symmetry

MDPI

Article

Relaxed Adaptive Lasso and Its Asymptotic Results

Rufei Zhang [1,2,3], Tong Zhao [1], Yajun Lu [1,4] and Xieting Xu [1,*]

[1] College of Economics, Hebei GEO University, Shijiazhuang 050031, China; zhangrufei@hgu.edu.cn (R.Z.); wenlei@hgu.edu.cn (T.Z.); luyajun@xiaomi.com; (Y.L.)
[2] Reaserch Center of Nutural Resources Assets, Hebei GEO University, Shijiazhuang 050031, China
[3] Hebei Province Mineral Resources Development and Management and the Transformation and Upgrading of Resources Industry Soft Science Resrarch Base, Shijiazhuang 050031, China
[4] Xiaomi Corporation, Beijing 100089, China
* Correspondence: xuxieting@163.com or liufei@hgu.edu.cn

Abstract: This article introduces a novel two-stage variable selection method to solve the common asymmetry problem between the response variable and its influencing factors. In practical applications, we cannot correctly extract important factors from a large amount of complex and redundant data. However, the proposed method based on the relaxed lasso and the adaptive lasso, namely, the relaxed adaptive lasso, can achieve information symmetry because the variables it selects contain all the important information about the response variables. The goal of this paper is to preserve the relaxed lasso's superior variable selection speed while imposing varying penalties on different coefficients. Additionally, the proposed method enjoys favorable asymptotic properties, that is, consistency with a fast rate of convergence with $O_p(n^{-1})$. The simulation demonstrates that the proper variable recovery, i.e., the number of significant variables selected, and prediction accuracy of the relaxed adaptive lasso in a limited sample is superior to the regular lasso, relaxed lasso and adaptive lasso estimators.

Keywords: variable selection; relaxed lasso; adaptive lasso; consistency

Citation: Zhang, R.; Zhao, T.; Lu, Y.; Xu, X. Relaxed Adaptive Lasso and Its Asymptotic Results. *Symmetry* **2022**, *14*, 1422. https://doi.org/10.3390/sym14071422

Academic Editor: Alexander Zaslavski

Received: 22 June 2022
Accepted: 8 July 2022
Published: 11 July 2022

1. Introduction

Rapid advancements in research and technology have resulted in enormous data in a variety of scientific domains. How to efficiently extract information from complex data and develop an ideal model that relates critical features to response variables has become a challenge for researchers in the data explosion era. Over the past two decades, statisticians have conducted substantial research on the subject of feature selection.

Tibshirani [1] first proposed lasso, a technique for screening high-dimensional variables that improves least squares estimation by including an L_1 penalty component. The penalty parameter of the lasso set some of the coefficients to zero, thus achieving the proposal of coefficient shrinkage and model selection. Lasso sacrifices unbiasedness for minimizing variance and solves the convex optimization problem to find the globally optimal solution. In the rare signal scenario, when the signal strength exceeds a certain level, lasso shows good performance, far outperforming other variable selection methods [2]. However, Meinshausen and Bühlmann [3] discovered a conflict in the lasso model between optimal prediction and consistent variable selection, which is one of the lasso's downsides. Due to the strong sensitivity of the lasso to the presence of correlation and multicollinearity in real data, insignificant noisy variables may be selected for the model. As a result, the noise variables in the model exacerbate the model fitting effect. Fan and Li [4] presented a more adaptive novel approach for maximizing the likelihood penalty function that applies to generalized linear models and other types of models. Moreover, Fan and Li [5] enhanced the preceding approach and stated that as long as the dimensionality of the model is not too large, the penalized likelihood technique can be used to estimate the model's parameters via

the penalty function. To address the issue of inconsistent lasso selection, Zou [6] reported an adaptive lasso estimator,

$$\hat{\beta}^{Alasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j^T \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|, \quad (1)$$

where $\hat{\omega} = 1/|\hat{\beta}|^{\gamma}, \gamma > 0$. The primary reason why adaptive lasso is superior to lasso is that it has an oracle quality that depends on the weight vector value $\hat{\omega}$. Without this quality, adaptive lasso's oracle property would be suboptimal. Fan and Peng [7] claimed that when the dimension p is less than the sample size n, the lasso and adaptive lasso can both be used to accelerate and optimize variable selection. The theory of Donoho and Johnstone can also be used to demonstrate the adaptive lasso's near-minimax optimality [8]. The non-negative garotte [9] is another regularization method. It can be considered as a special case of the adaptive lasso and was proved to have the property of consistent variable screening [10].

Meinshausen [11] defined the relaxed lasso estimator on the set $M \subseteq \{1, \ldots, p\}$, where p is the number of nonzero variables selected into the true model,

$$\hat{\beta}^{Rlasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j^T \{\beta_j \cdot 1_M\} \right\|_2^2 + \phi\lambda \sum_{j=1}^{p} |\beta_j|, \quad (2)$$

where $\lambda \in [0, \infty], \phi \in (0, 1]$, 1_M is an indicator function, $1_M = \begin{cases} 0, & k \in M \\ 1, & k \notin M \end{cases}$, for all $k \in \{1, \ldots, p\}$. Hastie et al. [12] compared the performance of lasso and forward stepwise regression across a range of signal-to-noise ratios (SNRs) and showed that it is extremely competitive in any environment. The relaxation parameter ϕ contributes to relaxed lasso's superior performance. By adjusting the control parameter ϕ appropriately, it can ensure that the sparse solution on the path does not experience excessive shrinkage. This is the primary reason we chose to expand the model using a relaxed lasso. In recent works, numerous studies have demonstrated that the relaxed lasso has excellent performances compared to other methods. Mentch and Zhou [13] showed that in high-dimensional settings, lasso, forward selection and randomized forward selection perform similarly at low SNRs, but for larger SNRs, relaxed lasso performs much better in terms of the relative test error. Bloise et al. [14] suggested that relaxed lasso is able to avoid overfitting by using two separate tuning parameters so as to obtain a more accurate model. Comparing relaxed Lasso to least squares and stepwise regression, He [15] came to the conclusion that relaxed Lasso improves the accuracy of the model by deleting insignificant variables. Kang et al. [16] proposed a new method that combines the relaxed lasso and a generalized multiclass support vector machine to obtain fewer feature variables and higher classification accuracy. Tay et al. [17] combined elastic net regularized regression with a simplified relaxed lasso model and built a prediction matrix to measure model performance, which speeds up the computational efficiency of the model.

We discuss the properties of different variable selection methods in the case of large samples. Consistency and asymptotic normality are two large sample properties of OLS; for consistency, we can assume a weaker overall zero-correlation assumption $Cov(x, \varepsilon) = 0$ and a zero-mean assumption of error $E(\varepsilon) = 0$. Fu and Knight [18] examined the consistency and asymptotic features of bridge estimation in convex and nonconvex scenarios and established that lasso is consistent when certain conditions are met. Thus, another disadvantage of lasso is that variable selection is conditional, which means that it does not work similarly to an oracle estimator. Zhao and Yu [19] claimed that the irrepresentable condition is a necessary and sufficient criterion for lasso to satisfy consistency. However, both adaptive lasso and relaxed lasso have been proved to be consistent without satisfying the strict condition. Zou [6] demonstrated that even with huge quantities of data, adaptive lasso can efficiently filter out the model's sparse solution while retaining oracle features. According to Meinshausen [11], relaxed lasso can still retain a high rate of convergence

with $O_p\left(n^{-1}\right)$ and can lead to consistent variable selection no matter what the asymptotic result is. To combine the advantages of the preceding two models, we propose a relaxed adaptive lasso and demonstrate that it has the same asymptotic properties and excellent convergence rate as the relaxed lasso. The Lars algorithm [20] and an improved algorithm have been shown to solve the relaxed adaptive lasso.

In this paper, we propose a novel variable filtering method named relaxed adaptive lasso, which can effectively address the model selection issue, and we demonstrate the method's asymptotic properties. We prove that the relaxed adaptive lasso estimator can achieve the same rate of convergence as the relaxed lasso, indicating that it can obtain the sparse solution at the optimal rate. The simulation of this study demonstrates that the relaxed adaptive lasso performs well in variable recovery and predictin accuracy. In particular, when sample size n and the number of variables p are varied, the performance of the relaxed adaptive lasso to filter true nonzero variables is superior to that of the lasso, relaxed lasso, and adaptive lasso. As the sample size n increases, the mean square error (MSE) of the model remains the best.

The rest of this article is structured as follows: Section 2 defines relaxed adaptive lasso, describes its computational algorithms, and establishes its asymptotic properties. Then, in Section 3, we compare the performance of the relaxed adaptive lasso to that of the lasso, adaptive lasso, and relaxed lasso using a simulation experiment. Section 4 discusses the application of real-world data. Section 5 makes a conclusion of the proposed method. The Appendixes A–G contains additional information about the proof.

2. Relaxed Adaptive Lasso and Asymptotic Results

2.1. Definition

Recall that adaptive lasso estimation improves the shrinkage force to equalize the coefficients in lasso by applying a weight vector. The adaptive lasso estimator's set of predictor variables $\hat{\beta}^{\lambda,\omega}$ is denoted by $S^{\lambda,\omega}$,

$$S^{\lambda,\omega} = \left\{1 \le k \le p \mid \hat{\beta}_k^{\lambda,\omega} \neq 0\right\}. \tag{3}$$

The solution of the relaxed adaptive lasso is obtained via the adaptive lasso estimator in $S^{\lambda,\omega}$ if and only if in the low-dimension case.

We now consider the linear regression model

$$Y = X^T\beta^* + \varepsilon, \tag{4}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a vector composed of i.i.d. random variables with mean 0 and variance σ^2. $X = \left(X_1, \ldots, X_p\right)$ is an $n \times p$ matrix with a normally distribution $X \sim N(0, \Sigma)$, where X_i is the ith column and Y is an $n \times 1$ vector of response variables. Now, we define relaxed adaptive lasso estimation. The variable selection and shrinkage are controlled by adding two constraints, λ and ϕ, and one weight vector, ω, to the L_1 penalty term. According to the setup of Zou [6], suppose that $\hat{\beta}$ is an $\sqrt{n}$-consistent estimator of β^*.

Definition 1. *Define the relaxed adaptive lasso estimator as* $\hat{\beta}^{\lambda,\omega}$ *denoted by* $S^{\lambda,\omega}$,

$$\hat{\beta}^* = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j^T \left\{\beta_j \cdot 1_{S^{\lambda,\omega}}\right\} \right\|_2^2 + \phi\lambda \sum_{j=1}^{p} \hat{\omega}_j |\beta_j|, \tag{5}$$

where $1_{S^{\lambda,\omega}}$ *is an indicator function* $\left\{1_{S^{\lambda,\omega}}\right\}_k = \begin{cases} 1, & k \in S^{\lambda,\omega} \\ 0, & k \notin S^{\lambda,\omega} \end{cases}$, *for all* $k \in \{1, \ldots, p\}$; $\phi \in [0,1]$; *given a* $\gamma > 0$, *define the weight vector* $\hat{\omega} = 1/|\hat{\beta}|^{\gamma}$.

Notably, only predictor variables in the set $S^{\lambda,\omega} \subseteq \{1, \ldots, p\}$ can be chosen as the relaxed adaptive lasso solution. In the following, we discuss different functions and value

ranges of parameters under the set $S^{\lambda,\omega}$. The parameter $\lambda \geq 0$ determines the number of variables retained in the model. For $\lambda = 0$ or $\phi = 0$, the problem of solving the estimators in Equation (5) is transformed into an ordinary least squares problem where $S_0^{\lambda,\omega} = \{1, \ldots, p\}$ so that the purpose of variable selection cannot be achieved. As λ increases, all coefficients of the variables selected by adaptive lasso are compressed towards 0, and some finally become exactly 0. However, for a large $\lambda \to \infty$, all estimators are shrunk to 0, where $S^{\lambda,\omega} = \emptyset$, leading to a null model. In addition, the relaxation parameter ϕ controls the amount of shrinkage applied to the coefficients in estimation. When $\phi = 1$, the adaptive lasso and relaxed adaptive lasso estimators are the same. When $\phi < 1$, the shrinkage force on the estimators is weaker than that of the adaptive lasso. The optimal tuning parameters λ and ϕ are chosen by cross-validation. The vector $\hat{\omega} = 1/|\hat{\beta}|^{\gamma}$ assigns different weights to the coefficients; hence, the relaxed adaptive lasso has consistency when the weight vector is correctly chosen.

2.2. Algorithm

We will discuss the algorithm for computing the estimator of the relaxed adaptive lasso in this section. Note that (5) is a convex optimization problem, which means that we can obtain the global optimal solution effectively. Unlike concave penalties, however, multiple minimal penalties, such as SCAD, suffer from the multiple minimal problem. In the following, we discuss a simplified version of the relaxed adaptive lasso estimator algorithm. An improved algorithm is then proposed based on the process of computation for the relaxed lasso estimator [11].

The simple algorithm for relaxed adaptive lasso

Step (1). For a given $\gamma > 0$, we use $\hat{\beta}^{OLS}$ to construct the weight in an adaptive lasso based on the definition from Zou [6]. We can also replace $\hat{\beta}^{OLS}$ with other consistent estimators, e.g., $\hat{\beta}^{Ridge}$.

Step (2). Define $X_j^* = X_j/\hat{\omega}_j, j = 1, \ldots, p$, where $\hat{\omega}_j = 1/|\hat{\beta}^{OLS}|^{\gamma}$

Step (3). Then, the process of computing relaxed adaptive lasso solutions is identical to that of solving the relaxed lasso solutions in Meinshausen [11]. The relaxed lasso estimator is defined as

$$\hat{\beta}^{**} = \arg\min \left\| Y - \sum_{j=1}^{p} \left(X_j^* \right)^T \{\beta_j \cdot 1_{S^{\lambda,\omega}}\} \right\|_2^2 + \phi\lambda \sum_{j=1}^{p} |\beta_j|. \tag{6}$$

The Lars algorithm is first used to compute all the adaptive lasso solutions. Select a total of h resulting models $S_1, \ldots, S_h$ attained with the sorted penalty parameters $\lambda_1 > \lambda_2 > \ldots > \lambda_h = 0$. When $\lambda_h = 0$, for example, all variables with nonzero coefficients are selected, which is identical to the OLS function. On the other hand, $\lambda_0 = \infty$ completely shrinks the estimators to zero, thus leading to a null model. Therefore, a moderate $\lambda_k, k = 1, \ldots, h$ in the sequence of $\{\lambda_1, \ldots, \lambda_h\}$ is chosen such that $S_k = S^{\lambda,\omega}$. Then, define the OLS estimator $\tilde{\beta} = \hat{\beta}^{\lambda_k} + \lambda_k \delta(k)$, where $\delta(k) = \left(\hat{\beta}^{\lambda_k} - \hat{\beta}^{\lambda_{k-1}}\right)/(\lambda_{k-1} - \lambda_k)$ is the direction of adaptive lasso solutions, which can be obtained from the last step. If there exists at least one component j such that $\operatorname{sgn}(\tilde{\beta}_j) \neq \operatorname{sgn}\left(\hat{\beta}_j^{\lambda_k}\right)$, then all the adaptive lasso solutions on the set S_k of variables are identical to the set of relaxed lasso estimators $\hat{\beta}^{**}$ for $\lambda \in \mathcal{L}_k$. Otherwise, $\hat{\beta}^{**}$ for $\lambda_k \in \mathcal{L}_k$ are computed by linear interpolation between $\hat{\beta}_j^{\lambda_k}$ and $\tilde{\beta}_j$.

Step (4). Output the relaxed adaptive lasso solutions: $\hat{\beta}_j^* = \hat{\beta}_j^{**}/\hat{\omega}_j, j = 1, \ldots, p$.

Simple algorithms have the same computational complexity as Lars-OLS hybrid algorithms. However, due to the high computing complexity, this approach is frequently not ideal. Then, we consider an improved algorithm introduced by Hastie et al. [12], which

uses the definition of the relaxed adaptive lasso estimators to solve a problem of high computational complexity.

The improved algorithm for relaxed adaptive lasso

Step (1). As before, $S^{\lambda,\omega}$ denotes the active set of the adaptive lasso. Let $\hat{\beta}^{Alasso}$ denote the adaptive lasso estimator. The relaxed adaptive lasso solution can be defined as

$$\hat{\beta}^{*} = \phi\hat{\beta}^{Alasso} + (1-\phi)\hat{\beta}^{OLS}, \tag{7}$$

where ϕ is a constant with a value between 0 and 1.

Step (2). The submatrix $X_{S^{\lambda,\omega}}$ of active predictors is a reversible matrix; thus, $\hat{\beta}^{OLS} = \left(X_{S^{\lambda,\omega}} X_{S^{\lambda,\omega}}^{T}\right)^{-1} X_{S^{\lambda,\omega}} Y$.

Step (3). Define $X^{\#} = X_{S^{\lambda,\omega}}/\hat{\omega}$, where $\hat{\omega} = 1/|\hat{\beta}^{OLS}|^{\gamma}$; then, the adaptive lasso solution $\hat{\beta}^{ALasso}$ is identical to solving the lasso problem

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} \left(X_j^{\#}\right)^{T} \beta \right\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{8}$$

By means of the Karush–Kuhn–Tucker (KKT) optimality condition, the lasso solution over its active set can be written as

$$\hat{\beta}^{Lasso} = \left(X_{S^{\lambda,\omega}}^{\#} \left(X_{S^{\lambda,\omega}}^{\#}\right)^{T}\right)^{-1} \left(X_{S^{\lambda,\omega}}^{\#} Y - \lambda \mathrm{sgn}\left(\hat{\beta}^{Lasso}\right)\right). \tag{9}$$

From the transformation of the predictor matrix in Step (2), it follows that the adaptive lasso estimator is $\hat{\beta}^{Alasso} = \hat{\beta}^{Lasso}/\hat{\omega}$.

Step (4). Thus the improved solution of the relaxed adaptive lasso can be written as

$$\hat{\beta}_j^{*} = \begin{cases} \frac{\phi}{\hat{\omega}} \left(X_{S^{\lambda,\omega}}^{\#} \left(X_{S^{\lambda,\omega}}^{\#}\right)^{T}\right)^{-1} \left(X_{S^{\lambda,\omega}}^{\#} Y - \lambda \mathrm{sgn}\left(\hat{\beta}^{Lasso}\right)\right) + (1-\phi)\left(X_{S^{\lambda,\omega}} X_{S^{\lambda,\omega}}^{T}\right)^{-1} X_{S^{\lambda,\omega}} Y, & j \in S^{\lambda,\omega}, \\ 0, & j \notin S^{\lambda,\omega}. \end{cases} \tag{10}$$

The computational complexity of Algorithm 1 in the best case is equivalent to the ordinary lasso. Specifically, in Step (3) of the simple algorithm, the relaxed adaptive lasso estimator can be solved in the same way as the relaxed lasso. The improved algorithm is computed from the adaptive lasso and lasso estimators. Given the weight vector, the computational cost of the relaxed adaptive lasso is the same as that of the lasso [21]. Therefore, the computational complexity of Algorithm 2 is equivalent to that of the lasso.

Now we compare the computational cost of the two algorithms. The relaxed lasso's computational cost in the worst scenario is $O(n^3 p)$, which is slightly more expensive than the cost of the regular lasso with $O(n^2 p)$ Meinshausen [11]. For this reason, we compute the relaxed adaptive lasso estimator using the improved algorithm.

Algorithm 1. The simple algorithm for relaxed adaptive lasso.

Input: a given constant $\gamma > 0$, the weight vector $\hat{\omega} = 1/|\hat{\beta}^{OLS}|^{\gamma}$,
Precompute: $X^* = X/\hat{\omega}$

Initialization: Let $\lambda_1 > \lambda_2 > \ldots > \lambda_h$ to be the optimal parameter

corresponding to the modified models $S_1, \ldots, S_h$.
Set $k = 1$ to an initial order number of λ_k
Define $Q(\beta) = \left\| Y - \sum_{j=1}^{p} \left(X_j^*\right)^T \left\{\beta_j \cdot 1_{S^{\lambda,\omega}}\right\} \right\|_2^2 + \phi\lambda \sum_{j=1}^{p} |\beta_j|$,
$\tilde{\beta} = \hat{\beta}^{\lambda_k} + \lambda_k \delta(k)$, where $\delta(k) = \left(\hat{\beta}^{\lambda_k} - \hat{\beta}^{\lambda_{k-1}}\right)/(\lambda_{k-1} - \lambda_k)$
for $j = 1, \ldots, p$ **do**
if $sgn\left(\tilde{\beta}_j\right) \neq sgn\left(\hat{\beta}_j^{\lambda_k}\right)$ **then**
$\hat{\beta}^{**} \leftarrow \hat{\beta}^{Alasso}$
else
$\hat{\beta}^{**} \leftarrow Q(\tilde{\beta}) + \frac{Q(\tilde{\beta}) - Q(\hat{\beta}^{\lambda_{k-1}})}{\tilde{\beta} - \hat{\beta}^{\lambda_{k-1}}} \left(\tilde{\beta} - \hat{\beta}^{\lambda_{k-1}}\right)$
Set $k = k + 1$
until $k = h$
Output: $\hat{\beta}_j^* = \hat{\beta}_j^{**}/\hat{\omega}_j$

Algorithm 2. The improved algorithm for the relaxed adaptive lasso.

Input: Adaptive lasso estimator $\hat{\beta}^{Alasso}$, OLS estimator $\hat{\beta}^{OLS}$,
weight vector $\hat{\omega} = 1/|\hat{\beta}^{OLS}|^{\gamma}$
Precompute: $X^{\#} = X_{S^{\lambda,\omega}}/\hat{\omega}$, Let $S^{\lambda,\omega}$ be the active set of the adaptive lasso
Initialization: Define $\hat{\beta}^{Lasso} = \arg\min_{\beta} \left\| Y - \sum_{j=1}^{p} \left(X_j^{\#}\right)^T \beta \right\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$
for $j = 1, \ldots, p$ **do**
if $j \in S^{\lambda,\omega}$ **then**
compute $\hat{\beta}^{OLS} = \left(X_{S^{\lambda,\omega}} X_{S^{\lambda,\omega}}^T\right)^{-1} X_{S^{\lambda,\omega}} Y$,
$\hat{\beta}^{Lasso} = \left(X_{S^{\lambda,\omega}}^{\#} \left(X_{S^{\lambda,\omega}}^{\#}\right)^T\right)^{-1} \left(X_{S^{\lambda,\omega}}^{\#} Y - \lambda \mathrm{sgn}(\hat{\beta}^{Lasso})\right)$
else
Stop iterations
until $j = p$
Output: $\hat{\beta}^{Alasso} = \hat{\beta}^{Lasso}/\hat{\omega}$, $\hat{\beta}^* = \phi\hat{\beta}^{Alasso} + (1 - \phi)\hat{\beta}^{OLS}$

2.3. Asymptotic Results

To investigate the asymptotic property, we make the following two assumptions about the architecture used in the setup of Fu and Knight [18]:

$$\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \to \Sigma, \tag{11}$$

where Σ is a positive definite matrix. Furthermore,

$$\frac{1}{n} \max_{1 \leq i \leq n} x_i^T x_i \to 0. \tag{12}$$

Without loss of generality, the sparse constant vector β is defined as the true coefficient of the model. We assume that the number of nonzero estimators selected into the real model is q, that is $\beta = (\beta_1, \ldots, \beta_q, 0, 0, \ldots)$, where $\beta_j \neq 0$ only for $j = 1, \ldots, q$ and $\beta_j = 0$ for $j = q + 1, \ldots, p$. The true model is, hence, $S_* = \{1, \ldots, q\}$. The covariance matrix

$\Sigma = \frac{1}{n}XX^T$ can be written in block-wise form, i.e., $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, where Σ_{11} is a $q \times q$ matrix. The random loss $L(\lambda, \omega)$ of the adaptive lasso is defined as

$$L(\lambda, \omega) = E\left(Y - X^T \hat{\beta}^{Alasso}\right)^2 - \sigma^2. \tag{13}$$

The loss $L(\lambda, \phi, \omega)$ of the relaxed adaptive lasso is analogously defined as

$$L(\lambda, \phi, \omega) = E\left(Y - X^T \hat{\beta}^{*}\right)^2 - \sigma^2. \tag{14}$$

We discover that the relaxed adaptive lasso estimator has the same rapid convergence rate as the relaxed lasso estimator when the exponential growth rate of the size p is ignored. Additionally, the adaptive lasso has a slower pace than both of them but is slightly faster than the lasso estimator. We make the following assumptions concerning asymptotic results for low-dimensional sparse solutions to demonstrate the above conclusion.

Assumption 1. *The number of predictors $p = p_n$ increases exponentially with the number of observations n, that is, there exist some $c > 0$, $0 < s < 1$ such that $p_n \sim se^{cn}$.*

We cannot rule out the possibility that the remaining $p_n - q$ noise factors are linked with the response. A square matrix is said to be diagonally dominant if the magnitude of the diagonal entry in each row of the matrix is greater than or equal to the sum of the magnitudes of all the other (nondiagonal) entries in that row.

Assumption 2. *Σ and Σ^{-1} are diagonally dominant at some constant $c < 0$, for all $n \in \mathbb{N}$.*

Notably, when the diagonal is positive, the diagonally dominating symmetric matrix is positive definite. Based on this premise, the inverse matrix of Σ can guarantee its existence.

Assumption 3. *We limit the penalty parameter λ to the range $\mathcal{L}$,*

$$\mathcal{L} = \{\lambda \geq 0 : ce^{p_n} \leq n\}, \tag{15}$$

if and only if there exists an arbitrarily large $c > 0$.

Assumption 3 holds true if the exponent of the number of variables in the selected model is less than the sample size n. Using λ values in the range $\mathcal{L}$, relaxed lasso, adaptive lasso, and relaxed adaptive lasso can obtain consistent variable selection and a specified number of nonzero coefficients.

Lemma 1. *Assume that predictor variables are independent of each other, λ_n, $n \in \mathbb{N}$ is the penalty parameter of the adaptive lasso, and its order is $\lambda_n = O\left(n^{\frac{s-1-2\gamma}{2}}\right)$ for $n \to \infty$. Under Assumptions 1–3,*

$$P(\exists k > q : k \in S_{\lambda_n}) \to 1, n \to \infty. \tag{16}$$

As a result of Lemma 1, the chance of at least one noise variable being evaluated as nonzero is close to one. We prove Theorem 1 by utilizing the conclusion of Lemma 1 on the order of the penalty parameter.

Lemma 2. *Let $\liminf_{n\to\infty} \frac{n^*}{n} \to \frac{1}{A}$ with $A \geq 2$, n^* being the number of observations. Then, under Assumptions 1–3,*

$$\sup_{\lambda \in \mathcal{L}, \gamma > 0} |L(\lambda, \phi, \omega) - L_{n^*}(\lambda, \phi, \omega)| = O_p\left(n^{-1} \log^2 n\right), n \to \infty. \tag{17}$$

We want to investigate the computational cost of the specified parameters by examining the order of the relaxed adaptive lasso loss function. Lemma 2 is a technique that will assist us in proving Theorem 3.

Lemma 3. *Assume that predictor variables are independent of each other, λ_n, $n \in \mathbb{N}$ is the penalty parameter of the relaxed adaptive lasso, and $n^{s+1}\lambda_n^3 \to \infty$ for $n \to \infty$. Under Assumptions 1–3,*

$$P(\exists k > q : k \in S_{\lambda_n}) \to 0. \tag{18}$$

As a result of Lemma 3, the noise variable can be predicted to be 0. If the penalty parameter ensures that λ_n^3 converges to 0 at a slower rate than n^{s+1}, the noise variable can be precisely evaluated as nonzero with a probability approaching 0. In addition, Lemma 3 helps to prove Theorem 3 by describing the order of the penalty parameter of the relaxed adaptive lasso.

Theorem 1 addresses the question of whether the adaptive lasso can sustain a faster convergence rate as the number of noise variables increases rapidly and the convergence speed exceeds that of the lasso. The addition of the weight parameter enables the adaptive lasso to gain oracle qualities while also increasing the algorithm's rate of convergence.

Theorem 1. *Assume that predictor variables are independent of each other. Under Assumptions 1–3, $\Sigma = 1$ for any $t > 0$ and $n \to \infty$. The convergence rate of the adaptive lasso is as follows:*

$$P\left(\inf_{\lambda \in \mathcal{L}} L(\lambda, \omega) > tn^{-r}\right) \to 1, \forall r > 1 + 2\gamma - s. \tag{19}$$

On the other hand, Theorem 2 establishes that the convergence rate of the relaxed adaptive lasso is equivalent to that of the relaxed lasso. Theorem 2 resolves the question of whether the convergence rate of the relaxed adaptive lasso is consistent with that of the relaxed lasso by establishing that the convergence rate of the relaxed adaptive lasso is not related to the noise variable's growth rate r or the parameter s that determines the growth rate.

Theorem 2. *Assume that predictor variables are independent of each other. Under Assumptions 1–3, for $n \to \infty$, the convergence rate of the relaxed adaptive lasso is as follows:*

$$\inf_{\lambda \in \mathcal{L}, \phi \in [0,1], \gamma > 0} L(\lambda, \phi, \omega) = O_p\left(n^{-1}\right). \tag{20}$$

The shade in Figure 1 represents the rate at which various models converge. The rate of the relaxed adaptive lasso is the same as that of the relaxed lasso; this indicates that the convergence rate of the relaxed adaptive lasso is unaffected by the rapid increase in the noise variable, and it can still retain a high rate. Although the adaptive lasso's convergence rate is suboptimal, it is faster than the lasso's due to the presence of the weight vector. The addition of an excessive number of noise variables slows the Lasso estimator, regardless of how the penalty parameter is chosen [11].

The convergence rate of the relaxed adaptive lasso is as robust as the rate of the relaxed lasso, i.e., it is unaffected by noise factors. Theorem 3 demonstrates that cross-validation selection of the parameters λ, ϕ can still maintain a rapid rate.

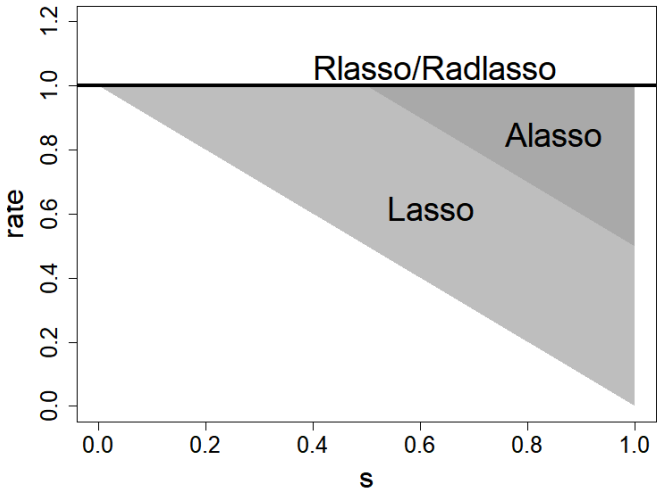

Figure 1. Comparison of convergence rates between the relaxed adaptive lasso, relaxed lasso, adaptive lasso, and ordinary lasso. Both the relaxed adaptive lasso and the relaxed lasso have the same rate $O_p(n^{-1})$, regardless of s. Adaptive lasso has a rate of $O_p(n^{-r})$ only if $r > 1 + 2\gamma - s$. Additionally, the rate of the lasso is $O_p(n^{-r})$ only if $r > 1 - s$.

Franklin [22] indicated that K-fold cross-validation includes K partitions and each partition consists of $\widetilde{n}$ observation data, where $\frac{\widetilde{n}}{n} \to \frac{1}{K}$ for $n \to \infty$. When building an estimator on a different set of observations than R, define the empirical loss of observations as $L_{R,\widetilde{n}}(\lambda, \phi, \omega)$ for $R = 1, \ldots, K$. Let $L_{cv}(\lambda, \phi, \omega)$ be the empirical loss function,

$$L_{cv}(\lambda, \phi, \omega) = K^{-1} \sum_{R=1}^{K} L_{R,\widetilde{n}}(\lambda, \phi, \omega). \tag{21}$$

The selection of $\hat{\lambda}$, $\hat{\phi}$ and $\hat{\omega}$ is performed by minimizing the loss function $L_{cv}(\lambda, \phi, \omega)$, that is,

$$(\hat{\lambda}, \hat{\phi}, \hat{\omega}) = \arg\min L_{cv}(\lambda, \phi, \omega). \tag{22}$$

This article uses five-fold cross-validation in the numerical study.

Theorem 3. *Under Assumptions 1–3, the convergence rate of K-fold cross-validation with* $2 \leq K < \infty$ *holds that*

$$L(\hat{\lambda}, \hat{\phi}, \hat{\omega}) = O_p\left(n^{-1} \log^2 n\right). \tag{23}$$

Therefore, when K-fold cross-validation is used to determine the relaxed adaptive lasso's penalty parameters λ, ϕ, the convergence speed may maintain a relatively ideal outcome. As a result, if using cross-validation to select the penalty parameters, the optimal rate and consistent variable selection under the oracle selection of penalty parameters may be nearly achieved.

Theorem 4. *If* $\frac{\lambda_n}{n} \to \lambda_0 \geq 0$, *then* $\beta \cdot 1_S \xrightarrow{p} \beta$ *in the relaxed adaptive lasso estimator; moreover, if* $\phi\lambda_n = o(n)$, $\hat{\beta}^*$ *is consistent.*

Theorem 4 indicates that the relaxed adaptive lasso estimator is consistent under the condition $\phi\lambda_n = o(n)$. $\hat{\beta}^*$ does not have to be root-n consistent; nonetheless, the consistency of the relaxed adaptive lasso is determined by the conclusion drawn from probability convergence.

3. Simulation

3.1. Setup

We present a numerical study in this section to compare the performance of the relaxed adaptive lasso to that of the lasso, relaxed lasso, and adaptive lasso. Based on the simulation setup of Meinshausen [11], we considered the linear model $y = x^T\beta + \varepsilon$, where $x = (x_1, \ldots, x_p)$ is the predictor vector and random error ε is an independent and identically distributed random variable with mean 0 and variance σ^2. The remaining parameter settings and procedures are as follows.

i. Given sample size $n = 100, 500, 1000$ and data dimension $p = 20, 50$.
ii. The true regression coefficient $\beta \in \mathbb{R}^p$ has its first $q = 10 (q \leq p)$ signal variables taking nonzero coefficients equally spaced from 0.5 to 10 in the sense that $\beta_j \neq 0$ for all $j \leq q$ and the remaining $p - q$ coefficients are zero.
iii. The design matrix $X \in \mathbb{R}^{n \times p}$ is generated from a normal distribution $N(0, \Sigma)$, where covariance matrix $\Sigma = \mathrm{cov}(x) = (c_{ij})_{p \times p}$ has entries $c_{ij} = 1, i = j = 1, \ldots, p$ and $c_{ij} = \rho^{|i-j|}, i \neq j$. The correlation between predictor variables is set to $\rho = 0.5$.
iv. The theoretical signal-to-noise ratio in this simulation is defined as $\mathrm{SNR} = \mathrm{Var}(x^T\beta)/\sigma^2$. We discuss either SNR=0.2 for low or SNR=0.8 for high to calculate the variance of ε so that the response variable Y generated from the linear regression model follows $N_n(x^T\beta, \sigma^2 I)$.
v. We compute the weight of the adaptive lasso via the ridge regression estimator with $\gamma = 1$. For each method, five-fold cross-validation is used to select the penalty parameters, and the loss function to apply for cross-validation is chosen by minimizing the prediction error on the test set. Furthermore, we pick the least complex model that is comparable in accuracy to the best model under the "one-standard-error" criterion Franklin [22].

For each of the settings above, this process is repeated a total of 100 times to compute the following evaluation metrics, and the average results are recorded.

3.2. Evaluation Metrics

The data are split randomly into a training set and a test set. Suppose that $x_0 \in \mathbb{R}^p$ is drawn from the row of the testing design matrix X, and $\hat{y}_0$ denotes its connected response value by fitting the model. Additionally, let $\hat{\beta}_0$ denote the corresponding estimated coefficient of the predictor variable x_0.

Mean-square error:

$$\mathrm{MSE} = E(y_{test} - \hat{y}_0)^2 = E\left(y_{test} - x_0^T\hat{\beta}_0\right)^2. \tag{24}$$

This value assesses the accuracy of the model prediction. A good model has the highest prediction accuracy in the sense that its prediction error, MSE, is minimized. The following metrics were developed by Hastie et al. [12].

Relative accuracy:

$$\mathrm{RA}(\hat{\beta}) = \frac{E(x_0^T\hat{\beta}_0 - x_0^T\beta)^2}{E(x_0^T\beta)^2} = \frac{(\hat{\beta}_0 - \beta)^T\Sigma(\hat{\beta}_0 - \beta)}{\beta^T\Sigma\beta}. \tag{25}$$

Relative test error:

$$\mathrm{RTE}(\hat{\beta}) = \frac{E(y_{test} - x_0^T\hat{\beta}_0)^2}{\sigma^2} = \frac{(\hat{\beta}_0 - \beta)^T\Sigma(\hat{\beta}_0 - \beta) + \sigma^2}{\sigma^2}. \tag{26}$$

Proportion of variance explained:

$$\mathrm{PVE}(\hat{\beta}) = 1 - \frac{E\left(y_{test} - x_0^T \hat{\beta}_0\right)^2}{Var(y_{test})} = 1 - \frac{\left(\hat{\beta}_0 - \beta\right)^T \Sigma \left(\hat{\beta}_0 - \beta\right) + \sigma^2}{\beta^T \Sigma \beta + \sigma^2}. \tag{27}$$

Number of nonzeros: The average number of nonzero estimated coefficients,

$$\left\|\hat{\beta}_{nonzero}\right\|_0 = \sum_{j=1}^{p} 1_{\{\hat{\beta}_j \neq 0\}}. \tag{28}$$

where $1_{\{\hat{\beta}_j \neq 0\}} = \begin{cases} 1, & \hat{\beta}_j \neq 0 \\ 0, & \hat{\beta}_j = 0 \end{cases}$. An ideal score should be close to the number of true nonzero coefficients q.

Furthermore, in addition to the assessment of prediction accuracy, we explore the last metric to measure the right variable recovery. This metric quantifies the degree to which the valid solution $\hat{\beta}$ to the convex optimization problem in Equation (5) matches the true coefficient β.

3.3. Summary of Results

Table 1 summarizes the average results of simulation for lasso, relaxed lasso, adaptive lasso and relaxed adaptive lasso with SNR = 0.2. We find that the relaxed adaptive lasso has the best RR, RTE, PVE and MSE scores on average. In other words, the proposed method achieves the maximum prediction accuracy in the majority of cases, despite occasions where the adaptive lasso's MSE is somewhat better than that of the relaxed adaptive lasso. Specifically, the adaptive lasso yields a much smaller MSE due to the small sample size (e.g., $n = 100$). However, when the sample size is increased to $n = 1000$, the relaxed adaptive lasso outperforms all other methods owing to the feature of large samples in which parametric estimators converge in probability to true parameters.

In Table 2, excellent performance is observed for all methods when the SNR is increased to 0.8. As expected, relaxed adaptive lasso maintains its competitive edge and achieves overall good performance. In particular, it roughly maintains the correct number of nonzero variables as the number of observations n increases. For $(n, p) = (100, 20)$ and $(n, p) = (100, 50)$, it holds up to five and four variables, respectively. For $(n, p) = (1000, 20)$ and $(n, p) = (1000, 50)$, up to nine and eight, respectively, approach the number of truly valid features $q = 10$. This illustrates that the sparsity pattern of estimators in the relaxed adaptive lasso achieves the proper variable recovery when n is quite large. In contrast, the relaxed lasso and adaptive lasso shrink too many coefficients toward zero; as a result, fewer variables remain in the resulting model. Therefore, we conclude that as the number of observations n grows rapidly, the number of variables preserved in the model grows as well, and it is possible to select the important variables approximately correctly, i.e., having proper variable recovery.

Table 1. Simulation results for SNR = 0.2.

p	*n*	Method	RR	RTE	PVE	MSE	Number of Nonzeros
20	100	Lasso	0.997	1.206	0.4	96.2	1
		Rlasso	0.997	1.205	0.6	95.4	1
		Alasso	0.995	1.205	0.8	94.5	1
		Radlasso	0.986	1.203	2.4	100.1	2
	500	Lasso	0.990	1.199	1.6	91.4	4
		Rlasso	0.987	1.198	2.1	90.3	2
		Alasso	0.989	1.198	1.9	90.5	3
		Radlasso	0.974	1.196	4.3	86.5	6
	1000	Lasso	0.987	1.197	2.1	90.2	5
		Rlasso	0.983	1.196	2.9	89.6	3
		Alasso	0.985	1.197	2.4	90.1	4
		Radlasso	0.974	1.195	4.4	86.8	7
50	100	Lasso	0.998	1.197	0.4	99.7	1
		Rlasso	0.997	1.197	0.5	99.9	1
		Alasso	0.993	1.196	1.2	98.8	2
		Radlasso	0.985	1.195	2.3	106.8	2
	500	Lasso	0.992	1.200	1.4	93.4	4
		Rlasso	0.986	1.199	2.3	92.5	2
		Alasso	0.988	1.199	1.9	91.6	3
		Radlasso	0.976	1.197	4.0	90.6	5
	1000	Lasso	0.987	1.195	2.1	88.8	5
		Rlasso	0.982	1.195	2.9	88.0	3
		Alasso	0.985	1.195	2.5	88.4	4
		Radlasso	0.974	1.193	4.3	86.5	6

Table 2. Simulation results for SNR = 0.8.

p	*n*	Method	RR	RTE	PVE	MSE	Number of Nonzeros
20	100	Lasso	0.980	1.789	8.8	75.1	5
		Rlasso	0.972	1.783	12.1	73.8	3
		Alasso	0.975	1.785	11.1	72.8	4
		Radlasso	0.960	1.773	17.8	75.2	5
	500	Lasso	0.969	1.781	13.8	61.5	7
		Rlasso	0.962	1.775	17.1	60.7	5
		Alasso	0.967	1.780	14.7	61.9	6
		Radlasso	0.956	1.771	19.7	58.8	9
	1000	Lasso	0.966	1.762	14.8	59.3	8
		Rlasso	0.959	1.756	17.8	58.5	6
		Alasso	0.964	1.760	15.9	59.3	7
		Radlasso	0.956	1.753	19.4	57.1	9
50	100	Lasso	0.985	1.784	6.7	75.5	4
		Rlasso	0.978	1.779	9.4	73.4	3
		Alasso	0.974	1.775	11.4	69.7	6
		Radlasso	0.963	1.766	16.2	83.3	4
	500	Lasso	0.970	1.773	13.1	62.9	7
		Rlasso	0.963	1.767	16.6	61.5	5
		Alasso	0.967	1.770	14.6	61.9	6
		Radlasso	0.958	1.763	18.7	60.4	7
	1000	Lasso	0.967	1.774	14.6	59.7	8
		Rlasso	0.960	1.768	17.9	58.7	6
		Alasso	0.964	1.772	15.8	59.4	7
		Radlasso	0.957	1.765	19.3	57.6	8

NOTE: The MSE and PVE values in the table are 100 and 1000 times larger to emphasize the distinction between these methods.

4. Application to Real Data

4.1. Dataset

The real dataset used in this study is from the CSMAR Database, which contains 11 research series on stocks, companies, funds, the economy, industries, etc. It is widely recognized as one of the most professional and accurate databases available for research purposes. Our data include a total of 2137 records, with each record corresponding to the financial data of one listed company in 2021. The training set is made up of the first 1496 observations, and the test set is made up of the rest. The response variable is the R&D investment of the company, and the predictor variables include 86 factors that may have an effect on the firm's R&D investment, such as fixed-assets depreciation, accounts receivable and payroll payable. To compare the model selection performance of the method proposed in this paper to that of the other three methods, the aforementioned methods are used to fit the model on the training set, and the prediction accuracy of these models is measured in terms of the MSE on the test set. It is shown in the following that the relaxed adaptive lasso has the highest prediction accuracy with the smallest MSE value.

4.2. Analysis Results

As can be seen in Table 3, the MSE values of the lasso and adaptive lasso are, respectively, as large as 0.521 and 0.575, indicating that they have the worst prediction accuracy. The relaxed lasso performs somewhat better than the lasso and the adaptive lasso in terms of MSE. As expected, the relaxed adaptive lasso estimator's prediction accuracy remains satisfactory, with the smallest MSE of 0.429. In Table 4, a total of 10 variables are selected by the relaxed adaptive lasso. It has shown that Cash Paid to and for Employees, Cash Paid for Commodities or Labor, Business Taxes and Surcharges are identified as the three most influential factors on R&D investment. As a result, we can conclude that relaxed adaptive lasso leads to the simplest model with the highest prediction accuracy among the four foregoing methods.

Table 3. Prediction accuracy for R&D investment study.

Method	Lasso	Rlasso	Alasso	Radlasso
MSE	0.521	0.485	0.575	0.429

Table 4. Variables selected by Radlasso.

Order Number	Explanatory Variable	Coefficient
x_{10}	Cash Flow from Operations	0.008
x_{13}	Net Increase in Cash and Cash Equivalents	0.048
x_{15}	Net Accounts Receivable	0.208
x_{26}	Non-Current Assets	−0.214
x_{48}	Business Taxes and Surcharges	−0.265
x_{67}	Interest Income	0.130
x_{70}	Profit and Loss from Asset Disposal	0.154
x_{73}	Cash Paid for Commodities or Labor	0.386
x_{74}	Cash Paid to and for Employees	0.569
x_{83}	Cash Flow from Financing Activities Net Amount	−0.080

Among the most important explanatory variables affecting R&D investment, Cash Paid to and for Employees measures the company's actual benefits and rewards; Cash Paid for Commodities or Labor measures the overall payment ability of the company; and Business Taxes and Surcharges measure the tax burden of the company's operation. According to the estimator coefficients estimated by the simplified model, firms with high Cash Paid to and for Employees and Cash Paid for Commodities or Labor tend to spend more on R&D (the positive coefficient on the response variable), whereas Business Taxes and Surcharges have a negative influence on the company's R&D investment. From the

results of the analysis, it is not surprising that companies focused on welfare take more advantage of innovative technology because generous compensation not only improves employees' work motivation but also helps to retain and recruit talent. Furthermore, strong payment ability implies the high profitability of companies with successful operations, allowing them to spend massive amounts of money on R&D. Note that a heavy tax burden may result in a lower investment cost for a company. In general, increasing R&D input is highly influenced by a few selected variables, the three most important of which are the company's welfare, payment ability and tax burden.

5. Conclusions

In this article, we have proposed a two-stage variable selection method called relaxed adaptive lasso as a combination of relaxed lasso and adaptive lasso estimation. From the proof of the theorem, we conclude that the relaxed adaptive lasso has the same convergence rate as the relaxed lasso with $O_p\left(n^{-1}\right)$ and that both are faster than adaptive lasso and ordinary lasso in the low-dimensional setting. Furthermore, the relaxed adaptive lasso has the property of consistency, which means that the probability of selecting the true model approaches one under the condition of $\phi\lambda_n = o(n)$. The simulation study has shown that the proposed method has comparable prediction accuracy and accurate variable recovery as the number of observations increases. In practical applications, the conclusion has been confirmed by the analysis of the financial data of the listed company.

We have shown the asymptotic property of the relaxed adaptive lasso in the linear model. For further research, it is suggested to extend the theory and methodology to the generalized linear model [23]. In addition, the model does not handle the high-dimensional case well, where the variable dimension is much larger than the sample size. We propose to combine the existing idea with two-stage variable selection methods such as Sure Independence Screening (SIS) [24] and Distance Correlation Based SIS (DC-SIS) [25] to overcome this challenge.

Author Contributions: Conceptualization, Y.L.; methodology, R.Z., T.Z., Y.L., X.X.; software, R.Z., T.Z., Y.L., X.X.; formal analysis, R.Z., T.Z., Y.L., X.X.; data curation, R.Z., T.Z., Y.L., X.X.; writing—original draft preparation, R.Z., T.Z., Y.L., X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Humanities and Social Science Research Project of Hebei Education Department (SQ201110), Hebei GEO University Science and Technology Innovation Team (KJCXTD-2022-02), Basic scientific research Funds of Universities in Hebei Province (QN202139) and S&T Program of Hebei (22557688D).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Appendix A. Appendix Proof of Lemma 1

Proof. First, define the adaptive lasso estimator $\hat{\beta}^{Alasso}$ on the set $S_* = \{1, \ldots, q\}$ as

$$\hat{\beta}^{Alasso} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} \left(Y_i - \sum_{k \in S_*} \beta_k X_i^k \right)^2 + \lambda_n \sum_{i=1}^{n} \hat{w}_j |\beta_j|,$$

where the estimator shrinks to 0 outside the interval S_*, $\hat{w} = 1/|\hat{\beta}|^{\gamma}$. According to Meinshausen [11], we similarly define the residuals under the adaptive lasso estimator $\hat{\beta}^{Alasso}$

$$D_i = Y_i - \sum_{b \in S_*} \hat{\beta}^{Alasso} X_i^b.$$

Thus,

$$P(\exists k > q : k \in S_{\lambda_n}) \geq P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} D_i X_i^k > \lambda_n |\hat{\beta}^k|^{-\gamma}\right). \tag{A1}$$

Consider the distribution of the gradient when $k > q$,

$$n^{-1} \sum_{i=1}^{n} D_i X_i^k \sim N\left(0, n^{-2} \sum_{i=1}^{n} D_i^2\right).$$

The expected value of the averaged squared residuals is larger than $\frac{\sigma^2(n-q)}{n}$ for any $\lambda > 0$, so

$$P\left(n^{-1} \sum_{i=1}^{n} D_i^2 > \frac{\sigma^2}{2}\right) \to 1, n \to \infty.$$

If $n^{-1} \sum_{i=1}^{n} D_i^2 = \frac{\sigma^2}{2}$, then $n^{-1} \sum_{i=1}^{n} D_i X_i^k \sim N\left(0, \frac{\sigma^2}{2n}\right)$; thus, for $c, d > 0$,

$$P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} D_i X_i^k > \lambda_n |\hat{\beta}^k|^{-\gamma}\right) \geq d\lambda_n^{-1} |\hat{\beta}^k|^{\gamma} \exp\left(-tn\lambda_n^2 \hat{\beta}_k^{-2\gamma}\right).$$

There are $p_n - q$ variables when $k > q$. Consider the boundary of the gradient for $p_n - q$ noise variables:

$$P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} D_i X_i^k > \lambda_n |\hat{\beta}^k|^{-\gamma}\right) \leq \exp\left(-(p_n - q)d\lambda_n^{-1} |\hat{\beta}^k|^{\gamma} \exp\left(-tn\lambda_n^2 \hat{\beta}_k^{-2\gamma}\right)\right).$$

Note that

$$n\lambda_n^2 \hat{\beta}_k^{-2\gamma} = n^{2\gamma+1} \lambda_n^2 O(1).$$

We set the order of the parameter λ_n in adaptive lasso to $\lambda_n = O\left(n^{\frac{s-1-2\gamma}{2}}\right)$, then $n^{2\gamma+1}\lambda_n^2 = O(n^s)$; so, we have $n^{2\gamma+1}\lambda_n^2 \to \infty$. According to Assumption 1, $p_n \sim se^{cn}$. Thus, for some $g > 0$,

$$\lambda_n^{-1} |\hat{\beta}^k|^{\gamma} \to \lambda_n^{-1} n^{-\gamma},$$

$$\lambda_n^{-1} n^{-\gamma} \sim n^{\frac{1-s}{2}} \to \infty,$$

so

$$P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} R_i X_i^k > \lambda_n |\hat{\beta}^k|^{-\gamma}\right) \to 0, n \to \infty.$$

which, using (A1), completes the proof. □

Appendix B. Appendix Proof of Lemma 2

Proof. Assume that $S_1, \ldots, S_h$ is the collection of models estimated by the adaptive lasso and let $\lambda_k, k = 1, \ldots, h (\lambda_1 < \ldots < \lambda_n)$ be the largest one such that $S_k = S_\lambda$. For all $k \in \{1, \ldots, h\}$, ϕ is a constant with a value between 0 and 1, according to the definition of a convex function, the relaxed adaptive lasso solution on the set $B_1, \ldots B_n$ is given as

$$B_k = \left\{\beta = \phi\hat{\beta}^{Alasso} + (1 - \phi)\hat{\beta}^{OLS}\right\}. \tag{A2}$$

The estimate $\hat{\beta}^{Alasso}$ is the adaptive lasso estimate for penalty parameter λ_k, and $\hat{\beta}^{OLS}$ is the corresponding OLS estimator. Give the loss function as follows,

$$L(\lambda, \phi, \omega) = E\left(Y - \sum_{k\in\{1,\ldots,p\}} \hat{\beta}_k^* X^k\right)^2.$$

Substituting into formula (A2) yields

$$L(\lambda, \phi, \omega) = E\left(Y - \sum_{k\in\{1,\ldots,p\}} \hat{\beta}^{OLS} X^k - \phi\left(\hat{\beta}^{Alasso} - \hat{\beta}^{OLS}\right) X^k\right)^2.$$

For any λ, set $M_\lambda = Y - \sum_{k\in\{1,\ldots,p\}} \hat{\beta}^{OLS} X^k$, $N_\lambda = \left(\sum_{k\in\{1,\ldots,p\}} \hat{\beta}^{Alasso} - \sum_{k\in\{1,\ldots,p\}} \hat{\beta}^{OLS}\right) X^k$.

Then

$$L(\lambda, \phi, \omega) = E\left(M_\lambda^2\right) - 2\phi E(M_\lambda N_\lambda) + \phi E\left(N_\lambda^2\right).$$

Let $M_\lambda^2 = x$. According to Bernstein's inequality, there are some $m > 0$. For any $\varepsilon > 0$,

$$P\left(\frac{1}{n}\sum x_i - Ex < \frac{d}{n}\log(1-\delta) + \sqrt{\frac{2\mathrm{var}(x)\log\left(\frac{1}{\delta}\right)}{n}}\right) \geq 1 - \delta.$$

Let $\delta = \frac{1}{n}$, we have

$$P\left(E_{n^*}\left(M_\lambda^2\right) - E\left(M_\lambda^2\right) > -m(n^*)^{-1}\log n\right) = P\left(\frac{n-n^*}{nn^*}\sum M_\lambda^2 > -m(n^*)^{-1}\log n\right) \geq 1 - \frac{1}{n},$$

so

$$\limsup_{n\to\infty} P\left(|E_{n^*}\left(M_\lambda^2\right) - E\left(M_\lambda^2\right)| > m(n^*)^{-1}\log n\right) < \varepsilon.$$

The same can be obtained:

$$\limsup_{n\to\infty} P\left(|E_{n^*}(M_\lambda N_\lambda) - E(M_\lambda N_\lambda)| > m(n^*)^{-1}\log n\right) < \varepsilon,$$

$$\limsup_{n\to\infty} P\left(|E_{n^*}\left(N_\lambda^2\right) - E\left(N_\lambda^2\right)| > m(n^*)^{-1}\log n\right) < \varepsilon.$$

Hence, there exists some $m > 0$ for every $\varepsilon > 0$ such that

$$\limsup_{n\to\infty} P\left(\sup_{\lambda,\omega}|L(\lambda,\phi,\omega) - L_{n^*}(\lambda,\phi,\omega)| < h \sup_{\lambda\in\{\lambda_1,\ldots,\lambda_h\},\omega}|L(\lambda_i,\phi,\omega) - L_{n^*}(\lambda_i,\phi,\omega)|\right) > 1 - \varepsilon,$$

so

$$\limsup_{n\to\infty} P\left(|L(\lambda,\phi,\omega) - L_{n^*}(\lambda,\phi,\omega)| > m(n^*)^{-1}\log^2 n\right) < \varepsilon,$$

which completes the proof. □

Appendix C. Appendix Proof of Lemma 3

Proof. Using Bonferroni's inequality, it can be written as

$$P(\exists k > q : k \in S_{\lambda_n}) = P\left(\sum_{k=q+1}^{p} \cup k \in S_{\lambda_n}\right) \leq \sum_{k=q+1}^{p} P(k \in S_{\lambda_n}).$$

By Lemma 1, it follows that

$$\sum_{k=q+1}^{p} P(k \in S_{\lambda_n}) \leq \sum_{k=q+1}^{p} d\lambda_n^{-1} \exp\left(-tn\lambda_n^2\right) = O\left(n^{-1-s}\lambda_n^{-3}\right), s > 0.$$

Let λ_n be a sequence with $n^{s+1}\lambda_n^3 \to \infty, n \to \infty$ and

$$\sum_{k=q+1}^{p} P(k \in S_{\lambda_n}) \leq O\left(n^{-1-s}\lambda_n^{-3}\right) \to 0,$$

which completes the proof. □

Appendix D. Appendix Proof of Theorem 1

Proof. Let $\theta = \beta - \hat{\beta}^{\lambda_*}$, $\delta^\lambda = \hat{\beta}^\lambda - \hat{\beta}^{\lambda_*}$ then

$$\left(\hat{\beta}_k^\lambda - \beta_k\right)^2 = \theta_k^2 - 2\theta_k\delta_k^\lambda + \left(\delta_k^\lambda\right)^2.$$

For $n \to \infty$ and any $\varepsilon > 0$, we have $|\theta_k| > (1-\varepsilon)\lambda_*$ with probability converging to 1; then, $|\theta_k| < (1+\varepsilon)\lambda_*$. Hence, for all $k \leq q$, there is

$$\left(\hat{\beta}_k^\lambda - \beta_k\right)^2 \geq (1-\varepsilon)^2\lambda_*^2 + 2(1+\varepsilon)\lambda_*\delta_k^\lambda + \left(\delta_k^\lambda\right)^2,$$

then

$$\left(\hat{\beta}_k^\lambda - \beta_k\right)^2 \geq (1-\varepsilon)^2\lambda_*^2 - 2\left(1-\varepsilon^2\right)\lambda_*(\lambda_* - \lambda) + (1-\varepsilon)^2(\lambda_* - \lambda)^2.$$

Therefore, with probability converging to 1 for $n \to \infty$, we can obtain

$$\inf_{\lambda \geq \lambda_*} L(\lambda) \geq \left[(1-\varepsilon)^2 + 2\sqrt{q}\left(1-\varepsilon^2\right) + q(1-\varepsilon)^2\right]^2\lambda_*^2.$$

According to Lemma 1: $\lambda_n \sim n^{\frac{s-1-2\gamma}{2}}$,

$$\inf_{\lambda \geq \lambda_*} L(\lambda) \sim O_p\left(n^{-r}\right), \forall r > 1 + 2\gamma - s,$$

which completes the proof. □

Appendix E. Appendix Proof of Theorem 2

Proof. Denote the set of nonzero coefficients of β by $S_* = \{1, \ldots, q\}$. Define event E as

$$\exists \lambda : S_\lambda = S_*.$$

Let $t > 0$, then

$$P\left(\inf_{\lambda,\phi,\omega} L(\lambda,\phi,\omega) > tn^{-1}\right) \leq P\left(\inf_{\lambda,\phi,\omega} L(\lambda,\phi,\omega) > tn^{-1} | E\right) P(E) + P(E^c).$$

Assume that λ_* is the smallest value for the penalty parameter that prevents any noise variable from entering the selected variable, for all $k > q$,

$$\lambda_* = \min_{\lambda \geq 0}\left\{\lambda | \hat{\beta}_k^\lambda = 0, \forall k > q\right\}.$$

Let L_* be the loss of the OLS estimator. It follows that

$$P\left(\inf_{\lambda,\phi,\omega} L(\lambda,\phi,\omega) > tn^{-1}\right) \leq P\left(L_* > tn^{-1}\right) + P(E^c).$$

We have $P(E^c) \to 0$ for $n \to \infty$. According to the properties of the OLS estimator,

$$\limsup_{n\to\infty} P\left(L_* > tn^{-1}\right) < \varepsilon,$$

which completes the proof. □

Appendix F. Appendix Proof of Theorem 3

Proof. For any $g > 0$, under $(\hat{\lambda},\hat{\phi},\hat{\omega})$, we obtain

$$P\left(L(\hat{\lambda},\hat{\phi},\hat{\omega}) > gn^{-1}\log^2 n\right) \leq 2\varepsilon.$$

Then, the loss function is

$$\begin{aligned} P\left(L(\hat{\lambda},\hat{\phi},\hat{\omega}) > gn^{-1}\log^2 n\right) &\leq P\left(L_{cv}(\hat{\lambda},\hat{\phi},\hat{\omega}) > gn^{-1}\log^2 n\right) \\ &\leq 2P\left(\sup|L(\hat{\lambda},\hat{\phi},\hat{\omega}) - L_{cv}(\hat{\lambda},\hat{\phi},\hat{\omega})| > \frac{1}{2}gn^{-1}\log^2 n\right) \\ &+ P\left(\inf L(\hat{\lambda},\hat{\phi},\hat{\omega}) > \frac{1}{2}gn^{-1}\log^2 n\right). \end{aligned}$$

By Lemma 2, for each $\varepsilon > 0$, there exists $g > 0$,

$$\limsup_{n\to\infty} P\left(L(\hat{\lambda},\hat{\phi},\hat{\omega}) > gn^{-1}\log^2 n\right) < \varepsilon,$$

which completes the proof. □

Appendix G. Appendix Proof of Theorem 4

Proof. According to Theorem 1 of Fu and Knight [18], we have $\beta\cdot 1_S \xrightarrow{p} \beta$.

Define $V_n(\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - x_i^T\{\beta\cdot 1_S\}\right)^2 + \frac{\phi\lambda_n}{n}\sum_{j=1}^{p}|\beta_j|$, note that

$$V_n(\hat{\beta}_n) \geq \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - x_i^T\beta_i\right)^2 = V_n^{(0)}(\hat{\beta}_n).$$

So $\arg\min\left(V_n^{(0)}(\hat{\beta}_n)\right) = O_p(1)$, also $V_n(\hat{\beta}_n) \geq V_n^{(0)}(\hat{\beta}_n)$, so

$$\arg\min\left(V_n^{(0)}(\hat{\beta}_n)\right) = \arg\min\left(V_n(\hat{\beta}_n)\right) = O_p(1).$$

We have $\hat{\beta}_n = O_p(1)$ and

$$V_n(\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i + x_i^T\left(\beta - \{\hat{\beta}_n\cdot 1_S\}\right)\right)^2 + \frac{\phi\lambda_n}{n}\sum_{j=1}^{p}|\beta_j|.$$

According to the point-by-point convergence principle and Lemma 3,

$$\lim_{n\to\infty} V(\hat{\beta}_n) = \frac{\phi\lambda_n}{n}\sum_{j=1}^{p}|\beta_j|,$$

then,

$$V_n(\hat{\beta}_n) = E\varepsilon_i^2 - 2\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_i^T(\{\hat{\beta}_n \cdot 1_S\} - \beta) + \lim_{n\to\infty} V(\hat{\beta}_n)$$
$$= \sigma^2 + V(\hat{\beta}_n),$$

so $\sup|V_n(\hat{\beta}_n) - V(\hat{\beta}_n) - \sigma^2| \xrightarrow{p} 0$. Then

$$\arg\min(V_n) \xrightarrow{p} \arg\min(V),$$

$$\hat{\beta}_n \xrightarrow{p} \beta,$$

which proves the consistency. □

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
2. Wang, S.; Weng, H.; Maleki, A. Which bridge estimator is optimal for variable selection? *arXiv* **2017**, arXiv:1705.08617.
3. Meinshausen, N.; Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **2006**, *34*, 1436–1462. [CrossRef]
4. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
5. Fan, J.; Li, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv* **2006**, arXiv:math/0602133. [CrossRef]
6. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
7. Fan, J.; Peng, H. Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **2004**, *32*, 928–961. [CrossRef]
8. Donoho, D.L.; Johnstone, J.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455. [CrossRef]
9. Breiman, L. Better subset regression using the nonnegative garrote. *Technometrics* **1995**, *37*, 373–384. [CrossRef]
10. Yuan, M.; Lin, Y. On the non-negative garrotte estimator. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2007**, *69*, 143–161. [CrossRef]
11. Meinshausen, N. Relaxed lasso. *Comput. Stat. Data Anal.* **2007**, *52*, 374–393. [CrossRef]
12. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv* **2017**, arXiv:1707.08692.
13. Mentch, L.; Zhou, S. Randomization as regularization: A degrees of freedom explanation for random forest success. *arXiv* **2019**, arXiv:1911.00190.
14. Bloise, F.; Brunori, P.; Piraino, P. Estimating intergenerational income mobility on sub-optimal data: A machine learning approach. *J. Econ. Inequal.* **2021**, *19*, 643–665. [CrossRef]
15. He, Y. The Analysis of Impact Factors of Foreign Investment Based on Relaxed Lasso. *J. Appl. Math. Phys.* **2017**, *5*, 693–699. [CrossRef]
16. Kang, C.; Huo, Y.; Xin, L.; Tian, B.; Yu, B. Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *J. Theor. Biol.* **2019**, *463*, 77–91. [CrossRef] [PubMed]
17. Tay, J.K.; Narasimhan, B.; Hastie, T. Elastic net regularization paths for all generalized linear models. *arXiv* **2021**, arXiv:2103.03475.
18. Fu, W.; Knight, K. Asymptotics for lasso-type estimators. *Ann. Stat.* **2000**, *28*, 1356–1378. [CrossRef]
19. Zhao, P.; Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
20. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499. [CrossRef]
21. Huang, J.; Ma, S.; Zhang, C.H. Adaptive Lasso for sparse high-dimensional regression models. *Stat. Sin.* **2008**, 1603–1618.
22. Franklin, J. The elements of statistical learning: Data mining, inference and prediction. *Math. Intell.* **2005**, *27*, 83–85. [CrossRef]
23. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Routledge: Oxfordshire, UK, 2019.
24. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 849–911. [CrossRef] [PubMed]
25. Li, R.; Zhong, W.; Zhu, L. Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **2012**, *107*, 1129–1139. [CrossRef]