Journal of Industrial and Management Optimization

doi:10.3934/jimo.2023094

VARIANCE ESTIMATION IN HIGH-DIMENSIONAL LINEAR REGRESSION VIA ADAPTIVE ELASTIC-NET

Xin Wang[1], Lingchen Kong[*1],
Xinying Zhuang[2] and Liqun Wang[3]

[1]School of Mathematics and Statistics, Beijing Jiaotong University, Beijing, China

[2]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

[3]Department of Statistics, University of Manitoba, Winnipeg, Canada

(Communicated by Xinmin Yang)

Abstract. Variance estimation in high-dimensional linear regression is a fundamental problem in statistical learning, and it plays a wide range of roles in signal processing, pattern recognition, and other fields. Because it is difficult to choose the true model precisely in high-dimensional regression, variance estimation remains a challenging problem, especially in scenarios where the true regression parameter has a large number of non-zero elements. In this paper, we develop a novel approach for variance estimation by solving a re-parameterized log-likelihood optimization problem with adaptive elastic-net regularization. It is called the natural adaptive elastic-net (NAEN). The relationship between NAEN and the naive adaptive elastic-net is established. The NAEN inherits the advantages of the naive adaptive elastic-net, that is, it can select and estimate the regression and variance parameters simultaneously. Moreover, we also give the asymptotic properties of NAEN for error variance. The simulation results show that the proposed NAEN is suitable for scenarios where the true regression parameter has many non-zero elements.

1. Introduction.

Consider linear regression model

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}$ is the response variable, $\boldsymbol{x} \in \mathbb{R}^p$ is the predictor variable, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown sparse regression parameter, $\varepsilon \in \mathbb{R}$ is the random error satisfying the normal distribution $N(0, (\sigma^*)^2)$. Given an $i.i.d.$ random sample $\{\boldsymbol{x}_i, y_i\}$, $i = 1, \cdots, n$, the regression model can be written in the following matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$. Here, suppose that $\|\boldsymbol{X}_j\|_2^2 = n$. The purpose of variance estimation is to estimate the variance of the random error ε.

2020 *Mathematics Subject Classification.* Primary: 65-11, 65C20; Secondary: 62J05.

Key words and phrases. Variance estimation, high-dimensional regression, natural adaptive elastic-net.

The work was supported by the National Natural Science Foundation of China (12071022) and the 111 Project of China (B16002).

*Corresponding author: Lingchen Kong.

In conventional linear models, the residual-based estimator plays an important role in statistical inferences and model checking. However, the residual-based estimator performs poorly in high-dimensional linear models, where $p > n$. The reason for this phenomenon is that model selection is a tedious and complex problem. In practice, the true model is difficult to select accurately, and often many irrelevant variables are selected for insurance. Consequently, residuals are actually predicted with many spurious variables so that the resulting estimator will seriously underestimate the error variance.

The following examples demonstrate that variance estimation is involved and plays an important role in statistical learning.

- (Model selection). Regularization estimation is one of the mainstream methods for model selection and parameter estimation in high-dimensional models. The efficiency of this method depends on tuning parameter that can be chosen by some criteria, such as Akaike's information criterion or the Bayesian information criterion. These criteria are closely related to the error variance.
- (Confidence intervals). Let $\widehat{\boldsymbol{\beta}} := (\widehat{\beta}_1, \cdots, \widehat{\beta}_p)^T$ be the regularization estimator with corresponding design matrix $\boldsymbol{X}_{\mathcal{A}}$, where $\mathcal{A} \subset \{1, \cdots, p\}$ is the index set corresponding to non-zero elements of $\widehat{\boldsymbol{\beta}}$, $\boldsymbol{X}_{\mathcal{A}}$ is the sub-matrix of $\boldsymbol{X}$ consisting of its columns associated with index set $\mathcal{A}$. If $\widehat{\boldsymbol{\beta}}$ has the oracle property ([9, 11, 18]), then the $(1-\alpha)$ confidence interval of non-zero element $\widehat{\beta}_i$, $i \in \{1, \cdots, p\}$, is given by

$$[\widehat{\beta}_i - z_{1-\alpha/2} c_i \sigma, \widehat{\beta}_i + z_{1-\alpha/2} c_i \sigma],$$

  where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the conditional normal distribution and c_i is the i-th diagonal element of $(\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}})^{-1}$.

1.1. **Literature review.** In the past ten years, variance estimation in high dimensional linear regression has attracted wide attention.

Here, we review some basic estimation methods and list some representative research. Let $\boldsymbol{\beta}^*$ denote the true regression parameter in (1). Then the ideal oracle estimator (OE) is defined as $\sigma^2_{\text{Oracle}} = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^*)^2/n$. Correspondingly, the naive estimator (NE) is computed by $\sigma^2_{\text{Naive}} = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})^2/n$, where $\widehat{\boldsymbol{\beta}}$ denotes some estimator for regression parameter. As is well-known that NE is biased and a modified unbiased estimator (MUE) is given by $\sigma^2_{\text{M}} = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}})^2/(n - \widehat{s})$, where $\widehat{s}$ is the number of non-zero elements of $\widehat{\boldsymbol{\beta}}$. However, when p is much larger than n and s^* is not very small, a small change in $\widehat{s}$ will lead to large change in the MUE.

For variance estimation in high-dimensional linear regression, Reference [13] constructed a re-parameterized likelihood with L_1 penalty to estimate the regression and variance parameters. Their model can be formulated as

$$(\widehat{\boldsymbol{\phi}}, \widehat{\rho}) = \arg\min_{\boldsymbol{\phi}, \rho} \left\{ \log(\rho) + \frac{\|\rho \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\phi}\|_2^2}{2n} + \lambda_n \|\phi\|_1 \right\},$$

where $\boldsymbol{\phi} = \boldsymbol{\beta}/\sigma$, $\rho = 1/\sigma$. A refitted cross-validation method was designed in [8] to attenuate the influence of irrelevant variables with high spurious correlations via a data-splitting technique. Reference [14] proposed the scaled lasso (SL), which can also estimate regression and variance parameters simultaneously like the L_1

penalized re-parameterized likelihood [13] and as follows:

$$(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}) = \arg\min_{\boldsymbol{\beta},\sigma} \left\{ \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2n\sigma} + \frac{(1-a)\sigma}{2} + \lambda_n \|\boldsymbol{\beta}\|_1 \right\}.$$

Reference [6] designed a moment estimation for error variance. A re-parameterized likelihood (See Section 2) different from [13] was developed in [17], who proposed two estimations: natural lasso (NL) and organic lasso (OL). Referring to [17], a natural adaptive lasso (NAL) for error variance was proposed in [16]. Essentially, NL, OL, and NAL for error variance are the respectively the optimal values of the optimization problems of the lasso [15], the exclusive lasso with a single group [20, 4], and the adaptive lasso [21] for regression parameters.

1.2. **Challenge and motivation.** Although there have been many studieson estimation problem of error variance in the past decade, it remains a challenging issue. We use an example to illustrate that the existing methods often overestimate or underestimate the error variance when the true regression parameter has many non-zero elements and the error variance is large. In this case, model selection is difficult to achieve accurately.

Example 1.1. The predictor matrix $\boldsymbol{X}$ is generated randomly from the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} := (\Sigma_{ij})_{i,j=1}^{p}$, where $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.3$, and the error variance $(\sigma^*)^2 = 1$. Set $n = 100$, $p = 200$ and sparsity $s = 30$. The true regression parameter vector $\boldsymbol{\beta}^* = (1, \cdots, 1, 0, \cdots, 0)$.

We apply SL, NL, OL, and NAL to estimate error variance and repeat 100 simulations for each method. The average mean squared error (AMSE) $\widehat{E}\{(\widehat{\sigma}/\sigma^* - 1)^2\}$ and the average relative ratio (ARR) $\widehat{E}\{\widehat{\sigma}/\sigma^*\}$ are used to evaluate the quality of these methods.

TABLE 1. The AMSE and ARR of SL, NL, OL and NAL

	SL	NL	OL	NAL
AMSE	0.734	14.187	0.114	0.103
ARR	0.147	4.765	1.333	0.731

Table 1 reports AMSE and ARR of four methods and Figure 1 shows all simulation results of these methods. These methods severely underestimate or overestimate the error variance. Specifically, SL and NAL underestimate the error variance, and NAL outperforms other methods. Moreover, OL and NL overestimate the error variance, and each estimate of NL is much larger than the true variance. Indeed, most existing methods only perform well in extremely sparse scenarios (See Section 4.4) or in cases where the dimension p is relatively small and the noise in (1) is small.

In order to obtain accurate estimator of variance parameter, this paper develops a novel approach for linear models where $\boldsymbol{\beta}^*$ has many non-zero elements, called the natural adaptive elastic-net (NAEN). The rest of this paper is organized as follows. The NAEN is introduced in Section 2 and its model is a re-parameterized log-likelihood with adaptive elastic-net. We analyze the relationship between the NAEN and the naive adaptive elastic-net in Section 3. The asymptotic properties of the NAEN for $\boldsymbol{\beta}$ and σ^2 are also analyzed in Section 3. In Section 4, we discuss the

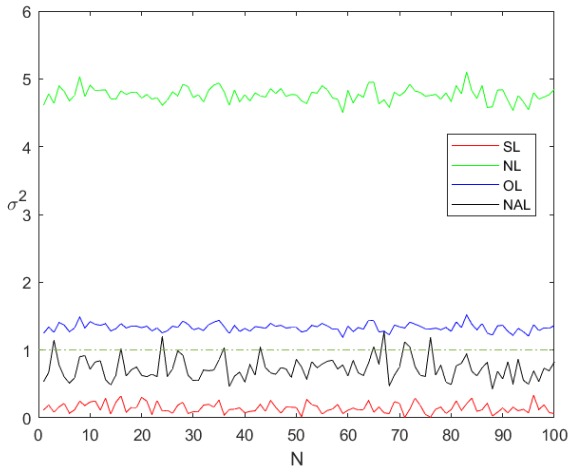

FIGURE 1. All results of SL, NL, OL and NAL

numerical optimization of the NAEN and its the finite-sample performance. The simulation results show that the proposed NAEN is suitable for the linear models where $\boldsymbol{\beta}^*$ has many non-zero elements.

1.3. **Notations.** Throughout the paper, we use the following notation. Let $\mathcal{A}_0$ be the index set of the non-zero elements of $\boldsymbol{\beta}^*$ and $\mathcal{A}_0^c$ is its complement. Without loss of generality, the true regression parameter can be written as

$$\boldsymbol{\beta}^* = ([\boldsymbol{\beta}^*_{\mathcal{A}_0}]^T, [\boldsymbol{\beta}^*_{\mathcal{A}_0^c}]^T)^T,$$

where $\boldsymbol{\beta}^*_{\mathcal{A}_0} \in \mathbb{R}^s$ is the sub-vector of $\boldsymbol{\beta}^*$ consisting of its non-zero elements and $\boldsymbol{\beta}^*_{\mathcal{A}_0^c} = (0, \cdots, 0)^T \in \mathbb{R}^{p-s}$. The number of non-zero elements of $\boldsymbol{\beta}^*$ is s. For vectors $\boldsymbol{v} := (v_1, \cdots, v_p)^T \in \mathbb{R}^p$ and $\boldsymbol{z} := (z_1, \cdots, z_p)^T \in \mathbb{R}^p$, $\|\boldsymbol{v}\|_1 := \sum_{i=1}^p |v_i|$ and $\|\boldsymbol{v}\|_2 := \sum_{i=1}^p v_i^2$ denotes the 1-norm and 2-norm of $\boldsymbol{v}$, respectively, and $\boldsymbol{v} \circ \boldsymbol{z} := (x_i y_i)_{i=1}^p$ denotes the Hadamard product between $\boldsymbol{v}$ and $\boldsymbol{z}$. $\boldsymbol{v}_{\mathcal{A}_0}$ denotes the sub-vector of $\boldsymbol{v}$ consisting of its elements associated with index set $\mathcal{A}_0$. In addition, $\text{sign}(\boldsymbol{v}) = (\text{sign}(v_1), \cdots, \text{sign}(v_p))^T$, where

$$\text{sign}(t) = \begin{cases} 1, & \text{if } t > 0, \\ 0, & \text{if } t = 0, \\ -1, & \text{if } t < 0. \end{cases}$$

Let $\partial\|\boldsymbol{v}\|_1 := (\partial|v_1|, \cdots, \partial|v_p|)^T$ denote the sub-differential set of $\|\cdot\|_1$ at $\boldsymbol{v}$, where

$$\partial|t| = \begin{cases} \{1\}, & \text{if } t > 0, \\ [-1, 1], & \text{if } t = 0, \\ \{-1\}, & \text{if } t < 0. \end{cases}$$

Each vector in $\partial\|\boldsymbol{v}\|_1$ is a sub-gradient of $\|\cdot\|_1$ at $\boldsymbol{v}$. For a $p \times p$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 := \sup_{\|\boldsymbol{v}\|_2 \leq 1} \|\boldsymbol{A}\boldsymbol{v}\|_2$ denotes the spectral norm.

2. **Natural adaptive elastic-net.** This section describes the NAEN for error variance and regression parameter. First, the negative log-likelihood function is given by

$$F(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \boldsymbol{y}) = \frac{n}{2} \log \sigma^2 + \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{2\sigma^2},$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$, $\boldsymbol{y} = (y_1, \cdots, y_n)^T$. Combing the re-parameterization $\boldsymbol{\beta} = \boldsymbol{\theta}/\phi$, $\sigma^2 = 1/\phi$ in [17] with the adaptive elastic-net in [22], the NAEN for $\boldsymbol{\theta}$ and ϕ are defined as the minimizer of the following problem:

$$(\boldsymbol{\theta}, \phi) \in \min_{\phi, \boldsymbol{\theta}} \left\{ L(\boldsymbol{\theta}, \phi) + \lambda_{n,1} \|\boldsymbol{w} \circ \boldsymbol{\theta}\|_1 + \frac{\lambda_{n,2} \|\boldsymbol{\theta}\|_2^2}{\phi} \right\}, \tag{2}$$

where $L(\boldsymbol{\theta}, \phi) := (1/n) F(\phi^{-1}\boldsymbol{\theta}, \phi^{-1})$ is the re-parameterized negative log-likelihood function, $\lambda_{n,1}$ and $\lambda_{n,2}$ are tuning parameters, $\boldsymbol{w} := (w_1, \cdots, w_p)^T$ is the adaptive weight vector. If $(\widehat{\boldsymbol{\theta}}, \widehat{\phi})$ is a minimizer of problem (2), then the NAEN estimators for $\boldsymbol{\beta}$ and σ^2 are defined as

$$\widehat{\boldsymbol{\beta}} = \frac{\widehat{\boldsymbol{\theta}}}{\widehat{\phi}}, \ \widehat{\sigma}^2 = \frac{1}{\widehat{\phi}}. \tag{3}$$

It is clear that the NL in [17] and the NAL in [16] are special cases of the NAEN.

Note that the quality of the NAEN for $\boldsymbol{\beta}$ and σ^2 relies on the choice of tuning parameters $\lambda_{n,1}$, $\lambda_{n,2}$ and weight vector $\boldsymbol{w}$. The orders of $\lambda_{n,1}$ and $\lambda_{n,2}$, in theory, are discussed in Section 3. The weight vector $\boldsymbol{w}$ is generated by the following two-step procedure:

- Step 1. Obtain a consistent estimator $\widetilde{\boldsymbol{\beta}}$ as the initial estimator for $\boldsymbol{\beta}$.
- Step 2. Set $\boldsymbol{w}$ with $w_j = p'_{\lambda_{n,1}}(|\widetilde{\beta}_j|)$, $j = 1, \cdots, p$, where $p_{\lambda_{n,1}}$ is a folded-concave penalty function (such as smoothly clipped absolute deviation (SCAD) in [9, 7] or minimax concave penalty (MCP) in [18]).

The Lasso estimator $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ in [15, 5, 19, 3] can be taken as the initial estimator for $\boldsymbol{\beta}$. Theoretical properties in next section show the effectiveness of this two-step procedure.

3. **Theoretical properties.** This section establishes the relationship between the naive adaptive elastic-net and the NAEN, and discusses the asymptotic properties of the NAEN for $\boldsymbol{\beta}$ and σ^2.

Recall that the naive adaptive elastic-net in [22] is defined as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_n(\boldsymbol{\beta}) := \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda_{n,1} \|\boldsymbol{w} \circ \boldsymbol{\beta}\|_1 + 2\lambda_{n,2} \|\boldsymbol{\beta}\|_2^2. \tag{4}$$

The next result establishes the relationship between the naive adaptive elastic net and the NAEN.

Proposition 3.1. *Let $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ be the NAEN estimator for $\boldsymbol{\beta}$ and σ^2 defined in (3), where $(\widehat{\boldsymbol{\theta}}, \widehat{\phi})$ is a solution of (2). Then,*

(i) $\widehat{\boldsymbol{\beta}}$ is the solution of problem (4);

(ii) $\widehat{\sigma}^2$ is the optimal value of problem (4).

Furthermore, we have $\widehat{\sigma}^2 = n^{-1}(\|\boldsymbol{y}\|_2^2 - \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2)$.

Proof. (i) Since $(\widehat{\boldsymbol{\theta}}, \widehat{\phi})$ is a solution of (2), $\widehat{\boldsymbol{\theta}}$ is the solution of the optimization

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} L(\boldsymbol{\theta}, \widehat{\phi}) + \lambda_{n,1} \|\boldsymbol{w} \circ \boldsymbol{\theta}\|_1 + \frac{\lambda_{n,2} \|\boldsymbol{\theta}\|_2^2}{\widehat{\phi}}.$$

By the first-order optimality of above optimization, we have

$$-\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{X}\frac{\widehat{\boldsymbol{\theta}}}{\widehat{\phi}}+n\lambda_{n,1}\boldsymbol{w}\circ\widehat{\boldsymbol{g}}+2n\lambda_{n,2}\frac{\widehat{\boldsymbol{\theta}}}{\widehat{\phi}}=0,$$

where $\widehat{\boldsymbol{g}}\in\partial(\|\widehat{\boldsymbol{\theta}}\|_1)$. Since $\text{sign}(\widehat{\boldsymbol{\theta}})=\text{sign}(\widehat{\boldsymbol{\beta}})$, we have $\partial(\|\widehat{\boldsymbol{\theta}}\|_1)=\partial(\|\widehat{\boldsymbol{\beta}}\|_1)$. It follows that

$$-\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+n\lambda_{n,1}\boldsymbol{w}\circ\widehat{\boldsymbol{g}}+2n\lambda_{n,2}\widehat{\boldsymbol{\beta}}=0,$$

which implies that $\widehat{\boldsymbol{\beta}}$ is the solution of problem (4). Here, we use the fact that the objective function of the problem (4) is strong convex.

(ii) Since $(\widehat{\boldsymbol{\theta}},\widehat{\phi})$ is a solution of (2), by the first-order optimality of problem (2), we have

$$\begin{aligned}-\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{X}\frac{\widehat{\boldsymbol{\theta}}}{\widehat{\phi}}+n\lambda_{n,1}\boldsymbol{w}\circ\widehat{\boldsymbol{g}}+2n\lambda_{n,2}\frac{\widehat{\boldsymbol{\theta}}}{\widehat{\phi}}&=0,\\ -\frac{1}{\widehat{\phi}}+\frac{1}{n}\|\boldsymbol{y}\|_2^2-\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2}{n\widehat{\phi}^2}-\frac{2\lambda_{n,2}\|\widehat{\boldsymbol{\theta}}\|_2^2}{\widehat{\phi}^2}&=0,\end{aligned}$$

where $\widehat{\boldsymbol{g}}\in\partial(\|\widehat{\boldsymbol{\theta}}\|_1)$. Therefore we have

$$-\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}+n\lambda_{n,1}\boldsymbol{w}\circ\widehat{\boldsymbol{g}}+2n\lambda_{n,2}\widehat{\boldsymbol{\beta}}=0, \tag{5}$$

$$-\frac{1}{\widehat{\phi}}+\frac{1}{n}\|\boldsymbol{y}\|_2^2-\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2^2}{n\widehat{\phi}^2}-\frac{2\lambda_{n,2}\|\widehat{\boldsymbol{\theta}}\|_2^2}{\widehat{\phi}^2}=0. \tag{6}$$

Since $\partial(\|\widehat{\boldsymbol{\theta}}\|_1)=\partial(\|\widehat{\boldsymbol{\beta}}\|_1)$, we have $\widehat{\boldsymbol{g}}\in\partial(\|\widehat{\boldsymbol{\beta}}\|_1)$. Thus,

$$\widehat{\boldsymbol{\beta}}^T(\boldsymbol{w}\circ\widehat{\boldsymbol{g}})=\sum_{i=1}^p w_i\widehat{\beta}_i\widehat{g}_i=\sum_{i=1}^p|w_i\widehat{\beta}_i|=\|\boldsymbol{w}\circ\widehat{\boldsymbol{\beta}}\|_1. \tag{7}$$

Combining (5)-(7), we have

$$0=-\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}+\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2+n\lambda_{n,1}\|\boldsymbol{w}\circ\widehat{\boldsymbol{\beta}}\|_1+2n\lambda_{n,2}\|\widehat{\boldsymbol{\beta}}\|_2^2, \tag{8}$$

$$\widehat{\sigma}^2=\frac{1}{n}\left(\|\boldsymbol{y}\|_2^2-\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2\right)-2\lambda_{n,2}\|\widehat{\boldsymbol{\beta}}\|_2^2. \tag{9}$$

Furthermore, by (8), it holds that

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2&=\|\boldsymbol{y}\|_2^2-\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2+2\left(\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2-\boldsymbol{y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\\&=\|\boldsymbol{y}\|_2^2-\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2-2n\lambda_{n,1}\|\boldsymbol{w}\circ\widehat{\boldsymbol{\beta}}\|_1-4n\lambda_{n,2}\|\widehat{\boldsymbol{\beta}}\|_2^2.\end{aligned}$$

Then, by (9),

$$\begin{aligned}\widehat{\sigma}^2&=\frac{1}{n}\left(\|\boldsymbol{y}\|_2^2-\|\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2\right)-2\lambda_{n,2}\|\widehat{\boldsymbol{\beta}}\|_2^2\\&=\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2+2\lambda_{n,1}\|\boldsymbol{w}\circ\widehat{\boldsymbol{\beta}}\|_1+2\lambda_{n,2}\|\widehat{\boldsymbol{\beta}}\|_2^2,\end{aligned}$$

which proves that $\widehat{\sigma}^2$ is the optimal value of problem (4). □

Proposition 3.1 is similar to Proposition 1 in [17] and Proposition 1 in [16]. Essentially, the NAEN for σ^2 is the optimal value of the optimization of the naive adaptive elastic-net. This fact not only makes NAEN easy to implement, but also establishes the following conclusions, which is a deterministic result and does not

rely on statistical assumptions for $\boldsymbol{X}$ and ε and is the core of the asymptotic properties of the NAEN.

Proposition 3.2. *Assume that* $\lambda_{n,1} \geq \|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X} - 2\lambda_{n,2}\boldsymbol{\beta}^*\|_\infty$. *Then,*

$$\left|\widehat{\sigma}^2 - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right| \leq 2\max\big\{\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + \lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2, \\ \big|\lambda_{n,1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 - \lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2\big|\big\}.$$

Proof. By Proposition 3.1, we have

$$\begin{aligned}\widehat{\sigma}^2 &\leq \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + 2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 \\ &= \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 + 2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2.\end{aligned} \tag{10}$$

On the other hand, since $\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{\beta}\|_2^2$ in (4) is convex, we have

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + 2\lambda_{n,1}\|\boldsymbol{w}\circ\widehat{\boldsymbol{\beta}}\|_1 + 2\lambda_{n,2}\|\widehat{\boldsymbol{\beta}}\|_2^2 \\ &\geq \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 \\ &\quad + \left[\frac{2}{n}\boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{\beta}^* - \boldsymbol{y}) + 4\lambda_{n,2}\boldsymbol{\beta}^*\right]^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\ &= \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 - \left[\frac{2}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X} - 4\lambda_{n,2}\boldsymbol{\beta}^*\right]^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\ &\geq \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 - \left\|\frac{2}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X} - 4\lambda_{n,2}\boldsymbol{\beta}^*\right\|_\infty\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \\ &\geq \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 - 2\lambda_{n,1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1.\end{aligned} \tag{11}$$

Combining inequalities (10) with (11), we obtain

$$\left|\widehat{\sigma}^2 - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right| \leq \max\big\{2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2, \\ |2\lambda_{n,1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 - 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2|\big\},$$

which completes the proof. □

Unlike the analysis of Lemma 1 in [17], Proposition 3.2 has nothing to do with duality. Since $\widehat{\sigma}^2$ is the optimal value of problem (4), the objective value of problem (4) at $\boldsymbol{\beta}^*$ provides an upper bound for $\widehat{\sigma}^2$. The convexity of the objective function in problem (4) provides a lower bound for $\widehat{\sigma}^2$.

We now discuss the asymptotic properties of the NAEN for $\boldsymbol{\beta}$ and σ^2 based on Propositions 3.1 and 3.2. We first give some regularity assumptions as follows.

Assumption 3.3. *With probability tending to one, the initial estimator satisfies* $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C_1\sqrt{s(\log p)/n}$, *where* $C_1 > 0$ *is constant.*

Assumption 3.4. $p'_{\lambda_{n,1}}(t)$ *is non-increasing in* $t \in (0,\infty)$ *and is Lipschitz with constant* C_2, *that is,*

$$|p'_{\lambda_{n,1}}(|t_1|) - p'_{\lambda_{n,1}}(|t_2|)| \leq C_2|t_1 - t_2|$$

for any $t_1, t_2 \in \mathbb{R}$. *Moreover,* $p'_{\lambda_{n,1}}(C_1\sqrt{s\log p/n}) > (1/2)p'_{\lambda_{n,1}}(0+)$ *for sufficiently large* n, *where* C_1 *is defined in Assumption 3.3.*

Assumption 3.5. *There exist positive constants* $2\lambda_{n,2} < c_{\min} < c_{\max} < \infty$ *such that*

$$c_{\min} \leq \lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0}\right) \leq \lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0}\right) \leq c_{\max}.$$

Assumption 3.6. *The following inequalities hold:*

$$\left\|\frac{1}{n}\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}_{\mathcal{A}_0}\right\|_{2,\infty} < \frac{\lambda_{n,1}}{4C_3\|\boldsymbol{w}_{\mathcal{A}_0^c}^{-1}\|_\infty a_n},$$

where $\|\boldsymbol{B}\|_{2,\infty} = \max_{\|\boldsymbol{v}\|_2\leq 1}\|\boldsymbol{B}\boldsymbol{v}\|_\infty$, $\boldsymbol{w}_{\mathcal{A}_0^c}^{-1} = (w_{s+1}^{-1},\ldots,w_p^{-1})^T$, C_3 *and* a_n *is defined in Theorem 3.7.*

Theorem 3.7. *Suppose Assumptions 3.3-3.6 hold,* $\min_{i\in\mathcal{A}_0^c} w_i^* > C_4^{-1}$, $\lambda_{n,1} \geq 4C_4\sigma^*\sqrt{2M\log p/n}$, *where* $M > 1$ *and* $C_4 > 0$ *are constants. Then, with probability tending to one, the minimizer* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}^T, \widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}^T)$ *of problem (4) satisfies* $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c} = \boldsymbol{0}$ *and* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq C_3 a_n$, *where* $a_n = \{\sigma^*\sqrt{2Ms\log p/n} + \lambda_{n,1}(C_1C_2\sqrt{s(\log p)/n} + \|\boldsymbol{w}_{\mathcal{A}_0}^*\|_2) + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2\}$, $C_3 > 0$ *is a sufficiently large constant,* C_1, C_2 *are defined in regularity assumptions.*

Proof. Since problem (4) is a convex optimization with a strong convex objective function, the minimizer of problem (4) is unique and we only need to show that there exists a $\widehat{\boldsymbol{\beta}}$ satisfying

$$\boldsymbol{X}_{\mathcal{A}_0}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - n\lambda_{n,1}\boldsymbol{w}_{\mathcal{A}_0} \circ \widehat{\boldsymbol{g}}_{\mathcal{A}_0} - 2n\lambda_{n,2}\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0} = \boldsymbol{0}, \tag{12}$$

$$\|\boldsymbol{X}_{\mathcal{A}_0^c}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - 2n\lambda_{n,2}\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}\|_\infty \leq n\lambda_{n,1}\boldsymbol{w}_{\mathcal{A}_0^c}, \tag{13}$$

where $\widehat{\boldsymbol{g}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$.

Since $\|\boldsymbol{X}_j\|_2^2 = n$ and $\boldsymbol{\varepsilon} \sim N(0,(\sigma^*)^2)$, it follows from Corollary 4.3 in [10] that for any $L > 0$,

$$\Pr\left\{\frac{\|\boldsymbol{X}^T\boldsymbol{\varepsilon}\|_\infty}{n\sigma^*} > \sqrt{\frac{2\log p + 2L}{n}}\right\} \leq e^{-L}.$$

Take $L = (M-1)\log p$, where $M > 1$ is constant. Then we have

$$\Pr\left\{\frac{\|\boldsymbol{X}^T\boldsymbol{\varepsilon}\|_\infty}{n\sigma^*} > \sqrt{\frac{2M\log p}{n}}\right\} \leq e^{-(M-1)\log p}. \tag{14}$$

Now we show that the minimizer $\widehat{\boldsymbol{\beta}}$ of problem (4) satisfies conditions (12)-(13).

We first consider the minimizer of problem (4) in the subspace $\{\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}_0}^T, \boldsymbol{\beta}_{\mathcal{A}_0^c}^T)^T : \boldsymbol{\beta}_{\mathcal{A}_0^c} = \boldsymbol{0}\}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}_0}^T, \boldsymbol{0}^T)^T$, where $\boldsymbol{\beta}_{\mathcal{A}_0} = \boldsymbol{\beta}_{\mathcal{A}_0}^* + a_n\boldsymbol{v}_{\mathcal{A}_0} \in \mathbb{R}^s$ with

$$a_n = \sigma^*\sqrt{\frac{2Ms\log p}{n}} + \lambda_{n,1}\left(C_1C_2\sqrt{\frac{s\log p}{n}} + \|\boldsymbol{w}_{\mathcal{A}_0}^*\|_2\right) + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2,$$

$\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 = C_3$ and $C_3 > 0$ is some large enough constant. Note that

$$Q_n(\boldsymbol{\beta}_{\mathcal{A}_0}^* + a_n\boldsymbol{v}_{\mathcal{A}_0}, \boldsymbol{0}) - Q_n(\boldsymbol{\beta}_{\mathcal{A}_0}^*, \boldsymbol{0}) = I_1(\boldsymbol{v}_{\mathcal{A}_0}) + I_2(\boldsymbol{v}_{\mathcal{A}_0}) + I_3(\boldsymbol{v}_{\mathcal{A}_0}), \tag{15}$$

where $I_1(\boldsymbol{v}_{\mathcal{A}_0}) = \frac{1}{n}\|\boldsymbol{X}(\boldsymbol{\beta}^* + a_n\boldsymbol{v}) - \boldsymbol{y}\|_2^2 - \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{y}\|_2^2$, $I_2(\boldsymbol{v}_{\mathcal{A}_0}) = 2\lambda_{n,1}\|\boldsymbol{w}_{\mathcal{A}_0} \circ (\boldsymbol{\beta}_{\mathcal{A}_0}^* + a_n\boldsymbol{v}_{\mathcal{A}_0})\|_1 - 2\lambda_{n,1}\|\boldsymbol{w}_{\mathcal{A}_0} \circ \boldsymbol{\beta}_{\mathcal{A}_0}^*\|_1$, $I_3(\boldsymbol{v}_{\mathcal{A}_0}) = 2\lambda_{n,2}\|\boldsymbol{\beta}_{\mathcal{A}_0}^* + a_n\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2\lambda_{n,2}\|\boldsymbol{\beta}_{\mathcal{A}_0}^*\|_2^2$.

For $I_1(\boldsymbol{v}_{\mathcal{A}_0})$, by (14), we have

$$\begin{aligned} I_1(\boldsymbol{v}_{\mathcal{A}_0}) &= \frac{1}{n}\|\boldsymbol{X}(\boldsymbol{\beta}^* + a_n\boldsymbol{v}) - \boldsymbol{y}\|_2^2 - \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{y}\|_2^2 \\ &= \frac{1}{n}a_n^2\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} + \frac{2}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}a_n\boldsymbol{v} \\ &= \frac{1}{n}a_n^2\boldsymbol{v}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0}\boldsymbol{v}_{\mathcal{A}_0} + \frac{2}{n}a_n\boldsymbol{\varepsilon}^T\boldsymbol{X}_{\mathcal{A}_0}\boldsymbol{v}_{\mathcal{A}_0} \\ &\geq c_{\min}a_n^2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2a_n\left\|\frac{\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{\varepsilon}}{n}\right\|_2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 \\ &\geq c_{\min}a_n^2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2a_n\sigma^*\sqrt{\frac{2Ms\log p}{n}}\|\boldsymbol{v}_{\mathcal{A}_0}\|_2, \end{aligned} \tag{16}$$

where the last inequality holds due to $\|\cdot\|_2 \leq \sqrt{s}\|\cdot\|_\infty$. For $I_2(\boldsymbol{v}_{\mathcal{A}_0})$, we have

$$\begin{aligned} |I_2(\boldsymbol{v}_{\mathcal{A}_0})| &= |2\lambda_{n,1}\|\boldsymbol{w}_{\mathcal{A}_0} \circ (\boldsymbol{\beta}^*_{\mathcal{A}_0} + a_n\boldsymbol{v}_{\mathcal{A}_0})\|_1 - 2\lambda_{n,1}\|\boldsymbol{w}_{\mathcal{A}_0} \circ \boldsymbol{\beta}^*_{\mathcal{A}_0}\|_1| \\ &\leq 2\lambda_{n,1}\|\boldsymbol{w}_{\mathcal{A}_0} \circ a_n\boldsymbol{v}_{\mathcal{A}_0}\|_1 \leq 2a_n\lambda_{n,1}\|\boldsymbol{w}_{\mathcal{A}_0}\|_2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2. \end{aligned} \tag{17}$$

From the two-step procedure and Assumptions 3.3, 3.4, it holds that

$$\begin{aligned} \|\boldsymbol{w}_{\mathcal{A}_0}\|_2 &\leq \|\boldsymbol{w}_{\mathcal{A}_0} - \boldsymbol{w}^*_{\mathcal{A}_0}\|_2 + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2 \leq C_2\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_0} - \boldsymbol{\beta}^*_{\mathcal{A}_0}\|_2 + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2 \\ &\leq C_1C_2\sqrt{\frac{s\log p}{n}} + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2. \end{aligned}$$

Combining the above inequality with (17), we have

$$|I_2(\boldsymbol{v}_{\mathcal{A}_0})| \leq 2a_n\lambda_{n,1}\left(C_1C_2\sqrt{\frac{s\log p}{n}} + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2\right)\|\boldsymbol{v}_{\mathcal{A}_0}\|_2. \tag{18}$$

For $I_3(\boldsymbol{v}_{\mathcal{A}_0})$,

$$\begin{aligned} |I_3(\boldsymbol{v}_{\mathcal{A}_0})| &= |2\lambda_{n,2}\|\boldsymbol{\beta}^*_{\mathcal{A}_0} + a_n\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2\lambda_{n,2}\|\boldsymbol{\beta}^*_{\mathcal{A}_0}\|_2^2| \\ &= \left|2\lambda_{n,2}\sum_{i\in\mathcal{A}_0}(2\beta_i^* + a_n\boldsymbol{v}_i)a_n\boldsymbol{v}_i\right| \\ &\leq 4a_n\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 + 2a_n^2\lambda_{n,2}\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2. \end{aligned} \tag{19}$$

Thus, by (15)-(19), it holds that

$$\begin{aligned} &Q_n(\boldsymbol{\beta}^*_{\mathcal{A}_0} + a_n\boldsymbol{v}_{\mathcal{A}_0}, \boldsymbol{0}) - Q_n(\boldsymbol{\beta}^*_{\mathcal{A}_0}, \boldsymbol{0}) \\ &\geq c_{\min}a_n^2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2a_n\|\boldsymbol{v}_{\mathcal{A}_0}\|_2\sigma^*\sqrt{\frac{2Ms\log p}{n}} \\ &\quad - 2a_n\lambda_{n,1}\left(C_1C_2\sqrt{\frac{s\log p}{n}} + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2\right)\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 \\ &\quad - 4a_n\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 - 2a_n^2\lambda_{n,2}\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 \\ &= (c_{\min} - 2\lambda_{n,2})a_n^2C_3^2 - 2C_3a_n\sigma^*\sqrt{\frac{2Ms\log p}{n}} \\ &\quad - C_3a_n\left[2\lambda_{n,1}\left(C_1C_2\sqrt{\frac{s\log p}{n}} + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2\right) + 4\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2\right]. \end{aligned}$$

Making C_3 large enough, we obtain that with probability tending to one,

$$Q_n(\boldsymbol{\beta}^*_{\mathcal{A}_0} + a_n\boldsymbol{v}_{\mathcal{A}_0}, \boldsymbol{0}) - Q_n(\boldsymbol{\beta}^*_{\mathcal{A}_0}, \boldsymbol{0}) > 0. \tag{20}$$

Thus, with probability tending to one, there exists a solution $\widehat{\boldsymbol{\beta}} := (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}^T, \mathbf{0}^T)^T$ to the problem (4) subject to subspace $\{\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}_0}^T, \boldsymbol{\beta}_{\mathcal{A}_0^c}^T)^T : \boldsymbol{\beta}_{\mathcal{A}_0^c} = \mathbf{0}\}$ and satisfies $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0} - \boldsymbol{\beta}_{\mathcal{A}_0}^*\|_2 \leq C_4 a_n$ with some constant $C_4 > 0$. Therefore, it follows that $\widehat{\boldsymbol{\beta}}$ satisfies equality (12) with probability tending to one.

It remains to prove that (13) holds for $\widehat{\boldsymbol{\beta}}$ with probability tending to one. By triangle inequality, we have

$$\|\boldsymbol{X}_{\mathcal{A}_0^c}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|_\infty \leq \|\boldsymbol{X}_{\mathcal{A}_0^c}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty + \|\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_\infty. \tag{21}$$

By Assumption 3.3, $|\widetilde{\boldsymbol{\beta}}_i| \leq C_1\sqrt{s(\log p)/n}$ with probability approaching one, where $i \in \mathcal{A}_0^c$. Then, by the features of fold-concave penalty function,

$$p'_{\lambda_{n,1}}(|\widetilde{\boldsymbol{\beta}}_i|) \geq p'_{\lambda_{n,1}}\left(C_1\sqrt{\frac{s\log p}{n}}\right). \tag{22}$$

Therefore, it follows from Assumption 3.4 and inequality (22) that

$$\begin{aligned}\|\boldsymbol{w}_{\mathcal{A}_0^c}^{-1}\|_\infty = [\min_{i\in\mathcal{A}_0^c} p'_{\lambda_{n,1}}(|\widetilde{\boldsymbol{\beta}}_i|)]^{-1} &\leq \left[p'_{\lambda_{n,1}}\left(C_1\sqrt{\frac{s\log p}{n}}\right)\right]^{-1} \\ &< \frac{2}{p'_{\lambda_{n,1}}(0+)} = 2\|(\boldsymbol{w}_{\mathcal{A}_0^c}^*)^{-1}\|_\infty.\end{aligned} \tag{23}$$

Thus, for the first term of the right-hand side of inequality (21), by (14), (23) and the condition that $\min_{i\in\mathcal{A}_0^c}\{w_i^*\} > C_4^{-1}$, it holds with probability tending to one that

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{X}_{\mathcal{A}_0^c}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty = \frac{1}{n}\|\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{\varepsilon}\|_\infty &< \sigma^*\sqrt{\frac{2M\log p}{n}} \\ \leq \frac{\lambda_{n,1}}{4C_4} < \frac{\lambda_{n,1}}{4\|(\boldsymbol{w}_{\mathcal{A}_0^c}^*)^{-1}\|_\infty} &< \frac{\lambda_{n,1}}{2\|\boldsymbol{w}_{\mathcal{A}_0^c}^{-1}\|_\infty}.\end{aligned} \tag{24}$$

For the second term of right hand of inequality (21), by Assumption 3.5, (14) and (22), it holds with probability tending to one that

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_\infty &\leq \frac{1}{n}\|\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}_{\mathcal{A}_0}\|_{2,\infty}\|\boldsymbol{\beta}_{\mathcal{A}_0}^* - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}\|_2 \\ &\leq \frac{\lambda_{n,1}}{4\|(\boldsymbol{w}_{\mathcal{A}_0^c}^*)^{-1}\|_\infty} < \frac{\lambda_{n,1}}{2\|\boldsymbol{w}_{\mathcal{A}_0^c}^{-1}\|_\infty}.\end{aligned}$$

Combining $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}^T = \mathbf{0}$ with (19), (23) and (24), we obtain inequality (13). Thus, $\widehat{\boldsymbol{\beta}}$ is the minimizer of problem (4). This completes the proof of Theorem 3.7. □

Based on Proposition 3.2 and Theorem 3.7, the mean squared error bound of the NAEN for σ^2 can be established.

Theorem 3.8. *Suppose conditions in Theorem 3.7 hold and* $\lambda_{n,1} \geq 4C_4\sigma^*\sqrt{2M\log p/n} + 2\|\lambda_{n,2}\boldsymbol{\beta}^*\|_\infty$, *where* $M > 1$ *is a constant. Then, with probability tending to one, it holds*

(i) $|\widehat{\sigma}^2 - \|\boldsymbol{\varepsilon}\|_2^2/n| \leq b_n$;

(ii) $\mathrm{E}\{(\widehat{\sigma}^2 - \|\boldsymbol{\varepsilon}\|_2^2/n|)^2\} \leq (M + p^{1-M}/\log p)b_n^2$;

(iii) $\mathrm{E}\{(\widehat{\sigma}^2/(\sigma^*)^2 - 1)^2\} \leq [b_n\sqrt{M + p^{1-M}/\log p}/(\sigma^*)^2 + \sqrt{2/n}]^2$,

where $b_n = \max\{2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2, |2C_3\sqrt{s}\lambda_{n,1}a_n - 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2|\}$.

Proof. (i) By (14), we obtain that $\|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\|_\infty \le 4C_4\sigma^*\sqrt{2M\log p/n}$ with probability tending to one. Therefore,

$$\begin{aligned}\left\|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X} - 2\lambda_{n,2}\boldsymbol{\beta}^*\right\|_\infty &\le \left\|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\right\|_\infty + 2\|\lambda_{n,2}\boldsymbol{\beta}^*\|_\infty \\ &\le 4C_4\sigma^*\sqrt{\frac{2M\log p}{n}} + 2\|\lambda_{n,2}\boldsymbol{\beta}^*\|_\infty \\ &\le \lambda_{n,1}.\end{aligned}$$

By Proposition 3.2 and the fact that $\|\cdot\|_1 \le \sqrt{s}\|\cdot\|_2$ for any vector with sparsity s, we have

$$\begin{aligned}&\left|\widehat{\sigma}^2 - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right| \\ &\le \max\{2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2, |2\lambda_{n,1}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 - 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2|\} \\ &\le \max\{2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2, |2C_3\sqrt{s}\lambda_{n,1}a_n - 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2|\}.\end{aligned}$$

Thus, (i) holds with probability tending to one.

(ii) It follows from (14) that for any constant $M > 1$,

$$\Pr\left(\left(\widehat{\sigma}^2 - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right)^2 > Mb_n^2\right) \le e^{-(M-1)\log p}.$$

Denote $Z_n = (\widehat{\sigma}^2 - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2)^2$. Then,

$$\begin{aligned}\mathrm{E}\left(\frac{Z_n}{b_n^2}\right) &= \int_0^\infty \Pr\left(\frac{Z_n}{b_n^2} > t\right)dt = \int_0^M \Pr\left(\frac{Z_n}{b_n^2} > t\right)dt + \int_M^\infty \Pr\left(\frac{Z_n}{b_n^2} > t\right)dt \\ &\le M + \int_M^\infty e^{-(t-1)\log p}\,dt = M + \frac{p^{1-M}}{\log p},\end{aligned} \tag{25}$$

which completes the proof of (ii).

(iii) Since $(\sigma^*)^{-2}\|\boldsymbol{\varepsilon}\|_2^2 \sim \chi^2(n)$, we have

$$\mathrm{E}\left(\frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right) = (\sigma^*)^2,\ \mathrm{Var}\left(\frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right) = \frac{2(\sigma^*)^4}{n}.$$

By the proof of Theorem 12 in [17], we have

$$\mathrm{E}\left\{(\widehat{\sigma}^2 - (\sigma^*)^2)^2\right\} \le \left\{\left[\mathrm{E}\left\{\left(\widehat{\sigma}^2 - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right)^2\right\}\right]^{\frac{1}{2}} + \left\{\mathrm{Var}(\frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2)\right\}^{\frac{1}{2}}\right\}^2. \tag{26}$$

By (25) and (26), it holds that

$$\mathrm{E}\left\{(\widehat{\sigma}^2 - (\sigma^*)^2)^2\right\} \le \left[\left(M + \frac{p^{1-M}}{\log p}\right)^{\frac{1}{2}} b_n + (\sigma^*)^2\left(\frac{2}{n}\right)^{\frac{1}{2}}\right]^2.$$

Thus, (iii) holds. □

When $\lambda_{n,2} = 0$, Theorems 3.7 and 3.8 coincide with Theorems 1 and 4 in [16] respectively. Theorems 3.7 and 3.8 are general results for the NAEN. Next, we give asymptotic properties of the NAEN with SCAD penalty for $\boldsymbol{\beta}$ and σ^2. The SCAD used in two-step procedure is defined as follows:

$$p'_{\lambda_{n,1}}(|t|) = \mathbf{1}\{|t| \le \lambda_{n,1}\} + \frac{(a\lambda_{n,1} - |t|)_+}{(a-1)\lambda_{n,1}}\mathbf{1}\{|t| > \lambda_{n,1}\},\ (a \ge 2),$$

where often $a = 3.7$ is used. It can be easily verified that Assumption 3.4 holds if $\lambda_{n,1} > 2(a+1)^{-1}C_1\sqrt{s\log p/n}$.

Corollary 3.9. *Assume* $\lambda_{n,1} = O(\sqrt{s\log p(\log\log n)/n})$, $s\log p = O(n^\gamma)$, $\gamma \in (0,1)$, $\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2 = O(\sqrt{s\log p/n})$, $\min_{i\in\mathcal{A}_0}\beta_i^* \geq 2a\lambda_{n,1}$. *Further assume that* $\|n^{-1}\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}_{\mathcal{A}_0}\|_{2,\infty} < C_5\sqrt{\log\log n}$ *with some positive constant* C_5. *Then, under Assumptions 3.3, 3.5, for sufficiently large* n, *with asymptotic probability one, the minimizer* $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}^T, \widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}^T)$ *of problem (4) satisfies* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq O(\sqrt{s\log p/n})$, $\mathrm{sign}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}) = \mathrm{sign}(\boldsymbol{\beta}_{\mathcal{A}_0}^*)$ *and* $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c} = \mathbf{0}$.

Proof. We need to verify that the conditions in Theorem 3.7 hold. Since the order of $\lambda_{n,1}$ is $\sqrt{s\log p(\log\log n)/n}$, Assumption 3.4 and the conditions related to $\lambda_{n,1}$ hold when n is sufficiently large. Conditions $\min_{i\in\mathcal{A}_0}\beta_i^* \geq 2a\lambda_{n,1}$ and $s\log p = O(n^\gamma)$, $\gamma \in (0,1)$ imply $\boldsymbol{w}_{\mathcal{A}_0}^* = \mathbf{0}$ and $\lambda_{n,1} = o(1)$ respectively. Combining these results with $\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 = O(\sqrt{s\log p/n})$, we have $a_n = O(\sqrt{s\log p/n})$. Moreover, by Assumption 3.3, we know that $|\widetilde{\beta}_i| < C_1\sqrt{s\log p/n}$ for any $i \in \mathcal{A}_0^c$. Then, $w_i = 1$ for any $i \in \mathcal{A}_0^c$ when n is sufficiently large. Thus, Assumption 3.6 follows from the conditions $\lambda_{n,1} = O(\sqrt{s\log p(\log\log n)/n})$ and $\|n^{-1}\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}_{\mathcal{A}_0}\|_{2,\infty} < C_5\sqrt{\log\log n}$. Hence, it following from Theorem 3.7 that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq O(\sqrt{s\log p/n})$, $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c} = \mathbf{0}$. The result $\mathrm{sign}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}) = \mathrm{sign}(\boldsymbol{\beta}_{\mathcal{A}_0}^*)$ holds due to condition $\min_{i\in\mathcal{A}_0}\beta_i^* \geq 2a\lambda_{n,1}$. □

Finally, the convergence rate of the mean squared error bound of NAEN with SCAD for σ^2 can be established as follows.

Corollary 3.10. *Assume that the conditions in Corollary 3.9 hold. Then, with probability tending to one,*

(i) $|\widehat{\sigma}^2 - \|\boldsymbol{\varepsilon}\|_2^2/n| \leq O(s\sqrt{\log p/n})$;
(ii) $\mathrm{E}\{(\widehat{\sigma}^2 - \|\boldsymbol{\varepsilon}\|_2^2/n|)^2\} \leq O(s^2\log p/n)$;
(iii) $\mathrm{E}\{(\widehat{\sigma}^2/(\sigma^*)^2 - 1)^2\} \leq O(s^2\log p/n)$.

Proof. Under the conditions in Corollary 3.9 and Assumption 3.3, with probability tending to one, $|\widetilde{\beta}_i| < \lambda_{n,1}$ for $i \in \mathcal{A}_0^c$ and $|\widetilde{\beta}_i| > a\lambda_{n,1}$ for $i \in \mathcal{A}_0$. Then, $\boldsymbol{w} = \boldsymbol{w}^*$ and $\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 = 0$ with probability tending to one. Thus,

$$2\lambda_{n,1}\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1 + 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 = 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2 = O\left(\sqrt{\frac{s\log p}{n}}\right). \tag{27}$$

Moreover, condition $s\log p = O(n^\gamma)$, $\gamma \in (0,1)$ implies that $\lambda_{n,1} = o(1)$. Thus,

$$|2C_3\sqrt{s}\lambda_{n,1}a_n - 2\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2^2| \leq O\left(s\sqrt{\frac{\log p}{n}}\right). \tag{28}$$

It follows from (27) and (28) that with probability tending to one, $b_n \leq O(s\sqrt{\log p/n})$. Thus,the results follows from Thereom 3.8. □

If $s^2\log p = o(n^\gamma)$, $\gamma \in (0,1)$, then error bounds in Corollary 3.10 converge to 0 with probability tending to one. Let λ_n denote the tuning parameter in SCAD penalty. We now discuss the convergence rate of the mean squared error bound of NAEN with SCAD for σ^2 in the case that λ_n and $\lambda_{n,1}$ are different. If $\lambda_n = O(\sqrt{s\log p(\log\log n)/n})$ and $\|n^{-1}\boldsymbol{X}_{\mathcal{A}_0^c}^T\boldsymbol{X}_{\mathcal{A}_0}\|_{2,\infty} < C_6\sqrt{1/s}$ with some positive constant C_6, then the condition on $\lambda_{n,1}$ can be weakened to $\lambda_{n,1} = O(\sqrt{\log p/n})$. Thus, $b_n \leq O(\max\{s\log p/n, \sqrt{s\log p/n}\})$ and the order of the

bounds of $\mathrm{E}\{(\widehat{\sigma}^2 - \|\boldsymbol{\varepsilon}\|_2^2/n|)^2\}$ and $\mathrm{E}\{(\widehat{\sigma}^2/(\sigma^*)^2 - 1)^2\}$ in Corollary 3.10 are all $b_n \leq O(\max\{s^2(\log p)^2/n^2, s\log p/n\})$. As long as $s\log p = o(n)$, these bounds converge to 0 with probability tending to one and the order of these bounds in Corollary 3.10 are $\sqrt{s\log p/n}$, $s\log p/n$, $s\log p/n$. Moreover, if $\lambda_{n,1} = O(1/\sqrt{s}) > O(\sqrt{\log p/n})$, then the order of these bounds in Corollary 3.10 are $\sqrt{s\log p/n}$, $s\log p/n$, $s\log p/n$ without the condition $s\log p = o(n)$. In fact, under different assumptions about the order of some parameters, the convergence rates in Corollary 3.10 are different.

All theoretical properties in this section can be generalized to the case where the model error ε_i is sub-Gaussian or sub-exponential.

4. Simulation study.

In this section, we study the finite-sample performance of the NAEN. The SCAD in [7] is applied to calculate the weight vector $\boldsymbol{w}$ in (2). The five methods mentioned in Section 1 are used for comparison, which are SL, NL, OL, NAL and MUE based on adaptive elastic-net. In addition, the oracle estimator (OE) $(1/n)\|\boldsymbol{\varepsilon}\|_2^2$ is included as the benchmark.

4.1. Numerical optimization of NAEN.

Proposition 3.1 establishes the relationship between the NAEN and the naive elastic-net, and it shows that the NAEN estimator for $\boldsymbol{\beta}$ is the solution of problem (4) and the NAEN estimator for σ^2 is the optimal value of problem (4). Thus, we can obtain the NAEN estimators for $\boldsymbol{\beta}$ and σ^2 by solving problem (4).

Clearly, the gradient of $\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{\beta}\|_2^2$ is globally Lipschitz continuous with modulus $\|(2/n)\boldsymbol{X}^T\boldsymbol{X} + 2\lambda_{n,2}\boldsymbol{I}_p\|_2$ and problem (4) is a convex optimization. We apply FISTA in [2] to solve problem (4). Theorem 4.4 in [2] shows that the complexity of FISTA is $O(1/k^2)$.

Let $\boldsymbol{G}_k := \nabla((1/n)\|\boldsymbol{y} - \boldsymbol{z}^k\|_2^2 + 2\lambda_{n,2}\|\boldsymbol{z}^k\|_2^2) = (2/n)\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{z}^k - \boldsymbol{y}) + 4\lambda_{n,2}\boldsymbol{z}^k$, $L = \|(2/n)\boldsymbol{X}^T\boldsymbol{X} + 4\lambda_{n,2}\boldsymbol{I}_p\|_2$. Then, the framework of FISTA for problem (4) is as follows.

Algorithm 1 Framework of FISTA

- **Initialize:** $\boldsymbol{z}^0 = \boldsymbol{\beta}_0$, L and $t_0 = 1$.
- **General step:** for any $k = 0, 1, 2, \cdots$ execute the following steps:
  (a) set $\boldsymbol{\beta}_{k+1} = \mathrm{prox}_{(2\lambda_{n,1}/L)\|\boldsymbol{w}\circ\boldsymbol{\beta}\|_1}(\boldsymbol{z}^k - \frac{1}{L}G_k)$.
  (b) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (c) compute $\boldsymbol{z}^{k+1} = \boldsymbol{\beta}_{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k)$.

By Definition 6.1 in [1],

$$\mathrm{prox}_{\frac{2\lambda_{n,1}}{L}\|\boldsymbol{w}\circ\boldsymbol{z}\|_1}\left(\boldsymbol{z}^k - \frac{1}{L}\boldsymbol{G}_k\right) := \mathrm{argmin}_{\boldsymbol{z}\in\mathbb{R}^p}\frac{1}{2}\left\|\boldsymbol{z} - \left(\boldsymbol{z}^k - \frac{1}{L}G_k\right)\right\|_2^2 + \frac{2\lambda_{n,1}}{L}\|\boldsymbol{w}\circ\boldsymbol{z}\|_1.$$

Analytical study shows that

$$\boldsymbol{z}^{k+1} = \left[\left|\boldsymbol{z}^k - \frac{1}{L}\boldsymbol{G}_k\right| - \frac{2\lambda_{n,1}}{L}\boldsymbol{w}\right]_+ \circ \mathrm{sign}\left(\left[\boldsymbol{z}^k - \frac{1}{L}\boldsymbol{G}_k\right]\right).$$

The stopping criterion of FISTA is

$$\frac{\|\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k\|_2}{\max\{1, \|\boldsymbol{\beta}^{k+1}\|_2\}} \leq \epsilon,$$

where $\epsilon > 0$ is a small constant, or the maximum number of iterations is reached.

4.2. **Selection of tuning parameters.** As we know that almost all regularization estimators depend on tuning parameters. The tuning parameter for SL is suggested as $\lambda = \sqrt{2\log p/n}$ in [14]. Reference [17] also developed a fixed choice of tuning parameter $\lambda = \log p/n$ for OL in scenarios with low sparsity. However, these fixed parameters are completely ineffective for SL and OL in our simulations. Notice that the five-fold cross-validation (CV) (See [8, 14, 17]) works well for SL, NL, OL and NAL in [14, 17, 16]. Therefore, we used the five-fold CV to select tuning parameters in all estimations. It should be mentioned that the five-fold CV is only used for the selection of $\lambda_{n,1}$ in NAEN. Since the order of tuning parameter $\lambda_{n,2}$ in NAEN is $O(1/n)$ under unified conditions that $s\log p = o(n)$ and $\lambda_{n,2}\|\boldsymbol{\beta}^*\|_2 = O(\sqrt{s}/n)$, for convenience we take $\lambda_{n,2} = 1/n$.

4.3. **Simulation in moderately sparse scenarios.** Let $n = 100$, $p = 200$, $s = 30$, and the design matrix $\boldsymbol{X}$ is generated from the normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}$, where $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.3$. Two types of sparse regression parameters are considered in simulations.

- Model 1 (Fixed $\boldsymbol{\beta}^*$). The non-zero sub-vector of the true parameter $\boldsymbol{\beta}^*$ is taken as fixed vector $\mathbf{1}_s := (1, \cdots, 1)^T \in \mathbb{R}^s$.
- Model 2 (Random $\boldsymbol{\beta}^*$). Each non-zero element β_i^*, $i \in \mathcal{A}_0$ of the true parameter generated from $\text{Uni}[0.5, 2] \times \text{Ber}(\{1, -1\}, 0.5)$, where $\text{Uni}[0.5, 2]$ is uniform distribution over $[0.5, 2]$, $\text{Ber}(\{1, -1\}, 0.5)$ is a Bernoulli distribution with probability 0.5 taking the value 1 or -1.

In each setting, $N = 100$ replications are used. Table 2 reports AMSE and ARR of various estimators for different models. Figures 2 and 3 show 100 simulation results of various estimators in different scenarios. Clearly, the NAEN outperforms other methods, and its simulation curve is almost the same as that of the OE. Indeed, NAEN can be regarded as a lifted version of NAL. As mentioned by Proposition 1 in [16], the NAL for σ^2 is the optimal value of the optimization of the adaptive lasso, which is the sum of the residual-based estimator and the adaptive regularization term in the adaptive Lasso. However, when $\boldsymbol{\beta}^*$ has many non-zero elements, the residual-based estimator seriously underestimates the error variance. The value of the pure adaptive regularization term is not enough to bridge the gap between the residual-based estimator and the true error variance. The idea of NAEN is to fill this gap again through ridge term in (4). Moreover, MUE slightly underestimates the error variance in model 1 and significantly underestimates the error variance in model 2.

TABLE 2. AMSE and ARR of various methods in 100 simulations.

	OE	SL	NL	OL	NAL	MUE	NAEN
(Model 1)							
AMSE	0.022	0.724	14.302	0.120	0.144	0.040	0.019
ARR	0.990	0.153	4.780	1.341	0.642	0.906	0.940
(Model 2)							
AMSE	0.019	0.647	26.319	1.260	0.306	0.075	0.008
ARR	0.983	0.202	6.129	2.120	0.453	0.759	0.965

Furthermore, while NAL and OL underestimate and overestimate error variance, respectively, the corresponding AMSE and ARR are very close. Moreover, the trend

of the curve of NAL is close to that of OL and the curve of OL is flatter. Finally, SL severely underestimates error variance and the estimated values of NL are much larger than the true variance. Both approaches almost fail. We now explain this phenomenon. In fact, the essence of SL is a residual-based estimator, so SL tends to underestimate error variance. Proposition 1 in [17] shows that NL for σ^2 is the optimal value of the optimization of the Lasso. However, the Lasso estimator has the sure screening property, i.e., the selected model includes the true model and the number of the selected variables is much more than true regression parameters. Then, when each true parameter is moderate or large and the sparsity of the true parameter vector is low, the value of the regularization function of NL is large.

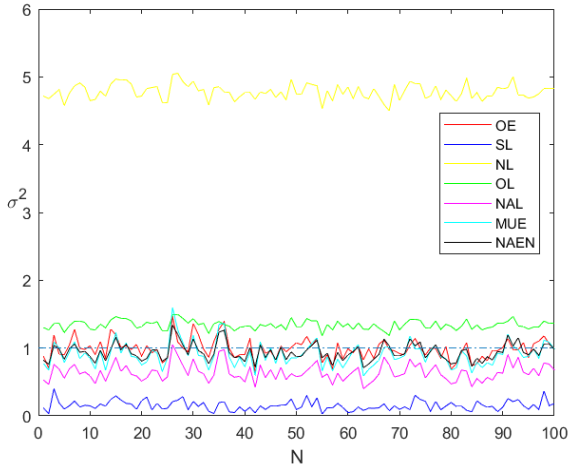

FIGURE 2. All results of various estimatiors for model 1.

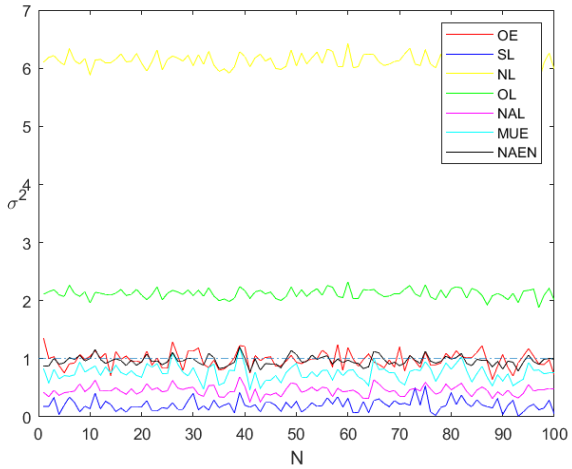

FIGURE 3. All results of various estimatiors for model 2.

4.4. **Simulation in extremely sparse scenarios.** The following simulation results demonstrate the limitation of the NAEN. Here, we consider the extremely sparse scenarios $s \in \{3, 5, 7, 10\}$. The remaining settings are the same as that in Section 4.3.

TABLE 3. AMSE and ARR of various methods in extremely sparse scenarios.

		OE	SL	NL	OL	NAL	MUE	NAEN
$s = 3$, Model 1	AMSE	0.027	0.040	0.019	0.025	0.027	0.027	0.025
	ARR	0.991	0.902	1.099	0.890	0.950	0.972	0.981
$s = 3$, Model 2	AMSE	0.019	0.037	0.071	0.031	0.054	0.029	0.066
	ARR	0.961	0.873	1.247	1.125	1.182	1.090	1.210
$s = 5$, Model 1	AMSE	0.021	0.028	0.095	0.058	0.020	0.019	0.018
	ARR	1.016	0.943	1.297	1.214	0.963	1.004	1.101
$s = 5$, Model 2	AMSE	0.022	0.035	0.159	0.018	0.040	0.039	0.067
	ARR	0.972	0.915	1.390	1.094	1.134	1.121	1.211
$s = 7$, Model 1	AMSE	0.020	0.028	0.060	0.015	0.028	0.022	0.027
	ARR	0.983	0.956	1.235	1.076	0.955	0.982	1.026
$s = 7$, Model 2	AMSE	0.020	0.030	0.126	0.016	0.023	0.021	0.035
	ARR	1.018	0.952	1.348	0.899	1.011	1.036	1.116
$s = 10$, Model 1	AMSE	0.022	0.035	0.016	0.019	0.032	0.022	0.017
	ARR	0.994	0.918	1.111	0.885	0.875	0.954	0.975
$s = 10$, Model 2	AMSE	0.015	0.094	0.481	0.083	0.021	0.018	0.058
	ARR	0.995	0.741	1.690	1.274	0.987	1.017	1.193

In extremely sparse scenarios, we are more likely to obtain the true model or approximately accurate models which includes the true model, and the number of selected variables is not significantly different from the true model. Table 3 lists AMSE and ARR for 100 simulations of several methods. Overall, most estimation methods perform well in extremely sparse scenarios. In general, the NAL or the NAEN and MUE are respectively the sub-optimal and optimal when the true model is accurately or approximately accurately selected. Intuitively, the NAEN does not have any advantages over the MUE and the NAEN is often greater than the MUE.

5. **Conclusion.** In this paper, we proposed the NAEN for variance estimation in high-dimensional linear models, which simultaneously selects and estimates regression and variance parameters. The established theory shows that the NAEN estimator for σ^2 is essentially the optimal value of the optimization problem of the naive adaptive elastic-net for regression coefficients. Furthermore, we established the asymptotic properties of the NAEN for $\boldsymbol{\beta}$ and σ^2. The FISTA was used to obtain estimators of NAEN for σ^2 and $\boldsymbol{\beta}$. The simulation results show that the NAEN is suitable for variance estimation in scenarios with moderate sparsity.

Essentially, the NAEN, NL, OL, and NAL are respectively the optimal values of the adaptive elastic-net, Lasso, organic Lasso, and adaptive Lasso. These methods involve the selection of tuning parameters, which is a costly process. Research on methods for variance estimation with fixed tuning parameters or easily controllable tuning parameters may be more attractive in the future.

Acknowledgments. The authors thank two anonymous reviewers and the editor for their constructive comments and suggestions that helped to improve the previous version of this paper.

REFERENCES

[1] A. Beck, *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, 2017.
[2] A. Beck and M. Teboulle, A Fast Iterative Shrinkage-Thresholding Algorithm for linear Inverse Problems, *SIAM J. Imaging Sciences*, **2** (2009), 183-202.
[3] P. J. Bickel, Y. A. Ritov and A. B. Tsybakov, Simultaneous analysis of lasso and dantzig selector, *Ann. Stat.*, **37** (2009), 1705-1732.
[4] F. Campbell and G. I. Allen, Within group variable selection through the exclusive lasso, *Electron. J. Statist.*, **11** (2017), 4220-4257.
[5] E. Candes and T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, *Ann. Stat.*, **35** (2007), 2313-2351.
[6] L. H. Dicker, Variance estimation in high-dimensional linear models, *Biometrika*, **101** (2014), 269-284.
[7] J. Fan, Y. Fan and E. Barut, Adaptive robust variable selection, *Ann. Stat.*, **42** (2014), 324-351.
[8] J. Fan, S. Guo and N. Hao, Variance estimation using refitted cross-validation in ultrahigh dimensional regression, *J. R. Stat. Soc. Ser. B*, **74** (2012), 37-65.
[9] J. Fan and Y. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.*, **106** (2001), 1348-1360.
[10] C. Giraud, *Introduction to High-Dimensional Statistics*, 1^{st} edition, Chapman and Hall/CRC, New York, 2014.
[11] J. Huang, J. L. Horowitz and S. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Ann. Stat.*, **36** (2008), 587-613.
[12] X. Liu, S. Zheng and X. Feng, Estimation of error variance via ridge regression, *Biometrika*, **107** (2020), 481-488.
[13] N. Städler and P. Bühlmann, ℓ_1-penalization for mixture regression models, *Test*, **19** (2010), 209-256.
[14] T. Sun and C.-H. Zhang, Scaled sparse linear regression, *Biometrika*, **99** (2012), 879-898.
[15] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B*, **73** (1996), 273-282.
[16] X. Wang, L. Kong and L. Wang, Estimation of error variance in regularized regression models via adaptive lasso, *Mathematics*, **10** (2022), 1937.
[17] G. Yu and J. Bien, Estimating the error variance in a high-dimensional linear model, *Biometrika*, **106** (2019), 533-546.
[18] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Ann. Stat.*, **38** (2010), 894-942.
[19] C.-H. Zhang and J. Huang, The sparsity and bias of the lasso selection in high-dimensional linear regression, *Ann. Stat.*, **36** (2008), 1567-1594.
[20] Y. Zhou, R. Jin and S. Hoi, Exclusive lasso for multi-task feature selection, *J. Mach. Learn. Res.*, **9** (2010), 988-995.
[21] H. Zou, The adaptive lasso and its oracle properties, *J. R. Stat. Soc. Ser. B*, **101** (2006), 1418-1429.
[22] H. Zou and H.-H. Zhang, On the adaptive elastic-net with a diverging number of parameters, *Ann. Stat.*, **37** (2009), 1733-1751.

Received September 2022; revised May 2023; early access July 2023.