Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2016, Article ID 6495417, 12 pages
http://dx.doi.org/10.1155/2016/6495417

Research Article

Variable Selection and Parameter Estimation with the Atan Regularization Method

Yanxin Wang[1] and Li Zhu[2]

[1] *School of Science, Ningbo University of Technology, Ningbo 315211, China*
[2] *School of Applied Mathematics, Xiamen University of Technology, Xiamen 361024, China*

Correspondence should be addressed to Yanxin Wang; wyxinbj@163.com

Received 23 September 2015; Revised 4 February 2016; Accepted 7 February 2016

Academic Editor: Steve Su

Variable selection is fundamental to high-dimensional statistical modeling. Many variable selection techniques may be implemented by penalized least squares using various penalty functions. In this paper, an arctangent type penalty which very closely resembles l_0 penalty is proposed; we call it Atan penalty. The Atan-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. The Atan procedure is efficiently implemented using an iteratively reweighted Lasso algorithm. Simulation results and data example show that the Atan procedure with BIC-type criterion performs very well in a variety of settings.

1. Introduction

High-dimensional data arise frequently in modern applications in biology, economics, chemometrics, neuroscience, and other scientific fields. To facilitate the analysis, it is often reasonable and useful to assume that only a small number of covariates are relevant for modeling the response variable. Under this sparsity assumption, a widely used approach for analyzing high-dimensional data is regularized or penalized regression. This approach estimates the unknown regression coefficients by solving the following penalized regression problem:

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\beta\|^2+\sum_{j=1}^{p}p_\lambda\left(\left|\beta_j\right|\right)\right\}, \qquad (1)$$

where $\mathbf{X}=(x_1,x_2,\ldots,x_n)^T$ is $n\times p$ design matrix, $\mathbf{y}=(y_1,y_2,\ldots,y_n)^T$ is an n-dimensional response vector, $\beta=(\beta_1,\ldots,\beta_p)^T$ is the vector of unknown regression coefficients, $\|\cdot\|$ denotes L_2 norm (Euclidean norm), and $p_\lambda(\cdot)$ is a penalty function which depends on a tuning parameter $\lambda>0$.

In the above regularization framework, various penalty functions are used to perform variable selection by putting relatively large penalties on small coefficients, such as the best subset selection, l_1 penalized regression or Lasso [1], Bridge regression [2], SCAD [3], MCP [4], SICA [5], SELO [6], Dantzig selector [7], and Bayesian variable selection method [8]. The mainstream methods are the Lasso, Dantzig selector, and the folded concave penalization [9] such as the SCAD and MCP. The best subset selection, namely, l_0 penalty, along with the traditional model selection criteria such as AIC, BIC, and RIC [10–12] 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 [13]. This results from the fact that l_0 penalty is not continuous at 0. Another perhaps more significant drawback of 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.

The Lasso penalized regression is computationally attractive and enjoys great performance in prediction. However, Lasso may not consistently select the correct model and is not necessarily asymptotically normal [14, 15]; a strong irrepresentable condition is necessary for the Lasso to be selection consistent [15, 16]. The folded concave penalization, unlike the Lasso, does not require the irrepresentable condition to achieve the variable selection consistency and can correct the intrinsic estimation bias of the Lasso. Fan and Li [3] first systematically studied nonconvex penalized likelihood for fixed finite dimension p. They recommended the SCAD penalty which enjoys the oracle property (a variable selection and estimation procedure is said to have the oracle property if it selects the true model A, with probability tending to one, and if the estimated coefficients are asymptotically normal, with the same asymptotic variance as the least squares estimator based on the true model) for variable selection. Fan and Peng [17] extended these results by allowing p to grow with n at the rate $p = o(n^{1/5})$ or $p = o(n^{1/3})$. Lv and Fan [5] introduced the weak oracle property, which means that the estimator enjoys the same sparsity as the oracle estimator with asymptotic probability one and has consistency, and established regularity conditions under which the PLS estimator given by folded concave penalties has a nonasymptotic weak oracle property when dimensionality p can grow nonpolynomially with sample size n. Theoretical properties enjoyed by SELO [6] estimators allow the number of predictors p to tend to infinity, along with the number of observations n, provided $p/n \to 0$. For high-dimensional nonconvex penalized regression with $p > n$, Kim et al. [18] proved that the oracle estimator itself is a local minimum of SCAD penalized least squares regression under very relaxed conditions; Zhang [4] proposed MCP and devised a novel PLUS algorithm which when used together can achieve the oracle property under certain regularity conditions. Recently, Fan et al. [9] have shown that the folded concave penalization methods enjoy the strong oracle property for high-dimensional sparse estimation. Important insight has also been gained through the recent work on theoretical analysis of the global solution [19–21].

The practical performance of PLS procedures depends heavily on the choice of a tuning parameter. The theoretically optimal tuning parameter does not have an explicit representation and depends on unknown factors such as the variance of the unobserved random noise. Cross-validation is commonly adopted in practice to select the tuning parameter but is observed to often result in overfitting. In the case of fixed p, Wang et al. [22] proposed that one selects tuning parameter by minimizing the generalized BIC tuning parameter selector. Wang et al. [23] extended those results to the setting of a diverging number of parameters. Recently, Dicker et al. [6] proposed a BIC-like tuning parameter selector. Wang et al. [21] extended the work of [24, 25] for BIC on high-dimensional least squares regression; they proposed a high-dimensional BIC for a nonconvex penalized solution path.

In this paper, we propose an arctangent type (Atan) penalty function which very closely approximates l_0 penalty. Because the Atan penalty is continuous, the Atan estimator may be more stable than the estimators obtained through l_0 penalized methods. The Atan penalty is a smooth function on $[0, \infty)$ and we use an iteratively reweighted Lasso algorithm. We formally establish the model selection oracle property enjoyed by the Atan estimator. In particular, the asymptotic normality of the Atan is formally established. Our asymptotic framework allows the number of predictors, $p \to \infty$, along with the number of observations n, provided $p/n \to 0$. Furthermore, a BIC-like tuning parameter selection procedure is implemented for Atan.

This paper is organized in the following way. In Section 2, we introduce PLS estimators and give a brief overview of existing nonconvex penalty terms, and the Atan penalty is then presented. Then, we discuss some of its theoretical properties in Section 3. In Section 4, we describe a simple and efficient algorithm for obtaining Atan estimator. Simulation studies and an application of the proposed methodology are presented in Section 5. Conclusions are given in Section 6. The proofs are relegated to the Appendix.

2. The Atan-Penalized Least Squares Method

2.1. Linear Model and Penalized Least Squares. Suppose that $\{(y_i, x_i)\}_{i=1}^{n}$ is a random sample from the linear regression model

$$\mathbf{y} = \mathbf{X}\beta^* + \varepsilon, \tag{2}$$

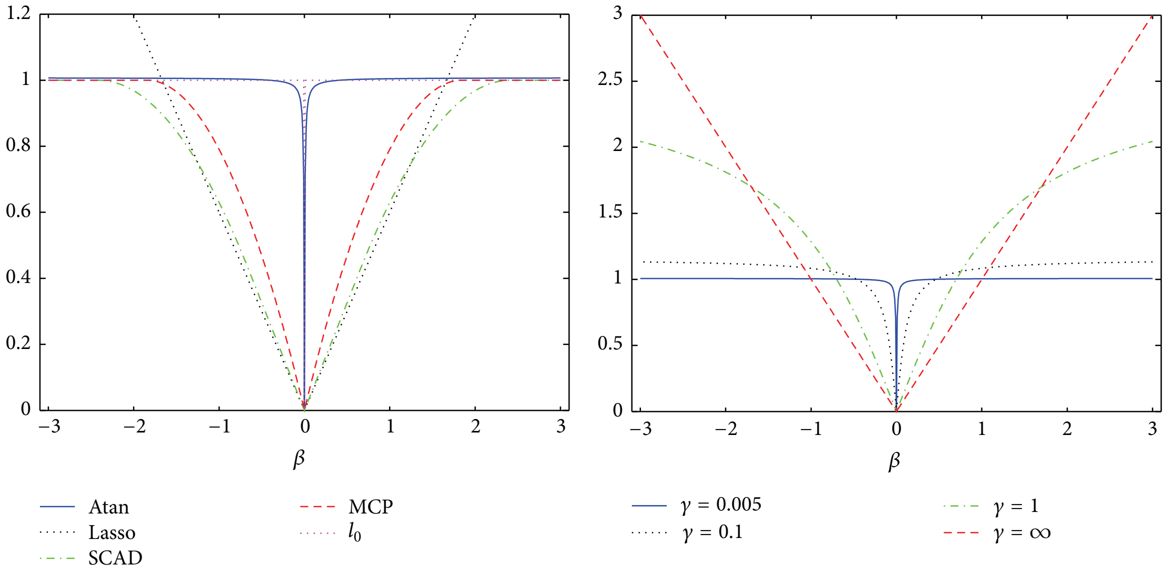
where $\mathbf{X} = (x_1, x_2, \ldots, x_n)^T$ is $n \times p$ design matrix, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is an n-dimensional response vector, and ε are the iid random errors with mean 0 and variance $\sigma^2 n$-dimensional noise vector.

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 subvector of β containing entries indexed by S and let $\mathbf{X}_S$ be $n \times |S|$ matrix obtained from $\mathbf{X}$ by extracting columns corresponding to S. Given $p \times p$ matrix C and subsets $S_1, S_2 \subseteq \{1, 2, \ldots, p\}$, let C_{S_1, S_2} be $|S_1| \times |S_2|$ submatrix of C with rows determined by S_1 and columns determined by S_2.

Various penalty functions have been used in the variable selection literature for linear regression model. Commonly used penalty functions include l_q, $0 \le q \le 2$, the nonnegative garrotte [26], elastic-net [27, 28], SCAD [3], and MCP [4]. In particular, l_1 penalized least squares procedure is called the Lasso. However, Lasso estimates may be biased and inconsistent for model selection [3, 15]. 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 [15]. Slightly abusing notation is that the adaptive Lasso penalty is defined by $p_\lambda(\beta) = \lambda \omega_j |\beta|$, where ω_j is a data-dependent weight.

Fan and Li [3] proposed a continuously differentiable penalty function called the SCAD penalty, which is defined by

$$p_\lambda'(|\beta|) = \lambda \left\{ I(|\beta| \le \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\}, \quad \text{for some } a > 2. \tag{3}$$



(a)

(b)

FIGURE 1: (a) Plots of the penalties (including l_0, Lasso, SCAD, MCP, and Atan). (b) Plots of the Atan penalties with different γ.

Authors of [3] suggested using $a = 3.7$ from a Bayesian perspective. The minimum concave penalty [4] translates the flat part of the derivative of the SCAD into the origin and is given by

$$p'_{\lambda}(|\beta|) = \frac{(a\lambda - |\beta|)_+}{a}, \tag{4}$$

which minimizes the maximum of the concavity. Zhang [4] proved that the MCP procedure may select the correct model with probability tending to 1 and that MCP estimator has good properties in terms of l_p-loss, provided λ and a satisfy certain conditions. Zhang's results in fact allow for $p \gg n$.

2.2. Atan Penalty. In this section we first propose a novel nonconvex penalty that we call Atan penalty. We then study its applications in sparse modeling.

The Atan penalty is defined by

$$p_{\lambda,\gamma}(|\beta|) = \lambda\left(\gamma + \frac{2}{\pi}\right)\arctan\left(\frac{|\beta|}{\gamma}\right), \tag{5}$$

for $\lambda \geq 0$ and $\gamma > 0$. It is clear that this penalty is concave in $|\beta|$. Moreover, we can establish its relationship with l_0 and l_1 penalties. In particular, we have the following propositions.

Proposition 1. *Let $p_{\lambda,\gamma}(|\beta|)$ be given in (5); then*

$$(a)\ \lim_{\gamma\to\infty} p_{\lambda,\gamma}(|\beta|) = \lambda|\beta|,$$

$$(b)\ \lim_{\gamma\to 0} p_{\lambda,\gamma}(|\beta|) = \begin{cases} \lambda, & \text{if } |\beta| \neq 0, \\ 0, & \text{if } |\beta| = 0. \end{cases} \tag{6}$$

The propositions show that the limits of Atan at 0 *and* ∞ *are l_0 penalty and l_1 penalty, respectively. The first-order derivative of $p_{\lambda,\gamma}(|\beta|)$ with respect to $|\beta|$ is*

$$p'_{\lambda,\gamma}(|\beta|) = \lambda\frac{\gamma(\gamma + 2/\pi)}{\gamma^2 + \beta^2}. \tag{7}$$

The Atan penalty function ($\gamma = 0.005$) is plotted in Figure 1(a), along with the SCAD, Lasso, MCP, and l_0 penalty. Figure 1(b) depicts the Atan with different γ.

3. Theoretical Properties

In this section we study the theoretical properties of the Atan estimator proposed in Section 2 in the situation where the number of parameters p tends to ∞ with increasing sample size n. We discuss some conditions of the penalty and loss functions in Section 3.1. Our main results are presented in Section 3.2.

3.1. Regularity Conditions. We need to place the following conditions on the penalty functions:

(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) $\lambda = O(1)$, $\lambda\sqrt{n/(p\sigma^2)} \to \infty$, and $\gamma = O(p^{1/2}\sigma^3 n^{-3/2})$.

(D) There exist constants $C_1, C_2 \in \mathbb{R}$ such that $C_1 < \lambda_{\min}((1/n)\mathbf{X}^T\mathbf{X}) < \lambda_{\max}((1/n)\mathbf{X}^T\mathbf{X}) < C_2$, where $\lambda_{\min}((1/n)\mathbf{X}^T\mathbf{X})$ and $\lambda_{\max}((1/n)\mathbf{X}^T\mathbf{X})$ are the smallest and largest eigenvalues of $(1/n)\mathbf{X}^T\mathbf{X}$, respectively.

(E) $\lim_{n\to\infty} n^{-1}\max_{1\leq i\leq n}\sum_{j=1}^{p} x_{ij}^2 = 0$.

(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 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 substantially weaker than that required in [17], which requires $p^5/n \to 0$, and slightly weaker than that required in [28], which requires $\log(p)/\log(n) \to \nu \in [0,1)$ and the same as that required in [6]. As mentioned in Section 1, other authors have studied PLS methods in settings where $p > n$; that is, their growth condition on p is weaker than condition (A) [18]. Condition (B) gives a lower bound on the size of the smallest nonzero entry of β^*. Notice that the smallest nonzero 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 [17]. Condition (C) restricts the rates of tuning parameters λ and γ. Note that condition (C) does not constrain the minimum size of γ. Indeed, no such constraint is required for our asymptotic results about the Atan estimator. Since the Atan penalty approaches l_0 penalty as $\gamma \to 0$, this suggests that the Atan 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 γ too small, in order to preserve stability of the Atan estimator. Condition (D) is an identifiability condition. Conditions (E) and (F) are used to prove asymptotic normality of Atan estimators and are related to the Lindeberg condition, of the Lindeberg-Feller central limit theorem. Conditions (A)–(F) imply that Atan has the oracle property and may correctly identify model A as we will see in Theorem 4.

3.2. Oracle Properties. Let

$$Q_n(\beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_{\lambda,\gamma}\left(|\beta_j|\right) \tag{8}$$

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

Theorem 2. *Suppose that conditions (A)–(D) hold; then, for every $r \in (0,1)$, there exists a constant $C_0 > 0$ such that*

$$\liminf_{n\to\infty} P\left[\arg\min_{\beta} Q_n(\beta) \subseteq \left\{\beta \in R^p;\ \|\beta - \beta^*\| \le C\sqrt{\frac{p\sigma^2}{n}}\right\}\right] > 1 - r, \tag{9}$$

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

Lemma 3. *Assume that (A)–(D) hold, and fix $C > 0$; then*

$$\lim_{n\to\infty} P\left[\arg\min_{\|\beta-\beta^*\| \le C\sqrt{p\sigma^2/n}} Q_n(\beta) \subseteq \{\beta \in R^p; \beta_{A^c} = 0\}\right] = 1, \tag{10}$$

where $A^c = \{1,\dots,p\} \setminus A$ is the complement of A in $\{1,\dots,p\}$.

Theorem 4 (oracle properties). *Suppose that (A)–(F) hold; then there exists a sequence of $\sqrt{n/p\sigma^2}$-consistent local minima of Atan, $\widehat{\beta}$, such that*

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

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

in distribution, where B_n is any arbitrary $q \times |A|$ matrix such that $B_n B_n^T \to G$, and G is $q \times q$ nonegative symmetric matrix.

4. Implementation

4.1. Iteratively Reweighted Lasso Algorithm. The computation for the Atan-penalized method is much more involved, because the resulting optimization problem is usually nonconvex and has multiple local minimizers. Several algorithms have been developed for computing the folded concave penalized estimators, such as the local quadratic approximation (LQA) algorithm and the local linear approximation (LLA) algorithm [3, 29]. Both LQA and LLA are related to the majorization-minorization (MM) principle [30]. Recently, coordinate descent algorithm was applied to solve the folded concave penalized least squares [31, 32]. Reference [4] devised a PLUS algorithm for solving the PLS using the MCP. Zhang [33] analyzed the capped-l_1 penalty for solving the PLS.

In this section, we present an iteratively reweighted Lasso (IRL) algorithm to solve the following minimization problem:

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\sum_{j=1}^{p}\left(\gamma + \frac{2}{\pi}\right)\arctan\left(\frac{|\beta_j|}{\gamma}\right)\right\}. \tag{11}$$

We show that the solution of the penalty least squares problem can be transformed into that of a series of convex weighted Lasso estimators, to which the existing Lasso algorithms can be efficiently applied.

Now, taking a first-order Taylor-series approximation of the Atan penalty about the current value $\beta_j = \beta_j^*$, we obtain the overapproximation of (11):

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\sum_{j=1}^{p}\frac{\gamma(\gamma + 2/\pi)}{\gamma^2 + \beta_j^{*2}}|\beta_j|\right\}. \tag{12}$$

Note that the linear approximation in (12) is analogous to one proposed in Zou and Li [29] and they argued strongly in favor of a one-step approximation. Instead, we offer the following algorithm for Atan-penalized least squares. Assume all the covariates have been standardized to have mean zero and unit variance. Without loss of generality, the span of regularization parameters is $\lambda_{\min} = \lambda_0 < \cdots < \lambda_L = \lambda_{\max}$, for $l = 0,\dots,L$ [34]. Here, we summarize the details of IRL algorithm as in Algorithm 1, where the active set is defined as $A = \{j \mid \beta_j \neq 0,\ j = 1,2,\dots,p\}$.

In the situation with $n > p$, we set $\lambda_{\min} = 0$. However, when $n < p$, one attains the saturated model long before the regularization parameter reaches zero. In this case, we suggest $\lambda_{\min} = 10^{-4}\lambda_{\max}$ [34], and we used $\tau = 10^{-4}$. In practice, we have found that if the columns of $\mathbf{X}$ are standardized so that $\|\mathbf{x}_j\|^2 = n$, for $j = 1,\dots,p$, then taking $\gamma = 0.005$ works well.

```
(1) Start with $\lambda_{\max} = \max_{1\le j\le p} |\mathbf{X}^T y|/n$. Set $l = L$ and $\beta_{(l)} = 0$.
  Outer loop:
(2) Set $k = 0$ and $\widetilde{\beta}^{(0)} = \beta_{(l)}$;
(3) Increment $k = k + 1$ and $\widetilde{\beta}^{(k+1)} = \widetilde{\beta}^{(k)}$;
(4) Update the weights: $\widetilde{\omega}_j = \gamma(\gamma + 2/\pi)/(\gamma^2 + (\widetilde{\beta}_j^{(k)})^2)$, $j = 1, \ldots, p$;
    Inner loop:
    Solve the Karush-Kuhn-Tucker (KKT) conditions for fixed $\widetilde{\omega}_j$:
    $x_j^T(y - \mathbf{X}\beta) - \lambda_l \widetilde{\omega}_j \operatorname{sgn}(\beta_j) = 0$, if $j \in A$,
    $|x_j^T(y - \mathbf{X}\beta)| < \lambda_l \widetilde{\omega}_j$, if $j \notin A$,
(5) Goto Step (3);
(6) Repeat Steps (3)–(5) until $\|\widetilde{\beta}^{(k+1)} - \widetilde{\beta}^{(k)}\|^2 < \tau$.
    The estimate $\beta_{(l)}$ is the limit point of the outer loop, $\widetilde{\beta}^{\infty}$;
(7) Decrement $l = l - 1$ and $\lambda_l = \lambda_{l-1}$. Return to (2) using $\beta_{(l)}$ as a warm start.
```

Algorithm 1: Iteratively reweighted Lasso (IRL) algorithm.

Remark 5. Algorithm 1 is very much like MM algorithms. Hunter and Li [30] were the first to advocate MM algorithms for variable selection when $n > p$. However, we should notice that Algorithm 1 differs from the algorithm described in [30]. The use of MM algorithms in [30] is to justify a quadratic overapproximation to penalty functions with singularities at the origin. In the inner loop of Algorithm 1, we avoid such approximation by solving the KKT conditions precisely and efficiently [35]. Our use of MM algorithms in Algorithm 1 is to justify the local linear approximation to Atan penalty in the outer loop.

Remark 6. LLA to SCAD penalty were also proposed by Zou and Li [29]. Algorithm 1 differs from that proposed in [29] in that it constructs the entire coefficient path, even when $p > n$, whereas the procedure in [29] computes the coefficient estimates for fixed regularization parameter λ starting with a root-n consistent estimator of true coefficient β^* at the initial step. Moreover, even with the same initial estimate, the limit point from Algorithm 1 will differ from [29] after one iteration.

Remark 7. In the more general case where $p > n$, we used coordinate-wise optimization to compute the entire regularized coefficient path via Algorithm 1. Our experience is that coordinate-wise optimization works well in practice and converges very quickly.

4.2. 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 [10] and BIC [11], 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 [13] and complicated stochastic properties. To overcome the deficiency of traditional methods, Fan and Li [3] proposed the SCAD method, which estimates parameters while simultaneously selecting important variables. They selected tuning parameter by minimizing GCV [1, 3, 26].

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 nonvanishing probability [22, 36]. 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 [22, 37, 38]. The BIC tuning parameter selector is defined by

$$\mathrm{BIC} = \log\left(\frac{\left\|\mathbf{y} - \mathbf{X}\widehat{\beta}\right\|^2}{n}\right) + \widehat{\mathrm{DF}}\frac{\log(n)}{n}, \tag{13}$$

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

$$\widehat{\mathrm{DF}} = \operatorname{tr}\left\{\mathbf{X}\left(\mathbf{X}^T + n\Sigma_\lambda\right)^T \mathbf{X}^T\right\}, \tag{14}$$

and $\Sigma_\lambda = \operatorname{diag}\{p'_\lambda(|\widehat{\beta}_1|)/|\widehat{\beta}_1|, \ldots, p'_\lambda(|\widehat{\beta}_p|)/|\widehat{\beta}_p|\}$. The diagonal elements of Σ_λ are coefficients of quadratic terms in the local quadratic approximation to SCAD penalty function $p_\lambda(\cdot)$ [3].

Dicker et al. [6] proposed BIC-like procedures implemented by minimizing

$$\mathrm{BIC}_0 = \log\left(\frac{\left\|\mathbf{y} - \mathbf{X}\widehat{\beta}\right\|^2}{n - \widehat{p}_0}\right) + \frac{\log(n)}{n}\widehat{p}_0. \tag{15}$$

To estimate the residual variance, they use $\widehat{\sigma}^2 = (n - \widehat{p}_0)^{-1}\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|^2$. This differs from other estimates of the residual variance used in PLS methods, where the denominator $n - \widehat{p}_0$ is replaced by n [22]; here, $n - \widehat{p}_0$ is used to account for degrees of freedom lost to estimation.

More works on the high-dimensional BIC for the least squares regression to tuning parameter selection for nonconvex penalized regression can be seen in [24, 25, 39]. Here we used BIC statistic as (15).

4.3. A Standard Error Formula. The standard errors for the estimated parameters can be obtained directly because we are estimating parameters and selecting variables at the same time. Let $\widehat{\beta} = \widehat{\beta}(\lambda, \gamma)$ be a local minimizer of Atan. Following [3, 17], standard errors of $\widehat{\beta}$ may be estimated by using quadratic approximations to Atan. Indeed, the approximation

$$\begin{aligned} p_{\lambda,a}\left(\left|\beta_j\right|\right) &\approx p_{\lambda,\gamma}\left(\left|\beta_{j0}\right|\right) \\ &\quad + \frac{1}{2\left|\beta_{j0}\right|} p'_{\lambda,\gamma}\left(\left|\beta_{j0}\right|\right)\left(\beta_j^2 - \beta_{j0}^2\right), \\ &\qquad \text{for } \beta_j \approx \beta_{j0}. \end{aligned} \tag{16}$$

suggests that Atan may be replaced by the quadratic minimization problem

$$\min\left\{\frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\beta\right\|^2 + \sum_{j=1}^{p} \frac{p'_{\lambda,a}\left(\left|\beta_{j0}\right|\right)}{\left|\beta_{j0}\right|}\beta_j^2\right\}, \tag{17}$$

at least for the purposes of obtaining standard errors. Using this expression, we obtain a sandwich formula for the estimated standard error of $\widehat{\beta}_{\widehat{A}}$, where $\widehat{A} = \{j;\ \widehat{\beta}_j \neq 0\}$. Consider

$$\begin{aligned} \widehat{\operatorname{cov}}\left(\widehat{\beta}_{\widehat{A}}\right) &= \widehat{\sigma}^2\left\{\mathbf{X}_{\widehat{A}}^T\mathbf{X}_{\widehat{A}} + n\Delta_{\widehat{A},\widehat{A}}\left(\widehat{\beta}\right)\right\}^{-1} \\ &\quad \cdot \mathbf{X}_{\widehat{A}}^T\mathbf{X}_{\widehat{A}}\left\{\mathbf{X}_{\widehat{A}}^T\mathbf{X}_{\widehat{A}} + n\Delta_{\widehat{A},\widehat{A}}\left(\widehat{\beta}\right)\right\}^{-1}, \end{aligned} \tag{18}$$

where $\Delta(\beta) = \operatorname{diag}\{p'_{\lambda,a}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda,a}(|\beta_p|)/|\beta_p|\}$, $\widehat{\sigma}^2 = n^{-1}\|\mathbf{y} - \mathbf{X}\widehat{\beta}\|^2$, and $\widehat{p}_0 = |\widehat{A}|$ is the number of elements in $|\widehat{A}|$. Under the conditions of Theorem 4,

$$\frac{B_n X_{\widehat{A}}^T X_{\widehat{A}} \widehat{\operatorname{cov}}\left(\widehat{\beta}_{\widehat{A}}\right) B_n^T}{\sigma^2} \longrightarrow G. \tag{19}$$

This is a consistency result for $\widehat{\operatorname{cov}}(\widehat{\beta}_{\widehat{A}})$.

5. Numerical Studies

5.1. Simulation Studies. We now investigate the sparsity recovery and estimation properties of the proposed estimator via numerical simulations. We compare the following estimators: the Lasso estimator (implemented using R package glmnet [40]); the adaptive Lasso estimator (denoted by Alasso [15]), the SCAD estimator from the CD algorithm without calibration [31]; the MCP estimator from the CD algorithm with $a = 1.5$ [31]; the Dantzig selector [7]; and Bayesian variable select method [8]. For the proposed Atan estimator, we take $\gamma = 0.005$ and BIC statistic (15) is used to select the tuning parameter λ. In the following, we report simulation results from four examples.

Example 8. In this example, simulation data are generated from the linear regression model:

$$y = \mathbf{x}^T\beta^* + \sigma\varepsilon, \tag{20}$$

where $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T$, $\varepsilon \sim N(0, 1)$, 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, sample size n is set to be 100 and 200, $\sigma = 2$. For each case, we repeated the simulation 200 times.

For linear model, model error for $\widehat{\mu} = \mathbf{x}^T\widehat{\beta}$ is $\mathrm{ME}(\widehat{\mu}) = (\widehat{\beta} - \beta)^T E(\mathbf{x}\mathbf{x}^T)(\widehat{\beta} - \beta)$. 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 unpenalized 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 "underfit," we present the proportion of excluding any nonzero coefficients in 200 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 "overfit," respectively.

As can be seen from Table 1, all variable selection procedures dramatically reduce model error. Atan has the smallest model error among all competitors, followed by Alasso, MCP, SCAD, Bayesian, and Dantzig. In terms of sparsity, Atan also has the highest probability of correct-fit. The Atan penalty performs better than the other penalties. Also, Atan has some advantages when dimensional p is high which can be seen in Example 9.

We now test the accuracy of our standard error formula (18). The median absolute deviation divided by 0.6745, denoted by SD in Table 2, of 3 estimated coefficients in the 200 simulations can be regarded as the true standard error. The median of the 200 estimated SDs, denoted by SD_{m}, and the median absolute deviation error of the 200 estimated standard errors divided by 0.6745, denoted by $\mathrm{SD}_{\mathrm{mad}}$, gauge the overall performance of standard error formula (18). Table 2 presents the results for nonzero coefficients when 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.

Example 9. The example is from Wang et al. [23]. More specifically, we take $\beta = (11/4, -23/6, 37/12, -13/9, 1/3, 0, 0, \ldots, 0)^T \in \mathbb{R}^p$ and $p = [4n^{1/4}] - 5$ and $[t]$ stands for the largest integer no larger than t. For this example, predictor dimension p is diverging but the dimension of the true model is fixed to be 5. Results from the simulation study are found in Table 3. A similar conclusion as in Example 8 can be found.

Example 10. 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 $p = 339$, $n = 500$, $\sigma^5 = 5$, and $\beta^* = (2I_{37}^T, -3I_{37}^T, I_{37}^T, 0_{228}^T)$, where $I^k \in R^k$ is the vector with all entries equal to 1. Thus, $p_0 = 111$. We simulated 200 independent datasets $\{(y_1, x_1^T), \ldots, (y_n, x_n^T)\}$ in this study and, for each dataset, we computed estimates of β^*. Results from this simulation study are found in Table 4.

Table 1: Simulation results for linear regression models of Example 8.

Method	MRME	Number of zeros		Proportion of		
		C	IC	Underfit	Correct-fit	Overfit
			$n = 100, \sigma = 2$			
Lasso	0.6213	3.0000	1.2050	0.0000	0.3700	0.6700
Alasso	0.3074	3.0000	0.3500	0.0000	0.7300	0.2700
SCAD	0.2715	3.0000	1.0650	0.0000	0.4400	0.5600
MCP	0.3041	3.0000	0.5650	0.0000	0.5800	0.4200
Dantzig	0.4623	3.0000	0.6546	0.0000	0.5700	0.4300
Bayesian	0.3548	3.0000	0.5732	0.0000	0.6300	0.3700
Atan	**0.2550**	3.0000	**0.1750**	0.0000	**0.8450**	**0.1550**
			$n = 200, \sigma = 2$			
Lasso	0.6027	3.0000	1.0700	0.0000	0.3550	0.6450
Alasso	0.2781	3.0000	0.1600	0.0000	0.8650	0.1350
SCAD	0.2900	3.0000	0.8550	0.0000	0.5250	0.4750
MCP	0.2752	3.0000	0.3650	0.0000	0.6850	0.3150
Dantzig	0.3863	3.0000	0.8576	0.0000	0.4920	0.5080
Bayesian	0.2563	3.0000	0.4754	0.0000	0.7150	0.2850
Atan	**0.2508**	3.0000	**0.1000**	0.0000	**0.9050**	**0.0950**

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

Method	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
	SD	SD_m (SD_mad)	SD	SD_m (SD_mad)	SD	SD_m (SD_mad)
Lasso	0.1753	0.1453 (0.0100)	0.1730	0.1688 (0.0094)	0.1591	0.1301 (0.0079)
Alasso	0.1483	0.1638 (0.0095)	0.1642	0.1636 (0.0098)	0.1475	0.1439 (0.0073)
SCAD	0.1797	0.1634 (0.0105)	0.1819	0.1608 (0.0104)	0.1398	0.1438 (0.0076)
MCP	0.1602	0.1643 (0.0096)	0.1861	0.1656 (0.0097)	0.1464	0.1435 (0.0069)
Dantzig	0.1734	0.1645 (0.0115)	0.1723	0.1665 (0.0094)	0.1581	0.1538 (0.0074)
Bayesian	0.1568	0.1635 (0.0089)	0.1678	0.1649 (0.0092)	0.1367	0.1375 (0.0078)
Atan	0.1510	0.1643 (0.0108)	0.1591	0.1658 (0.0096)	0.1609	0.1434 (0.0083)

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, Atan selects the most parsimonious models of all methods and has the smallest model error. Atan'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 Atan's. Since $p_0 = 111$, it is clear that Atan 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 nonzero entries in β^* are small relative to the noise level $\sigma^2 = 5$.

Example 11. As an extension of this method, we consider the problem of simultaneous variable selection and estimation in the partially linear model:

$$Y = X'\beta + g(T) + \varepsilon, \qquad (21)$$

where Y is a scalar response variate, X is a p-vector covariate, T is a scalar covariate and takes values in a compact interval (for simplicity, we assume this interval to be $[0, 1]$), β is $p \times 1$ column vector of unknown regression parameter, function $g(\cdot)$ is unknown, and model error ε is independent of (X, T) with mean 0. Traditionally, it has generally been assumed that β is finite dimension; several standard approaches, such as the kernel method, the spline method [41], and the local linear estimation [17], have been proposed.

In this study, we simulate $n = 100, 200$ points T_i, $i = 1, \ldots, 100(200)$, from the uniform distribution on $[0, 1]$. For each i, $e'_{ij}s$ are simulated to be normally distributed with autocorrelated variance structure $AR(\rho)$, such that

$$\operatorname{cov}\left(e_{ij}, e_{il}\right) = \rho^{|i-j|}, \quad 1 \le i,\ j \le 10. \qquad (22)$$

X_{ij}'s are then formed as follows:

$$X_{i1} = \sin\left(2T_i\right) + e_{i1},$$

$$X_{i2} = \left(0.5 + T_i\right)^{-2} + e_{i2},$$

Table 3: Simulation results for linear regression models of Example 9.

Method	MRME	Number of zeros			Proportion of	
		C	IC	Underfit	Correct-fit	Overfit
			$n = 200$, $p = 10$			
Lasso	0.9955	4.9900	2.0100	0.0100	0.1150	0.8750
Alasso	0.5740	4.8650	0.1100	0.1350	0.8150	0.0500
SCAD	0.5659	4.9800	0.5600	0.0200	0.5600	0.4200
MCP	0.6177	4.9100	0.1200	0.0900	0.8250	0.0850
Dantzig	0.6987	4.8900	0.6700	0.1100	0.4360	0.4540
Bayesian	0.5656	4.8650	0.2500	0.1350	0.6340	0.2310
Atan	**0.5447**	4.8900	0.1150	0.1100	**0.8250**	0.0650
			$n = 400$, $p = 12$			
Lasso	1.2197	5.0000	2.0650	0.0000	0.1400	0.8600
Alasso	0.4458	4.9950	0.0900	0.0050	0.9250	0.0700
SCAD	0.4481	5.0000	0.4850	0.0000	0.6350	0.3650
MCP	0.4828	5.0000	0.1150	0.0000	0.8950	0.1050
Dantzig	0.7879	5.0000	0.5670	0.0000	0.3200	0.6800
Bayesian	0.4237	5.0000	0.1800	0.0000	0.7550	0.2450
Atan	**0.4125**	4.9950	**0.0250**	0.0050	**0.9700**	**0.0250**
			$n = 800$, $p = 16$			
Lasso	1.2004	5.0000	2.5700	0.0000	0.0900	0.9100
Alasso	0.3156	5.0000	0.0700	0.0000	0.9300	0.0700
SCAD	0.3219	5.0000	0.6550	0.0000	0.5950	0.4050
MCP	0.3220	5.0000	0.0750	0.0000	0.9300	0.0700
Dantzig	0.5791	5.0000	0.5470	0.0000	0.3400	0.6600
Bayesian	0.3275	5.0000	0.2800	0.0000	0.6750	0.3250
Atan	0.3239	5.0000	0.0750	0.0000	**0.9350**	**0.0650**

Table 4: Simulation results for linear regression models of Example 10.

Method	MRME	Number of zeros			Proportion of	
		C	IC	Underfit	Correct-fit	Overfit
			$n = 500$, $\sigma = 5$			
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
Dantzig	0.3120	108.3400	18.4650	0.7570	0.0000	0.2430
Bayesian	0.2876	104.4350	7.4700	0.9800	0.0000	0.0200
Atan	0.2794	**101.4000**	**4.0600**	1.0000	0.0000	0.0000

$$\begin{aligned}
X_{i3} &= \exp\left(T_i\right) + e_{i3}, \\
X_{i5} &= \left(T_i - 0.7\right)^4 + e_{i5}, \\
X_{i6} &= T_i\left(1 + T_i^2\right)^{-1} + e_{i6}, \\
X_{i5} &= \sqrt{1 + T_i} + e_{i7}, \\
X_{i8} &= \log\left(3T_i + 8\right) + e_{i8}.
\end{aligned} \qquad (23)$$

We investigate the scenario: $p = 10$, $X_{ij} = e_{ij}$, $j = 4, 9, 10$. In the scenario, we have $\beta_j = j$, $1 \leq j \leq 4$, and $\beta_j = 0$ with others, $\varepsilon \sim N(0, 1)$. For each case, we repeated the simulation 200 times. We investigate $g(\cdot)$ functions: $g(t) = \cos(2\pi t)$. For comparison, we apply the different penalties in the parametric component with the B-spline in the nonparametric component.

The results are summarized in Table 5. Columns 2–5 in Table 5 are the averages of the estimates of β_j, $j = 1, \ldots, 4$, respectively. Column 6 is the number of estimates of β_j,

Table 5: Example 11: comparison of estimators.

Estimator	β_1	β_2	β_3	β_4	$\overline{C}$	$\widetilde{C}$	MME (SD)	RASE ($g(T)$)
				$n = 100,\ p = 10$				
SCAD	0.9617	2.0128	2.8839	4.0611	4.7450	5.0000	0.0695 (0.0625)	0.6269
Lasso	0.8278	1.9517	2.7984	3.9274	5.5200	6.0000	0.1727 (0.0790)	0.8079
Dantzig	0.8767	1.9544	2.7894	3.9302	5.4590	6.0000	0.1696 (0.0680)	0.7278
Atan	0.9710	2.0121	2.9865	4.0472	4.9820	5.0000	0.0658 (0.0630)	0.6214
				$n = 200,\ p = 10$				
SCAD	0.9836	1.9989	2.9283	4.0278	5.2000	5.0000	0.0230 (0.0210)	0.3604
Lasso	0.9219	1.9529	2.9124	3.9580	5.1450	5.0000	0.0500 (0.0315)	0.4220
Dantzig	0.9534	1.9675	2.9096	3.9345	5.1460	5.0000	0.0510 (0.0290)	0.4154
Atan	0.9904	1.9946	2.9548	4.0189	5.1250	5.0000	0.0190 (0.0245)	0.3460

$5 \le j \le p$, which are 0, averaged over 200 simulations, and their medians are given in column 7. Model errors are computed as $\mathrm{ME}(\widehat{\mu}) = (\widehat{\beta} - \beta)^T E(\mathbf{x}\mathbf{x}^T)(\widehat{\beta} - \beta)$. Their medians are listed in the 8th column, followed by the model errors standard deviations in parentheses.

The performance of $\widehat{g}(T)$ is assessed by the square root of average squared errors (RASE):

$$\mathrm{RASE}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\widehat{g}\left(T_i\right) - g\left(T_i\right)\right)^2, \tag{24}$$

where $\{T_i,\ i = 1, \ldots, n\}$ are the observed data points at which function g is estimated. Column 9 summarizes the RASE values for the different situations.

We can see the following from Table 5: (1) Comparing with the Lasso and Dantzig estimator, the SCAD and Atan estimators have a smaller model error, which is due to the fact that the SCAD and Atan estimators are unbiased estimators while the Lasso and Dantzig are biased especially for the larger coefficient. Moreover, the Atan and SCAD estimators are more stable. (2) Each method is able to select important variables, but it is obvious that the Atan estimator has slightly stronger sparsity. (3) For the nonparametric component, Atan and SCAD estimator have smaller RASE values.

5.2. Real Data Analysis. In this section, we apply the Atan regularization scheme to a prostate cancer example. The dataset in this example is derived from a study of prostate cancer in [42]. The dataset consists of the medical records of 97 patients who were 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 Atan are all applied to the data. We also compute the OLS estimate of the prostate cancer data. Results are summarized in Table 6. The OLS estimator does not perform variable selection. Lasso selects five variables in the final model; SCAD selects lcavol, lweight, lbph, and svi in the final model, while Alasso, MCP, and Atan select lcavol, lweight, and svi. Thus, Atan selects a substantially simpler model than Lasso, SCAD. Furthermore, as indicated by the columns labeled R^2 (R^2 is equal to one minus the residual sum of squares divided by the total sum of squares) and R^2/R^2_{OLS} in Table 6, the Atan estimator describes more variability in the data than Alasso and MCP and nearly as much as OLS estimator.

Table 6: Prostate cancer data: comparing different methods.

Method	R^2	R^2/R^2_{OLS}	Variables selected
OLS	0.6615	1.0000	All
Lasso	0.5867	0.8870	(1, 2, 4, 5, 8)
Alasso	0.5991	0.9058	(1, 2, 5)
SCAD	0.6140	0.9283	(1, 2, 4, 5)
MCP	0.5999	0.9069	(1, 2, 5)
Atan	0.6057	0.9158	(1, 2, 5)

6. Conclusion and Discussion

In this paper, a new Atan penalty which very closely resembles L_0 penalty is proposed. First, we establish the theory of the Atan estimator under mild conditions. The theory indicates that the Atan-penalized least squares procedure enjoys oracle properties even when the number of variables grows slower than the number of observations. Second, the iteratively reweighted Lasso algorithm makes our proposed estimator implementation fast. Third, we suggest a BIC-like tuning parameter selector to identify the true model consistently. Numerical studies further endorse our theoretical results and the advantage of the Atan estimator for model selection.

We do not address the situation where $p \gg n$ in this paper. In fact, the proposed Atan method can be easily extended for variable selection in the situation of $p \gg n$. Also, as it is shown in Example 11, the Atan method can be applied to semiparametric model and nonparametric model [43, 44]. Furthermore, there is a recent field of applications of variable selection which is to look for impact points in functional data analysis [45, 46]. The possible combination of Atan estimator with the questions in [45, 46] would be interesting. These problems are beyond the scope of this paper and will be interesting topics for future research.

Appendix

Proof of Theorem 2. 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

$$Q_n(\beta^*) < \inf_{\|\mu\|=C} Q_n(\beta^* + \alpha_n\mu) \tag{A.1}$$

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

$$\begin{aligned} D_n(\mu) &= \frac{1}{2n}\left(\alpha_n^2\|\mathbf{X}\mu\|^2 - 2\alpha_n\varepsilon^T\mathbf{X}\mu\right) \\ &\quad + \sum_{j=1}^{p}\left\{p_{\lambda,a}\left(\left|\beta_j^* + \alpha_n\mu_j\right|\right) - p_{\lambda,\gamma}\left(\left|\beta_j^*\right|\right)\right\}. \end{aligned} \tag{A.2}$$

The fact that $p_{\lambda,\gamma}$ is concave on $[0,\infty)$ implies that

$$\begin{aligned} &p_{\lambda,\gamma}\left(\left|\beta_j^* + \alpha_n\mu_j\right|\right) - p_{\lambda,\gamma}\left(\left|\beta_j^*\right|\right) \\ &\quad \geq p'_{\lambda,\gamma}\left(\left|\beta_j^* + \alpha_n\mu_j\right|\right)\left(\left|\beta_j^* + \alpha_n\mu_j\right| - \left|\beta_j^*\right|\right) \\ &\quad \geq p'_{\lambda,\gamma}\left(\left|\beta_j^* + \alpha_n\mu_j\right|\right)\left(-\alpha_n\left|\mu_j\right|\right) \\ &\quad = -\lambda\alpha_n\left|\mu_j\right|\frac{\gamma(\gamma+2/\pi)}{\gamma^2 + \left(\beta_j^* + \alpha_n\mu_j\right)^2}, \end{aligned} \tag{A.3}$$

when n is sufficiently large.

Condition (B) implies that

$$\frac{\gamma(\gamma+2/\pi)}{\gamma^2 + \left(\beta_j^* + \alpha_n\mu_j\right)^2} \leq \frac{\gamma(\gamma+2/\pi)}{\gamma^2+\rho^2}. \tag{A.4}$$

Thus, for n big enough,

$$\begin{aligned} D_n(\mu) &\geq \frac{1}{2n}\left(\alpha_n^2\|\mathbf{X}\mu\|^2 - 2\alpha_n\varepsilon^T\mathbf{X}\mu\right) \\ &\quad - \frac{Cp\lambda\alpha_n\gamma(\gamma+2/\pi)}{\gamma^2+\rho^2}. \end{aligned} \tag{A.5}$$

By (D),

$$\frac{1}{2n}\alpha_n^2\|\mathbf{X}\mu\|^2 \geq \frac{\lambda_{\min}}{2}C^2\alpha_n^2. \tag{A.6}$$

On the other hand (D) implies

$$\frac{1}{n}\alpha_n\left|\varepsilon^T\mathbf{X}\mu\right| \leq \frac{C\alpha_n}{\sqrt{n}}\left\|\frac{1}{\sqrt{n}}\mathbf{X}^T\varepsilon\right\| = O_P\left(C\alpha_n^2\right). \tag{A.7}$$

Furthermore, (C) and (B) imply

$$\frac{Cp\lambda\alpha_n\gamma(\gamma+2/\pi)}{\gamma^2+\rho^2} = o\left(C\alpha_n^2\right). \tag{A.8}$$

From (A.5)–(A.8), 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 Theorem 2. □

Proof of Lemma 3. Suppose that $\beta \in R^p$ and that $\|\beta - \beta^*\| \leq C\sqrt{p\sigma^2/n}$. Define $\widetilde{\beta} \in R^p$ by $\widetilde{\beta}_{A^c} = 0$ and $\widetilde{\beta}_A = \beta_A$. Similar to the proof of Theorem 2, let

$$D_n\left(\beta, \widetilde{\beta}\right) = Q_n(\beta) - Q_n\left(\widetilde{\beta}\right), \tag{A.9}$$

where $Q_n(\beta)$ is defined in (8). Then

$$\begin{aligned} D_n\left(\beta,\widetilde{\beta}\right) &= \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 - \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\widetilde{\beta}\right\|^2 \\ &\quad + \sum_{j\in A^c} p_{\lambda,\gamma}\left(\left|\beta_j\right|\right) \\ &= \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\widetilde{\beta} - \mathbf{X}\left(\beta - \widetilde{\beta}\right)\right\|^2 \\ &\quad - \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}\widetilde{\beta}\right\|^2 + \sum_{j\in A^c} p_{\lambda,\gamma}\left(\left|\beta_j\right|\right) \\ &= \frac{1}{2n}\left(\beta - \widetilde{\beta}\right)^T\mathbf{X}^T\mathbf{X}\left(\beta - \widetilde{\beta}\right) \\ &\quad - \frac{1}{2n}\left(\beta - \widetilde{\beta}\right)^T\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\widetilde{\beta}\right) \\ &\quad + \sum_{j\in A^c} p_{\lambda,\gamma}\left(\left|\beta_j\right|\right) \\ &= O_P\left(\left\|\beta - \widetilde{\beta}\right\|\sqrt{\frac{p\sigma^2}{n}}\right) \\ &\quad + \sum_{j\in A^c} p_{\lambda,\gamma}\left(\left|\beta_j\right|\right). \end{aligned} \tag{A.10}$$

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

$$p_{\lambda,\gamma}\left(\left|\beta_j\right|\right) \geq \lambda\left(\gamma + \frac{2}{\pi}\right)\arctan\left(\frac{C}{\gamma\sqrt{n/p\sigma^2}}\right)\left|\beta_j\right|, \tag{A.11}$$

for $j \in A^c$. Thus,

$$\begin{aligned} &\sum_{j\in A^c} p_{\lambda,\gamma}\left(\left|\beta_j\right|\right) \\ &\quad \geq \lambda\left(\gamma + \frac{2}{\pi}\right)\arctan\left(\frac{C}{\gamma\sqrt{n/p\sigma^2}}\right)\left\|\beta - \widetilde{\beta}\right\|. \end{aligned} \tag{A.12}$$

By (C),

$$\liminf_{n\to\infty}\arctan\left(\frac{C}{\gamma\sqrt{n/p\sigma^2}}\right) > 0. \tag{A.13}$$

It follows from (A.10)–(A.12) that there is constant $K > 0$ such that

$$\frac{D_n\left(\beta,\widetilde{\beta}\right)}{\left\|\beta - \widetilde{\beta}\right\|} \geq K\lambda + O_P\left(\sqrt{\frac{p\sigma^2}{n}}\right). \tag{A.14}$$

Since $\lambda\sqrt{p\sigma^2/n} \to \infty$, the result follows. □

Proof of Theorem 4. Taken together, Theorem 2 and Lemma 3 imply that there exists a sequence of local minima $\widehat{\beta}$ of (8) such that $\|\widehat{\beta}-\beta^*\| = O_P(\sqrt{p\sigma^2/n})$ and $\widehat{\beta}_{A^c} = 0$. Part (i) of the theorem follows immediately.

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

$$\widehat{\beta}_A = \beta_A^* + \left(\mathbf{X}_A^T\mathbf{X}_A\right)^{-1}\mathbf{X}_A^T\varepsilon - \left(n^{-1}\mathbf{X}_A^T\mathbf{X}_A\right)^{-1} p_A', \tag{A.15}$$

where $p_A' = (p_{\lambda,\gamma}'(\widehat{\beta}_j))_{j\in A}$. It follows that

$$\begin{aligned}\sqrt{n}B_n\left(\frac{n_{-1}\mathbf{X}_A^T\mathbf{X}_A}{\sigma^2}\right)^{1/2}\left(\widehat{\beta}_A-\beta_A^*\right) &= B_n\left(\sigma^2\mathbf{X}_A^T\mathbf{X}_A\right)^{-1/2}\mathbf{X}_A^T\varepsilon \\ &\quad - nB_n\left(\sigma^2\mathbf{X}_A^T\mathbf{X}_A\right)^{-1/2}p_A',\end{aligned} \tag{A.16}$$

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

$$\begin{aligned}&\left\|nB_n\left(\sigma^2\mathbf{X}_A^T\mathbf{X}_A\right)^{-1/2}p_A'\right\| \\ &\quad = O_P\left(\sqrt{\frac{np}{\sigma^2}}\frac{\lambda\gamma\left(\gamma+2/\pi\right)}{\gamma^2+\rho^2}\right) = o_P\left(1\right),\end{aligned} \tag{A.17}$$

and, thus,

$$\begin{aligned}&\sqrt{n}B_n\left(\frac{n_{-1}\mathbf{X}_A^T\mathbf{X}_A}{\sigma^2}\right)^{1/2}\left(\widehat{\beta}_A-\beta_A^*\right) \\ &\quad = B_n\left(\sigma^2\mathbf{X}_A^T\mathbf{X}_A\right)^{-1/2}\mathbf{X}_A^T\varepsilon + o_P\left(1\right).\end{aligned} \tag{A.18}$$

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

$$B_n\left(\sigma^2\mathbf{X}_A^T\mathbf{X}_A\right)^{-1/2}\mathbf{X}_A^T\varepsilon \longrightarrow N\left(0,G\right) \tag{A.19}$$

in distribution. Observe that

$$B_n\left(\sigma^2\mathbf{X}_A^T\mathbf{X}_A\right)^{-1/2}\mathbf{X}_A^T\varepsilon = \sum_{i=1}^n\omega_{i,n}, \tag{A.20}$$

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

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

$$\begin{aligned}E\left[\left\|\omega_{i,n}\right\|^2;\left\|\omega_{i,n}\right\|^2>\delta_0\right] &= \eta_{i,n}E\left[\frac{\varepsilon_i^2}{\sigma^2};\frac{\eta_{i,n}\varepsilon_i^2}{\sigma^2}>\delta_0\right] \\ &\le \eta_{i,n}E\left(\left|\frac{\varepsilon_i}{\sigma}\right|^{2+\delta}\right)^{2/(2+\delta)}P\left(\frac{\eta_{i,n}\varepsilon_i^2}{\sigma^2}>\delta_0\right)^{\delta/(2+\delta)} \\ &\le \eta_{i,n}^{1+\delta/2+\delta}\delta_0^{-1}E\left(\left|\frac{\varepsilon_i}{\sigma}\right|^{2+\delta}\right)^{2/(2+\delta)}.\end{aligned} \tag{A.21}$$

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

$$\begin{aligned}&\max_{1\le i\le n}\eta_{i,n}\lambda_{\min}\left(n^{-1}\mathbf{X}^T\mathbf{X}\right)\lambda_{\max}\left(B_n^TB_n\right)\max_{1\le i\le n}\frac{1}{n}\sum_{j=1}^p x_{ij}^2 \\ &\quad \longrightarrow 0,\end{aligned} \tag{A.22}$$

we must have

$$\begin{aligned}&\sum_{i=1}^n E\left[\left\|\omega_{i,n}\right\|^2;\left\|\omega_{i,n}\right\|^2>\delta_0\right] \\ &\quad = \delta_0^{-1}E\left(\left|\frac{\varepsilon_i}{\sigma}\right|^{2+\delta}\right)^{2/(2+\delta)}\sum_{i=1}^n\eta_{i,n}^{1+\delta/(2+\delta)} \\ &\quad \le \delta_0^{-1}E\left(\left|\frac{\varepsilon_i}{\sigma}\right|^{2+\delta}\right)^{2/(2+\delta)}\mathrm{tr}\left(B_n^TB_n\right)\max_{1\le i\le n}\eta_{i,n}^{\delta/(2+\delta)} \\ &\quad \longrightarrow 0.\end{aligned} \tag{A.23}$$

Thus, the Lindeberg condition is satisfied and (A.19) holds. □

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Tian Yuan Special Funds of the National Natural Science Foundation of China (Grant no. 11426192) and the Project of Education of Zhejiang Province (no. Y201533324).

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[2] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–148, 1993.

[3] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[4] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[5] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, vol. 37, no. 6, pp. 3498–3528, 2009.

[6] L. Dicker, B. Huang, and X. Lin, "Variable selection and estimation with the seamless-L_0 penalty," *Statistica Sinica*, vol. 23, no. 2, pp. 929–962, 2013.

[7] E. Candes and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[8] J. Ghosh and A. E. Ghattas, "Bayesian variable selection under collinearity," *The American Statistician*, vol. 69, no. 3, pp. 165–173, 2015.

[9] J. Fan, L. Z. Xue, and H. Zou, "Strong oracle optimality of folded concave penalized estimation," *Annals of Statistics*, vol. 42, no. 3, pp. 819–849, 2014.

[10] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki, Eds., pp. 267–281, Akademiai Kiado, Budapest, Hungary, 1973.

[11] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[12] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *The Annals of Statistics*, vol. 22, no. 4, pp. 1947–1975, 1994.

[13] L. Breiman, "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

[14] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[15] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[16] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[17] J. Fan and H. Peng, "Nonconcave penalized likehood with a diverging number parameters," *Annals of Statistics*, vol. 32, pp. 928–961, 2004.

[18] Y. Kim, H. Choi, and H.-S. Oh, "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1665–1673, 2008.

[19] Y. Kim and S. Kwon, "Global optimality of nonconvex penalized estimators," *Biometrika*, vol. 99, no. 2, pp. 315–325, 2012.

[20] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576–593, 2012.

[21] L. Wang, Y. Kim, and R. Li, "Calibrating nonconvex penalized regression in ultra-high dimension," *The Annals of Statistics*, vol. 41, no. 5, pp. 2505–2536, 2013.

[22] H. Wang, R. Li, and C.-L. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.

[23] H. Wang, B. Li, and C. Leng, "Shrinkage tuning parameter selection with a diverging number of parameters," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 71, no. 3, pp. 671–683, 2009.

[24] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[25] Y. Kim, S. Kwon, and H. Choi, "Consistent model selection criteria on high dimensions," *Journal of Machine Learning Research*, vol. 13, pp. 1037–1057, 2012.

[26] L. Breiman, "Better subset regression using the non-negative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[28] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *The Annals of Statistics*, vol. 37, no. 4, pp. 1733–1751, 2009.

[29] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.

[30] D. R. Hunter and R. Li, "Variable selection using MM algorithms," *The Annals of Statistics*, vol. 33, no. 4, pp. 1617–1642, 2005.

[31] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The Annals of Applied Statistics*, vol. 5, no. 1, pp. 232–253, 2011.

[32] J. Fan and J. Lv, "Nonconcave penalized likelihood with NP-dimensionality," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5467–5484, 2011.

[33] T. Zhang, "Multi-stage convex relaxation for feature selection," *Bernoulli*, vol. 19, no. 5, pp. 2277–2293, 2013.

[34] J. H. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[35] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[36] J. Shao, "Linear model selection by cross-validation," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[37] H. Wang and C. Leng, "Unified LASSO estimation by least squares approximation," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 1039–1048, 2007.

[38] H. Zou and T. Hastie, "On the 'degrees of freedom' of lasso," *Annals of Statistics*, vol. 35, pp. 2173–2192, 2007.

[39] E. R. Lee, H. Noh, and B. U. Park, "Model selection via Bayesian information criterion for quantile regression models," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 216–229, 2014.

[40] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[41] N. E. Heckman, "Spline smoothing in a partly linear model," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 48, no. 2, pp. 244–248, 1986.

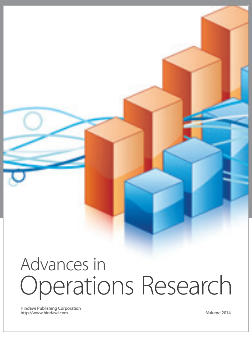
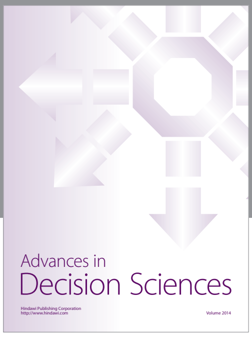
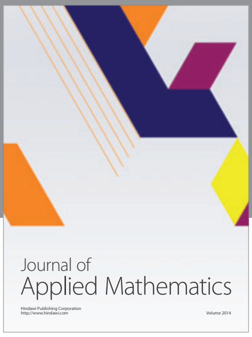
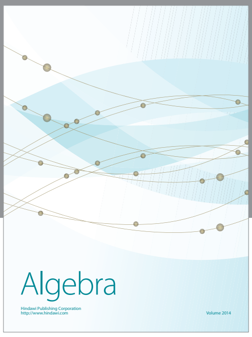
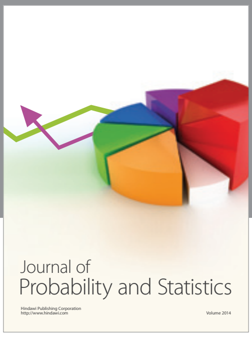
[42] T. A. Stamey, J. N. Kabalin, J. E. McNeal et al., "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients," *The Journal of Urology*, vol. 141, no. 5, pp. 1076–1083, 1989.

[43] J. Huang, J. L. Horowitz, and F. Wei, "Variable selection in nonparametric additive models," *The Annals of Statistics*, vol. 38, no. 4, pp. 2282–2313, 2010.

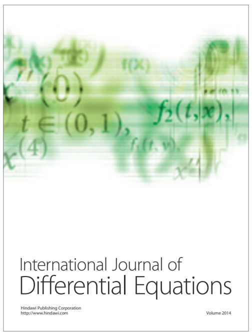
[44] P. Radchenko, "High dimensional single index models," *Journal of Multivariate Analysis*, vol. 139, pp. 266–282, 2015.

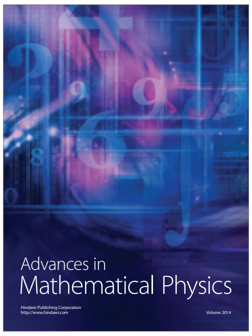
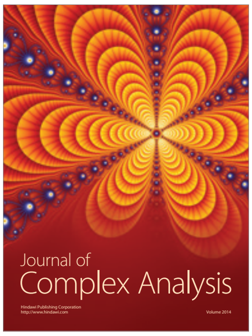
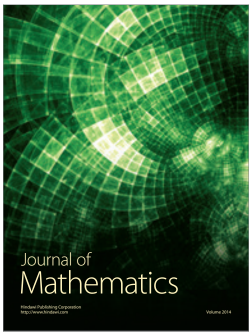
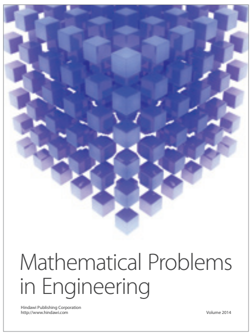
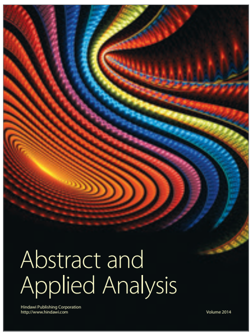
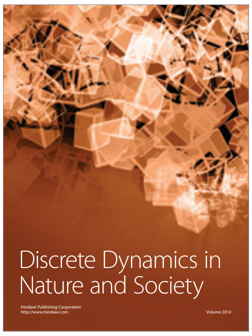
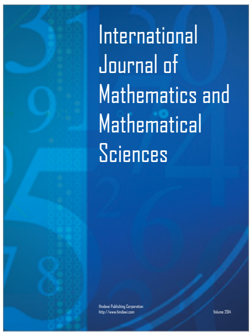
[45] G. Aneiros and P. Vieu, "Variable selection in infinite-dimensional problems," *Statistics & Probability Letters*, vol. 94, pp. 12–20, 2014.

[46] G. Aneiros and P. Vieu, "Partial linear modelling with multi-functional covariates," *Computational Statistics*, vol. 30, no. 3, pp. 647–671, 2015.

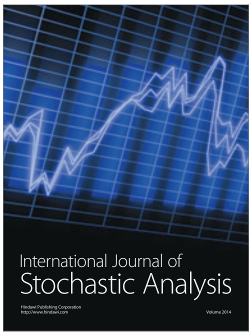

Advances in
Operations Research

Advances in
Decision Sciences

Journal of
Applied Mathematics

Algebra

Journal of
Probability and Statistics

The Scientific
World Journal


International Journal of
Differential Equations

International Journal of
Combinatorics

Advances in
Mathematical Physics

Journal of
Complex Analysis

Journal of
Mathematics

Mathematical Problems
in Engineering

Abstract and
Applied Analysis

Discrete Dynamics in
Nature and Society

International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces

International Journal of
Stochastic Analysis

Journal of
Optimization