Applied Intelligence (2019) 49:592–604
https://doi.org/10.1007/s10489-018-1295-y

E-ENDPP: a safe feature selection rule for speeding up Elastic Net

Yitian Xu[1] · Ying Tian[1] · Xianli Pan[1] · Hongmei Wang[2]

Published online: 7 September 2018

Abstract

Lasso is a popular regression model, which can do automatic variable selection and continuous shrinkage simultaneously. The Elastic Net is one of the corrective methods of Lasso, which selects groups of correlated variables. It is particularly useful when the number of features p is much bigger than the number of observations n. However, the training efficiency of the Elastic Net for high-dimensional data remains a challenge. Therefore, in this paper, we propose a new safe screening rule, i.e., E-ENDPP, for the Elastic Net problem which can identify the inactive features prior to training. Then, the inactive features or predictors can be removed to reduce the size of problem and accelerate the training speed. Since this E-ENDPP is derived from the optimality conditions of the model, it can be guaranteed in theory that E-ENDPP will give identical solutions with the original model. Simulation studies and real data examples show that our proposed E-ENDPP can substantially accelerate the training speed of the Elastic Net without affecting its accuracy.

Keywords Elastic Net · Lasso · Screening rule · Feature selection

1 Introduction

With the popularization of the internet and the rapid development of data collection methods, high-dimensional data becomes common in our real life, such as bioinformatics, online education and hyper-spectral data analysis. The rapid growth of data dimension brings a great challenge in terms of data analysis. That is to say, when we use some existing algorithms to handle these high-dimensional data directly, an important issue named "curse of dimensionality" will appear. It refers to the phenomenon that data becomes sparser in high-dimensional space, adversely affecting algorithms designed for low-dimensional space [1–4]. The learning algorithms tend to over-fitting with a lot of features and the performance of regression may be disappointed. What's more, for that "curse of dimensionality" is a more crucial issue for hyper-spectral data analysis and genetic data analysis, so how to overcome it in these domains becomes more and more important in machine learning.

So far, one of the important ways to deal with high-dimensional data is variable selection or feature selection. Different from the feature extraction techniques, such as factor analysis and so on [5, 6], feature selection methods find the important features from data rather than produce new features which are the linear combinations of existing features. One popular method of identifying important explanatory features is sparse regularization, which is among the embedded feature selection methods. The L_1 regularized least squares problem known as Lasso (Least absolute shrinkage and selection operator) [7] is widely used. By taking both empirical error and L_1-norm penalty into account, Lasso not only fits well but also obtains a sparse solution. That is to say, a majority of coefficients of related features are zero, and then these features have no influence on shaping the regression model. So it can do variable selection automatically. Besides, the computational complexity of Lasso has been proven to be the same as linear regression. Lasso has gained great success in many applicatons [8–11] and a lot of algorithms have been

✉ Yitian Xu
xytshuxue@126.com

[1] College of Science, China Agricultural University, Beijing, 100083, China

[2] College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China

proposed to solve the Lasso problem [12–17]. However, for large-scale problems, solving the Lasso problem becomes difficult because of the computational burden.

In 2010, Ghaoui et al. [18] first introduced a safe screening rule in the context of L_1 sparse regularization. This rule allows to eliminate features whose coefficients are zero at the optimum [19]. Thus, the problem will be greatly simplified because we can work on a reduced feature matrix when solving the Lasso problem, which can lead to substantial savings in computational cost and memory usage [19]. This rule has been extended to other models [20–22], including Lasso, group Lasso, logistic regression, and other l_1-regularized models [19]. Motivated by the work of Ghaoui, Enhanced Dual Polytope Projections method (EDPP) has been proposed in [23]. It can identify a large portion of redundant features before actually solving a problem. So the scale of Lasso problem can be cut down a lot to accelerate the solving process. The most important advantage of EDPP is its safety in the sense that no active features will be mistakenly discarded. And experimental results have shown that EDPP is more effective in identifying inactive features than existing state-of-the-art screening rules for Lasso.

Although the Lasso has shown success in many situations [8, 24], and many algorithms have been developed to efficiently solve the Lasso problem in recent years [12, 13], it has some limitations. So, the Elastic Net is introduced as a corrective method to do feature selection. Elastic Net maintains the advantages of Lasso. It can do continuous shrinkage and automatic variable selection simultaneously and it selects groups of correlated variables. It is like a stretchable fishing net that retains 'all the big fish' [25]. Elastic Net outperforms Lasso for that when $p > n$ (n observations), the Lasso can select at most n features (even when more are associated with the outcome) and it tends to select only one feature from any set of highly correlated variables. Additionally, even when $n > p$, if the variables are strongly correlated, ridge regression tends to perform better. But Elastic Net can potentially select all p predictors in all situations. Besides, studies and real data examples show that the Elastic Net often outperforms the Lasso in terms of prediction accuracy [25].

This paper is organized as follows: Section 2 reviews the Elastic Net problem. Screening rules via dual polytope projection are presented in Section 3. In Section 4, we apply the EDPP method to the Elastic Net problem to derive our screening rule for the Elastic Net(E-ENDPP). In Section 5, both synthetic data sets and real-world data sets are used to verify the efficiency of E-ENDPP. The last section is the conclusion drawn from this paper.

2 Related works

Suppose that we have an input vector $X = (x_1, x_2, \cdots, x_p)$ and want to predict a real-valued output y. The linear regression model has the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon.$$

Here, β_0 is a constant term and $(\beta_1, \cdots, \beta_p)$ are coefficients for each feature. There are many methods to estimate the vector of coefficients $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_p)$. As we know, the traditional ordinary least square (OLS) estimates, which are obtained by minimizing the residual sum of squares, have many good properties, but they often do poorly in both prediction and interpretation.

Penalization techniques have been proposed to improve OLS [26]. For example, ridge regression [27] shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares:

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin} \left\{ (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \quad (1)$$

where β is the unknown regression coefficient. $\lambda > 0$ is a parameter. $X \in R^{n \times p}$ represents a matrix of all samples, where n and p are the number of samples and features, respectively. $y \in R^n$ is a real-value output vector for all samples. However, ridge regression cannot produce a sparse model, for it always keeps all the predictors in the model. In contrast, the best subset selection produces a sparse model, but it is extremely variable because of its inherent discreteness, as addressed by Breiman(1996) [28].

Another technique called the Lasso was proposed by Tibshirani (1996) [7]. It is a penalized least squares method by imposing an L_1 norm penalty on the regression coefficients:

$$\min_{\beta \in R^p} \quad \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \quad (2)$$

where $\lambda > 0$ is also a non-negative parameter, and all the other notations are the same as mentioned in (1). $\hat{\beta}^{Lasso}$ is the optimal solution of (2). Because of the introduction of L_1 norm, the Lasso does continuous shrinkage and automatic variable selection simultaneously.

2.1 Elastic Net problem

In 2005, Zou and Hastie [25] introduced the Elastic Net penalty. For notational convenience, we restate the definition of the Elastic Net problem of regression.

Definition 1 Suppose that we have n observations with p features. Let y be the n dimensional response vector and $X = [x_1, x_2, \cdots, x_p]$ denotes the $n \times p$ feature matrix. For any fixed non-negative γ and λ, we define the Elastic Net problem as follows:

$$\min_{\beta\in R^p} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\gamma}{2}\|\beta\|_2^2 + \lambda\|\beta\|_1, \tag{3}$$

where $\|\beta\|_2^2 = \sum_{i=1}^p \beta_i^2$ and $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$.

Note that this definition differs from the original form of the naive Elastic Net in Zou and Hastie [25] by the factors of $\frac{1}{2}$ (just for notational convenience). If $\lambda = 0$, the Elastic Net becomes ridge regression, and if $\gamma = 0$, it turns into the Lasso problem. So the Elastic Net is a convex combination of the Lasso and ridge penalty.

In fact, Elastic Net can be transformed into a simplified form which is similar to Lasso problem (2) by a series of derivation. The conclusion is given as follows:

Lemma 1 *Given data set X and the corresponding output vector y, define the augmented vector $\bar{Y}$ and augmented matrix $\bar{X}$ by*

$$\bar{Y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \bar{X} = \begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}.$$

Then, the formula (3) can be written as:

$$\min_{\beta\in R^p} \frac{1}{2}\|\bar{Y} - \bar{X}\beta\|_2^2 + \lambda\|\beta\|_1. \tag{4}$$

For brevity, the proof of Lemma 1 is displayed in Appendix A. Lemma 1 shows that we can transform the Elastic Net problem into an equivalent Lasso problem on augmented data. Note that the sample size in the augmented problem is $n + p$ and $\bar{X}$ has rank p, which means that the Elastic Net can potentially select all p predictors in all situations. This important property overcomes the limitations of the Lasso. Lemma 1 also shows that the Elastic Net can perform an automatic variable selection similar to the Lasso.

2.2 The solution of the Elastic Net problem

By Lemma 1, we develop a method to solve the Elastic Net problem. Let $L(\lambda, \beta) = \frac{1}{2}\|\bar{Y} - \bar{X}\beta\|_2^2 + \lambda\|\beta\|_1$, minimizing $L(\lambda, \beta)$ is equivalent to a Lasso optimization problem. Thus we can get the solution of the Elastic Net problem $\hat{\beta}^{ENnaive}$ by minimizing $L(\lambda, \beta)$.

$$\begin{aligned}\hat{\beta}^{ENnaive} &= \underset{\beta\in R^p}{argmin} L(\lambda, \beta) = (\bar{X}^T\bar{X})^{-1}(\bar{X}^T\bar{Y} - \lambda\frac{\partial\|\beta\|_1}{\partial\beta}) \\ &= (X^TX + \gamma I)^{-1}(X^Ty - \lambda\frac{\partial\|\beta\|_1}{\partial\beta}).\end{aligned}$$

Here, γ and λ are non-negative parameters chosen a priori. In the case of orthogonal columns of X, the solution of the Elastic Net with parameters (γ, β) is

$$\hat{\beta}_i^{ENnaive} = \frac{(|\hat{\beta}_i^{LS}| - \lambda)_+}{1+\gamma} sgn(\hat{\beta}_i^{LS}),$$

where $\hat{\beta}^{LS} = X^Ty$ is the ordinary least square estimator and $(Z)_+$ denotes the positive part, i.e., $(Z)_+ = Z$ if $Z > 0$ and $(Z)_+ = 0$ otherwise. And $sgn(x)$ is the classical sign function as follows:

$$sgn(x) = \begin{cases} +1, \text{if } x > 0, \\ 0, \text{if } x = 0, \\ -1, \text{otherwise.} \end{cases}$$

Also in the case of orthogonal columns of X, the solution of the ridge regression with parameter γ is

$$\hat{\beta}^{ridge} = \frac{(\hat{\beta}^{LS})}{1+\gamma},$$

and the Lasso solution is

$$\hat{\beta}_i^{Lasso} = (|\hat{\beta}_i^{LS}| - \lambda)_+ sgn(\hat{\beta}_i^{LS}).$$

We can see that the Elastic Net estimator is a two-stage procedure: for each fixed γ, we first get the ridge regression coefficients, and then do the Lasso shrinkage along the Lasso coefficient solution paths. Compared with the pure Lasso or ridge shrinkage, the Elastic Net appears to do double shrinkage. Double shrinkage does not help to reduce the variance a lot and introduces unnecessary extra bias. So, we correct the double shrinkage to improve the prediction performance of the Elastic Net. The corrected Elastic Net estimates are defined by

$$\hat{\beta}^{EN} = (1+\gamma)\hat{\beta}^{ENnaive}.$$

So, in Section 3, all the Elastic Net estimators are the corrected estimator.

2.3 The dual problem of the Elastic Net problem

In this subsection, we directly present the dual problems for formula (2) and formula (4). Here, formula (2) and formula (4) are corresponding to Lasso problem and Elastic Net, respectively. Research in other paper has shown that the dual problem of (2) is as follows:

Lemma 2 *The dual formulation of original Lasso formula (2) can be derived as follows:*

$$\begin{aligned} \max_{\theta} \quad & \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{y}{\lambda}\|_2^2, \\ s.t. \quad & |X_i^T\theta| \le 1 \quad i = 1, 2, \cdots p, \end{aligned} \tag{5}$$

where θ is the dual variable, λ is a parameter chosen a priori.

For the sake of simplicity, we omit the proof here. One can refer to Appendix B in this paper to get more information about the proof of Lemma 2. Note that, let $\bar{Y} = \begin{pmatrix} y \\ 0 \end{pmatrix}$, $\bar{X} = \begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}$ and $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$, then Elastic Net (3) can be rewritten as formula (4), which is very similar with Lasso problem (2). And then dual formulation of Elastic Net (4) can also be derived in a similar way as follows:

Theorem 1 *The dual formulation of formula (4) is:*

$$\begin{aligned} &\underset{\theta_1,\theta_2}{max} \quad \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}(\|\theta_1 - \frac{y}{\lambda}\|_2^2 + \|\theta_2\|_2^2), && (6)\\ &s.t. \quad \mid X_i^T\theta_1 + \sqrt{\gamma}(\theta_2)_i \mid \leq 1 \quad i = 1, 2, \cdots p, \end{aligned}$$

where $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ *is the dual variable,* λ *and* γ *are parameters chosen a priori.*

2.4 Karush-Kuhn-Tucker (KKT) conditions of the dual problem

For notational convenience, let the optimal solution of problem (5) be $\theta^*(\gamma, \lambda)$, and the optimal solution of problem (2) with parameters γ and λ is denoted by $\beta^*(\gamma, \lambda)$. Then, the KKT conditions are given by

$$y = X\beta^*(\gamma, \lambda) + \lambda\theta^*(\gamma, \lambda), \tag{7}$$

$$\begin{aligned} &X_i^T\theta^*(\gamma, \lambda) \in \begin{cases} sign([\beta^*(\gamma, \lambda)]_i), & if\ [\beta^*(\gamma, \lambda)]_i \neq 0 \\ [-1, 1], & if\ [\beta^*(\gamma, \lambda)]_i = 0 \end{cases} \\ &i = 1, 2, \cdots, p, \end{aligned} \tag{8}$$

where $[Z]_k$ denotes the k^{th} component of Z. From the KKT condition (8), we have

$$\begin{aligned} |X_i^T\theta^*(\gamma, \lambda)| < 1 \quad &\Rightarrow [\beta^*(\gamma, \lambda)]_i = 0 \\ &\Rightarrow X_i\ is\ an\ inactive\ feature. \end{aligned} \tag{9}$$

That is to say, we can use (9) to identify the inactive features for the Elastic Net problem. However, since $\theta^*(\gamma, \lambda)$ is usually unknown, we cannot directly apply (9) to identify the inactive features. Motivated by the Safe Feature Elimination (SAFE) rules [18], we can first find a region Θ which contains $\theta^*(\gamma, \lambda)$. Then (9) can be relaxed as follows:

$$\begin{aligned} \underset{\theta\in\Theta}{sup}|X_i^T\theta| < 1 \quad &\Rightarrow [\beta^*(\gamma, \lambda)]_i = 0 \\ &\Rightarrow X_i\ is\ an\ inactive\ feature. \end{aligned} \tag{10}$$

By geometric properties, we know the solution of the dual problem (5) is the projection of $\frac{y}{\lambda}$ onto the feasible set F of problem (5), which is clearly a closed and convex polytope [23].

3 Screening rules via dual polytope projection (DPP)

For Lasso problem, there are several screening rules such as SAFE, spherical dome region test (DOME), strong rule and EDPP. In this paper, we mainly discuss the method of EDPP [23] and apply it to Elastic Net problem to draw our screening rule. In 2015, Wang et al. [23] proposed basic Dual Polytope Projections (DPP) screening rule and two improved methods for discarding the inactive features for Lasso, and a more effective EDPP rule. In this section, we focus on reviewing DPP and EDPP rules which have been proposed in [23]. And some of the related theoretical basis, such as non-expansiveness in [29], is also given in this section. One can refer to [23] to get more detailed information.

Lemma 3 *Let C be a nonempty closed convex subset of Hilbert space $\mathcal{H}$. Then the projection operator is continuous and nonexpansive, i.e.,*

$$\|P_C(w_2) - P_C(w_1)\|_2 \leq \|w_2 - w_1\|_2, \forall w_1, w_2 \in \mathcal{H},$$

where the projection operator is defined as $P_C(w) = \underset{u\in C}{argmin}\|u - w\|_2$.

According to [23], the dual optimal solution $\theta^*(\gamma, \lambda)$ of (5) can be expressed by

$$\theta^*(\gamma, \lambda) = P_F(y/\lambda) = \underset{\theta\in F}{argmin}\|\theta - \frac{y}{\lambda}\|_2.$$

Here, F is the feasible region of formula (5). For that $\theta^*(\gamma, \lambda)$ can not be obtained before actually solving the above problem, so a spherical region which contains $\theta^*(\gamma, \lambda)$ should be found as follows to construct DPP rule by using Lemma 3 directly.

Corollary 1 *For dual Lasso formulation (5), let $\lambda, \lambda_0 > 0$ be two regularization parameters. Then,*

$$\left\|\theta^*(\gamma, \lambda) - \theta^*(\gamma, \lambda_0)\right\|_2 \leq \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2. \tag{11}$$

Corollary 1 implies that the dual solution must be inside a ball centered at $\theta^*\left(\gamma, \lambda_0\right)$ with radius $\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2$, i.e.,

$$\theta^*(\gamma, \lambda) \in B\left(\theta^*\left(\gamma, \lambda_0\right), \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2\right). \tag{12}$$

Based on Corollary 1, the following Corollary 2 for Lasso problem, i.e., DPP rule in [23] could be obtained.

Corollary 2 ***(DPP):*** *Let $\lambda_{max} = max|X_i^T y|$, if $\lambda \geq \lambda_{max}$, then $[\beta^*(\gamma, \lambda)]_i = 0, \forall i \in I$. Otherwise, $[\beta^*(\gamma, \lambda)]_i = 0$, if*

$$\left|X_i^T \frac{y}{\lambda_{max}}\right| < 1 - \left(\frac{1}{\lambda} - \frac{1}{\lambda_{max}}\right) \|X_i\|_2 \|y\|_2. \qquad (13)$$

Wang et al. [23] also proposed two better methods to improve the estimation of the dual optimal solution. In the following, we briefly describe these two improved methods and the related theoretical basis.

Lemma 4 *For Lasso formulation (5), assume that the dual optimal solution $\theta^*(.)$ at γ and $\lambda_0 \in (0, \lambda_{max}]$ is known. Then, for each $\lambda \in (0, \lambda_0]$, we have $[\beta^*(\gamma, \lambda)]_i = 0$, if*

$$\left|X_i^T \theta^*(\gamma, \lambda_0)\right| < 1 - \|V_2^{\perp}(\lambda, \lambda_0)\|_2 \|X_i\|_2,$$

where

$$V_2^{\perp}(\lambda, \lambda_0) = V_2(\lambda, \lambda_0) - \frac{< V_1(\lambda_0), V_2(\lambda, \lambda_0) >}{\| V_1(\lambda_0) \|} V_1(\lambda_0),$$

$$V_1(\lambda_0) = \begin{cases} \frac{y}{\lambda_0} - \theta^*(\gamma, \lambda_0), & if\ \lambda_0 \in (0, \lambda_{max}), \\ sign(X_*^T y) X_*, & if\ \lambda_0 = \lambda_{max}, \end{cases}$$
$$where\ X_* = \underset{X_i}{argmax}|X_i^T y|.$$

and

$$V_2(\lambda, \lambda_0) = \frac{y}{\lambda} - \theta^*(\gamma, \lambda_0).$$

And it has been proved in [23] that Lemma 4 is better than Corollary 2 in feature selection.

Another approach of improving the feature screening ability of DPP is to use the so called firmly nonexpansiveness of the projections onto closed convex subset of a Hilbert space. It is displayed as follows:

Lemma 5 *(Bauschke and Combettes, 2011 [30]) Let C be a nonempty closed convex subset of Hilbert space $\mathcal{H}$. Then the projection operator is continuous and firmly nonexpansive. In other words, for any $w_1, w_2 \in \mathcal{H}$, we have*

$$\|P_C(w_2) - P_C(w_1)\|_2^2 + \|(Id - Pc)(w_2) - (Id - Pc)(w_1)\|_2^2 \leq \|w_2 - w_1\|_2^2,$$

where Id is the identity operator.

According to Lemma 5, we will obtain a much smaller spherical region which contains optimal solution $\theta^*(\gamma, \lambda_0)$ compared with original DPP in the following corollary,

Corollary 3 *For the dual formulation (5) of Lasso, λ and λ_0 are two parameters chosen a priori. Then*

$$\left\|\theta^*(\gamma, \lambda) - \left(\theta^*(\gamma, \lambda_0) + \frac{1}{2}(\frac{1}{\lambda} - \frac{1}{\lambda_0})y\right)\right\|_2 \leq \frac{1}{2}\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2, \qquad (14)$$

i.e. the dual solution must be inside a ball centered at $\theta^*(\gamma, \lambda_0) + \frac{1}{2}(\frac{1}{\lambda} - \frac{1}{\lambda_0})y$ with radius $\frac{1}{2}\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2$. And we have

$$\theta^*(\gamma, \lambda) \in B\left(\theta^*(\gamma, \lambda_0) + \frac{1}{2}(\frac{1}{\lambda} - \frac{1}{\lambda_0})y, \frac{1}{2}\left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2\right) \subset B\left(\theta^*(\gamma, \lambda_0), \left|\frac{1}{\lambda} - \frac{1}{\lambda_0}\right| \|y\|_2\right). \qquad (15)$$

In view of Corollary 3, the more effective feature screening rule compared with DPP will be obtained as follows:

Corollary 4 *For dual Lasso formula (5), assume that the dual optimal solution $\theta^*(.)$ at $\lambda_0 \in (0, \lambda_{max}]$ is known. Then, for each $\lambda \in (0, \lambda_0]$, we have $[\beta^*(\gamma, \lambda)]_i = 0$, if*

$$\left|X_i^T \left(\theta^*(\gamma, \lambda_0) + \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)y\right)\right| < 1 - \frac{1}{2}\left(\frac{1}{\lambda} - \frac{1}{\lambda_0}\right)\|y\|_2 \|X_i\|_2.$$

By combining above two approaches, we can further derive a much smaller spherical region which contains dual optimal solution and obtain a more effective feature screening rule compared with all the above rules. The region and the rule are displayed in the following lemma and corollary, respectively.

Lemma 6 *For the dual formulation (5) of Lasso, let λ and λ_0 be two parameter values. Then*

$$\left\|\theta^*(\lambda) - \left(\theta^*(\lambda_0) + \frac{1}{2}V_2^{\perp}(\lambda, \lambda_0)\right)\right\|_2 \leq \frac{1}{2}\|V_2^{\perp}(\lambda, \lambda_0)\|_2,$$

i.e. the dual solution must be inside a ball centered at $\theta^*(\lambda_0) + \frac{1}{2}V_2^{\perp}(\lambda, \lambda_0)$ with radius $\frac{1}{2}\|V_2^{\perp}(\lambda, \lambda_0)\|_2$,

$$\theta^*(\lambda) \in B\left(\theta^*(\lambda_0) + \frac{1}{2}V_2^{\perp}(\lambda, \lambda_0), \frac{1}{2}\|V_2^{\perp}(\lambda, \lambda_0)\|_2\right). \qquad (16)$$

In view of Lemma 6, we can get the following conclusion.

Corollary 5 ***(EDPP):*** *For the Lasso formula (5), assume that the dual optimal solution $\theta^*(.)$ at $\lambda_0 \in (0, \lambda_{max}]$ is known. Then, for each $\lambda \in (0, \lambda_0]$, we have $[\beta^*(\gamma, \lambda)]_i = 0$, if*

$$\left|X_i^T \left(\theta^*(\gamma, \lambda_0) + \frac{1}{2}V_2^{\perp}(\lambda, \lambda_0)\right)\right| < 1 - \frac{1}{2}\|V_2^{\perp}(\lambda, \lambda_0)\|_2 \|X_i\|_2.$$

4 E-ENDPP for the Elastic Net problem

To the best of our knowledge, DPP rules for the Elastic Net problem are not yet available. In this section, we propose our feature selection approach for the Elastic Net problem, thus we draw our screening rules for the Elastic Net. All the rules include: ENDPP (Dual Polytope Projections for Elastic Net problem), sequential ENDPP, E-ENDPP (Enhanced ENDPP) and sequential E-ENDPP. Among them, E-ENDPP and sequential E-ENDPP are based on Improvement 1 and Improvement 2 which are also proposed in this section.

According to Lemma 1, for that $\bar{Y} = \begin{pmatrix} y \\ 0 \end{pmatrix}$, $\bar{X} = \begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}$, so $|\bar{X}_i^T \bar{Y}| = |x_i^T y|$, $\|\bar{X}_i\|_2 = \sqrt{\|x_i\|_2^2 + \gamma}$, $\|\bar{Y}\|_2 = \|y\|_2$.

Substituting these equalities into (13), we can easily obtain the first feature screening rule for Elastic Net in the following:

Theorem 2 ***(ENDPP):*** *For the formula* (4) *of Elastic Net problem and a given* γ*, if* $\lambda > max \mid x_i^T y \mid$*, then* $[\beta^*(\gamma, \lambda)]_i = 0, \forall i \in I$*. Otherwise, if*

$$\frac{\mid x_i^T y \mid}{max \mid x_i^T y \mid} < 1 - \left(\frac{1}{\lambda} - \frac{1}{max \mid x_i^T y \mid} \right) \|y\|_2 \sqrt{\|x_i\|_2^2 + \gamma}, \tag{17}$$

then $[\beta^*(\gamma, \lambda)]_i = 0$.

Let $\lambda_{max} = max \mid x_i^T y \mid$*, then* (17) *can be written as:*

$$\mid x_i^T y \mid < \lambda_{max} - \lambda_{max} \left(\frac{1}{\lambda} - \frac{1}{\lambda_{max}} \right) \|y\|_2 \sqrt{\|x_i\|_2^2 + \gamma}. \tag{18}$$

Notice that, ENDPP is different from the following ENSAFE rule (i.e., Safe Feature Elimination rule for Elastic Net), which discards the predictor i if

$$\mid x_i^T y \mid < \lambda - \frac{\lambda_{max} - \lambda}{\lambda_{max}} \|y\|_2 \sqrt{\|x_i\|_2^2 + \gamma}. \tag{19}$$

Note that, ENSAFE means applying SAFE rule for Lasso problem into Elastic Net to get a new feature screening for Elastic Net. The radius of inequality (19) is the same as that of (18), and their centers are y/λ and y/λ_{max}, which are different.

However, we should solve a series of Elastic Net problems with different λ, so we further embed our ENDPP into parameter tuning process, termed as Sequential ENDPP, to accelerate the whole solving process.

Theorem 3 ***(Sequential ENDPP)*** *For the simplified formulation* (4) *of Elastic Net problem, fixing* γ*, suppose we are given a sequence of parameter values* $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$*. Then for any integer* k $(0 \leq k \leq m)$*, we have* $\beta^*(\gamma, \lambda_{k+1}) = 0$*, if* $\beta^*(\gamma, \lambda_k)$ *satisfies the following inequality:*

$$\frac{1}{\lambda_k} |x_i^T \left(y - X\beta^*(\gamma, \lambda_k) \right) - \gamma \beta_i^*(\gamma, \lambda_k)|$$
$$< 1 - \left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k} \right) \sqrt{\|x_i\|_2^2 + \gamma} \ \ \|y\|_2. \tag{20}$$

To improve the feature screening ability of ENDPP, we further substitute $|\bar{X}_i^T \bar{Y}| = |x_i^T y|$, $\|\bar{X}_i\|_2 = \sqrt{\|x_i\|_2^2 + \gamma}$, $\|\bar{Y}\|_2 = \|y\|_2$ in Lemma 4 and Corollary 3 to obtain the following two better feature screening rules for Elastic Net problem, termed as Improvement 1 and Improvement 2.

Improvement 1 *For the Elastic Net formula* (4)*, fixing* γ*, suppose we are given a sequence of parameter values* $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$*. Then for any integer* k $(0 \leq k \leq m)$*, we have* $\beta^*(\gamma, \lambda_{k+1}) = 0$ *, if* $\beta^*(\gamma, \lambda_k)$ *satisfies the following inequality:*

$$\left| \bar{X}_i^T \frac{\bar{Y} - \bar{X}\beta^*(\gamma, \lambda_k)}{\lambda_k} \right| < 1 - \|V_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2 \|\bar{X}_i\|_2^2. \tag{21}$$

where

$$V_2^{\perp}(\lambda_{k+1}, \lambda_k) = V_2(\lambda_{k+1}, \lambda_k) - \frac{< V_1(\lambda_k), V_2(\lambda_{k+1}, \lambda_k) >}{\| V_1(\lambda_k) \|} V_1(\lambda_k),$$

$$V_1(\lambda_k) = \begin{cases} \frac{\bar{Y}}{\lambda_k} - \theta^*(\gamma, \lambda_k), & if \ \lambda_k \in (0, \lambda_{max}), \\ sign(X_*^T y) X_*, & if \ \lambda_k = \lambda_{max}, \end{cases}$$
$$where \ X_* = \underset{X_i}{argmax} |X_i^T y|.$$

Improvement 2 *For the simplified formulation* (4) *of Elastic Net problem, fixing* γ*, suppose we are given a sequence of parameter values* $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$*. Then for any integer* k $(0 \leq k \leq m)$*, we have* $\beta^*(\gamma, \lambda_{k+1}) = 0$ *, if* $\beta^*(\gamma, \lambda_k)$ *satisfies the following inequality:*

$$\left| \bar{X}_i^T \left(\frac{\bar{Y} - \bar{X}\beta^*(\gamma, \lambda_k)}{\lambda_k} + \frac{1}{2}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)\bar{Y} \right) \right|$$
$$< 1 - \frac{1}{2}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)\|\bar{Y}\|_2 \|\bar{X}_i\|_2. \tag{22}$$

Similarly, we can further substitute $|\bar{X}_i^T \bar{Y}| = |x_i^T y|$, $\|\bar{X}_i\|_2 = \sqrt{\|x_i\|_2^2 + \gamma}$, $\|\bar{Y}\|_2 = \|y\|_2$ in Corollary 5 to improve the estimation of the dual optimal solution and derive the more effective screening rules: Enhanced ENDPP (E-ENDPP) and Sequential E-ENDPP.

Theorem 4 *(E-ENDPP): For the formulation (4) of Elastic Net problem, fixing γ, assume that the dual optimal solution $\theta^*(\gamma, \lambda_0)$ at $\lambda_0 \in (0, \lambda_{max}]$ is known. Then, for each $\lambda \in (0, \lambda_0]$, we have $[\beta^*(\gamma, \lambda)]_i = 0$, if*

$$\left|\bar{X}_i^T\left(\theta^*(\lambda_0) + \frac{1}{2}V_2^{\perp}(\lambda, \lambda_0)\right)\right| < 1 - \frac{1}{2}\|V_2^{\perp}(\lambda, \lambda_0)\|_2\|\bar{X}_i\|_2.$$

According to Theorem 4, once we know the optimal dual solution $\theta^*(\gamma, \lambda_0)$, we can get a new screening rule by setting $\lambda = \lambda_1$ to identify inactive features for the Elastic Net problem (4). By repeating the above process, we can obtain the sequential version of the ENDPP rule as in the following theorem.

Theorem 5 *(Sequential E-ENDPP): For the formulation (4) of Elastic Net problem, fixing γ, suppose we are given a sequence of parameter values $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$, we have $\beta^*(\gamma, \lambda_{k+1}) = 0$, if $\beta^*(\gamma, \lambda_k)$ satisfies the following inequality:*

$$\left|\bar{X}_i^T\left(\frac{\bar{Y} - \bar{X}\beta^*(\gamma, \lambda_k)}{\lambda_k} + \frac{1}{2}V_2^{\perp}(\lambda_{k+1}, \lambda_k\right)\right| < 1 - \frac{1}{2}\|V_2^{\perp}(\lambda_{k+1}, \lambda_k)\|_2\sqrt{\|x_i\|_2^2 + \gamma}. \quad (23)$$

In general, the pseudo code of the sequential enhanced ENDPP is given in Algorithm 1.

Algorithm 1 Sequential Enhanced ENDPP

```
Input: Training data $(X, y)$, parameters $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$ and $\gamma$
Output: Solution $\beta^*(\lambda_i), i = 0, 1, 2, \ldots, m$
$\beta^*(\lambda_0) \leftarrow 0$
for k=0: m-1 do
    $\Gamma = []$
    for i=1: do
        if $x_i$ satisfies (23) then
            $\beta^*(\lambda_{k+1})_i = 0$
            $\Gamma = \Gamma \bigcup i$
            $X \leftarrow X/X(:, i)$
        end if
    end for
    $\beta^*(\lambda_{k+1})_\Gamma \leftarrow$ solving Lasso with $(X, y)$
    return $\beta^*(\lambda_{k+1})$
end for
```

5 Computational complexity analysis

For the Elastic Net model, with the transformation in Lemma 1, we have $\bar{X} \in R^{(n+p)\times p}$ and $\bar{Y} \in R^{(n+p)\times 1}$. Suppose parameter γ is fixed and λ has m values for parameter selection. Suppose LARS is used to solve this model, then the computational complexity of the Elastic Net is $O(p^3 + p^2(n+p))$ [12]. To solve it with a sequence of λ, the overall computational complexity is $O(mp^3 + mp^2(n+p))$.

Now, we analyze the whole complexity of applying our E-ENDPP before solving the model. From (23) we can see that the calculation of the E-ENDPP rule itself is determined by the term on the left of the inequality, it is $O(p(n+p))$. For all parameters, the overall time complexity of the proposed E-ENDPP is $O(mp(n+p))$. Compared with the calculation of the model, this is very small and can be ignored. After applying the rule, most inactive features in $\bar{X}$ will be deleted and the calculation of solving the Elastic Net will be substantially reduced. Therefore, the computational complexity of our proposed E-ENDPP is much smaller than the original model without any screening rules.

6 Numerical experiments

In this section, to verify the validity of our algorithm E-ENDPP, we conduct the numerical experiments on one artificial data set and four benchmark data sets. We compare the E-ENDPP with other efficient screening algorithms. All experiments are carried out in MATLAB R2014a on Windows 7 running on a PC with system configuration Intel(R) Core(TM) i5-4590 CPU(3.30 GHz) with 8.00 GB of RAM.

6.1 Artificial data set

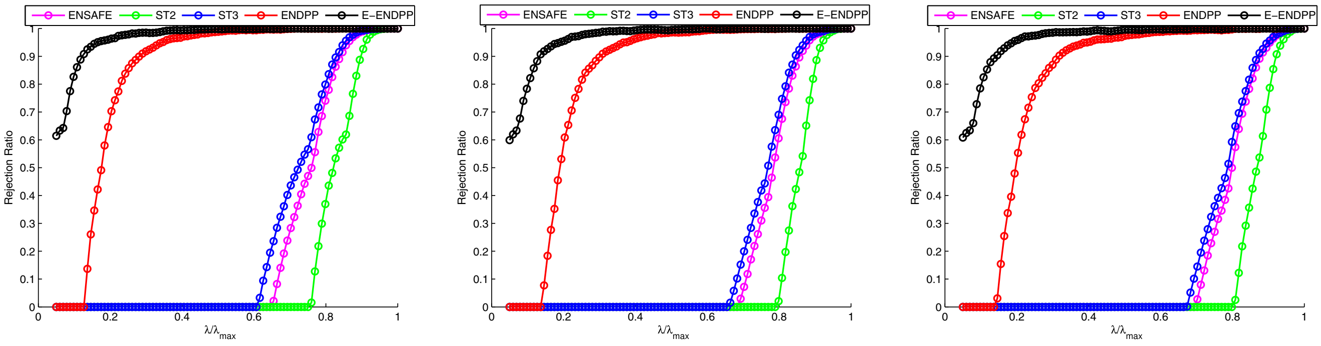
In this experiment, we want to evaluate the acceleration performance of the safe screening rules and study the influence of different parameter values.

We simulate data from the model

$$y = x\beta + \sigma\varepsilon, \ \varepsilon \sim N(0, 1).$$

The data matrix of X is 50 × 1000 entry and drawn from i.i.d standard Gaussian, and σ is set to be 0.1. We randomly choose $\overline{p}$ components of β from a uniform [-1,1] distribution, and set the remaining ones as 0. This has been commonly used in the sparse learning literature [5].

After we generate the data matrix X and the response vector y, we run the solver with or without screening rules to solve the Elastic Net problems along a sequence of 100 parameter values equally spaced on the λ/λ_{max} scale from 0.05 to 1.0. Another parameter γ is set as [0.01, 1, 100]. For feature screening, E-ENDPP is compared with the efficient ENSAFE, ENDPP, ST2 and ST3.



(a) $\bar{p} = 10$ (b) $\bar{p} = 100$ (c) $\bar{p} = 500$

Fig. 1 Rejection ratios of ENSAFE, ST2, ST3, ENDPP and E-ENDPP on three synthetic data sets

Running time of solving the Elastic Net problem along a sequence of 100 tuning parameter values equally spaced on the scale λ/λ_{max} from 0.05 to 1 by: the solver (Liu et al., 2009) (reported in the third column) without screening; the solver combined with different screening methods (reported from the fourth to the eighth columns, ST2 and ST3 are from literature [21]). By comparison, we can see that in three cases, E-ENDPP is always the fastest rule and saves the most computational time. ENDPP follows.

To further observe the performance of five compared screening rules with different λ, the rejection ratios (the proposition of identified zero components in the solution compared with the real zero components) are displayed along different λ/λ_{max} in Fig. 1. These rejection ratios are the average values of different γ. From Fig. 1, we can find that the efficiency of five methods is $E\text{-}ENDPP > ENDPP > ST3 > ENSAFE > ST2$, which is identical with the results in Table 1. Furthermore, we can find that screening rules give better performance with large λ. The reason may lies in that when λ is small, the solution of the model will be less sparse. As for another parameter γ, we can see from Table 1 that our proposed E-ENDPP makes similar acceleration with different γ. Thus, E-ENDPP keeps stable for different γ.

6.2 Experiments on benchmark data sets

To verify the efficiency and safety of our proposed E-ENDPP, we compare its performance with the efficient ENSAFE, ENDPP, ST2, ST3, BLITZ [31] and Laplacian Score (LS) [32] on four real data sets. BLITZ is a principled meta-algorithm, which is also regarded as a working set method, for scaling sparse optimization. LS is a classical feature selection method based on Laplacian Eigenmaps and Locality Preserving Projection. Statistics of the four data sets are given in Table 2. Details of them are as follows:
Pyrimidibes and **Triazines** are two regression data sets from LIBSVM.

Pulmon is collected using a chemical sensing system based on an array of 16 metal-oxide gas sensors and an external mechanical ventilator to simulate the biological respiration cycle. The primary data contains two responses, where

Table 1 Time comparison (seconds) of six algorithms on the artificial data sets

γ	$\bar{p}$	solver	ENSAFE	ST2	ST3	ENDPP	E-ENDPP
	10	94.20	79.62	81.81	74.69	14.83	1.78
0.1	100	249.09	227.16	245.26	220.41	45.61	1.81
	500	278.84	242.21	254.75	233.24	47.01	2.86
	10	93.20	80.06	83.39	75.91	16.03	1.84
1	100	268.65	218.71	244.18	218.85	44.37	2.30
	500	271.71	238.12	252.46	228.61	47.41	2.92
	10	39.65	26.49	27.17	23.98	6.42	3.31
100	100	80.32	68.65	72.39	69.81	15.13	6.57
	500	83.40	73.55	31.01	27.96	17.08	8.00

Table 2 Statistics of four real data sets

Data sets	$^\#$Samples	$^\#$Features	Data type	Source
Pyrimidibes	74	27	regression	LIBSVM
Triazines	186	60	regression	LIBSVM
Riboflavin	71	4088	regression	[3][1]
Pulmon	58	120432	regression	UCI

[1] https://www.annualreviews.org/doi/suppl/10.1146/annurev-statistics-022513-115545

the predictors were the features extracted from the sensor signals and the responses were the concentrations of two analytes, acetone and ethanol. Here, we randomly choose one response to produce a regression data set.

Riboflavin This data set is about riboflavin (vitamin B2) production in Bacillus subtilis. The log-transformed riboflavin production rate is the single real-valued response variable, and there are $p = 4088$ (co)variables measuring the logarithm of the expression level of 4088 genes. There are 71 samples in total.

In this experiment, parameter λ is set as a sequence of 100 values equally spaced on the λ/λ_{max} scale from 0.05 to 1.0. This range is commonly used in related papers. Another parameter γ is set empirically as [0.01, 1, 100].

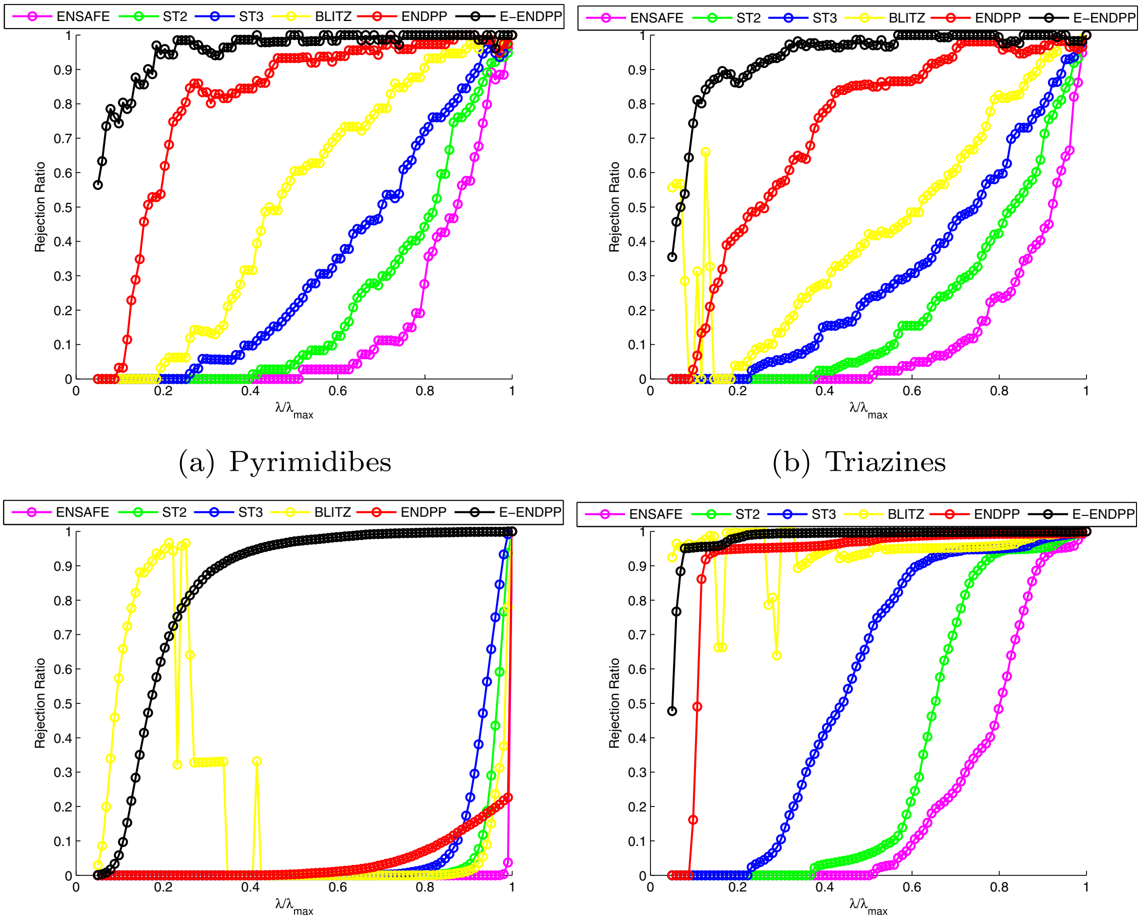
For each data set and for each parameter γ, running time of eight methods for solving the Elastic Net problems with 100 tuning parameter values is given in Table 3. Results of the LS method on the Pulmon data set is missing since the training time is over 24 hours which is much longer than other methods. The reason is that calculating k nearest neighbours in LS is extremely time-consuming for Pulmon. The best mean squared error (MSE) of all methods for each data set is also displayed. The eight methods are the solver (Liu et al.,2009) (reported in the third and fourth columns) without screening and the solver combined with different screening methods (reported from the fifth to the last columns).

Obviously, for small data sets with low dimension, i.e., Pyrimidibes and Triazines, the training efficiencies of all screening rules are not so good. The LS method gives the best performance on training time. However, its corresponding MSE is larger than other safe screening rules. For high-dimensional data sets, i.e., Riboflavin and Pulmon, ENDPP and E-ENDPP are more efficient than other methods. LS becomes much slow since it needs to calculate the KNNs of each training sample. This is time consuming for high-dimensional data. Also, its corresponding MSEs are not stable which may affect the performance of Lasso. Our proposed enhanced ENDPP (E-ENDPP) is the most effective rule in accelerating the computation of the Elastic Net especially for large-scale problems. As for the safety of E-ENDPP, it gives the same MSEs as Lasso for all data sets. This is identical with our theoretical analysis. Therefore, E-ENDPP is verified to be safe.

In addition, the average rejection ratios of each safe screening rule with different γ on four data sets are shown in Fig. 2. From it we can see that the rejection ratios of these methods are $E\text{-}ENDPP > ENDPP > BLITZ > ST3 > ST2 > ENSAFE$, which is identical with the results in Table 3. Compared with other four rules, E-ENDPP is able to identify far more inactive features and leads to much higher speedup.

Table 3 Performance comparison of eight methods with different γ on four benchmark data sets

Data sets	γ	Solver		ENSAFE		ST2		ST3		BLITZ		LS		ENDPP		E-ENDPP	
		time(s)	MSE	time(s)	MSE	time(s)	MSE	time(s)	MSE	time(s)	MSE	time(s)	MSE	time(s)	MSE	time(s)	MSE
Pyrimidibes	0.1	1.25	0.02	1.13	0.02	1.12	0.02	1.11	0.02	1.11	0.02	0.74	0.12	1.08	0.02	1.08	0.02
	1	1.24	0.02	1.22	0.02	1.22	0.02	1.24	0.02	1.20	0.02	0.29	0.12	1.14	0.02	1.13	0.02
	100	0.30	0.02	0.30	0.02	0.31	0.02	0.31	0.02	0.31	0.02	0.10	0.09	0.30	0.02	0.30	0.02
Triazines	0.1	3.80	0.03	3.72	0.03	3.57	0.03	3.42	0.03	3.78	0.03	1.17	0.06	2.97	0.03	2.64	0.03
	1	3.77	0.03	3.69	0.03	3.56	0.03	3.44	0.03	3.77	0.03	1.13	0.06	2.98	0.03	2.66	0.03
	100	1.62	0.03	1.67	0.03	1.66	0.03	1.63	0.03	1.54	0.03	0.35	0.20	1.45	0.03	1.32	0.03
Riboflavin	0.1	39.39	5.33	41.65	5.33	39.30	5.33	39.14	5.33	44.71	5.33	302.18	1.50	40.40	5.33	11.21	5.33
	1	39.24	5.33	40.39	5.33	40.38	5.33	38.99	5.33	44.79	5.33	311.74	3.59	39.99	5.33	11.79	5.33
	100	41.82	0.25	44.00	0.26	40.29	0.26	37.96	0.26	44.65	0.26	296.04	0.94	37.84	0.26	10.44	0.26
Pulmon	0.1	2472.83	0.36e-2	2037.88	0.36e-2	1732.54	0.36e-2	1124.95	0.36e-2	964.41	0.36e-2	>24h	–	392.44	0.36e-2	207.77	0.36e-2
	1	2633.05	0.93e-2	2124.40	0.93e-2	1663.21	0.93e-2	1109.53	0.93e-2	993.94	0.93e-2	>24h	–	345.19	0.93e-2	152.22	0.93e-2
	100	2599.98	0.01	1873.36	0.01	1964.51	0.01	2195.43	0.01	1220.71	0.01	>24h	–	348.89	0.01	156.62	0.01



(a) Pyrimidibes

(b) Triazines

(c) Riboflavin

(d) Pulmon

Fig. 2 Rejection ratios of ENSAFE, ST2, ST3, ENDPP and E-ENDPP on four real data sets

7 Conclusion

The Elastic Net problem produces good prediction accuracy. However, it costs a lot of time when dealing with the large-scale problem. To improve the computational speed, in this paper, we propose a novel screening rule for the Elastic Net problem. Our proposed rule can effectively identify inactive predictors of the Elastic Net problem, thus it greatly reduces the scale of the optimization problem. Moreover, the solution obtained by the proposed rule is exactly the same as the original problem, i.e., our rule is safe. Both synthetic and real data sets demonstrate the effectiveness of the proposed rule.

Acknowledgments The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. This work was supported in part by the Beijing Natural Science Foundation (No. 4172035) and National Natural Science Foundation of China (No. 11671010).

Appendix A: Proof of Lemma 1

First of all, we reintroduce problem (3) in the following:

$$\min_{\beta \in R^p} \frac{1}{2}\|y - X\beta\|_2^2 + \frac{\gamma}{2}\|\beta\|_2^2 + \lambda\|\beta\|_1.$$

For the above problem, let

$$\bar{Y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \bar{X} = \begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}.$$

Then we will obtain that

$$\begin{aligned} &\frac{1}{2}\|y - X\beta\|_2^2 + \frac{\gamma}{2}\|\beta\|_2^2 + \lambda\|\beta\|_1 \\ &= \frac{1}{2}(y - X\beta)^T(y - X\beta) + \frac{\gamma}{2}\beta^T\beta + \lambda\|\beta\|_1 \\ &= \frac{1}{2}y^T y - y^T X\beta + \frac{1}{2}\beta^T(X^T X + \gamma I)\beta + \lambda\|\beta\|_1 \end{aligned}$$

$$\begin{aligned} &= \frac{1}{2}(y^T \;\; 0)\begin{pmatrix} y \\ 0 \end{pmatrix} - (y^T \;\; 0)\begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}\beta \\ &+ \frac{1}{2}\beta^T (X^T \;\; \sqrt{\gamma} I)\begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}\beta + \lambda\|\beta\|_1 \\ &= \frac{1}{2}\bar{Y}^T\bar{Y} + \frac{1}{2}\beta^T\bar{X}^T\bar{X}\beta - \bar{Y}^T\bar{X}\beta + \lambda\|\beta\|_1 \\ &= \frac{1}{2}\|\bar{Y} - \bar{X}\beta\|_2^2 + \lambda\|\beta\|_1, \end{aligned} \tag{24}$$

therefore, problem (3) can be rewritten as:

$$\min_{\beta \in R^p} \frac{1}{2}\|\bar{Y} - \bar{X}\beta\|_2^2 + \lambda\|\beta\|_1,$$

Lemma 1 has been proved.

Appendix B: The dual problem of the Elastic Net

For the problem (2), according to (Boyd and Vandenberghe, 2004)[4], by introducing a new set of variables $w = y - X\beta$, the problem (2) becomes

$$\begin{aligned} \min_{w,\beta} \quad & \frac{1}{2}\|w\|_2^2 + \lambda\|\beta\|_1, \\ \text{s.t.} \quad & w = y - X\beta. \end{aligned} \tag{25}$$

By introducing the multipliers $\eta \in R^n$, we get the Lagrangian function as follows:

$$L(\beta, w, \eta) = \frac{1}{2}\|w\|_2^2 + \lambda\|\beta\|_1 + \eta^T(y - X\beta - w).$$

The dual function $g(\eta)$ is

$$\begin{aligned} g(\eta) &= \inf_{\beta,w} L(\beta, w, \eta) = \eta^T y + \inf_{w} \left(\frac{1}{2}\|w\|_2^2 - \eta^T w\right) \\ &+ \inf_{\beta} \left(-\eta^T X\beta + \lambda\|\beta\|_1\right). \end{aligned} \tag{26}$$

Note that the right side of $g(\eta)$ has three items. In order to get $g(\eta)$, we need to solve the following two optimization problems.

$$\inf_{w} \quad \frac{1}{2}\|w\|_2^2 - \eta^T w, \tag{27}$$

$$\inf_{\beta} \quad -\eta^T X\beta + \lambda\|\beta\|_1. \tag{28}$$

Let us first consider (27). Denote the objective function as

$$f_1(w) = \frac{1}{2}\|w\|_2^2 - \eta^T w,$$

then, let

$$\frac{\partial f_1(w)}{\partial w} = w - \eta = 0 \Rightarrow w = \eta,$$

so

$$\inf_{w} f_1(w) = -\frac{1}{2}\eta^T\eta = -\frac{1}{2}\|\eta\|_2^2.$$

Next, let us consider the problem (28). Denote the objective function as

$$f_2(\beta) = -\eta^T X\beta + \lambda\|\beta\|_1,$$

$f_2(\beta)$ is convex but not smooth. So let us consider its subgradient

$$\frac{\partial f_2(\beta)}{\partial \beta} = -X^T\eta + \lambda\frac{\partial\|\beta\|_1}{\partial \beta} = 0,$$

where

$$\frac{\partial\|\beta\|_1}{\partial \beta_i} = \begin{cases} sign([\beta^*]_i), & if\ [\beta^*]_i \neq 0 \\ [-1, 1], & if\ [\beta^*]_i = 0 \end{cases} \quad i = 1, 2, \cdots p.$$

According to the necessary condition for $f_2(\beta)$ to attain an optimum, we have $|X_i^T\eta| \leq \lambda, \quad i = 1, 2, \cdots p$. Therefore, the optimum value of problem (28) is 0. Combining the equations above, we can get the dual problem:

$$\begin{aligned} \max_{\eta} \quad & g(\eta) = \eta^T y - \frac{1}{2}\|\eta\|_2^2, \\ \text{s.t.} \quad & |X_i^T\eta| \leq \lambda, \quad i = 1, 2, \cdots p, \end{aligned}$$

which is equivalent to the following optimization problem:

$$\begin{aligned} \max_{\eta} \quad & \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|\eta - y\|_2^2, \\ \text{s.t.} \quad & |X_i^T\eta| \leq \lambda, \quad i = 1, 2, \cdots p. \end{aligned} \tag{29}$$

By a simple re-scaling of the dual variable η, that is, let $\theta = \frac{\eta}{\lambda}$, we get the following result. The dual problem of (2) is:

$$\begin{aligned} \max_{\theta} \quad & \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}\|\theta - \frac{y}{\lambda}\|_2^2, \\ \text{s.t.} \quad & \mid X_i^T\theta \mid \leq 1 \quad i = 1, 2, \cdots p, \end{aligned} \tag{30}$$

where θ is the dual variable. For notational convenience, let the optimal solution of problem (30) be $\theta^*(\gamma, \lambda)$, and the optimal solution of problem (2) with parameters γ and λ is denoted by $\beta^*(\gamma, \lambda)$. Then, the KKT conditions are given by:

$$y = X\beta^*(\gamma, \lambda) + \lambda\theta^*(\gamma, \lambda), \tag{31}$$

$$X_i^T\theta^*(\gamma, \lambda) \in \begin{cases} sign([\beta^*(\gamma, \lambda)]_i), & if\ [\beta^*(\gamma, \lambda)]_i \neq 0 \\ [-1, 1], & if\ [\beta^*(\gamma, \lambda)]_i = 0 \end{cases} \quad i = 1, 2, \cdots p, \tag{32}$$

Note that $\bar{Y} = \begin{pmatrix} y \\ 0 \end{pmatrix}$, $\bar{X} = \begin{pmatrix} X \\ \sqrt{\gamma} I \end{pmatrix}$. Let $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$, then we have the dual problem of (3) as follows,

$$\begin{aligned} \max_{\theta} \quad & \frac{1}{2}\|y\|_2^2 - \frac{\lambda^2}{2}(\|\theta_1 - \frac{y}{\lambda}\|_2^2 + \|\theta_2\|_2^2), \\ \text{s.t.} \quad & \mid \bar{X}_i^T\theta \mid = \mid x_i^T\theta_1 + \sqrt{\gamma}(\theta_2)_i \mid \leq 1, \ i = 1, 2, \cdots p, \end{aligned} \tag{33}$$

where $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ is the dual variable. So, Theorem 1 has been proved.

References

1. Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2):83–85
2. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H (2017) Feature selection: a data perspective. Acm Computing Surveys 50(6):1–45
3. Bühlmann P., Kalisch M, Meier L (2014) High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application 1(1):255–278
4. Boyd S, Vandenberghe L (2004) Convex optimization, Cambridge University Press, New York
5. Bondell H, Reich B (2010) Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. Biometrics 64(1):115–123
6. Xu Y, Zhong P, Wang L (2010) Support vector machine-based embedded approach feature selection algorithm. Journal of Information and Computational Science 7(5):1155–1163
7. Tibshirani R (1996) Regression shrinkage and subset selection with the lasso. Journal of the Royal Statistical Society
8. Zhao P, Yu B (2006) On model selection consistency of lasso. J Mach Learn Res 7(12):2541–2563
9. Candès E (2006) Compressive sampling. In: Proceedings of the international congress of mathematics
10. Chen S, Donoho D, Saunders M (2001) Atomic decomposition by basis pursuit. SIAM Journal on Scientific Computing (SISC) 58(1):33–61
11. Wright J, Ma Y, Mairal J, Sapiro G, Huang T, Yan S (2010) Sparse representation for computer vision and pattern recognition. In: Proceedings of IEEE, 98(6): pp 1031–1044
12. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499
13. Kim S, Koh K, Lustig M, Boyd S, Gorinevsky D (2008) An interior-point method for large scale l1-regularized least squares. IEEE Journal on Selected Topics in Signal Processing 1(4):606–617
14. Friedman J, Hastie T, Hëfling H, Tibshirani R (2007) Pathwise coordinate optimization. Ann Appl Stat 1(2):302–332
15. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22
16. Park M, Hastie T (2007) L1-regularized path algorithm for generalized linear models. Journal of the Royal Statistical Society Series B 69(4):659–677
17. Donoho D, Tsaig Y (2008) Fast solution of l-1 norm minimization problems when the solution may be sparse. IEEE Trans Inf Theory 54(11):4789–4812
18. El Ghaoui L, Viallon V, Rabbani T (2012) Safe feature elimination in sparse supervised learning. Pacific Journal of Optimization 8(4):667–698
19. Pan X, Yang Z, Xu Y, Wang L (2018) Safe screening rules for accelerating twin support vector machine classification. IEEE Transactions on Neural Networks and Learning Systems 29(5):1876–1887
20. Xiang Z, Ramadge P (2012) Fast lasso screening tests based on correlations. In: 2012 IEEE international conference on acoustics. Speech and Signal Processing (ICASSP) 22(10):2137–2140
21. Xiang Z, Xu H, Ramadge P (2011) Learning sparse representations of high dimensional data on large scale dictionaries. International conference on neural information processing systems 24:900–908
22. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N (2012) Strong rules for discarding predictors in lasso-type problems. J R Stat Soc 74(2):245–266
23. Wang J, Wonka P, Ye J (2012) Lasso screening rules via dual polytope projection. J Mach Learn Res 16(1):1063–1101
24. Bruckstein A, Donoho D, Elad M (2009) From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev 51:34–81
25. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67(2):301–320
26. Hastie T, Tibshirani R (2009) The elements of statistical learning. Technometrics 45(3):267–268
27. Hoerl A, Kennard R (1988) Ridge regression. In: Encyclopedia of statistical sciences, 8: 129–136. Wiley, New York
28. Breiman L (1996) Heuristics of instability in model selection. The Annals of Statistics 24
29. Bertsekas D (2003) Convex analysis and opitimization. Athena scientific
30. Bauschke H, Combettes P (2011) Convex analysis and monotone operator theory in hilbert spaces, Springer, New York
31. Johnson T, Guestrin C (2015) BLITZ: a principled meta-algorithm for scaling sparse optimization. In: International conference on international conference on machine learning 18(12): 1171–1179
32. He X, Cai D, Niyogi P (2005) Laplacian score for feature selection. In: International conference on neural information processing systems 18:507–514

Yitian Xu received the Ph.D. degree from the College of Science, China Agricultural University, Beijing, China, in 2007. He was a Visiting Scholar with the Department of Computer Science and Engineering, Arizona State University, Tempe, AZ, USA, from 2013 to 2014. He is currently a Professor and Supervisor for the Ph.D. candidates with the College of Science, China Agricultural University. He has authored about 50 papers. His current research interests include machine learning and data mining.

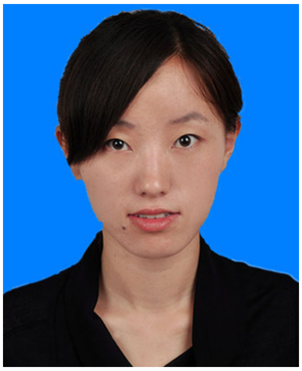
Prof. Xu's research has appeared in IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, Information Science, Pattern Recognition, Knowledge-Based Systems, Neurocomputing, and so on.

Ying Tian received her BS degree from Beijing University of Technology. She is currently a lecturer and pursuing the Ph.D.degree in College of Science at China Agriculture University. Her research interests include support vector machine, statistical models and data mining.

Xianli Pan was born in China in 1991. She received the B.S. and M.S. degrees from the College of Science, China Agricultural University, Beijing, China, in 2014 and 2016, respectively, where she is currently pursuing the Ph.D. degree.

Her current research interests include support vector machine, machine learning, and data mining.

Ms. Pan's research has appeared in IEEE Transactions on Neural Networks and Learning Systems, Knowledge-Based Systems, Journal of Intelligent and Fuzzy Systems, and Neurocomputing.



Hongmei Wang was born in China in 1991. She received the B.S. degree from the college of science, Jining University, Shandong, China, in 2014, and the M.S. degree from the college of science, Beijing University of Technology, Beijing, China, in 2017. She is currently pursuing the Ph.D. degree in College of information and electrical engineering, China Agricultural University, Beijing, China.

Her current research interests include support vector machine, machine learning, and data mining.

Ms. Wang's research has appeared in Information Science and Neurocomputing.