Journal of Mathematical Research with Applications
Nov., 2014, Vol. 34, No. 6, pp. 743–753
DOI:10.3770/j.issn:2095-2651.2014.06.013
Http://jmre.dlut.edu.cn

Combination of Nonconvex Penalties and Ridge Regression for High-Dimensional Linear Models

Xiuli WANG, Mingqiu WANG*
School of Mathematical Sciences, Qufu Normal University, Shandong 273165*, P. R. China*

Abstract Nonconvex penalties including the smoothly clipped absolute deviation penalty and the minimax concave penalty enjoy the properties of unbiasedness, continuity and sparsity, and the ridge regression can deal with the collinearity problem. Combining the strengths of nonconvex penalties and ridge regression (abbreviated as NPR), we study the oracle property of the NPR estimator in high dimensional settings with highly correlated predictors, where the dimensionality of covariates p_n is allowed to increase exponentially with the sample size n. Simulation studies and a real data example are presented to verify the performance of the NPR method.

Keywords high dimension; nonconvex penalties; oracle property; ridge regression; variable selection.

MR(2010) Subject Classification 62F12; 62J05; 62J07

1. Introduction

High dimensional data are frequently encountered in modern applications [1–3]. A common feature of high dimensional data is that the dimensionality of covariates p_n is large compared with the number of observations n. In addition, the collinearity problem is unavoidable in high dimensional data analysis. For example, a typical microarray data has many thousands of predictors and often fewer than 100 samples. The correlations between those genes sharing the same biological 'pathway' can be high [4]. As pointed in [5,6], the ridge regression can overcome the collinearity with better prediction performance.

Much progress has been made to overcome these challenges simultaneously. Elastic net (Enet) proposed by [7] is the first try to understand variable selection with highly correlated variables. Yuan and Lin [8] gave a necessary and sufficient condition for the Enet to select the true model in the classical setting when p_n is fixed. Jia and Yu [9] demonstrated that the Enet is selection consistent under an Elastic Irrepresentable Condition (EIC) and certain other conditions in the case of $p_n \gg n$. Their results generalize those of [10] on the selection consistency of the Lasso. However, the Enet estimator cannot obtain the selection consistency and prediction

Received January 20, 2014; Accepted March 23, 2014
Supported by the National Natural Science Foundation of China (Grant No. 11401340), China Postdoctoral Science Foundation (Grant No. 2014M561892) and the Foundation of Qufu Normal University (Grant Nos. bsqd2012041; xkj201304).
* Corresponding author
E-mail address: wang.sawh@gmail.com (Xiuli WANG); wmq0829@gmail.com (Mingqiu WANG)

accuracy simultaneously due to the L_1 part [11,12]. Zuo and Zhang [13] proposed the adaptive Enet estimator, which is oracle under some regular conditions. However, they assumed that the design matrix is nonsingular, which excludes the case of $p_n > n$. Due to the unbiasedness, sparsity and continuity properties of the smoothly clipped absolute deviation (SCAD) proposed in [14], several authors studied the selection consistency by combining the SCAD and L_2 penalties [15,16]. However, they did not give the theoretical properties when p_n is much larger than n. Furthermore, they only gave the asymptotic results of a local minimizer due to the non-concave of the SCAD. Kim, Choi and Oh [17] studied the asymptotic properties of the SCAD estimator when $p_n > n$, but it will perform bad when predictors are highly correlated. Huang, Ma, Li and et al [18] studied the minimax concave penalty (MCP)[19] to deal with the collinearity problem.

In this paper, based on the definitions of SCAD and MCP, we propose a class of nonconvex penalties, and study the theoretical properties of the combination of nonconvex penalties and ridge regression (abbreviated as NPR) for high-dimensional linear models. Under certain conditions, the oracle ridge estimator given in subsection 2.1 is a local minimizer of the penalized objective function (2.1) even when the number of parameters is much larger than the sample size. In addition, the oracle ridge estimator becomes the global minimizer of (2.1) asymptotically in subsection 2.2. That is, this method can correctly select predictors with nonzero coefficients and estimate the selected coefficients using the ridge regression. Simulation studies and a real data indicate that the NPR works much better than the nonconvex penalties in high dimensionality and highly correlated.

The rest of the article is organized as follows. Section 2 gives the theoretical properties of the NPR estimator. In Section 3, we compare the finite sample performance of the NPR estimate by simulation studies. A real data example is analyzed to illustrate the application of our method in Section 4. Some concluding remarks are given in Section 5.

2. Theoretical properties

Consider the linear regression model

$$Y_i = \beta_0 + \sum_{j=1}^{p_n} X_{ij}\beta_j + \varepsilon_i, \quad i = 1, \ldots, n,$$

where Y_i is the response variable, X_{ij}'s are covariates or design variables and ε_i's are independent and identically distributed random error terms. Without loss of generality, let $\beta_0 = 0$ which can be obtained by centering the response and covariates. Throughout this paper, we assume that the outcome is centered and the predictors are standardized; that is,

$$\sum_{i=1}^{n} Y_i = 0, \ \sum_{i=1}^{n} X_{ij} = 0 \text{ and } \frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1, j = 1, \ldots, p_n.$$

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})^T$ and $\mathbf{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{p_n})$. To indicate the dependence of parameters on the sample size, denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^T$ by $\boldsymbol{\beta}_n$.

Without loss of generality, let the true parameter be $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$,

where $\boldsymbol{\beta}_{10} \neq \mathbf{0}$ and $\boldsymbol{\beta}_{20} = \mathbf{0}$. Denote $J_{n1} = \{j : \beta_{0j} \neq 0\}$ and $b_{n1} = \min_{j \in J_{n1}} |\beta_{0j}|$. The cardinality of J_{n1} is denoted by $|J_{n1}|$ and $|J_{n1}| = k_n$. $\mathbf{X}_1 = (\boldsymbol{X}_j,\ j \in J_{n1})$ and $\mathbf{X}_2 = (\boldsymbol{X}_j,\ j \notin J_{n1})$. Define $\Sigma_{11} = n^{-1}\mathbf{X}_1^T\mathbf{X}_1$ with the smallest and largest eigenvalues being τ_{n1} and τ_{n2}.

Combining the nonconvex penalties and ridge regression, we consider the following penalized objective function

$$Q_n(\boldsymbol{\beta}_n) = \frac{1}{2n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_n\|^2 + \sum_{j=1}^{p_n} p_{\lambda_1}(|\beta_{nj}|) + \frac{1}{2}\lambda_2 \sum_{j=1}^{p_n} \beta_{nj}^2, \tag{2.1}$$

where $p_\lambda(\cdot)$ belongs to a class of nonconvex penalties which satisfy the following conditions: (a) Let $p'_\lambda(\cdot)$ be nonnegative, nonincreasing and continuous over $(0, \infty)$; (b) There exists a constant $a > 0$ such that $\lim_{t \to 0^+} p'_\lambda(t) = \lambda$ and $p'_\lambda(t) \geqslant \lambda - t/a$ $(0 < t < a\lambda)$; (c) $p_\lambda(t) = O(\lambda^2)$ $(t \geqslant a\lambda)$. This class includes the SCAD and MCP as special cases.

The value $\hat{\boldsymbol{\beta}}_n$ that minimizes (2.1) is called the NPR estimator. λ_1 and λ_2 are tuning parameters. Note that both of them are dependent on n, and we write them without n when there is no confusion. Let

$$\hat{\boldsymbol{\beta}}_n^0 = \arg\min_{\boldsymbol{\beta}_n} \Big\{ \frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_n\|^2 + \lambda_2 \sum_{j=1}^{p_n} \beta_{nj}^2,\ \beta_{nj} = 0 \text{ for } j \notin J_{n1} \Big\}.$$

Then $\hat{\boldsymbol{\beta}}_n^0$ is the oracle ridge estimator, which only contains all important variables. In practice, the oracle set with nonzero coefficients cannot be known in advance. Here, we use the oracle ridge estimator to facilitate the derivation about the oracle property of the NPR estimator.

2.1. Asymptotic properties of the oracle ridge estimator

In this section, we give sufficient conditions such that the oracle ridge estimator is a local minimizer of the objective function (2.1) even when the number of parameters grows at an exponential rate of the sample size. We now state the following conditions.

(A1) ε_i's are independent and identically distributed random variables with $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. For a constant $1 \leqslant d \leqslant 2$, the tail probabilities of ε_i satisfy $\Pr(|\varepsilon_i| > t) \leqslant K \exp\{-Ct^d\}$ for all $t > 0$ and $i = 1, \ldots, n$, where C and K are positive constants.

(A2) There exist constants $C_1 > 0$ and $C_2 > 0$ such that $C_1 \leqslant \tau_{n1} \leqslant \tau_{n2} \leqslant C_2$ for all n.

(A3) $(\log n)^{I(d=1)}[\frac{(\log k_n)^{\frac{1}{d}}}{\sqrt{n}(C_1+\lambda_2)(b_{n1}-a\lambda_1)} + \frac{(\log p_n)^{\frac{1}{d}}}{\sqrt{n}\lambda_1}] \to 0$.

(A4) For any constant $M > 0$, $M\frac{\lambda_2}{C_1+\lambda_2}\|\boldsymbol{\beta}_{10}\| < b_{n1} - a\lambda_1$.

(A5) $\frac{\lambda_2}{C_1+\lambda_2}\sqrt{C_2}\|\boldsymbol{\beta}_{10}\| < \frac{\lambda_1}{2}$.

Condition (A1) is commonly used in linear regression models ([20]). Condition (A2) implies that the matrix Σ_{11} is nonsingular, which is regular for the true model. Although we consider the collinearity among all variables, the correlations between important variables are small. Now, we discuss the rationality of conditions (A3–A5). Condition (A3) states the number of coefficients. Condition (A4) requires that the minimum of nonzero coefficients should not be too small to be identified from the initial model. Condition (A5) assumes that λ_1 should be at least in the same order as λ_2. Similar to [10], suppose $k_n = O(n^{d_1})$ for some $0 < d_1 < 1$ and

$b_{n1} \geqslant Mn^{(d_2-1)/2}$ for some $d_1 < d_2 < 1$. We choose $\lambda_1 = O(n^{(d_3-1)/2})$, with $0 < d_3 < d_2 - d_1$. If $\lambda_2 = o(n^{(d_3-d_1-1)/2})$, then conditions (A4) and (A5) hold. For condition (A3), if $d = 2$, i.e., the tails of the error distributions are sub-Gaussian, then the number of parameters p_n can be as large as $\exp(o(n^{d_3}))$, which can be much larger than the sample size n.

We first give a lemma that will be used in the proof of Theorem 2.2.

Lemma 2.1 ([20]) *Under condition (A1), for all constants a_i satisfying $\sum_{i=1}^n a_i^2 = 1$, we have $\Pr(\sum_{i=1}^n a_i\varepsilon_i > t) \leqslant \exp\{-t^d/M(1+\log n)^{I(d=1)}\}$ for certain M depending on $\{d, K, C\}$.*

Theorem 2.2 *Let $A_n(\lambda_1, \lambda_2)$ be the set of local minimizer of (2.1). Under conditions (A1–A5), we have $\Pr(\hat{\boldsymbol{\beta}}_n^0 \in A_n(\lambda_1, \lambda_2)) \to 1$.*

Proof By Karush-Kunh-Tucker (KKT) conditions (see, e.g., Bertsekas 1999 [21], p. 320), $\boldsymbol{\beta}_n$ is the solution of (2.1) if

$$-\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_n) + \lambda_2\beta_{nj} + p'_{\lambda_1}(|\beta_{nj}|)\mathrm{sgn}(\beta_{nj}) = 0, \quad \beta_{nj} \neq 0,$$
$$|\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_n)| \leqslant \lambda_1, \quad \beta_{nj} = 0. \tag{2.2}$$

Thus it suffices to show that $\hat{\boldsymbol{\beta}}_n^0$ satisfies (2.2). By the definition of $\hat{\boldsymbol{\beta}}_n^0$, we only need to show

$$|\hat{\beta}_{nj}^0| \geqslant a\lambda_1, \quad j \in J_{n1},$$
$$|\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^0)| \leqslant \lambda_1, \quad j \notin J_{n1}.$$

Let $\hat{\boldsymbol{\beta}}_{1n}^0 = (\hat{\beta}_{nj}^0,\ j \in J_{n1})^T$. It is easy to get $\hat{\boldsymbol{\beta}}_{1n}^0 = (\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\Sigma_{11}\boldsymbol{\beta}_{10}+\frac{1}{n}(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\boldsymbol{\varepsilon}$, where $\mathbf{I}_{11}$ is a $k_n \times k_n$ identity matrix. Thus, for $j \in J_{n1}$, we have

$$\hat{\beta}_{nj}^0 = \boldsymbol{e}_j^T(\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\Sigma_{11}\boldsymbol{\beta}_{10} + \frac{1}{n}\boldsymbol{e}_j^T(\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\boldsymbol{\varepsilon},$$
$$\hat{\beta}_{nj}^0 - \beta_{0j} = \frac{1}{n}\boldsymbol{e}_j^T(\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\boldsymbol{\varepsilon} - \lambda_2\boldsymbol{e}_j^T(\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\boldsymbol{\beta}_{10}, \tag{2.3}$$

where $\boldsymbol{e}_j$ is a unit vector in the direction of the j-th coordinate. For $j \notin J_{n1}$, we have

$$\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^0) = \frac{1}{n}\boldsymbol{X}_j^T\mathbf{X}_1\left[\mathbf{I}_{11} - (\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\Sigma_{11}\right]\boldsymbol{\beta}_{10} +$$
$$\frac{1}{n}\boldsymbol{X}_j^T\left[\mathbf{I}_n - \frac{1}{n}\mathbf{X}_1(\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\right]\boldsymbol{\varepsilon}. \tag{2.4}$$

Here $\mathbf{I}_n$ is an $n \times n$ unit matrix. Note that $|\hat{\beta}_{nj}^0| \geqslant b_{n1} - \max_{j\in J_{n1}}|\hat{\beta}_{nj}^0 - \beta_{0j}|$, for $j \in J_{n1}$. Hence

$$\Pr\left(\hat{\boldsymbol{\beta}}_n^0 \notin A_n(\lambda_1, \lambda_2)\right) \leqslant \Pr\left(\exists j \in J_{n1},\ |\hat{\beta}_{nj}^0| < a\lambda_1\right) + \Pr\left(\exists j \notin J_{n1},\ |\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^0)| > \lambda_1\right)$$
$$\leqslant \Pr\left(\exists j \in J_{n1}, \max_{j\in J_{n1}}|\hat{\beta}_{nj}^0 - \beta_{0j}| > b_{n1} - a\lambda_1\right) +$$
$$\Pr\left(\exists j \notin J_{n1},\ |\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^0)| > \lambda_1\right).$$

Combining (2.3) and (2.4), we obtain

$$\Pr\left(\hat{\boldsymbol{\beta}}_n^0 \notin A_n(\lambda_1, \lambda_2)\right) \leqslant \Pr\left(\exists j \in J_{n1},\ |\frac{1}{n}\boldsymbol{e}_j^T(\Sigma_{11} + \lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\boldsymbol{\varepsilon}| > \frac{1}{2}\left(b_{n1} - a\lambda_1\right)\right) +$$

$$\begin{aligned}
&\Pr\Big(\exists j \in J_{n1},\ |\lambda_2 \boldsymbol{e}_j^T(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\boldsymbol{\beta}_{10}| > \frac{1}{2}(b_{n1}-a\lambda_1)\Big)+\\
&\Pr\Big(\exists j \notin J_{n1},\ |\frac{1}{n}\boldsymbol{X}_j^T\mathbf{X}_1(\mathbf{I}_{11}-(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\Sigma_{11})\boldsymbol{\beta}_{10}| > \frac{\lambda_1}{2}\Big)+\\
&\Pr\Big(\exists j \notin J_{n1},\ |\frac{1}{n}\boldsymbol{X}_j^T[\mathbf{I}_n-\frac{1}{n}\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T]\boldsymbol{\varepsilon}| > \frac{\lambda_1}{2}\Big)\\
\triangleq&\mathrm{T}_1+\mathrm{T}_2+\mathrm{T}_3+\mathrm{T}_4.
\end{aligned}$$

Note that $\|\boldsymbol{e}_j\|^2=1$. Let $\boldsymbol{a}_j=n^{-1}[\boldsymbol{e}_j^T(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T]^T$. By condition (A2), we have

$$\begin{aligned}
\|\boldsymbol{a}_j\|^2=&\frac{1}{n^2}\boldsymbol{e}_j^T(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\boldsymbol{e}_j\\
\leqslant&\frac{\rho_{\max}(\Sigma_{11})}{(\rho_{\min}(\Sigma_{11})+\lambda_2)^2}\frac{1}{n}\boldsymbol{e}_j^T\boldsymbol{e}_j\leqslant\frac{C_2}{(C_1+\lambda_2)^2}\frac{1}{n}\boldsymbol{e}_j^T\boldsymbol{e}_j=\frac{C_2}{n(C_1+\lambda_2)^2},
\end{aligned}$$

where $\rho_{\min}$ and $\rho_{\max}$ are the minimum and maximum eigenvalues, respectively. So by Lemma 2.1 and condition (A3), we have

$$\begin{aligned}
\mathrm{T}_1\leqslant&k_n\Pr\Big(|\frac{1}{\|\boldsymbol{a}_j\|}\boldsymbol{a}_j^T\boldsymbol{\varepsilon}|>\frac{1}{2}(b_{n1}-a\lambda_1)/\|\boldsymbol{a}_j\|\Big)\\
\leqslant&k_n\Pr\Big(|\frac{1}{\|\boldsymbol{a}_j\|}\boldsymbol{a}_j^T\boldsymbol{\varepsilon}|>\frac{\sqrt{n}(C_1+\lambda_2)}{2\sqrt{C_2}}(b_{n1}-a\lambda_1)\Big)\\
\leqslant&k_n\cdot\exp\Big\{-\frac{1}{M(\log n+1)^{I(d=1)}}[\frac{\sqrt{n}(C_1+\lambda_2)}{2\sqrt{C_1}}(b_{n1}-a\lambda_1)]^d\Big\}\\
=&\exp\Big\{-\log k_n[\frac{1}{M(\log n+1)^{I(d=1)}}(\frac{\sqrt{n}(C_1+\lambda_2)(b_{n1}-a\lambda_1)}{2\sqrt{C_1}(\log k_n)^{1/d}})^d-1]\Big\}\to 0.
\end{aligned}$$

For T_2, by the Cauchy-Schwartz inequality, we have

$$|\lambda_2\boldsymbol{e}_j^T(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\boldsymbol{\beta}_{10}|\leqslant\lambda_2\|\boldsymbol{\beta}_{10}\|/(C_1+\lambda_2).$$

Then under condition (A4), $\mathrm{T}_2\to 0$.

Use the Cauchy-Schwartz inequality,

$$\begin{aligned}
&\Big|\frac{1}{n}\boldsymbol{X}_j^T\mathbf{X}_1\left[\mathbf{I}_{11}-(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\Sigma_{11}\right]\boldsymbol{\beta}_{10}\Big|\\
=&\Big|\frac{1}{n}\boldsymbol{X}_j^T\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\lambda_2\boldsymbol{\beta}_{10}\Big|\leqslant\frac{\lambda_2}{\rho_{\min}(\Sigma_{11})+\lambda_2}|\frac{1}{n}\boldsymbol{X}_j^T\mathbf{X}_1\boldsymbol{\beta}_{10}|\\
\leqslant&\frac{\lambda_2\rho_{\max}^{1/2}(\Sigma_{11})}{\rho_{\min}(\Sigma_{11})+\lambda_2}\frac{1}{n^{1/2}}\|\boldsymbol{X}_j\|\cdot\|\boldsymbol{\beta}_{10}\|\leqslant\frac{\lambda_2}{C_1+\lambda_2}\sqrt{C_2}\|\boldsymbol{\beta}_{10}\|,
\end{aligned}$$

by condition (A5), $\mathrm{T}_3\to 0$.

Let $\boldsymbol{b}_j^T=n^{-1/2}\boldsymbol{X}_j^T[\mathbf{I}_n-n^{-1}\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T]$. Then

$$\begin{aligned}
\|\boldsymbol{b}_j\|^2=&\frac{1}{n}\boldsymbol{X}_j^T\Big[\mathbf{I}_n-\frac{1}{n}\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\Big]\Big[\mathbf{I}_n-\frac{1}{n}\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\Big]\boldsymbol{X}_j\\
\leqslant&\rho_{\max}^2\Big\{\mathbf{I}_n-\frac{1}{n}\mathbf{X}_1(\Sigma_{11}+\lambda_2\mathbf{I}_{11})^{-1}\mathbf{X}_1^T\Big\}\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{X}_j\leqslant 1.
\end{aligned}$$

By Lemma 2.1 and condition (A2), we have

$$\mathrm{T}_4\leqslant p_n\cdot\Pr\Big(|\boldsymbol{b}_j^T\boldsymbol{\varepsilon}|>\frac{\sqrt{n}\lambda_1}{2}\Big)$$

$$\leqslant p_n \cdot \exp\Big\{ -\frac{1}{M(\log n+1)^{I(d=1)}}(\frac{\sqrt{n}\lambda_1}{2})^d\Big\}$$
$$= \exp\Big\{ -\log p_n[\frac{1}{M(\log n+1)^{I(d=1)}}(\frac{\sqrt{n}\lambda_1}{2(\log p_n)^{1/d}})^d - 1]\Big\} \to 0.$$

The proof is completed. □

2.2. Asymptotic equivalence of the NPR estimator and oracle ridge estimator

In this section, under the condition about the smallest eigenvalue of Σ, the following theorem shows that the NPR estimator is exactly the same as the oracle ridge estimator asymptotically.

Theorem 2.3 *Let $\Sigma = n^{-1}\mathbf{X}^T\mathbf{X}$ and ρ be the smallest eigenvalue of Σ with $\rho + \lambda_2 > 0$. Let $\hat{\boldsymbol{\beta}}_n$ be the global minimizer of (2.1). Then under conditions (A1–A5),* $\Pr(\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n^0) \to 1$.

Proof It suffices to show that $\Pr(\inf_{\boldsymbol{\beta}_n \in \mathbb{R}^{p_n}} Q_n(\boldsymbol{\beta}_n) \geqslant Q_n(\hat{\boldsymbol{\beta}}_n^0)) \to 1$ as $n \to \infty$. By the definition of $\hat{\boldsymbol{\beta}}_n^0$, we have

$$\begin{aligned}
&Q_n(\boldsymbol{\beta}_n) - Q_n(\hat{\boldsymbol{\beta}}_n^0)\\
&\quad = -\sum_{j\notin J_{n1}} \Big[\frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n^0)^T\boldsymbol{X}_j\beta_{nj}\Big] + \frac{1}{2n}(\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_n)^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_n) +\\
&\qquad \frac{1}{2}\lambda_2\sum_{j=1}^{p_n}(\beta_{nj} - \hat{\beta}_{nj}^0)^2 + \sum_{j=1}^{p_n}[p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|)]\\
&\quad = -\sum_{j\notin J_{n1}} o_P(\lambda_1)|\beta_{nj}| + \frac{1}{2}(\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_n)^T(\Sigma + \lambda_2\mathbf{I})(\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_n) + \sum_{j=1}^{p_n}[p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|)]\\
&\quad \geqslant -\sum_{j\notin J_{n1}} o_P(\lambda_1)|\beta_{nj}| + \frac{1}{2}(\rho + \lambda_2)\|\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_n\|^2 + \sum_{j=1}^{p_n}[p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|)].
\end{aligned}$$

Define $J_{n1}^+ = \{j : |\beta_{nj}| \geqslant a\lambda_1, j \in J_{n1}\}$, $J_{n1}^- = \{j : |\beta_{nj}| < a\lambda_1, j \in J_{n1}\}$, $J_{n2}^+ = \{j : |\beta_{nj}| < a\lambda_1, j \notin J_{n1}\}$ and $J_{n2}^- = \{j : |\beta_{nj}| \geqslant a\lambda_1, j \notin J_{n1}\}$. Then, by the definition of nonconvex penalties, for $j \in J_{n1}^+$, $p'_{\lambda_1}(|\beta_{nj}|) = 0$, and for $j \in J_{n2}^+$, $p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|) = p'_{\lambda_1}(s)|\beta_{nj}| \geqslant (\lambda_1 - s/a)\,|\beta_{nj}| \geqslant (\lambda_1 - |\beta_{nj}|/a)\,|\beta_{nj}|$, where s is between $|\beta_{nj}|$ and 0. So we have

$$\begin{aligned}
&\sum_{j=1}^{p_n}[p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|)]\\
&\quad \geqslant \sum_{j\in J_{n1}^-}[p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|)] + \sum_{j\in J_{n2}^+}(\lambda_1 - |\beta_{nj}|/a)\,|\beta_{nj}| + \sum_{j\in J_{n2}^-} p_{\lambda_1}(|\beta_{nj}|)\\
&\quad \geqslant -\sum_{j\in J_{n1}^-} p_{\lambda_1}(|\hat{\beta}_{nj}^0|) + \sum_{j\in J_{n2}^+}(\lambda_1 - |\beta_{nj}|/a)\,|\beta_{nj}|.
\end{aligned}$$

According to the proof of Theorem 2.2, $|\hat{\beta}_{nj}^0| \geqslant a\lambda_1$, then $p_{\lambda_1}(|\hat{\beta}_{nj}^0|) = O(\lambda_1^2)$. Hence,

$$\sum_{j=1}^{p_n}[p_{\lambda_1}(|\beta_{nj}|) - p_{\lambda_1}(|\hat{\beta}_{nj}^0|)] \geqslant -|J_{n1}^-|O(\lambda_1^2) + \sum_{j\in J_{n2}^+}(\lambda_1 - |\beta_{nj}|/a)\,|\beta_{nj}|.$$

In addition, by the definition of $\hat{\boldsymbol{\beta}}_n^0$, it is easy to show that

$$\|\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_0\| = O_P\left(\lambda_2\|\boldsymbol{\beta}_0\| + C_2\sigma^2 k_n/[n(C_1+\lambda_2)]\right).$$

For $j \in J_{n1}^-$, we have

$$|\hat{\beta}_{nj}^0 - \beta_{nj}| \geqslant \min_{j\in J_{n1}^-}|\beta_{0j}| - |\hat{\beta}_{nj}^0 - \beta_{0j}| - |\beta_{nj}| \geqslant b_{n1} - O_P\big(\frac{\lambda_2\|\boldsymbol{\beta}_0\| + C_2\sigma^2 k_n/n}{C_1+\lambda_2}\big) - a\lambda_1.$$

Let $\zeta_n = b_{n1} - O_P(\frac{\lambda_2\|\boldsymbol{\beta}_0\|+C_2\sigma^2 k_n/n}{C_1+\lambda_2}) - a\lambda_1$. Then

$$\begin{aligned}\frac{1}{2}(\rho+\lambda_2)\|\hat{\boldsymbol{\beta}}_n^0 - \boldsymbol{\beta}_n\|^2 &\geqslant \sum_{j\in J_{n1}^-}\frac{1}{2}(\rho+\lambda_2)(\hat{\beta}_{nj}^0-\beta_{nj})^2 + \sum_{j\in J_{n2}^-}\frac{1}{2}(\rho+\lambda_2)a\lambda_1|\beta_{nj}| \\ &\geqslant |J_{n1}^-|\frac{1}{2}(\rho+\lambda_2)\zeta_n^2 + \sum_{j\in J_{n2}^-}\frac{1}{2}(\rho+\lambda_2)a\lambda_1|\beta_{nj}|.\end{aligned}$$

Hence, we have

$$\begin{aligned}Q_n(\boldsymbol{\beta}_n) - Q_n(\hat{\boldsymbol{\beta}}_n^0) \geqslant &-\sum_{j\notin J_{n1}} o_P(\lambda_1)|\beta_{nj}| + |J_{n1}^-|\frac{1}{2}(\rho+\lambda_2)\zeta_n^2 + \sum_{j\in J_{n2}^-}\frac{1}{2}(\rho+\lambda_2)a\lambda_1|\beta_{nj}| - \\ &|J_{n1}^-|O(\lambda_1^2) + \sum_{j\in J_{n2}^+}\left(\lambda_1 - \max_{j\in J_{n2}^+}|\beta_{nj}|/a\right)|\beta_{nj}| \\ \geqslant &\Big[\min\{\frac{1}{2}(\rho+\lambda_2)a, 1 - \max_{j\in J_{n2}^+}|\beta_{nj}|/(a\lambda_1)\}\lambda_1 - o_P(\lambda_1)\Big]\sum_{j\notin J_{n1}}|\beta_{nj}| + \\ &|J_{n1}^-|\big[\frac{1}{2}(\rho+\lambda_2)\zeta_n^2 - O(\lambda_1^2)\big].\end{aligned}$$

By conditions (A3) and (A4), we have $\Pr(\inf_{\boldsymbol{\beta}\in\mathbb{R}^{p_n}} Q_n(\boldsymbol{\beta}_n) \geqslant Q_n(\hat{\boldsymbol{\beta}}_n^0)) \to 1$ as $n\to\infty$. □

Remark 2.4 In Theorem 2.3, we assume that the matrix Σ may be singular, but $\rho+\lambda_2 > 0$. Hence, for ultra-high dimensional data, if we choose $\lambda_2 > 0$, the oracle ridge estimator is the global minimizer of (2.1).

3. Numerical studies

In this section, we present simulation studies to examine the finite sample performance of the NPR estimate. The nonconvex penalized method includes the SCAD and MCP, so we consider the combinations of these two methods and ridge regression, denoted by Snet and Mnet, respectively. The simulation studies in [21] suggested that $a = 3$ is a reasonable choice for the MCP, so we set $a = 3$ for the MCP. As suggested in [14], we choose $a = 3.7$ for the SCAD. We consider six methods in simulations: the Lasso, the Enet, the MCP, the Mnet, the SCAD and the Snet. All these estimates are computed using the ncvreg package [22]. The regularization parameters λ_1 and λ_2 are selected by ten-fold cross validation.

To evaluate the performance of these methods, we give four summary statistics: the average number of zero coefficients that are correctly estimated by zero (C), the average number of nonzero coefficients that are incorrectly estimated by zero (IC), the median of L_2 loss $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$,

and the median of the prediction error (PE) defined as $E(Y-\boldsymbol{X}^T\hat{\boldsymbol{\beta}})^2$. We compute the prediction error on the basis of 1000 independent test data sets. Summary statistics are given based on 100 replications. The numbers in parentheses of all tables are the corresponding standard errors.

We simulate data from the linear model

$$Y=\sum_{j=1}^{p_n} x_j\beta_{0j}+\varepsilon,$$

where $\boldsymbol{\beta}_0=(\beta_{01},\ldots,\beta_{0p_n})^T$ is a p_n-vector, $\boldsymbol{X}=(x_1,\ldots,x_{p_n})^T$ is a multivariate normal distribution with zero mean and the covariance of x_j and x_k being $r^{|j-k|}$, $1\leqslant j,k\leqslant p_n$, and $\varepsilon\sim N(0,1)$. The true coefficients are $\beta_{0j}=3$, for $j=1,\ldots,15$, the rest are zero. We consider $r=0.5$ and $r=0.9$, which represent moderate and strong correlation, respectively. We investigate $p_n=40,600$ for $n=100,300$.

r	n	Methods	C	IC	L_2 loss	PMSE
0.5	100	Lasso	17.51 (3.486)	0 (0)	0.615 (0.113)	1.292 (0.110)
		Enet	15.53 (4.565)	0 (0)	0.631 (0.137)	1.311 (0.149)
		MCP	24.71 (0.998)	0 (0)	0.516 (0.117)	1.165 (0.092)
		Mnet	24.14 (2.531)	0 (0)	0.514 (0.140)	1.179 (0.132)
		SCAD	24.41 (1.408)	0 (0)	0.527 (0.110)	1.168 (0.085)
		Snet	23.97 (2.380)	0 (0)	0.515 (0.133)	1.179 (0.101)
		Truth	25.00 (0.000)	0 (0)	0.501 (0.115)	1.160 (0.074)
	300	Lasso	16.63 (4.223)	0 (0)	0.343 (0.065)	1.089 (0.053)
		Enet	15.36 (5.252)	0 (0)	0.354 (0.068)	1.102 (0.066)
		MCP	24.46 (1.629)	0 (0)	0.311 (0.075)	1.065 (0.051)
		Mnet	24.02 (1.682)	0 (0)	0.312 (0.072)	1.052 (0.063)
		SCAD	24.19 (1.774)	0 (0)	0.307 (0.069)	1.045 (0.052)
		Snet	23.92 (1.942)	0 (0)	0.306 (0.070)	1.060 (0.049)
		Truth	25.00 (0.000)	0 (0)	0.299 (0.067)	1.053 (0.054)
0.9	100	Lasso	22.02 (2.318)	0 (0)	1.263 (0.328)	1.210 (0.100)
		Enet	20.47 (3.341)	0 (0)	1.149 (0.313)	1.205 (0.109)
		MCP	23.52 (1.817)	0 (0)	1.246 (0.386)	1.205 (0.165)
		Mnet	24.56 (1.358)	0 (0)	0.849 (0.289)	1.125 (0.086)
		SCAD	23.43 (1.677)	0 (0)	1.227 (0.370)	1.177 (0.139)
		Snet	24.25 (1.629)	0 (0)	0.890 (0.377)	1.136 (0.120)
		Truth	25.00 (0.000)	0 (0)	1.184 (0.311)	1.159 (0.088)
	300	Lasso	21.93 (2.289)	0 (0)	0.677 (0.168)	1.065 (0.049)
		Enet	21.33 (2.305)	0 (0)	0.655 (0.159)	1.074 (0.057)
		MCP	24.65 (0.936)	0 (0)	0.662 (0.176)	1.057 (0.049)
		Mnet	24.70 (0.916)	0 (0)	0.549 (0.167)	1.051 (0.062)
		SCAD	24.45 (1.123)	0 (0)	0.675 (0.170)	1.048 (0.052)
		Snet	24.61 (1.053)	0 (0)	0.557 (0.169)	1.047 (0.053)
		Truth	25.00 (0.000)	0 (0)	0.667 (0.166)	1.053 (0.053)

Table 1 Simulation results for $p_n=40$

The results on variable selection and prediction errors are shown in Tables 1–2. It can be seen that (1) Table 1 shows the prediction and selection performance in low dimensional case. When the correlation is moderate, the Lasso, the MCP and the SCAD perform better than the Enet, the Mnet and the Snet, respectively. The four methods with oracle property outperform the lasso and the Enet which do not have the oracle property, in terms of variable selection and prediction. Both the L_2 loss and the prediction error decrease as the sample size increases. However, when the correlation is high, the three combination approaches are better than the three methods without L_2 penalty. This is in line with our theoretical results.

(2) In the settings of high dimension (Table 2), when the correlation is moderate, the four oracle-like methods perform much better than the Lasso and the Enet, especially for large sample size. However, the SCAD and the MCP are much worse in the case of high correlation. The Snet and the Mnet are similar in all cases. All these findings confirm the theoretical results obtained in Section 2.

r	n	Methods	C	IC	L_2 loss	PMSE
0.5	100	Lasso	554.80 (15.475)	0 (0)	0.856 (0.152)	1.856 (0.268)
		Enet	553.74 (15.640)	0 (0)	0.866 (0.165)	1.795 (0.312)
		MCP	584.07 (2.253)	0 (0)	0.541 (0.136)	1.184 (0.102)
		Mnet	583.00 (3.088)	0 (0)	0.539 (0.149)	1.218 (0.160)
		SCAD	581.71 (6.323)	0 (0)	0.547 (0.135)	1.170 (0.096)
		Snet	580.36 (6.774)	0 (0)	0.550 (0.136)	1.199 (0.114)
		Truth	585.00 (0.000)	0 (0)		
	300	Lasso	554.89 (20.680)	0 (0)	0.413 (0.072)	1.191 (0.073)
		Enet	553.62 (22.994)	0 (0)	0.411 (0.079)	1.187 (0.089)
		MCP	583.58 (3.540)	0 (0)	0.291 (0.068)	1.063 (0.057)
		Mnet	583.24 (3.937)	0 (0)	0.288 (0.066)	1.057 (0.051)
		SCAD	581.93 (6.612)	0 (0)	0.291 (0.064)	1.059 (0.052)
		Snet	580.74 (8.816)	0 (0)	0.289 (0.068)	1.059 (0.055)
		Truth	585.00 (0.000)	0 (0)		
0.9	100	Lasso	573.99 (11.230)	0.00 (0.000)	1.340 (0.305)	1.325 (0.168)
		Enet	572.51 (11.076)	0.00 (0.000)	1.234 (0.308)	1.315 (0.155)
		MCP	577.88 (3.560)	7.68 (1.213)	12.803 (1.934)	13.481 (5.353)
		Mnet	582.75 (2.746)	0.02 (0.200)	0.858 (0.469)	1.148 (0.338)
		SCAD	569.66 (6.046)	7.39 (1.270)	12.411 (1.946)	13.415 (7.067)
		Snet	578.05 (5.068)	0.04 (0.400)	0.913 (0.700)	1.148 (0.527)
		Truth	585.00 (0.000)	0.00 (0.000)		
	300	Lasso	574.59 (10.937)	0 (0)	0.682 (0.165)	1.093 (0.065)
		Enet	572.84 (11.939)	0 (0)	0.674 (0.162)	1.109 (0.070)
		MCP	584.02 (1.864)	0 (0)	0.658 (0.158)	1.069 (0.055)
		Mnet	584.26 (1.952)	0 (0)	0.515 (0.163)	1.048 (0.050)
		SCAD	582.55 (4.314)	0 (0)	0.657 (0.160)	1.066 (0.043)
		Snet	582.47 (4.747)	0 (0)	0.556 (0.154)	1.058 (0.054)
		Truth	585.00 (0.000)	0 (0)		

Table 2 Simulation results for $p_n = 600$

4. Data analysis

In this section, we apply the NPR method to the Boston housing data, which were previously used to evaluate the performance of different regression methods for example in [23]. Originally, the data set contains 13 variables which may have influence over the house prices. As in [23], we include the interaction terms between the variables in the analysis such that the data has 91 variables and 506 observations. Note that, because of the way that the variables are produced, there are large sample correlations across the columns of the design matrix. We randomly partition the data into a training set with 60 observations and a test set with 446 observations. We fit the model with the training data, then calculate the prediction error (residual sum of squares) for the test data. Note that this partition results in high dimensionality. Random splitting of the data is repeated 200 times, and each time a new training data and test data are made. The median of the number of selected variables and the median of prediction errors are reported in Table 3.

As shown in Table 3, the Lasso, the MCP and the SCAD, which cannot deal with the collinearity problem, produce smaller models than the Enet, the Mnet and the Snet, respectively. Although the Lasso gives the smallest prediction error, it may miss some important variables and include some irrelevant variables. Specifically, the Snet performs much better than the SCAD. Compared to the Mnet, the Snet has smaller perdition error with two more variables.

Methods	Number of nonzero coefficients	Prediction errors
Lasso	9.5	8.169
Enet	13.0	8.656
MCP	5.0	10.136
Mnet	7.0	8.828
SCAD	6.0	9.726
Snet	9.0	8.459

Table 3 Boston housing data: the median of the prediction errors and the number of nonzero coefficients

5. Discussion

We have adopted the nonconvex penalties and ridge regression to deal with high-dimensional variable selection for linear models with high correlation. In theory, we demonstrate the oracle property of the NPR estimator under certain conditions. In simulation, we find that the NPR performs much better than the nonconvex penalties in high dimensionality. In addition, we focus on the NPR in the context of linear models. It seems possible to extend our results to some other complex parametric models and semiparametric models. However, these extensions are by no means straightforward.

Acknowledgements

The authors would like to thank the anonymous referees and the editor for constructive comments and helpful suggestions.

References

[1] T. R. GOLUB, D. K. SLONIM, P. TAMAYO, et al. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999, **286**(5493): 531–537.
[2] R. JAGANNATHAN, Tongshu MA. *Risk reduction in large portfolios: Why imposing the wrong constraints helps.* The Journal of Finance, 2003, **58**(4): 1651–1684.
[3] R. VAN DER HILST, M. DE HOOP, P. WANG, et al. *Seismo-stratigraphy and thermal structure of earth's core-mantle boundary region.* Science, 2007, **315**(5820): 1813–1817.
[4] M. SEGAL, K. DAHLQUIST, B. CONKLIN. *Regression approach for microarray data analysis.* J. Comput. Biol., 2003, **10**(6): 961–980.
[5] A. E. HOERL, R. W. KENNARD. *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, 1970, **12**(1): 55–67.
[6] R. J. TIBSHIRANI. *Regression shrinkage and selection via the Lasso.* J. Roy. Statist. Soc. Ser. B, 1996, **58**(1): 267–288.
[7] Hui ZOU, T. HASTIE. *Regularization and variable selection via the elastic net.* J. Roy. Statist. Soc. Ser. B, 2005, **67**(2): 301–320.
[8] Ming YUAN, Yi LIN. *On the nonnegative garrote estimator.* J. Roy. Statist. Soc. Ser. B, 2007, **69**(2): 143–161.
[9] Jinzhu JIA, Bin YU. *On model selection consistency of elastic net when $p \gg n$.* Statist. Sinica, 2010, **20**(2): 595–611.
[10] Peng ZHAO, Bin YU. *On model selection consistency of Lasso.* J. Mach. Learn. Res., 2006, **7**(11): 2541–2563.
[11] N. MEINSHAUSEN, P. BÜHLMANN. *High-dimensional graphs and variable selection with the Lasso.* Ann. Statist., 2006, **34**(3): 1436–1462.
[12] Hui ZOU. *The adaptive Lasso and its oracle properties.* J. Amer. Statist. Assoc., 2006, **101**(476): 1418–1429.
[13] Hui ZOU, Hao Helen ZHANG. *On the adaptive elastic-net with a diverging number of parameters.* Ann. Statist., 2009, **37**(4): 1733–1751.
[14] Jianqing FAN, Runze LI. *Variable selection via nonconcave penalized likelihood and its oracle properties.* J. Amer. Statist. Assoc., 2001, **96**(456): 1348–1360.
[15] Xiaoming WANG, T. PARK, K. C. CARRIERE. *Variable selection via combined penalization for high-dimensional data analysis.* Comput. Stat. Data Anal., 2010, **54**(10): 2230–2243.
[16] Lingmin ZENG, Jun XIE. *Group variable selection via SCAD-L_2.* Statistics, 2014, **48**(1): 49–66.
[17] Y. KIM, H. CHOI, H. OH. *Smoothly clipped absolute deviation on high dimensions.* J. Amer. Statist. Assoc., 2008, **103**(484): 1665–1673.
[18] Jian HUANG, Shuangge MA, Hongzhe LI, et al. *The sparse laplacian shrinkage estimator for high-dimensional regression.* Ann. Statist., 2011, **39**(4): 2021–2046.
[19] Cunhui ZHANG. *Nearly unbiased variable selection under minimax concave penalty.* Ann. Statist., 2010, **38**(2): 894–942.
[20] Jian HUANG, Shuangge MA, Cunhui ZHANG. *Adaptive Lasso for high dimensional regression models.* Statist. Sinica, 2008, **18**(4): 1603–1618.
[21] D. P. BERTSEKAS. *Nonlinear Programming* (2nd ed.). Athena Scientific, Belmount. 1999.
[22] P. BREHENY, Jian HUANG. *Coordinate descent algorithms for nonconvex penalized regression, with application to biological feature selection.* Ann. Appl. Stat., 2011, **5**(1): 232–253.
[23] P. RADCHENKO, G. M. JAMES. *Improved variable selection with forward-lasso adaptive shrinkage.* Ann. Appl. Stat., 2011, **5**(1): 427–448.