SCALED AND SQUARE-ROOT ELASTIC NET

Elias Raninen and Esa Ollila

Aalto University, Dept. of Signal Processing and Acoustics, P.O.Box 13000, FI-00076 Aalto, Finland

ABSTRACT

In scaled lasso, the unknown regression coefficients and the scale parameter of the error distribution are estimated jointly. In lasso, the optimal penalty parameter is well-known to depend on the error scale, and it is therefore typically chosen using cross-validation. The main benefit of scaled lasso is that the penalty parameter is scale-free and can be predetermined from pure theoretical considerations. Nevertheless, scaled lasso performs poorly when there exist strong correlations between the predictors. As a remedy, we propose two different scaled elastic net (EN) formulations and derive convergent algorithms for their computation. The first formulation uses a conventional EN penalty whereas the second formulation differs from the former in that the ℓ_2-loss is not squared. The former approach is referred to as the scaled EN estimator and the latter as the square-root EN estimator. We illustrate via numerical examples and simulations that the proposed methods outperform the scaled lasso, especially in the presence of high mutual coherence in the feature space.

Index Terms— Scaled lasso, square-root lasso, penalized linear regression, scale invariance, elastic net.

1. INTRODUCTION

We consider a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the observed n-dimensional response (measurement) vector, $\mathbf{X} = \begin{pmatrix}\mathbf{x}_1 \cdots \mathbf{x}_p\end{pmatrix}$ is the fixed $n \times p$ design matrix, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients, and $\boldsymbol{\varepsilon}$ is an unobserved n-vector of i.i.d. random variables from a symmetric distribution with an unknown *error scale* parameter σ. As is common in penalized regression, we standardize the columns of $\mathbf{X}$ to unit norm, i.e., $\|\mathbf{x}_j\|_2 = 1$. The popular lasso (Least Absolute Shrinkage and Selection Operator) [1] estimator is defined as the minimizer of the criterion $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$, where $\lambda \geq 0$ is a *penalty parameter*, or threshold level, chosen by the user. For more references on earlier work on ℓ_1-regularization and its usage and applications in other fields, see, e.g., references [2, 3, 4, 5] to mention only a few. The theoretically optimal values of λ depend on the scale parameter σ, the estimation of which is a non-trivial task, especially in high-dimensional problems. Consequently, the penalty parameter λ is typically chosen using data adaptive methods such as cross-validation. Although cross-validation often performs well in choosing a model for prediction, in addition to being computationally costly, it does not provide consistent model selection [6].

In *scaled lasso* [7], the penalty parameter is no longer dependent on the error scale and therefore optimal universal penalty levels can be selected based on theoretical properties of the estimator. It has also proven to be an accurate method in the estimation of the error variance in high-dimensional settings [8]. In scaled lasso, one estimates the unknown regression coefficients and the scale simultaneously by solving Huber's [9] jointly convex concomitant criterion function:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\sigma>0}{\text{minimize}} \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma} + \frac{n\sigma}{2} + \lambda\|\boldsymbol{\beta}\|_1 . \tag{1}$$

The problem was first studied in [10] and a theoretical analysis was provided in [7]. Interestingly, the solution path of the scaled lasso is 1-to-1 with the lasso path, the benefit being that in scaled lasso the penalty parameter is now scale-independent. The scaled lasso solution $\hat{\boldsymbol{\beta}}$ is also a solution of the optimization program

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + (\lambda/\sqrt{n})\|\boldsymbol{\beta}\|_1 , \tag{2}$$

which is called the *square-root lasso* [11].

A well-known deficiency of the lasso (and scaled lasso) is its poor performance in case of high mutual coherence (or multicollinearity), i.e., when the basis vectors $\mathbf{x}_j$ are highly correlated. Moreover, lasso picks at most n variables in the $p > n$ case. The *elastic net* (EN) [12] is a popular regularization and variable selection method that overcomes the above shortcomings by utilizing a penalty function that is a combination of the ℓ_1 and ℓ_2-norm penalties, defined as

$$\mathcal{P}_{\text{EN}}(\boldsymbol{\beta};\alpha) = \frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1 , \tag{3}$$

where $\alpha \in [0,1]$ is an additional *EN tuning parameter*. The EN penalty reduces to the lasso penalty when $\alpha = 1$ and to the ridge regression [13] penalty when $\alpha = 0$.

In this paper, we extend the idea of scaled lasso to the elastic net and solve the optimization programs

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\sigma>0}{\text{minimize}} \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma} + \frac{n\sigma}{2} + \lambda\mathcal{P}(\boldsymbol{\beta};\alpha), \tag{4}$$

where $\mathcal{P}(\boldsymbol{\beta};\alpha)$ is either the conventional EN penalty (3), or

$$\mathcal{P}_{\sqrt{\text{EN}}}(\boldsymbol{\beta};\alpha) = (1-\alpha)\left\|\boldsymbol{\beta}\right\|_2 + \alpha\left\|\boldsymbol{\beta}\right\|_1, \tag{5}$$

referred to as the *square-root EN penalty* since it utilizes a non-squared ℓ_2-norm, as does the square-root lasso in (2). When $\alpha = 1$, both estimators reduce to the conventional scaled (or square-root) lasso, but for intermediate values they differ; and for $\alpha = 0$, they yield different scaled ridge regression estimators. Both approaches are potentially interesting EN penalties to be used in scaled sparse regression. The former approach is referred to as the *scaled EN estimator* and the latter as the *square-root EN estimator*.

The paper is organized as follows. In Section 2, we study the scaled EN estimator and derive a cyclic coordinate-wise descent (CCD) algorithm to find its solution. Section 3 studies the square-root EN estimator and derives the corresponding CCD algorithm for its computation. Finally, Section 4 provides numerical examples as well as a simulation study. Section 5 concludes.

Notations: For an n-dimensional vector $\mathbf{a}$, the ℓ_2-norm and ℓ_1-norm are defined as $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^\top\mathbf{a}}$ and $\|\mathbf{a}\|_1 = \sum_{i=1}^n |a_i|$, respectively. The pseudo-norm $\|\mathbf{a}\|_0 = \sum_{i=1}^n \mathsf{I}_{a_i\neq 0}$, where the symbol I_A denotes an indicator function, counts the number of non-zero elements in the vector. Notations such as $\text{sign}(\mathbf{a})$ imply that the univariate function $\text{sign}(\cdot)$ acts coordinate-wise to vector $\mathbf{a}$, so that $[\text{sign}(\mathbf{a})]_i = \text{sign}(a_i)$. The *soft-thresholding operator* is defined as $[\mathcal{S}(\mathbf{a},\lambda)]_i = \text{sign}\,(a_i)\,(|a_i|-\lambda)_+$, where $(\cdot)_+ = \max\{\cdot, 0\}$ denotes the *subplus operator*.

Relations to prior work: The square-root lasso was extended to the group square-root lasso in [14]. Our work continues with extending the scaled (or square-root) lasso to the elastic net.

2. SCALED EN ESTIMATOR

Let us first note that the minimizer of both estimators in (4) with respect to the noise scale σ satisfies $-\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2/(2\hat{\sigma}^2)+n/2=0$, which gives

$$\hat{\sigma}(\boldsymbol{\beta}) = \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2}{\sqrt{n}}, \tag{6}$$

whereas the minimizer with respect to $\boldsymbol{\beta}$ differs between the two variants.

The scaled EN estimators of regression and scale, $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$, are defined as the minimizers of the criterion

$$\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma} + \frac{n\sigma}{2} + \lambda\left\{\frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right\} \tag{7}$$

over $(\boldsymbol{\beta},\sigma) \in \mathbb{R}^p \times (0,\infty)$. The criterion function in (7) is *separable* as it can be written in the form: $f(\boldsymbol{\beta},\sigma) = g(\boldsymbol{\beta},\sigma) + \sum_{j=1}^p h_j(\beta_j)$, where $g : \mathbb{R}^p \times (0,\infty) \to \mathbb{R}$ is convex and differentiable and $h_j : \mathbb{R} \to \mathbb{R}$ are convex [3]. A CCD procedure, in which the function is minimized cyclically with respect to one coordinate at a time, is therefore guaranteed to converge [15]. Hence, we will next derive a CCD algorithm for the problem.

Let $\mathbf{X}_{-j}$ and $\boldsymbol{\beta}_{-j}$ denote the matrix $\mathbf{X}$ and vector $\boldsymbol{\beta}$ with the j^{th} column and element excluded, respectively, and let $\mathbf{x}_j$ denote the j^{th} column of $\mathbf{X}$. We rewrite the objective function in (7) as

$$\frac{\left\|\boldsymbol{r}^{(j)} - \mathbf{x}_j\beta_j\right\|_2^2}{2\sigma} + \frac{n\sigma}{2} + \lambda\left\{\frac{(1-\alpha)}{2}\beta_j^2 + \alpha|\beta_j|\right\} + \lambda\left\{\frac{(1-\alpha)}{2}\left\|\boldsymbol{\beta}_{-j}\right\|_2^2 + \alpha\left\|\boldsymbol{\beta}_{-j}\right\|_1\right\}, \tag{8}$$

where $\boldsymbol{r}^{(j)} = \mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j}$ denotes the partial residual vector. Using the following notation:

$$\lambda_1 = \hat{\sigma}\alpha\lambda, \quad \lambda_2 = \hat{\sigma}(1-\alpha)\lambda, \tag{9}$$

the minimizer of (8) with respect to β_j, when considering the other coefficients β_k, $k \neq j$, and σ fixed at their current iterates, needs to verify the zero subgradient equation

$$-\mathbf{x}_j^\top(\boldsymbol{r}^{(j)} - \mathbf{x}_j\hat{\beta}_j) + \lambda_2\hat{\beta}_j + \lambda_1\hat{t}_j = 0,$$

where $\hat{t}_j$ is a subgradient of $|\beta_j|$ evaluated at $\hat{\beta}_j$, i.e., equal to $\hat{\beta}_j/|\hat{\beta}_j|$ if $\hat{\beta}_j \neq 0$ and some number in $[-1,1]$ otherwise. If $\hat{\beta}_j \neq 0$, then $\hat{t}_j = \hat{\beta}_j/|\hat{\beta}_j|$, and we have

$$\left(1 + \lambda_2 + \frac{\lambda_1}{|\hat{\beta}_j|}\right)\hat{\beta}_j = \mathbf{x}_j^\top\boldsymbol{r}^{(j)}. \tag{10}$$

Taking the absolute value of both sides of (10) and some algebra yields

$$|\hat{\beta}_j| = \frac{|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}| - \lambda_1}{1+\lambda_2}. \tag{11}$$

Since we assumed $\hat{\beta}_j \neq 0$, the term $|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}| - \lambda_1$ has to be positive. Substituting (11) into (10) and solving for $\hat{\beta}_j$ yields

$$\hat{\beta}_j = \frac{\mathbf{x}_j^\top\boldsymbol{r}^{(j)}}{|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}|}\frac{\left(|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}| - \lambda_1\right)_+}{1+\lambda_2} = \frac{\mathcal{S}(\hat{\beta}_j + \mathbf{x}_j^\top\boldsymbol{r}, \lambda_1)}{1+\lambda_2}. \tag{12}$$

In obtaining the last identity, we used that $\boldsymbol{r} = \boldsymbol{r}^{(j)} - \mathbf{x}_j\hat{\beta}_j$.

The algorithm for solving the scaled lasso (1) in [7] is an iterative procedure in which the scale and the coefficients are updated sequentially. A same type of approach can be adopted for the scaled EN following from the CCD principle. Thus, the scale is first updated using (6) and the current full iterate $\hat{\boldsymbol{\beta}}$ in place of $\boldsymbol{\beta}$. Thereafter, one performs coordinate-wise minimization over all predictors $j \in \{1,\ldots,p\}$ according to (12) holding $\boldsymbol{\beta}_{-j}$ and σ fixed at their current iterates $\hat{\beta}_k$, $k \neq j$, and $\hat{\sigma}$. These two steps are alternated until convergence as described in Algorithm 1.

Algorithm 1: Scaled EN and square-root EN

Input : $\mathbf{X}, \mathbf{y}, \lambda, \alpha, \hat{\boldsymbol{\beta}} \leftarrow \mathbf{0}$
while *not converged* **do**
  $\hat{\sigma} \leftarrow \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2/\sqrt{n}$;
  $\lambda_1 \leftarrow \hat{\sigma}\alpha\lambda, \quad \lambda_2 \leftarrow \hat{\sigma}(1-\alpha)\lambda$;
  for $j = 1$ *to* p **do**
    $\boldsymbol{r} \leftarrow \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$;
    if *Scaled EN* **then**
      $\hat{\beta}_j \leftarrow \dfrac{\mathcal{S}\big(\hat{\beta}_j + \mathbf{x}_j^\top \boldsymbol{r}\,,\, \lambda_1\big)}{1+\lambda_2}$
    else if *Square-root EN* **then**
      if *condition* (14) *is met* **then**
        $\hat{\boldsymbol{\beta}} \leftarrow \mathbf{0}$;
      else
        $\hat{\beta}_j \leftarrow \dfrac{\mathcal{S}\big(\hat{\beta}_j + \mathbf{x}_j^\top \boldsymbol{r}\,,\, \lambda_1\big)}{1+\lambda_2/\|\hat{\boldsymbol{\beta}}\|_2}$
Output: $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$

It is instructive to consider the orthonormal design matrix case, i.e., $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$ and $n = p$. With a little algebra, it is easy to show that $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ then satisfies

$$\hat{\boldsymbol{\beta}} = \frac{\mathcal{S}\big(\mathbf{X}^\top\mathbf{y}, \lambda_1\big)}{1+\lambda_2} \quad \text{and} \quad \hat{\sigma} = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2}{\sqrt{n}}.$$

As with the conventional elastic net, one can argue that also the scaled EN can suffer from a double shrinkage effect due to both ridge and lasso style shrinkage which introduces unwanted excess bias. As a consequence, the estimator can loose its predictive power. To remedy for the double shrinkage effect, we define, similar to [12], the corrected scaled EN estimates of regression and scale, $(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}^*)$, as $\hat{\boldsymbol{\beta}}^* = (1+\lambda_2)\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^* = \hat{\sigma}(\hat{\boldsymbol{\beta}}^*)$.

3. SQUARE-ROOT EN ESTIMATOR

The square-root EN estimators, $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$, are defined as the minimizers of the criterion

$$\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2}{2\sigma} + \frac{n\sigma}{2} + \lambda\Big\{(1-\alpha)\,\|\boldsymbol{\beta}\|_2 + \alpha\,\|\boldsymbol{\beta}\|_1\Big\} \quad (13)$$

over $(\boldsymbol{\beta}, \sigma) \in \mathbb{R}^p \times (0, \infty)$.

The zero subgradient equation of (13) with respect to $\boldsymbol{\beta}$ is

$$-\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda_2\hat{\mathbf{s}} + \lambda_1\hat{\mathbf{t}} = 0,$$

where (λ_1, λ_2) are defined in (9), $\hat{\mathbf{t}}$ is a p-vector whose j^{th} element belongs to the subdifferential of $|\beta_j|$ evaluated at $\hat{\beta}_j$, and $\hat{\mathbf{s}}$ belongs to the subdifferential of $\|\boldsymbol{\beta}\|_2$ evaluated at $\hat{\boldsymbol{\beta}}$, i.e., $\hat{\mathbf{s}} = \hat{\boldsymbol{\beta}}/\|\hat{\boldsymbol{\beta}}\|_2$ if $\hat{\boldsymbol{\beta}} \neq \mathbf{0}$ and $\hat{\mathbf{s}} \in \{\mathbf{s} \in \mathbb{R}^p : \|\mathbf{s}\|_2 \leq 1\}$ if $\hat{\boldsymbol{\beta}} = \mathbf{0}$. It can be shown that the zero subgradient equation is satisfied with $\hat{\boldsymbol{\beta}} = \mathbf{0}$ if and only if

$$\big\|\mathcal{S}\big(\mathbf{X}^\top\mathbf{y}, \lambda\alpha\|\mathbf{y}\|_2/\sqrt{n}\big)\big\|_2 \leq \lambda(1-\alpha)\|\mathbf{y}\|_2/\sqrt{n}. \quad (14)$$

If we assume $\hat{\beta}_j \neq 0$, then

$$\left(1 + \frac{\lambda_2}{\|\hat{\boldsymbol{\beta}}\|_2} + \frac{\lambda_1}{|\hat{\beta}_j|}\right)\hat{\beta}_j = \mathbf{x}_j^\top\boldsymbol{r}^{(j)}. \quad (15)$$

Taking the modulus of both sides and solving for $|\hat{\beta}_j|$ yields

$$|\hat{\beta}_j| = \left(1 + \frac{\lambda_2}{\|\hat{\boldsymbol{\beta}}\|_2}\right)^{-1}\Big(|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}| - \lambda_1\Big)_+.$$

Plugging this back into (15) and solving for $\hat{\beta}_j$ gives

$$\hat{\beta}_j = \frac{\mathbf{x}_j^\top\boldsymbol{r}^{(j)}}{|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}|}\frac{(|\mathbf{x}_j^\top\boldsymbol{r}^{(j)}| - \lambda_1)_+}{1+\lambda_2/\|\hat{\boldsymbol{\beta}}\|_2} = \frac{\mathcal{S}(\hat{\beta}_j + \mathbf{x}_j^\top\boldsymbol{r}, \lambda_1)}{1+\lambda_2/\|\hat{\boldsymbol{\beta}}\|_2}.$$

As a consequence of non-separability, the formula depends on the norm of the optimal coefficients. In order to use the formula, we simply use the norm of the previous full iterate $\hat{\boldsymbol{\beta}}$. The procedure is given in Algorithm 1. For this problem, one could also have utilized a theoretically better justified generalized gradient descent scheme as in [16].

It is again instructive to consider the orthonormal design matrix case, i.e., $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$ and $n = p$. With a little algebra, it is easy to show that $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ then satisfies

$$\hat{\boldsymbol{\beta}} = \left(1 - \frac{\lambda_2}{\|\mathcal{S}(\mathbf{X}^\top\mathbf{y}, \lambda_1)\|_2}\right)_+ \mathcal{S}(\mathbf{X}^\top\mathbf{y}, \lambda_1),$$
$$\hat{\sigma} = (1/\sqrt{n})\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2.$$

To remedy for the double shrinkage effect, we define the corrected square-root EN estimates of regression and scale, $(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}^*)$, as

$$\hat{\boldsymbol{\beta}}^* = \left(1 - \frac{\lambda_2}{\|\mathcal{S}(\mathbf{X}^\top\mathbf{y}, \lambda_1)\|_2}\right)_+^{-1}\hat{\boldsymbol{\beta}}$$

and $\hat{\sigma}^* = \hat{\sigma}(\hat{\boldsymbol{\beta}}^*)$.

4. NUMERICAL RESULTS

The asymptotic results for the scaled lasso when $n \to \infty$ (including the case $p \geq n \geq \|\boldsymbol{\beta}\|_0 \to \infty$) suggest $\lambda \propto \sqrt{\log(p)}$ for the optimal tuning parameter [7]. When the sample size is finite, the performance will depend on the chosen proportionality constant. In the original paper [7], three different values for λ are considered; namely $\sqrt{2^{j-1}\log(p)}$, where $j = 1, 2$, and 3. Herein we consider those same values. In the simulations, we use the uncorrected EN estimators.

4.1. Example 1: Grouping effect of collinear variables

The first set-up illustrates the superiority of the EN penalties to lasso in situations of high mutual coherence as well as the grouping effect. The set-up is as in [3], where the linear model consists of two groups of three highly correlated predictor variables. The data is generated as $\mathbf{y} = 3\mathbf{z}_1 - 1.5\mathbf{z}_2 + 2\boldsymbol{\varepsilon}$, where $\mathbf{z}_1, \mathbf{z}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$, where $n = 100$; and the design matrix is generated as follows: $\mathbf{x}_j = \mathbf{z}_1 \mathbf{I}_{j\in\{1,2,3\}} + \mathbf{z}_2 \mathbf{I}_{j\in\{4,5,6\}} + (1/5)\boldsymbol{\varepsilon}_j$, where $\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$, $j \in \{1, \ldots, 6\}$. The predictors are then standardized to unit norm. As the tuning parameter λ is varied, the estimated regression coefficients trace a path in $\mathbb{R}^p$, referred to as the *solution path*, shown in Figure 1. Even with a very mild EN parameter value $\alpha = 0.95$, the scaled and square-root EN estimators are able to identify the two groups of correlated variables and connect them by setting them to zero at the same value of λ. By contrast, the scaled lasso fails to do so.

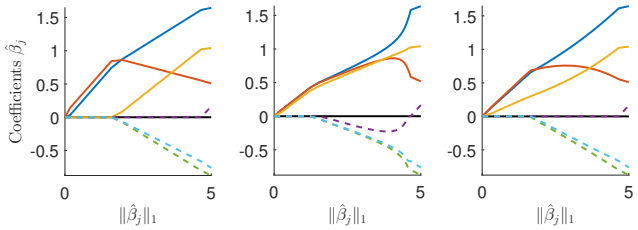


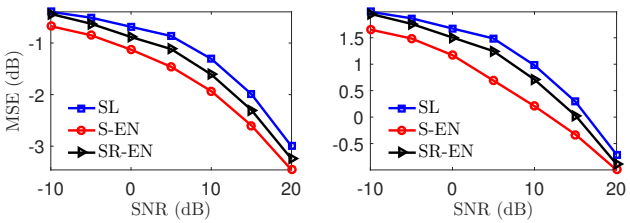
Fig. 1. The solution path of scaled lasso (left panel), scaled EN and square-root EN (right panel) with $\alpha = 0.95$.

4.2. Example 2: Performance vs. SNR

In the second set-up, we consider the following linear model: $(n, p) = (50, 10)$, the row vectors of $\mathbf{X}$ are normally distributed with mean vector $\mathbf{0}_{p\times 1}$ and covariance matrix $\boldsymbol{\Sigma}$, $\Sigma_{ij} = \text{corr}(i, j) = 0.9^{|i-j|}$ for $i, j = 1, \ldots, p$. In Figure 2 (a), we have $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)^\top = \boldsymbol{\beta}^{(1)}$, and in (b), we have $\boldsymbol{\beta} = (1, 2, 3, 4, 5, 0, 0, 0, 0, 0)^\top = \boldsymbol{\beta}^{(2)}$. For all methods, the tuning parameter value is set to $\lambda = \sqrt{2\log(p)}$. The EN tuning parameter is set to $\alpha = 0.9$ for both the scaled EN and the square-root EN. Figure 2 depicts the (empirical) mean squared error (MSE) versus the signal to noise ratio (SNR). As can be seen, both the scaled EN and the square-root EN outperform the scaled lasso. Here, the MSE is defined as $\text{MSE}(\hat{\boldsymbol{\beta}}) = \text{Ave}\{\frac{1}{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2\}$, where the average is over 200 Monte-Carlo trials, and the SNR (dB) is defined as $10\log_{10}(\sigma_{\boldsymbol{\beta}}^2/\sigma^2)$, where $\sigma_{\boldsymbol{\beta}}^2 = \sum_j |\beta_j|^2 / \|\boldsymbol{\beta}\|_0$.

4.3. Example 3: A high-dimensional setting

Next, we consider a high-dimensional problem where $(n, p) = (30, 150)$. The design matrix is generated as in example 2, SNR $= 0$ dB, the true coefficient vector is $\boldsymbol{\beta} = (1, \ldots, 1, \mathbf{0}_{1\times 130})^\top$, and the EN parameter is set to $\alpha = 0.9$. Three different values for λ are considered, namely $\lambda_j = \sqrt{2^{j-1}\log(p)}$, $j \in \{1, 2, 3\}$. Table 1 tabulates the $\text{MSE}(\hat{\boldsymbol{\beta}})$, the ratio of the estimated and true error scale, $\hat{\sigma}/\sigma$, the mean false positive rate (FPR), and the mean false negative rate (FNR). The reported results are averages over 100 Monte-Carlo trials. The standard deviation ($\times 10$) is given in the parenthesis. Based on Table 1, $\lambda_2 = \sqrt{2\log(p)}$ appears to be the best compromise giving the best estimates of the error scale, and the best mean squared error. With the chosen EN tuning parameter $\alpha = 0.9$, the scaled EN performs the best.



(a) $\boldsymbol{\beta}^{(1)}$ (b) $\boldsymbol{\beta}^{(2)}$

Fig. 2. $\text{MSE}(\hat{\boldsymbol{\beta}})$ vs. SNR for scaled lasso (SL), scaled elastic net (S-EN), and square-root elastic net (SR-EN).

Table 1. Results of example 3. The standard deviation ($\times 10$) is given in the parenthesis.

		$\text{MSE}(\hat{\boldsymbol{\beta}})$	$\hat{\sigma}/\sigma$	FPR	FNR
SL	λ_1	0.23 (0.7)	0.85 (1.7)	0.01 (0.1)	0.59 (0.8)
	λ_2	0.21 (0.6)	1.23 (2.1)	0.00 (0.0)	0.67 (0.8)
	λ_3	0.13 (0.2)	2.58 (2.9)	0.00 (0.0)	0.93 (0.6)
S-EN	λ_1	0.08 (0.2)	0.87 (1.7)	0.03 (0.2)	0.26 (1.0)
	λ_2	0.07 (0.1)	1.25 (2.0)	0.01 (0.1)	0.28 (1.2)
	λ_3	0.10 (0.2)	2.38 (2.4)	0.00 (0.0)	0.63 (1.8)
SR-EN	λ_1	0.18 (0.5)	0.79 (1.7)	0.02 (0.2)	0.49 (0.9)
	λ_2	0.14 (0.4)	1.11 (1.9)	0.00 (0.1)	0.51 (1.0)
	λ_3	0.11 (0.1)	2.21 (3.3)	0.00 (0.0)	0.71 (1.6)

5. DISCUSSION AND CONCLUSIONS

We proposed two EN extensions of the scaled lasso. The methods were shown to outperform the scaled lasso in the case of high correlations between the basis vectors as well as to encourage the grouping effect. In future research, we will investigate the theoretical properties of the scaled and square-root EN estimators. The performance of the scaled and square-root EN estimators are in general somewhat sensitive to the magnitude of the coefficients and sparsity level, which was also noted in [8] for the scaled lasso. For small n, the performance of the scaled estimators depends on the selected λ. This can be alleviated by choosing among a set of tuning parameters.

6. REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[2] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[3] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, 2015.

[4] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[5] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal of selected topics in signal processing*, vol. 1, no. 4, pp. 586–597, 2007.

[6] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[7] T. Sun and C.-H. Zhang, "Scaled sparse linear regression," *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.

[8] S. Reid, R. Tibshirani, and J. Friedman, "A study of error variance estimation in lasso regression," *Preprint* `arXiv:1311.5274v2 [stat.ME]`, 2013.

[9] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[10] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemporary Mathematics*, vol. 443, pp. 59–72, 2007.

[11] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: Pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[12] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[13] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[14] F. Bunea, J. Lederer, and Y. She, "The group square-root lasso: Theoretical properties and fast algorithms," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1313–1325, 2014.

[15] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[16] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.