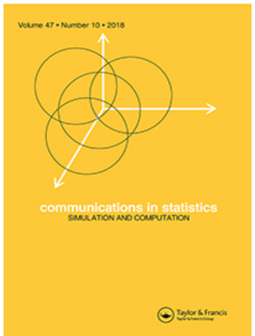


Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Nonnegative estimation and variable selection via adaptive elastic-net for high-dimensional data

Ning Li, Hu Yang & Jing Yang

To cite this article: Ning Li, Hu Yang & Jing Yang (2019): Nonnegative estimation and variable selection via adaptive elastic-net for high-dimensional data, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1642484

To link to this article: https://doi.org/10.1080/03610918.2019.1642484

Published online: 24 Jul 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1642484

Taylor & Francis
Taylor & Francis Group

Nonnegative estimation and variable selection via adaptive elastic-net for high-dimensional data

Ning Li[a,b], Hu Yang[a], and Jing Yang[c]

[a]College of Mathematics and Statistics, Chongqing University, Chongqing, P. R. China; [b]Department of Mathematics and Physics, Hefei University, Hefei, P. R. China; [c]Key Laboratory of High Performance Computing and Stochastic Information Processing (Ministry of Education of China), College of Mathematics and Statistics, Hunan Normal University, Changsha, P. R. China

ABSTRACT

This paper proposes the nonnegative adaptive elastic-net for simultaneous nonnegative estimation and variable selection in sparse high-dimensional linear regression models. By inheriting the good features of adaptive elastic-net, the nonnegative adaptive elastic-net enjoys the oracle property even in high-dimensional settings where the dimension of covariates can be much larger than the sample size. Through the simulation, we show that the newly proposed method deals with the collinearity problem better than alternative procedures in the literature. To make the proposed method practically feasible, we extend the multiplicative updates algorithm for implementation. Finally, we illustrate the favorable finite-sample performance of the proposed method through tracking the CSI 300 index, an important stock market index in China.

ARTICLE HISTORY
Received 9 January 2019
Accepted 3 July 2019

KEYWORDS
Nonnegative estimation; variable selection; adaptive elastic-net; oracle property; high-dimensional data; collinearity

1. Introduction

Consider the classical linear model

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

where $\mathbf{y} = (y_1, ..., y_n)^T$ is the response vector and $\mathbf{X} = (\mathbf{x}_1 ..., \mathbf{x}_{p_n})$ is the design matrix, where $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T, j = 1, ..., p_n$. Throughout this paper, we assume $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ is a vector of independent identically distributed Gaussian variables with mean 0 and finite variance σ^2. Without loss of generality, we assume that both $\mathbf{X}$ and $\mathbf{y}$ are centered, so that the intercept is not included in the regression function.

Variable selection is important when the actual model has a sparse representation and it could greatly improve the prediction performance of the fitted model. Traditional variable selection procedures such as AIC (Akaike 1973), BIC (Schwarz 1978), and k-fold Cross-Validation (Breiman 1995) all suffered from the problems of instability and intensive computation. To address these deficiencies, statisticians have recently proposed various penalized methods to perform simultaneous estimation and variable selection. In particular, the lasso proposed by Tibshirani (1996) is one of the most popular

CONTACT Ning Li ninglicqu@163.com College of Mathematics and Statistics, Chongqing University, Chongqing 401331, P. R. China; Department of Mathematics and Physics, Hefei University, Hefei 230601, P. R. China.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

methods due to its good computational and statistical properties. But just as Fan and Li (2001) conjectured, the oracle property does not hold for the lasso. Meanwhile, Zou and Hastie (2005) showed that collinearity can severely degrade the performance of the lasso for high-dimensional data. Later, the elastic-net (Zou and Hastie 2005) and the adaptive lasso (Zou 2006) have been proposed to improve the lasso in these two different aspects. That is, the adaptive lasso achieves the oracle property and the elastic-net handles the problem of collinearity. However, Zou and Hastie (2005) also pointed out that the adaptive lasso inherits the instability of the lasso for high-dimensional data, while Zou (2006) argued that the elastic-net lacks the oracle property. Thus, Zou and Zhang (2009) proposed the adaptive elastic-net that penalizes the squared error loss using a combination of the quadratic regularization and the adaptive lasso shrinkage. They showed that the adaptive elastic-net not only enjoys the oracle property but also deals with the collinearity problem better than the other oracle-like procedures.

In this paper, we are concerned with the case that $\boldsymbol{\beta}^*$ is nonnegative. This constraint is particularly essential if we are to tackle the nonnegative data, which is often derived from economical quantities such as prices, incomes and growth rates. However, there are still difficulties for the penalized methods to deal with this type of data. To our best knowledge, several literatures have studied the nonnegative problem of the penalized methods. Breiman (1995) proposed the nonnegative garrote estimator, which is showed to be more stable than the subset regression and ridge regression. Slawski and Hein (2013) showed that the nonnegative least squares performs better than the lasso in prediction and estimation. Meanwhile, Meinshausen (2013) showed that sign-constrained least squares is an effective regularization technique for sparse high-dimensional data under certain conditions. Recently, Wu et al. (2014) proposed the nonnegative lasso and proved that it enjoys oracle property in high-dimensional settings. Wu and Yang (2014) proposed the nonnegative elastic-net that can still select the true variables even when the nonnegative lasso fails. Yang and Wu (2016) proposed the nonnegative adaptive lasso for sparse high-dimensional linear regression models.

However, there is no literature discussed about the adaptive elastic-net under the nonnegative constraint before, although it combines the strengths of the adaptive lasso and the elastic-net. Thus, we propose the nonnegative adaptive elastic-net for simultaneous nonnegative estimation and variable selection in sparse high-dimensional linear regression models. The proposed method is an extension of the adaptive elastic-net with nonnegative constraint on the coefficients. Under certain appropriate conditions, we prove that the nonnegative adaptive elastic-net has nice properties of variable selection consistency and asymptotic normality, thus enjoying the oracle property. The main contribution of this paper are threefold. Firstly, the oracle property of the nonnegative adaptive elastic-net holds when p_n grows much faster than n. Secondly, we show by simulation studies that the nonnegative adaptive elastic-net deals with the collinearity problem better than the other nonnegative methods for high-dimensional data. Thirdly, the results of the real data example illustrate that the nonnegative adaptive elastic-net is an appropriate method for index tracking.

The rest of this paper is organized as follows. In Sec. 2, we introduce the nonnegative adaptive elastic-net. The oracle property, including variable selection consistency and asymptotic normality, of the nonnegative adaptive elastic-net is established in Sec. 3. In

Sec. 4, we extend the multiplicative updates algorithm for implementation. In Sec. 5, we present simulations to compare the finite performance of the nonnegative adaptive elastic-net with the existing nonnegative methods. In Sec. 6, the nonnegative adaptive elastic-net is applied in the financial modeling for tracking the CSI 300 index in China. We conclude with a few remarks in Sec. 7. All the technical proofs are given in the Appendix.

2. Methodology

Without loss of generality, we assume that the response is centered and the covariates are standardized,

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \qquad \text{and} \qquad \sum_{i=1}^{n} x_{ij}^2/n = 1, \qquad \text{for } j = 1, \dots, p_{n.} \tag{2}$$

For fixed regularization parameters λ_1 and λ_2, we define the nonnegative adaptive elastic-net estimator as follows:

$$\hat{\boldsymbol{\beta}} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\boldsymbol{\beta} \succcurlyeq \mathbf{0}}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p_n} \hat{\omega}_j \beta_j \right\} \tag{3}$$

where $\hat{\boldsymbol{\omega}} = (\hat{\omega}_1, \dots, \hat{\omega}_{p_n})^T$ is the weight vector. Throughout this paper, the symbols $\preccurlyeq, \succcurlyeq$ and $\prec, \succ$ denote element-wise inequalities and element-wise strict inequalities, respectively. In particular, when $\lambda_2 = 0$, the nonnegative adaptive elastic-net is reduced to the nonnegative adaptive lasso (Yang and Wu 2016); when $\hat{\omega}_j = 1$ for all $i = 1, \dots, n$, the nonnegative adaptive elastic-net is same as the nonnegative elastic-net (Wu and Yang 2014). Without loss of generality, we assume that the true parameters $\beta_j^* \neq 0$, for $j = 1, \dots, q_n$, and $\beta_j^* = 0$, for $j = q_n + 1, \dots, p_n$. We decompose the vector $\boldsymbol{\beta}^*$ into two parts, that is $\boldsymbol{\beta}_1^* = (\beta_1^*, \dots, \beta_{q_n}^*)^T$, and $\boldsymbol{\beta}_2^* = (\beta_{q_n+1}^*, \dots, \beta_{p_n}^*)^T$. In the same way, we let $\mathbf{X}_1$ and $\mathbf{X}_2$ as the first q_n and last $p_n - q_n$ columns of $\mathbf{X}$, respectively. Let $S = \{j \in \{1, \dots, p_n\} : \beta_j \neq 0\}$ with cardinality $|S| = q_n$, and $\hat{S} = \{j \in \{1, \dots, p_n\} : \hat{\beta}_j \neq 0\}$. Then we denote $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, $\mathbf{C}_{11} = \frac{1}{n}\mathbf{X}_1^T\mathbf{X}_1$, $\mathbf{C}_{12} = \frac{1}{n}\mathbf{X}_1^T\mathbf{X}_2$, $\mathbf{C}_{21} = \frac{1}{n}\mathbf{X}_2^T\mathbf{X}_1$, and $\mathbf{C}_{22} = \frac{1}{n}\mathbf{X}_2^T\mathbf{X}_2$, where $\mathbf{C}_{11}$ is assumed to be invertible.

Suppose that an initial estimator $\hat{\boldsymbol{\beta}}^0$ is available, then we can set the weights $\hat{\omega}_j = |\hat{\beta}_j^0|^{-1}, j = 1, \dots, p_n$. By putting relatively bigger weights on zero coefficients and smaller weights on nonzero coefficients, the nonnegative adaptive elastic-net hopes to reduce the estimation bias and improve variable selection accuracy, compared with the standard nonnegative elastic-net. For the nonnegative adaptive elastic-net estimator to be variable selection consistent, it is crucial to have a suitable initial estimator. When $p_n \leq n$, the least squares estimator is consistent and we can use the least squares estimator as the initial estimator for the weights. However, when $p_n > n$, the least squares estimator is no longer feasible. In this case, we choose the marginal regression estimator as the initial estimator. With the centering and scaling given in (2), the marginal regression estimator is defined as follows:

$$\hat{\beta}_j^0 = \frac{\mathbf{x}_j^T \mathbf{y}}{n}, \tag{4}$$

where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. According to Huang et al. (2008), we need certain conditions to ensure the consistency of the marginal regression estimator.

Consider the following conditions:
(C1) (Partial orthogonality) The covariates with zero coefficients and those with nonzero coefficients are weakly correlated

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}\right| = \left|\frac{\mathbf{x}_j^T\mathbf{x}_k}{n}\right| \le \rho_n, \qquad j \in S,\ k \in S^c,$$

where, for certain $0 < \kappa < 1$ and $g_j = \frac{E(\mathbf{x}_j^T\mathbf{y})}{n}$, ρ_n satisfies

$$c_n = \left(\max_{j\in S^c}|g_j|\right)\left(\sum_{j\in S}\frac{|g_j|^{-2}}{q_n}\right)^{\frac{1}{2}} \le \frac{\kappa\Lambda_{\min}(\mathbf{C}_{11})}{q_n\rho_n},$$

and $\Lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of $\mathbf{C}_{11}$.
(C2) The minimum $b_n = \min\{|g_j|, j \in S\}$ satisfies

$$\frac{q_n^{\frac{1}{2}}(1+c_n)}{b_n r_n} \longrightarrow 0, \qquad r_n = \frac{n^{\frac{1}{2}}}{\left(\log\left(p_n - q_n\right)\right)^{\frac{1}{2}}}.$$

Condition (C1) is the partial orthogonality assumption in which the covariates with zero coefficients have weaker correlation with those with nonzero coefficients. Condition (C2) assumes that the non-zero coefficients are bounded away from zero at certain rate depending on the growth of p_n and q_n.

Definition 1. We say that $\hat{\boldsymbol{\beta}}^0 = (\hat{\beta}_1^0, \ldots, \hat{\beta}_{p_n}^0)^T$ is $\mathrm{r_n}$-consistent for the estimation of $\mathbf{g} = (g_1, \ldots, g_{p_n})^T$ if

$$r_n \max_{1\le j\le p_n} |\hat{\beta}_j^0 - g_j| = O_p(1), \qquad r_n \to \infty.$$

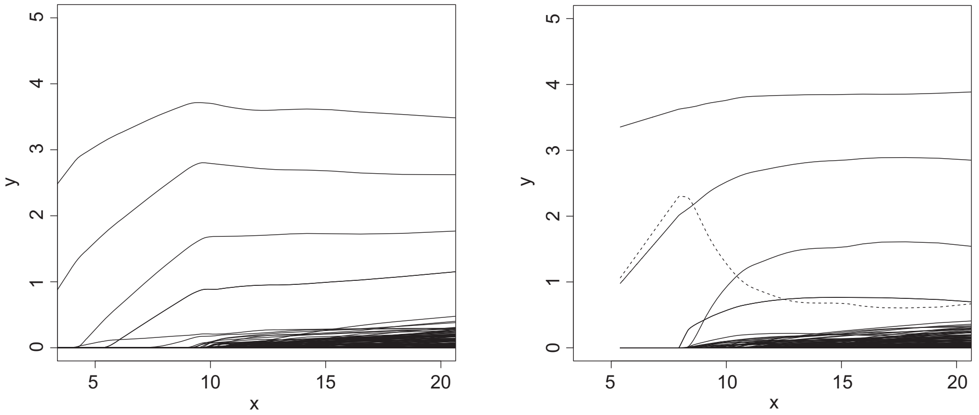
Definition 2. Nonnegative adaptive Elastic Irrepresentable Condition (NAEI condition): For $\mathbf{s}_1 = (|g_j|^{-1}, j \in S)^T$, and $\mathbf{s}_2 = (|g_j|^{-1}, j \in S^c)^T$, there exists a positive constant vector $\boldsymbol{\eta}$ such that:

$$\mathbf{C}_{21}\left(\mathbf{C}_{11} + \frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{s}_1 - \mathbf{s}_2 \leqslant \boldsymbol{\eta}, \tag{5}$$

where $\mathbf{I}$ $q_n \times q_n$*identity matrix.*

Theorem 1. *Suppose that conditions (C1)-(C2) hold. Then the* $\hat{\boldsymbol{\beta}}^0$ *given in (4) is* r_n*-consistent for* $\mathbf{g}$ *and the NAEI condition holds.*

Remark 1. Theorem 1 provides justification for using marginal regression estimator as the initial estimator. This theorem also shows that the NAEI condition can be satisfied under the partial orthogonality condition on the covariates. Similar irrepresentable conditions are existed in other regularization methods, like lasso, nonnegative lasso, etc. The proof of this theorem follows from the adaptations of the proof of Theorem 3 in Huang et al. (2008), so the detail is omitted in this paper.



(a) case 1.

(b) case 2.

Figure 1. The nonnegative adaptive elastic-net's solution paths. Each line represents a solution path, plotted by $||\hat{\boldsymbol{\beta}}||_1$ on the horizontal axis and the value of $\hat{\beta}_i$ on the vertical axis. The dashed line in each subfigure is the solution path for $\hat{\beta}_{400}$.

3. Statistical theory

In our theoretical analysis, we assume the following regularity conditions throughout: there exist positive constants M_1, M_2, M_3 and $0 \leq c_1 < c_2 \leq 1$,

(C3) $M_1 \leq \boldsymbol{\alpha}^T \mathbf{C}_{11} \boldsymbol{\alpha} \leq M_2$, where $\boldsymbol{\alpha}$ is any vector such that $||\boldsymbol{\alpha}||_2^2 = 1$,
(C4) $n^{\frac{1-c_2}{2}} \min_{i \in S} |\beta_i^*| \geq M_3$.

Condition (C4) requires the eigenvalue of $\mathbf{C}_{11}$ is bounded. Condition (C5) restricts the decay rate of $\boldsymbol{\beta}_1^*$ to prevent the estimation to be dominated by the noise terms.

Theorem 2. *Under conditions (C1)-(C4), if there exists a positive constant c_3 such that $c_3 < c_2 - c_1$ for which $p_n = \mathrm{O}(e^{n^{c_3}})$ and $q_n = \mathrm{O}(n^{c_1})$. If the regularization parameters λ_1, λ_2 are chosen such that $\lambda_1 \propto n^{\frac{1+c_4}{2}}$, for $c_3 < c_4 < c_2 - c_1, c_1 + c_2 > 1 + c_4$, and $\lambda_2/n \to 0, \frac{\lambda_2}{\sqrt{n}}\sqrt{\sum_{j \in S} \beta_j^{*2}} \to 0$, as $n \to \infty$, then the nonnegative adaptive elastic-net has variable selection consistency. That is*

$$\lim_{n \to \infty} P(\hat{S} = S) = 1. \tag{6}$$

Remark 2. Theorem 2 suggests that the variable selection consistency of the nonnegative adaptive elastic-net is still valid when $\mathrm{p_n}$ grows much faster than n (up to exponentially fast). The complete proof of Theorem 2 can be found in the Appendix.

Theorem 3. *Under the same settings in Theorem 2, the nonnegative adaptive elastic-net is consistent and asymptotically normal,*

$$\boldsymbol{\alpha}^T \left\{ \sqrt{n} \frac{\mathbf{I} + (\lambda_2/n)\mathbf{C}_{11}^{-1}}{1 + \lambda_2/n} \mathbf{C}_{11}^{1/2} \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^* \right) \right\} \to_d N(0, \sigma^2), \tag{7}$$

where $\boldsymbol{\alpha}$ is a q_n-dimensional vector of norm 1.

Remark 3. Theorem 3 indicates that the nonnegative adaptive elastic-net enjoys the asymptotic normality. Thus, the nonnegative adaptive elastic-net has the oracle property in the sense of Fan and Li (2001).

Remark 4. As a special case, if we let $\lambda_2 = 0$, Theorem 2 and Theorem 3 also suggest that the nonnegative adaptive lasso enjoys the variable selection consistency and the asymptotical normality,

$$\boldsymbol{\alpha}^T\left\{\sqrt{n}\mathbf{C}_{11}^{1/2}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right)\right\} \to_d N\left(0, \sigma^2\right),$$

which is coincides with the theoretical results of Yang and Wu (2016). Otherwise, λ_2 serves as the ridge parameter to tackle the problem of collinearity and, hence, improves the prediction.

4. Estimation algorithm

In this section, we extend the multiplicative updates algorithm, proposed by Sha et al. (2007), to compute the solution of the nonnegative adaptive elastic-net.

For simplicity, we first define the naive nonnegative adaptive elastic-net estimator

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \geqslant \mathbf{0}}{\arg\min}\left\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\sum_{j=1}^{p_n}\hat{\omega}_j\beta_j\right\}, \tag{8}$$

for nonnegative parameters λ_1 and λ_2. It is easy to see that, $\hat{\boldsymbol{\beta}} = (1 + \lambda_2/n)\tilde{\boldsymbol{\beta}}$.

For given λ_1 and λ_2, the naive nonnegative adaptive elastic-net is the solution of the following quadratic programing problem:

$$\begin{cases} \text{minimize } F(\boldsymbol{\beta}) = \boldsymbol{\beta}^T\left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)\boldsymbol{\beta} + \left(\lambda_1\hat{\boldsymbol{\omega}} - 2\mathbf{X}^T\mathbf{y}\right)^T\boldsymbol{\beta} \\ \text{subject to } \boldsymbol{\beta} \geqslant \mathbf{0}. \end{cases} \tag{9}$$

We first denote

$$\left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)_{ij}^+ = \begin{cases} \left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)_{ij} & \text{if } \left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

and

$$\left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)_{ij}^- = \begin{cases} \left|\left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)_{ij}\right| & \text{if } \left(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}\right)_{ij} < 0 \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Let $F_a(\boldsymbol{\beta}) = \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^+\boldsymbol{\beta}, F_b(\boldsymbol{\beta}) = (\lambda_1\hat{\boldsymbol{\omega}} - 2\mathbf{X}^T\mathbf{y})^T\boldsymbol{\beta}$, and $F_c(\boldsymbol{\beta}) = \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^-\boldsymbol{\beta}$. Furthermore, we denote $a_i = \frac{\partial F_a}{\partial \beta_i} = 2((\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^+\boldsymbol{\beta})_i, b_i = \frac{\partial F_b}{\partial \beta_i} = (\lambda_1\hat{\boldsymbol{\omega}} - 2\mathbf{X}^T\mathbf{y})_i$, and $c_i = \frac{\partial F_c}{\partial \beta_i} = 2((\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^-\boldsymbol{\beta})_i$. Then, the iterative steps are:

$$\left[\frac{-b_i + \left(b_i^2 + 4a_ic_i\right)^{\frac{1}{2}}}{2a_i}\right]\beta_i^{(m)} \to \beta_i^{(m+1)}. \tag{12}$$

Table 1. Simulation results for Example 2.

Example	Method	PMSE	C
	NALasso	2.865 (0.540)	162
1	NEnet	3.476 (0.485)	104
	NAEnet	2.808 (0.507)	164
	NALasso	2.412 (0.340)	185
2	NEnet	2.552 (0.382)	168
	NAEnet	2.410 (0.332)	185
	NALasso	2.980 (0.470)	356
3	NEnet	4.102 (0.534)	203
	NAEnet	2.917 (0.459)	356
	NALasso	2.458 (0.376)	384
4	NEnet	2.738 (0.410)	345
	NAEnet	2.377 (0.373)	385
	NALasso	3.207 (0.652)	139
5	NEnet	3.964 (0.615)	85
	NAEnet	3.151 (0.630)	143
	NALasso	3.428 (0.672)	324
6	NEnet	4.398 (0.834)	116
	NAEnet	3.313 (0.616)	330
	NALasso	2.845 (0.430)	161
7	NEnet	3.529 (0.499)	105
	NAEnet	2.828 (0.426)	163
	NALasso	2.591 (0.389)	177
8	NEnet	2.719 (0.436)	159
	NAEnet	2.569 (0.387)	178

We iterate the above steps until convergence, and denote the final solution as $\tilde{\boldsymbol{\beta}}$, then $\hat{\boldsymbol{\beta}} = (1+\lambda_2/n)\tilde{\boldsymbol{\beta}}$ is the solution of the nonnegative adaptive elastic-net.

5. Simulation studies

5.1. Example 1

In this subsection, our aim is to illustrate how the NAEI condition affects the consistency result of the nonnegative adaptive elastic-net. Consider the following two cases:

case 1: we simulated 100 data sets consisting of 400 predictors, where $x_1, \ldots, x_{399}$ are i.i.d. random variables from Gaussian distribution with mean 0 and variance 1, and x_{400} is generated according to

$$x_{400} = \frac{1}{8}x_1 + \frac{1}{8}x_2 + \frac{1}{8}x_3 + \frac{1}{8}x_4 + \frac{1}{8}x_5 + \frac{1}{8}x_6 + \frac{3}{8}x_7 + \frac{7}{8}e, \tag{13}$$

where $e \sim N(0,1)$.

case 2: we considered the same setup as in case 1, except that x_{400} is generated in a different way

$$x_{400} = \frac{7}{8}x_1 + \frac{3}{8}x_2 + \frac{1}{8}x_3 + \frac{1}{8}x_4 + \frac{1}{8}x_5 + \frac{1}{8}x_6 + \frac{1}{8}x_7 + \frac{1}{8}e. \tag{14}$$

In both cases, the response is generated by $y = \sum_{j=1}^{400} x_j\beta_j + \varepsilon$, where $\varepsilon \sim N(0,1)$ and the true parameters are given by $\beta_1^* = 2, \beta_2^* = 3, \beta_3^* = 4, \beta_4^* = 1$, and $\beta_j^* = 0$, for $j = 5, \ldots, 400$. By simple algebra calculation, if we choose $\lambda_2 = 0.01$, it is easy to check that the NAEI condition holds for case 1 and fails for case 2, respectively. If we varying the

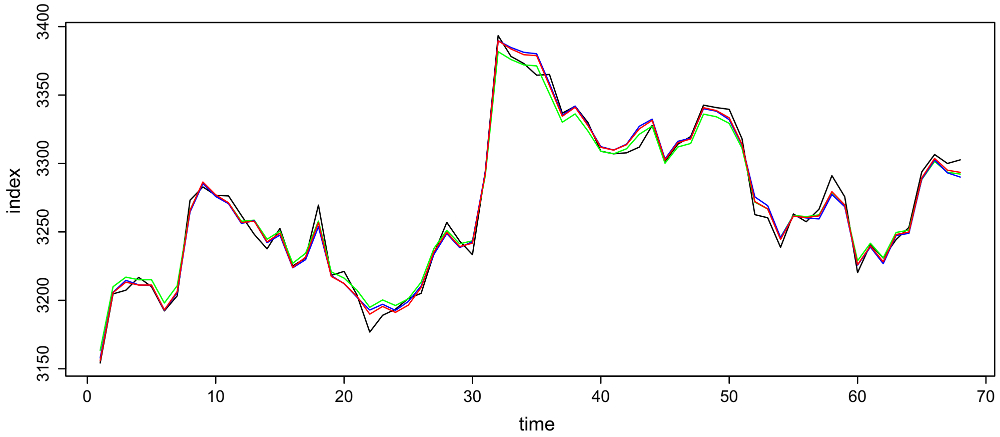


Figure 2. Fitted results for real data: CSI 300 index (black line), nonnegative adaptive lasso (green line), nonnegative elastic-net (blue line), nonnegative adaptive elastic-net (red line).

amount of regularization $||\hat{\boldsymbol{\beta}}||_1$, which is controlled by λ_1, from 0 to 1000 by step 0.5, we get the the nonnegative adaptive elastic-net solution paths for these two cases in Figure 1.

Clearly, in case 1 when the NAEI condition holds, the nonnegative adaptive elastic-net correctly selects the true variables, and shrink other irrelevant variables to zeros. But, in case 2 when the NAEI condition fails, the nonnegative adaptive elastic-net not only selects the true variables but also selects other irrelevant variables like x_{400}, hence it does not satisfy variable selection consistence.

5.2. Example 2

In this subsection, we present simulations to study the finite performance of the nonnegative adaptive elastic-net. We consider three methods in the simulation study: nonnegative adaptive lasso (NALasso), nonnegative elastic-net (NEnet) and nonnegative adaptive elastic-net (NAEnet). In our implementation, we let $\lambda_2 = 0$ in the nonnegative adaptive elastic-net to get the nonnegative adaptive lasso fit. There are several commonly used tuning parameter selection methods, such as k-fold cross-validation (CV), generalized cross-validation (GCV), AIC and BIC. In this work, we use 5-fold CV to select the tuning parameters for each method. For sake of evaluation, the estimation accuracy is measured by PMSE (prediction mean squared error). The variable selection performance is gauged by (C, IC), where C is the number of zero coefficients that are correctly estimated by zero and IC is the number of nonzero coefficients that are incorrectly estimated by zero.

We simulate data from the linear regression model

$$y = \mathbf{x}^T \boldsymbol{\beta}^* + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2\right). \tag{15}$$

Eight different examples with $p_n > n$ are considered, representing eight commonly encountered scenarios. In each example, the covariate vector $\mathbf{x}$ is generated from a multivariate normal distribution whose marginal distributions are $N(0,\ 1)$ and whose

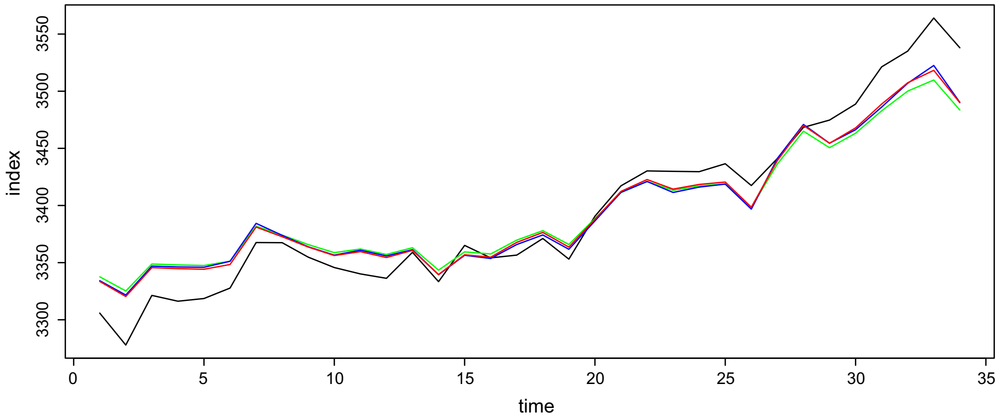


Figure 3. Predicted results for real data: CSI 300 index (black line), nonnegative adaptive lasso (green line), nonnegative elastic-net (blue line), nonnegative adaptive elastic-net (red line).

covariance matrix is given in the description below. Similar models were also considered in Huang et al. (2008). We consider following eight examples.

Example 1. $p = 200$ and $\sigma = 1.5$. The first 15 covariates $(x_1, ..., x_{15})$ and the remaining 185 covariates $(x_{16}, ..., x_{200})$ are independent. The pairwise correlation between the ith and the jth components of $(x_1, ..., x_{15})$ is $r^{|i-j|}$ with $r = 0.5$, $i, j = 1, ..., 15$. The pairwise correlation between the ith and the jth components of $(x_{16}, ..., x_{200})$ is $r^{|i-j|}$ with $r = 0.5$, $i, j = 16, ..., 200$. Components 1–5 of β^* are 2.5, components 6–10 are 1.5, components 11–15 are 0.5, and the rest are 0.

Example 2. The same as Example 1, except that $r = 0.95$.

Example 3. The same as Example 1, except that $p_n = 400$.

Example 4. The same as Example 2, except that $p_n = 400$.

Example 5. $p = 200$ and $\sigma = 1.5$. The predictors are generated as follows:

$$\begin{aligned}
&x_i = z_1 + e_i, \quad z_1 \sim N(0, 1), \quad i = 1, ..., 5;\\
&x_i = z_2 + e_i, \quad z_2 \sim N(0, 1), \quad i = 6, ..., 10;\\
&x_i = z_3 + e_i, \quad z_3 \sim N(0, 1), \quad i = 11, ..., 15;\\
&x_i \sim N(0, 1), \ \ x_i \ \ \text{i.i.d.} \qquad\quad i = 16, ..., 200,
\end{aligned}$$

where e_i are i.i.d. $N(0, 0.01), i = 1, ..., 15$. The first 15 components of β^* are 1.5, the rest are 0.

Example 6. The same as Example 5, except that $p = 400$.

Example 7. $p = 200$ and $\sigma = 1.5$. The pairwise correlation between the ith and the jth components is $r^{|i-j|}$ with $r = 0.5$, $i, j = 1, ..., 200$. Components 1–5 of β^* are 2.5, components 6–10 are 1.5, components 11–15 are 0.5, and the rest are 0.

Example 8. The same as Example 7, except that $r = 0.95$.

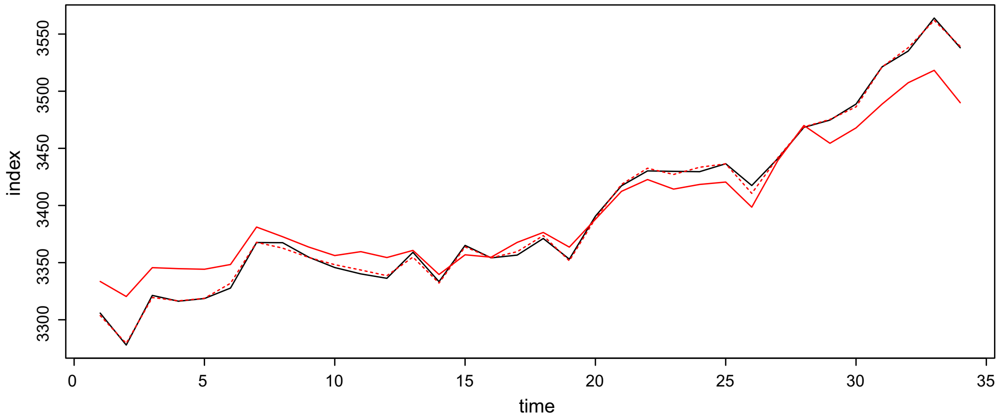


Figure 4. The comparison of the two-stage method and the nonnegative adaptive elastic-net: CSI 300 index (black line), nonnegative adaptive elastic-net (red line), nonnegative adaptive elastic-net + nonnegative least squares (red dotted line).

Among the eight examples, the partial orthogonality assumption in condition (A2) is satisfied in Examples 1–6. To be specific, the correlation among covariates is moderate in Examples 1 and 3; Examples 2 and 4 have strongly correlated covariates; while Examples 5 and 6 are follow from the grouping effects in Zou and Hastie (2005) with three equally important groups, where the correlation among the same group is as high as 0.99. Examples 7 and 8 are the cases where the partial orthogonality assumption is violated; that is to say, covariates with nonzero coefficients are correlated with those with zero coefficients. From Theorem 1, we can further conclude that the NAEI condition is satisfied in Examples 1–6, and violated in Examples 7 and 8.

The simulations are repeated 200 times randomly. Within each replication, our simulated data consists of a training set and a test set, each of size 100, and tuning parameters are selected on the training set only. After tuning parameter selection, the estimators are also computed on the training set. We then compute the PMSE, $\sum_{i=1}^{100}(\hat{y}_i - y_i)^2/100$, for the test set, based on the training set estimators. The medians of PMSE's, C's, and IC's from 200 replications are displayed in Table 1, and the corresponding estimated standard deviations are given in the parentheses. Since the medians of IC's are all zeros, we opt to not present these values for convenience. From the results, we see that for Examples 1–6, nonnegative adaptive elastic-net yields smaller PMSE with better predictive performance than the other two methods. Furthermore, nonnegative adaptive elastic-net has almost same variable selection result as nonnegative adaptive lasso, mainly because that the sparsity of both methods is only controlled by the the adaptive lasso shrinkage. Meanwhile, nonnegative adaptive elastic-net has bigger 'C' which leads to better variable selection results than nonnegative elastic-net, which do not have the oracle property in high-dimensional settings. It is an encouraging result, even when the NAEI condition is violated, nonnegative adaptive elastic-net is still better than the other two methods for both variable selection and estimation. Overall, nonnegative adaptive elastic-net deals with the collinearity problem better than nonnegative adaptive lasso and nonnegative elastic-net in high-dimensional settings.

6. A real data example

In this section, we focus on the application of the nonnegative adaptive elastic-net in financial market. The performance of the nonnegative adaptive elastic-net is tested to track the CSI 300 index, which is a important stock market index designed to replicate the performance of 300 A-shares in the Shanghai and Shenzhen stock exchanges.

Index tracking is a popular form of passive fund management (Connor and Leland 1995; Franks 1992; Jacobs and Levy 1996; Jobst et al. 2001; Roll 1992; Toy and Zurack 1989, etc.), which aims to replicate the performance of an target index, but without purchasing all of the stocks that make up the index. Meanwhile, the weights in the prices relationship between the target index and the constituent stocks are always positive. Thus, the nonnegative adaptive elastic-net is an appropriate method to achieve these goals.

Our data consists of the prices of stocks in the CSI 300 index, from 1 July 2016 to 30 November 2016. There are 102 observations, and 300 predictors in this data. In this work, we let x_{ij} be the prices of the jth constituent stock and y_i be the CSI 300 index. Furthermore, we split the data into two parts, the first 2/3 observations are used as a training set and the remaining 1/3 observations are used as a test set. Thus, the statistical model built on the training set is a typical high-dimensional model. Then we can describe the relationship between x_{ij} and y_i by a linear regression model:

$$y_i = \sum_{j=1}^{300} x_{ij}\beta_j^* + \varepsilon_i, \quad i = 1, \ldots, 68, \quad s.t. \quad \beta_j^* \geq 0. \tag{16}$$

Wu et al. (2014) and Yang and Wu (2016) have already tested the performance of nonnegative elastic-net and nonnegative adaptive lasso for index tracking. In their works, however, x_{ij} and y_i represent the returns of the jth constituent stock and the index respectively. Since the prices is more simple, straightforward and intuitive than the returns, we prefer to directly use the prices of the constituent stocks and the index itself in model (16). Meanwhile, it is important to note that the constituent stocks of the CSI 300 index change regularly every half a year. Consequently, our data only cover half year' observations.

In this work, we want to select a small subset (say size 30) of the constituent stocks to match the performance of index. Thus, we need to tune the two regularization parameters λ_1 and λ_2 for the nonnegative adaptive elastic-net. Typically, we first pick a (relatively small) grid of values for λ_2, say (0.0001, 0.001, 0.01, 0.1, 1, 10, 100). Then, for each λ_2, we adapt a strategy like bisection method to find a λ_1 that can select suitable number of constituent stocks. Finally, the chosen λ_2 is the one giving the smallest mean squared error between the fitted values and the responses in the training set. Using the chosen regularization parameters and the model estimated based on the training set, we compute the PMSE for the test set.

For comparison, we also apply nonnegative adaptive lasso and nonnegative elastic-net for index tracking. We first show the fitted results with 30 selected stocks in Figure 2. It can be seen that the nonnegative adaptive elastic-net outperforms the other two methods in fitting the index. Similar conclusions can be drawn from the predicted results in Figure 3. Furthermore, we calculate the PMSE's to accurately compare the prediction performance and find that the PMSE's of nonnegative adaptive lasso, nonnegative elastic-net and nonnegative adaptive elastic-net are 573.593, 457.488, and 432.094, respectively. However, we observe that the fitted results are much better than the predicted results. To improve the prediction accuracy, we suggest to use the nonnegative adaptive elastic-net only to select important variables and then apply the nonnegative least squares (Slawski and Hein 2013) to estimate the weights of the chosen stocks. Figure 4 shows prediction results of the this two-stage method, which obviously improves the performance of the nonnegative adaptive elastic-net. As an example, the results obtained have well demonstrated the utilization and usefulness of our new method for tracking the CSI 300 index.

7. Concluding remarks

In this paper, we propose the nonnegative adaptive elastic-net for sparse high-dimensional linear regression models. Under certain appropriate conditions, we prove that the nonnegative adaptive elastic-net enjoys the oracle property. Some simulation results and real data analysis confirm that the proposed method works competitively compared to other existing methods.

At present, we have studied on the nonnegative adaptive elastic-net in the context of linear regression models. It is an important and challenging problem to generalize the results of this paper to more complicated models such as generalized linear models and Cox models. But such an extension is by no means of trivial and needs additional investigations in the future.

Appendix: Proofs of theorems

Proof of Theorem 2. As stated in Sec. 4, we have $\hat{\boldsymbol{\beta}} = (1+\lambda_2/n)\tilde{\boldsymbol{\beta}}$. Thus, we only need to show the variable selection consistency of the naive nonnegative adaptive elastic-net.

Recalling that the naive nonnegative adaptive elastic-net estimator is defined as follow:

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \geqslant \mathbf{0}}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{j=1}^{p_n} \hat{\omega}_j \beta_j \right\}. \tag{A.1}$$

By KKT (Karush-Kuhn-Tucker) conditions, $\tilde{\boldsymbol{\beta}}$ is the naive nonnegative adaptive elastic-net estimator for given λ_1 and λ_2 if and only if

$$\begin{cases} -2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right) + \lambda_1 \hat{\boldsymbol{\omega}} + 2\lambda_2 \tilde{\boldsymbol{\beta}} - \boldsymbol{\gamma} = 0, \\ \boldsymbol{\gamma} \geq \mathbf{0}, \tilde{\boldsymbol{\beta}} \geqslant \mathbf{0}, \boldsymbol{\gamma}^T \tilde{\boldsymbol{\beta}} = 0. \end{cases} \tag{A.2}$$

Replace $\mathbf{y}$ by $\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, then (A.2) can be rewritten as

$$\begin{cases} -2\mathbf{X}^T\left[\mathbf{X}\left(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}\right) + \boldsymbol{\varepsilon}\right] + \lambda_1 \hat{\boldsymbol{\omega}} + 2\lambda_2 \tilde{\boldsymbol{\beta}} - \boldsymbol{\gamma} = 0, \\ \boldsymbol{\gamma} \geqslant \mathbf{0}, \tilde{\boldsymbol{\beta}} \geqslant \mathbf{0}, \boldsymbol{\gamma}^T \tilde{\boldsymbol{\beta}} = 0. \end{cases} \tag{A.3}$$

If there exists $\tilde{\boldsymbol{\beta}}$ satisfies (A.3) and $\tilde{\boldsymbol{\beta}}_1 \succ \mathbf{0}, \tilde{\boldsymbol{\beta}}_2 = \mathbf{0}$, then $\hat{S} = S$. Thus, the existence of such $\hat{\boldsymbol{\beta}}$ satisfies variable selection consistency is implied by

$$\begin{cases} -2\mathbf{X}_1^T[\mathbf{X}_1(\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1)+\boldsymbol{\varepsilon}]+\lambda_1\hat{\boldsymbol{\omega}}_1-2\lambda_2\tilde{\boldsymbol{\beta}}_1=\mathbf{0},\\ -2\mathbf{X}_2^T[\mathbf{X}_1(\boldsymbol{\beta}_1^*-\tilde{\boldsymbol{\beta}}_1)+\boldsymbol{\varepsilon}]+\lambda_1\hat{\boldsymbol{\omega}}_2-\gamma_2=\mathbf{0},\\ \gamma_1=\mathbf{0},\ \gamma_2=\mathbf{0},\ \tilde{\boldsymbol{\beta}}_1>\mathbf{0},\ \tilde{\boldsymbol{\beta}}_2=\mathbf{0},\ \gamma^T\boldsymbol{\beta}=0. \end{cases} \tag{A.4}$$

From the first equation in (A.4), we have

$$\tilde{\boldsymbol{\beta}}_1=\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\left(\mathbf{C}_{11}\boldsymbol{\beta}_1^*+\frac{1}{n}\mathbf{X}_1^T\boldsymbol{\varepsilon}-\frac{\lambda_1}{2n}\hat{\boldsymbol{\omega}}_1\right). \tag{A.5}$$

It follows that the existence of such $\tilde{\boldsymbol{\beta}}$ satisfies variable selection consistency is implied by

$$\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{W}_1\succ\frac{\lambda_1}{2\sqrt{n}}\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\hat{\boldsymbol{\omega}}_1-\sqrt{n}\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{C}_{11}\boldsymbol{\beta}_1^*,$$

and

$$\begin{aligned}\mathbf{C}_{21}\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{W}_1-\mathbf{W}_2\succcurlyeq{}&\frac{\lambda_1}{2\sqrt{n}}\left[\mathbf{C}_{21}\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\hat{\boldsymbol{\omega}}_1-\hat{\boldsymbol{\omega}}_2\right]\\ &+\mathbf{C}_{21}\left[I-\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{C}_{11}\right]\boldsymbol{\beta}_1^*,\end{aligned}$$

where $\mathbf{W}=\mathbf{X}^T\boldsymbol{\varepsilon}/\sqrt{n}$, and $\mathbf{W}_1,\mathbf{W}_2$ are as the first q_n and last p_n-q_n columns of $\mathbf{W}$, respectively. Under the conditions (C3)-(C5), we have

$$g_j=\frac{E\left(\mathbf{x}_j^T\mathbf{y}\right)}{n}=\frac{\mathbf{x}_j^T\left(\mathbf{X}_1\boldsymbol{\beta}_1^*\right)}{n}\ge M_1M_3n^{(c_1+c_2-1)/2},\qquad j=1,\dots,p_n. \tag{A.6}$$

Then, by the r_n-consistency of $\hat{\boldsymbol{\beta}}^0$, we have

$$\max_{j\le p_n}\left|\frac{|\hat{\beta}_j^0|}{|g_j|}-1\right|\le(M_1M_3)^{-1}n^{(1-c_1-c_2)/2}O(1/r_n)=o_p(1), \tag{A.7}$$

where $r_n=O(n^{\frac{1-c_3}{2}})$ can be derived from condition (C2) and the orders of p_n and q_n. It follows that

$$|\hat{\beta}_j^0|=\left(1+o_p(1)\right)|g_j|,\quad j=1,\dots,p_n, \tag{A.8}$$

which leads to

$$|\hat{\omega}_j|=|\hat{\beta}_j^0|^{-1}=\left(1+o_p(1)\right)|g_j|^{-1},\quad j=1,\dots,p_n. \tag{A.9}$$

For simplicity, we set

$$\boldsymbol{\xi}=\left(\xi_1,\dots,\xi_{q_n}\right)^T=\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{W}_1, \tag{A.10}$$

and

$$\boldsymbol{\varsigma}=\left(\varsigma_1,\dots,\varsigma_{p_n-q_n}\right)^T=\mathbf{C}_{21}\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{W}_1-\mathbf{W}_2. \tag{A.11}$$

Recalling the NAEI condition and $\lambda_2/n\to0$ as $n\to\infty$, we then have

$$\begin{aligned}&\left\{\boldsymbol{\varsigma}\preccurlyeq\frac{\lambda_1}{2\sqrt{n}}\left[\mathbf{C}_{21}\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\hat{\boldsymbol{\omega}}_1-\hat{\boldsymbol{\omega}}_2\right]+\mathbf{C}_{21}\left[I-\left(\mathbf{C}_{11}+\frac{\lambda_2}{n}\mathbf{I}\right)^{-1}\mathbf{C}_{11}\right]\boldsymbol{\beta}_1^*\right\}\\ &\subset\left\{\boldsymbol{\varsigma}\preccurlyeq\frac{\lambda_1}{2\sqrt{n}}\boldsymbol{\eta}^*\right\}\end{aligned}$$

where $\eta_j^*=\eta_j+o(1)>0$, for $j=1,2,\dots,p_n-q_n$.

Then through probability theory knowledge we have

$$P(\hat{S} \neq S) \leq \sum_{i=1}^{q_n} P\left\{ \xi_i \leq -\sqrt{n} e_i^T \left[\left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \mathbf{C}_{11} \boldsymbol{\beta}_1^* \right] + \frac{\lambda_1}{2\sqrt{n}} e_i^T \left[\left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \hat{\boldsymbol{\omega}}_1 \right] \right\}$$
$$+ \sum_{i=1}^{p_n - q_n} P\left\{ \varsigma_i \leq \frac{\lambda_1}{2\sqrt{n}} \eta_i^* \right\},$$

where e_i denotes the vector with ith entry 1 and otherwise 0.

Suppose that $H_a^T = (h_1^a, \ldots, h_{q_n}^a)^T = n^{-\frac{1}{2}} (\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I})^{-1} \mathbf{X}_1^T$, then $\boldsymbol{\xi} = H_a^T \boldsymbol{\varepsilon}$. Thus we have:

$$\|h_i^a\|_2^2 = e_i^T H_a^T H_a e_i = e_i^T \left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \mathbf{C}_{11} \left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} e_i \leq e_i^T \mathbf{C}_{11}^{-1} e_i \leq \frac{1}{M_1}, \tag{A.12}$$

where the last inequality comes from the condition (C4).

Similarly, suppose that $H_b^T = (h_1^b, \ldots, h_{p_n - q_n}^b)^T = n^{-\frac{1}{2}} \mathbf{C}_{21} (\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I})^{-1} \mathbf{X}_1^T - n^{-\frac{1}{2}} \mathbf{X}_2^T$, then $\varsigma = H_b^T \boldsymbol{\varepsilon}$. From the condition (C3), we can get that:

$$\begin{aligned} \|h_i^b\|_2^2 = e_i^T H_b^T H_b e_i &\leq e_i^T \left\{ \frac{1}{n} \mathbf{X}_2^T \left(\mathbf{I} - \mathbf{X}_1 \left(\mathbf{X}_1^T \mathbf{X}_1 + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \mathbf{X}_1^T \right) \mathbf{X}_2 \right\} e_i \\ &\leq \frac{1}{n} \|\mathbf{X}_2 e_i\|_2^2 \leq \frac{1}{n} \|X^i\|_2^2 \leq 1, \end{aligned} \tag{A.13}$$

where X^i is the ith column of $\mathbf{X}_2$.

Since ε_i's are independent identically distributed Gaussian variables, therefore by (A.12) and (A.13), ξ_i's, ς_i's are Gaussian variables with bounded second moments. Let $\Phi(t)$ denote the distribution of the standard Gaussian variable, then for any $t > 0$,

$$1 - \Phi(t) \leq t^{-1} e^{-\frac{1}{2} t^2}. \tag{A.14}$$

By (A.6), (A.9), and $\lambda_1 \propto n^{\frac{1+c_4}{2}}$, we have

$$\frac{\lambda_1}{2\sqrt{n}} e_i^T \left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \hat{\boldsymbol{\omega}}_1 \leq \frac{\lambda_1}{2\sqrt{n} M_1} (M_1 M_3)^{-1} n^{(1 - c_1 - c_2)/2} = o\left(n^{\frac{c_2}{2}}\right). \tag{A.15}$$

Therefore,

$$\begin{aligned} &\sum_{i=1}^{q_n} P\left\{ \xi_i \leq -\sqrt{n} e_i^T \left[\left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \mathbf{C}_{11} \boldsymbol{\beta}_1^* \right] + \frac{\lambda_1}{2\sqrt{n}} e_i^T \left[\left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \hat{\boldsymbol{\omega}}_1 \right] \right\} \\ &\quad \leq \sum_{i=1}^{q_n} P\left\{ |\xi_i| \geq \sqrt{n} e_i^T \left[\left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \mathbf{C}_{11} \boldsymbol{\beta}_1^* \right] - \frac{\lambda_1}{2\sqrt{n}} e_i^T \left[\left(\mathbf{C}_{11} + \frac{\lambda_2}{n} \mathbf{I} \right)^{-1} \hat{\boldsymbol{\omega}}_1 \right] \right\} \\ &\quad = q_n \cdot O\left(1 - \Phi\left((1 + o(1)) \sqrt{M_1} M_3 n^{\frac{c_2}{2}} \right) \right) \\ &\quad = o\left(e^{-n^{c_3}}\right), \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{p_n - q_n} P\left\{ \varsigma_i \leq \frac{\lambda_1}{2\sqrt{n}} \eta_i^* \right\} &\leq \sum_{i=1}^{p_n - q_n} P\left\{ |\varsigma_i| \geq \frac{\lambda_1}{2\sqrt{n}} \eta_i^* \right\} \\ &= (p_n - q_n) \cdot O\left(1 - \Phi\left(\frac{\lambda_1}{2\sqrt{n}} \eta_i^* \right) \right) \\ &= o\left(e^{-n^{c_3}}\right). \end{aligned}$$

Sum the above two terms, we have $P(\hat{S} \neq S) = 0$, as $\rightarrow \infty$. This completes the proof for variable selection consistency.

Proof of Theorem 3. We now prove the asymptotic normality. For convenience, we write

$$z_n = \boldsymbol{\alpha}^T\left\{\sqrt{n}\frac{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}}{1+\lambda_2/n}\mathbf{C}_{11}^{1/2}\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right)\right\}. \tag{A.16}$$

Recalling that $\hat{\boldsymbol{\beta}} = (1+\lambda_2/n)\tilde{\boldsymbol{\beta}}$, then we have

$$\begin{aligned} z_n &= \sqrt{n}\boldsymbol{\alpha}^T\{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}\}\mathbf{C}_{11}^{1/2}\tilde{\boldsymbol{\beta}}_1 - \sqrt{n}\boldsymbol{\alpha}^T\frac{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}}{1+\lambda_2/n}\mathbf{C}_{11}^{1/2}\boldsymbol{\beta}_1^* \\ &= \sqrt{n}\boldsymbol{\alpha}^T\{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}\}\mathbf{C}_{11}^{1/2}\left(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right) + \sqrt{n}\boldsymbol{\alpha}^T\{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}\}\frac{\lambda_2}{n+\lambda_2}\mathbf{C}_{11}^{1/2}\boldsymbol{\beta}_1^*. \end{aligned} \tag{A.17}$$

In addition, by (A.5) in Theorem 2, which follows that $z_n = T_1 + T_2 + T_3$, where

$$T_1 = -\boldsymbol{\alpha}^T\left\{\frac{\lambda_2}{\sqrt{n}}\mathbf{C}_{11}^{-1/2}\boldsymbol{\beta}_1^*\right\}, \tag{A.18}$$

$$T_2 = \boldsymbol{\alpha}^T\left\{\mathbf{C}_{11}^{-1/2}\left(\mathbf{W}_1 - \frac{\lambda_1}{2\sqrt{n}}\hat{\boldsymbol{\omega}}_1\right)\right\}, \tag{A.19}$$

$$T_3 = \sqrt{n}\boldsymbol{\alpha}^T\{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}\}\frac{\lambda_2}{n+\lambda_2}\mathbf{C}_{11}^{1/2}\boldsymbol{\beta}_1^*. \tag{A.20}$$

We now show that $T_1 \to 0, T_3 \to 0$, as $n \to \infty$, and $T_2 \to_d N(0,\sigma^2)$. Then, by Slutsky's theorem, we know $z_n \to_d N(0,\sigma^2)$.

By condition (C4), we have

$$T_1^T T_1 \le \frac{1}{M_1}\frac{\lambda_2^2}{n}\sum_{j\in S}\beta_j^{*2}. \tag{A.21}$$

Hence, it follows the choice of λ_2 that $T_1 \to 0$ as $n \to \infty$. Similarly, we can bound T_3 as follows:

$$\begin{aligned} T_3^T T_3 &\le 2\|\sqrt{n}\boldsymbol{\alpha}^T\{\mathbf{I}+(\lambda_2/n)\mathbf{C}_{11}^{-1}\}\frac{\lambda_2}{n+\lambda_2}\mathbf{C}_{11}^{1/2}\boldsymbol{\beta}_1^*\|_2^2 \\ &\le 2nM_2\left(\frac{\lambda_2}{n+\lambda_2}\right)^2\left(1+\frac{\lambda_2}{nM_1}\right)^2\|\boldsymbol{\beta}_1^*\|_2^2 = O\left(\frac{\lambda_2^2}{n}\right)\sum_{j\in S}\beta_j^{*2}, \end{aligned} \tag{A.22}$$

which also tells us that $T_3 \to 0$ as $n \to \infty$.

Since $\varepsilon_i \sim N(0,\sigma^2)$, then

$$\boldsymbol{\alpha}^T\mathbf{C}_{11}^{-1/2}\mathbf{W}_1 = \boldsymbol{\alpha}^T\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1/2}\mathbf{X}_1^T\boldsymbol{\varepsilon}\sim N\left(0,\sigma^2\right). \tag{A.23}$$

Moreover, by (A.6), (A.9), and $\lambda_1 \propto n^{\frac{1+c_4}{2}}$, we have

$$\left(\frac{\lambda_1}{\sqrt{n}}\boldsymbol{\alpha}^T\mathbf{C}_{11}^{-1/2}\hat{\boldsymbol{\omega}}_1\right)^T\left(\frac{\lambda_1}{\sqrt{n}}\boldsymbol{\alpha}^T\mathbf{C}_{11}^{-1/2}\hat{\boldsymbol{\omega}}_1\right) \le \left(M_1^3M_3^2\right)^{-1}n^{(1-c_1-c_2+c_4)}. \tag{A.24}$$

Since $c_1 + c_2 > 1 + c_4$, we have

$$\frac{\lambda_1}{\sqrt{n}}\boldsymbol{\alpha}^T\mathbf{C}_{11}^{-1/2}\hat{\boldsymbol{\omega}}_1 \to 0, \quad \text{as } n\to\infty, \tag{A.25}$$

then by Slutsky's theorem

$$T_2 \to_d N\left(0,\sigma^2\right). \tag{A.26}$$

This completes the proof for Theorem 3.

Funding

Ning Li and Hu Yang's research is supported by the National Natural Science Foundation of China (Grant No. 11671059), and Jing Yang's research is supported by the National Natural Science Foundation of China (Grant No. 11801168).

References

Akaike, H. 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* 60 (2):255–65. doi:10.1093/biomet/60.2.255.

Breiman, L. 1995. Better subset regression using the nonnegative garrote. *Technometrics* 37 (4): 373–84. doi:10.1080/00401706.1995.10484371.

Connor, G., and H. Leland. 1995. Cash management for index tracking. *Financial Analysts Journal* 51 (6):75–80. doi:10.2469/faj.v51.n6.1952.

Fan, J., and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 (456):1348–60. doi:10.1198/016214501753382273.

Franks, E. 1992. Targeting excess-of-benchmark returns. *The Journal of Portfolio Management* 18 (4):6–12. doi:10.3905/jpm.1992.409419.

Huang, J., S. G. Ma, and C. H. Zhang. 2008. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* 18:1603–18.

Jacobs, B., and K. Levy. 1996. Residual risk: How much is too much. *The Journal of Portfolio Management* 22 (3):10–5. doi:10.3905/jpm.1996.10.

Jobst, N., M. Horniman, C. Lucas, and G. Mitra. 2001. Computational aspects of alternative portfolio selection models in the presence of discrete asset choice constraints. *Quantitative Finance* 1 (5):1–13. doi:10.1088/1469-7688/1/5/301.

Meinshausen, N. 2013. Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics* 7:1607–31. doi:10.1214/13-EJS818.

Roll, R. 1992. A mean/variance analysis of tracking error. *The Journal of Portfolio Management* 18 (4):13–22. doi:10.3905/jpm.1992.701922.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461–4. doi:10.1214/aos/1176344136.

Sha, F., Y. Lin, L. K. Saul, and D. D. Lee. 2007. Multiplicative updates for nonnegative quadratic programming. *Neural Computation* 19 (8):2004–31. doi:10.1162/neco.2007.19.8.2004.

Slawski, M., and M. Hein. 2013. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics* 7: 3004–56. doi:10.1214/13-EJS868.

Tibshirani, R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B* 58:267–88. doi:10.1111/j.2517-6161.1996.tb02080.x.

Toy, W., and M. Zurack. 1989. Tracking the Euro-Pac index. *The Journal of Portfolio Management* 15 (2):55–8. doi:10.3905/jpm.1989.409186.

Wu, L., and Y. H. Yang. 2014. Nonnegative elastic net and application in index tracking. *Applied Mathematics and Computation* 227:541–52. doi:10.1016/j.amc.2013.11.049.

Wu, L., Y. H. Yang, and H. Z. Liu. 2014. Nonnegative-lasso and application in index trackin. *Computational Statistics & Data Analysis* 70:116–26. doi:10.1016/j.csda.2013.08.012.

Yang, Y. H., and L. Wu. 2016. Nonnegative adaptive lasso for ultra-high dimensionalm regression models and a two-stage method applied in financial modeling. *Journal of Statistical Planning and Inference* 174:52–67. doi:10.1016/j.jspi.2016.01.011.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101 (476):1418–29. doi:10.1198/016214506000000735.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301–20. doi:10.1111/j.1467-9868.2005.00503.x.

Zou, H., and H. Zhang. 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* 37 (4):1733–51. doi:10.1214/08-AOS625.