Ann Inst Stat Math (2018) 70:191–214
https://doi.org/10.1007/s10463-016-0588-3

Variable selection and estimation using a continuous approximation to the L_0 penalty

Yanxin Wang[1,2] · **Qibin Fan**[3] · **Li Zhu**[4]

Received: 11 May 2015 / Revised: 18 September 2016 / Published online: 19 October 2016

Abstract Variable selection problems are typically addressed under the regularization framework. In this paper, an exponential type penalty which very closely resembles the L_0 penalty is proposed, we called it EXP penalty. The EXP penalized least squares procedure is shown to consistently select the correct model and is asymptotically normal, provided the number of variables grows slower than the number of observations. EXP is efficiently implemented using a coordinate descent algorithm. Furthermore, we propose a modified BIC tuning parameter selection method for EXP and show that it consistently identifies the correct model, while allowing the number of variables to diverge. Simulation results and data example show that the EXP procedure performs very well in a variety of settings.

Keywords Penalized least squares · Coordinate descent algorithm · Variable selection · MBIC · Oracle property

This work was supported by the K. C. Wong Education Foundation, Hong Kong, Program of National Natural Science Foundation of China (No. 61179039) and the Project of Education of Zhejiang Province (No. Y201533324). The authors gratefully acknowledges the support of K. C. Wong Education Foundation, Hong Kong. And the authors are grateful to the editor, the associate editor and the anonymous referees for their constructive and helpful comments.

✉ Yanxin Wang
wyxinbj@163.com

[1] School of Science, Ningbo University of Technology, Ningbo 315211, China

[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[3] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

[4] School of Applied Mathematics, Xiamen University of Technology, Xiamen 361024, China

1 Introduction

Variable selection is an important aspect of high dimensional statistical modelling, particularly in regression and classification. In the regularization framework, various penalty functions are used to perform variable selection by putting relatively large penalties on small coefficients. The best subset selection, namely, the L_0 penalty, along with the traditional model selection criteria such as AIC, BIC and RIC (Akaike 1973; Schwarz 1978; Foster and George 1994) is attractive for variable selection since it directly penalizes the number of nonzero coefficients. However, one drawback of L_0 penalized least squares (PLS) procedure is instability of the resulting estimators (Breiman 1996). This results from the fact that the L_0 penalty is not continuous at 0. Another perhaps more significant drawback of the L_0 penalty is that implementing L_0 PLS procedures is NP-hard and may involve an exhaustive search over all possible models. Thus, implementing these procedures is computationally infeasible when the number of potential predictors is even moderately large, let along the high dimensional data.

Such computational difficulty motivated various continuous relaxations. For example, the bridge regression (Frank and Friedman 1993) uses the L_q penalty, $0 < q \leq 2$. In particular, the use of the L_2 penalty is called the ridge regression. The nonnegative garrote was introduced by Breiman (1995) for variable selection and shrinkage estimation. The L_1 PLS method was termed LASSO by Tibshirani (1996), which is also collectively referred to as the L_1 penalization methods in other contexts. However, LASSO may not consistently select the correct model and is not necessarily asymptotically normal (Knight and Fu 2000; Zou 2006). Other commonly used penalty functions include the SCAD (Fan and Li 2001), MCP (Zhang 2010), SICA (Lv and Fan 2009), Elastic net (Zou and Hastie 2005), SELO (Dicker et al. 2013) and so on.

In this paper, we propose a exponential type penalty function named EXP penalty which very closely approximates the L_0 penalty. EXP penalty is continuous, so the EXP estimators may be more stable than those obtained through L_0 penalized methods. Furthermore, the EXP penalty is a smooth function on $[0, \infty)$ and we use a coordinate descent (CD) algorithm (Friedman et al. 2010; Fan and Lv 2011). Also, we formally establish the model selection oracle property enjoyed by EXP estimators. In particular, the asymptotic normality of the EXP is formally established. Our asymptotic framework allows the number of predictors $p \to \infty$, along with the number of observations n, provided $p/n \to 0$.

The practical performance of PLS procedures depends heavily on the choice of a tuning parameter. In our study, we propose a modified BIC (MBIC) tuning parameter selector that accounts for lost degrees of freedom and performs very well when used in conjunction with EXP penalty. Furthermore, we also prove that this EXP/MBIC procedure consistently identifies the correct model, if $p/n \to 0$ and other regularity conditions are met.

This paper is organized in the following way: In Sect. 2, we introduce PLS estimators and give a brief overview of existing nonconvex penalty terms, and a new nonconvex penalty function is presented. Then, we discuss some of its theoretical properties in Sect. 3. In Sect. 4, we propose an MBIC tuning parameter selection method for EXP. In Sect. 5, we describe a simple and efficient algorithm for obtaining the EXP estimator.

Simulation studies and an application of the proposed methodology are presented in Sect. 6. The proofs are relegated to the Appendix.

2 Variable selection and estimation with EXP penalty

2.1 Linear models and penalized least squares

We start with the linear regression model

$$\mathbf{y} = \mathbf{X}\beta^* + \varepsilon, \tag{1}$$

where $\mathbf{X} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ is an $n \times p$ design matrix, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ is an n-dimensional response vector, and ε are the iid random errors with mean 0 and variance σ^2 n-dimensional noise vector, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}}$ are the regression parameter.

The problem of interest involves estimating a vector of coefficients β defined by minimizing an objective function $Q(\beta)$ composed of a loss function combined with a penalty that encourages sparsity and prevents overfitting:

$$Q(\beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2}$$

where $\|\cdot\|^2$ denotes the L_2 norm, and $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$. By regularizing the conventional least-squares estimation, we hope to simultaneously select important variables and estimate their regression coefficients with sparse estimates.

Various penalty functions have been used in the variable selection literature for linear regression models. Tibshirani (1996) proposed the LASSO. However, LASSO estimates may be biased and inconsistent for model selection (Fan and Li 2001; Zou 2006). This implies that the LASSO does not have the oracle property. The adaptive LASSO is a weighted version of LASSO which has the oracle property (Zou 2006). Other commonly used penalty functions include the SCAD and MCP. The SCAD penalty is the continuous function, whose derivative is given by

$$p'_\lambda(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}, \tag{3}$$

where $a > 2$. Fan and Li (2001) recommended taking $a = 3.7$ and we follow this recommendation throughout. A closely related minimax concave penalty (MCP) was proposed by Zhang (2010), the MCP is the continuous function defined by

$$p'_\lambda(t) = \frac{(a\lambda - t)_+}{a}, \tag{4}$$

where the parameter $a > 0$ determines the concavity of $p_\lambda(\cdot)$. Zhang (2010) proved that the MCP procedure may select the correct model with probability tending to 1 and that MCP estimators have good properties in terms of L_q-loss, provided λ and a satisfy certain conditions. Zhang's results in fact allow for $p \gg n$.

2.2 EXP penalty

The EXP penalty is defined by

$$p_{\lambda,a}(|\theta|) = \lambda \left\{1 - \mathrm{e}^{-\frac{|\theta|}{a}}\right\}, \tag{5}$$

The EXP penalty has a tuning parameter $a > 0$, in addition to λ, and when a is small, $p_\lambda(|\theta|) \approx \lambda I(\theta \neq 0)$. In fact, we have

$$\lim_{a\to 0} p_{\lambda,a}(|\theta|) = \begin{cases} \lambda, & \text{if } |\theta| \neq 0, \\ 0, & \text{if } |\theta| = 0. \end{cases}$$

It is clearly that the EXP penalty is a continuous approximation to the L_0 penalty. Since the EXP penalty is continuous, the associated PLS procedure is more stable than L_0 procedure. The EXP penalty function ($a = 0.5$) is plotted in Fig. 1, along with the SCAD, LASSO, MCP and L_0 penalties. Notice that the EXP penalty mimics the L_0 penalty much more closely than the L_1, MCP or SCAD penalties.

Also, we can see MCP and EXP penalty begin by applying the same rate of penalization as the LASSO, but continuously relax that penalization. MCP relaxes the rate of penalization linearly, and thus results in $p'(|\theta|) = 0$ for all $|\theta| > a\lambda$, where a is an additional tuning parameter of MCP playing a role similar to a in (5). The EXP penalty, on the other hand, allows the penalty to decay exponentially, approaching $p'_{\lambda,a}(|\theta|) = 0$ asymptotically but never reaching it. The diminishing rate of penalization is an attractive property; as discussed in Fan and Li (2001), it leads to the estimator $\hat{\theta}$ being nearly unbiased given a large enough sample size.

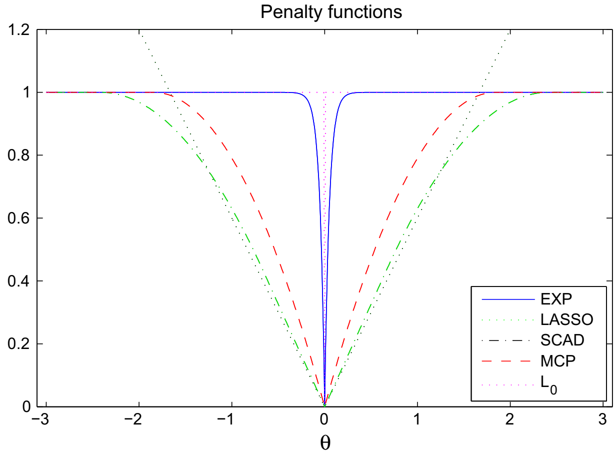

Fig. 1 Plots of the penalties (including L_0, LASSO, SCAD, MCP and EXP)

It is worth mentioning that Douglas (2011) also proposed a continuous function which very closely resembles the L_0 penalty. The penalty function is of the form:

$$p_\lambda(|\theta|) = \lambda \left\{ 1 - \mathrm{e}^{-\frac{\theta^2}{a^2}} \right\}.$$

Similar to the EXP penalty proposed by us, the penalty function is continuous and can have arbitrarily steep slope in a neighborhood near zero, thus mimicking the L_0 penalty. However the function is not singular at zero, θ will have no zero-valued components, although some will have been shrunk arbitrarily close thereto. The EXP function we proposed is singular at zero and satisfies the condition of sparsity (Fan and Li 2001). Also, we noticed that the EXP penalty is similar to that proposed by Breheny (2015)

$$p_{\lambda,a}(|\theta|) = \frac{\lambda^2}{a} \left\{ 1 - \mathrm{e}^{-\frac{a|\theta|}{\lambda}} \right\}. \tag{6}$$

Despite the penalties in Eqs. (5) and (6) have the different forms, but they have similar properties. The two penalties allow the penalty to decay exponentially, approaching $p'_{\lambda,a}(|\theta|) = 0$ asymptotically but never reaching it. The EXP penalty in (5) is infinitely differentiable with a smooth derivative that decays exponentially at a rate regulated by the parameter a, while the exponential penalty in (6) is controlled by the parameter a and λ. For example, the EXP penalty in (5) approaches the L_0 penalty as $a \to 0$, the parameter a controlled the degree of approximation, while the the exponential penalty in (6) approaches the L_0 penalty as $\frac{\lambda}{a} \to 0$. Furthermore, Breheny (2015) only focus on using the exponential penalty for grouped regularization, and detailed study of the properties of the estimator for the ungrouped case is not be done.

Like the exponential penalty in (6), the EXP penalty in (5) is not convex, neither the coordinate descent algorithms nor the LLA algorithm are guaranteed to converge to a global minimum in general. However, it is possible for the objective function to be convex with respect to β even though it contains a nonconvex penalty component. These conditions are presented in the following proposition. Let

$$Q_n(\beta) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_{\lambda,a}(|\beta_j|), \tag{7}$$

be the objective function, and $p_{\lambda,a}(|\beta_j|)$ is the EXP penalty function.

Proposition 1 *Let* $\lambda_{\min}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X})$ *denote the minimum eigenvalue of* $\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$*, Then objective function* (7) *is strictly convex if* $\lambda_{\min}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}) > \frac{\lambda}{a^2}$.

Remark 1 Proposition 1 indicates that if $Q_n(\beta_j)$ denote (7) considered as a function only of β_j, with all other coefficients fixed. Then $Q_n(\beta_j)$ is strictly convex if $\lambda < a^2$. However, the conclusion is different with Breheny (2015) which required $a < 1$. In fact, the $\frac{\lambda}{a^2}$ in EXP penalty (5) is equal to a in the exponential penalty (6). The proof is similar, so we omit it here.

Remark 2 Since the penalty function is separable and the objective function is convex in each coordinate dimension, we may apply a coordinate descent approach to solve for β and the approach is guaranteed to converge to the minimum.

Remark 3 In high dimensions ($p > n$), the minimum eigenvalue $\lambda_{\min}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X})$ will be zero, so the strict global convexity is not possible. However, local convexity may still apply.

3 Theoretical properties of the EXP estimator

In this section we study the theoretical properties of the EXP estimator proposed in Sect. 2 in the situation where the number of parameters p tends to ∞ with increasing sample size n. When discussing variable selection, it is convenient to have concise notation. Denote the columns of $\mathbf{X}$ by $\mathbf{x}_1, \ldots, \mathbf{x}_p \in R^n$ and the rows of $\mathbf{X}$ by $x_1, \ldots, x_n \in R^p$. Let $A = \{j; \beta_j^* \neq 0\}$ be the true model and suppose that p_0 is the size of the true model. That is, suppose that $|A| = p_0$, where $|A|$ denotes the cardinality of A. In addition, for $S \subseteq \{1, 2, \ldots, p\}$, let $\beta_S = (\beta_j)_{j \in S}$ be the $|S|$-dimensional sub-vector of β containing entries indexed by S and let $\mathbf{X}_S$ be the $n \times |S|$ matrix obtained from $\mathbf{X}$ by extracting columns corresponding to S. Given a $p \times p$ matrix C and subsets $S_1, S_2 \subseteq \{1, 2, \ldots, p\}$, let C_{S_1,S_2} be the $|S_1| \times |S_2|$ sub-matrix of C with rows determined by S_1 and columns determined by S_2.

3.1 Regularity conditions

We need to place the following conditions:

(A) $n \to \infty$ and $p\sigma^2/n \to 0$.
(B) $\rho\sqrt{n/(p\sigma^2)} \to \infty$, where $\rho = \min_{j \in A} |\beta_j^*|$.
(C) $\lambda = O(n^{-\gamma})$, $\lambda\sqrt{n/(p\sigma^2)} \to \infty$, $a = O(p^{1/2}\sigma n^{-\gamma/2})$ for some positive constant $0 < \gamma < 1$.
(D) There exist constants $C_1, C_2 \in R$ such that $C_1 < \lambda_{\min}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}) < \lambda_{\max}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}) < C_2$,
where $\lambda_{\min}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X})$ and $\lambda_{\max}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X})$ are the smallest and largest eigenvalues of $\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$, respectively.
(E) $\lim_{n\to\infty} n^{-1} \max_{1 \le i \le n} \sum_{j=1}^{p} x_{ij}^2 = 0$.
(F) $E(|\varepsilon_i/\sigma|^{2+\delta}) < M$ for some δ and $M < \infty$.

Since p may vary with n, it is implicit that β^* may vary with n. Additionally, we allow the model A and the distribution of ε (in particular, σ^2) to change with n. Condition (A) limits how p and σ^2 may grow with n. This condition is the same as that required in Dicker et al. (2013) and substantially weaker than that required in Fan and Peng (2004), who require $p^5/n \to 0$, and slightly weaker than that required in Zou and Zhang (2009), who require $\log(p)/\log(n) \to \nu \in [0, 1)$. As mentioned in Sect. 1, other authors have studied PLS methods in settings where $p > n$, i.e. their growth condition on p is weaker than Condition (A). However, if Condition

(A) is relaxed, additional stronger conditions are typically required in order to obtain desirable theoretical properties. For instance, Kim et al. (2008) require a stronger moment condition on ε_i and an additional condition on p_0. Fan and Lv (2011) require stronger conditions on $n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Condition (B) gives a lower bound on the size of the smallest nonzero entry of β^*. Notice that the smallest non-zero entry of β^* is allowed to vanish asymptotically, provided it does not do so faster than $\sqrt{p\sigma^2/n}$. Similar conditions are found in Fan and Peng (2004) and Zou and Zhang (2009). Condition (C) restricts the rates of the tuning parameters λ and a. Note that condition (C) does not constrain the minimum size of a. Indeed, no such constraint is required for our asymptotic results about the EXP estimator. Since the EXP penalty approaches the L_0 penalty as $a \to 0$, this suggests that the EXP and L_0-penalized least squares estimator have similar asymptotic properties. On the other hand, in practice, we have found that one should not take a too small, in order to preserve stability of the EXP estimator. Condition (D) is an identifiability condition. Conditions (E) and (F) are used to prove asymptotic normality of EXP estimators and are related to the Lindeberg condition of the Lindeberg–Feller central limit theorem. As we will see in Theorem 2, which is stated below, Conditions (A)–(F) imply that EXP has the oracle property and may correctly identify the model A.

3.2 Oracle properties

Theorem 1 *Suppose that conditions* (A)–(D) *hold, then for every* $r \in (0, 1)$*, there exists a constant* $C_0 > 0$ *such that*

$$\liminf_{n\to\infty} P\left[\arg\min_{\beta} Q_n(\beta) \subseteq \left\{\beta \in R^p \|\beta - \beta^*\| \le C\sqrt{p\sigma^2/n}\right\}\right] > 1 - r. \quad (8)$$

whenever $C \geq C_0$*. Consequently, there exists a sequence of local minimizers of* $Q_n(\beta)$*,* $\hat{\beta}$*, such that* $\|\hat{\beta} - \beta^*\| = O_P(\sqrt{p\sigma^2/n})$*.*

Theorem 2 (Oracle properties) *Suppose that* (A)–(F) *hold, then there exists a sequence of* $\sqrt{n/p\sigma^2}$*-consistent local minima of EXP,* $\hat{\beta}$*, such that:*

(i) (*Model selection consistency*)

$$\lim_{n\to\infty} P(\{j; \hat{\beta}_j \neq 0\} = A) = 1.$$

(ii) (*Asymptotic normality*)

$$\sqrt{n}B_n(n^{-1}X_A^{\mathrm{T}}X_A/\sigma^2)^{1/2}(\hat{\beta}_A - \beta_A^*) \to N(0, G),$$

in distribution, where B_n *is any arbitrary* $q \times |A|$ *matrix such that* $B_n B_n^{\mathrm{T}} \to G$*, and* G *is a* $q \times q$ *nonnegative symmetric matrix.*

It is worth pointing out that we do not make any assumptions about the sparsity level of β^* in Theorem 2. In other words, we do not make any assumptions about

p_0, other than those implied by our assumptions about p. In any implementation of EXP, concrete values of the tuning parameters λ and a must be selected. In Sect. 4 we propose an MBIC tuning parameter selection procedure and prove that when EXP is implemented with MBIC tuning parameter selection, the resulting estimator consistently selects the correct model.

4 Regularity parameter selection

Tuning parameter selection is an important issue in most PLS procedures. There are relatively few studies on the choice of penalty parameters. Traditional model selection criteria, such as AIC (Akaike 1973) and BIC (Schwarz 1978), suffer from a number of limitations. Their major drawback arises because parameter estimation and model selection are two different processes, which can result in instability (Breiman 1996) and complicated stochastic properties. To overcome the deficiency of traditional methods, Fan and Li (2001) proposed the SCAD method, which estimates parameters while simultaneously selecting important variables. They selected tuning parameter by minimizing the generalized cross-validation criterion (Breiman 1995; Tibshirani 1996; Fan and Li 2001).

However, it is well known that GCV and AIC-based methods are not consistent for model selection in the sense that, as $n \to \infty$, they may select irrelevant predictors with non-vanishing probability (Shao 1993; Wang et al. 2007). On the other hand, BIC-based tuning parameter selection roughly corresponds to maximizing the posterior probability of selecting the true model in an appropriate Bayesian formulation and has been shown to be consistent for model selection in several settings (Wang et al. 2007; Wang and Leng 2007; Zou and Hastie 2007; Lee et al. 2014). The BIC tuning parameter selector is defined by

$$\mathrm{BIC} = \log\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{n}\right) + \widehat{DF}\frac{\log(n)}{n}. \tag{9}$$

where $\widehat{DF}$ is the generalized degrees of freedom given by

$$\widehat{DF} = \mathrm{tr}\left\{\mathbf{X}(\mathbf{X}^{\mathrm{T}} + n\Sigma_{\lambda})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right\},$$

and $\Sigma_{\lambda} = \mathrm{diag}\{p'_{\lambda}(|\hat{\beta}_1|)/|\hat{\beta}_1|, \ldots, p'_{\lambda}(|\hat{\beta}_p|)/|\hat{\beta}_p|\}$. The diagonal elements of Σ_{λ} are coefficients of quadratic terms in the local quadratic approximation to the SCAD penalty function $p_{\lambda}(\cdot)$ (Fan and Li 2001). Dicker et al. (2013) pointed out that the consistency results for BIC tuning parameter selection assume that the number of predictors is fixed, they proposed a BIC-like procedures which implemented by minimizing

$$\mathrm{BIC}_0 = \log\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{n - \hat{p}_0}\right) + \frac{\log(n)}{n}\hat{p}_0.$$

They proposed estimating the degrees of freedom by the number of selected coefficients: $\widehat{DF} = \hat{p}_0$, where $\hat{p}_0 = |\{j : \hat{\beta}_j \neq 0\}|$. To estimate the residual variance, they use $\hat{\sigma}^2 = (n - \hat{p}_0)^{-1}\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2$. This differs from other estimates of the residual variance used in PLS methods, where the denominator $n - \hat{p}_0$ is replaced by n (Wang et al. 2007); here, $n - \hat{p}_0$ is used to account for degrees of freedom lost to estimation.

In this section, motivated by Zou and Hastie (2007) and Dicker et al. (2013), we propose an MBIC tuning parameter selector for the EXP procedure and show that the EXP/MBIC procedure is consistent for model selection, provided $p\sigma^2/n \to 0$ and other regularity conditions hold. On the basis of the above notation, we defined an MBIC as

$$\mathrm{MBIC} = \log\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{n - \hat{p}_0}\right) + \frac{\log(n)\hat{p}_0}{n} C_n, \tag{10}$$

where C_n is some positive constant to be discussed more carefully. If $C_n = 1$, the modified BIC (10) reduces to the BIC_0. Moreover, p is allowed to diverge to ∞ as $n \to \infty$. Slightly stronger conditions on p and $\rho = \min_{j\in A} |\beta_j^*|$ than those required for Theorem 2 are needed for the next theorem, which implies that the EXP/MBIC procedure is consistent for model selection.
(A$'$) $n \to \infty$ and $C_n\sigma^2 p\log(n)/n \to 0$.
(B$'$) $\rho\sqrt{n/C_n\sigma^2 p\log(n)} \to \infty$, where $\rho = \min_{j\in A} |\beta_j^*|$.

Theorem 3 *Suppose that conditions* (A$'$–B$'$), (C) *and* (E–F) *hold and* $C_n \to \infty$. *Suppose further that* $\Omega \subseteq R^2$ *is a subset which contains a sequence* $(\lambda, a) = (\lambda_n^*, a_n^*)$ *such that condition* (C) *holds. Let* $\hat{\beta}^* = \hat{\beta}(\lambda_n^*, a_n^*)$ *be the local minima of EXP described in Theorem* 2 *and let* $\mathrm{MBIC}^- = \inf\{\mathrm{MBIC}\ \{\hat{\beta}(\lambda, a)\}; (\lambda, a) \in \Omega, \hat{A} \neq A\}$. *Then*

$$\lim_{n\to\infty} P\left\{\mathrm{MBIC}\{\hat{\beta}(\lambda_n^*, a_n^*)\} < \mathrm{MBIC}^-\right\} = 1,$$

where $\hat{A} = \{j; \hat{\beta}_j \neq 0\}$.

Theorem 3 implies that if $\hat{\beta}(\hat{\lambda}, \hat{a})$ is chosen to minimize MBIC, then $\hat{\beta}(\hat{\lambda}, \hat{a})$ is consistent for model selection. In other words, if $\mathrm{MBIC}\{\hat{\beta}(\hat{\lambda}, \hat{a})\} = \mathrm{MBIC}^-$, then $\lim_{n\to\infty} P\{\{j; \hat{\beta}_j(\hat{\lambda}, \hat{a}) \neq 0\} = A\} = 1$.

Although in theory we require $C_n \to \infty$, its rate of divergence can be arbitrarily slow. For example, $C_n = \log(\log(p))$ is used for all our numerical experiments and the simulation results are quite encouraging.

5 Implementation: algorithm

Finding the estimator of β that minimizes the objective function (7) poses a number of interesting challenges because the penalized functions are nondifferentiable at the origin and nonconcave with respect to β (Breheny 2015).

Coordinate optimization has been widely used to solve regularization problems. For example, for the PLS problem, Fu (1998), Daubechies et al. (2004) and Wu and Lange (2008) proposed a coordinate descent (CD) algorithm that iteratively optimizes Eq. (7) one component at a time. The algorithm was also recently proposed for a very general class of penalized likelihood methods by Fan and Lv (2011), who refer to the algorithm as "iterative coordinate ascent" (ICA). Coordinate descent algorithms for fitting LASSO-penalized models have also been described by Friedman et al. (2007, 2010) and Breheny and Huang (2011). Recently, Peng and Wang (2014) proposed a new iterative coordinate descent algorithm (QICD) for solving nonconvex penalized quantile regression in high dimension. In this section, we consider the CD algorithm for obtaining EXP estimators for a range of tuning parameter values. The EXP tuning parameter selection procedure is described above.

Next we describe CD algorithms for least squares regression penalized by EXP. For $\beta > 0$, the derivative of the EXP penalty is

$$p'_{\lambda,a} = \frac{\lambda}{a} \mathrm{e}^{-\frac{\beta}{a}},$$

for $\lambda > 0, a > 0$. The rationale behind the penalty can be understood by considering its derivative: EXP penalty allows the penalty to decay exponentially, approaching 0 asymptotically but never reaching it.

The rationale behind the EXP can also be understood by considering its univariate solution. Consider the simple linear regression of y upon x, with unpenalized least squares solution $z = n^{-1}x^{\mathrm{T}}y$ (recall that x has been standardized so that $x^{\mathrm{T}}x = n$). In fact, for this simple linear regression problem, it is easy to show that the EXP estimator has the following closed form:

$$\hat{\beta} = \begin{cases} \mathrm{sign}(z)\beta_s, & \text{if } |z| > \frac{\lambda}{a}, \\ 0, & \text{if } |z| \le \frac{\lambda}{a}. \end{cases} \tag{11}$$

where β_s is the solution of equation $\beta - |z| + \mathrm{sign}(\beta)\frac{\lambda}{a}\mathrm{e}^{-\frac{|\beta|}{a}} = 0$, which may be done very rapidly using Newton iterative formula or some other procedure.

The idea of the CD algorithm is to find local optima of a multivariate optimization problem by solving a sequence of univariate optimization problems. Consider a penalized residual sum squares as (7). Without loss of generality, let us assume that the predictors are standardized: $\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$. For each fixed λ, a, cyclic coordinate descent can be easily implemented for solving the EXP. Let $r_i = y_i - x_i^{\mathrm{T}}\tilde{\beta}$ be the current residual. To update the estimate for β_j we need to solve a univariate EXP problem

$$\hat{\beta}_j = \arg\min_{\beta_j} Q_n(\beta_j|\tilde{\beta}),$$

where

$$Q_n(\beta_j|\tilde{\beta}) = \frac{1}{2}(\beta_j - \tilde{\beta}_j)^2 - \frac{1}{n}\sum_{i=1}^{n} r_i x_{ij}(\beta_j - \tilde{\beta}_j) + \lambda p_{\lambda,a}(\beta_j).$$

Indeed, one observes that the solution can be obtained by Eq. (11).

The CD algorithm is implemented by minimizing $Q_n(\beta_j|\tilde{\beta})$ and using the solution to update β; we next set $\tilde{\beta}_j = \hat{\beta}_j$ as the new estimate. In this way, we cycle through the indices $j = 1, \ldots, p$. The operation is sequentially conducted on each coordinate β_j till convergence.

The CD algorithm returns the minimum of the EXP PLS, for a fixed pair of tuning parameters, (λ, a). To obtain a EXP solution path, we repeatedly implement CD algorithm for a range of value (λ, a). The details are given in Algorithm 1. In practice, we have found that if the columns of $\mathbf{X}$ are standardized so that $\|\mathbf{x}_j\|^2 = n$, for $j = 1, \ldots, p$, then taking $a = 0.01$ or selecting a from a relatively small range of possible values works well.

Algorithm 1 The coordinate descent algorithm for EXP PLS

```
1. Input a grid of increasing $\lambda$ values $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$, and a grid of increasing $a$
   values $\Gamma = \{a_1, \ldots, a_K\}$. Define $\lambda_{L+1}$, such that $\hat{\beta}(\lambda_{L+1}, a_K) = 0$.
2. For each value of $l \in \{L, L-1, \ldots, 1\}$ repeat the following:
   (1) Initialize $\tilde{\beta} = \hat{\beta}(\lambda_{l+1}, a_K)$
   (2) For each value of $k \in \{K, K-1, \ldots, 1\}$ repeat the following:
      (a) Cyclic coordinate descent, for $j = 1, 2, \ldots, p$
          Calculate $r_i = y_i - x_i^T\tilde{\beta}$ and $\hat{\beta}_j = \arg\min_{\beta_j} Q_n(\beta_j|\tilde{\beta})$
          where
           $Q_n(\beta_j|\tilde{\beta}) = \frac{1}{2}(\beta_j - \tilde{\beta}_j)^2 - \frac{1}{n}\sum_{i=1}^{n} r_i x_{ij}(\beta_j - \tilde{\beta}_j) + \lambda p_{\lambda,a}(\beta_j)$
      (b) Set $\tilde{\beta}_j = \hat{\beta}_j$
      (c) Repeat steps (a)-(b) until the updates $\hat{\beta}$ converge to $\hat{\beta}^*$.
      (d) Assign $\hat{\beta}(\lambda_l, a_k) \leftarrow \hat{\beta}^*$.
   (3) Decrement $k$.
3. Decrement $l$.
4. Return the two-dimensional solution surface $\hat{\beta}(\lambda, a)$, $(\lambda, a) \in \Lambda \times \Gamma$.
```

6 Simulation studies and a data example

The standard errors for the estimated parameters can be obtained directly because we are estimating parameters and selecting variables at the same time. Let $\hat{\beta} = \hat{\beta}(\lambda, a)$ be a local minimizer of EXP. Following Fan and Li (2001) and Fan and Peng (2004), standard errors of $\hat{\beta}$ may be estimated by using quadratic approximations to EXP. Indeed, the approximation

$$p_{\lambda,a}(|\beta_j|) \approx p_{\lambda,a}(|\beta_{j0}|) + \frac{1}{2|\beta_{j0}|}p'_{\lambda,a}(|\beta_{j0}|)(\beta_j^2 - \beta_{j0}^2), \quad \text{for } \beta_j \approx \beta_{j0} \qquad (12)$$

suggests that EXP may be replaced by the quadratic minimization problem

$$\min\left\{\frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p}\frac{p'_{\lambda,a}(|\beta_{j0}|)}{|\beta_{j0}|}\beta_j^2\right\}, \qquad (13)$$

at least for the purposes of obtaining standard errors. Using this expression, we obtain a sandwich formula for the estimated standard error of $\hat{\beta}_{\hat{A}}$,

$$\widehat{\mathrm{cov}}(\hat{\beta}_{\hat{A}}) = \hat{\sigma}^2\{\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}} + n\Delta_{\hat{A},\hat{A}}(\hat{\beta})\}^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}}\{\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}} + n\Delta_{\hat{A},\hat{A}}(\hat{\beta})\} \tag{14}$$

where $\hat{A} = \{j; \hat{\beta}_j \neq 0\}$, $\Delta(\beta) = \mathrm{diag}\{p'_{\lambda,a}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda,a}(|\beta_p|)/|\beta_p|\}$, $\hat{\sigma}^2 = (n - \hat{p}_0)^{-1}\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2$, and $\hat{p}_0 = |\hat{A}|$ is the number of elements in $|\hat{A}|$.

6.1 Simulation studies I

In this example, simulation data are generated from the linear regression model,

$$y = \mathbf{x}^{\mathrm{T}}\beta + \sigma\varepsilon,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$, $\varepsilon \sim N(0, 1)$ and $\mathbf{x}$ is multivariate normal distribution with zero mean and covariance between the ith and jth elements being $\rho^{|i-j|}$ with $\rho = 0.5$. In our simulation, the sample size n is set to be 100 and 200, $\sigma = 3$. For each case, we repeated the simulation 1000 times.

In addition to EXP, we consider the LASSO, adaptive LASSO (ALASSO) (with weights $\omega_j = |\hat{\beta}^{(0)}|^{-1}$, where $\hat{\beta}^{(0)}$ is the OLS estimator), SCAD and MCP procedures. Covariates were standardized to have $\|\mathbf{x}_j\| = n, \ j = 1, \ldots, p$, prior to obtaining estimates; however, all summary statistics discussed below pertain to estimators transformed to the original scale. In our simulations, EXP, LASSO, ALASSO, MCP and SCAD solution paths are all computed using CD algorithms. The tuning parameter selection is all performed with MBIC as (10). For EXP tuning parameter selection, we find that $a = 0.01$ works well.

The model error for $\hat{\mu} = \mathbf{x}^{\mathrm{T}}\hat{\beta}$ is $ME(\hat{\mu}) = (\hat{\beta} - \beta)^{\mathrm{T}}E(\mathbf{x}\mathbf{x}^{\mathrm{T}})(\hat{\beta} - \beta)$ for linear model. Simulation results are summarized in Table 1, in which MRME stands for median of ratios of ME of a selected model to that of the un-penalized minimum square estimate under the full model. Both the columns of "C" and "IC" are measures of model complexity. Column "C" shows the average number of nonzero coefficients correctly estimated to be nonzero, and column "IC" presents the average number of zero coefficients incorrectly estimated to be nonzero. In the column labeled "Under-fit", we present the proportion of excluding any nonzero coefficients in 1000 replications. Likewise, we report the probability of selecting the exact subset model and the probability of including all three significant variables and some noise variables in the columns "Correct-fit" and "Over-fit", respectively.

As can be seen from Table 1, all variable selection procedures dramatically reduce model error. EXP has the smaller model error among all competitors, followed by the MCP and SCAD. In terms of sparsity, EXP also has the higher probability of correct fit. The EXP penalty performs better than the other penalties. Also, EXP has some advantages when the dimensionality p is high which can be seen in simulation III.

We now test the accuracy of our standard error formula (14). The median absolute deviation divided by 0.6745, denoted by SD in Table 2, of 1000 estimated coefficients

Table 1 Simulation results for linear regression models of simulation I

Method	MRME	No. of zeros		Proportion of		
		C	IC	Under-fit	Correct-fit	Over-fit
$n = 100, \sigma = 3$						
LASSO	0.6503	2.9990	0.9980	0.0010	0.3840	0.6150
ALASSO	0.3668	2.9430	0.2940	0.0570	0.7290	0.2140
SCAD	0.3653	2.9150	0.4910	0.0850	0.5890	0.3260
MCP	0.2920	2.9180	0.2520	0.0810	0.7390	0.1800
EXP	0.2839	2.9600	0.2440	0.0400	0.7640	0.1960
$n = 200, \sigma = 3$						
LASSO	0.6738	3.0000	0.8540	0.0000	0.4520	0.5480
ALASSO	0.3290	3.0000	0.1280	0.0000	0.8870	0.1130
SCAD	0.2777	2.9950	0.1900	0.0050	0.8530	0.1420
MCP	0.2496	2.9980	0.1190	0.0020	0.9050	0.0930
EXP	0.2488	3.0000	0.1450	0.0000	0.8710	0.1290

Table 2 Standard deviations of estimators for the linear regression model ($n = 200$)

Method	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
	SD	SD_m $(\mathrm{SD}_{\mathrm{mad}})$	SD	SD_m (SD_{mad})	SD	SD_m $(\mathrm{SD}_{\mathrm{mad}})$
$n = 200$, $p = 12$						
LASSO	0.2585	0.2065 (0.0128)	0.2407	0.1847 (0.0186)	0.2297	0.1827 (0.0126)
ALASSO	0.2670	0.2334 (0.0134)	0.2693	0.2196 (0.0178)	0.2397	0.2053 (0.0112)
SCAD	0.2738	0.2435 (0.0160)	0.2767	0.2415 (0.0201)	0.2240	0.2149 (0.0114)
MCP	0.2604	0.2457 (0.0161)	0.2378	0.2468 (0.0170)	0.2238	0.2153 (0.0117)
EXP	0.2575	0.2460 (0.0162)	0.2354	0.2477 (0.0167)	0.2256	0.2155 (0.0120)
$n = 200$, $p = 20$						
LASSO	0.2655	0.1981 (0.0115)	0.2526	0.1724 (0.0168)	0.2487	0.1743 (0.0119)
ALASSO	0.2753	0.2295 (0.0134)	0.2788	0.2124 (0.0188)	0.2525	0.2017 (0.0115)
SCAD	0.2907	0.2400 (0.0155)	0.3510	0.2328 (0.0295)	0.2328	0.2140 (0.0117)
MCP	0.2646	0.2445 (0.0155)	0.2587	0.2450 (0.0168)	0.2286	0.2147 (0.0115)
EXP	0.2594	0.2450 (0.0156)	0.2512	0.2469 (0.0164)	0.2311	0.2149 (0.0116)

in the 1000 simulations can be regarded as the true standard error. The median of the 1000 estimated SD's, denoted by SD_m, and the median absolute deviation error of the 1000 estimated standard errors divided by 0.6745, denoted by $\mathrm{SD}_{\mathrm{mad}}$, gauge the overall performance of the standard error formula (13). Table 2 presents the results for nonzero coefficients when the sample size $n = 200$. The results for the other case with $n = 100$ are similar. Table 2 suggests that the sandwich formula performs surprisingly well.

Table 3 Simulation results for linear regression models of simulation I with common BIC

Method	MRME	No. of zeros		Proportion of		
		C	IC	Under-fit	Correct-fit	Over-fit
$n = 100, \sigma = 3$						
LASSO	0.6104	2.9990	1.2210	0.0010	0.3260	0.6730
ALASSO	0.3698	2.9500	0.3720	0.0500	0.6880	0.2620
SCAD	0.3513	2.9840	1.2820	0.0160	0.3390	0.6450
MCP	0.3403	2.9740	0.8090	0.0260	0.4820	0.4920
EXP	0.3129	2.9650	0.3170	0.0350	0.7240	0.2410
$n = 200, \sigma = 3$						
LASSO	0.6604	3.0000	0.9580	0.0000	0.4170	0.5830
ALASSO	0.3294	3.0000	0.1530	0.0000	0.8680	0.1320
SCAD	0.2880	3.0000	0.7790	0.0000	0.5510	0.4490
MCP	0.2900	3.0000	0.5240	0.0000	0.6140	0.3860
EXP	0.2829	3.0000	0.2020	0.0000	0.8350	0.1650

The simulation results described above give an indication of the performance of the proposed EXP methods, in comparison with current recommended implementations of other PLS methods. In particular, we focus on MBIC criterion (10) implementations when implementing the alternative PLS methods. Table 1 summarizes the results of a simulation study where the same MBIC criterion for (7) was used for each PLS method.

Then, when the common BIC criterion (9) is utilized, the results indicate that EXP performs well when compared to the alternative methods. The details can be seen in Table 3. Compare Tables 1 with 3, we can see: (1) EXP/MBIC perform better than EXP/BIC which indicate that MBIC criteria is more suitable for EXP estimator; (2) LASSO, ALASSO, SCAD estimation method have the smaller model error under the BIC criterion; (3) MCP estimator perform better with MBIC than with BIC, the reason need to be further explained; (4) When $n \to \infty$, all kinds of variable selection methods can identify the correct model consistently and reduce the model error.

6.2 Simulation study II

The example is from Wang et al. (2009). In this example, we consider the situation where the dimension of the full model and the dimension of the true model are all diverging. More specifically, we take $p = [7n^{1/4}]$ and the dimension of the true model to be $p_0 = [p/3]$, $[t]$ stands for the largest integer no larger than t. To summarize the simulation results, we computed the median of the relative model error MRME, the average model size (i.e. the number of non-zero coefficients), MS, and also the percentage of the correctly identified true models CM. Intuitively, a better model selection procedure should produce more accurate prediction results (i.e. smaller MRME-value), more correct model sizes (i.e. MS $\approx p_0$) and better model selection capability (i.e.

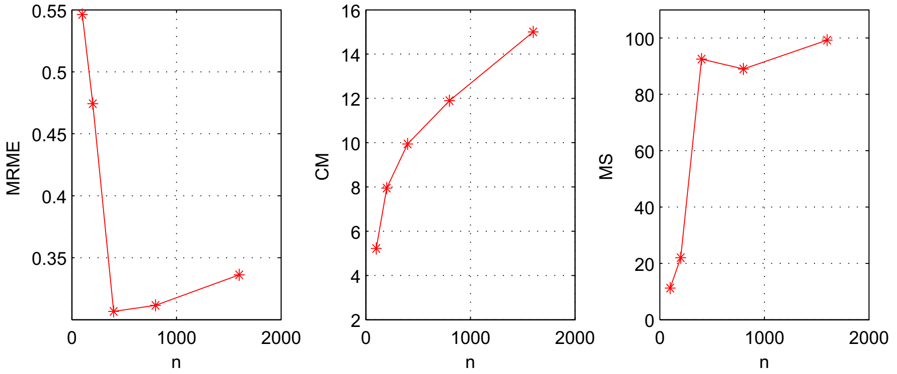

Fig. 2 Detailed simulation results with normal ε: MRME (*left*); MS (*middle*); CM (*right*)

larger CM-value). For a more detailed explanation of MRME, MS and CM, we refer to Fan and Li (2001) and Wang and Leng (2007). The detailed results are reported in Fig. 2. As one can see that the CM-value of EXP/MBIC approaches 100 % quickly, which clearly confirms the consistency of the MBIC proposed. As a direct consequence, the MRME-values are consistently smaller and the MS-values approximately equal p_0.

6.3 Simulation study III

In this simulation study presented here, we examined the performance of the various PLS methods for p substantially larger than in the previous studies. In particular, we took $p = 339$, $n = 500$, $\sigma^2 = 5$, and $\beta^* = (2I_{37}^{\mathrm{T}}, -3I_{37}^{\mathrm{T}}, I_{37}^{\mathrm{T}}, 0_{228}^{\mathrm{T}})$, where $I^k \in R^k$ is the vector with all entries equal to 1. Thus, $p_0 = 111$. We simulated 200 independent datasets $\{(y_1, x_1^{\mathrm{T}}), \ldots, (y_n, x_n^{\mathrm{T}})\}$ in this study and, for each dataset, we computed estimates of β^*. Results from this simulation study are found in Table 4.

Perhaps the most striking aspect of the results presented in Table 4 is that hardly no method ever selected the correct model in this simulation study. However, given that p, p_0, and β^* are substantially larger in this study than in the previous simulation studies, this may not be too surprising. Notice that on average, EXP selects the most parsimonious models of all methods and has the smaller model error. EXP's nearest competitor in terms of model error is ALASSO. This implementation of ALASSO has mean model error 0.2783, but its average selected model size is 103.5250 larger than EXP's. Since $p_0 = 111$, it is clear that EXP underfits in some instances. In fact, all of the methods in this study underfit to some extent. This may be due to the fact that many of the non-zero entries in β^* are small relative to the noise level $\sigma^2 = 5$.

6.4 Application

In this section, we apply the EXP regularization scheme to a prostate cancer example. The dataset in this example is derived from a study of prostate cancer by Stamey et al. (1989). The dataset consists of the medical records of 97 patients who were

Table 4 Simulation results for linear regression models of Example 3

Method	MRME	No. of zeros		Proportion of		
		C	IC	Under-fit	Correct-fit	Over-fit
$n = 500, \sigma = 5$						
LASSO	0.3492	110.2000	24.7450	0.5650	0.0000	0.4350
ALASSO	0.2783	103.5250	6.1250	1.0000	0.0000	0.0000
SCAD	0.3012	106.0400	35.7150	0.9900	0.0000	0.0150
MCP	0.2791	103.3600	8.1100	1.0000	0.0000	0.0900
EXP	0.2792	101.6200	5.0200	1.0000	0.0000	0.0000

Table 5 Prostate cancer data: comparing different methods

Method	R^2	R^2/R^2_{OLS}	Variables selected	Total times (s)
OLS	0.6615	1.0000	All	0.0110
LASSO	0.5867	0.8870	(1, 2, 4, 5,8)	0.0045
ALASSO	0.5925	0.8950	(1, 2, 5)	0.0060
SCAD	0.6194	0.9364	(1, 2, 4, 5)	0.0235
MCP	0.6087	0.9202	(1, 2, 4, 5)	0.0092
EXP	0.6074	0.9184	(1, 2, 5)	0.0210

about to receive a radical prostatectomy. The predictors are eight clinical measures: log (cancer volume) (lcavol), log (prostate weight) (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log (capsular penetration) (lcp), Gleason score (Gleason) and percentage Gleason score 4 or 5 (pgg45). The response is the logarithm of prostate-specific antigen (lpsa). One of the main aims here is to identify which predictors are more important in predicting the response.

The LASSO, ALASSO, SCAD, MCP and EXP are all applied to the data. The BIC and MBIC are used to select tuning parameters. We also compute the OLS estimate of the prostate cancer data. Results are summarized in Table 5. The OLS estimator does not perform variable selection. LASSO selects five variables in the final model, SCAD and MCP select lcavol, lweight lbph and svi in the final model. While ALASSO and EXP select lcavol, lweight and svi. Thus, EXP selects a substantially simpler model than LASSO, SCAD and MCP. Furthermore, as indicated by the columns labeled R^2 (R^2 is equal to one minus the residual sum of squares divided the total sum of squares) and R^2/R^2_{OLS} in Table 5, the EXP estimator describes more variability in the data than LASSO and ALASSO, and nearly as much as OLS estimator.

7 Conclusion

In this paper, a new EXP penalty which very closely resembles the L_0 penalty is proposed. The model selection oracle property is investigated and the EXP is implemented by using CD algorithm. Moreover, an MBIC tuning parameter selection method for EXP is proposed and it is shown that it consistently identifies the correct model. Numerical studies further endorse our theoretical results and the advantage of our new methods for model selection.

It would be interesting to extend the results to regularization methods for the generalized linear models (GLMs) and more general models and loss functions. Also, motivated by the elastic net (Zou and Hastie 2005) and the adaptive elastic net (Zou and Zhang 2009), one could consider a mixed penalty involving EXP and an L_2-norm penalty. In this paper, we do not address the situation where $p \gg n$, in fact, the proposed EXP method can be easily extended for variable selection in the situation $p \gg n$. These problems are beyond the scope of this paper and will be interesting topics for future research.

Appendix

Proof of Theorem 1 Let $\alpha_n = \sqrt{p\sigma^2/n}$ and fix $r \in (0, 1)$. To prove the Theorem, it suffices to show that if $C > 0$ is large enough, then

$$Q_n(\beta^*) < \inf_{\|\mu\|=C} Q_n(\beta^* + \alpha_n \mu)$$

holds for all n sufficiently large, with probability at least $1 - r$. Define $D_n(\mu) = Q_n(\beta^* + \alpha_n\mu) - Q_n(\beta^*)$ and note that

$$\begin{aligned} D_n(\mu) &= \frac{1}{2n}(\alpha_n^2\|\mathbf{X}\mu\|^2 - 2\alpha_n\varepsilon^{\mathrm{T}}\mathbf{X}\mu) + \sum_{j=1}^{p}\{p_{\lambda,a}(|\beta_j^* + \alpha_n\mu_j|) - p_{\lambda,a}(|\beta_j^*|)\} \\ &\geq \frac{1}{2n}(\alpha_n^2\|\mathbf{X}\mu\|^2 - 2\alpha_n\varepsilon^{\mathrm{T}}\mathbf{X}\mu) + \sum_{j\in K(\mu)}\{p_{\lambda,a}(|\beta_j^* + \alpha_n\mu_j|) - p_{\lambda,a}(|\beta_j^*|)\}, \end{aligned}$$

where $K(\mu) = \{j;\, p_{\lambda,a}(|\beta_j^* + \alpha_n\mu_j|) - p_{\lambda,a}(|\beta_j^*|) < 0\}$. The fact that $p_{\lambda,a}$ is concave on $[0, \infty)$ implies that

$$\begin{aligned} &p_{\lambda,a}(|\beta_j^* + \alpha_n\mu_j|) - p_{\lambda,a}(|\beta_j^*|) \geq p'_{\lambda,a}(|\beta_j^* + \alpha_n\mu_j|)(|\beta_j^* + \alpha_n\mu_j| - |\beta_j^*|) \\ &\geq p'_{\lambda,a}(|\beta_j^* + \alpha_n\mu_j|)(-\alpha_n|\mu_j|) = -\frac{\lambda\alpha_n|\mu_j|}{a}\mathrm{e}^{-\frac{|\beta_j^*+\alpha_n\mu_j|}{a}}. \end{aligned}$$

when n is sufficiently large.

Condition (B) implies that

$$\mathrm{e}^{-\frac{|\beta_j^*+\alpha_n\mu_j|}{a}} \leq \mathrm{e}^{-\frac{\rho}{a}}.$$

Thus, for n big enough,

$$D_n(\mu) \geq \frac{1}{2n}(\alpha_n^2\|\mathbf{X}\mu\|^2 - 2\alpha_n\varepsilon^{\mathrm{T}}\mathbf{X}\mu) - \frac{Cp\lambda\alpha_n}{a}\mathrm{e}^{-\frac{\rho}{a}}. \tag{15}$$

By (D),

$$\frac{1}{2n}\alpha_n^2\|\mathbf{X}\mu\|^2 \geq \frac{\lambda_{\min}}{2}C^2\alpha_n^2. \tag{16}$$

On the other hand (D) implies,

$$\frac{1}{n}\alpha_n|\varepsilon^{\mathrm{T}}\mathbf{X}\mu| \leq \frac{C\alpha_n}{\sqrt{n}}\|\frac{1}{\sqrt{n}}\mathbf{X}^{\mathrm{T}}\varepsilon\| = O_P(C\alpha_n^2). \tag{17}$$

Furthermore, (C) and (B) imply

$$\frac{Cp\lambda\alpha_n}{a}\mathrm{e}^{-\frac{\rho}{a}} = o(C\alpha_n^2). \tag{18}$$

From (15)–(18), we conclude that if $C > 0$ is large enough, then $\inf_{\|\mu\|=C} D_n(\mu) > 0$ holds for all n sufficiently large, with probability at least $1 - r$. This proves the Theorem 1. □

To prove Theorem 2, we first show that the EXP penalized estimator possesses the sparsity property by following lemma.

Lemma 1 *Assume that* (A)–(D) *hold, and fix* $C > 0$. *Then*

$$\lim_{n\to\infty} P\left[\mathop{\arg\min}_{\|\beta-\beta^*\|\leq C\sqrt{p\sigma^2/n}} Q_n(\beta) \subseteq \left\{\beta \in R^p; \beta_{A^c} = 0\right\}\right] = 1.$$

where $A^c = \{1, \ldots, p\}\backslash A$ *is the complement of* A *in* $\{1, \ldots, p\}$.

Proof Suppose that $\beta \in R^p$ and that $\|\beta - \beta^*\| \leq C\sqrt{p\sigma^2/n}$. Define $\tilde{\beta} \in R^p$ by $\tilde{\beta}_{A^c} = 0$ and $\tilde{\beta}_A = \beta_A$. Similar to the proof of Theorem 1, let

$$D_n(\beta, \tilde{\beta}) = Q_n(\beta) - Q_n(\tilde{\beta}),$$

where $Q_n(\beta)$ is defined in (7). Then

$$\begin{aligned}
&D_n(\beta, \tilde{\beta}) \\
&\quad= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 - \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\tilde{\beta}\|^2 + \sum_{j\in A^c} p_{\lambda,a}(|\beta_j|) \\
&\quad= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\tilde{\beta} - \mathbf{X}(\beta - \tilde{\beta})\|^2 - \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\tilde{\beta}\|^2 + \sum_{j\in A^c} p_{\lambda,a}(|\beta_j|) \\
&\quad= \frac{1}{2n}(\beta - \tilde{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\beta - \tilde{\beta}) - \frac{1}{n}(\beta - \tilde{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\tilde{\beta}) + \sum_{j\in A^c} p_{\lambda,a}(|\beta_j|) \\
&\quad= O_p(\|\beta - \tilde{\beta}\|\sqrt{p\sigma^2/n}) + \sum_{j\in A^c} p_{\lambda,a}(|\beta_j|). \qquad (19)
\end{aligned}$$

On the other hand, since the EXP penalty is concave on $[0, \infty)$,

$$p_{\lambda,a}(|\beta_j|) \geq p'_{\lambda,a}(|\beta_j|)|\beta_j| = \frac{\lambda}{a}\mathrm{e}^{-\frac{|\beta_j|}{a}}|\beta_j| \geq \frac{\lambda}{a}\mathrm{e}^{-\frac{C\sqrt{p\sigma^2/n}}{a}}|\beta_j|.$$

Thus,

$$\sum_{j\in A^c} p_{\lambda,a}(|\beta_j|) \geq \frac{\lambda}{a}\mathrm{e}^{-\frac{C\sqrt{p\sigma^2/n}}{a}}\|\beta - \tilde{\beta}\|. \qquad (20)$$

By (C), it is clear that

$$\liminf_{n\to\infty}\left(\frac{\lambda}{a}\mathrm{e}^{-\frac{C\sqrt{p\sigma^2/n}}{a}}\right) > 0.$$

and $\lambda\sqrt{n/(p\sigma^2)} \to \infty$. Combining these observations with (19) and (20) gives $D_n(\beta, \tilde{\beta}) > 0$ with probability tending to 1, as $n \to \infty$. The result follows. □

Proof of Theorem 2 Taken together, Theorem 1 and Lemma 1 imply that there exist a sequence of local minima $\hat{\beta}$ of (7) such that $\|\hat{\beta} - \beta^*\| = O_P(\sqrt{p\sigma^2/n})$ and $\hat{\beta}_{A^c} = 0$. Part (i) of the theorem follows immediately.

To prove part (ii), observe that on the event $\{j;\, \hat{\beta}_j \neq 0\} = A$, we must have

$$\hat{\beta}_A = \beta_A^* + (\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}\mathbf{X}_A^{\mathrm{T}}\varepsilon - (n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}p'_A,$$

where $p'_A = (p'_{\lambda,a}(\hat{\beta}_j))_{j\in A}$. It follows that

$$\begin{aligned}
&\sqrt{n}B_n(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A/\sigma^2)^{1/2}(\hat{\beta}_A - \beta_A^*) \\
&\quad= B_n(\sigma^2\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}\mathbf{X}_A^{\mathrm{T}}\varepsilon - nB_n(\sigma^2\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}p'_A,
\end{aligned}$$

whenever $\{j;\, \hat{\beta}_j \neq 0\} = A$. Now note that conditions (A)–(D) imply

$$\|nB_n(\sigma^2 \mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2} p_A'\| = O_P(\sqrt{np/\sigma^2}\frac{\lambda}{a}\mathrm{e}^{-\frac{\rho}{a}}) = o_P(1),$$

Thus,

$$\sqrt{n}B_n(n^{-1}\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A/\sigma^2)^{1/2}(\hat{\beta}_A - \beta_A^*) = B_n(\sigma^2\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}\mathbf{X}_A^{\mathrm{T}}\varepsilon + o_P(1).$$

To complete the proof of (ii), we use the Lindeberg–Feller central limit theorem to show that

$$B_n(\sigma^2\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}\mathbf{X}_A^{\mathrm{T}}\varepsilon \to N(0, G), \tag{21}$$

in distribution. Observe that

$$B_n(\sigma^2\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}\mathbf{X}_A^{\mathrm{T}}\varepsilon = \sum_{i=1}^{n}\omega_{i,n},$$

where $\omega_{i,n} = B_n(\sigma^2\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}x_{i,A}\varepsilon_i$.

Fix $\delta_0 > 0$ and let $\eta_{i,n} = x_{i,A}^{\mathrm{T}}(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}B_n^{\mathrm{T}}B_n(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1/2}x_{i,A}$ Then

$$\begin{aligned}
&E[\|\omega_{i,n}\|^2;\, \|\omega_{i,n}\|^2 > \delta_0] \\
&\quad = \eta_{i,n}E[\varepsilon_i^2/\sigma^2;\, \eta_{i,n}\varepsilon_i^2/\sigma^2 > \delta_0] \\
&\quad \leq \eta_{i,n}E(|\varepsilon_i/\sigma|^{2+\delta})^{2/(2+\delta)}P(\eta_{i,n}\varepsilon_i^2/\sigma^2 > \delta_0)^{\delta/(2+\delta)} \\
&\quad \leq \eta_{i,n}^{1+\delta/2+\delta}\delta_0^{-1}E(|\varepsilon_i/\sigma|^{2+\delta})^{2/(2+\delta)}.
\end{aligned}$$

Since $\sum_{i=1}^{n}\eta_{i,n} = tr(B_n^{\mathrm{T}}B_n) \to tr(G) < \infty$ and since (E) implies

$$\max_{1\leq i\leq n}\eta_{i,n}\lambda_{\min}(n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})\lambda_{\max}(B_n^{\mathrm{T}}B_n)\max_{1\leq i\leq n}\frac{1}{n}\sum_{j=1}^{p}x_{ij}^2 \to 0,$$

we must have

$$\begin{aligned}
&\sum_{i=1}^{n}E[\|\omega_{i,n}\|^2;\, \|\omega_{i,n}\|^2 > \delta_0] \\
&\quad \leq \delta_0^{-1}E(|\varepsilon_i/\sigma|^{2+\delta})^{2/(2+\delta)}\sum_{i=1}^{n}\eta_{i,n}^{1+\delta/(2+\delta)} \\
&\quad \leq \delta_0^{-1}E(|\varepsilon_i/\sigma|^{2+\delta})^{2/(2+\delta)}tr(B_n^{\mathrm{T}}B_n)\max_{1\leq i\leq n}\eta_{i,n}^{\delta/(2+\delta)} \\
&\quad \to 0.
\end{aligned}$$

Thus, the Lindeberg condition is satisfied and (21) holds. □

Proof of Theorem 3 Suppose we are on the event $\{j; \hat{\beta}^*_j \neq 0\} = A$. The first order optimality conditions for (7) imply that

$$\hat{\beta}^*_A = (\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}\mathbf{X}_A^{\mathrm{T}}y - n(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}p'_A(\hat{\beta}^*),$$

where $p'_A(\beta) = (p'_{\lambda,a}(\beta_j))_{j\in A}$. Thus,

$$\begin{aligned}
&\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2 \\
&\quad = \varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_A(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}\mathbf{X}_A^{\mathrm{T}}\}\varepsilon + n^2 p'_A(\hat{\beta}^*)^{\mathrm{T}}(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}p'_A(\hat{\beta}^*) \\
&\quad = \varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_A(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{-1}\mathbf{X}_A^{\mathrm{T}}\}\varepsilon + o_P(\sigma^2).
\end{aligned}$$

Now let $\hat{\beta} = \hat{\beta}(\lambda, a)$ be a local minimizer of (7) with $(\lambda, a) \in \Omega$ and let $\hat{A} = \{j; \hat{\beta}_j \neq 0\}$. Note that

$$\begin{aligned}
&\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 \\
&= \mathbf{y}^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\mathbf{y} + n^2 p'_{\hat{A}}(\hat{\beta})^{\mathrm{T}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}p'_{\hat{A}}(\hat{\beta}) \\
&= (\mathbf{X}_{A\backslash\hat{A}}\beta^*_{A\backslash\hat{A}} + \varepsilon)^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}(\mathbf{X}_{A\backslash\hat{A}}\beta^*_{A\backslash\hat{A}} + \varepsilon) \\
&\quad + n^2 p'_{\hat{A}}(\hat{\beta})^{\mathrm{T}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}p'_{\hat{A}}(\hat{\beta}) \\
&= (\beta^*_{A\backslash\hat{A}})^{\mathrm{T}}\mathbf{X}_{A\backslash\hat{A}}{}^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\mathbf{X}_{A\backslash\hat{A}}\beta^*_{A\backslash\hat{A}} \\
&\quad + 2\varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\mathbf{X}_{A\backslash\hat{A}}\beta^*_{A\backslash\hat{A}} \\
&\quad + \varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\varepsilon \\
&\quad + n^2 p'_{\hat{A}}(\hat{\beta})^{\mathrm{T}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}p'_{\hat{A}}(\hat{\beta}).
\end{aligned}$$

Thus, if $A\backslash\hat{A} = \Phi$, then

$$\begin{aligned}
&\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 \\
&\quad \geq \|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2 + (\beta^*_{A\backslash\hat{A}})^{\mathrm{T}}\mathbf{X}_{A\backslash\hat{A}}{}^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\mathbf{X}_{A\backslash\hat{A}} \\
&\qquad \beta^*_{A\backslash\hat{A}} + 2\varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\mathbf{X}_{A\backslash\hat{A}}\beta^*_{A\backslash\hat{A}} + O_P(p\sigma^2) \\
&\quad \geq \|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2 + nr\rho^2 + O_P(\sigma\rho\sqrt{n}) + O_P(p\sigma^2) \\
&\quad = \|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2 + nr\rho^2(1 + o_P(1))
\end{aligned}$$

where $0 < r < \lambda_{\min}(n^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})$ is a positive constant. Furthermore, whenever $\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 - \|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2 > 0$, we have

$$\begin{aligned}
&\mathrm{MBIC}(\hat{\beta}) - \mathrm{MBIC}(\hat{\beta}^*)\\
&\quad = \log\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2}\right) + \log\left(\frac{n - p_0}{n - \hat{p}_0}\right) + \frac{C_n \log(n)}{n}(\hat{p}_0 - p)\\
&\quad \geq 1 - \frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2}{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2} + \log\left(\frac{n - p_0}{n - \hat{p}_0}\right) + \frac{C_n \log(n)}{n}(\hat{p}_0 - p)\\
&\quad \geq \frac{1}{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}\left[\left(1 - \frac{2C_n p \log(n)}{n}\right)\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 - \|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2\right]\\
&\quad \geq \frac{1}{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}[O_P(\sigma^2 C_n p \log(n)) + nr\rho^2(1 + o_P(1))],
\end{aligned}$$

where $\hat{p}_0 = |\hat{A}|$. By Condition (B'), it follows that

$$\lim_{n\to\infty} P\left\{\inf\{\mathrm{MBIC}(\hat{\beta});\, A\backslash\hat{A} \neq \Phi\} > \mathrm{MBIC}(\hat{\beta}^*)\right\} = 1. \tag{22}$$

It remains to consider $\hat{\beta}$, where A is a proper subset of $\hat{A}$. Suppose that $A\backslash\hat{A}$. Then

$$\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 = \varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\varepsilon + n^2 p'_{\hat{A}}(\hat{\beta})^{\mathrm{T}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1} p'_{\hat{A}}(\hat{\beta})$$

and

$$\log\left(\frac{\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2}{\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2}\right) \geq \log\left(\frac{\varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\varepsilon}{\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2}\right).$$

Since

$$\begin{aligned}
&\varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\varepsilon - \|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2\\
&\quad = \varepsilon^{\mathrm{T}}\mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\varepsilon - \varepsilon^{\mathrm{T}}\mathbf{X}_{A}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{A})^{-1}\mathbf{X}_{A}^{\mathrm{T}}\varepsilon + o_P(\sigma^2)
\end{aligned}$$

it follows that

$$\log\left(\frac{\varepsilon^{\mathrm{T}}\{\mathbf{I} - \mathbf{X}_{\hat{A}}(\mathbf{X}_{\hat{A}}^{\mathrm{T}}\mathbf{X}_{\hat{A}})^{-1}\mathbf{X}_{\hat{A}}^{\mathrm{T}}\}\varepsilon}{\|\mathbf{y} - \mathbf{X}\hat{\beta}^*\|^2}\right) = O_P((\hat{p}_0 - p_0)/n).$$

Thus,

$$\mathrm{MBIC}(\hat{\beta}) - \mathrm{MBIC}(\hat{\beta}^*) \geq (\hat{p}_0 - p_0)(C_n \log(n)/n - O_P(1/n)).$$

We conclude that

$$\lim_{n\to\infty} P\left\{\inf\{\mathrm{MBIC}(\hat{\beta});\, A \subset \hat{A}\} > \mathrm{MBIC}(\hat{\beta}^*)\right\} = 1. \tag{23}$$

Combining this with (22) proves the proposition. □

References

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

Breiman, L. (1995). Better subset regression using the non-negative garrote. *Technometrics*, *37*, 373–384.

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

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, *71*(3), 731–740.

Breheny, P., Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression with applications to biological feature selection. *Annals of Applied Statistics*, *5*(1), 232–253.

Daubechies, I., Defrise, M., De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, *57*, 1413–1457.

Dicker, L., Huang, B., Lin, X. (2013). Variable selection and estimation with the seamless-L_0 penalty. *Statistica Sinica, 23*, 929–962.

Douglas, N. VanDerwerken (2011). Variable selection and parameter estimation using a continuous and differentiable approximation to the L_0 penalty function. All Theses and Dissertations, Paper 2486.

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

Fan, J., Lv, J. (2011). Non-concave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, *57*, 5467–5484.

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

Frank, I. E., Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*, 109–148.

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

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

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

Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, *7*, 397–416.

Kim, Y., Choi, H., Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, *103*, 1665–1673.

Knight, K., Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *Annals of Statistics*, *28*, 1356–1378.

Lee, E. R., Noh, H., Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, *109*, 216–229.

Lv, J., Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, *37*, 3498–3528.

Peng, B., Wang, L. (2014). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, *24*, 00–00.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

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

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., et al. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. *The Journal of Urology*, *16*, 1076–1083.

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

Wang, H., Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, *102*, 1039–1048.

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

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

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

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

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

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B Statistical Methodology*, *67*, 301–320.

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

Zou, H., Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, *37*, 1733–1751.