matched controls. The original GWAS study, as reported in Hunter et al. (2007), identified several loci as potentially associated with breast cancer using individual SNP analysis and found several SNPs in FGFR2 to be highly associated with breast cancer. After removing subjects with missing values, we have 1,141 women with postmenopausal invasive breast cancer and 1,141 individually matched controls. For each woman, the 41 SNPs associated with the gene FGFR2 were measured on the Illumina array.

In this section, we analyzed the 41 SNP simultaneously using logistic regression to jointly select a subset of SNPs that are associated with breast cancer risk, i.e., the response is the postmenopausal invasive breast cancer status (yes or no), and the predictors are 41 SNP measures. We applied our proposed SELO-GLM and compare its performance to LASSO, adaptive LASSO, SCAD and MCP.

For all methods, the tuning parameters were selected using BIC. The adaptive LASSO failed to identify any SNP. This may be because the signals were weak, and the weights based on unpenalized logistic regression did not well reflect the importance of SNPs, and hence downgraded the performance of adaptive LASSO. Except adaptive LASSO, all other methods identified one SNP: rs1219648, which has been previously shown to be highly associated with increased risk of breast cancer (Hunter et al. 2007). We also observed that there were two SNPs rs2420946 and rs2981579 that were highly correlated with SNP rs1219648 (correlation coefficients were 0.91 and 0.93, respectively). We removed rs2420946 and rs2981579 and redid the analysis. The adaptive LASSO again failed to identify any SNP, and the SELO-GLM as well as other methods identified the SNP rs1219648 only.

6 Discussions

In this paper, we extend the SELO procedure from linear models to generalized linear models for variable selection. We show that, the SELO-GLM procedure gives regression coefficient parameters that have the "oracle" property when the tuning parameters are properly selected. We also show that the BIC in combination with SELO-GLM procedure consistently identifies the correct model. The simulation studies demonstrate that the SELO-GLM procedure has a better performance in identifying correct models than several existing methods, especially when the number of parameters is high and signals weak. We applied the SELO-GLM method to a breast cancer genetic study and obtained clinically meaningful results that were independently validated.

In this paper, our numerical study focuses on evaluating the performance of the SELO method using logistic regression. Small scale simulation studies (not reported here) show that the proposed SELO-GLM method performs well for poisson regressions and multinomial logistic regressions. The SELO method can also be applied to the Cox proportional hazard regression for censored data as well. We expect similar theoretical results as presented in this paper to hold under the Cox model.

We allow in this paper both the number of regression parameters p and the sample size n diverge, but p goes to infinity more slowly than n. Future research is needed to extend the results to situations when p is much larger than n.

Acknowledgements

The research of Li was partially supported by NCFC (NO.11071137 and NO.10731010) and by the National Natural Science Foundation of China (No. 11071137). The research of Lin was partially supported by R37-CA076404 and P01-CA134294.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22:555–574.

Dicker, L., Huang, B., and Lin, X. (2012). Variable selection and estimation with the seamless-l0 penalty. *Statistica Sinica*. To appear.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2009). *The elements of statistical learning*. Springer, second edition.

Hunter, D., Kraft, P., Jacobs, K., Cox, D., Yeager, M., Hankinson, S., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., et al. (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, 39(7):870–874.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378.

Mallows, C. (1973). Some comments on cp. *Technometrics*, 15(4):661–675.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC, second edition.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, Y., Li, R., and Tsai, C. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733.

Appendix

Due to space limitation, we only present a sketch of proofs here. The detailed proofs can be found in the attached Supplemental Materials.

A.1. Proof of Theorem 1

First we prove estimation consistency of the SELO-GLM, namely (i) in Theorem 1.

Proof Let $\delta_n = \sqrt{\frac{p}{n}}$, fix $\epsilon > 0$. To prove (i), it suffices to show that, it exists $C > 0$ s.t.

$$P\Big(Q_n(\boldsymbol{\beta}_n^*) < \inf_{||\boldsymbol{u}||=1} Q_n(\boldsymbol{\beta}_n^* + C\delta_n \boldsymbol{u})\Big) \geq 1 - \epsilon$$

holds for all n sufficiently large. Define

$$\begin{aligned} & D_n(u) = Q_n(\boldsymbol{\beta}_n^* + C\delta_n \boldsymbol{u}) - Q_n(\boldsymbol{\beta}_n^*) \\ = & \Big(-\frac{1}{n}\ell_n(\boldsymbol{\beta}_n^* + C\delta_n \boldsymbol{u})\Big) + \frac{1}{n}\ell_n(\boldsymbol{\beta}_n^*)) + \sum_{j=1}^{p} p_{SELO}(\beta_{nj}^* + C\delta_n \boldsymbol{u}) - p_{SELO}(\beta_{nj}^*) \triangleq R_1 + R_2 \end{aligned}$$

By a Taylor expansion, we have

$$R_1 = -\frac{1}{n}\Big(\nabla \ell_n(\boldsymbol{\beta}_n^*)^T \boldsymbol{u} C\delta_n + \frac{1}{2}(\boldsymbol{u}^T \nabla^2 \ell_n(\boldsymbol{\beta}_n^*)\boldsymbol{u})C^2\delta_n^2 + \frac{1}{6}\nabla^T\{\boldsymbol{u}^T \nabla^2 \ell_n(\tilde{\boldsymbol{\beta}}_n)\boldsymbol{u}\}C^3\delta_n^3\Big) \triangleq I_1 + I_2 + I_3$$

where $\tilde{\boldsymbol{\beta}}_n \in [\boldsymbol{\beta}_n^*, \boldsymbol{\beta}_n^* + C\delta_n \boldsymbol{u}]$. As Fan and Peng (2004) have shown in their paper, we can get

$$|I_1| = CO_p(\delta_n^2), \qquad I_2 \geq \frac{1}{2}C^2\delta_n^2 C_1 - \frac{1}{2}C^2\delta_n^2 o_p(1), \qquad I_3 = C^2\delta_n^2 o_p(1). \tag{A. 1}$$

$$||\frac{1}{n}\nabla^2 l_n(\boldsymbol{\beta}_n^*) + \boldsymbol{I}_n(\boldsymbol{\beta}_n^*)|| = o_p(\frac{1}{p}) \tag{A. 2}$$

For R_2, we have

$$R_2 = \sum_{j=1}^{p}[p_{SELO}(\beta_{nj}^* + C\delta_n u_j) - p_{SELO}(\beta_{nj}^*)] \geq \sum_{j\in K(u)}[p_{SELO}(\beta_{nj}^* + C\delta_n u_j) - p_{SELO}(\beta_{nj}^*)]$$

where $K(u) = \{j | p_{SELO}(\beta_{nj}^* + C\delta_n u_j) - p_{SELO}(\beta_{nj}^*) < 0\}$. Condition (D) and the fact that p_{SELO} is concave on $[0, \infty]$ implies that, for each C,

$$p_{SELO}(\beta_{nj}^* + C\delta_n u_j) - p_{SELO}(\beta_{nj}^*) \geq -C\delta_n |u_j| p'_{SELO}(\beta_{nj}^* + C\delta_n u_j)$$

when n is sufficiently large. Thus, for n big enough, we have

$$\sum_{j\in K(u)} C\delta_n |u_j| p'_{SELO}(\beta_{nj}^* + C\delta_n u_j) \leq \frac{C\delta_n \lambda_n \tau_n}{\rho_n^2 \log 2}\sqrt{p} = C^2\delta_n^2 \frac{1}{C\sqrt{p}\log 2}\frac{\lambda_n}{\rho_n^2}(\tau_n\sqrt{pn}) = o(C^2\delta_n^2) \tag{A. 3}$$

By (A. 1)-(A. 3), allowing C to be large enough, all terms I_1, I_3, R_2 are dominated by I_2 which is negative. This proves (i) in Theorem 1.

To prove (ii), we prove a lemma first.

Lemma 1 *Assume conditions (A)-(D) hold and let $\hat{\boldsymbol{\beta}}_n$ is a sequence of local minima of SELO which satisfied $||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*|| = O_p(\sqrt{\frac{p}{n}})$. Then,*

$$P(\hat{\boldsymbol{\beta}}_{n,A_n^c} \neq 0) \to 0$$

Proof It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any constant $C > 0$, we have

$$\operatorname{sgn}(\beta_{nj}) \frac{\partial}{\partial \beta_{nj}} Q_n(\boldsymbol{\beta}_n) > 0$$

holds for all $\beta_{nj} \neq 0$, $j \in A_n^c$ and all $\boldsymbol{\beta}_n \in \mathbb{R}^{p+1}$ such that $||\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^*|| < C\sqrt{\frac{p}{n}}$.

Let $\boldsymbol{u} = \boldsymbol{\beta}_n - \boldsymbol{\beta}_n^*$, then we have

$$\begin{aligned}
&\frac{\partial}{\partial \beta_{nj}} Q_n(\boldsymbol{\beta}_n) = -\frac{1}{n}\frac{\partial \ell_n(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} + \frac{\lambda_n \tau_n \operatorname{sgn}(\beta_{nj})}{\log 2(2|\beta_{nj}| + \tau_n)(|\beta_{nj}| + \tau_n)} \\
= \; & -\frac{1}{n}\frac{\partial \ell_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}} - \frac{1}{n}\sum_{l=0}^{p} \frac{\partial^2 \ell_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}\partial \beta_{nl}} u_l - \frac{1}{n}\sum_{l,k=0}^{p} \frac{\partial^3 \ell_n(\tilde{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}\partial \beta_{nl}\partial \beta_{nk}} u_l u_k + \frac{\lambda_n \tau_n \operatorname{sgn}(\beta_{nj})}{\log 2(2|\beta_{nj}| + \tau_n)(|\beta_{nj}| + \tau_n)} \qquad \text{(A. 4)} \\
\triangleq \; & I_1 + I_2 + I_3 + I_4, \qquad \text{(A. 5)}
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_n \in [\boldsymbol{\beta}_n, \boldsymbol{\beta}_n^*]$. As Fan and Peng (2004) have shown in their paper, we can get

$$I_1 = O_p(\sqrt{\frac{p}{n}}), \qquad I_2 = O_p(\sqrt{\frac{p}{n}}), \qquad I_3 = O_p(\sqrt{\frac{p}{n}})$$

Finally we consider I_4

$$\begin{aligned}
\operatorname{sgn}(\beta_{nj}) I_4 / \sqrt{\frac{p}{n}} &= \frac{\lambda_n \tau_n}{\log 2\sqrt{\frac{p}{n}}(2|\beta_{nj}| + \tau_n)(|\beta_{nj}| + \tau_n)} = \frac{\lambda_n \tau_n (\frac{n}{p})^{\frac{3}{2}}}{\log 2(\frac{2|\beta_{nj}| + \tau_n}{\sqrt{\frac{p}{n}}} +)(\frac{|\beta_{nj}| + \tau_n}{\sqrt{\frac{p}{n}}})} \\
&\geq \frac{\lambda_n \tau_n (\frac{n}{p})^{\frac{3}{2}}}{\log 2(2C + \tau_n)(C + \tau_n)} \to \infty
\end{aligned}$$

It follows that the sign of β_{nj} completely determines the sign of $\frac{\partial}{\partial \beta_{nj}} Q_n(\boldsymbol{\beta}_n)$. Then we complete the proof of lemma 1.

Let $\hat{A}_n = \{j \geq 1 | \hat{\beta}_{nj} \neq 0\}$. On one hand, by lemma 1, we can get $P(A_n \subseteq \hat{A}_n) \to 1$. On the other hand, for $\frac{\rho_n}{\delta_n} \to \infty$, where $\delta_n = \sqrt{\frac{p}{n}}$, using (i) we have $P(\hat{A}_n \subseteq A_n) \to 1$. So we get (ii) in Theorem 1, namely,

$$P(\hat{A}_n = A_n) \to 1$$

Finally, we proof (iii) in Theorem 1,

Proof Denote $A_n^0 = \{0\} \bigcup A_n$. By Theorem 1(ii), we can let $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_{n,A_n^0}, 0)$, then we get

$$0 = \frac{\partial}{\partial \beta_{nj}} Q_n(\hat{\boldsymbol{\beta}}_n) = -\frac{1}{n}\frac{\partial \ell_n(\hat{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}} + \frac{\lambda_n \tau_n \operatorname{sgn}(\hat{\beta}_{nj})}{\log 2(2|\hat{\beta}_{nj}| + \tau_n)(|\hat{\beta}_{nj}| + \tau_n)}$$

for any $j = 1, \cdots, |A_n|$. By a Taylor expansion, we have

$$\begin{aligned} 0 &= -\frac{1}{n}[\frac{\partial \ell_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}} + \sum_{l=0}^{|A_n|} \frac{\partial^2 \ell_n(\boldsymbol{\beta}_n^*)}{\partial \beta_{nj}\partial \beta_{nl}}(\hat{\beta_{nl}} - \beta_{nl}^*) + \sum_{l,k=0}^{|A_n|} \frac{\partial^3 \ell_n(\tilde{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}\partial \beta_{nl}\partial \beta_{nk}}(\hat{\beta}_{nl} - \beta_{nk}^*)(\hat{\beta}_{nk} - \beta_{nk}^*)] \\ &\quad + \frac{\lambda_n \tau_n \operatorname{sgn}(\hat{\beta}_{nj})}{\log 2(2|\hat{\beta}_{nj}| + \tau_n)(|\hat{\beta}_{nj}| + \tau_n)} \end{aligned}$$

Let $\nabla_{A_n^0} = (\frac{\partial}{\partial \beta_{n0}}, \cdots, \frac{\partial}{\partial \beta_{n|A_n|}})$, then we get,

$$\begin{aligned} \sqrt{n}\boldsymbol{B}_n \boldsymbol{I}^{\frac{1}{2}}_{n,A_n^0}(\hat{\boldsymbol{\beta}}_{n,A_n^0} - \boldsymbol{\beta}^*_{n,A_n^0}) &= \frac{1}{\sqrt{n}}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0} \nabla l_n(\boldsymbol{\beta}^*_{n,A_n^0}) \\ &\quad + \frac{1}{\sqrt{n}}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0}(\nabla^2 l_n(\boldsymbol{\beta}^*_{n,A_n^0}) + n\boldsymbol{I}_{n,A_n^0})(\hat{\boldsymbol{\beta}}_{n,A_n^0} - \boldsymbol{\beta}^*_{n,A_n^0}) \\ &\quad + \frac{1}{\sqrt{n}}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0}(b_0, \cdots, b_{|A_n|})^T - \sqrt{n}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0}(0, p'(\hat{\beta_{n1}}), \cdots, p'(\hat{\beta}_{n,|A_n|}))^T \end{aligned} \quad \text{(A. 6)}$$

where $b_i = \sum_{l,k=0}^{|A_n|} \frac{\partial^3 \ell(\tilde{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}\partial \beta_{nl}\partial \beta_{nk}}(\hat{\beta}_{nl} - \beta_{nk}^*)(\hat{\beta}_{nk} - \beta_{nk}^*)$. Following Fan and Peng (2004), we can get

$$\frac{1}{\sqrt{n}}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0} \nabla \ell_n(\boldsymbol{\beta}^*_{n,A_n^0}) \longrightarrow N(0, G) \quad \text{(A. 7)}$$

Let $\boldsymbol{Z} = \frac{1}{\sqrt{n}}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0}(b_0, \cdots, b_{|A_n|})^T$. By condition (C), we have

$$\begin{aligned} |b_i| &= \sum_{l,k=0}^{|A_n|} |\frac{\partial^3 \ell_n(\tilde{\boldsymbol{\beta}}_n)}{\partial \beta_{nj}\partial \beta_{nl}\partial \beta_{nk}}(\hat{\beta}_{nl} - \beta_{nk}^*)(\hat{\beta}_{nk} - \beta_{nk}^*|) \le n \sum_{l,k=0}^{|A_n|} M_{njkl}(\boldsymbol{V})|\hat{\beta}_{nl} - \beta_{nk}^*||\hat{\beta}_{nk} - \beta_{nk}^*)| \\ &\le n(\sum_{l,k=0}^{|A_n|} M_{njkl}^2(\boldsymbol{D}))^{\frac{1}{2}}||u||^2 \le O_p(np\frac{p}{n}) = O_p(p^2) \end{aligned}$$

Then we get,

$$Z = O_p(\frac{1}{\sqrt{n}}p^2\sqrt{p}) = o_p(1) \quad \text{(A. 8)}$$

Let $A_1 = \frac{1}{\sqrt{n}}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0}(\nabla^2 l_n(\boldsymbol{\beta}^*_{n,A_n^0}) + n\boldsymbol{I}_{n,A_n^0})(\hat{\boldsymbol{\beta}}_{n,A_n^0} - \boldsymbol{\beta}^*_{n,A_n^0})$, then by (2) we get

$$A_1 = \frac{1}{\sqrt{n}}O_p(\sqrt{\frac{p}{n}})o_p(\frac{n}{p}) = o_p(\frac{1}{\sqrt{n}}\sqrt{\frac{p}{n}}\frac{n}{p}) = o_p(1) \quad \text{(A. 9)}$$

Let $A_2 = \sqrt{n}\boldsymbol{B}_n \boldsymbol{I}^{-\frac{1}{2}}_{n,A_n^0}(0, p'(\hat{\beta}_{n1}), \cdots, p'(\hat{\beta}_{n,|A_n|}))^T$, then because of

$$\begin{aligned} EA_2^T A_2 &\le n\frac{\lambda_{\max}(\boldsymbol{B}_n\boldsymbol{B}_n^T)}{C_1}E\sum_{j=1}^{|A_n|} p'(\hat{\beta}_{nj})^2 \le np\frac{\lambda_{\max}(\boldsymbol{B}_n\boldsymbol{B}_n^T)}{C_1}\frac{\lambda_n^2\tau_n^2}{\rho_n^4(\log 2)^2} \\ &= (\tau_n\sqrt{np})^2(\frac{\lambda_n}{\rho_n^2})^2\frac{\lambda_{\max}(\boldsymbol{B}_n\boldsymbol{B}_n^T)}{C_1(\log 2)^2} \to 0 \end{aligned}$$

So we finally get,

$$A_2 = o_p(1) \tag{A. 10}$$

Then for (A. 7)-(A. 10), we proof (iii) in Theorem 1.

A.2. Proof of Theorem 2

Lemma 2 *Let Y_i, $i = 1, 2, \cdots n$, be independent random variables following exponential family distributions of form $f(Y;\theta) = \exp\{\theta Y - b(\theta)\}$ with natural parameters θ_i. Let μ_i and σ_i^2 denote the mean and variance of Y_i respectively. Suppose that $\{\theta_i : i = 1, 2, \cdots, n\}$ is contained in a compact subset of the natural parameter space Θ. Let a_{ni}, $i = 1, 2, \cdots, n$ be real numbers such that $\sum_{i=1}^n a_{ni}^2\sigma_i^2 = 1$, $\sum_{i=1}^n a_{ni}^2 \leq M$. Then, for any m_n satisfying $\max_{1\leq i\leq n} a_{ni}^2 m_n = o(1)$ and positive ϵ, we have*

$$P(\sum_{i=1}^n a_{ni}(Y_i - \mu_i) > \sqrt{2m_n}) \leq \exp\{-m_n(1-\epsilon)\}$$

This lemma has shown in Chen and Chen (2012) , details are provided in the Supplementary Material. Denote $\alpha_n^0 = \alpha_n \bigcup \{0\}$. Let $\boldsymbol{\beta}(\alpha_n)$ be the vector of the components in $\boldsymbol{\beta}$ that correspond to the features in α_n^0 and $\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n)$ be the MLE when we use model α_n^0.

Lemma 3 *Under conditions $(A) - (I)$,*

$$\max_{\alpha_n \in A_{1n}\bigcup\{\alpha_n^*\}} ||\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n) - \boldsymbol{\beta}^*(\alpha_n)|| = O_p(n^{-\frac{1}{3}}).$$

Chen and Chen (2012) has a similar lemma, but our lemma and its proof are a little bit different from theirs.

Proof For any unit vector $\boldsymbol{u}$, let $\boldsymbol{\beta}(\alpha_n, \boldsymbol{u}) = \boldsymbol{\beta}^*(\alpha_n) + n^{-\frac{1}{3}}\boldsymbol{u}$. Clearly, there exists $N_1 > 0$, when $n \geq N_1$, this $\boldsymbol{\beta}(\alpha_n, \boldsymbol{u})$ falls into the neighborhood of $\boldsymbol{\beta}^*(\alpha_n)$ so that $\epsilon = \frac{1}{2}$ in condition (E) and (F) become applicable. Thus by a Taylor expansion, for all $\alpha_n \in A_{1n}\bigcup\{\alpha_n^*\}$,

$$\ell_n(\boldsymbol{\beta}(\alpha_n, \boldsymbol{u})) - \ell_n(\boldsymbol{\beta}^*(\alpha_n)) = n^{-\frac{1}{3}}\boldsymbol{u}^T\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) - \frac{1}{2}n^{\frac{1}{3}}\boldsymbol{u}^T\{\frac{1}{n}\boldsymbol{H}_n(\tilde{\boldsymbol{\beta}}(\alpha_n, \boldsymbol{u}))\}\boldsymbol{u} < n^{-\frac{1}{3}}\boldsymbol{u}^T\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) - \frac{1}{4}C_1 n^{\frac{1}{3}}$$

where $\tilde{\boldsymbol{\beta}}(\alpha_n, \boldsymbol{u}) \in [\boldsymbol{\beta}^*(\alpha_n), \boldsymbol{\beta}(\alpha_n, \boldsymbol{u})]$. Then we have

$$\begin{aligned}
& P(\bigcup_{\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}} \{\ell_n(\boldsymbol{\beta}(\alpha_n)) - \ell_n(\boldsymbol{\beta}^*(\alpha_n)) > 0;\ for\ some\ u\}) \\
\leq\ & P(\bigcup_{\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}} \{\boldsymbol{u}^T \boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) \geq \frac{1}{4} C_1 n^{\frac{2}{3}};\ for\ some\ u\}) \\
\leq\ & P(\bigcup_{\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}} \{s_n(\boldsymbol{\beta}^*(\alpha_n))^T s_n(\boldsymbol{\beta}^*(\alpha_n)) \geq \frac{1}{16} C_1^2 n^{\frac{4}{3}}\}) \\
\leq\ & P(\bigcup_{\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}} \bigcup_{j \in \alpha_n} \{s_{nj}(\boldsymbol{\beta}^*(\alpha_n))^2 \geq \frac{1}{16} C_1^2 n^{\frac{4}{3}} \frac{1}{p+1}\}) = P(\bigcup_{j=0}^{p} \{s_{nj}(\boldsymbol{\beta}_n^*)^2 \geq \frac{1}{16} C_1^2 n^{\frac{4}{3}} \frac{1}{p+1}\}) \\
\leq\ & \sum_{j=0}^{p} P(\{s_{nj}(\boldsymbol{\beta}_n^*) \geq \frac{1}{4} C_1 n^{\frac{2}{3}} (p+1)^{-\frac{1}{2}}\}) + \sum_{j=0}^{p} P(\{-s_{nj}(\boldsymbol{\beta}_n^*) \geq \frac{1}{4} C_1 n^{\frac{2}{3}} (p+1)^{-\frac{1}{2}}\})
\end{aligned}$$

The second inequality is form Cauchy-Schwarz inequality. Note that for all $\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}$

$$s_{nj}(\boldsymbol{\beta}^*(\alpha_n)) = \sum_{i=1}^{n} [Y_i - b^{'}(\boldsymbol{X}_i^T \boldsymbol{\beta}^*(\alpha_n))] X_{ij} = \sum_{i=1}^{n} (Y_i - \mu_i) X_{ij}$$

and the condition (E) implies $\max_{0 \leq j \leq p} \sum_{i=1}^{n} X_{ij}^2 \sigma_i^2 \leq n C_2$. Thus we have

$$[\sum_{i=1}^{n} X_{ij}^2 \sigma_i^2]^{\frac{1}{2}} 2\sqrt{\log n} = O((n \log n)^{\frac{1}{2}})$$

Let $a_{ni} = \frac{X_{ij}}{[\sum_{i=1}^{n} X_{ij}^2 \sigma_i^2]^{\frac{1}{2}}}$, for $|a_{ni}| 2\sqrt{\log n} \to 0$, by applying Lemma 2, there exists $N_2 > 0$, for any $n \geq N_2$, we have

$$\begin{aligned}
P(\{s_{nj}(\boldsymbol{\beta}_n^*) \geq \frac{1}{4} C_1 n^{\frac{2}{3}} (p+1)^{-\frac{1}{2}}\}) &\leq P(\{s_{nj}(\boldsymbol{\beta}_n^*) \geq [\sum_{i=1}^{n} X_{ij}^2 \sigma_i^2]^{\frac{1}{2}} (4 \log n)^{\frac{1}{2}}\}) \\
&\leq P(\sum_{i=1}^{n} a_{ni} (Y_i - \mu_i) > (4 \log n)^{\frac{1}{2}}) \leq \exp\{-2(1 - \frac{1}{2}) \log n\} = \frac{1}{n}
\end{aligned}$$

uniformly for all $\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}$. Let $N = \max\{N_1, N_2\}$, for any $n \geq N$ we have,

$$\sum_{j=0}^{p} P(\{s_{nj}(\boldsymbol{\beta}_n^*) \geq \frac{1}{4} C_1 n^{\frac{2}{3}} (p+1)^{-\frac{1}{2}}\}) \leq \frac{p+1}{n} \to 0$$

By replacing $Y_i - \mu_i$ with $-(Y_i - \mu_i)$ in the above argument, for any $n \geq N$, we also have

$$\sum_{j=0}^{p} P(\{-s_{nj}(\boldsymbol{\beta}_n^*) \geq \frac{1}{4} C_1 n^{\frac{2}{3}} (p+1)^{-\frac{1}{2}}\}) \leq \frac{p+1}{n} \to 0$$

Because $\ell_n(\boldsymbol{\beta}(\alpha_n))$ is a concave function for any α_n, the above result implies that with probability tending to 1 as $n \to \infty$, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n)$ exists and falls within a $n^{-\frac{1}{3}}$-neighborhood of $\boldsymbol{\beta}^*(\alpha_n)$ uniformly for $\alpha_n \in A_{1n} \bigcup \{\alpha_n^*\}$. The lemma is proved.

Now we consider the consistency of important variable selection, say (a) in Theorem 2.

Proof For any $\epsilon > 0$, Theorem 1(ii) implies that there exists $N_1 > 0$, for any $n \geq N_1$, we have $P(\alpha_{\lambda_n,\tau_n} = \alpha_n^*) \geq 1 - \frac{\epsilon}{4}$. Let $R_{1n} = \{\alpha_{\lambda_n,\tau_n} = \alpha_n^*\}$ and $\hat{\boldsymbol{\beta}}_{\text{mle}}$ be $\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n^*)$ augmented with zeros corresponding to the elements $\{1, \cdots, p\} \backslash \alpha_n^*$. Then on R_{1n}, by the definition of $\hat{\boldsymbol{\beta}}_{\lambda_n,\tau_n}$, we obtain

$$-\frac{2}{n} l_n(\hat{\boldsymbol{\beta}}_{\lambda_n,\tau_n}) - (-\frac{2}{n} l_n(\hat{\boldsymbol{\beta}}_{\text{mle}})) \leq \sum_{j \in \alpha_n^*} \Big(2p_{SELO,\lambda_n,\tau_n}(|\hat{\boldsymbol{\beta}}_{\text{mle},j}(\alpha_n^*)|)) - 2p_{SELO,\lambda_n,\tau_n}(|\hat{\boldsymbol{\beta}}_{\lambda_n,\tau_n,j}|) \Big) \quad \text{(A. 11)}$$

For $||\hat{\boldsymbol{\beta}}_{mle}(\alpha_n^*) - \boldsymbol{\beta}^*(\alpha_n^*)|| = O_p(\sqrt{\frac{p}{n}})$, $||\hat{\boldsymbol{\beta}}_{\lambda_n,\tau_n}(\alpha_n^*) - \boldsymbol{\beta}^*(\alpha_n^*)|| = O_p(\sqrt{\frac{p}{n}})$ and $\frac{\rho_n}{\sqrt{\frac{p}{n}}} \to \infty$, let $E_{1n} = \{\text{sgn}(\hat{\beta}_{\text{mle},j}(\alpha_n^*)) \,\text{sgn}(\hat{\beta}_{\lambda_n,\tau_n,j}(\alpha_n^*)) > 0 \; for \; all \; j \in \alpha_n^*\}, E_{2n} = \{||\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n^*) - \boldsymbol{\beta}^*(\alpha_n^*)|| \leq C\sqrt{\frac{p}{n}}\}, E_{3n} = \{||\hat{\boldsymbol{\beta}}_{\lambda_n,\tau_n}(\alpha_n^*) - \boldsymbol{\beta}^*(\alpha_n^*)|| \leq C\sqrt{\frac{p}{n}}\}$, there exists $N_2 > 0, C > 0$, for any $n \geq N_2$, we have $P(E_{1n} \bigcap E_{2n} \bigcap E_{3n}) \geq 1 - \frac{\epsilon}{4}$. Let $R_{2n} = E_{1n} \bigcap E_{2n} \bigcap E_{3n}$, then on $R_{1n} \bigcap R_{2n}$ we have,

$$\begin{aligned} (12) \quad &\leq \quad 2 \sum_{j \in \alpha_n^*} |p'_{SELO,\lambda_n,\tau_n}(\tilde{\beta}_{nj})| * |\hat{\beta}_{\text{mle},j}(\alpha_n^*) - \hat{\beta}_{\lambda_n,\tau_n,j}| \quad where \quad \tilde{\beta}_{nj} \in [\hat{\beta}_{\text{mle},j}(\alpha_n^*), \hat{\beta}_{\lambda_n,\tau_n,j}] \\ &\leq \quad 2 \sum_{j \in \alpha_n^*} \frac{\lambda_n \tau_n}{\log 2} \frac{1}{(|\tilde{\beta}_j| + \tau)(2|\tilde{\beta}_j| + \tau)} * 2C\sqrt{\frac{p}{n}} \leq \frac{4C}{\log 2} (\tau_n \sqrt{pn}) \frac{\lambda_n p n^{\frac{2}{5}}}{\log n} \frac{1}{\rho_n^2 n^{\frac{2}{5}}} \frac{\log n}{n} = o(\frac{\log n}{n}) \end{aligned}$$

Now on $R_{1n} \bigcap R_{2n}$, using the definition of mle we have,

$$\begin{aligned} & \inf_{\alpha_{\lambda,\tau} \in A_{2n}} BIC(\lambda, \tau) - BIC(\lambda_n, \tau_n) \\ \geq \quad & \inf_{\alpha_{\lambda,\tau} \in A_{2n}} -\frac{2}{n} \ell_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_{\lambda,\tau})) - [-\frac{2}{n} \ell_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n^*))] + \frac{[m(\alpha_{\lambda,\tau}) - m(\alpha_n^*)] \log n}{n} + o(\frac{\log n}{n}) \\ \geq \quad & \inf_{\alpha_n \in A_{2n}} -\frac{2}{n} \ell_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n)) - [-\frac{2}{n} \ell_n(\boldsymbol{\beta}^*(\alpha_n^*))] + \frac{[m(\alpha_{\lambda,\tau}) - m(\alpha_n^*)] \log n}{n} + o(\frac{\log n}{n}) \end{aligned}$$

For any $\alpha_n \in A_{2n}$, let $\tilde{\alpha}_n = \alpha_n \bigcup \alpha_n^*$. Now consider those $\boldsymbol{\beta}(\tilde{\alpha}_n)$ near $\boldsymbol{\beta}^*(\tilde{\alpha}_n)$. We have

$$\ell_n(\boldsymbol{\beta}(\tilde{\alpha}_n)) - \ell_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) = [\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)]^T \boldsymbol{s}_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) - \frac{1}{2} [\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)]^T \boldsymbol{H}_n(\tilde{\boldsymbol{\beta}}(\tilde{\alpha}_n)) [\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)]$$

where $\tilde{\boldsymbol{\beta}}(\tilde{\alpha}_n) \in [\boldsymbol{\beta}(\tilde{\alpha_n}), \boldsymbol{\beta}^*(\tilde{\alpha}_n)]$. By conditions (E) and (F),

$$\frac{1}{2} [\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)]^T \boldsymbol{H}_n(\tilde{\boldsymbol{\beta}}(\tilde{\alpha}_n)) [\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)] \geq \frac{1}{4} C_1 ||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)||^2$$

Therefore,

$$\ell_n(\boldsymbol{\beta}(\tilde{\alpha}_n)) - \ell_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) \leq [\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)]^T \boldsymbol{s}_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) - \frac{C_1}{4} n ||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)||^2$$

Hence, for any $\boldsymbol{\beta}(\tilde{\alpha}_n)$ such that $||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)|| = n^{-\frac{1}{4}}$, we have

$$l_n(\boldsymbol{\beta}(\tilde{\alpha}_n)) - l_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) \leq n^{-\frac{1}{4}} ||\boldsymbol{s}_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n))|| - \frac{C_1}{2} n^{\frac{1}{4}} \quad \text{(A. 12)}$$

By Lemma 2, we can show that there exists $N_3 > 0$, for any $n \geq N_3$, we have

$$P\Big(\sup_{\alpha_n \in A_{1n} \bigcup \alpha_n^*} ||\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n))|| \leq [(p+1)n]^{\frac{1}{2}} \log n\Big) \geq 1 - \frac{\epsilon}{4}$$

Let $R_{3n} = \{\sup_{\alpha_n \in A_{1n} \bigcup \alpha_n^*} ||\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n))|| \leq [(p+1)n]^{\frac{1}{2}} \log n\}$. Therefore for R_{3n}, there exists $N_4 > 0$, for any $n \geq N_4$, we have

$$(13) \leq c(n^{\frac{1}{4}}(p+1)^{\frac{1}{2}} \log n - n^{\frac{1}{2}}) = -cn^{\frac{1}{2}}(1 - (p+1)^{\frac{1}{2}} n^{-\frac{1}{4}} \log n) \leq -cn^{\frac{1}{2}}$$

uniformly over $\tilde{\alpha}_n$ and $\boldsymbol{\beta}(\tilde{\alpha}_n)$ such that $||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)|| = n^{-\frac{1}{4}}$ for a generic positive constant c. Because $\ell_n(\boldsymbol{\beta}(\tilde{\alpha}_n))$ is concave in $\boldsymbol{\beta}(\tilde{\alpha}_n)$, the above result implies that maximum of $\ell(\boldsymbol{\beta}(\tilde{\alpha}_n))$ is attained inside $||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)|| \leq n^{-\frac{1}{4}}$.

The concavity also implies that $\sup\{\ell_n(\boldsymbol{\beta}(\tilde{\alpha}_n)) - \ell_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) : ||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)|| \geq n^{-\frac{1}{4}}\} \leq \sup\{\ell_n(\boldsymbol{\beta}(\tilde{\alpha}_n)) - \ell_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) : ||\boldsymbol{\beta}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)|| = n^{-\frac{1}{4}}\} \leq -cn^{\frac{1}{2}}$ uniformly over $\tilde{\alpha}_n \in A_{1n} \bigcup \{\alpha_n^*\}$. Now let $\breve{\boldsymbol{\beta}}(\tilde{\alpha}_n)$ be $\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n)$ augmented with zeros corresponding to the elements in $\tilde{\alpha_n} \backslash \alpha_n$. It can be seen that

$$||\breve{\boldsymbol{\beta}}(\tilde{\alpha}_n) - \boldsymbol{\beta}^*(\tilde{\alpha}_n)|| \geq ||\boldsymbol{\beta}^*(\alpha_n^* \backslash \alpha_n)|| > n^{-\frac{1}{4}}$$

Therefore,

$$\ell_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n)) - \ell_n(\boldsymbol{\beta}^*(\alpha_n^*)) = \ell_n(\breve{\boldsymbol{\beta}}(\tilde{\alpha}_n)) - \ell_n(\boldsymbol{\beta}^*(\tilde{\alpha}_n)) \leq -cn^{\frac{1}{2}}$$

Finally let $N = \max\{N_1, N_2, N_3, N_4\}$, for any $n \geq N$, then on $R_{1n} \bigcap R_{2n} \bigcap R_{3n}$ we have

$$\inf_{\alpha_{\lambda,\tau} \in A_{2n}} BIC(\lambda, \tau) - BIC(\lambda_n, \tau_n) \geq \frac{1}{n}(2cn^{\frac{1}{2}} - m(\alpha_n^*) + o(1)) \log n > 0$$

Note that $P(R_{1n} \bigcap R_{2n} \bigcap R_{3n}) > 1 - \epsilon$, for the arbitrariness of ϵ, we have

$$P(\{\inf_{\alpha_{\lambda,\tau} \in A_{2n}} BIC(\lambda, \tau) - BIC(\lambda_n, \tau_n) > 0\}) \to 1$$

Finally we proof the consistency of unimportant variable removal, say (b) in Theorem 2.

Proof For R_{1n} we have,

$$\begin{aligned}
& BIC(\lambda, \tau) - BIC(\lambda_n, \tau_n) \\
\geq\ & -\frac{2}{n} l_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_{\lambda,\tau})) - [-\frac{2}{n} l(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n^*))] + o(\frac{\log n}{n}) + \frac{(m(\alpha_{\lambda,\tau}) - m(\alpha_n^*)) \log n}{n} \\
\geq\ & -\frac{2}{n} \ell_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_{\lambda,\tau})) - [-\frac{2}{n} \ell(\boldsymbol{\beta}^*(\alpha_n^*))] + o(\frac{\log n}{n}) + \frac{(m(\alpha_{\lambda,\tau}) - m(\alpha_n^*)) \log n}{n}
\end{aligned}$$

Let $\boldsymbol{u} = \hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n) - \boldsymbol{\beta}^*(\alpha_n)$, where $\alpha_n \in A_{1n}$ Then using lemma 2, there exists $N_5 > 0$, for any $n \geq N_5$, $\epsilon = \frac{1}{100}$ in (F) become applicable. By Taylor expansion,

$$\begin{aligned}\ell_n(\hat{\boldsymbol{\beta}}_{\text{mle}}(\alpha_n)) - \ell_n(\boldsymbol{\beta}^*(\alpha_n)) &= \boldsymbol{u}^T \boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) - \frac{1}{2}\boldsymbol{u}^T(\boldsymbol{H}_n(\tilde{\boldsymbol{\beta}}(\alpha_n)))\boldsymbol{u} \\ &\leq \frac{1}{2(1-\frac{1}{100})}\boldsymbol{s}_n^T(\boldsymbol{\beta}^*(\alpha_n))\boldsymbol{H}_n^{-1}(\boldsymbol{\beta}^*(\alpha_n)))\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n))\end{aligned}$$

Thus for R_{1n}, if we want to get $\inf_{\alpha_{\lambda,\tau}\in A_{1n}} BIC(\lambda,\tau) - BIC(\lambda_n,\tau_n) \leq 0$, it sufficient to get

$$\sup_{\alpha_n \in A_{1n}} \boldsymbol{s}_n^T(\boldsymbol{\beta}^*(\alpha_n))\boldsymbol{H}_n^{-1}(\boldsymbol{\beta}^*(\alpha_n)))\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) \geq (1-\frac{1}{100})^2 \log n[m(\alpha_n) - m(\alpha_n^*)]$$

Note that

$$\begin{aligned}&\sup_{\alpha_n \in A_{1n}} \boldsymbol{s}_n^T(\boldsymbol{\beta}^*(\alpha_n))\boldsymbol{H}_n^{-1}(\boldsymbol{\beta}^*(\alpha_n)))\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) \\ \leq\ & \sup_{\alpha_n \in A_{1n}} \frac{1}{n}\frac{1}{C_1}\boldsymbol{s}_n^T(\boldsymbol{\beta}^*(\alpha_n))\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) = \sup_{\alpha_n \in A_{1n}} \frac{1}{nC_1}\Big(\sum_{j\in\alpha_n^*\bigcup\{0\}} s_{nj}^2(\boldsymbol{\beta}^*(\alpha_n)) + \sum_{j\in\alpha_n\setminus\alpha_n^*} s_{nj}^2(\boldsymbol{\beta}^*(\alpha_n))\Big) \\ =\ & \frac{1}{nC_1}\sum_{j\in\alpha_n^*\bigcup\{0\}} s_{nj}^2(\boldsymbol{\beta}^*(\alpha_n)) + \sup_{\alpha_n\in A_{1n}} \sum_{j\in\alpha_n\setminus\alpha_n^*} s_{nj}^2(\boldsymbol{\beta}^*(\alpha_n))\end{aligned}$$

So we can get

$$\begin{aligned}&\{\sup_{\alpha_n \in A_{1n}} \boldsymbol{s}_n^T(\boldsymbol{\beta}^*(\alpha_n))\boldsymbol{H}_n^{-1}(\boldsymbol{\beta}^*(\alpha_n)))\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) \geq (1-\frac{1}{100})^2 \log n[m(\alpha_n) - m(\alpha_n^*)]\} \\ \subseteq\ & \{\sum_{j\in\alpha_n^*\bigcup\{0\}} s_{nj}^2(\boldsymbol{\beta}_n^*) \geq \frac{1}{100}(1-\frac{1}{100})^2 C_1 \log n\}\bigcup \bigcup_{j\in\alpha_n\setminus\alpha_n^*}\{\frac{1}{n}s_{nj}^2(\boldsymbol{\beta}^*(\alpha_n)) \geq (1-\frac{1}{100})^3 C_1 \log n\}\end{aligned}$$

Note that

$$\begin{aligned}&\lim_{n\to\infty} P\Big(\sup_{\alpha_n\in A_{1n}} \frac{1}{nC_1}\boldsymbol{s}_n^T(\boldsymbol{\beta}^*(\alpha_n))\boldsymbol{s}_n(\boldsymbol{\beta}^*(\alpha_n)) \geq (1-\frac{1}{100})^2 \log n[m(\alpha_n) - m(\alpha_n^*)]\Big) \\ \leq\ & \lim_{n\to\infty} \frac{E\sum_{j\in\alpha_n^*\bigcup\{0\}}\frac{1}{n}s_{nj}^2(\boldsymbol{\beta}_n^*)}{\frac{1}{100}(1-\frac{1}{100})^2 C_1\log n} + \sum_{j\in\alpha_n^{*c}} P\Big(\{\frac{1}{n}s_{nj}^2(\boldsymbol{\beta}_n^*) \geq (1-\frac{1}{100})^3 C_1\log n\}\Big) + P(W_n^c) \quad (13)\end{aligned}$$

As we have shown before, using lemma 1 we have,

$$P\Big(\{\frac{1}{n}s_{nj}^2(\boldsymbol{\beta}_n^*) \geq (1-\frac{1}{100})^3 C_1\log n\}\Big) \leq \exp\{-\frac{1}{2}(1-\frac{1}{100})^4\frac{C_1}{C_2}\log n\}$$

It is easy to show that $E\sum_{j\in\alpha_n^*\bigcup\{0\}}\frac{1}{n}s_{nj}^2(\boldsymbol{\beta}_n^*) \leq C_2(m(\alpha_n^*)+1)$. So we can get,

$$\begin{aligned}(14) \quad &\leq \lim_{n\to\infty}\frac{C_2(m(\alpha_n^*)+1)}{\frac{1}{100}(1-\frac{1}{100})^2 C_1\log n} + \sum_{j\in\alpha_n^{*c}}\exp\{-\frac{1}{2}(1-\frac{1}{100})^4\frac{C_1}{C_2}\log n\} + \frac{\epsilon}{4} \\ &= \lim_{n\to\infty}\exp\{-\frac{1}{2}(1-\frac{1}{100})^4\frac{C_1}{C_2}\log n + \log(p-m(\alpha_n^*))\} + \frac{\epsilon}{4} = \frac{\epsilon}{4}\end{aligned}$$

Then for the arbitrariness of ϵ, we have,

$$P(\{\inf_{\alpha_{\lambda,\tau}\in A_{1n}} BIC(\lambda,\tau) - BIC(\lambda_n,\tau_n) > 0\}) \to 1$$

Table 1: Simulation results for logistic regression in Example 1. "Model size" indicates the averaged model size, "Correct model" indicates the proportion of times the correct model was selected over $1,000$ independent datasets, "FPR" indicates the averaged false positive rate for variable selection, "FNR" indicates the false negative rate for variable selection, "Square error" indicated the averaged square error between the estimated coefficients and true coefficients, "Misclassification rate" indicates the averages misclassification rates on independent test sets. The numbers within parenthesis are corresponding standard errors.

Method	Model size	Correct model	FPR	FNR	Square error	Misclassification rate
$n = 200$, BIC tuning						
LASSO	2.10 (0.04)	0.33	0.17 (0.01)	0.07 (0.003)	0.226 (0.005)	0.384 (0.001)
ALASSO	1.93 (0.03)	0.50	0.11 (0.01)	0.06 (0.003)	0.195 (0.005)	0.378 (0.001)
SCAD	2.98 (0.03)	0.25	0.31 (0.01)	0.02 (0.002)	0.203 (0.005)	0.377 (0.001)
MCP	1.88 (0.02)	0.56	0.10 (0.01)	0.05 (0.003)	0.202 (0.006)	0.375 (0.001)
SELO	1.79 (0.02)	0.56	0.10 (0.01)	0.06 (0.003)	0.214 (0.007)	0.375 (0.001)
$n = 300$, BIC tuning						
LASSO	2.60 (0.03)	0.43	0.20 (0.01)	0.02 (0.002)	0.139 (0.003)	0.370 (0.001)
ALASSO	2.22 (0.02)	0.64	0.11(0.01)	0.02 (0.002)	0.112 (0.003)	0.367 (0.001)
SCAD	3.01 (0.03)	0.28	0.29 (0.01)	0.004 (0.001)	0.114 (0.004)	0.366 (0.001)
MCP	2.07 (0.02)	0.76	0.07 (0.01)	0.02 (0.002)	0.109 (0.004)	0.365 (0.001)
SELO	2.01 (0.01)	0.79	0.06 (0.01)	0.02 (0.002)	0.110 (0.004)	0.365 (0.001)
$n = 200$, DV tuning						
LASSO	4.32 (0.07)	0.18	0.44 (0.01)	0.02 (0.002)	0.216 (0.004)	0.374 (0.001)
ALASSO	3.61 (0.06)	0.27	0.35 (0.01)	0.02 (0.002)	0.215(0.005)	0.374 (0.001)
SCAD	3.96 (0.06)	0.19	0.41 (0.01)	0.02 (0.002)	0.227 (0.005)	0.374 (0.001)
MCP	3.22 (0.06)	0.33	0.30 (0.01)	0.03 (0.002)	0.222 (0.006)	0.373 (0.001)
SELO	2.89 (0.05)	0.36	0.62 (0.01)	0.01 (0.002)	0.233 (0.006)	0.374 (0.001)
$n = 300$, DV tuning						
LASSO	4.30 (0.06)	0.21	0.43 (0.01)	0.01 (0.001)	0.158 (0.03)	0.368 (0.0003)
ALASSO	3.58 (0.06)	0.35	0.33 (0.01)	0.01 (0.001)	0.143 (0.003)	0.366 (0.0003)
SCAD	3.86 (0.06)	0.26	0.38 (0.01)	0.01 (0.001)	0.151 (0.004)	0.367 (0.0003)
MCP	3.21 (0.05)	0.42	0.27 (0.01)	0.01 (0.001)	0.139 (0.004)	0.364 (0.0003)
SELO	2.88 (0.04)	0.48	0.61 (0.01)	0.003 (0.001)	0.134 (0.004)	0.364 (0.0004)

Table 2: Simulation results for logistic regression in Example 2. "Model size" indicates the averaged model size, "Correct model" indicates the proportion of times the correct model was selected over $1,000$ simulated datasets, "FPR" indicates the averaged false positive rate for variable selection, "FNR" indicates the false negative rate for variable selection, "Square error" indicated the averaged square error between the estimated coefficients and true coefficients, "Misclassification rate" indicates the averages misclassification rates on independent test sets. The numbers within parenthesis are corresponding standard errors.

Method	Model size	Correct model	FPR	FNR	Square error	Misclassification rate
$n = 200$, BIC tuning						
LASSO	1.82 (0.05)	0.25	0.16 (0.01)	0.036 (0.001)	0.286 (0.005)	0.394 (0.001)
ALASSO	2.02 (0.04)	0.30	0.18 (0.01)	0.029 (0.001)	0.248 (0.006)	0.386 (0.001)
SCAD	4.08 (0.05)	0.06	0.49 (0.01)	0.008 (0.001)	0.237 (0.005)	0.381 (0.001)
MCP	2.17 (0.04)	0.36	0.20 (0.01)	0.024 (0.001)	0.278 (0.008)	0.381 (0.001)
SELO	2.04 (0.03)	0.42	0.19 (0.01)	0.025 (0.001)	0.302 (0.009)	0.380 (0.0007)
$n = 300$, BIC tuning						
LASSO	2.45 (0.04)	0.41	0.19 (0.01)	0.013 (0.001)	0.194 (0.004)	0.379 (0.0007)
ALASSO	2.37 (0.03)	0.51	0.16 (0.01)	0.01 (0.001)	0.146 (0.004)	0.372 (0.001)
SCAD	4.35 (0.04)	0.05	0.50 (0.01)	0.002 (0.001)	0.144 (0.004)	0.371 (0.001)
MCP	2.28 (0.03)	0.58	0.13 (0.01)	0.009 (0.001)	0.147 (0.005)	0.370 (0.001)
SELO	2.19 (0.02)	0.64	0.12 (0.01)	0.009 (0.001)	0.142 (0.006)	0.368 (0.001)
$n = 200$, DV tuning						
LASSO	7.17 (0.16)	0.10	0.58 (0.01)	0.007(0.001)	0.30 (0.01)	0.381 (0.001)
ALASSO	5.58 (0.14)	0.10	0.48 (0.01)	0.013 (0.001)	0.31 (0.01)	0.382 (0.001)
SCAD	7.35 (0.14)	0.06	0.55 (0.01)	0.009 (0.001)	0.34 (0.01)	0.381 (0.001)
MCP	4.62 (0.12)	0.21	0.42 (0.01)	0.014 (0.001)	0.36 (0.01)	0.381 (0.001)
SELO	4.60 (0.14)	0.28	0.35 (0.01)	0.018 (0.001)	0.45 (0.01)	0.383 (0.0007)
$n = 300$, DV tuning						
LASSO	7.36 (0.16)	0.13	0.56 (0.01)	0.003 (0.0004)	0.215 (0.005)	0.373 (0.001)
ALASSO	5.64 (0.14)	0.22	0.46 (0.01)	0.005 (0.0005)	0.203 (0.005)	0.372 (0.001)
SCAD	6.39 (0.14)	0.14	0.53 (0.01)	0.003 (0.0004)	0.225 (0.006)	0.372 (0.001)
MCP	4.71 (0.12)	0.30	0.38 (0.01)	0.005 (0.0005)	0.216 (0.006)	0.372 (0.001)
SELO	4.63 (0.14)	0.41	0.32 (0.01)	0.007 (0.0006)	0.251 (0.008)	0.372 (0.001)