SCALED AND SQUARE-ROOT ELASTIC NET

Aalto University

Elias Raninen and Esa Ollila
Department of Signal Processing and Acoustics, Aalto University, Finland
Contact: elias.raninen@aalto.fi and esa.ollila@aalto.fi

INTRODUCTION

Consider a **linear model** $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where
- $\mathbf{y}$ is n-vector of response variables, i.e., received signal,
- $\mathbf{X} = (\mathbf{x}_1 \cdots \mathbf{x}_p)$ is a fixed $n \times p$ design matrix ($\|\mathbf{x}_j\|_2 = 1$),
- $\boldsymbol{\beta}$ is a p-vector of unknown regression coefficients, and
- $\boldsymbol{\varepsilon}$ is an n-vector of i.i.d. random variables with zero mean and *error scale parameter* σ.

- The popular **lasso** [1] estimate is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2}{2} + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda \geq 0$ is a *penalty parameter*.
- The optimal value for λ depends on the unknown error scale σ.
- This dependency is eliminated in **scaled lasso** [2], which solves

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

- *Note*: Both methods perform poorly in the presence of **high correlations** in the feature space.
- *We propose* two elastic net [3] (EN) extensions to scaled lasso: the **scaled elastic net** and the **square-root elastic net**.
- *Gains:* Lower MSE and better estimate of the scale.

SCALED ELASTIC NET

The scaled EN estimators of regression and scale, $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$, solve

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

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

SQUARE-ROOT ELASTIC NET

The square-root EN estimators, $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$, solve

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

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

ALGORITHM

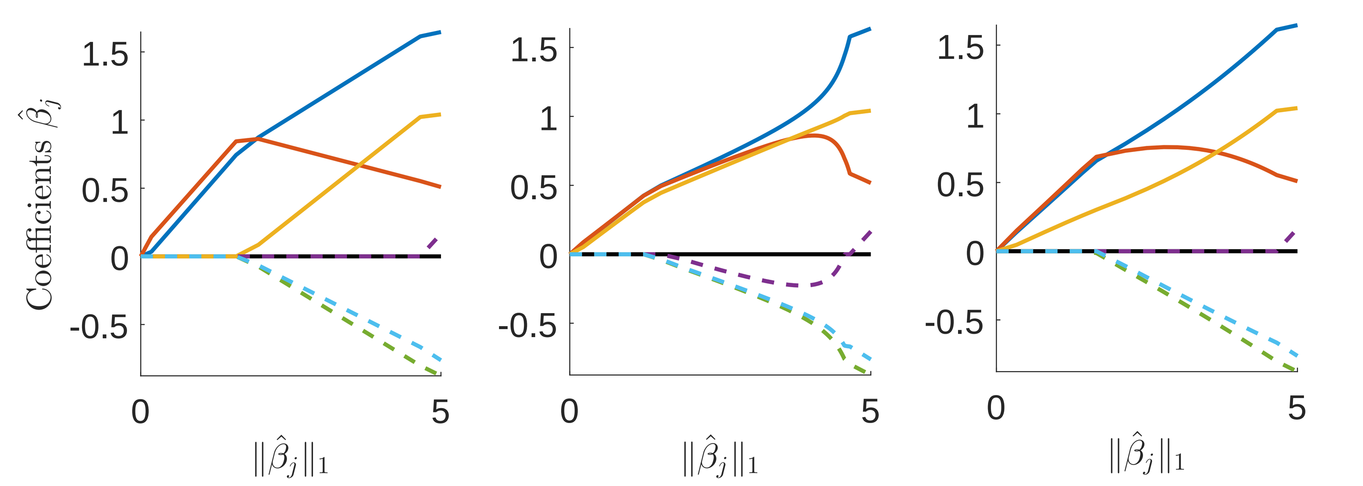
Input : $\mathbf{X}, \mathbf{y}, \lambda, \alpha, \hat{\boldsymbol{\beta}} \leftarrow \mathbf{0}$
while *not converged* **do**
  $\hat{\sigma} \leftarrow \|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|_2/\sqrt{n}$;
  $\lambda_1 \leftarrow \hat{\sigma}\alpha\lambda, \quad \lambda_2 \leftarrow \hat{\sigma}(1-\alpha)\lambda$;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

Output: $(\hat{\boldsymbol{\beta}}, \hat{\sigma})$

Notation:
$\mathcal{S}(z,\lambda) \triangleq \text{sign}(z)(|z|-\lambda)_+$, for $z \in \mathbb{R}$, and $(z)_+ \triangleq \max(0,z)$.

GROUPING OF COLLINEAR VARIABLES

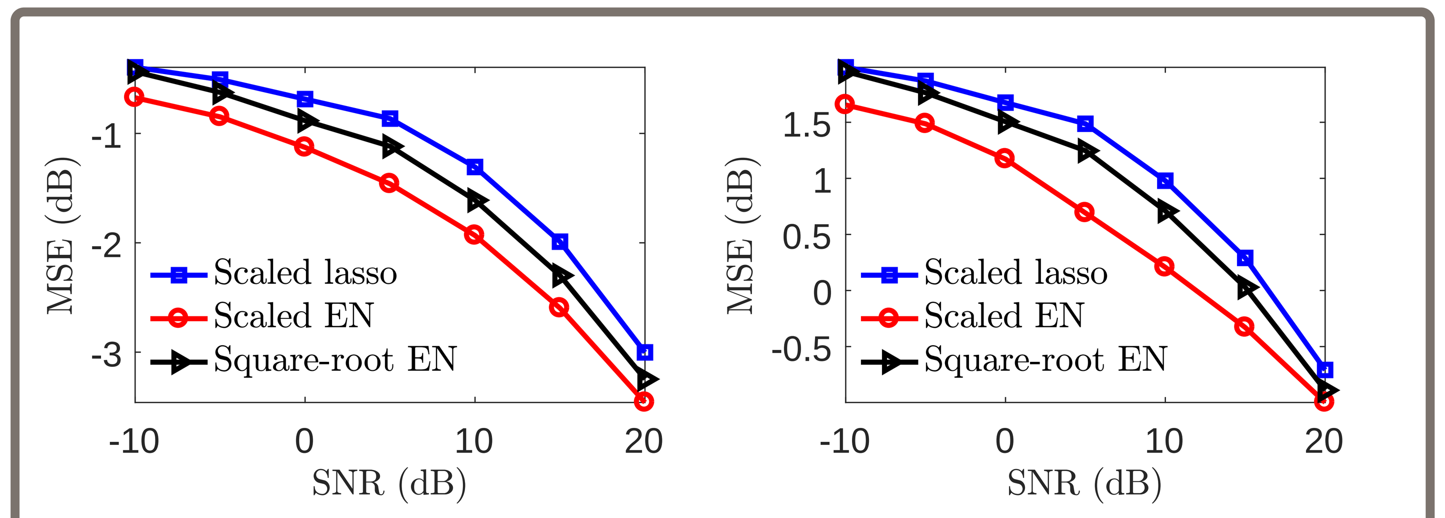
- Two groups of three highly correlated predictor variables.
- The proposed estimators are able to correctly group together the correlated variables, while the scaled lasso is not.



(a) Scaled lasso (b) Scaled EN (c) Square-root EN

PERFORMANCE VS. SNR

- Dimensions: $(n,p) = (50, 10)$
- Penalty parameter: $\lambda = \sqrt{2\log(p)}$
- EN parameter: $\alpha = 0.9$
- Correlation: $\Sigma_{ij} = \text{corr}(i,j) = 0.9^{|i-j|}$ for $i,j = 1,\ldots,p$
- $\text{MSE}(\hat{\boldsymbol{\beta}}) \triangleq \text{Ave}\left\{\frac{1}{p}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2\right\}$ (200 Monte Carlo trials)



(a) $\boldsymbol{\beta}^{(1)} = (1,1,1,1,1,0,0,0,0,0)^\top$ (b) $\boldsymbol{\beta}^{(2)} = (1,2,3,4,5,0,0,0,0,0)^\top$

A HIGH-DIMENSIONAL SETTING

- Dimensions: $(n,p) = (30, 150)$
- Penalty parameter: $\lambda = \sqrt{2\log(p)}$
- EN parameter: $\alpha = 0.9$
- Correlation: $\Sigma_{ij} = \text{corr}(i,j) = 0.9^{|i-j|}$ for $i,j = 1,\ldots,p$
- SNR $= 0$ dB
- $\text{MSE}(\hat{\boldsymbol{\beta}}) \triangleq \text{Ave}\{\frac{1}{p}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2\}$ (100 Monte Carlo trials)

	$\text{MSE}(\hat{\boldsymbol{\beta}})$	$\hat{\sigma}/\sigma$
Scaled lasso	0.21 (0.6)	1.23 (2.1)
Scaled EN	**0.07** (0.1)	1.25 (2.0)
Square-root EN	0.14 (0.4)	**1.11** (1.9)

(std × 10 is given in the parenthesis.)

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

[3] 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.

ICASSP'17, New Orleans, Louisiana, U.S.