Received: 12 August 2017 | Revised: 30 January 2018 | Accepted: 3 February 2018

DOI: 10.1002/asmb.2322

SPECIAL ISSUE PAPER

WILEY

# Likelihood adaptively modified penalties

Yang Feng[1] | Tengfei Li[2] | Zhiliang Ying[1]

[1]Department of Statistics, Columbia University, New York, NY 10027 USA

[2]Biostatistics Department, The University of Texas MD Anderson Cancer Center, Houston, TX 77030 USA

**Correspondence**
Yang Feng, Department of Statistics, Columbia University, New York, NY 10027 USA.
Email: yang.feng@columbia.edu

**Funding information**
National Science Foundation (NSF), Grant/Award Number: DMS-1554804

JEL Classification: Primary 62G05, secondary 62P10

A new family of penalty functions, ie, adaptive to likelihood, is introduced for model selection in general regression models. It arises naturally through assuming certain types of prior distribution on the regression parameters. To study the stability properties of the penalized maximum-likelihood estimator, 2 types of asymptotic stability are defined. Theoretical properties, including the parameter estimation consistency, model selection consistency, and asymptotic stability, are established under suitable regularity conditions. An efficient coordinate-descent algorithm is proposed. Simulation results and real data analysis show that the proposed approach has competitive performance in comparison with the existing methods.

**KEYWORDS**

asymptotic stability, consistency, convexity, negative absolute prior, penalized likelihood, Poisson penalty, sigmoid penalty

## 1 | INTRODUCTION

Classical work on variable selection dates back to Akaike,[1] who proposed to choose a model that minimizes the Kullback-Leibler divergence of the fitted model from the true model, leading to the well-known Akaike Information Criterion (AIC). Schwarz and Gideon[2] took a Bayesian approach by assuming prior distributions with nonzero probabilities on lower-dimensional subspaces. They proposed what is known as the Bayesian information criterion (BIC) method for model selection. Other types of $L_0$ penalties include $C_p$,[3] AIC$_\text{C}$,[4] risk inflation criterion,[5] and extended BIC (EBIC),[6] among others.

The $L_0$ regularization has a natural interpretation in the form of best subset selection. It also exhibits good sampling properties.[7] However, in a high-dimensional setting, the combinatorial problem has NP-complexity, which is computationally prohibitive. As a result, numerous attempts have been made to modify the $L_0$-type regularization to alleviate the computational burden. They include bridge regression,[8] nonnegative garrote,[9] LASSO,[10] smoothly clipped absolute deviation (SCAD),[11] elastic net,[12] adaptive LASSO (ALASSO),[13] Dantzig selector,[14] smooth integration of counting and absolute deviation (SICA),[15] and minimax concave penalty (MCP),[16] among others.

To a certain extent, existing penalties can be classified into one of the following 2 categories: convex penalty and nonconvex penalty. Convex penalties, such as LASSO,[10] can lead to a sparse solution and are stable as the induced optimization problems are convex. Nonconvex penalties, such as SCAD[11] and MCP,[16] can, on the other hand, lead to sparser solutions and the so-called oracle properties (the estimator works as if the identities of nonzero regression coefficients were known beforehand). However, the nonconvexity of the penalty could make the entire optimization problem nonconvex, which, in turn, could lead to a local minimizer, and the solution may not be as stable as the one if instead a convex penalty is used. Therefore, an important issue for nonconvex penalties is a good balance between sparsity and stability. For example, both SCAD and MCP have an extra tuning parameter, which regulates the concavity of the penalty so that, when it exceeds a threshold, the optimization problem becomes convex.

It is well known that penalty functions have Bayesian interpretation. The classical $L_2$ penalty (ridge regression) is equivalent to the Bayesian estimator with a normal prior. The $L_1$-type penalties, such as LASSO and ALASSO, also have Bayesian counterparts (cf. the works of Park and Casella,[17] Griffin and Brown,[18] and Hara and Sillanp[19]).

Breiman[20] initiated the discussion about the issue of stability in model selection. He demonstrated that many model selection methods are unstable but can be stabilized by perturbing the data and averaging over many predictors. Breiman[21] introduced the random forest, providing a way to stabilize the selection process. Bühlmann and Yu[22] derived the theoretical results to analyze the variance reduction effect of bagging in hard decision problems. Meinshausen and Bühlmann[23] proposed what they called stability selection, which combines the subsampling with high-dimensional variable selection methods.

Despite efforts to deal with stability, there are many important and fundamental issues that remain to be addressed. In particular, there is lack of consensus on the precise definition of stability for the high-dimensional penalized likelihood estimators. To address this issue and to evaluate stability for the proposed method, we introduce herein the 2 types of asymptotic stability for the maximum penalized likelihood estimators. The new concepts cover a wide range of situations and provide a mathematical framework under which the issue of stability can be studied rigorously.

The main objective of this paper is to introduce a family of penalty functions for generalized linear models that can achieve a proper balance between sparsity and stability. Because for the generalized linear models, the loss function is often chosen to be the negative log-likelihood, and it is conceivable to take into consideration the form of the likelihood in the construction of penalty functions. The Bayesian connection to the penalty construction and to the likelihood function makes it natural to introduce penalty functions through suitable prior distributions. To this end, we introduce the family of negative absolute priors (NAP) and use it to develop what to be called likelihood adaptively modified penalties (LAMPs).

The rest of this paper is organized as follows. Section 2 introduces the LAMP family with motivations from its Bayesian and likelihood connections. Specific examples are given for the commonly encountered generalized linear models. In Section 3, we introduce the 2 types of asymptotic stability and study the asymptotic properties of LAMP family. The choice of the tuning parameters and an efficient algorithm are discussed in Section 4. In Section 5, we present simulation results and applied the proposed method to the 2 real data sets. We conclude with a short discussion in Section 6. All the technical proofs are relegated to the Appendix.

## 2 | LIKELIHOOD ADAPTIVELY MODIFIED PENALTY

To introduce our approach, we will first focus on the generalized linear models.[24] It will be clear that the approach also works for other types of regression models, including nonlinear regression. Indeed, our simulation results presented in Section 5 also include the probit model, which does not fall into the exponential family induced generalized linear models.

Throughout this paper, we shall use $(\boldsymbol{X}, Y)$ and $(\boldsymbol{X}_i, Y_i)$ to denote the independent and identically distributed random variables for $i = 1, 2, \dots, n$, where $Y_i$ is the response observation of $Y$ and $\boldsymbol{X}_i$ is the $p+1$-dimensional covariate observation of $\boldsymbol{X}$, $\boldsymbol{X}_i = (X_{i0}, X_{i1}, \dots, X_{ip})^\tau$, $X_{i0} \equiv 1$, and $\mathcal{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n)^\tau$. Moreover, we use $\boldsymbol{\theta}$ for $(\alpha, \boldsymbol{\beta}^T)^T$. Following the work of Nelder and Wedderburn,[24] we assume that the conditional density of $Y_i$ given covariates $\boldsymbol{X}_i$ has the following form:

$$f(Y_i, \boldsymbol{\theta}|\boldsymbol{X}_i) = h(Y_i)\exp\left[\frac{Y_i\xi_i - g(\xi_i)}{\phi}\right], \tag{1}$$

where $\xi_i = \boldsymbol{X}_i^T\boldsymbol{\theta}$, $\phi$ is the dispersion parameter, and $g$ is a smooth convex function. Then, up to an affine transformation, the log-likelihood function is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[Y_i\xi_i - g(\xi_i)\right]. \tag{2}$$

Note that the form of $l$ is uniquely determined by $g$.

For a given $g$, we propose the induced penalty function

$$p_\lambda(\beta) = \frac{\lambda^2}{g'(\alpha_1)\lambda_0}\left[g(\alpha_1) - g\left(\alpha_1 - \frac{\lambda_0}{\lambda}\beta\right)\right], \tag{3}$$

which contains 3 parameters $\alpha_1 \leq 0$, $\lambda_0 > 0$, and $\lambda > 0$. The corresponding penalized log-likelihood function is

$$\tilde{l}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{4}$$

Because the penalty function defined by (3) is likelihood specific, we will call such a penalty "LAMP."

Taking $\beta = 0$, we clearly have $p_\lambda(0) = 0$ and $p'_\lambda(0) = \lambda$. Furthermore, taking the first- and second-order derivatives, we have

$$p'_\lambda(\beta) = \lambda \frac{g'\left(\alpha_1 - \frac{\lambda_0}{\lambda}\beta\right)}{g'(\alpha_1)} \text{ and } p''_\lambda(\beta) = -\lambda_0 \frac{g''\left(\alpha_1 - \frac{\lambda_0}{\lambda}\beta\right)}{g'(\alpha_1)}.$$

*Remark* 1. The parameters have clear interpretations: $\lambda$ is the usual tuning parameter to control the overall penalty level; $\alpha_1$ is a location parameter, which may be fixed as a constant; $\lambda_0$ controls the concavity of the penalty in the same way as $a$ in SCAD and $\gamma$ in MCP.

Like many other penalty functions, the family of LAMPs also has a Bayesian interpretation. To see this, we introduce the following prior for any given $g$ that defines the exponential family (1).

**Definition 1.** Let $a_j \leq 0, b_j < 0$, and $c_j, j = 1, \dots, p$, be parameters. Define the following prior density to be called NAP:

$$p(\boldsymbol{\beta}) \propto \prod_{j=1}^{p} \exp\left[-c_j \left\{g(a_j) - g(a_j + b_j|\beta_j|)\right\}\right]. \tag{5}$$

Let

$$a_j = \alpha_1, b_j = -\lambda_0/\lambda, c_j = -\frac{\lambda^2}{g'(\alpha_1)\lambda_0},$$

and we will get the posterior mode exactly the same as optimizing the penalty form of LAMP. We form this to achieve the good asymptotic property as in the next section. Note that $c_j \geq 0$ if $g'(\xi) > 0$.

*Remark* 2. We know that $g'(\xi) = E(Y|x)$. When the response is nonnegative (eg, logistic and Poisson), we typically have $g'(\xi) > 0$, implying that the corresponding LAMP must be concave.

*Remark* 3. The additive form for the penalty function entails that the prior must be of product form. The parameter $b_i$ scales $\beta_i$, whereas $c_i$ represents the rate of decay. For the $i$th parameter $\beta_i$, the larger the values of $b_i$ and $c_i$ are, the more information the prior has for $\beta_i$. Translating it into the penalty function, it means that the values of $b_i$ and $c_i$ represent the level of penalty and they can be adjusted separately for each component.

*Remark* 4. Note that the form of the prior is similar to the form of the conjugate prior, both in their similar shapes to the likelihood function, whereas NAP differs at several points: negative when $g'(\cdot) > 0$, absolute, from separate dimension, and with a LASSO term taken away. Unlike conjugate prior, which assumes additional samples, this can be seen as to take away the absolute value of redundant sample information from each dimension.

It is worth looking into the commonly encountered generalized linear models and examine the properties of the corresponding LAMPs.

*Linear regression*. In this case, $g(\xi) = \xi^2/2$. Thus, LAMP reduces to the elastic net[12]

$$p_\lambda(\theta) = \lambda\theta - \alpha_1^{-1}\frac{\lambda_0}{2}\theta^2.$$

*Logistic regression*. Here, $g(\xi) = \log(1 + e^\xi)$. Consequently, the penalty function

$$p_\lambda(\theta) = \frac{\lambda^2(1+\rho)}{\lambda_0\rho}\log\left[\frac{1+\rho}{1+\rho\exp(-\lambda_0/\lambda\theta)}\right], \tag{6}$$

where $\rho = \exp(\alpha_1) > 0$. This penalty will be called sigmoid penalty.

*Poisson regression*. Since $g(\xi) = e^\xi$, we have

$$p_\lambda(\theta) = \frac{\lambda^2}{\lambda_0}\left[1 - \exp\left(-\frac{\lambda_0}{\lambda}\theta\right)\right].$$

This will be called the Poisson penalty.

*Gamma regression*. For the gamma regression, we have $g(\xi) = -\log(-\xi)$. Then, the penalty has the following form:

$$p_\lambda(\theta) = \frac{\lambda^2\alpha_1}{\lambda_0}\left[\log(-\alpha_1) - \log\left(\frac{\lambda_0}{\lambda}\theta - \alpha_1\right)\right].$$

*Inverse Gaussian regression*. For the inverse Gaussian, we have $g(\xi) = -\sqrt{-2\xi}$. The resulting penalty

$$p_\lambda(\theta) = \frac{2\lambda^2}{\lambda_0}(-\alpha_1)^{1/2}\left[\left(\frac{\lambda_0}{\lambda}\theta - \alpha_1\right)^{\frac{1}{2}} - (-\alpha_1)^{1/2}\right].$$

*Probit regression*. As we mentioned earlier, LAMP approach can also accommodate nonexponential family–based regression models. In particular, it is applicable to the probit regression for binary outcomes. In this case, $g(\xi) = -\log(\Phi(-\xi))$, which leads to the following penalty form:

$$p_\lambda(\theta) = \frac{\lambda^2}{\lambda_0}\frac{\phi(-\alpha_1)}{\Phi(-\alpha_1)}\log\left[\Phi\left(-\alpha_1 + \frac{\lambda_0\theta}{\lambda}\right) - \Phi(-\alpha_1)\right],$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and the density function of the standard normal distribution, respectively.

*Remark* 5. For the above examples, the effect of the parameters on the penalty function is very different across settings. For example, $\alpha_1$ does not play a role in Poisson regression. In addition, we have a natural choice for $\alpha_1$ for all penalty functions. Specifically, we can choose $\alpha_1 = -1$ for the cases of linear, gamma, and inverse Gaussian, and $\alpha_1 = 0$ for the cases of logistic and probit.

The above examples show that the proposed LAMP family is fairly rich. They also differ from the commonly used penalties. Figure 1 contains plots of penalty functions (left panel) along with their derivatives (right panel) that include LASSO, SCAD, and MCP and the 2 members of the LAMP family (sigmoid and Poisson). Here, $\gamma = 1.1$ for MCP, $\gamma = 2.1$ for SCAD, $\lambda_0 = 2/1.1$ for sigmoid and $1/1.1$ for Poisson penalty, and $\rho = 1$ for sigmoid penalty. The parameters are chosen to keep the maximum concavity of these penalties the same. Figure 1 shows that sigmoid and Poisson penalties lie between MCP and SCAD when the maximum concavity is the same. In addition, from the graphs of the derivatives, it is easy to identify that the penalties of the LAMP family have continuous derivatives (actually, they have continuous derivatives of any order for most common generalized linear models) as compared with the discontinuous ones for SCAD and MCP. It will be shown that this feature can make the optimization problem easier and the estimation more stable.

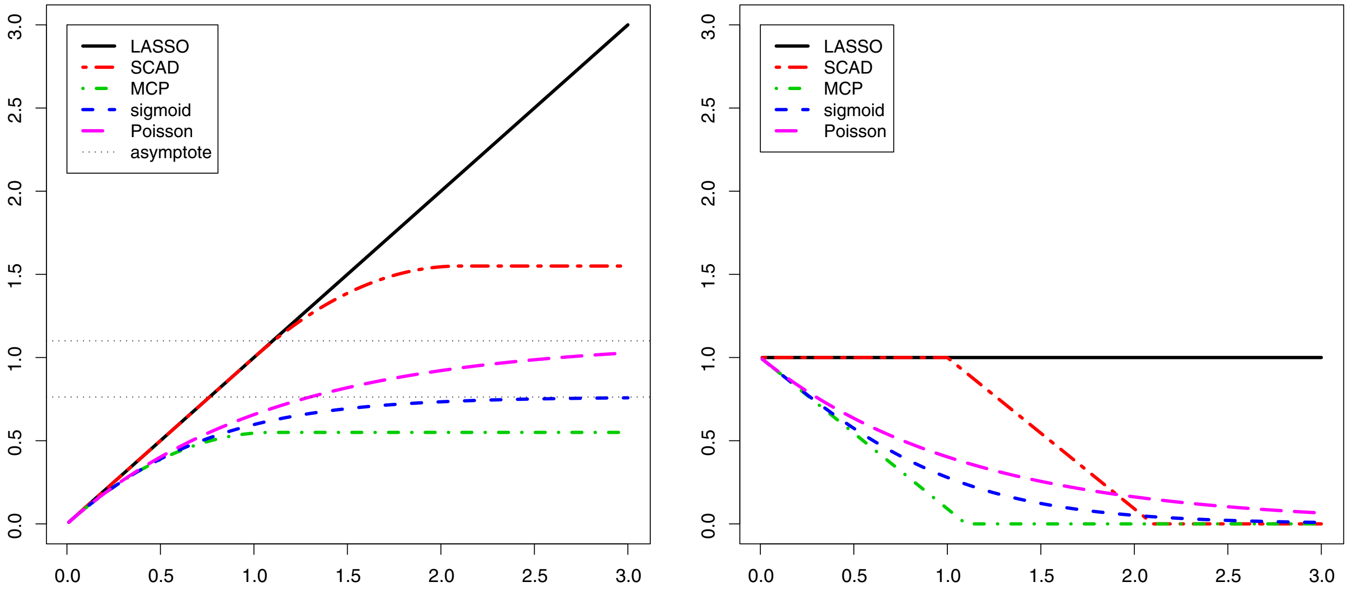

**FIGURE 1** Penalties and the derivative of the penalties. MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation [Colour figure can be viewed at wileyonlinelibrary.com]

## 3 | THEORETICAL PROPERTIES

### 3.1 | Asymptotic stability

Recall that the log-likelihood considered here is concave, and the maximum likelihood estimator (MLE) is uniquely defined and stable. By adding a nonconvex penalty, the concavity may be destroyed. To study stability of the penalized MLE, it is necessary to study the impact of the penalty on the concavity, especially when $n$ is large.

For the penalized maximum log-likelihood estimation procedure (4), if nonconvex penalties are used, the solution to the maximization problem may not be unique. Therefore, it is natural to study the behavior of the local maximizers in penalized likelihood estimates when the observations are perturbed. Here, we introduce a new concept of asymptotic stability to describe the asymptotic performances of local maximizers in penalized likelihood estimates. Note that even for the penalized maximum-likelihood estimators with convex penalty where the unique maximizer exists, such asymptotic stability concept is still useful in characterizing the behavior of the global maximizer.

Suppose that we want to minimize with respect to $\boldsymbol{\theta}$ a criterion function $M_n(\mathcal{Z}_n, \boldsymbol{\theta})$, where $\mathcal{Z}_n = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^T$ and $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, Y_i)^T$ is the $i$th observation of $\boldsymbol{Z} = (\boldsymbol{X}, Y)$. Denote by $\boldsymbol{S}_Z, S_{\mathcal{Z}_n}$, and $\Theta$ the support for $\boldsymbol{Z}, \mathcal{Z}_n$, and domain for $\boldsymbol{\theta}$, respectively. We say that $\boldsymbol{\theta}^*$ is a local minimizer if there exists a neighborhood in which $M_n(\mathcal{Z}_n, \cdot)$ attains its minimum within that neighborhood. More precisely, the set of the local minimizers is defined as

$$\underset{\Theta}{\operatorname{arglmin}} \quad M_n(\mathcal{Z}_n, \boldsymbol{\theta}) \triangleq \left\{ \boldsymbol{\theta}^* \in \Theta | \exists \epsilon > 0, M_n(\mathcal{Z}_n, \boldsymbol{\theta}^*) = \min_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \le \epsilon} M_n(\mathcal{Z}_n, \boldsymbol{\theta}) \right\}.$$

Throughout this paper, $\| \cdot \|$ indicates the square root of sum of square of each element of a matrix or a vector.

It is clear from the definition that the set of local maximizers is random. We characterize its asymptotic behavior in terms of whether or not the set converges to a single point as $n \to \infty$. For a set $A$, define its diameter as $\operatorname{diam}(A) \triangleq \sup_{\boldsymbol{x}\boldsymbol{y} \in A} \|\boldsymbol{x} - \boldsymbol{y}\|$.

**Definition 2.** (Weak asymptotic stability)
We say that the set of local minimizers of $M_n(\boldsymbol{Z}_n, \cdot)$ satisfies weak asymptotic stability if $\forall \delta > 0$,

$$\lim_{n\to\infty} \mathrm{P} \left( \overline{\lim_{\epsilon \to 0}} \operatorname{diam} \left[ \bigcup_{\substack{\|\mathcal{E}_n\|/\sqrt{n} < \epsilon \\ \mathcal{Z}_n + \mathcal{E}_n \in S_{\mathcal{Z}_n}}} \left\{ \underset{\boldsymbol{\theta}}{\operatorname{arglmin}}\, M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}) \right\} \right] > \delta \right) = 0. \tag{7}$$

The weak asymptotic stability characterizes the asymptotic behavior of local minimizers when the data are perturbed slightly. It shows that for large $n$ and small perturbation, the local minimizers stay sufficiently close to each other with high probability. Defined below is a stronger version, which guarantees the uniqueness of the minimizer.

**Definition 3.** (Strong asymptotic stability)
We say that the set of local minimizers of $M_n(\boldsymbol{Z}_n, \cdot)$ satisfies strong asymptotic stability if

$$\lim_{n\to\infty} \mathrm{P} \left( \overline{\lim_{\epsilon \to 0}} \operatorname{diam} \left[ \bigcup_{\substack{\|\mathcal{E}_n\|/\sqrt{n} < \epsilon \\ \mathcal{Z}_n + \mathcal{E}_n \in S_{\mathcal{Z}_n}}} \left\{ \underset{\boldsymbol{\theta}}{\operatorname{arglmin}}\, M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}) \right\} \right] = 0 \right) = 1. \tag{8}$$

*Remark* 6. Under the weak asymptotic stability, multiple minimizers, although shrinking to 0, may exist with high probability for any finite $n$. The strong asymptotic stability, on the other hand, entails that for sufficiently large $n$, the probability of having multiple minimizers must converge to 0, implying that there must be a unique minimizer with high probability.

*Remark* 7. $L_0$ penalties, although may have the weak asymptotic property if we adjust its tuning parameter such as AIC, will never possess the strong asymptotic property because then the solution of each submodel with the number of parameters constrained will be a local optimizer, and with no probability, all solutions will coincide, as $n \to \infty$.

*Remark* 8. Fan and Li (2001) discussed that a good penalty function should result in an estimator with 3 properties: unbiasedness, sparsity, and continuity. Here, we propose to add a fourth property, namely, "asymptotic stability," for a desirable penalty function.

We now consider the situation in which $M_n(\mathcal{Z}_n, \boldsymbol{\theta})$ can be approximated by an independent and identically distributed sum with a remainder term, ie,

$$M_n(\mathcal{Z}_n, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{Z}_i, \boldsymbol{\theta}) + r_n(\boldsymbol{\theta}). \tag{9}$$

This general form includes the form of a negative log-likelihood plus a penalty for us to study. Let $m_n(\mathcal{Z}_n, \boldsymbol{\theta}) \triangleq n^{-1}\sum_{i=1}^{n} m(\boldsymbol{Z}_i, \boldsymbol{\theta})$. We assume throughout the rest that $\Theta$ is compact with the true parameter value, denoted by $\boldsymbol{\theta}_0$, lying in its interior and that $\bar{m}(\boldsymbol{\theta}) = \mathrm{E}m(\boldsymbol{Z}, \boldsymbol{\theta})$ is finite. We need the following regularity conditions.

(C1) $\forall x \in \boldsymbol{S}_Z, m(x, \boldsymbol{\theta})$ is convex about $\boldsymbol{\theta}$, and $\bar{m}(\boldsymbol{\theta})$ is continuous and strictly convex.

(C2) Lipchitz-type condition: there exists projection $K : \boldsymbol{S}_Z \mapsto \mathbb{R}$ such that $\mathrm{E}K^2(\boldsymbol{Z}) < \infty$ and, for arbitrary $\boldsymbol{\delta}_1$,$\boldsymbol{\delta}_2$, and $\boldsymbol{x} : \boldsymbol{x} + \boldsymbol{\delta}_i \in \boldsymbol{S}_Z, i = 1, 2,$

$$\sup_{\boldsymbol{\theta}\in\Theta} |m(\boldsymbol{x} + \boldsymbol{\delta}_1, \boldsymbol{\theta}) - m(\boldsymbol{x} + \boldsymbol{\delta}_2, \boldsymbol{\theta})| \leq K(\boldsymbol{x}) \|\boldsymbol{\delta}_1 - \boldsymbol{\delta}_2\| .$$

(C3) The remainder term is asymptotically flat:

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\Theta} \frac{|r_n(\boldsymbol{\theta}_1) - r_n(\boldsymbol{\theta}_2)|}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|} = 0.$$

(C4) There exists a local minimizer of (9) that is consistent to $\boldsymbol{\theta}_0$.

(C5) There exists $\delta_0 > 0$ such that

$$\lim_{n\to\infty} P\left(M_n(\mathcal{Z}_n, \boldsymbol{\theta}) \text{ is strictly convex within } o(\boldsymbol{\theta}_0, \delta_0) \cap \Theta\right) = 1.$$

*Remark* 9. When $m$ and $r$ are both smooth functions, Conditions (C2) and (C3) can be guaranteed by assuming that the derivative of $m$ is bounded uniformly and that the derivative of $r_n$ tends to 0 uniformly. Condition (C5) is guaranteed by the convexity around the true parameters, uniformly in $n$.

The next 2 lemmas provide sufficient conditions for the 2 types of asymptotic stability.

**Lemma 1.** *If Conditions (C1)-(C3) are satisfied, then we have weak asymptotic stability.*

It is straightforward to verify that for generalized linear models, SCAD, MCP, sigmoid penalty and Poisson penalty, all satisfy Conditions (C1)-(C3), leading to the weak asymptotic stability.

**Lemma 2.** *If Conditions (C1)-(C5) are satisfied, then strong asymptotic stability holds.*

With the 2 kinds of asymptotic stabilities, it makes sense to design an algorithm that converges to any element within that local optimizer set from the view of large sample theory. While in reality with a finite "$n$", it is also important for us to choose a solution that is the closest to the truth or have a smooth enough solution path as the tuning parameter $\lambda$ changes.

## 3.2 | Asymptotic properties

In this section, we study asymptotic properties for the proposed LAMP, including parameter estimation consistency, model selection consistency, and asymptotic stability. Let $q \triangleq \sum_{j=1}^{p} \mathrm{I}_{\beta_j \neq 0}$ denote the number of signals and $m = \inf\{|\beta_1|, \ldots, |\beta_q|\}$ denote the minimum signal level. Here, we consider the setting of fixed $p, q, m$ when $n \to \infty$ although some results may be extended to the case of $p = o(\sqrt{n})$.

We first introduce certain regularity conditions that are needed for establishing asymptotic properties.

(C6) Let $g^{(k)}$ denotes the $k$th derivative of $g$. Then,

$$E|Y| + \sup_{\substack{\boldsymbol{\theta}\in\Theta \\ 0\leq k\leq 2}} E\left(\|\boldsymbol{X}\|^k \left|g^{(k)}\left(\boldsymbol{X}^T\boldsymbol{\theta}\right)\right|\right) + .$$

For any $\boldsymbol{\theta}$, $E|Y\boldsymbol{X}^T\boldsymbol{\theta}-g(\boldsymbol{X}^T\boldsymbol{\theta})| < \infty$ and $\lambda_{\min}(\mathrm{E}\boldsymbol{X}g''(\boldsymbol{X}^T\boldsymbol{\theta})\boldsymbol{X}^T) > 0$, where $\lambda_{\min}(D)$ of a symmetric matrix $D$ denotes its minimum eigenvalue. $g'(x) > 0$ at $x \in [-\infty, \alpha_1)$ and $\lim_{\xi \to -\infty} g'(\xi) = 0$; $g''(\xi)$ is increasing at $[-\infty, \alpha_1)$, where $\alpha_1 \leq 0$ is a constant. There exists $\delta > 0$ such that

$$E\left[\|\boldsymbol{X}\|_2^2 g''\left(\boldsymbol{X}^T\boldsymbol{\theta}_0\right) + \|\boldsymbol{X}\|_1^3 \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| \leq \delta} g'''\left(\boldsymbol{X}^T\boldsymbol{\theta}\right)\right] < \infty.$$

(C7) $\lambda \to 0$, $\sqrt{n}\lambda \to +\infty$, and $\sqrt{n}\lambda g'(\alpha_1 - \lambda_0 m/\lambda) \to 0$.

(C8) There exists a neighborhood of $\boldsymbol{\theta}_0$, denoted by $\Theta_1$, such that

$$\lambda_{\min}\left\{\mathrm{E}[\boldsymbol{X}\boldsymbol{X}^T]\right\} > \frac{\lambda_0 g''(\alpha_1)}{|g'(\alpha_1)|} \frac{1}{\min_{\boldsymbol{\theta} \in \Theta_1} g''(\boldsymbol{X}^T\boldsymbol{\theta})}.$$

**Theorem 1.** *Suppose that Conditions (C6)-(C8) are satisfied. Then, the penalized maximum-likelihood estimator based on the LAMP family is consistent and asymptotically normal and achieves model selection consistency and strong asymptotic stability.*

**Lemma 3.** *Suppose that (C6) is satisfied and* $\log[g'(-x)] = x^u L(x)$, *where constant* $u > 0$ *and function* $L(x)$ *is negative and slowly varying at* $\infty$, *ie, for any* $a > 0$, $\lim_{x \to \infty}[L(ax)/L(x)] = 1$, *and* $\lim_{x \to +\infty} L(x) \in [-\infty, 0)$. *Then,* $n^{-1/2} \ll \lambda \ll \lambda_0(\log n)^{-1/u} \ll 1$ *implies (C7).*

It can be verified that the logistic regression, the Poisson regression, and the probit model all satisfy the conditions of Lemma 3.

*Remark* 10. To achieve model selection consistency, we can choose $\lambda_0 = o(1)$ and $\lambda = o((\log n)^{-1/u}\lambda_0)$.

*Remark* 11. For the case $p = p_n \gg n, q = q_n \to \infty, m = m_n \to 0$, we can directly apply the results in the work of Fan and Lv.[25] The corresponding conditions (B1)-(B10) on the rate of $p_n, q_n, m_n$ and the covarite matrix $\boldsymbol{X}$, response $\boldsymbol{Y}$ can be found in the appendix. If conditions (B1)-(B10) are satisfied, there exists $\hat{\boldsymbol{\theta}} \in \Theta$ such that

$$\mathrm{P}\left[\left\{\hat{\boldsymbol{\beta}}_2 = 0\right\} \cap \left\{\left\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right\|_\infty = O\left(n^{-\gamma_1}\log(n)\right)\right\}\right] \geq 1 - 2\left[q_n n^{-1} + (p_n - q_n)(1/n)^{n^{1-2\gamma_4}}\right].$$

Now, we present the corresponding results for the case of sigmoid penalty. For sigmoid penalty, $\log(g'(-x)) = xL(x)$, where $L(x) = [-\log(1 + e^x)/x]$, is a slowly varying function. The following conditions replace (C6)-(C8).

(C6') $E\|\boldsymbol{X}\|^3 < +\infty$.

(C7') Constant $0 < \rho < +\infty$, $n^{-\frac{1}{2}} \ll \lambda \ll \lambda_0/\log n \ll 1$.

(C8') There exists a neighborhood of $\boldsymbol{\theta}_0$, denoted by $\Theta_1$, such that

$$\lambda_0 < \min_{\theta \in \Theta_1} \lambda_{\min}\left[\mathrm{E}\left(\boldsymbol{X}\boldsymbol{X}^T \frac{e^{\boldsymbol{X}\theta}}{\left(1 + e^{\boldsymbol{X}^T\theta}\right)^2}\right)\right](1 + \rho)/\rho.$$

The following proposition shows that the results of Theorem 1 carry over to the penalized logistic regression when conditions are replaced.

**Proposition 1.** *For penalized maximum-likelihood estimator of logistic regression, where the sigmoid penalty (6) is applied, we have parameter estimation consistency, model selection consistency, and strong asymptotic stability under conditions (C6')-(C8').*

## 4 | ALGORITHMS

An important aspect of a penalized likelihood estimation method is the computational efficiency. For the LASSO penalty, Efron et al[26] proposed the path-following LARS algorithm. In the work of Friedman et al,[27] the coordinatewise descent method was proposed. It optimizes a target function with respect to a single parameter at a time, cycling through all parameters until convergence is reached. For nonconvex penalties, Fan and Li[11] used the LQA approximation approach. In the work of Zou and Li,[28] a local linear approximation–type method was proposed for maximizing the penalized likelihood for a broad class of penalty functions. In the work of Fan and Lv,[25] the coordinatewise descent method was implemented for nonconvex penalties as well. Yu and Feng[29] proposed a hybrid approach of Newton-Raphson and coordinate descent for calculating the approximate path for penalized likelihood estimators with both convex and nonconvex penalties.

### 4.1 | Iteratively reweighted least squares

We apply quadratic approximation and use the coordinate decent algorithm similar to the works of Keerthi and Shevade.[30,31] Recall that our objective function is the penalized log-likelihood $\tilde{l}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$. Following the work of Keerthi and Shevade,[30] let $F_j = \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}$, and define the violation function

$$\text{viol}_j(\boldsymbol{\theta}) \triangleq \begin{cases} |F_j|, & \text{if } j = 0, \\ \max\{0, -n\lambda - F_j, -n\lambda + F_j\}, & \text{if } \theta_j = 0,\ j > 0, \\ \left|F_j - n\text{sgn}(\theta_j)p_\lambda'\left(|\theta_j|\right)\right|, & \text{if } \theta_j \neq 0,\ j > 0. \end{cases}$$

We see that the objective function achieves its maximum value if and only if $\text{viol}_j = 0$, for all $j$. Thus, we use $\max_j\{\text{viol}_j\} < \tau$ as the stop condition of our iteration, with $\tau > 0$ being the chosen tolerance threshold.

In each step, we use quadratic approximation to the log-likelihood function $l(\boldsymbol{\theta}) \approx -\frac{1}{2}(\tilde{\boldsymbol{y}} - \boldsymbol{X\theta})^T \mathcal{W}(\tilde{\boldsymbol{y}} - \boldsymbol{X\theta}) + \text{constant}$, where $\mathcal{W}$ and $\tilde{\boldsymbol{y}}$ depend on the current value of $\boldsymbol{\theta}$. The algorithm is summarized as follows.

**Algorithm:** Set values for $\tau > 0, \lambda, \lambda_0, \rho > 0$. Denote by $\boldsymbol{X}_{\cdot j}$ the $(j+1)$th column of $\mathcal{X}, j = 0, \ldots, p$.

S1. Standardize $\boldsymbol{X}_i, i = 1, 2, \ldots, n$.
S2. Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}^{(0)}$. Calculate $\text{viol} = \max\{\text{viol}_j(\boldsymbol{\theta})\}$ and go to S3 if $\text{viol} > \tau$ or else go to S5.
S3. Choose $j^* \in \arg\max_j \text{viol}_j(\boldsymbol{\theta})$. Calculate $\mathcal{W}, v = \frac{1}{n}\boldsymbol{X}'_{\cdot j^*}\mathcal{W}\boldsymbol{X}_{\cdot j^*}, z = \frac{1}{n}\boldsymbol{X}'_{\cdot j^*}\mathcal{W}(\tilde{\boldsymbol{y}} - \boldsymbol{X\theta}) + v\theta_{j^*}$. If $j^* \neq 0$, let $r = p_\lambda'(|\theta_j^*|)$; else $r = 0$.
S4. If $j^* = 0, \theta_{j^*} = z/v$; else $\theta_{j^*} = \text{sign}(z)(|z| - r)_+$. Calculate $\text{viol} = \max_j\{\text{viol}_\text{j}\}$ and go to S3 if $\text{viol} > \tau$, else go to S5.
S5. Do the transformation of the coefficients $\theta$ due to standardization.

For S2, the initial solution $\boldsymbol{\theta}^{(0)}$ can be taken as the zero solution or the MLE or the estimate calculated using a parameter $\lambda^* \in o(\lambda, \epsilon)$ from the previous steps.

For S4, we first perform the iterations for the variables in the current active set until convergence, then check whether additional variables should join the active set. Alternatively, we may speed up the calculation by using "warm start." Readers are referred to see the works of Keerthi and Shevade[30,32] for details of the strategies to speed up calculation in coordinate descent algorithms.

For example, the logistic regression with sigmoid penalty is $\sum_{i=1}^{n}\log(1 + e^{-y_i\boldsymbol{X}_i^T\theta}) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$, where $p_\lambda(\theta) = \frac{\lambda_n^2(1+\rho)}{\lambda_0}\log[(1+\rho)e^{\lambda_0/\lambda\theta}/(1+\rho e^{\lambda_0/\lambda\theta})]$, for $\theta > 0$. We define $r_i = \exp(-y_i\boldsymbol{X}_i^T\boldsymbol{\theta}), i = 1, \ldots, n$, and $F_j(\theta) = \sum_{i=1}^{n} r_i y_i X_{i,j}/(1 + r_i), j = 0, \ldots, d$. In S3, we have the following 2 different approximation methods for updating.

1. Quadratic approximation from iteratively reweighted least squares.[33] Let $\mathcal{W} = \frac{1}{2}\text{diag}\{\tanh(\frac{\pi_1}{2})/\pi_1, \ldots, \tanh(\frac{\pi_n}{2})/\pi_n\}$ and $\tilde{\boldsymbol{y}} = \frac{1}{2}\mathcal{W}^{-1}\boldsymbol{y}$, where $\pi_i = \boldsymbol{X}_i^T\boldsymbol{\theta}^{(0)}y_i$.
2. Quadratic approximation using Taylor expansion. Let $\mathcal{W} = \text{diag}\{\pi_1(1-\pi_1), \ldots, \pi_n(1-\pi_n)\}$ and $\tilde{\boldsymbol{y}} = \boldsymbol{X\theta}^{(0)} + \mathcal{W}^{-1}[\boldsymbol{y} \circ (1 - \boldsymbol{\pi})]$, where $\pi_i = 1/[1 + \exp(-y_i\boldsymbol{X}_i^T\boldsymbol{\theta}_0)]$ and $\circ$ is the componentwise product operator.

For an initial estimator $\boldsymbol{\theta}^{(0)}$, denote by $a(\boldsymbol{\theta}, \boldsymbol{\theta}^{(0)})$ a quadratic approximation of $-n^{-1}l(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(0)}$, ie,

$$a\left(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(0)}\right) = -n^{-1}l\left(\boldsymbol{\theta}^{(0)}\right), \quad \frac{\partial a\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(0)}\right)}{\partial \boldsymbol{\theta}}|_{\theta = \theta^{(0)}} = -n^{-1}l'\left(\boldsymbol{\theta}^{(0)}\right).$$

**Theorem 2.** *Let $C \subset R^d$ be a closed set and the objective function $M_n(\boldsymbol{\theta}) = -n^{-1}l(\boldsymbol{\theta}) + n\sum_{j=1}^{p} p_\lambda(|\theta_j|)$ is strictly convex. In addition, assuming that the quadratic approximation at $\boldsymbol{\theta}^{(0)}$ satisfies $a(\boldsymbol{\theta}, \boldsymbol{\theta}^{(0)}) \geq -n^{-1}l(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in C$, then the algorithm constrained in C (achieve the minimum within C) converges to the true minimum $\theta^* = \text{argmin}_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta})$. In addition, method 1 for logistic regression satisfies the conditions on quadratic approximation.*

## 4.2 | Balance between stability and parsimony

It is important to address the issue of choosing tuning parameters in the penalized likelihood estimation. For the LAMP family, there are 3 tuning parameters, ie, $\lambda$, $\lambda_0$, and $\alpha_1$. Our numerical experiences show that the resulting estimator is not sensitive to the choice of $\alpha_1$. In most cases, we may simply take $\alpha_1 = -1$ or $\alpha = 0$, depending on the type of regression (See Remark 5). For $\lambda$ and $\lambda_0$, we recommend using cross-validation (CV) or BIC, so long as the solutions are stable enough. The ncvreg package described in the work of Breheny and Huang[34] to determine a stable area or do local diagnosis is recommended. There are 2 approaches to get a stable area: to control the smoothness of the $\lambda$-estimate curve and calculate the smallest eigenvalue of the penalized likelihood at each point of the path as stated in Theorem 1. Here, we take the second approach in all numerical analysis.

Our algorithm differs from ncvreg in the following 2 aspects: we use the "viol" function as the convergence criteria; we do not use the adaptive-scale. Both algorithms 1 and 2 use the linear approximation (suppose at $\theta^{(0)}$) of the penalty term. $p_\lambda(|\theta|) \approx p_\lambda(|\theta^{(0)}|)+p'_\lambda(|\theta^{(0)}|)(|\theta|-|\theta_0|)$. For algorithm 1, from concavity, we have $p_\lambda(|\theta|) < p_\lambda(|\theta^{(0)}|)+p'_\lambda(|\theta^{(0)}|)(|\theta|-|\theta_0|)$, which naturally falls into the minorization-maximization algorithm framework.

To choose the $(\lambda_0, \lambda)$ pair, we use the hybrid approach introduced by the work of Breheny and Huang,[34] ie, combining BIC, CV, and convexity diagnostics. For a path of solutions with a given value of $\lambda_0$ large enough, use AIC/BIC to select $\lambda$ and use the convexity diagnostics to determine the locally convex regions of the solution path. If the chosen solution lies outside the stable region, one can lower $\lambda_0$ to make the penalty more convex. Once this process has been iterated a few times, we can find a value of $\lambda_0$ that produces a good balance between sparsity and convexity. Then, we can fix $\lambda_0$ and use BIC or CV to choose the best $\lambda$.

# 5 | SIMULATION RESULTS AND 2 EXAMPLES

Simulation studies cover logistic, Poisson, and probit cases. The performance of the LAMP family is compared with those of LASSO, SCAD, and MCP. Particular attention is given to the logistic regression to demonstrate how sparsity and stability are properly balanced. Two classification examples from microarray experiments involving persons with cancer are presented.

**TABLE 1** Model selection and estimation results for different penalties over 100 simulations. Mean values are presented with the standard error in parentheses. Here, "sig" represents the sigmoid penalty

| Penalties | TP | FP | cf | of | uf | L1 | L2 |
|---|---|---|---|---|---|---|---|
| sig(0)/LASSO | 1.87 (0.053) | 0.12 (0.041) | 0.05 | 0.01 | 0.94 | 2.49 | 1.46 |
| sig(.02) | 1.87 (0.053) | 0.11 (0.040) | 0.05 | 0.01 | 0.94 | 2.48 | 1.45 |
| sig(.03) | 1.90 (0.052) | 0.13 (0.042) | 0.06 | 0.01 | 0.93 | 2.41 | 1.41 |
| sig(.05) | 1.95 (0.05) | 0.15 (0.041) | 0.08 | 0.01 | 0.91 | 2.34 | 1.36 |
| sig(.07) | 1.99 (0.048) | 0.20 (0.047) | 0.09 | 0.02 | 0.89 | 2.25 | 1.31 |
| sig(.09) | 2.02 (0.047) | 0.20 (0.049) | 0.1 | 0.02 | 0.88 | 2.17 | 1.26 |
| sig(.15) | 2.21 (0.057) | 1.86 (0.43) | 0.1 | 0.19 | 0.71 | 3.18 | 1.36 |
| sig(.38) | 2.61 (0.049) | 4.65 (0.40) | 0.1 | 0.51 | 0.39 | 6.23 | 2.13 |
| SCAD(300) | 1.69 (0.073) | 0.14 (0.043) | 0.04 | 0.01 | 0.95 | 2.94 | 1.74 |
| SCAD(7) | 1.69 (0.073) | 0.14 (0.043) | 0.04 | 0.01 | 0.95 | 2.94 | 1.74 |
| SCAD(5) | 1.84 (0.077) | 3.20 (0.69) | 0.03 | 0.11 | 0.86 | 2.94 | 1.74 |
| SCAD(4) | 2.25 (0.089) | 12 (0.75) | 0.01 | 0.45 | 0.54 | 91.9 | 23.5 |
| MCP(300) | 1.69 (0.073) | 0.14 (0.043) | 0.04 | 0.01 | 0.95 | 2.94 | 1.74 |
| MCP(50) | 1.70 (0.073) | 0.16 (0.047) | 0.04 | 0.01 | 0.95 | 2.94 | 1.74 |
| MCP(15) | 1.75 (0.073) | 0.14 (0.043) | 0.06 | 0.01 | 0.93 | 2.45 | 1.45 |
| MCP(7) | 1.79 (0.071) | 0.14 (0.040) | 0.07 | 0.01 | 0.92 | 2.21 | 1.31 |
| MCP(5) | 1.91 (0.081) | 2.98 (0.69) | 0.09 | 0.1 | 0.81 | 2.07 | 1.23 |
| MCP(4) | 2.29 (0.092) | 11.89 (0.80) | 0.02 | 0.50 | 0.48 | 109.3 | 28.7 |

Abbreviations: MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

## 5.1 | Logistic regression

We simulate from logistic regression model with $n = 200, p = 1000, \alpha = 0, \boldsymbol{\beta} = (1.5, 1, -0.7, \mathbf{0}_{997}^T)^T$, and $\boldsymbol{x} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma_{i,j} = \rho + (1-\rho)1_{i=j}$ with $\rho = 0.5$. The number of replications is 100 for this and all the subsequent simulations.

Table 1 reports true positive (TP), false positive (FP), proportion of correct fit (cf), proportion of over fit (of), proportion of under fit (uf), $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_1$ (L1 loss), and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|_2^2$ (L2 loss). To compare performances among LASSO, SCAD, MCP and the sigmoid penalty, we use glmnet to calculate the LASSO solution path, ncvreg to calculate the SCAD and MCP. For all penalties, we use EBIC[6] to choose the tuning parameter $\lambda$ with other parameters fixed. The EBIC parameter $\eta = 1$.

From Table 1, it is clear that the sigmoid penalty outperforms SCAD and MCP in the sense that at a similar level of TP, the sigmoid penalty results in a smaller FP. Somewhat surprisingly, the LASSO has a competitive performance, which may be attributed to the use of the EBIC selection criterion.

## 5.2 | Smoothness

The same logistic regression model as in Section 5.1 except $\alpha_0 = -3$, $\boldsymbol{\beta}_0 = (1.5, 1, -0.7, 0, 0, 0, 0, 0)^T$ and $\Sigma_{i,j} = 0.5^{|i-j|}$ is used. In Figure 2, we compare the smoothness of solution paths generated from the sigmoid penalty, SCAD, and MCP with the same concavity at 0. It can be seen that this choice will also lead to similar sparse level as in Table 2. SCAD and MCP use the same algorithm as sigmoid does (not adaptive scale as in ncvreg package). The shorter vertical line is the BIC choice of $\lambda$, and the longer one uses a 10-fold cross validation. To avoid variation due to the random division in the CV, the result in Table 2 uses BIC to choose $\lambda$ with the other tuning parameter fixed. Figures 2C and 2D for SCAD and

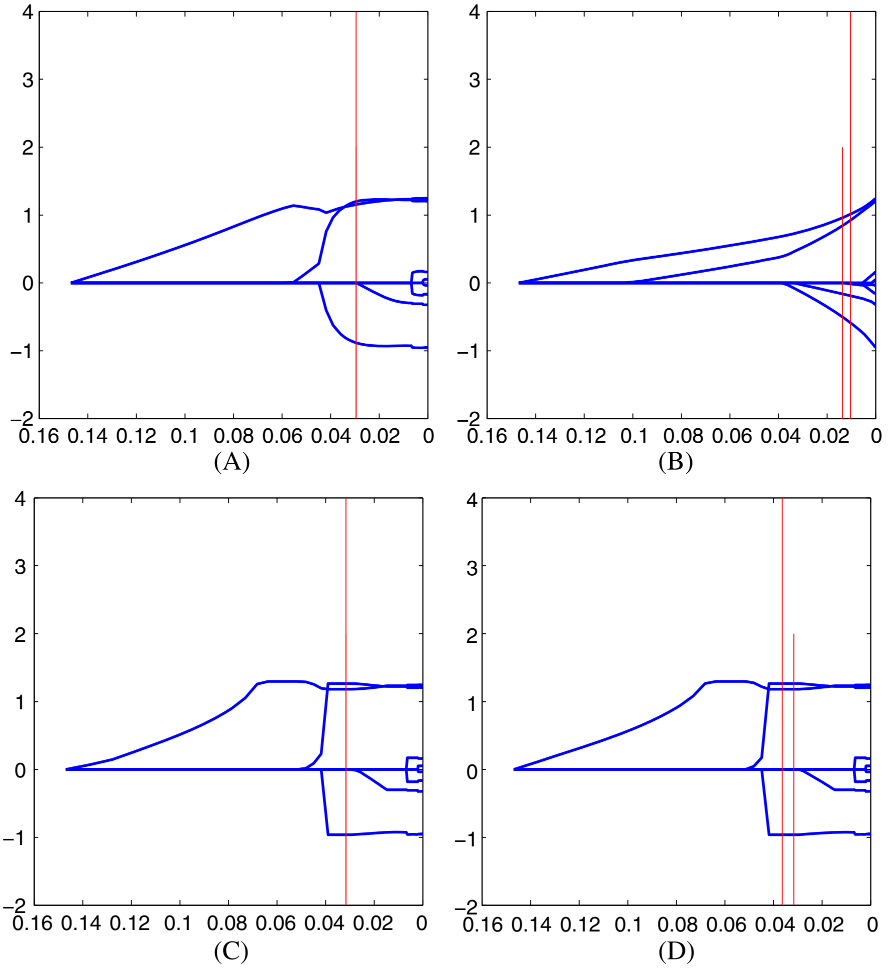

**FIGURE 2** Solution paths of the sigmoid penalty, LASSO, smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP) for the setting described in Section 5.2. A, Sigmoid(0.1); B, LASSO; C, SCAD(20); D, MCP(20) [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 2** The average true positive (TP) and false positive (FP) over 100 simulations with the tuning parameter selected by Bayesian information criterion for the sigmoid penalty, smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP)

| Sigmoid | | | SCAD | | | MCP | | |
|---|---|---|---|---|---|---|---|---|
| $\lambda_0$ | TP | FP | $\gamma$ | TP | FP | $\gamma$ | TP | FP |
| .04 | 1.86 | 0.28 | 1.1 | 1.89 | 0.21 | 1.1 | 1.90 | 0.21 |
| .05 | 1.81 | 0.22 | 7 | 1.89 | 0.21 | 7 | 1.90 | 0.21 |
| .06 | 1.84 | 0.24 | 14 | 1.89 | 0.21 | 14 | 1.90 | 0.21 |
| .08 | 1.89 | 0.22 | 20 | 1.89 | 0.21 | 20 | 1.90 | 0.21 |
| .10 | 1.91 | 0.22 | 27 | 1.88 | 0.23 | 27 | 1.89 | 0.22 |
| .11 | 1.89 | 0.21 | 34 | 1.83 | 0.23 | 34 | 1.86 | 0.24 |
| .13 | 1.89 | 0.21 | 41 | 1.83 | 0.24 | 41 | 1.83 | 0.23 |
| .15 | 1.89 | 0.21 | 54 | 1.83 | 0.28 | 54 | 1.83 | 0.29 |
| .16 | 1.89 | 0.21 | 60 | 1.86 | 0.32 | 60 | 1.86 | 0.32 |
| .18 | 1.89 | 0.21 | 67 | 1.88 | 0.35 | 67 | 1.88 | 0.36 |

MCP are both are less smooth than Figure 2A for the sigmoid penalty, which is of similar smoothness as Figure 2B for LASSO. The sigmoid penalty appears to outperform SCAD and MCP in terms of smoothness of the solution path.

## 5.3 | Stability

The data are generated in the same way as in the preceding section. For each replication, we repeat 100 times CV to select $\lambda$ and calculate its mean sample standard deviation. The box plots for the mean sample standard deviations are generated. In addition, to evaluate the asymptotic stability as introduced in Section 5.3, we add a small random perturbation generated from $N(0, 0.1)$ to all the observations before conducting the analysis. The box plot results are presented in Figure 3.

The result without the random error term evaluates the stability regarding the randomness of CV for each penalty. The result with the random error evaluates the stability toward the random perturbation of the data. To ensure a fair comparison, we choose the same level of concavity at 0 for SCAD, MCP, and the sigmoid penalty. It is seen from the box plots that LASSO is the most stable one, whereas the sigmoid penalty outperforms SCAD and MCP in terms of both the median and 75% quantile in the case without error and 75% quantile in the case with the error added.

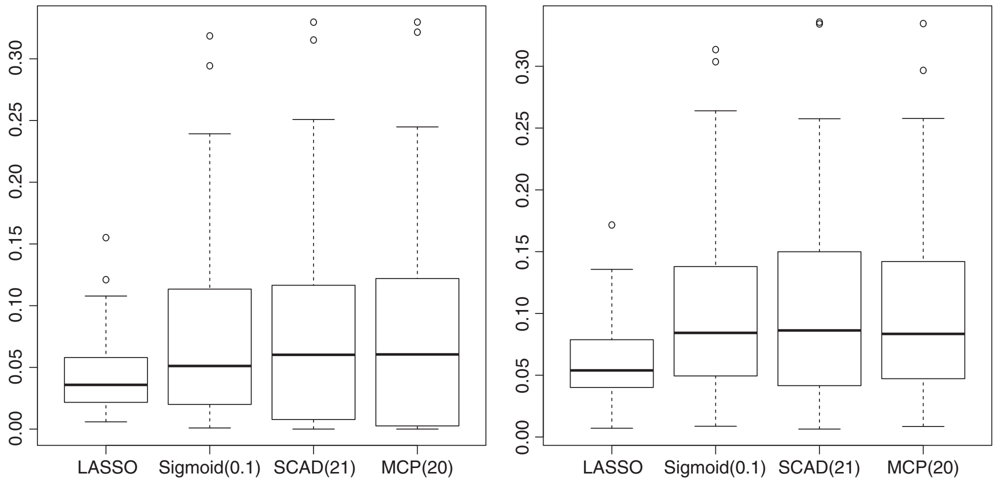

**FIGURE 3** Box plots for the mean standard deviations in Section 5.3. The left panel shows the box plot without perturbation, and the right one is that with perturbation. MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation

## 5.4 | Ultra-high-dimensional logistic regression with weak effects

In this section, we consider an ultra-high-dimensional logistic regression model with a mixture of strong and weak signals. In particular, we set $n = 300$, $p = 5000$, $\alpha = 0$, $\beta = (2, 1.5, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, \mathbf{0}_{4988}^T)^T$, and $\boldsymbol{x} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma_{i,j} = \rho + (1-\rho)1_{i=j}$ with $\rho = 0.5$. Here, we use EBIC to find the optimal parameters for all penalties.

The results are presented in Table 3. It is clear that LAMP leads to the smallest L1 and L2 losses and its edge over SCAD and MCP is very obvious in this scenario with weak signals. On the other hand, we observe that LASSO has the best performance in terms of TP and FP; however, it has a significantly larger L1 and L2 losses than LAMP.

## 5.5 | Poisson and probit

We simulate from the Poisson regression model with $n = 200$, $p = 5000$, $\alpha = -1$, $\boldsymbol{\beta} = (.7, .6, .5, .4, .3, .2, .1, \mathbf{0}_{4993}^T)^T$, and $\boldsymbol{x} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma_{i,j} = \rho + (1-\rho)1_{i=j}$ with $\rho = 0.5$. For probit regression, we simulate $\boldsymbol{x}$ the same way and set $\alpha = 0$ and $\boldsymbol{\beta} = (3, 2, 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, , \mathbf{0}_{4988}^T)^T$.

The results for Poisson regression and probit regression are summarized in Tables 4 and 5, respectively. For Poisson regression, it is clear that by using Poisson penalty, we have the smallest L1 and L2 losses. SCAD and MCP are a bit too aggressive in the sense that they miss many weak signals. For the probit model, we observe similar results as the ultra-high-dimensional logistic regression model, with Probit penalty having a competitive performance over all the criteria we considered.

**TABLE 3** Model selection and estimation results for different penalties over 100 repetitions. Mean values are presented with the standard error in parentheses

| Penalties | TP | FP | L1 | L2 |
|---|---|---|---|---|
| LASSO | 7.62(0.10) | 1.29(0.13) | 6.81(0.03) | 5.19(0.05) |
| Sigmoid | 5.50(0.07) | 3.52(0.25) | 5.73(0.07) | 3.50(0.09) |
| SCAD | 4.68(0.07) | 3.65(0.18) | 13.20(0.44) | 15.12(0.93) |
| MCP | 4.69(0.07) | 4.18(0.16) | 14.57(0.39) | 17.73(0.93) |

Abbreviations: MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

**TABLE 4** Model selection and estimation results for different penalties over 100 repetitions under Poisson regression. Mean values are presented with the standard error in parentheses

| | TP | FP | L1 | L2 |
|---|---|---|---|---|
| LASSO | 5.61(0.08) | 9.78(0.44) | 2.04(0.03) | 0.69(0.02) |
| Poisson | 5.60(0.07) | 9.24(0.50) | 1.83(0.03) | 0.53(0.02) |
| SCAD | 3.31(0.07) | 3.78(0.21) | 2.97(0.09) | 1.01(0.05) |
| MCP | 3.30(0.07) | 7.53(1.04) | 5.05(0.65) | 3.03(0.79) |

Abbreviations: MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

**TABLE 5** Model selection and estimation results for different penalties over 100 repetitions under probit regression. Mean values are presented with the standard error in parentheses

| | TP | FP | L1 | L2 |
|---|---|---|---|---|
| LASSO | 6.37(0.11) | 1.00(0.12) | 9.13(0.02) | 11.72(0.05) |
| Probit | 5.67(0.12) | 1.41(0.15) | 8.44(0.03) | 9.65(0.09) |
| SCAD | 3.89(0.07) | 1.52(0.10) | 8.87(0.24) | 8.11(0.44) |
| MCP | 3.91(0.07) | 1.59(0.12) | 9.54(0.43) | 10.42(1.48) |

Abbreviations: MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

**TABLE 6** Classification errors of lung cancer data

| Penalties | $\lambda_0/\gamma$ | Training Error | Test Error | Number of Selected Genes |
|---|---|---|---|---|
| LASSO | — | 0/32 | 7/149 | 15 |
| SCAD | (3∼6) | 0/32 | 6/149 | 13 |
| MCP | (3∼6) | 0/32 | 7/149 | 3 |
| Sigmoid | (.022∼0.03) | 0/32 | 7/149 | 3 |
| Sigmoid | (.021) | 0/32 | 6/149 | 5 |

Abbreviations: MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

**TABLE 7** Classification errors of prostate cancer data

| Penalties | $\lambda_0/\gamma$ | Training Error | Test Error | Number of Selected Genes |
|---|---|---|---|---|
| LASSO | — | 0/102 | 2/34 | 30 |
| SCAD | (20∼25) | 0/102 | 2/34 | 26 |
| MCP | (35∼50) | 0/102 | 2/34 | 24 |
| Sigmoid | (.001) | 0/102 | 1/34 | 26 |
| Sigmoid | (.002) | 0/102 | 2/34 | 23 |

Abbreviations: MCP, minimax concave penalty; SCAD, smoothly clipped absolute deviation.

### 5.6 | Examples

We apply the proposed LAMP to 2 gene expression data sets: lung cancer data[35] and prostate cancer data.[36] The 2 data sets are available at http://www.chestsurg.org and http://www.broad.mit.edu. The response variable in each data set is binary.

We aim to use the lung cancer data to classify malignant pleural mesothelioma from adenocarcinoma of the lung. The data consist of 181 tissue samples, 32 of them for training and the remaining 149 for testing. Each sample is described by 12 533 genes.

First, the predictors are standardized into mean zero and variance one. We then apply LASSO, SCAD, MCP, and the sigmoid penalty using glmnet for LASSO and ncvreg for SCAD and MCP. For each method, a 10-fold CV is used to select the best $\lambda$. We repeat 10 times to make different divisions to calculate the CV error. For SCAD and MCP, we evaluate the performance when $\gamma \in [3, 6]$, whereas for sigmoid penalty, $\lambda_0 \in (.005, 0.03)$. The results are summarized in Table 6. The result for sigmoid penalty is quite similar to MCP when $\lambda_0 \in (.022, 0.03)$. When $\lambda_0 = 0.021$, we have 6 test errors with only 5 genes selected, which is better compared with SCAD.

For prostate cancer data, the goal is to classify prostate tumor samples from the normal samples. There are 102 patient samples for training, and 34 patient samples for testing with 12 600 genes in total. The result are reported in Table 7. Here, we see the test errors are similar across methods, although the sigmoid penalty leads to the most sparse solution.

## 6 | DISCUSSION

Penalty-based regularization methods have received much attention in recent years. This paper proposes a family of penalty functions (ie, LAMP) that is adaptive and specific to the shapes of the log-likelihood functions. The proposed LAMP family is different from the well-known LASSO, SCAD, and MCP. It can be argued that the new approach provides a good balance between sparsity and stability, the 2 important aspects in model selection. It is shown that the resulting penalized estimation achieves model selection consistency and strong asymptotic stability, in addition to the usual consistency and asymptotic normality.

An important issue is how to choose the 3 parameters imbedded in a LAMP. The "location" parameter $\alpha_1$ can be chosen in an obvious way for the standard generalized linear models, whereas $\lambda$, which represents the penalty level, can be chosen through standard CV, the modified CV,[37,38] or information criteria. For $\lambda_0$, which controls the concavity level, it is computationally intensive to use CV. It is desirable to develop more effective ways to select $\lambda_0$. It is also important to study another type of asymptotic stability, ie, the stability of the solution path.

When there are weak signals, we have shown that the LAMP-based estimators could miss some of the weak signals. To address this issue, there is a recent surge of interest in the post model selection shrinkage estimation.[39] It is vital to study the corresponding results using the LAMP penalized estimators.

Another interesting future work is to generalize the current family of penalties to the Cox model. As the partial likelihood of the Cox model could not be written in the form of (1), we need to develop some other family of penalty functions that is adaptive to the Cox model.

## ACKNOWLEDGEMENT

This work was supported by the National Science Foundation (NSF) CAREER grant DMS-1554804.

## REFERENCES

1. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F, eds. *Second International Symposium on Information Theory*. Budapest, Hungary: Akademiai Kiado; 1973:267-281.
2. Schwarz GE. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
3. Mallows CL. Some comments on $C_p$. *Technometrics*. 1973;15(1):661-675.
4. Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. *Biometrika*. 1989;76(2):297-307.
5. Foster D, George E. The risk inflation criterion for multiple regression. *Ann Stat*. 1994;22:1947-1975.
6. Chen J, Chen Z. Extended Bayesian information criterion for model selection with large model space. *Biometrica*. 2008;94(3):759-771.
7. Barron A, Birge L, Massart P. Risk bounds for model selection via penalization. *Probab Theory Relat Fields*. 1999;113(3):301-413.
8. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics*. 1993;35(2):109-148.
9. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics*. 1995;37(4):373-384.
10. Tibshrani R. Regression shrinkage and selection via the lasso. *J R Stat Soc*. 1996;58:267-288.
11. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Am Stat Assoc*. 2001;96(456):1348-1360.
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc*. 2005;67(2):301-320.
13. Zou H. The adaptive lasso and its oracle properties. *Am Stat Ass*. 2006;101(476):1418-1429.
14. Candes E, Tao T. The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *Ann Stat*. 2007;35(6):2313-2404.
15. Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. *Ann Stat*. 2009;37:3498-3528.
16. Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.
17. Park T, Casella G. The Bayesian lasso. *J Am Stat Assoc*. 2008;103(482):681-686.
18. Griffin JE, Brown PJ. Bayesian hyper-lassos with non-convex penalization. *Australian New Zealand J Stat*. 2010. To appear.
19. Hara RBO, Sillanp MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal*. 2009;4(1):85-118.
20. Breiman L. Heuristics of instability and stabilization in model selection. *Ann Stat*. 1996;24(6):2350-2383.
21. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
22. Bühlmann P, Yu B. Analyzing bagging. *Ann Stat*. 2002;30(4):927-961.
23. Meinshausen N, Bühlmann P. Stability selection. *J Roy Stat Soc Ser B: Stat Methodol*. 2010;72(4):417-473.
24. Nelder JA, Wedderburn RWM. Generalized linear models. *J Roy Stat Soc Ser A: General*. 1972;135(3):370-384.
25. Fan J, Lv J. Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans Inf Theory*. 2011;57(8):5467-5484.
26. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407-499.
27. Friedman J, Hastie T, Hoefling H, Tibshirani R. Pathwise coordinate optimization. *Ann Appl Stat*. 2007;1(2):302-332.
28. Zou H, Li R. One step sparse estimates in non-concave penalized likelihood models. *Ann Stat*. 2008;36(4):1509-1533.
29. Yu Y, Feng Y. APPLE: approximate Path for Penalized Likelihood Estimators. *Stat Comput*. 2014;24(4):803-819.
30. Keerthi SS, Shevade SK. A fast tracking algorithm for generalized LARS/LASSO. *IEEE Trans Neut Netw*. 2007;18(6):1826-1830.
31. Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*. 2003;19(17):2246-2253.
32. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22.
33. Munk A, Bissantz N, Dumbgen L, Stratmann B. Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces. *SIAM J Optim*. 2006;19(4):1828-1845.
34. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann Appl Stat*. 2011;5(1):232-253.
35. Gordon GJ, Jensen RV, Hsiao L-L, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*. 2002;62(17):4963-4967.
36. Singh D, Febbo P, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002;1(2):203-209.
37. Yu Y, Feng Y. Modified cross-validation for penalized high-dimensional linear regression models. *J Comput Graph Stat*. 2014;23(4):1009-1027.
38. Feng Y, Yu Y. Restricted consistent cross-validation for tuning parameter selection in high-dimensional variable selection. Manuscript. arXiv:1308.5390; 2016.

39. Gao X, Ahmed SE, Feng Y. Post selection shrinkage estimation for high-dimensional data analysis. *Appl Stoch Model Bus Ind*. 2017;33(2):97-120.
40. Knight K, Fu W. Asymptotics for LASSO-type estimators. *Ann Stat*. 2000;28:1356-1378.
41. Geyer CJ. On the asymptotics of convex stochastic optimization. Unpublished Manuscript; 1996.

**How to cite this article:** Feng Y, Li T, Ying Z. Likelihood adaptively modified penalties. *Appl Stochastic Models Bus Ind*. 2019;35:330–353. https://doi.org/10.1002/asmb.2322

## APPENDIX A : GENERAL RESULTS

We first state a general result about estimation consistency, model selection consistency, and asymptotic normality.

Consider the penalized log-likelihood function

$$\tilde{l}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_{\lambda,j}\left(|\beta_j|\right),$$

where we assume that $l$ is a smooth function. Let $\boldsymbol{V}_i = (\boldsymbol{X}_i, Y_i) \overset{\text{i.i.d.}}{\sim} f(\boldsymbol{V}, \boldsymbol{\beta})$ be observations with a common support, $i = 1, \dots, n$. Regularity conditions on the identifiable model are satisfied (see the work of Fan and Li[11]).

- $\mathrm{E}_{\beta}[\partial \log f(\boldsymbol{V}, \boldsymbol{\beta})/\partial \beta_j] = 0$,

  $$\mathrm{E}_{\beta}\left[\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\beta})}{\partial \beta_k}\right] = \mathrm{E}_{\beta}\left[\frac{\partial^2 \log f(\boldsymbol{V}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right],$$

  where $j, k = 1, 2, \dots, p$.
- $$0 < \mathrm{E}\left[\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\beta=\beta_0}\right]^{\otimes 2} < \infty.$$
- $\exists$ functions $M_{jkl}$, a neighborhood $w$ of $\boldsymbol{\beta}_0$ such that for almost all $\boldsymbol{V}$ and any $\boldsymbol{\beta} \in w$, the third derivatives $\partial f(\boldsymbol{V}, \boldsymbol{\beta})/(\partial \beta_j \partial \beta_k \partial \beta_l)$ exists, and

  $$\left|\frac{\partial^3 \log f(\boldsymbol{V}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l}\right| \leq M_{jkl}(\boldsymbol{V}),$$

  for any $j$, $k$, and $l$.

Next let $\boldsymbol{\beta}_{10} = (\beta_{1,0}, \dots, \beta_{q,0})^T$ be the nonzero true parameters of $\boldsymbol{\beta}_0 = (\beta_{1,0}, \dots, \beta_{p,0})^T$, whereas $\boldsymbol{\beta}_{20} = (\beta_{(q+1),0}, \dots, \beta_{p,0})^T$; $q$ and $p$ are fixed as $n$ changes. Let

$$m = \min_{1\leq i\leq q} |\beta_{i,0}|; M = \max_{1\leq i\leq q} |\beta_{i,0}|.$$

For convenience, we define the following notations:

$$p_{1,n} \triangleq \sup_{\theta\in[m,M],1\leq j\leq q} \left|p'_{\lambda,j}(\theta)\right|; p_{2,n}(\theta) \triangleq \inf_{q<j\leq p} p'_{\lambda,j}(\theta); p_{2,n} \triangleq p_{2,n}(0)$$

$$p_{3,n}(\theta) \triangleq \sup_{\beta\in(0,\theta),q<j\leq p} \left|p''_{\lambda,j}(\beta)\right|; p_{4,n} \triangleq \sup_{\theta\in[m,M],1\leq j\leq q} \left|p''_{\lambda,j}(\theta)\right|;$$

$$p_{5,n} \triangleq \inf_{\theta\in[m,M],1\leq j\leq q} p''_{\lambda,j}(\theta); \Sigma_1 = \mathrm{diag}\left\{p''_{\lambda,1}(|\beta_{10}|), \dots, p''_{\lambda,q}(|\beta_{q0}|)\right\};$$

$$\boldsymbol{b} = \left(p'_{\lambda,1}\left(|\beta_{10}|\right)\mathrm{sgn}(\beta_{10}), \dots, p'_{\lambda,q}\left(|\beta_{q0}|\right)\mathrm{sgn}(\beta_{q0})\right)^T;$$

$\mathcal{R} = \begin{pmatrix} \mathcal{R}_{11} & \mathcal{R}_{12} \\ \mathcal{R}_{21} & \mathcal{R}_{22} \end{pmatrix}$ is the Fisher information matrix at $\boldsymbol{\beta}_0$ partitioned by the zero and nonzero part.

**Lemma 4.** *Suppose $-l(\boldsymbol{\beta})$ and $-El(\boldsymbol{\beta})$ are both strictly convex for $\boldsymbol{\beta} \in \Theta$; $p_{\lambda,j}, j = 1, \dots, p$ are p functions: $[0,+\infty) \to [0,+\infty)$, continuous at 0. There exist $(\epsilon_n^{(k)})_n < m, k = 1,2,\ n^{-\frac{1}{2}} \vee p_{1,n} = o(\epsilon_n^{(1)})$, such that the kth derivative of $p_{\lambda,j}(\theta)$ exists and is continuous for $\theta \in (0,\epsilon_n^{(k)}) \bigcup [m,M]$, and*

$$p_{\lambda,j}(\theta) \geq 0, \forall 1 \leq j \leq p, \theta > 0; p_{\lambda,j}(0) = 0.$$

$p'_{\lambda,j}(0) \triangleq p'_{\lambda,j}(0_+) \in \bar{\mathbb{R}} = [-\infty,+\infty], p''_{\lambda,j}(0) \triangleq p''_{\lambda,j}(0_+) \in \bar{\mathbb{R}} = [-\infty,+\infty]$.

*Then, we have the following results.*

1. *(Parameter estimation consistency) If either of the following 2 conditions holds, there exists a consistent local maximizer $\hat{\boldsymbol{\beta}}_n$.*

   *(1.a)* $p_{1,n} \to 0, p_{4,n} \to 0$.
   *(1.b)* $p_{1,n} \to 0, \underline{\lim}_n p_{5,n} > 0$.

2. *(Model selection consistency) If any of the following 3 conditions holds, any consistent local maximizer (eg, the one specified in 1) will be model selection consistent.*

   *(2.a)*

$$p_{2,n} > 0, \sqrt{n}p_{2,n} \to \infty, \frac{p_{2,n}}{p_{1,n}} \to \infty.$$

   *For any $u_n = O(p_{1,n} + \frac{1}{\sqrt{n}})$, $p_{2,n} = O(p_{2,n}(u_n)) < \infty$.*

   *(2.b)*

$$p_{2,n} > 0, \sqrt{n}p_{2,n} \to \infty, \frac{p_{2,n}}{p_{1,n}} \to C,\ \textit{where } 0 < C < \infty, \|\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\|_\infty < C;$$

$$\epsilon_n^{(1)} = \epsilon_n^{(2)}, p_{3,n}\left(\epsilon_n^{(2)}\right) + p_{4,n} \to 0.$$

   *(2.c)*

$$p_{2,n} > 0;\ \textit{for any } u_n = O\left(p_{1,n} + 1/\sqrt{n}\right), \sqrt{n}p_{2,n}(u_n) \to \infty, \frac{p_{2,n}(u_n)}{p_{1,n}} \to \infty.$$

3. *(Asymptotic normality) Assume that the estimator has the model selection consistency stated in 2. If in addition $\sqrt{n}p_{1,n} \to 0$, we have the asymptotic normality for the nonzero part of the estimator $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ as follows:*

$$\sqrt{n}(\mathcal{R}_{11} + \Sigma_1)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathcal{R}_{11} + \Sigma_1)^{-1}\boldsymbol{b}\right\} \to N(0, \mathcal{R}_{11}).$$

*Remark* 12. In this lemma, we intend to include as many penalties as possible, such as LASSO, SCAD, MCP, adaptive LASSO, hard thresholding, bridge, and the LAMP family. Notice that the conditions on smoothness are weaker compared with conditions in the work of Fan and Li.[11] Conditions (2.a), (2.b), and (2.c) are for SCAD(MCP), LASSO, and bridge penalty, respectively.

*Remark* 13. $\hat{\boldsymbol{\beta}}_n$ will not have the mode selection consistency if

$$\overline{\lim_n}\sqrt{n}p_{2,n} < \infty, p_{3,n} + p_{4,n} \to 0.$$

## APPENDIX B: CONDITIONS

The following conditions are needed for the weak oracle property described in Remark 11. Define$G_1$ and $G_2$ and set $\mathcal{U}$ as follows:

$$G_1(\boldsymbol{\theta}) \triangleq (\alpha, \boldsymbol{I}_{\beta_1 \neq 0}, \boldsymbol{I}_{\beta_2 \neq 0}, \dots, \boldsymbol{I}_{\beta_p \neq 0})^T,$$

$$G_2(\boldsymbol{\theta}) \triangleq (\alpha, \boldsymbol{I}_{\beta_1 \neq 0}\beta_1, \boldsymbol{I}_{\beta_2 \neq 0}\beta_2, \dots, \boldsymbol{I}_{\beta_p \neq 0}\beta_p)^T,$$

$$\mathcal{U} \triangleq \{\boldsymbol{\theta} | \max(\|G_2(\boldsymbol{\theta}) - G_2(\boldsymbol{\theta}_0)\|_\infty) < 0.5m_n\}.$$

(B1)

$$m_n \geq 2n^{-\gamma_1}\log(n), \text{where } 0 < \gamma_1 \leq 0.5;$$

$$q_n = O(n^{\gamma_2}), \text{where } 0 < \gamma_2 < 1.$$

(B2)

$$\left\|\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\right\|_\infty = O(n^{\gamma_3}), \text{where } 0 \leq \gamma_3 \leq .5.$$

(B3)

$$\log(p_n) = O(n^{1-2\gamma_4}), \lambda \gg n^{-\gamma_4}\big(\log(n)\big)^2,$$

where $\gamma_4 = \min(0.5, 2\gamma_1 - \gamma_2) - \gamma_3 \geq 0$.

(B4)

$$\lambda_0 = o\left(\min_{\theta\in\mathcal{U}} \lambda_{\min}\left[\frac{1}{n}G_1(\boldsymbol{\theta})^T\mathcal{X}^T\text{diag}\left(\boldsymbol{g}''\left(\boldsymbol{X}G_2(\boldsymbol{\theta})\right)\right)\mathcal{X}G_1(\boldsymbol{\theta})\right]\right).$$

(B5)

$$\max_{\substack{j=1,\dots,p\\ \theta\in\mathcal{U}}}\left\{\lambda_{\max}\left[G_1(\boldsymbol{\theta})^T\mathcal{X}^T\operatorname*{diag}_{i=1,2,\dots,n}\left(\left|X_{ij}g''\left(\boldsymbol{X}_i^TG_2(\boldsymbol{\theta})\right)\right|\right)\mathcal{X}G_1(\boldsymbol{\theta})\right]\right\} = O(n).$$

(B6)

$$\left\|\mathcal{R}_{11}^{-1}\right\|_\infty = O(\gamma_5), \text{where } \gamma_5 = o\left(\min\left[n^{0.5-\gamma_1}\sqrt{\log(n)}, q_n^{-1}n^{\gamma_1}/\log(n)\right]\right);$$

$$\left\|\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\right\|_\infty < \frac{g'(\alpha_1)}{g'\left[(2\lambda)^{-1}\lambda_0 m_n + \alpha_1\right]}.$$

(B7)

$$\lambda g'\left(\frac{\lambda_0}{2\lambda}m_n + \alpha_1\right) = o\left(\gamma_5^{-1}n^{-\gamma_1}\log(n)\right).$$

(B8) If $\boldsymbol{Y} \in [c, d]$ is bounded, for any $\boldsymbol{a} \in \mathbb{R}^n$ and $\epsilon > 0$,

$$\mathrm{P}\left(\left|\boldsymbol{a}^T\boldsymbol{Y} - \boldsymbol{a}^Tg'(\boldsymbol{X}^T\boldsymbol{\theta}_0)\right| > \|\boldsymbol{a}\|_2\epsilon\right) \leq 2e^{-c_1\epsilon^2},$$

where $c_1 = 2(d-c)^{-2}$.

(B9) If $\boldsymbol{Y}$ is unbounded, there exists $v_0, M_0 > 0$,

$$\mathrm{E}\left\{\exp\left[\frac{T(Y) - g'(\boldsymbol{X}^T\boldsymbol{\theta}_0)}{M_0}\right] - 1 - \frac{T(Y) - g'(\boldsymbol{X}^T\boldsymbol{\theta}_0)}{M_0}\right\}M_0^2 \leq \frac{v_0}{2}.$$

(B10) If $\boldsymbol{Y}$ is unbounded,

$$\forall\boldsymbol{a} \in \mathbb{R}^n, 0 < \epsilon \leq \frac{\|\boldsymbol{a}\|_2}{\|\boldsymbol{a}\|_\infty}, \mathrm{P}\left(\left|\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{a}^Tg'(\boldsymbol{X}^T\boldsymbol{\theta}_0)\right| > \|\boldsymbol{a}\|_2\epsilon\right) \leq 2e^{-c_2\epsilon^2},$$

$$\text{where } c_2 = \frac{1}{2v_0 + 2M_0}; \max_j \left\|(X_{1j}, \dots, X_{nj})\right\|_\infty = o\left(n^{\gamma_4}/\sqrt{\log(n)}\right).$$

## APPENDIX C: PROOFS

*Proof of Lemma 1.* For any $\epsilon > 0$, we have $\mathcal{E}_1$ and $\mathcal{E}_2$ with $\|\mathcal{E}_i\|/\sqrt{n} < \epsilon$. Now, select

$$\boldsymbol{u}_{i,n} \in \operatorname*{arglmin}_{\theta} M_n(\mathcal{Z}_n + \mathcal{E}_i, \boldsymbol{\theta}), i = 1, 2.$$

Denote function

$$f_i(\boldsymbol{u}) = m_n(\mathcal{Z}_n + \mathcal{E}_i, \boldsymbol{u}), i = 1, 2.$$

Throughout the proof, $o(1)$ indicates a sequence that approaches 0 as $n \to \infty$ and $\epsilon \to 0$ simultaneously, $o(1, n)$ that approaches 0 as $n \to \infty$, $o(1, \epsilon; n)$ that approaches 0 as $\epsilon \to 0$ with $n$ fixed, and $O_p(1, n)$ bounded by a square integrable function in probability with $n \to \infty$.

First, from Condition (C2),

$$\begin{aligned}\sup_{\boldsymbol{u}\in\Theta}|f_2(\boldsymbol{u}) - f_1(\boldsymbol{u})| &= \sup_{\boldsymbol{u}\in\Theta}\frac{|m_n(\mathcal{Z}_n + \mathcal{E}_1, \boldsymbol{u}) - m_n(\mathcal{Z}_n + \mathcal{E}_2, \boldsymbol{u})|}{\|\mathcal{E}_1 - \mathcal{E}_2\|_d/\sqrt{n}}\|\mathcal{E}_1 - \mathcal{E}_2\|_d/\sqrt{n}\\ &= O_p(1, n)o(1, \epsilon; n).\end{aligned}$$

Similarly, we have

$$\sup_{\boldsymbol{u}\in\Theta} |m_n(\mathcal{Z}_n, \boldsymbol{u}) - f_1(\boldsymbol{u})| = O_p(1)o(1, \epsilon; n).$$

Second, from Condition (C1), $f_i(\boldsymbol{u})$ is convex, $i = 1, 2$. We will prove that

$$\forall 0 \le r \le 1, f_i\left(\boldsymbol{u}'_{2,n}\right) \ge f_i(\boldsymbol{u}_{i,n}) + o_p(1, n), i = 1, 2, \qquad \text{(C1)}$$

where $\boldsymbol{u}'_{2,n} = r\boldsymbol{u}_{1,n} + (1-r)\boldsymbol{u}_{2,n}$. Since $\boldsymbol{u}_{1,n}$ is the local minimum of $f_1(\boldsymbol{u}) + r_n(\boldsymbol{u})$, there exists $0 < \sigma_{1,n} < 1$, such that

$$\begin{aligned} & f_1(\boldsymbol{u}_{1,n}) + r_n(\boldsymbol{u}_{1,n}) - r_n\left(\boldsymbol{u}_{1,n} + \sigma_{1,n}\left(\boldsymbol{u}'_{2,n} - \boldsymbol{u}_{1,n}\right)\right) \\ & \quad \le f_1\left(\boldsymbol{u}_{1,n} + \sigma_{1,n}\left(\boldsymbol{u}'_{2,n} - \boldsymbol{u}_{1,n}\right)\right) \\ & \quad = f_1\left(\sigma_{1,n}\boldsymbol{u}'_{2,n} + (1-\sigma_{1,n})\boldsymbol{u}_{1,n}\right) \\ & \quad \le \sigma_{1,n} f_1\left(\boldsymbol{u}'_{2,n}\right) + (1-\sigma_{1,n}) f_1(\boldsymbol{u}_{1,n}). \end{aligned}$$

After the simplification, we have

$$f_1\left(\boldsymbol{u}'_{2,n}\right) \ge f_1(\boldsymbol{u}_{1,n}) - \frac{r_n\left(\boldsymbol{u}_{1,n} + \sigma_{1,n}\left(\boldsymbol{u}'_{2,n} - \boldsymbol{u}_{1,n}\right)\right) - r_n(\boldsymbol{u}_{1,n})}{\sigma_{1,n}\left\|\boldsymbol{u}'_{2,n} - \boldsymbol{u}_{1,n}\right\|} \left\|\boldsymbol{u}'_{2,n} - \boldsymbol{u}_{1,n}\right\|.$$

Therefore, we have $f_1(\boldsymbol{u}'_{2,n}) \ge f_1(\boldsymbol{u}_{1,n}) + o_p(1, n)$ from Condition (C3).

Third, $\mathrm{E}m_n(\mathcal{Z}_n, \boldsymbol{u}'_{2,n}) - m_n(\mathcal{Z}_n, \boldsymbol{u}'_{2,n}) = o_p(1, n)$ from the law of large number.

Note that

$$\begin{aligned} \bar{m}\left(\boldsymbol{u}'_{2,n}\right) - \bar{m}(\boldsymbol{u}_{1,n}) &= \mathrm{E}m_n\left(\mathcal{Z}_n, \boldsymbol{u}'_{2,n}\right) - \mathrm{E}m_n(\mathcal{Z}_n, \boldsymbol{u}_{1,n}) \\ &= \left[\mathrm{E}m_n\left(\mathcal{Z}_n, \boldsymbol{u}'_{2,n}\right) - m_n\left(\mathcal{Z}_n, \boldsymbol{u}'_{2,n}\right)\right] + \left[m_n\left(\mathcal{Z}_n, \boldsymbol{u}'_{2,n}\right) - f_2\left(\boldsymbol{u}'_{2,n}\right)\right] \\ &\quad - \left[\mathrm{E}m_n(\mathcal{Z}_n, \boldsymbol{u}_{1,n}) - m_n(\mathcal{Z}_n, \boldsymbol{u}_{1,n})\right] - \left[m_n(\mathcal{Z}_n, \boldsymbol{u}_{1,n}) - f_1(\boldsymbol{u}_{1,n})\right] \\ &\quad + \left[f_2\left(\boldsymbol{u}'_{2,n}\right) - f_1\left(\boldsymbol{u}'_{2,n}\right)\right] + \left[f_1\left(\boldsymbol{u}'_{2,n}\right) - f_1(\boldsymbol{u}_{1,n})\right] \\ &\ge o_p(1, n) + o(1, \epsilon; n)O_p(1, n) \triangleq o_p(1, \epsilon, n), \end{aligned}$$

which is the same as

$$\lim_{n\to\infty} \mathrm{P}\left(\lim_{\epsilon\to 0} \bar{m}\left(\boldsymbol{u}'_{2,n}\right) - \bar{m}(\boldsymbol{u}_{1,n}) < -\delta\right) = 0, \forall \delta > 0.$$

Similarly,

$$\lim_{n\to\infty} \mathrm{P}\left(\lim_{\epsilon\to 0} \bar{m}\left(\boldsymbol{u}'_{2,n}\right) - \bar{m}(\boldsymbol{u}_{2,n}) < -\delta\right) = 0, \forall \delta > 0.$$

Combine these 2 together, and we have

$$\lim_{n\to\infty} \mathrm{P}\left(\lim_{\epsilon\to 0} \bar{m}\left(\boldsymbol{u}'_{2,n}\right) - \left[r\bar{m}(\boldsymbol{u}_{1,n}) + (1-r)\bar{m}(\boldsymbol{u}_{2,n})\right] < -\delta\right) = 0, \forall \delta > 0, 0 \le r \le 1.$$

Define

$$\begin{aligned} \Delta(\boldsymbol{u}_1, \boldsymbol{u}_2) &\triangleq \max_{0\le r\le 1} r\bar{m}(\boldsymbol{u}_1) + (1-r)\bar{m}(\boldsymbol{u}_2) - \bar{m}\left(r\boldsymbol{u}_1 + (1-r)\boldsymbol{u}_2\right); \\ C_\delta &= \inf\left\{C \ge 0 | \forall \boldsymbol{u}_1, \boldsymbol{u}_2 \in \Theta \text{ and } \|\boldsymbol{u}_1 - \boldsymbol{u}_2\| \ge C, \Delta(\boldsymbol{u}_1, \boldsymbol{u}_2) > \delta\right\}. \end{aligned}$$

Since $\Theta$ is a compact set, $C_\delta$ exists for any $\delta > 0$; since $\bar{m}$ is strictly convex, $C_\delta \to 0$, as $\delta \to 0$. We then conclude that

$$\lim_{n\to\infty} \mathrm{P}\left(\lim_{\epsilon\to 0} \|\boldsymbol{u}_{1,n} - \boldsymbol{u}_{2,n}\| > C_\delta\right) = 0, \forall C_\delta > 0.$$

*Proof of Lemma 2.* From Condition (C2),

$$\begin{aligned} \sup_{\substack{\lambda\in[0,1] \\ \boldsymbol{\theta}_i\in\Theta, i=1,2}} & \left\{\left[\lambda M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}_1) + (1-\lambda)M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}_2) - M_n\left(\mathcal{Z}_n + \mathcal{E}_n, \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2\right)\right]\right. \\ & \left. - \left[\lambda M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}_1) + (1-\lambda)M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}_2) - M_n\left(\mathcal{Z}_n + \mathcal{E}_n, \lambda\boldsymbol{\theta}_1 + (1-\lambda)\boldsymbol{\theta}_2\right)\right]\right\} \\ = {} & O_p(1, n)o(1, \epsilon; n). \end{aligned}$$

From Condition (C5), there exists $\epsilon_0 > 0$, such that $\forall \|\mathcal{E}_n\| \le \sqrt{n}\epsilon_0$,

$$\lim_{n\to\infty} P\left(M_n(\mathcal{Z}_n + \mathcal{E}_n, \boldsymbol{\theta}) \text{is strictly convex within } o(\boldsymbol{\theta}_0, \delta_0) \cap \Theta\right) = 1.$$

From weak asymptotic stability and consistency in (C4),

$$\overline{\lim_{n\to\infty}} \mathrm{P}\left( \underline{\lim_{\epsilon\to 0}} \operatorname{diam} \bigcup_{\substack{\|\mathcal{E}\|<\sqrt{n}\epsilon \\ \mathcal{Z}_n+\mathcal{E}_n\in S_{\mathcal{Z}_n}}} \left\{ \operatorname{arglmin} M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}) \right\} > \delta_0/2 \right) = 0,$$

$$\overline{\lim_{n\to\infty}} \mathrm{P}\left( \underline{\lim_{\epsilon\to 0}} d\left( \boldsymbol{\theta}_0, \bigcup_{\substack{\|\mathcal{E}\|<\sqrt{n}\epsilon \\ \mathcal{Z}_n+\mathcal{E}_n\in S_{\mathcal{Z}_n}}} \left\{ \operatorname{arglmin} M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}) \right\} \right) > \delta_0/2 \right) = 0,$$

then

$$\underline{\lim_{n\to\infty}} \mathrm{P}\left( \overline{\lim_{\epsilon\to 0}} \bigcup_{\substack{\|\mathcal{E}_n\|<\sqrt{n}\epsilon \\ \mathcal{Z}_n+\mathcal{E}_n\in S_{\mathcal{Z}_n}}} \left\{ \operatorname{arglmin} M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}) \right\} \in o(\boldsymbol{\theta}_0,\delta_0) \right) = 1. \tag{C2}$$

Denote $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ the minimizer of $M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta})$ and $M_n(\mathcal{Z}_n,\boldsymbol{\theta})$ for $\boldsymbol{\theta}\in o(\boldsymbol{\theta}_0,\delta_0)\cap\Theta$, respectively. We have

$$\begin{aligned} 0 &\geq M_n(\mathcal{Z}_n,\boldsymbol{\theta}^*) - M_n(\mathcal{Z}_n,\tilde{\boldsymbol{\theta}}) \\ &= M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}^*) - M_n(\mathcal{Z}_n+\mathcal{E}_n,\tilde{\boldsymbol{\theta}}) + O_p(1)o(1,\epsilon;n) \\ &\geq O_p(1)o(1,\epsilon;n). \end{aligned}$$

Then, the strict convexity of $M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta})$ together with the case

$$\overline{\lim_{\epsilon\to 0}} \bigcup_{\substack{\|\mathcal{E}_n\|<\sqrt{n}\epsilon \\ \mathcal{Z}_n+\mathcal{E}_n\in S_{\mathcal{Z}_n}}} \left\{ \operatorname{arglmin} M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}) \right\} \in o(\boldsymbol{\theta}_0,\delta_0)$$

implies

$$\overline{\lim_{\epsilon\to 0}} \bigcup_{\substack{\|\mathcal{E}_n\|<\sqrt{n}\epsilon \\ \mathcal{Z}_n+\mathcal{E}_n\in S_{\mathcal{Z}_n}}} \left\{ \operatorname{arglmin} M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}) \right\} \in o(\boldsymbol{\theta}_0,\delta_0) = \boldsymbol{\theta}^*,$$

which means that

$$\underline{\lim_{n\to\infty}} \mathrm{P}\left( \overline{\lim_{\epsilon\to 0}} \operatorname{diam} \bigcup_{\substack{\|\mathcal{E}_n\|<\sqrt{n}\epsilon \\ \mathcal{Z}_n+\mathcal{E}_n\in S_{\mathcal{Z}_n}}} \left\{ \operatorname{arglmin} M_n(\mathcal{Z}_n+\mathcal{E}_n,\boldsymbol{\theta}) \right\} = 0 \right) = 1.$$

*Proof of Lemma 3.* First, it is obvious that $\lambda\to 0$ and $\sqrt{n}\lambda\to+\infty$.

$$\begin{aligned} \sqrt{n}\lambda e^{\left(\frac{\lambda_0}{\lambda}m-\alpha_1\right)^u L\left(\frac{\lambda_0}{\lambda}m-\alpha_1\right)} \to 0 &\iff \log\left(\sqrt{n}\lambda\right) + \left(\frac{\lambda_0}{\lambda}m-\alpha_1\right)^u L\left(\frac{\lambda_0}{\lambda}m-\alpha_1\right) \to -\infty \\ &\Longleftarrow \log\left(\sqrt{n}\lambda\right) \ll \left(\frac{\lambda_0}{\lambda}\right)^u L\left(\frac{\lambda_0}{\lambda}\right) \\ &\Longleftarrow \log\left(\sqrt{n}\lambda\right) \ll \left(\frac{\lambda_0}{\lambda}\right)^u \\ &\iff \lambda\left[\log\left(\sqrt{n}\lambda\right)\right]^{1/u} \ll \lambda_0 \\ &\Longleftarrow \lambda(\log n)^{1/u} \ll \lambda_0, \end{aligned}$$

where $m$ is assumed to be fixed.

*Proof of Lemma 4. Part 1 (Parameter estimation consistency).*

It is sufficient to show that for any given $\epsilon>0$, there exists a positive constant $C$, such that

$$\mathrm{P}\left( \sup_{\|\boldsymbol{u}\|=C} \tilde{l}\left(\boldsymbol{\beta}_0 + \boldsymbol{u}\left(n^{-\frac{1}{2}}+p_{1,n}\right)\right) - \tilde{l}(\boldsymbol{\beta}_0) < 0 \right) > 1-\epsilon.$$

Then, there exists a local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\alpha_n)$, where $\alpha_n = n^{-\frac{1}{2}} + p_{1,n}$. Consistency holds as $p_{1,n} \to 0$.

Using Taylor expansion, we have

$$\begin{aligned}
&\tilde{l}\left(\boldsymbol{\beta}_0 + \left(n^{-\frac{1}{2}} + p_{1,n}\right)\boldsymbol{u}\right) - \tilde{l}(\boldsymbol{\beta}_0)\\
&\quad = l\left(\boldsymbol{\beta}_0 + \left(n^{-\frac{1}{2}} + p_{1,n}\right)\boldsymbol{u}\right) - l(\boldsymbol{\beta}_0) - n\sum_{j=1}^{p}\left[p_{\lambda,j}\left(\left|\beta_{j0} + (n^{-\frac{1}{2}} + p_{1,n})u_j\right|\right) - p_{\lambda,j}\left(|\beta_{j0}|\right)\right]\\
&\quad \le l\left(\boldsymbol{\beta}_0 + \left(n^{-\frac{1}{2}} + p_{1,n}\right)\boldsymbol{u}\right) - l(\boldsymbol{\beta}_0) - n\sum_{j=1}^{q}\left[p_{\lambda,j}\left(\left|\beta_{j0} + \left(n^{-\frac{1}{2}} + p_{1,n}\right)u_j\right|\right) - p_{\lambda,j}\left(|\beta_{j0}|\right)\right]\\
&\quad = \left(n^{-\frac{1}{2}} + p_{1,n}\right)\boldsymbol{u}^T l'(\boldsymbol{\beta}_0) - \frac{1}{2}n\boldsymbol{u}^T\mathcal{R}\boldsymbol{u}\left(n^{-\frac{1}{2}} + p_{1,n}\right)^2\left(1 + o_p(1)\right)\\
&\qquad - n\sum_{j=1}^{q}\left[p'_{\lambda,j}\left(|\beta_{j,0}|\right)\operatorname{sgn}(\beta_{j0})\left(n^{-\frac{1}{2}} + p_{1,n}\right)u_j + \frac{1}{2}p''_{\lambda,j}(r_j)\left(n^{-\frac{1}{2}} + p_{1,n}\right)^2 u_j^2\right],
\end{aligned} \tag{C3}$$

where

$$|\beta_{j,0}| \le r_j \le |\beta_{j,0}| + \left(n^{-\frac{1}{2}} + p_{1,n}\right)C, \mathcal{R} = -\mathrm{E}\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T\partial\boldsymbol{\beta}}\bigg|_{\beta=\beta_0} > 0.$$

We now analyze the order of the 4 terms in (C3). For Condition (1.a), the first term is of the order $1 + n^{\frac{1}{2}}p_{1,n}$, whereas the second term is of the order $(1 + n^{\frac{1}{2}}p_{1,n})^2$, which is of the same or higher order compared with the first term. Using Cauchy inequality,

$$n\sum_{j=1}^{q}p'_{\lambda,j}\left(|\beta_{j0}|\right)\operatorname{sgn}(\beta_{j0})\left(n^{-\frac{1}{2}} + p_{1,n}\right)u_j \le n\sqrt{q}p_{1,n}\left(n^{-\frac{1}{2}} + p_{1,n}\right)\|u\|.$$

For fixed $q$, it is also of the same or a smaller order compared with the second term. When $p_{4,n} \to 0$, $p''_{\lambda,j}(\cdot)$ in the fourth term vanishes, it is also controlled by the second term. Regarding the constant involving $\boldsymbol{u}$, the second term contains $\|\boldsymbol{u}\|^2$, whereas both the first and third has $\|\boldsymbol{u}\|$. Thus, the whole expression is controlled by its second term as long as we choose a sufficiently large $C$ since $\mathcal{R}$ is positive definite.

For Condition (1.b), the fourth term is negative and of the same sign as the second term. The conclusion still holds.

*Part 2 (Model selection consistency).* Now, we have an $\alpha_n$-consistent local minimizer $\hat{\boldsymbol{\beta}}_n$. If the model selection consistency does not hold, there exists a $j \in \{q+1, q+2, \dots, p\}$ such that $\hat{\beta}_j \neq 0$, then there is contradiction if we can show that there exists a small enough $\epsilon_n \ll \alpha_n$, and a neighborhood $O(\hat{\beta}_j, \epsilon_n)$, within which the sign of $\frac{\partial\tilde{l}(\beta)}{\partial\beta_j}$ does not change. If $\hat{\beta}_j$ is the optimum solution, the sign of left derivative at the optimal value should be different from the sign of right derivative at the optimal value. Therefore, the nonzero $\hat{\beta}_j$ does not exist. Using Taylor's expansion,

$$\frac{\partial\tilde{l}(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial l(\boldsymbol{\beta}_0)}{\partial\beta_j} + \sum_{l=1}^{p}\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_l\partial\beta_j}(\beta_l - \beta_{l0})\left(1 + o_p(1)\right) - np'_{\lambda,j}\left(|\beta_j|\right)\operatorname{sgn}(\beta_j), \tag{C4}$$

for $j = q+1, q+2, \dots, p$. We see that the third term of (C4) does not depend on $\boldsymbol{\beta}_0$ and is only related to the sign of $\beta_j$, which remains constant in $O(\hat{\beta}_j, \epsilon_n)$ since $\hat{\beta}_j \neq 0$. Therefore, if the first 2 terms are controlled by the third one, we can derive the sparsity using the method above. The coefficient of the sign function in the third term should be positive to control the direction of the derivative. That is, $\inf_{q<j\le p}p'_{\lambda,j}(0) > 0$.

For Condition (2.a), the orders of the 3 term are $\sqrt{n}$, $\sqrt{n} + np_{1,n}$, $p_{2,n}(u_n)$, where $u_n$ is the sequence in Condition (2.a). Condition (2.a) guarantees that the first 2 terms are controlled by the third one. Condition (2.c) will lead to the same reasoning.

For Condition (2.b), we can first prove

$$\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0}{p_{1,n}} \xrightarrow{p} \arg\min_{\boldsymbol{v}}\frac{1}{2}\boldsymbol{v}\mathcal{R}\boldsymbol{v}^T + \sum_{j=1}^{q}\operatorname{sgn}(\beta_{j0})\frac{p'_{\lambda,j}\left(|\beta_{j0}|\right)}{p_{1,n}} \times v_j + \sum_{j=q+1}^{p}\frac{p'_{\lambda,j}(0)}{p_{1,n}} \times |v_j|.$$

(similarly, see the work of Knight and Fu[40]) Applying Karush-Kuhn-Tucker conditions to

$$\frac{\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{2,0}}{p_{1,n}} \xrightarrow{p} 0,$$

which is weaker than sparsity, we get the necessary condition $\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{a} + \boldsymbol{\Lambda}_2\boldsymbol{z} = 0$, where

$$\boldsymbol{\Lambda}_1 = p_{1,n}^{-1}\text{diag}\left\{p'_{\lambda,1}\left(|\beta_{10}|\right), p'_{\lambda,2}\left(|\beta_{20}|\right), \dots, p'_{\lambda,q}\left(|\beta_{q0}|\right)\right\},$$
$$\boldsymbol{\Lambda}_2 = p_{1,n}^{-1}\text{diag}\left\{p'_{\lambda,q+1}(0), p'_{\lambda,q+2}(0), \dots, p'_{\lambda,p}(0)\right\},$$

$\boldsymbol{a} = \text{sgn}(\boldsymbol{\beta}_{1,0})$ and $|\boldsymbol{z}| \leq 1$, which is equivalent to $\boldsymbol{\Lambda}_2^{-1}\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{a} \leq \mathbf{1}$. We will show that a sufficient condition is

$$\left\|\boldsymbol{\Lambda}_2^{-1}\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\right\|_\infty < 1,$$

or equivalently,

$$\left\|\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\right\|_\infty < \left(\left\|\boldsymbol{\Lambda}_2^{-1}\right\|_\infty \|\boldsymbol{\Lambda}_1\|_\infty\right)^{-1} \to C.$$

Then, it will be obvious that the first 2 terms are controlled by the third one. Let

$$\boldsymbol{v}_n = \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0}{p_{1,n}} = \arg\max_{\boldsymbol{v}} \tilde{l}(\boldsymbol{\beta}_0 + p_{1,n}\boldsymbol{v}).$$

Then, from $p_{3,n}(\epsilon_n^{(2)}) \to 0$, we have $\sup_{q<j\leq p}|p'_{\lambda,j}(0)| < \infty$ and

$$\forall u_n \to 0, p_{2,n}\left(|u_n|\right) = p_{2,n} + o(1). \tag{C5}$$

Together with $p_{4,n} \to 0$,

$$\begin{aligned}\tilde{l}(\boldsymbol{\beta}_0 + p_{1,n}\boldsymbol{v}) &= \tilde{l}(\boldsymbol{\beta}_0) + p_{1,n}\boldsymbol{v}^T\frac{\partial l(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} + \frac{1}{2}p_{1,n}^2\boldsymbol{v}^T\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T\partial \boldsymbol{\beta}}\boldsymbol{v}\left(1 + o_p(1)\right)\\ &\quad - np_{1,n}^2\left(1 + o(1)\right)\sum_{j=1}^{p}\left[\text{sgn}(\boldsymbol{\beta}_{j0})v_j\frac{p'_{\lambda,j}\left(|\beta_{j0}|\right)}{p_{1,n}}I_{\beta_{j0}\neq 0} + \frac{p'_{\lambda,j}\left(|\beta_{j0}|\right)}{p_{1,n}}|v_j|I_{\beta_{j0}=0}\right]\\ &= \tilde{l}(\boldsymbol{\beta}_0) + np_{1,n}^2 \times \left\{-\frac{1}{2}\boldsymbol{v}\mathcal{R}\boldsymbol{v}^T - \sum_{j=1}^{q}\text{sgn}(\beta_{j0})\frac{p'_{\lambda,j}\left(|\beta_{j0}|\right)}{p_{1,n}} \times v_j - \sum_{j=q+1}^{p}\frac{p'_{\lambda,j}(0)}{p_{1,n}} \times |v_j|\right\}\left(1 + o_p(1)\right).\end{aligned}$$

Following the work of Geyer,[41] we get

$$\boldsymbol{v}_n \xrightarrow{p} \arg\min_{\boldsymbol{v}}\left\{\frac{1}{2}\boldsymbol{v}\mathcal{R}\boldsymbol{v}^T + \sum_{j=1}^{q}\text{sgn}(\beta_{j0})\frac{p'_{\lambda,j}\left(|\beta_{j0}|\right)}{p_{1,n}} \times v_j + \sum_{j=q+1}^{p}\frac{p'_{\lambda,j}(0)}{p_{1,n}} \times |v_j|\right\}.$$

Let $\boldsymbol{v}_n \triangleq (\boldsymbol{v}_{1,n}^T, \boldsymbol{v}_{2,n}^T)^T$, where $\boldsymbol{v}_{1,n}^T$ is a $q \times 1$ vector. We see that $\boldsymbol{v}_{2,n} \to \boldsymbol{v}_2 = 0$ in probability is a necessary condition for sparsity. From conclusions above,

$$\left(\boldsymbol{v}_{1,n}^T, \boldsymbol{v}_{2,n}^T\right) \to \arg\min_{\boldsymbol{v}_1,\boldsymbol{v}_2}\frac{1}{2}\left(\boldsymbol{v}_1^T\mathcal{R}_{11}\boldsymbol{v}_1 + 2\boldsymbol{v}_2^T\mathcal{R}_{21}\boldsymbol{v}_1 + \boldsymbol{v}_2^T\mathcal{R}_{22}\boldsymbol{v}_2\right) + \boldsymbol{v}_1^T\boldsymbol{\Lambda}_1\boldsymbol{a} + \left|\boldsymbol{v}_2^T\right|\boldsymbol{\Lambda}_2\boldsymbol{1}. \tag{C6}$$

Karush-Kuhn-Tucker conditions lead to the following equations:

$$\mathcal{R}_{11}\boldsymbol{v}_1 + \mathcal{R}_{12}\boldsymbol{v}_2 + \boldsymbol{\Lambda}_1\boldsymbol{a} = 0,$$
$$\mathcal{R}_{21}\boldsymbol{v}_1 + \mathcal{R}_{22}\boldsymbol{v}_2 + \boldsymbol{\Lambda}_2\boldsymbol{z} = 0,$$

where $\boldsymbol{z} = (z_1, \dots, z_{p-q})^T$ and

$$z_j \in \left\{z|\text{if } v_{q+j} \neq 0, z = \text{sgn}(v_{q+j});\ \text{ else } |z| \leq 1\right\}.$$

Then,

$$\begin{aligned}\boldsymbol{v}_2 = 0 &\Longleftrightarrow \mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{a} + \boldsymbol{\Lambda}_2\boldsymbol{z} = 0, |\boldsymbol{z}| \leq \mathbf{1}\\ &\Longleftrightarrow \left|\boldsymbol{\Lambda}_2^{-1}\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{a}\right| \leq \mathbf{1},\end{aligned}$$

which is a necessary condition for sparsity. Now, let

$$\boldsymbol{\Lambda}_2(\boldsymbol{v}_2) \triangleq p_{1,n}^{-1}\text{diag}\left\{p'_{\lambda,q+1}(v_{q+1}), p'_{\lambda,q+2}(v_{q+2}), \dots, p'_{\lambda,p}(v_p)\right\}.$$

Then,

$$\sqrt{n}\lambda \to \infty, \frac{\partial \tilde{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2} = np_{1,n}\left[(\mathcal{R}_{21}, \mathcal{R}_{22})\left(\boldsymbol{v}_1^T, \boldsymbol{v}_2^T\right)^T\left(1 + o_p(1)\right) - \boldsymbol{\Lambda}_2(\boldsymbol{v}_2)\text{sgn}(\boldsymbol{\beta_2})\right].$$

From (C5), $\boldsymbol{\Lambda}_2(\boldsymbol{v}_2) = \boldsymbol{\Lambda}_2 + o(1)$. Therefore, a sufficient condition for sparsity is that $(\mathcal{R}_{21}, \mathcal{R}_{22})(\boldsymbol{v}_1^T, \boldsymbol{v}_2^T)^T$ is controlled by $\boldsymbol{\Lambda}_2$ coordinate wisely. That is,

$$\begin{aligned} & \mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{a} + \boldsymbol{\Lambda}_2\boldsymbol{z} = 0, |\boldsymbol{z}| < 1 \\ \Longleftarrow\ & \left|\boldsymbol{\Lambda}_2^{-1}\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{a}\right| < \mathbf{1} \\ \Longleftarrow\ & \left\|\mathcal{R}_{21}\mathcal{R}_{11}^{-1}\right\|_\infty < \left(\left\|\boldsymbol{\Lambda}_2^{-1}\right\|_\infty \|\boldsymbol{\Lambda}_1\|_\infty\right)^{-1} \to C. \end{aligned}$$

*Part 3 (Asymptotic normality and oracle property).* Suppose $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, 0)^T$ is the maximizer. Using parameter estimation consistency and model selection consistency property, for $j = 1, 2, \dots, q$,

$$\begin{aligned} 0 = \frac{\partial \tilde{l}(\hat{\boldsymbol{\beta}})}{\partial \beta_j} &= \frac{\partial l(\hat{\boldsymbol{\beta}})}{\partial \beta_j} - np'_{\lambda,j}\left(|\hat{\beta}_j|\right)\operatorname{sgn}\left(|\hat{\beta}_j|\right) \\ &= \frac{\partial l(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{i=1}^{q}\left(\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_i}\right)(\hat{\beta}_i - \beta_{i0})\left(1 + o_p(1)\right) \\ &\quad - np'_{\lambda,j}\left(|\beta_{j0}|\right)\operatorname{sgn}(\beta_{j0}) - np''_{\lambda,j}\left(|\beta_j|\right)(\hat{\beta}_j - \beta_{j0})\left(1 + o_p(1)\right). \end{aligned}$$

Thus,

$$\sqrt{n}(\mathcal{R}_{11} + \Sigma_1)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathcal{R}_{11} + \Sigma_1)^{-1}\boldsymbol{b}\right\} \to N(0, \mathcal{R}_{11}).$$

If the order of $\boldsymbol{b}$ is controlled by $1/\sqrt{n}$, the oracle property holds.

*Proof of Theorem 1.* Using the notations of Lemma 4 and the form of the LAMP penalty, $p_{1,n} = \lambda g'(\alpha_1 - \frac{\lambda_0}{\lambda}m)/g'(\alpha_1)$; $p_{2,n} = \lambda$; $p_{2,n}(u_n) = \lambda g'(\alpha_1 - \frac{\lambda_0}{\lambda}u_n)/g'(\alpha_1)$; $p_{3,n} = \lambda_0 g''(\alpha_1)/|g'(\alpha_1)|$; $p_{4,n} = \lambda_0 g''(\alpha_1 - \frac{\lambda_0}{\lambda}m)/|g'(\alpha_1)|$.

From (C6),

$$\mathrm{E}\left|Y\boldsymbol{X}^T\boldsymbol{\theta} - g(\boldsymbol{X}^T\boldsymbol{\theta})\right| < \infty, \forall \boldsymbol{\theta} \in \Theta,$$

there exists $\bar{m}(\boldsymbol{\theta}) = -\mathrm{E}l(\boldsymbol{\theta})$ such that $m_n(\boldsymbol{\theta}) = -l_n(\boldsymbol{\theta}) \to \bar{m}(\boldsymbol{\theta})$ almost surely. From

$$\lambda_{\min}\left(\mathrm{E}\boldsymbol{X}g''(\boldsymbol{X}^T\boldsymbol{\theta})\boldsymbol{X}^T\right) > 0, \forall \boldsymbol{\theta} \in \Theta,$$

$-El(\boldsymbol{\theta})$ is a strictly convex function of $\boldsymbol{\theta}$. Condition (C1) holds.

From smoothness of the function $g$ and compactness of $\Theta$, Condition (C2) holds.

From (C7), $\lambda_0/\lambda \to \infty$, together with $\lim_{\xi\to\infty} g'(\xi) = 0$, $\alpha_1$ is a constant and $\lambda \to 0$, we have

$$\sup_{\theta\in\Theta} p'_\lambda(\theta) = \sup_{\theta\in\Theta} \lambda \frac{g'(\alpha_1 - \lambda_0/\lambda\theta)}{g'(\alpha_1)} \to 0.$$

Condition (C3) holds, and $p_{1,n} \to 0, p_{4,n} \to 0$.

From (C6)

$$E\left[\|\boldsymbol{X}\|_2^2 g''(\boldsymbol{X}^T\boldsymbol{\theta}_0) + \|\boldsymbol{X}\|_1^3 \sup_{\|\theta-\theta_0\|\le\delta} g'''(\boldsymbol{X}^T\boldsymbol{\theta})\right] < \infty,$$

regularity conditions in Lemma 4 holds.

From $p_{1,n} \to 0, p_{4,n} \to 0$, Condition (1.a) in Lemma 4 holds. Condition (C4) holds.

Given $u_n = O(p_{1,n}, 1/\sqrt{n})$ from $\sqrt{n}p_{1,n} \to 0$, we get $u_n = O(1/\sqrt{n})$. From $\sqrt{n}\lambda \to \infty$, we get $u_n/\lambda \to 0$. Together with the smoothness of the function $g$, $p_{2,n}(u_n)/p_{2,n} \to 1$. From Condition (C7), it is obvious that

$$p_{2,n} > 0, \frac{p_{2,n}}{p_{1,n}} \to \infty, \sqrt{n}p_{2,n} \to \infty.$$

Condition (2.a) and conditions for asymptotic normality in Lemma 4 hold.

In conclusion, with Conditions (C6)-(C8), the penalized maximum-likelihood estimator based on the LAMP family is consistent and asymptotically normal and achieves model selection consistency and strong asymptotic stability. □

*Proof of Theorem 2.* The idea of the proof is adapted from the works of Munk et al[33] and Zou and Li.[28] For convenience, we define the following notations:

$$
\begin{aligned}
M_n(\boldsymbol{\theta}) &\triangleq M_{(1)}(\boldsymbol{\theta}) + M_{(2)}(\boldsymbol{\theta}),\\
M_{(1)}(\boldsymbol{\theta}) &\triangleq -\frac{1}{n}l(\boldsymbol{\theta}), M_{(2)}(\boldsymbol{\theta}) \triangleq \sum_{j=1}^{p} p_\lambda\left(|\theta_j|\right),\\
\boldsymbol{\theta}^{(\boldsymbol{m})} &\triangleq \left(\theta_0^{(m_0)}, \theta_1^{(m_1)}, \dots, \theta_p^{(m_p)}\right)^T,\\
\boldsymbol{\theta}_{(-j)}^{(\boldsymbol{m})}(x) &\triangleq \left(\theta_0^{(m_0)}, \theta_1^{(m_1)}, \dots, \theta_{j-1}^{(m_{j-1})}, x, \theta_{j+1}^{(m_{j+1})}, \theta_p^{(m_p)}\right)^T,\\
Q\left(x, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right) &\triangleq a\left(\boldsymbol{\theta}_{(-j)}^{(\boldsymbol{m})}(x), \boldsymbol{\theta}^{(\boldsymbol{m})}\right),\\
\phi\left(x, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right) &\triangleq \sum_{1\le k\le p} p_\lambda\left(\left|\theta_k^{(m_k)}\right|\right) + p'_\lambda\left(\left|\theta_j^{(m_j)}\right|\right)\left(|x| - \left|\theta_j^{(m_j)}\right|\right),\\
R\left(x, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right) &\triangleq Q\left(x, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right) + \phi\left(x, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right),
\end{aligned}
$$

where $m_0, m_1, \dots, m_p$ are the iteration times of $\theta_0, \theta_1, \dots, \theta_p$ respectively and $\boldsymbol{m} = (m_0, m_1, \dots, m_p)$. From the concavity of the penalty on the positive part, we have

$$M_{(2)}\left(\boldsymbol{\theta}^{(\boldsymbol{m}+\boldsymbol{e}_j)}\right) \le \phi\left(\theta_j^{(m_j+1)}, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right),$$

where $\boldsymbol{e}_j$ is a length-$(p+1)$ vector with the $j$th element 1 and all the others 0; from conditions of the theorem, we see

$$M_{(1)}\left(\boldsymbol{\theta}^{(\boldsymbol{m}+\boldsymbol{e}_j)}\right) \le Q\left(\theta_j^{(m_j)}, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right),$$

and

$$M\left(\boldsymbol{\theta}^{(\boldsymbol{m}+\boldsymbol{e}_j)}\right) \le R\left(\theta_j^{(m_j+1)}, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right).$$

In the algorithm,

$$\theta_j^{(m_j+1)} = \text{argmin}_\theta R\left(\theta, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right).$$

Then,

$$
\begin{aligned}
M\left(\boldsymbol{\theta}^{(\boldsymbol{m})}\right) &= R\left(\theta_j^{(m_j)}, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right)\\
&> R\left(\theta_j^{(m_j+1)}, j, \boldsymbol{\theta}^{(\boldsymbol{m})}\right)\\
&\ge M\left(\boldsymbol{\theta}^{(\boldsymbol{m}+\boldsymbol{e}_j)}\right).
\end{aligned}
$$

Since function $M$ decreases as iteration continues and has a lower bound, it converges.

By the monotonicity of $M(\boldsymbol{\theta}^{(t)})$ ($t$ represents the number of iterations), all points $\boldsymbol{\theta}^{(\boldsymbol{m})}$ are in a compact set

$$\left\{\boldsymbol{\theta} \in C | M(\boldsymbol{\theta}) \le M\left(\boldsymbol{\theta}^{(\boldsymbol{0}_{p+1})}\right)\right\}.$$

It is compact because function $M(\boldsymbol{\theta})$ is continuous and coercive. Then, there exists a convergent subsequence $\boldsymbol{\theta}^{(t_l)}$, $\boldsymbol{\theta}^* \in C, j_0 \in \{0, 1, \dots, p\}$ such that

$$\lim_{l\to\infty} \boldsymbol{\theta}^{(t_l)} = \boldsymbol{\theta}^*; j^{(t_l+1)} \equiv j_0.$$

Next, denote $m_l^* \triangleq m_{j_0}^{(t_l+1)}$. For any $v \in C$, we have

$$M\left(\boldsymbol{\theta}^{(t_{l+1})}\right) \le M\left(\boldsymbol{\theta}^{(t_l+1)}\right) \le R\left(\theta_{j_0}^{(m_l^*)}, j_0 | \boldsymbol{\theta}^{(t_l)}\right) \le R\left(v_{j_0}, j_0 | \boldsymbol{\theta}^{(t_l)}\right).$$

Assume $M(\boldsymbol{\theta}^{t_{l+1}}) \to M(\boldsymbol{\theta}^*) = R(\boldsymbol{\theta}^*, j | \boldsymbol{\theta}^*), j = 0, \dots, p$. Taking limit $l \to \infty$ on both sides of the above equation, we have

$$M(\boldsymbol{\theta}^*) \le \lim_{l\to\infty} R\left(v_{j_0}, j_0 | \boldsymbol{\theta}^{(t_l)}\right) = R(v_{j_0}, j_0 | \boldsymbol{\theta}^*).$$

Thus, the subgradient of $R(\cdot, j_0 | \boldsymbol{\theta}^*)$ at $\boldsymbol{\theta}^*$ is 0, which is exactly the derivative of $M(\boldsymbol{\theta})$ with respect to $\theta_{j_0}$ at $\boldsymbol{\theta}^*$, because it can be easily verified that the smooth approximation keeps the first-order derivative the same.

From the algorithm and the definition of the "viol" function, we have

$$\forall j \in \{0, \dots, p\}, \left|\dot{R}\left(\cdot, j | \boldsymbol{\theta}^{(t_l)}\right)\right| \le \left|\dot{R}\left(\cdot, j_0 | \boldsymbol{\theta}^{(t_l)}\right)\right|.$$

Taking the derivative of both sides of the above equation, the subgradient of $R(\cdot, j|\boldsymbol{\theta}^*)$ at $\boldsymbol{\theta}^*$ is 0, which is exactly the same as the partial derivative of $M(\boldsymbol{\theta})$ with respect to $\theta_j$ at $\boldsymbol{\theta}^*, \forall j \in \{0, \dots, p\}$. From the strict convexity, $\boldsymbol{\theta}^*$ is the unique local minimum of $M(\boldsymbol{\theta})$.

Method 1 for logistic regression uses lemma 1 in the work of Munk et al.[33] That is,

$$\log(1+e^{z}) \leq \frac{z}{2} + \log\left(e^{-z_0/2} + e^{z_0/2}\right) + \frac{1}{2}\frac{\tanh(.5z_0)}{z_0}\left(z^2 - z_0^2\right).$$

□