mathematics

MDPI

Article

Estimation of Error Variance in Regularized Regression Models via Adaptive Lasso

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

[1] Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China; lchkong@bjtu.edu.cn
[2] Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada; liqun.wang@umanitoba.ca
* Correspondence: 18118020@bjtu.edu.cn

Abstract: Estimation of error variance in a regression model is a fundamental problem in statistical modeling and inference. In high-dimensional linear models, variance estimation is a difficult problem, due to the issue of model selection. In this paper, we propose a novel approach for variance estimation that combines the reparameterization technique and the adaptive lasso, which is called the natural adaptive lasso. This method can, simultaneously, select and estimate the regression and variance parameters. Moreover, we show that the natural adaptive lasso, for regression parameters, is equivalent to the adaptive lasso. We establish the asymptotic properties of the natural adaptive lasso, for regression parameters, and derive the mean squared error bound for the variance estimator. Our theoretical results show that under appropriate regularity conditions, the natural adaptive lasso for error variance is closer to the so-called oracle estimator than some other existing methods. Finally, Monte Carlo simulations are presented, to demonstrate the superiority of the proposed method.

Keywords: high-dimensional linear model; variance estimation; natural adaptive lasso; mean squared error bound; regularized regression

MSC: 62F10; 62J05; 62J10

Citation: Wang, X.; Kong, L.; Wang, L. Estimation of Error Variance in Regularized Regression Models via Adaptive Lasso. *Mathematics* **2022**, *10*, 1937. https://doi.org/10.3390/math10111937

Academic Editor: Andreas Artemiou

Received: 24 April 2022
Accepted: 30 May 2022
Published: 6 June 2022

1. Introduction

Consider linear regression model $y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon$, 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 regression parameter and $\varepsilon \in \mathbb{R}$ is the random error satisfying $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$. Given an *i.i.d.* random sample $(y_i, \boldsymbol{x}_i^T)^T \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, the model can be written in the matrix form as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$. In this paper, we are mainly interested in the high-dimensional sparse model, where $p \gg n$.

Regularized methods for simultaneous model selection and parameter estimation have been intensively studied in the literature, e.g., the lasso [1], smoothly clipped absolute deviation (SCAD) [2], adaptive lasso [3], bridge [4], adaptive elastic net [5], and minimax concave penalty (MCP) [6], as well as the Dantzig selector [7]. In addition, screening rules for dimension reduction are proposed, e.g., the sure independent screening method and iteratively sure independent screening method [8], lasso-based screening rules [9–11], etc.

However, most of these works focus on selection and estimation, with respect to regression parameters, and few studies deal with estimation of error variance, although it is a fundamental and crucial problem in statistical inference and regression analysis. In conventional linear models, the common estimator, based on residuals, plays an important role in statistical inferences and model checking. In high-dimensional models, however, variance estimation becomes a difficult problem, mainly due to two reasons. One is that the traditional residual-based methods may perform poorly or, even, fail, as, for example, the ordinary least squares method does not work when the number of covariates is greater than the sample size. The other reason is that it is difficult to select the true model,

accurately, since in practice the selected model, often, contains spurious variables that are correlated with the residuals, resulting in significant underestimation of error variance (e.g., [12,13]).

Next, we provide some examples, where model error variance is involved and plays an important role.

Example 1 (Model selection)**.** *Penalization is a common approach to model selection and parameter estimation, in high-dimensional linear models. The efficiency and accuracy of such methods depend on certain tuning parameters that are chosen using some criteria, such as Mallows's C_p, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). For example, the AIC and BIC for the lasso [14] are given by*

$$\mathrm{AIC}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}, \sigma^2) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2}{n\sigma^2} + \frac{2}{n}\mathrm{df}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n})$$

and

$$\mathrm{BIC}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}, \sigma^2) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2}{n\sigma^2} + \frac{\log(n)}{n}\mathrm{df}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n})$$

respectively, where $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ is the lasso estimator with tuning parameter λ_n and the degrees of freedom $\mathrm{df}(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n})$ is equal to the number of non-zero elements in $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. It is easy to see that these criteria rely on error variance.

Example 2 (Confidence intervals)**.** *For a least-squares-based penalized estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$, let $\widehat{\mathcal{A}}$ be its index set, corresponding to non-vanishing parameters. If $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ has the oracle property, then for each $i \in \widehat{\mathcal{A}}$, the $1-\alpha$ confidence interval for β_i is given by*

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

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard normal distribution and c_i is the i-th diagonal element of the matrix $(\boldsymbol{X}_{\widehat{\mathcal{A}}}^T\boldsymbol{X}_{\widehat{\mathcal{A}}})^{-1}$. It is clear that the above intervals depend on the variance parameter.

Example 3 (Penalized second-order least squares estimation)**.** *The second-order least squares method, in [15], extends the ordinary least squares method by, simultaneously, minimizing the first two order distances*

$$\rho_i(\boldsymbol{\beta}, \sigma^2) = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta},\ y_i^2 - (\boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \sigma^2)^T$$

and yields the joint estimators for the regression and variance parameters. Under general conditions, the second-order least squares estimator has been shown to be, asymptotically, more efficient than the ordinary least squares estimator, if the model error has a nonzero third moment, and they are equivalent otherwise. The regularized version of this method can be used in high-dimensional models.

1.1. Literature Review

Variance estimation in high-dimensional models has attracted increasing attention in recent years. Here, we briefly review some important advances in this area. First, if the true parameter vector $\boldsymbol{\beta}^*$ was known, then the ideal variance estimator, called the oracle estimator, is $\sigma^2_{\text{oracle}} = (1/n)\sum_{i=1}^n(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*)^2$. Correspondingly, the estimator $\sigma^2_{\text{naive}} = \sum_{i=1}^n(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2/n$, based on some estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, is called a naive estimator. Since the naive estimator is downward biased, a modified unbiased estimator is given by $\widehat{\sigma}^2 = \sum_{i=1}^n(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2/(n - \widehat{s})$, where $\widehat{s} := \#\{i : \widehat{\beta}_i \neq 0\}$ is the number of nonzero elements in $\widehat{\boldsymbol{\beta}}$. Unfortunately, when p is much larger than n, even a small change in $\widehat{s}$ will cause huge fluctuation in $\widehat{\sigma}^2$, if $\widehat{s} \approx n$.

To overcome this problem, Ref. [16] estimated the mean and variance parameters jointly, by maximizing a reparameterized likelihood with ℓ_1 penalty:

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

where $\boldsymbol{\phi} = \boldsymbol{\beta}/\sigma$, $\rho = 1/\sigma$ and $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$. Moreover, they proposed a generalized EM algorithm for the numerical optimization.

A refitted cross-validation (RCV) method, to derive a variance estimator, was proposed in [12], and its asymptotic properties were studied. The main idea of RCV is to attenuate the influence of irrelevant variables with high spurious correlations, via a data-splitting technique. Ref. [12], also, discussed the asymptotic properties of the lasso-based estimator $\widehat{\sigma}^2_{\text{lasso}} = \sum_{i=1}^n (\boldsymbol{y} - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{\text{lasso}} s)^2/(n - \widehat{s}_{\text{lasso}})$ and SCAD-based estimator $\widehat{\sigma}^2_{\text{SCAD}} = \sum_{i=1}^n (\boldsymbol{y} - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{\text{SCAD}})^2/(n - \widehat{s}_{\text{SCAD}})$, where $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$ and $\widehat{\boldsymbol{\beta}}_{\text{SCAD}}$ are the least squares estimator, with ℓ_1 penalty [1] and SCAD penalty [2], respectively; $\widehat{s}_{\text{lasso}} = \#\{i : (\widehat{\boldsymbol{\beta}}_{\text{lasso}})_i \neq 0\}$ and $\widehat{s}_{\text{SCAD}} = \#\{i : (\widehat{\boldsymbol{\beta}}_{\text{SCAD}})_i \neq 0\}$.

Further, a scaled lasso was proposed in [17], for simultaneous estimation of regression and variance parameters. Their model can be written as

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

Under some regularity conditions, Ref. [17] proved the oracle inequalities for prediction and their estimators.

A moment estimator for the error variance, based on the covariance matrix Σ of the predictor variables, was studied in [18], where three cases were considered: $\Sigma = I$, estimable Σ and non-estimable Σ. A maximum likelihood method for the normally distributed noise was developed in [19].

Moreover, Ref. [13] considered another re-parameterized likelihood, with lasso penalty

$$(\widehat{\boldsymbol{\theta}}_{\lambda_n}, \widehat{\phi}_{\lambda_n}) \in \arg \min_{\boldsymbol{\theta}\in\mathbb{R}^p, \phi\in\mathbb{R}_{++}} \left\{ -\frac{1}{2}\log\phi + \phi\frac{\|\boldsymbol{y}\|_2^2}{2n} - \frac{1}{n}\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\theta} + \frac{\|\boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2n\phi} + \lambda_n \Omega(\boldsymbol{\theta}, \phi) \right\}, \quad (1)$$

where $\phi_{\lambda_n} = 1/\sigma^2_{\lambda_n}$, $\boldsymbol{\theta}_{\lambda_n} = \phi_{\lambda_n}\boldsymbol{\beta}_{\lambda_n}$. In particular, they proposed two estimators: the natural lasso with $\Omega(\boldsymbol{\theta}, \phi) = \|\boldsymbol{\theta}\|_1$ and the organic lasso with $\Omega(\boldsymbol{\theta}, \phi) = \phi^{-1}\|\boldsymbol{\theta}\|_1^2$.

Finally, Ref. [20] proposed a ridge-based method to estimate the error variance, under certain conditions, which is defined as follows:

$$\widehat{\sigma}^2 = \{1 - n^{-1}\text{tr}(\boldsymbol{A}_{1n})\}^{-1}\check{\sigma}^2,$$

where $\check{\sigma}^2 = n^{-1}\boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{A}_{1n})\boldsymbol{y}$, $\boldsymbol{A}_{1n} = n^{-1}\boldsymbol{X}(n^{-1}\boldsymbol{X}^T\boldsymbol{X} + \eta\boldsymbol{I}_p)\boldsymbol{X}^T$ and η is the tuning parameter. This method performs well in low-dimensional cases, with weak signals, and it is suitable for sparse as well as non-sparse models.

1.2. Notation and Outline

Throughout the paper, let $\mathcal{A}_0 := \{i : \beta_i^* \neq 0\}$ be the index set and $s := \#\{i : \beta_i^* \neq 0\}$ be the number of the nonzero elements of $\boldsymbol{\beta}^*$, respectively. Given a design matrix $\boldsymbol{X}$ and a subset $\mathcal{A}$ of $\{1, \ldots, p\}$, $\boldsymbol{X}_i$ denotes the i-th column vector of $\boldsymbol{X}$, and $\boldsymbol{X}_\mathcal{A}$ denotes the sub-matrix, consisting of the columns with indices in $\mathcal{A}$. For vectors $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{x} \circ \boldsymbol{y} := (x_1y_1, \ldots, x_py_p)^T$ denotes the Hadamard product. Moreover, let $1/|\boldsymbol{x}|$ or $|\boldsymbol{x}|^{-1} =$

$(1/|x_1|,\ldots,1/|x_p|)^T$, $\mathrm{sgn}(\boldsymbol{x}) = (\mathrm{sgn}(x_1),\ldots,\mathrm{sgn}(x_p))^T$, $\mathrm{sign}(\boldsymbol{x}) = (\mathrm{sign}(x_1),\ldots,\mathrm{sign}(x_p))^T$, $\partial\|\boldsymbol{x}\|_1 = \partial|x_1| \times \ldots \times \partial|x_p|$, where

$$\mathrm{sgn}(t) = \begin{cases} 0, & t \neq 0, \\ 1, & t = 0, \end{cases} \quad \mathrm{sign}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0, \end{cases} \quad \text{and } \partial|t| = \begin{cases} \{1\}, & t > 0, \\ [-1,1], & t = 0, \\ \{-1\}, & t < 0. \end{cases}$$

The rest of this paper is organized as follows: Section 2 defines and describes the proposed natural adaptive lasso, and Section 3 gives its asymptotic properties. Section 4 deals with the numerical optimization of the proposed estimators. Monte Carlo simulation studies of finite sample properties are provided in Section 5. The conclusions and discussion are given in Section 6, while the mathematical proofs are given in Section 7.

2. Natural Adaptive Lasso (NAL)

Some researchers, e.g., Refs. [13,16], used reparameterized likelihood to jointly estimate the mean and variance parameters in high-dimensional linear models. In particular, the method of [13] has good performance, and the associated numerical computation can be converted to some simple optimization procedures. However, the natural lasso in [13] always overestimates error variance, due to the over-selection of the covariates. This motivates us to consider the more generally adaptive lasso penalty, to further improve the properties of the estimators. Consider the following adaptively weighted ℓ_1-penalized likelihood

$$(\widehat{\boldsymbol{\theta}}_{\lambda_n}, \widehat{\phi}_{\lambda_n}) \in \arg\min_{\theta,\phi\in\mathbb{R}_+} \left\{ L(\boldsymbol{\theta}_{\lambda_n}, \phi_{\lambda_n}) + \lambda_n \|\boldsymbol{w} \circ \boldsymbol{\theta}\|_1 \right\}, \tag{2}$$

where $L(\boldsymbol{\theta}_{\lambda_n}, \phi_{\lambda_n})$ is the reparameterized likelihood as (1), λ_n is the tuning parameter and $\boldsymbol{w} := (w_1,\ldots,w_p)^T$ is the adaptive weight vector. Given a solution $(\widehat{\boldsymbol{\theta}}_{\lambda_n}, \widehat{\phi}_{\lambda_n})$ of problem (2), the natural adaptive lasso estimators (NALE) for $\boldsymbol{\beta}$ and σ^2 are given by

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

It is easy to see that, when $\boldsymbol{w} = \mathbf{1}$, the NALE reduces to the natural lasso estimator (NLE) of [13].

Note that the quality of the NALE depends on the weight vector $\boldsymbol{w}$. It follows from Proposition 1 in Section 3, that the weight $\boldsymbol{w}$ in problem (2) plays the same role as in the adaptive lasso estimation of the regression coefficients only, which solves the following convex optimization problem:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ada}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + 2\lambda_n \|\boldsymbol{w} \circ \boldsymbol{\beta}\|_1 \right\}, \tag{4}$$

where the weight $\boldsymbol{w}$ depends on the initial estimator $\widetilde{\boldsymbol{\beta}}^{\mathrm{ini}}$. As indicated by [3], any root-n consistent estimator can be used as the initial estimator for $\boldsymbol{\beta}$. For example, the least squares estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{ols}} := (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ can be used, and the weight vector is calculated as $\boldsymbol{w} = 1/|\widehat{\boldsymbol{\beta}}_{\mathrm{ols}}|^\gamma$, $\gamma > 0$. Ref. [4] discusses the selection of the initial estimators in linear models, with $\log p = O(n^a)$ for some $a \in (0,1)$. They show that their marginal regression estimator can be used in the adaptive lasso, to yield the desirable selection and estimation properties. In addition, the weight $\boldsymbol{w}$ in adaptive elastic-net [5], for moderate dimensional models ($\log p = O(\log n)$), can be constructed as $\boldsymbol{w} = 1/|\widehat{\boldsymbol{\beta}}_{\mathrm{net}} + (1/n)\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{\mathrm{net}})|^\gamma$, $\gamma > 0$, where $\widehat{\boldsymbol{\beta}}_{\mathrm{net}}$ is the elastic-net estimator. In this paper, we use the following two-step procedure to calculate the weight vector.

Step 1: Solve the lasso problem to obtain the NLE $\widehat{\boldsymbol{\beta}}_{\text{lasso}}$*, which is used as the initial estimator* $\widetilde{\boldsymbol{\beta}}^{\text{ini}}$.

Step 2: Set $\boldsymbol{w}$ *with* $w_j = p'_{\lambda_n}(|\widetilde{\beta}_j^{\text{ini}}|)$*, where* $j = 1, \ldots, p$ *and* p_{λ_n} *is a folded-concave penalty function (such as SCAD, MCP or bridge).*

Remark 1. *From [7,21,22], under some regularity conditions, the lasso is consistent with a near-oracle rate* $\sqrt{s \log p/n}$ *and has the sure-screening property, i.e.,*

$$\|\widehat{\boldsymbol{\beta}}_{\text{lasso}} - \boldsymbol{\beta}^*\|_2 \leq O(\sqrt{s \log(p)/n}),\ \text{supp}(\widehat{\boldsymbol{\beta}}_{\text{lasso}}) \supseteq \text{supp}(\boldsymbol{\beta}^*).$$

Further, based on the order of the bias of the lasso, under suitable conditions for the minimum signal strength (dee the first part of Condition 4 in Section 7) and the choice of tuning parameter, $\boldsymbol{w}_{\mathcal{A}_0}$ *will be close, or even equal, to zero vector, when* n *is sufficiently larger, if a folded-concave penalty, such as SCAD, is used. These properties play an important role in some of the conclusions that follow.*

3. Asymptotic Properties

In this section, we, first, establish the relationship between the NALE and the adaptive lasso, then analyze the asymptotic properties of the NALE for σ^2.

Proposition 1. *The NALE estimator* $(\widehat{\boldsymbol{\beta}}_{\lambda_n}, \widehat{\sigma}^2_{\lambda_n})$*, defined in (3), where* $(\widehat{\boldsymbol{\theta}}_{\lambda_n}, \widehat{\phi}_{\lambda_n})$ *is a solution of (2), satisfies the following properties:*

(i) $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ *is a solution of the adaptive lasso (4);*

(ii) $\widehat{\sigma}^2_{\lambda_n}$ *is the optimal value, of the objective function of the adaptive lasso (4). Furthermore, we have* $\widehat{\sigma}^2_{\lambda_n} = n^{-1}(\|\boldsymbol{y}\|_2^2 - \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2)$.

The results of Proposition 1 are instrumental in the derivation of the other theoretical results in this paper. Moreover, they, also, provide a method for calculating the NALE for $\boldsymbol{\beta}$ and σ^2. It is well known that the adaptive lasso (4) is a convex optimization, and many existing optimization tools can be used to compute this problem.

Note that, since

$$\begin{aligned} \widehat{\sigma}^2_{\lambda_n} &= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2 + 2\lambda_n\|\boldsymbol{w} \circ \widehat{\boldsymbol{\beta}}_{\lambda_n}\|_1 \\ &= \widehat{\sigma}^2_{\text{naive}}(\widehat{\boldsymbol{\beta}}_{\lambda_n}) + 2\lambda_n\|\boldsymbol{w} \circ \widehat{\boldsymbol{\beta}}_{\lambda_n}\|_1 \end{aligned} \tag{5}$$

and $\|\boldsymbol{w} \circ \widehat{\boldsymbol{\beta}}_{\lambda_n}\|_1 = \|\boldsymbol{w}_{\mathcal{A}_0} \circ \widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}\|_1$ will be close or even equal to zero, for suitably chosen $\boldsymbol{w}$, the NALE for σ^2 will be close to the naive estimator, if $\lambda_n \to 0$. As mentioned before, the naive estimator for σ^2, based on the adaptive lasso estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$, may work well when non-zero variables are selected, accurately. However, when more irrelevant variables are selected, the value of the penalty term will not be close to 0 in the finite sample, so that the naive estimator for σ^2 will, always, underestimate the true error variance. In this case, the penalty term will mitigate the difference between the naive estimator and the true variance. Although the form of the natural lasso estimator of [13] is similar to (5), their method often tends to over-select predictors, due to the use of a lasso penalty. In addition, the value of the penalty term in [13] remains large because it is not controlled by the weight vector. These facts explain why the natural lasso estimator for σ^2 tends to be larger than the true error variance in the simulation studies, in [13].

Next, we establish a key inequality for the NALE for σ^2.

Lemma 1. *If* $\lambda_n \geq \frac{1}{n}\|\boldsymbol{X}^T\boldsymbol{\varepsilon}\|_\infty$*, then*

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

The above inequality is deterministic, in that it does not rely on any statistical assumptions for $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$. Unlike Lemma 1 in [13], the proof of this result uses the fact that any vector $\boldsymbol{\beta}$ provides an upper bound on $\widehat{\sigma}^2_{\lambda_n}$ and the convexity of the loss function. In addition, if $\boldsymbol{w} = (1, \ldots, 1)^T$ and $O(\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_1) \leq O(\|\boldsymbol{\beta}^*\|_1)$, Lemma 1 reduces to Lemma 1 in [13]. If $\boldsymbol{w}$ is close or equal to zero, and $O(\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_1) \leq O(\|\boldsymbol{\beta}^*\|_1)$, then the bound on the right-hand side of the inequality in Lemma 1 is lower than that for the natural lasso and organic lasso in [13].

3.1. Adaptive Lasso

It follows, from Lemma 1, that the error bound of the NALE for σ^2 is controlled by the convergence rate of the adaptive-lasso estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. Therefore, it is necessary to establish the asymptotic properties for $\widehat{\boldsymbol{\beta}}_{\lambda_n}$. The results in this subsection are similar to that in [23]. All regularity conditions and proofs are given in Section 7.

Theorem 1. *Suppose Conditions 1–3 hold. Assume that*

$$\min_{i \in \mathcal{A}_0^c} w_i^* > C_1^{-1}, \ \lambda_n = 4C_1\sigma\sqrt{(2\log p + 2L)/n}$$

and $s(\log(p) + L)/n \to 0$, *where* C_1 *is some positive constant and* $L > 0$. *Then, with probability at least* $1 - e^{-L}$, *there exists unique minimizer* $\widehat{\boldsymbol{\beta}}_{\lambda_n} = (\widehat{\boldsymbol{\beta}}^T_{\mathcal{A}_0}, \widehat{\boldsymbol{\beta}}^T_{\mathcal{A}_0^c})$ *of problem* (4), *such that* $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c} = \mathbf{0}$ *and* $\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_2 \leq a_n$, *where*

$$a_n = C_4(\sqrt{s(2\log p + 2L)/n} + 2\lambda_n(\|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2 + C_2C_3\sqrt{s(\log p)/n}))$$

with some constant $C_4 > 0$, C_2 *and* C_3 *are defined in the regularity conditions.*

It follows from inequality (18) that the extra term $\lambda_n\sqrt{s(\log p)/n}$ in a_n is due to the bias of the initial estimator $\widetilde{\boldsymbol{\beta}}^{\text{ini}}$. When λ_n tends to zero, the order of the extra term is $o(\sqrt{s(\log p)/n})$. Thus, under some general conditions, the convergence rate of $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ is $O(\sqrt{s(\log(p) + L)/n})$. Usually, the order of L is $O(\log p)$.

We, now, present the asymptotic normality of the adaptive-lasso estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$.

Theorem 2. *Assume that conditions of Theorem 1 hold. Let* $s_n^2 = (1/n)\sigma^2\boldsymbol{\alpha}_n^T\boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{X}_{\mathcal{A}_0}\boldsymbol{\alpha}_n$ *for any* $\boldsymbol{\alpha}_n \in \mathbb{R}^s$ *satisfying* $\|\boldsymbol{\alpha}\|_2 \leq 1$. *Then, under Conditions 1–4, with probability at least* $1 - e^{-L}$, *the minimizer* $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ *in Theorem 1 satisfies*

$$\begin{aligned} n^{\frac{1}{2}}s_n^{-1}\boldsymbol{\alpha}_n^T\Big[(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0} - \boldsymbol{\beta}^*_{\mathcal{A}_0}) + n\boldsymbol{\alpha}_n^T(\boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{X}_{\mathcal{A}_0})^{-1}\lambda_n\boldsymbol{w}^*_{\mathcal{A}_0} \circ \boldsymbol{g}^*_{\mathcal{A}_0}\Big] \\ = n^{1/2}s_n^{-1}\boldsymbol{\alpha}_n^T(\boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{X}_{\mathcal{A}_0})^{-1}\boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{\varepsilon} + o_p(1) \xrightarrow{\mathcal{D}} N(0,1), \end{aligned}$$

where $g^*_{\mathcal{A}_0} \in \partial\|\boldsymbol{\beta}^*_{\mathcal{A}_0}\|_1$.

The result of Theorem 2 is consistent with the asymptotic normality, for the bridge estimator of $\boldsymbol{\beta}$ in [4]. The only difference is in the form of the penalty function.

Next, we consider the convergence performance of the specific adaptive-lasso estimator $\widehat{\boldsymbol{\beta}}_{\lambda_n}$, with a weight vector decided by the SCAD penalty [23], which is defined by

$$p'_{\lambda_n}(|t|) = \mathbf{1}\{|t| \leq \lambda_n^{\text{SCAD}}\} + \frac{(a\lambda_n^{\text{SCAD}} - |t|)_+}{(a-1)\lambda_n^{\text{SCAD}}}\mathbf{1}\{|t| > \lambda_n^{\text{SCAD}}\},$$

where $a > 2$ is a given constant and $(\cdot)_+ := \max\{0, \cdot\}$. Usually, the order of $\lambda_n^{\mathrm{SCAD}}$ is $O(\sqrt{s(\log p + L)/n})$. By definition, it holds $w^*_{\mathcal{A}_0} = \mathbf{0}$, and Condition 4 is satisfied when $\min_{i\in\mathcal{A}_0} |\beta_i^*| \geq 2a\lambda_n^{\mathrm{SCAD}}$. Thus, we have the following result.

Corollary 1. *Assume that the conditions of Theorem 1 hold. Then, under Conditions 1–4, with probability at least* $1 - e^{-L}$*, there exists unique minimizer* $\widehat{\boldsymbol{\beta}}_{\lambda_n} = (\widehat{\boldsymbol{\beta}}^T_{\mathcal{A}_0}, \widehat{\boldsymbol{\beta}}^T_{\mathcal{A}_0^c})$ *of problem (4), such that*

$$\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_2 \leq O(\sqrt{s(\log p + L)/n}),$$
$$\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}) = \mathrm{sgn}(\boldsymbol{\beta}^*_{\mathcal{A}_0}) \text{ and } \widehat{\boldsymbol{\beta}}_{\mathcal{A}_0^c} = \mathbf{0}.$$

Furthermore, $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ *satisfies*

$$n^{\frac{1}{2}} s_n^{-1} \boldsymbol{\alpha}_n^T (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0} - \boldsymbol{\beta}^*_{\mathcal{A}_0}) = n^{1/2} \boldsymbol{\alpha}_n^T (\boldsymbol{X}^T_{\mathcal{A}_0} \boldsymbol{X}_{\mathcal{A}_0})^{-1} \boldsymbol{X}^T_{\mathcal{A}_0} \boldsymbol{\varepsilon} \xrightarrow{\mathcal{D}} N(0,1),$$

where $s_n^2 = (1/n)\sigma^2 \boldsymbol{\alpha}_n^T \boldsymbol{X}^T_{\mathcal{A}_0} \boldsymbol{X}_{\mathcal{A}_0} \boldsymbol{\alpha}_n$ *for any* $\boldsymbol{\alpha}_n \in \mathbb{R}^s$ *satisfying* $\|\boldsymbol{\alpha}_n\|_2 \leq 1$.

The rate of convergence of the estimators in Theorem 1 and Corollary 1 is controlled by the distribution of random error and predictor matrix. Moreover, these results can be generalized for other situations, where random error follows sub-Gaussian or sub-exponential distributions.

3.2. Error Bounds of NALE

In this subsection, we establish the error bound for the NALE of σ^2. It follows from (14) that, under the conditions of Theorem 1, $\lambda_n \geq (1/n)\|\boldsymbol{X}^T\boldsymbol{\varepsilon}\|_\infty$ holds, with probability $1 - e^{-L}$. Since $s(\log(p) + L)/n \to 0$, we have $a_n \to 0$. Thus, in order to establish the asymptotic properties of NALE for σ^2, we still need to determine the order of $\lambda_n \|\boldsymbol{w} \circ \boldsymbol{\beta}^*\|_1$. By Condition 2 and Theorem 1, we have

$$\begin{aligned} \|\boldsymbol{w} \circ \boldsymbol{\beta}^*\|_1 &= \sum_{i\in\mathcal{A}_0} w_i |\beta_i^*| \leq \sum_{i\in\mathcal{A}_0} (C_3 |\widehat{\beta}_i - \beta_i^*| + w_i^*) |\beta_i^*| \\ &\leq C_3 a_n \|\boldsymbol{\beta}^*\|_1 + \|\boldsymbol{w}^* \circ \boldsymbol{\beta}^*\|_1. \end{aligned} \tag{6}$$

Thus, we have the following result on the error bound of the NALE for σ^2.

Theorem 3. *Under the conditions of Theorem 1, the NALE for* σ^2 *has the following error bound, with probability at least* $1 - e^{-L}$*:*

$$\left| \widetilde{\sigma}^2_{\lambda_n} - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 \right| \leq b_n,$$

where $b_n = 2\lambda_n \max\left\{ C_3 a_n \|\boldsymbol{\beta}^*\|_1 + \|\boldsymbol{w}^* \circ \boldsymbol{\beta}^*\|_1, \sqrt{s} a_n \right\}$.

The proof of the above result follows, straightforwardly, from Lemma 1 and Theorem 1, so it is omitted. Since $a_n \to 0$, $\|\boldsymbol{w}^* \circ \boldsymbol{\beta}^*\|_1$ is close or equal to zero, and the order of λ_n for the adaptive lasso is $O(\sqrt{(\log(p) + L)/n})$, we have $b_n = o(\sqrt{s(\log(p) + L)/n})$. It follows that when $L = O(\log(p))$, the error bound of NALE for σ^2 is smaller than that of the NLE, OLE and SLE, when n is sufficiently large. In the following, we analyze the mean squared error bound for the NALE of σ^2.

Theorem 4. *Under the conditions of Theorem 1, for any $M > 1$ and $\lambda_n = 4C_1\sigma\sqrt{(2M\log p)/n}$, the NALE for σ^2 satisfies*

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

Note that the above mean squared error bound of NALE for σ^2 is lower than that for the NLE, OLS and SLE estimators. Finally, we consider the case using the SCAD penalty. Then, by Theorem 3 and the fact that $\|\boldsymbol{w} \circ \boldsymbol{\beta}^*\|_1 = 0$, under the condition on minimum signal strength, we have the following result.

Corollary 2. *Under the conditions of Corollary 1, the NALE for σ^2 using the SCAD has the following error bound, with probability at least $1 - e^{-L}$:*

$$\left|\widehat{\sigma}^2_{\lambda_n} - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2\right| \leq 2\sqrt{s}\lambda_n a_n.$$

Further, by Theorem 4 and Corollary 2, we have the mean squared error bound of the NALE for σ^2 using the SCAD.

Corollary 3. *Under the conditions of Corollary 1, for any $M > 1$, the NALE for σ^2 using SCAD with $\lambda_n = 4C_1\sigma\sqrt{(2M\log p)/n}$ satisfies the following relative mean squared error bound:*

$$E\left\{\left(\frac{\widehat{\sigma}^2_{\lambda_n}}{\sigma^2} - 1\right)^2\right\} \leq \left[\left(M + \frac{p^{1-M}}{\log p}\right)^{\frac{1}{2}} \frac{4s\lambda_n^2 a_n^2}{\sigma^2} + \left(\frac{2}{n}\right)^{\frac{1}{2}}\right]^2.$$

4. Numerical Optimization

In this section, we study the optimization method for the NALE. Proposition 1 provides an easy way to calculate the NALE for σ^2, through existing optimization tools, to compute the adaptive lasso (4). Given the tuning parameter λ_n, we consider the proximal gradient algorithm (PGA) to calculate this problem, which has the following steps:

Initialization: *take initial value* $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, $\tau \in (0, \tau^*)$.
Iterative step: $\boldsymbol{\beta}^{k+1} = \text{prox}_{\tau\lambda_n\|\boldsymbol{\beta}\|_1}(\boldsymbol{\beta}^k - \frac{2\tau}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}^k - \boldsymbol{y}))$.

In the above framework, $1/\tau^*$ is taken to be the Lipschitz constant of $\nabla Q_n(\boldsymbol{\beta})$, $Q_n = (1/n)\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$, such that for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$,

$$\|\nabla Q_n(\boldsymbol{\beta}_1) - \nabla Q_n(\boldsymbol{\beta}_2)\|_2 \leq \frac{1}{\tau^*}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

Usually, $\tau = (2/n)\lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})$. In addition, by the definition of proximal mapping,

$$\text{prox}_{\tau\lambda_n\|\boldsymbol{\beta}\|_1}(\boldsymbol{\beta}^k - \frac{2\tau}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}^k - \boldsymbol{y})) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{\beta} - \left[\boldsymbol{\beta}^k - \frac{2\tau}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}^k - \boldsymbol{y})\right]\right\|_2^2 + \tau\lambda_n\|\boldsymbol{\beta}\|_1.$$

By simple calculation,

$$\boldsymbol{\beta}^{k+1} = \left[\left|\boldsymbol{\beta}^k - \frac{2\tau}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}^k - \boldsymbol{y})\right| - \tau\lambda_n\mathbf{1}\right]_+ \circ \text{sign}\left(\left[\boldsymbol{\beta}^k - \frac{2\tau}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}^k - \boldsymbol{y})\right]\right).$$

Finally, the PGA is terminated, when either the sequence $\{\boldsymbol{\beta}^k\}$ meets the criterion

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

or the maximum number of iterations is reached.

5. Numerical Simulations

In this section, we carry out Monte Carlo simulations to study the finite-sample performance of the NALE, with the weight calculated by using the SCAD penalty. Further, we compare the NALE with the square-root/scaled lasso (SLE) [17], the natural lasso (NLE) [13], the organic lasso (OLE) [13] and the ridge-based estimator (RBE) [20]. We, also, include the oracle estimator (OE) $(1/n)\|\boldsymbol{\varepsilon}\|_2^2$, as a benchmark in the comparisons. All numerical computation was done using Matlab. The programs are available upon request, from the first author of this paper or Supplementary Materials.

5.1. Simulation Settings

Following [23], throughout the simulations we use the sample size $n = 100$ and parameter dimension $p = 400$. Further, each row of the design matrix $\boldsymbol{X}$ is generated from the multivariate normal distribution $N(0, \boldsymbol{\Sigma})$, with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho \in (0, 1)$. The sparsity of $\boldsymbol{\beta}^*$ is set to be the largest integer less than or equal to n^α, and the locations of the nonzero elements in $\boldsymbol{\beta}^*$ are determined randomly. We consider various parameter values, $\rho \in \{0.1, 0.3, 0.5\}$, $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\sigma^2 \in \{0.5, 1\}$, and use the following true regression parameter vectors

$$\begin{aligned}
&\alpha = 0.1,\ \boldsymbol{\beta}^* = (1.2, -0.8, 0, \ldots, 0)^T;\\
&\alpha = 0.2,\ \boldsymbol{\beta}^* = (1.2, -0.8, 1, 0, \ldots, 0)^T;\\
&\alpha = 0.3,\ \boldsymbol{\beta}^* = (1.2, -0.8, 1, -0.6, 0, \ldots, 0)^T;\\
&\alpha = 0.4,\ \boldsymbol{\beta}^* = (1.2, -0.8, 1, -0.6, 0.8, -0.9, 1.2, 0, \ldots, 0)^T;\\
&\alpha = 0.5,\ \boldsymbol{\beta}^* = (1.2, -0.8, 1, -0.6, 0.8, -0.9, 1.2, 0.4, 0.9, -1.1, 0, \ldots, 0)^T.
\end{aligned}$$

We have, also, considered other variance settings, such as $\sigma^2 \in \{3, 5\}$, however, the simulation results are similar to that of the above settings and, therefore, are not included. To assess the performances of the estimators, we calculate the average mean squared error (MSE) $\hat{E}\{(\sigma^{-1}\widehat{\sigma} - 1)^2\}$ and the average relative error (RE) $\hat{E}(\sigma^{-1}\widehat{\sigma})$, based on 100 Monte Carlo runs.

5.2. Selection of Tuning Parameters

Usually, five-fold cross-validation can be used, to select tuning parameters for each estimation, which is fairly expensive. In order to reduce the computational cost, we consider the following methods, with a fixed choice of tuning parameters for all estimators, except for the NLE and NALE.

For the SLE, we consider three penalty levels $\lambda_{n,i} = \sqrt{2^{i-1}(\log p)/n}$, $i = 1, 2, 3$, which is similar to Example 1 in [17]. Then, the best estimator is selected as the final SLE estimator. Indeed, Ref. [17] found that $\lambda_{n,2}$ works very well for SLE. By the simulation results of [13], the OLE with $\lambda_{n,1} = \log(p)/n$ and $\lambda_{n,2} = E(n^{-2}\|\boldsymbol{X}^T\boldsymbol{\epsilon}\|_\infty^2)$ performed very well, where $\boldsymbol{\epsilon} \in N(0, \boldsymbol{I}_n)$. From [20,24], the tuning parameter used in RBE is calculated by setting $\eta = \alpha \max_{1 \le i \le p} |\boldsymbol{X}_i^T \boldsymbol{y}|/(np)$ with $\alpha = 0.1$.

5.3. Simulation Results

In each simulation, 100 runs are carried out to calculate the average of the performance measures. The results are presented in Tables 1–4. These results show that, overall, both the MSE and RE of the NALE are very close to that of the OE, and are remarkably better than the other estimators, in most of the cases. However, in a few cases, such as $\rho = 0.3$ and $\alpha = 0.5$ with both $\sigma^2 = 0.5$ and $\sigma^2 = 1$, the NALE has a slightly larger MSE than the NLE, although it has smaller RE than the latter. As expected, the NLE often overestimates the true value, due to the bias and over-selection of the lasso. Moreover, in the cases where the NLE has a relatively large MSE, the NALE tends to have a large MSE as well, indicating that the poor performance of the NLE will impact the performance of the NALE, since it is used as the initial estimator. Finally, Ref. [20] reported that the RBE performs well in the

cases with relatively small p and weak signals, however, it performs poorly and is, even, ineffective in the settings of our simulations.

We, further, summarize the performances of various methods using boxplots, in Figures 1 and 2. As one can easily see, the NALE is accurate and stable in all cases, while the OLE is less accurate, although it is, still, fairly stable. Further, the NALE performs well in extremely sparse scenarios. Another interesting point is that the NALE inherits the variable selection and parameter estimation of the adaptive lasso. Although we focus on the variance estimation in this work, the method performs well in estimating the regression coefficients as well.

Table 1. Average RE of various estimators, true $\sigma^2 = 0.5$.

α	OE	NALE	$\text{SLE}_{\lambda_{n,1}}$	$\text{SLE}_{\lambda_{n,2}}$	$\text{SLE}_{\lambda_{n,3}}$	NLE	$\text{OLE}_{\lambda_{n,1}}$	$\text{OLE}_{\lambda_{n,2}}$	RBE
$(\rho = 0.1)$									
0.1	0.004	0.004	0.692	0.081	0.013	0.052	0.098	0.152	1.034
0.2	0.005	0.005	0.723	0.080	0.011	0.115	0.283	0.097	1.799
0.3	0.006	0.006	0.758	0.076	0.010	0.170	0.474	0.067	2.043
0.4	0.005	0.005	0.817	0.059	0.018	0.503	1.998	0.001	3.758
0.5	0.005	0.006	0.730	0.022	0.164	0.830	3.405	0.044	5.769
$(\rho = 0.3)$									
0.1	0.005	0.005	0.621	0.069	0.008	0.051	0.086	0.150	0.655
0.2	0.005	0.004	0.642	0.053	0.005	0.111	0.242	0.098	0.765
0.3	0.004	0.004	0.585	0.043	0.014	0.161	0.383	0.067	1.249
0.4	0.006	0.006	0.598	0.024	0.081	0.483	1.616	0.002	3.431
0.5	0.006	0.006	0.721	0.015	0.536	0.790	2.848	0.003	4.896
$(\rho = 0.5)$									
0.1	0.004	0.005	0.458	0.056	0.008	0.049	0.079	0.147	0.485
0.2	0.005	0.013	0.422	0.035	0.015	0.099	0.184	0.099	1.214
0.3	0.005	0.005	0.425	0.023	0.043	0.142	0.283	0.070	0.878
0.4	0.005	0.004	0.392	0.013	0.219	0.438	1.228	0.003	3.062
0.5	0.017	0.018	0.810	0.066	1.703	2.854	6.952	0.040	15.446

Table 2. Average RE of various estimators, true $\sigma^2 = 0.5$.

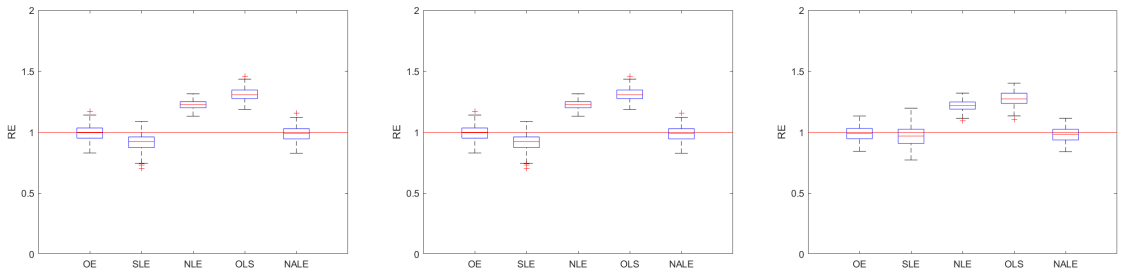
α	OE	NALE	$\text{SLE}_{\lambda_{n,1}}$	$\text{SLE}_{\lambda_{n,2}}$	$\text{SLE}_{\lambda_{n,3}}$	NLE	$\text{OLE}_{\lambda_{n,1}}$	$\text{OLE}_{\lambda_{n,2}}$	RBE
$(\rho = 0.1)$									
0.1	0.993	0.988	0.175	0.727	0.914	1.225	1.309	0.612	2.013
0.2	0.983	0.979	0.156	0.737	0.953	1.336	1.528	0.690	2.337
0.3	0.992	0.992	0.133	0.743	0.982	1.410	1.685	0.745	2.426
0.4	1.003	0.994	0.100	0.775	1.094	1.708	2.412	0.998	2.937
0.5	1.003	0.961	0.152	0.900	1.388	1.910	2.844	1.206	3.400
$(\rho = 0.3)$									
0.1	1.007	0.998	0.219	0.754	0.963	1.222	1.287	0.614	1.803
0.2	0.996	0.988	0.205	0.783	1.015	1.330	1.487	0.689	1.870
0.3	0.998	0.992	0.242	0.815	1.071	1.399	1.615	0.745	2.113
0.4	1.000	0.993	0.239	0.887	1.259	1.694	2.269	0.987	2.850
0.5	1.001	1.017	0.159	0.976	1.717	1.999	2.686	1.182	3.211
$(\rho = 0.5)$									
0.1	0.990	0.985	0.331	0.779	0.965	1.217	1.274	0.619	1.688
0.2	0.994	1.043	0.359	0.845	1.077	1.311	1.423	0.688	2.097
0.3	1.000	1.001	0.357	0.881	1.183	1.374	0.528	0.737	1.932
0.4	0.999	1.014	0.384	1.029	1.453	1.660	2.105	0.966	2.747
0.5	1.011	1.025	0.297	1.058	1.503	1.638	1.901	0.900	2.251

Table 3. Average MSE of various estimators, true $\sigma^2 = 1$.

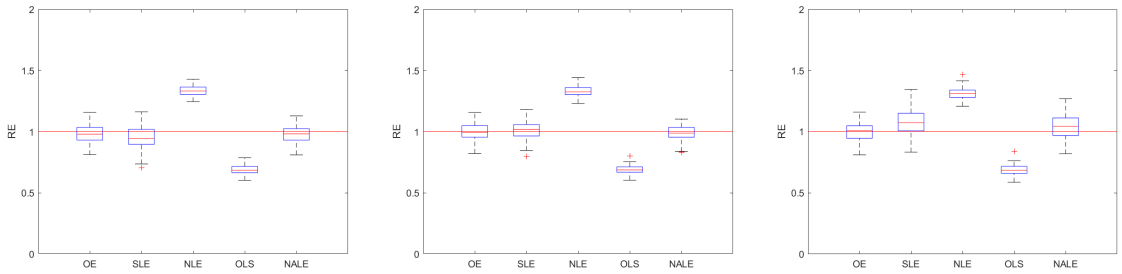
α	OE	NALE	$SLE_{\lambda_{n,1}}$	$SLE_{\lambda_{n,2}}$	$SLE_{\lambda_{n,3}}$	NLE	$OLE_{\lambda_{n,1}}$	$OLE_{\lambda_{n,2}}$	RBE
$(\rho = 0.1)$									
0.1	0.004	0.004	0.740	0.090	0.013	0.019	0.025	0.200	0.347
0.2	0.005	0.005	0.731	0.074	0.006	0.005	0.083	0.155	0.455
0.3	0.005	0.005	0.759	0.081	0.008	0.074	0.161	0.127	0.600
0.4	0.004	0.008	0.748	0.043	0.038	0.043	0.592	0.041	1.479
0.5	0.005	0.009	0.834	0.028	0.223	0.091	1.021	0.009	1.934
$(\rho = 0.3)$									
0.1	0.005	0.005	0.655	0.087	0.015	0.018	0.023	0.201	0.266
0.2	0.005	0.006	0.642	0.063	0.008	0.045	0.074	0.153	0.457
0.3	0.005	0.005	0.597	0.048	0.012	0.065	0.116	0.131	0.478
0.4	0.005	0.015	0.621	0.015	0.128	0.192	0.404	0.049	1.101
0.5	0.004	0.016	0.667	0.026	0.356	0.349	0.856	0.010	1.811
$(\rho = 0.5)$									
0.1	0.004	0.005	0.476	0.054	0.007	0.016	0.017	0.193	0.141
0.2	0.005	0.007	0.414	0.028	0.012	0.036	0.044	0.156	0.275
0.3	0.006	0.009	0.357	0.017	0.030	0.049	0.064	0.134	0.345
0.4	0.005	0.004	0.398	0.015	0.108	0.169	0.325	0.055	0.868
0.5	0.004	0.004	0.490	0.023	0.237	0.306	0.722	0.016	1.328

Table 4. Average RE of various estimators, true $\sigma^2 = 1$.

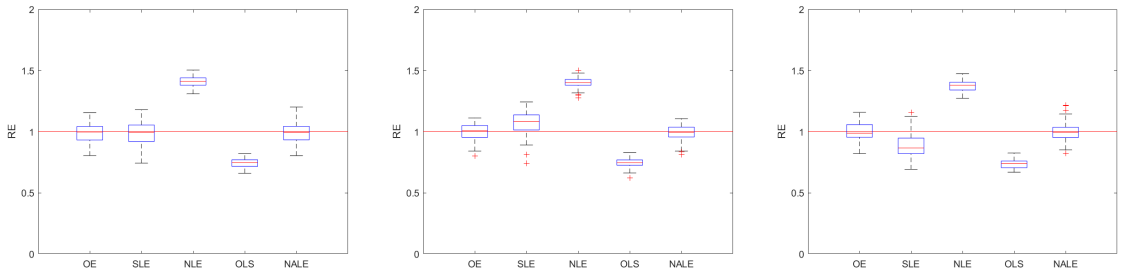
α	OE	NALE	$SLE_{\lambda_{n,1}}$	$SLE_{\lambda_{n,2}}$	$SLE_{\lambda_{n,3}}$	NLE	$OLE_{\lambda_{n,1}}$	$OLE_{\lambda_{n,2}}$	RBE
$(\rho = 0.1)$									
0.1	1.002	0.977	0.144	0.711	0.914	1.131	1.148	0.555	1.581
0.2	0.999	0.964	0.149	0.742	0.969	1.211	1.279	0.608	1.667
0.3	1.003	0.964	0.133	0.726	0.963	1.268	1.394	0.636	1.767
0.4	1.003	0.940	0.141	0.818	1.176	1.205	1.766	0.801	2.212
0.5	0.981	0.937	0.091	0.891	1.459	1.300	2.007	0.913	2.387
$(\rho = 0.3)$									
0.1	0.994	0.971	0.197	0.720	0.913	1.125	1.137	0.553	1.506
0.2	1.005	0.987	0.207	0.767	0.997	1.207	1.263	0.610	1.669
0.3	1.007	1.000	0.234	0.804	1.054	1.251	1.331	0.640	1.684
0.4	1.001	1.068	0.219	0.982	1.342	1.436	1.632	0.783	2.048
0.5	0.993	1.068	0.193	1.111	1.584	1.589	1.922	0.912	2.341
$(\rho = 0.5)$									
0.1	1.001	0.980	0.316	0.783	0.977	1.118	1.113	0.562	1.364
0.2	1.000	1.016	0.362	0.862	1.065	1.183	1.200	0.607	1.514
0.3	0.987	1.028	0.412	0.911	1.146	1.215	1.242	0.636	1.580
0.4	0.990	0.996	0.381	1.026	1.310	1.408	1.566	0.771	1.926
0.5	1.000	1.011	0.310	1.100	1.473	1.551	1.846	0.882	2.148



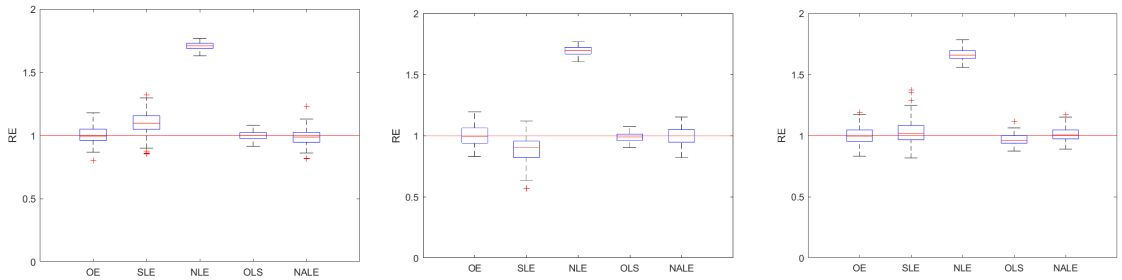
$\rho = 0.1, \alpha = 0.1$ $\rho = 0.3, \alpha = 0.1$ $\rho = 0.5, \alpha = 0.1$



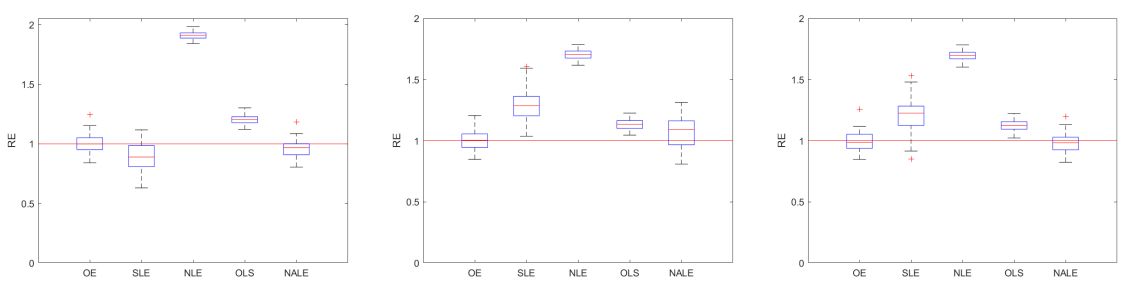
$\rho = 0.1, \alpha = 0.2$ $\rho = 0.3, \alpha = 0.2$ $\rho = 0.5, \alpha = 0.2$



$\rho = 0.1, \alpha = 0.3$ $\rho = 0.3, \alpha = 0.3$ $\rho = 0.5, \alpha = 0.3$

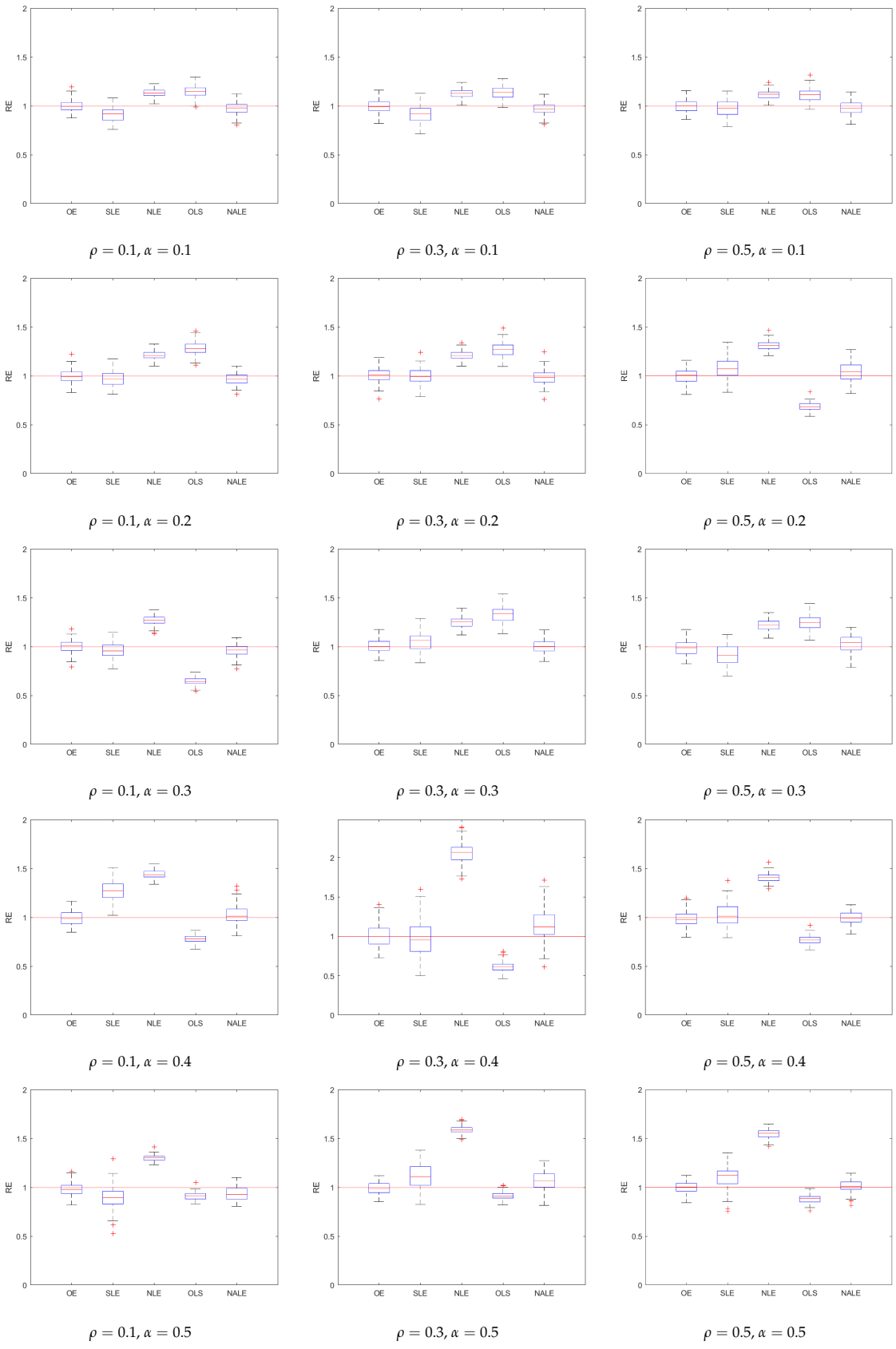


$\rho = 0.1, \alpha = 0.4$ $\rho = 0.3, \alpha = 0.4$ $\rho = 0.5, \alpha = 0.4$



$\rho = 0.1, \alpha = 0.5$ $\rho = 0.3, \alpha = 0.5$ $\rho = 0.5, \alpha = 0.5$

Figure 1. Boxplots of 100 RE values for five estimators, true $\sigma^2 = 0.5$.



$\rho = 0.1, \alpha = 0.1$ $\rho = 0.3, \alpha = 0.1$ $\rho = 0.5, \alpha = 0.1$

$\rho = 0.1, \alpha = 0.2$ $\rho = 0.3, \alpha = 0.2$ $\rho = 0.5, \alpha = 0.2$

$\rho = 0.1, \alpha = 0.3$ $\rho = 0.3, \alpha = 0.3$ $\rho = 0.5, \alpha = 0.3$

$\rho = 0.1, \alpha = 0.4$ $\rho = 0.3, \alpha = 0.4$ $\rho = 0.5, \alpha = 0.4$

$\rho = 0.1, \alpha = 0.5$ $\rho = 0.3, \alpha = 0.5$ $\rho = 0.5, \alpha = 0.5$

Figure 2. Boxplots of 100 RE values for five estimators, true $\sigma^2 = 1$.

6. Conclusions and Discussion

We proposed a novel approach for variance estimation that combines the reparameterized log-likelihood function and adaptive-lasso penalization. We have established the asymptotic properties of the NALE. The theory in this paper shows that the NALE converges at a faster rate than some other existing estimators, including the NLE, OLE and SLE. In addition, the NAL is closely related to the adaptive lasso, which makes its numerical calculation straightforward. We have used the PGA to obtain the NALE in numerical simulations. Our simulation results show that the NALE performs well and favorably against other existing methods, in most finite sample situations, especially in extremely sparse scenarios. However, the quality of the NALE depends on that of the initial estimator used in its numerical optimization, and the poor performance of the initial estimator may result in the poor performance of the NALE.

7. Regularity Conditions and Proofs

This section provides theoretical proofs. We first state the following regularity conditions.

Condition 1. *With probability approaching one, the initial estimator satisfies* $\|\widetilde{\boldsymbol{\beta}}^{\text{ini}} - \boldsymbol{\beta}^*\|_2 \leq C_2\sqrt{s(\log p)/n}$.

Condition 2. $p'_\lambda(t)$ *is non-increasing in* $t \in (0,\infty)$ *and is Lipschitz with constant* C_3, *that is,*

$$|p'_{\lambda_n}(|t_1|) - p'_{\lambda_n}(|t_2|)| \leq C_3|t_1 - t_2|$$

for any $t_1, t_2 \in \mathbb{R}$. *Moreover,* $p'_{\lambda_n}(C_2\sqrt{s\log p/n}) > (1/2)p'_{\lambda_n}(0+)$ *for sufficiently large* n, *where* C_2 *is defined in Condition 1.*

Condition 3. *There exist positive constants* $0 < c_{\min} < c_{\max} < \infty$, *such that*

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

and

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

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

Condition 4. *The true coefficients satisfy* $\min_{i\in\mathcal{A}_0}|\boldsymbol{\beta}_i^*| \gg \sqrt{(s(2\log p + 2L))/n}$. *Moreover, it holds* $p''_{\lambda_n}(t) = o(s^{-1}\lambda_n^{-1}(2\log p + 2L)^{-1/2})$ *for any* $t > \min_{i\in\mathcal{A}_0}|\boldsymbol{\beta}_i^*|/2$ *and* $L > 0$.

As we pointed out in Remark 1, the lasso estimator $\boldsymbol{\beta}_{\text{lasso}}$ satisfies Condition 1. Condition 2 affects the bound between $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ and $\boldsymbol{\beta}^*$ and is used in the proof of Theorem 1. Further, it determines the bound between $\widehat{\sigma}^2_{\lambda_n}$ and σ^2_{oracle}. The first part of Condition 3 is a very common regularity condition (see [4,12,23]) in high-dimensional regression. The remaining part is similar to Condition 3 in [23], which is used in the proofs of Theorems 1 and 2. Condition 4 is needed in the analysis of Corollary 1.

Proof of Proposition 1. (i) Since $(\widehat{\boldsymbol{\theta}}_{\lambda_n}, \widehat{\phi}_{\lambda_n})$ is a solution of (2), $\widehat{\boldsymbol{\theta}}_{\lambda_n}$ is a solution of the problem

$$\min_{\boldsymbol{\theta}_{\lambda_n}\in\mathbb{R}^p} L(\boldsymbol{\theta}, \widehat{\phi}_{\lambda_n}) + \lambda_n\|\boldsymbol{w}\circ\boldsymbol{\theta}\|_1.$$

Hence, by optimality of the above problem, we have

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

where $\widehat{\boldsymbol{g}} \in \partial(\|\widehat{\boldsymbol{\theta}}_{\lambda_n}\|_1)$. It follows that

$$-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n} + n\lambda_n \boldsymbol{w} \circ \widehat{\boldsymbol{g}} = 0.$$

Since $\operatorname{sign}(\widehat{\boldsymbol{\theta}}_{\lambda_n}) = \operatorname{sign}(\widehat{\boldsymbol{\beta}}_{\lambda_n})$, we have $\partial(\|\widehat{\boldsymbol{\theta}}_{\lambda_n}\|_1) = \partial(\|\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_1)$, which, further, implies that $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ is a solution of the adaptive lasso (4).

(ii) Since $(\widehat{\boldsymbol{\theta}}_{\lambda_n}, \widehat{\phi}_{\lambda_n})$ is a solution of (2), by optimality of problem (2), we have

$$-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\frac{\widehat{\boldsymbol{\theta}}_{\lambda_n}}{\widehat{\phi}_{\lambda_n}} + n\lambda \boldsymbol{w} \circ \widehat{\boldsymbol{g}} = 0, \ -\frac{1}{\widehat{\phi}_{\lambda_n}} + \frac{1}{n}\|\boldsymbol{y}\|_2^2 - \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_{\lambda_n}\|_2^2}{n\widehat{\phi}_{\lambda_n}^2} = 0,$$

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

$$-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n} + n\lambda_n \boldsymbol{w} \circ \widehat{\boldsymbol{g}} = 0, \ -\frac{1}{\widehat{\phi}_{\lambda_n}} + \frac{1}{n}\|\boldsymbol{y}\|_2^2 - \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\theta}}_{\lambda_n}\|_2^2}{n\widehat{\phi}_{\lambda_n}^2} = 0. \tag{7}$$

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

$$\widehat{\boldsymbol{\beta}}_{\lambda_n}^T(\boldsymbol{w} \circ \widehat{\boldsymbol{g}}) = \sum_{i=1}^{p} w_i\widehat{\beta}_i\widehat{g}_i = \sum_{i=1}^{p} |w_i\widehat{\beta}_i| = \|\boldsymbol{w} \circ \widehat{\boldsymbol{\beta}}\|_1. \tag{8}$$

Combining (7) and (8), we obtain

$$0 = -\widehat{\boldsymbol{\beta}}_{\lambda_n}^T\boldsymbol{X}^T\boldsymbol{y} + \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2 + \lambda_n\|\boldsymbol{w} \circ \widehat{\boldsymbol{\beta}}_{\lambda_n}\|_1, \ \widehat{\sigma}_{\lambda_n}^2 = \frac{1}{n}\left(\|\boldsymbol{y}\|_2^2 - \|\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2\right).$$

Further, by the first term in (7),

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

Combining the above equality and the second term in (7), we have

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

which implies that $\widehat{\sigma}_{\lambda_n}^2$ is the optimal value of the adaptive lasso (4). □

Proof of Lemma 1. From Proposition 1, we have

$$\widehat{\sigma}_{\lambda_n}^2 \leq \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + 2\lambda_n\|\boldsymbol{w} \circ \boldsymbol{\beta}^*\|_1 = \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 + 2\lambda_n\|\boldsymbol{w} \circ \boldsymbol{\beta}^*\|_1. \tag{9}$$

Since the loss function in the adaptive lasso is convex, we have

$$\begin{aligned} \widehat{\sigma}_{\lambda_n}^2 &= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}\|_2^2 + 2\lambda_n\|\boldsymbol{w} \circ \widehat{\boldsymbol{\beta}}_{\lambda_n}\|_1 \\ &\geq \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*\|_2^2 + [\frac{2}{n}\boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{\beta}^* - \boldsymbol{y})]^T(\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*) \\ &= \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 - \frac{2}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*) \\ &\geq \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 - 2\|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\|_\infty\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_1 \\ &\geq \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2 - 2\lambda_n\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_1. \end{aligned} \tag{10}$$

Combining inequalities (9) and (10), we obtain

$$|\widehat{\sigma}^2_{\lambda_n} - \frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2| \leq \max\{2\lambda_n\|\boldsymbol{w}\circ\boldsymbol{\beta}^*\|_1, 2\|\frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\|_\infty\|\widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\|_1\},$$

which completes the proof. □

Proof of Theorem 1. Since problem (4) is a convex optimization, by Theorem 1 of [25], it suffices to show that, with probability tending to 1, there exists a minimizer $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ of problem (4) that satisfies

$$\boldsymbol{X}^T_{\mathcal{A}_0}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n}) - n\lambda_n\boldsymbol{w}_{\mathcal{A}_0}\circ\widehat{\boldsymbol{g}}_{\mathcal{A}_0} = \boldsymbol{0}, \tag{11}$$

$$\|\boldsymbol{X}^T_{\mathcal{A}_0^c}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n})\|_\infty < n\lambda_n\boldsymbol{w}_{\mathcal{A}_0^c}, \tag{12}$$

$$\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{X}_{\mathcal{A}_0}\right) \geq c_{\min}. \tag{13}$$

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

Let $\boldsymbol{\xi}_1 = \boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{\varepsilon}$ and $\boldsymbol{\xi}_2 = \boldsymbol{X}^T_{\mathcal{A}_0^c}\boldsymbol{\varepsilon}$. Since $\|\boldsymbol{X}_j\|_2^2 = n$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$, it follows from Corollary 4.3 in [26] that, for any $L > 0$,

$$P\left\{\frac{\|\boldsymbol{X}^T\boldsymbol{\varepsilon}\|_\infty}{n\sigma} > \sqrt{\frac{2\log p + 2L}{n}}\right\} \leq e^{-L}. \tag{14}$$

Now we show that there exists a minimizer $\widehat{\boldsymbol{\beta}}$ of problem (4) satisfies conditions (11)–(13).

Equation (11): Consider the minimizer of problem (4) in the subspace $\{\boldsymbol{\beta} = (\boldsymbol{\beta}^T_{\mathcal{A}_0}, \boldsymbol{\beta}^T_{\mathcal{A}_0^c})^T : \boldsymbol{\beta}_{\mathcal{A}_0^c} = \boldsymbol{0}\}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^T_{\mathcal{A}_0}, \boldsymbol{0}^T)^T$, where $\boldsymbol{\beta}_{\mathcal{A}_0} = \boldsymbol{\beta}^*_{\mathcal{A}_0} + \widetilde{a}_n\boldsymbol{v}_{\mathcal{A}_0} \in \mathbb{R}^s$ with $\widetilde{a}_n = \sqrt{s(2\log p + 2L)/n} + 2\lambda_n(\|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2 + C_2C_3\sqrt{s(\log p)/n})$, $\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 = C$, and $C > 0$ is some large enough constant. Note that

$$L_n(\boldsymbol{\beta}^*_{\mathcal{A}_0} + \widetilde{a}_n\boldsymbol{v}_{\mathcal{A}_0}, \boldsymbol{0}) - L_n(\boldsymbol{\beta}^*_{\mathcal{A}_0}, \boldsymbol{0}) = I_1(\boldsymbol{v}_{\mathcal{A}_0}) + I_2(\boldsymbol{v}_{\mathcal{A}_0}), \tag{15}$$

where $I_1(\boldsymbol{v}_{\mathcal{A}_0}) = \frac{1}{n}\|\boldsymbol{X}(\boldsymbol{\beta}^* + \widetilde{a}_n\boldsymbol{v}) - \boldsymbol{y}\|_2^2 - \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{y}\|_2^2$, $I_2(\boldsymbol{v}_{\mathcal{A}_0}) = 2\lambda_n\|\boldsymbol{w}_{\mathcal{A}_0}\circ(\boldsymbol{\beta}^*_{\mathcal{A}_0} + \widetilde{a}_n\boldsymbol{v}_{\mathcal{A}_0})\|_1 - 2\lambda_n\|\boldsymbol{w}_{\mathcal{A}_0}\circ\boldsymbol{\beta}^*_{\mathcal{A}_0}\|_1$. For $I_1(\boldsymbol{v}_{\mathcal{A}_0})$, by (14), we have

$$\begin{aligned} I_1(\boldsymbol{v}_{\mathcal{A}_0}) &= \frac{1}{n}\|\boldsymbol{X}(\boldsymbol{\beta}^* + \widetilde{a}_n\boldsymbol{v}) - \boldsymbol{y}\|_2^2 - \frac{1}{n}\|\boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{y}\|_2^2 \\ &= \frac{1}{n}\widetilde{a}_n^2\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v} + \frac{2}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\widetilde{a}_n\boldsymbol{v} \\ &= \frac{1}{n}\widetilde{a}_n^2\boldsymbol{v}^T_{\mathcal{A}_0}\boldsymbol{X}^T_{\mathcal{A}_0}\boldsymbol{X}_{\mathcal{A}_0}\boldsymbol{v}_{\mathcal{A}_0} + \frac{2}{n}\widetilde{a}_n\boldsymbol{\varepsilon}^T\boldsymbol{X}_{\mathcal{A}_0}\boldsymbol{v}_{\mathcal{A}_0} \\ &\geq c_{\min}\widetilde{a}_n^2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2\widetilde{a}_n\left\|\frac{\boldsymbol{\xi}_1}{n}\right\|_2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2 \\ &\geq c_{\min}\widetilde{a}_n^2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2\sigma\widetilde{a}_n\sqrt{s(2\log p + 2L)/n}\|\boldsymbol{v}_{\mathcal{A}_0}\|_2, \end{aligned} \tag{16}$$

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

$$\begin{aligned} I_2(\boldsymbol{v}_{\mathcal{A}_0}) &= 2\lambda_n\|\boldsymbol{w}_{\mathcal{A}_0}\circ(\boldsymbol{\beta}^*_{\mathcal{A}_0} + \widetilde{a}_n\boldsymbol{v}_{\mathcal{A}_0})\|_1 - 2\lambda_n\|\boldsymbol{w}_{\mathcal{A}_0}\circ\boldsymbol{\beta}^*_{\mathcal{A}_0}\|_1 \\ &\leq 2\lambda_n\|\boldsymbol{w}_{\mathcal{A}_0}\circ(\widetilde{a}_n\boldsymbol{v}_{\mathcal{A}_0})\|_1 \leq 2\widetilde{a}_n\lambda_n\|\boldsymbol{w}_{\mathcal{A}_0}\|_2\|\boldsymbol{v}_{\mathcal{A}_0}\|_2. \end{aligned} \tag{17}$$

By the two-steps procedure of weight vector and Condition 2, it holds that

$$\begin{aligned} \|\boldsymbol{w}_{\mathcal{A}_0}\|_2 &\leq \|\boldsymbol{w}_{\mathcal{A}_0} - \boldsymbol{w}^*_{\mathcal{A}_0}\|_2 + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2 \leq C_3\|\widetilde{\boldsymbol{\beta}}^{\text{ini}}_{\mathcal{A}_0} - \boldsymbol{\beta}^*_{\mathcal{A}_0}\|_2 + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2 \\ &\leq C_2C_3\sqrt{s(\log p)/n} + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2. \end{aligned} \tag{18}$$

Hence, combining (15)–(18) yields

$$\begin{aligned} L_n(\boldsymbol{\beta}^*_{\mathcal{A}_0} + \widetilde{a}_n \boldsymbol{v}_{\mathcal{A}_0}, \mathbf{0}) - L_n(\boldsymbol{\beta}^*_{\mathcal{A}_0}, \mathbf{0}) \quad &\geq \quad c_{\min} \widetilde{a}_n^2 \|\boldsymbol{v}_{\mathcal{A}_0}\|_2^2 - 2\sigma \widetilde{a}_n \sqrt{s(2\log p + 2L)/n} \|\boldsymbol{v}_{\mathcal{A}_0}\|_2 \\ &\quad -2\widetilde{a}_n \lambda_n (C_2 C_3 \sqrt{s(\log p)/n} + \|\boldsymbol{w}^*_{\mathcal{A}_0}\|_2) \|\boldsymbol{v}_{\mathcal{A}_0}\|_2. \end{aligned}$$

Taking a large enough C, we have obtained, with probability tending to one,

$$L_n(\boldsymbol{\beta}^*_{\mathcal{A}_0} + \widetilde{a}_n \boldsymbol{v}_{\mathcal{A}_0}, \mathbf{0}) - L_n(\boldsymbol{\beta}^*_{\mathcal{A}_0}, \mathbf{0}) > 0.$$

It follows, immediately, that, with probability approaching one, there exists a minimizer $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}$ of problem (4), subject to subspace $\{\boldsymbol{\beta} = (\boldsymbol{\beta}^T_{\mathcal{A}_0}, \boldsymbol{\beta}^T_{\mathcal{A}_0^c})^T : \boldsymbol{\beta}_{\mathcal{A}_0^c} = \mathbf{0}\}$, such that $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0} - \boldsymbol{\beta}^*_{\mathcal{A}_0}\|_2 \leq C_4 \widetilde{a}_n \equiv a_n$, with some constant $C_4 > 0$. Therefore, equality (11) holds, by the optimality theory.

Inequality (12): It remains to be proven that with asymptotic probability 1, (12) holds. Then, by optimality theory, $\widehat{\boldsymbol{\beta}}_{\lambda_n} = (\widehat{\boldsymbol{\beta}}^T_{\mathcal{A}_0}, \mathbf{0}^T)^T$ is the unique global minimizer of problem (4).

By triangle inequality, we have

$$\|\boldsymbol{X}^T_{\mathcal{A}_0^c}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|_\infty \leq \|\boldsymbol{X}^T_{\mathcal{A}_0^c}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty + \|\boldsymbol{X}^T_{\mathcal{A}_0^c}\boldsymbol{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_\infty. \tag{19}$$

Further, by Condition 1, we have $|\widetilde{\boldsymbol{\beta}}^{\text{ini}}_i| \leq C_2\sqrt{s(\log p)/n}$ with probability approaching one, where $i \in \mathcal{A}_0^c$. Moreover, by the definition of the fold-concave penalty function,

$$p'_{\lambda_n}(|\widetilde{\boldsymbol{\beta}}^{\text{ini}}_i|) \geq p'_{\lambda_n}(C_2\sqrt{s(\log p)/n}). \tag{20}$$

Therefore, by Condition 2 and inequality (20), we conclude that

$$\|\boldsymbol{w}^{-1}_{\mathcal{A}_0^c}\|_\infty = \min_{i \in \mathcal{A}_0^c} p'_{\lambda_n}(|\widetilde{\boldsymbol{\beta}}^{\text{ini}}_i|)^{-1} < \frac{2}{p'_{\lambda_n}(0+)} = 2\|(\boldsymbol{w}^*_{\mathcal{A}_0^c})^{-1}\|_\infty. \tag{21}$$

Thus, for the first term of the right hand of inequality (19), by (14) and the condition that $\min_{i \in \mathcal{A}_0^c}\{w^*_i\} > C_1^{-1}$, with probability approaching one,

$$\begin{aligned} \frac{1}{n}\|\boldsymbol{X}^T_{\mathcal{A}_0^c}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^*)\|_\infty \quad &= \quad \frac{1}{n}\|\boldsymbol{X}^T_{\mathcal{A}_0^c}\boldsymbol{\varepsilon}\|_\infty \\ &< \quad \sigma\sqrt{\frac{2\log p + 2L}{n}} = \frac{\lambda_n}{4C_1} < \frac{\lambda_n}{4\|(\boldsymbol{w}^*_{\mathcal{A}_0^c})^{-1}\|_\infty} < \frac{\lambda_n}{2\|\boldsymbol{w}^{-1}_{\mathcal{A}_0^c}\|_\infty}. \end{aligned} \tag{22}$$

As for the second term of right hand of inequality (19), by Condition 3, inequality (21) and inequality (14), with probability approaching one, we have

$$\begin{aligned} \frac{1}{n}\|\boldsymbol{X}^T_{\mathcal{A}_0^c}\boldsymbol{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_\infty \quad &\leq \quad \frac{1}{n}\|\boldsymbol{X}^T_{\mathcal{A}_0^c}\boldsymbol{X}_{\mathcal{A}_0}\|_{2,\infty}\|\boldsymbol{\beta}^*_{\mathcal{A}_0} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}\|_2 \\ &\leq \quad \frac{\lambda_n}{4\|(\boldsymbol{w}^*_{\mathcal{A}_0^c})^{-1}\|_\infty} < \frac{\lambda_n}{2\|\boldsymbol{w}^{-1}_{\mathcal{A}_0^c}\|_\infty}. \end{aligned} \tag{23}$$

Combining (19), (22) and (23), we obtain inequality (12).

Inequality (13): it follows from Condition 3 that with an asymptotic probability of one, inequality (13) holds. This completes the proof of Theorem 1. □

Proof of Theorem 2. By equality (11), $(1/n)\boldsymbol{X}_{\mathcal{A}_0}^T(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_n})-\lambda_n\boldsymbol{w}_{\mathcal{A}_0}\circ\widehat{\boldsymbol{g}}_{\mathcal{A}_0}=\boldsymbol{0}$. Since $\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}^*=\boldsymbol{\varepsilon}$, we have

$$\frac{1}{n}\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0})=-\lambda_n\boldsymbol{w}_{\mathcal{A}_0}\circ\widehat{\boldsymbol{g}}_{\mathcal{A}_0}+\frac{1}{n}\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{\varepsilon}.$$

Therefore,

$$n^{1/2}\boldsymbol{\alpha}_n^T(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0})+n^{3/2}\lambda_n\boldsymbol{\alpha}_n^T(\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0})^{-1}\boldsymbol{w}_{\mathcal{A}_0}\circ\widehat{\boldsymbol{g}}_{\mathcal{A}_0}=n^{1/2}\boldsymbol{\alpha}_n^T(\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0})^{-1}\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{\varepsilon}. \quad (24)$$

By the first part of Condition 4 and the bound of $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ in Theoroem 1, we have $\operatorname{sign}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0})=\operatorname{sign}(\boldsymbol{\beta}^*_{\mathcal{A}_0})$. Then, $\widehat{\boldsymbol{g}}_{\mathcal{A}_0}=\boldsymbol{g}^*_{\mathcal{A}_0}$. In addition, by the second part in condition 4,

$$\boldsymbol{w}_{\mathcal{A}_0}=\boldsymbol{w}^*_{\mathcal{A}_0}+\boldsymbol{\zeta}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0}),$$

where $\boldsymbol{\zeta}=\operatorname{diag}(p''_{\lambda_n}(\widetilde{\beta}_1),\ldots,p''_{\lambda_n}(\widetilde{\beta}_s))^T\in\mathbb{R}^{s\times s}$ and $\widetilde{\boldsymbol{\beta}}\in\mathbb{R}^s$ lie on the line segment $[\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0},\boldsymbol{\beta}^*_{\mathcal{A}_0}]$. It follows that $\|\boldsymbol{\zeta}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0})\circ\boldsymbol{g}^*_{\mathcal{A}_0}\|_2=\|\boldsymbol{\zeta}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0})\|_2=o(\lambda_n^{-1}\sqrt{1/n})$. Further, since

$$|n^{3/2}\lambda_n\boldsymbol{\alpha}_n^T(\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0})^{-1}[\boldsymbol{\zeta}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0})\circ\boldsymbol{g}^*_{\mathcal{A}_0}]|\le n^{1/2}\lambda_n c_{\max}\|\boldsymbol{\alpha}_n\|_2\|\boldsymbol{\zeta}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_0}-\boldsymbol{\beta}^*_{\mathcal{A}_0})\|_2=o(1),$$

we have, for the second term of the left hand of (24),

$$n^{3/2}\lambda_n\boldsymbol{\alpha}_n^T(\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0})^{-1}\boldsymbol{w}_{\mathcal{A}_0}\circ\widehat{\boldsymbol{g}}_{\mathcal{A}_0}=n^{3/2}\lambda_n\boldsymbol{\alpha}_n^T(\boldsymbol{X}_{\mathcal{A}_0}^T\boldsymbol{X}_{\mathcal{A}_0})^{-1}\boldsymbol{w}_{\mathcal{A}_0}\circ\boldsymbol{g}^*_{\mathcal{A}_0}+o(1).$$

Finally, the result follows, by verifying the conditions of the Lindeberg–Feller central limit theorem, in the same way as in the proof of Theorem 2 in [4]. □

Proof of Theorem 4. For any $M>1$, take $L=(M-1)\log p$ and denote $Z_n=(\widehat{\sigma}^2_{\lambda_n}-\frac{1}{n}\|\boldsymbol{\varepsilon}\|_2^2)^2$. Then, by Theorems 1 and 3, we have

$$P(Z_n>Mb_n^2)\le e^{-(M-1)\log p}.$$

It follows that

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

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

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

By the proof of Theorem 12 in [13], we have

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

Combining (25) and (26), we obtain

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

The proof is complete. □

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math10111937/s1, The programs in numerical simulations are available in supplementary materials.

Author Contributions: Methodology, X.W., L.K. and L.W.; software, X.W.; writing—original draft, X.W., L.K. and L.W.; validation, X.W., L.K. and L.W. All authors have read and agreed to the published version of the article.

Funding: The National Natural Science Foundation of China (12071022) and the 111 Project of China (B16002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the editor and the three anonymous reviewers, for their helpful comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *73*, 273–282. [CrossRef]
2. Fan, J.; Li, Y. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
3. Zou, H. The adaptive lasso and its oracle properties. *J. R. Stat. Soc. Ser. B* **2006**, *101*, 1418–1429. [CrossRef]
4. Huang, J.; Horowitz, J.L.; Ma, S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Stat.* **2008**, *36*, 587–613. [CrossRef]
5. Zou, H.; Zhang, H.H. On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **2009**, *37*, 1733–1751. [CrossRef]
6. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
7. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351.
8. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 849–911. [CrossRef]
9. Ghaoui, L.E.; Viallon, V.; Rabbani, T. Safe feature elimination in sparse supervised learning. *Pac. J. Optim.* **2012**, *8*, 667–698.
10. Wang, J.; Wonka, P.; Ye, j. Lasso screening rules via dual polytope projection. *J. Mach. Learn. Res.* **2015**, *16*, 1063–1101.
11. Xiang, Z.J.; Wang, Y.; Ramadge, P.J. Safe feature elimination in sparse supervised learning. *IEEE Trans. Pattern Anal.* **2017**, *39*, 1008–1027. [CrossRef] [PubMed]
12. Fan, J.; Guo, S.; Hao, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B* **2012**, *74*, 37–65. [CrossRef] [PubMed]
13. Yu, G.; Bien, J. Estimating the error variance in a high-dimensional linear model. *Biometrika* **2019**, *106*, 533–546. [CrossRef]
14. Zou, H.; Hastie, T.; Tibshirani, R. On the "Degrees of freedom" of the lasso. *Ann. Stat.* **2007**, *35*, 2173–2192. [CrossRef]
15. Wang, L.; Leblanc, A. Second-order nonlinear least squares estimation. *Ann. Inst. Stat. Math.* **2008**, *60*, 883–900. [CrossRef]
16. Stadler, N.; Bühlmann, P. ℓ_1-penalization for mixture regression models. *Test* **2010**, *19*, 209–256. [CrossRef]
17. Sun, T.; Zhang, C.H. Scaled sparse linear regression. *Biometrika* **2012**, *99*, 879–898. [CrossRef]
18. Dicker, L.H. Variance estimation in high-dimensional linear models. *Biometrika* **2014**, *101*, 269–284. [CrossRef]
19. Dicker, L.H.; Erdogdu, M.A. Maximum likelihood for variance estimation in high-dimensional linear models. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Cadiz, Spain, 9–11 May 2016; pp. 159–167.
20. Liu, X.; Zheng, S.; Feng, X. Estimation of error variance via ridge regression. *Biometrika* **2020**, *107*, 481–488. [CrossRef]
21. Zhang, C.H.; Huang, J. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Stat.* **2008**, *36*, 1567–1594. [CrossRef]
22. Bickel, P.J.; Ritov, Y.A.; Tsybakov, A.B. Simultaneous analysis of lasso and dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732. [CrossRef]
23. Fan, J.; Fan, Y.; Barut, E. Adaptive robust variable selection. *Ann. Stat.* **2014**, *42*, 324–351. [CrossRef] [PubMed]
24. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef] [PubMed]
25. Fan, J.; Lv, J. Non-concave penalized likelihood with np-dimensionality. *IEEE Trans. Inform. Theory* **2011**, *57*, 5467–5484. [CrossRef] [PubMed]
26. Giraud, C. *Introduction to High-Dimensional Statistics*, 1st ed.; Chapman and Hall/CRC: New York, NY, USA, 2014.