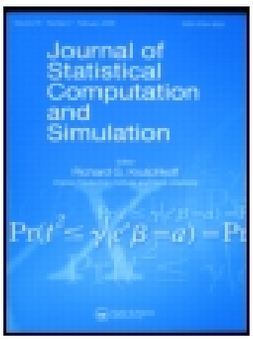

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

SLASSO: a scaled LASSO for multicollinear situations

Mohammad Arashi, Yasin Asar & Bahadır Yüzbaşı

To cite this article: Mohammad Arashi, Yasin Asar & Bahadır Yüzbaşı (2021): SLASSO: a scaled LASSO for multicollinear situations, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2021.1924174

To link to this article: https://doi.org/10.1080/00949655.2021.1924174

Published online: 11 May 2021.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2021.1924174

SLASSO: a scaled LASSO for multicollinear situations

Mohammad Arashi [a], Yasin Asar [b] and Bahadır Yüzbaşı [c]

[a]Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Azadi Square Iran; [b]Department of Mathematics–Computer, Necmettin Erbakan University, Konya, Turkey; [c]Department of Econometric, Inonu University, Malatya, Turkey

ABSTRACT

We propose a re-scaled LASSO by pre-multiplying the LASSO with a matrix term, namely, scaled LASSO (SLASSO), for multicollinear situations. Our numerical study has shown that the SLASSO is comparable with other sparse modeling techniques and often outperforms the LASSO and elastic net. Our findings open new visions about using the LASSO still for sparse modeling and variable selection. We conclude our study by pointing that the same efficient algorithm can solve the SLASSO for solving the LASSO and suggest following the same construction technique for other penalized estimators

ARTICLE HISTORY
Received 28 November 2020
Accepted 27 April 2021

KEYWORDS
Biasing parameter; L_1-penalty; LASSO; Liu estimation; Multicollinearity; variable selection

AMS CLASSIFICATIONS
Primary: 62F15; Secondary: 62H05

1. Introduction

Let $\{(\boldsymbol{x}_i, Y_i), i = 1, \ldots, n\}$ be a random sample from the linear regression model

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $Y_i \in \mathbb{R}$ is the response, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ is the covariate vector and ϵ_i is the random error with $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2 \in \mathbb{R}^+$, $i = 1, \ldots, n$.

The ordinary least squares (OLS) estimator has the form $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{C}_n^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$, $\boldsymbol{C}_n = \boldsymbol{X}^\top\boldsymbol{X}$ with $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$. For sparse modeling, the OLS estimator is not practical, and in contrast, one may use a regularization method to find a few non-zero elements of $\boldsymbol{\beta}$. Under the L_1-penalty, Tibshirani [1] proposed the least absolute penalty and selection operator (LASSO) given by

$$\widehat{\boldsymbol{\beta}}_n^L = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\lambda > 0$ is the threshold, and $\|\boldsymbol{v}\|_q = (\sum_{j=1}^d |v_j|^q)^{1/q}$ for $\boldsymbol{v} = (v_1, \ldots, v_d)^\top$, with $q > 0$.

The LASSO has tractable theoretical and computational properties. Although it is a successful technique, two drawbacks of the LASSO are given as

(1) In the high-dimensional regime ($p > n$), the LASSO can select at most n variables, which is a limiting feature of a variable selection method.

CONTACT Mohammad Arashi arashi@um.ac.ir

(2) If there is a group of highly correlated variables, the LASSO tends to arbitrarily select only one from the group.

Zou and Hastie [2] introduced the Elastic Net (E-net) approach using a weighted combination of the L_1 and L_2 norms to alleviate these two issues. It can deal with the strongly correlated variables effectively. Like LASSO, the E-net also has some good properties. The E-net is given by

$$\widehat{\boldsymbol{\beta}}_n^{En} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}, \tag{3}$$

where λ_1 and λ_2 are non-negative tuning parameters.

Indeed the E-net is an improved LASSO, in which the penalty term in the ridge approach is taken into account in the optimization problem. Zou and Hastie [2] formulated the naïve E-net in such a way that the solution to the optimization problem connected with that of the LASSO. In the same line, we have a different concern which is motivated in below.

1.1. Motivation

Under a multicollinear situation, apart from the sparsity, the OLS estimator $\widehat{\boldsymbol{\beta}}_n$ is far away from the true value $\boldsymbol{\beta}$. Hence, it is of major importance to find a closer estimator. Based on the regularization approach of Tikhonov [3], Hoerl and Kennard [4] proposed to minimize the sum of squares error (SSE) plus the L_2-penalty as follows

$$\widehat{\boldsymbol{\beta}}_n^{RR} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}, \tag{4}$$

where $\lambda > 0$ is a biasing parameter. The ridge regression (RR) estimator is a non-linear function concerning the tuning (biasing, here) parameter in nature. We refer to Yuzbasi et al. [5] for the recent advancement of RR in high-dimensional analysis.

Another approach to combat multicollinearity is to minimize the SSE subject to

$$\widehat{\boldsymbol{\beta}}_n^{d} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|d\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2 \right\}, \tag{5}$$

where $0 < d < 1$, due to Mayer and Wilke [6] and Liu [7]. The idea is the estimator $\widehat{\boldsymbol{\beta}}_n^{d}$ can be closer to the true value $\boldsymbol{\beta}$ than $\widehat{\boldsymbol{\beta}}_n$, if we can appropriately select the regularization (biasing, here) parameter d. The resulting estimator is called the linear unified (Liu) estimator given by $\widehat{\boldsymbol{\beta}}_n^{d} = \boldsymbol{F}_n(d)\widehat{\boldsymbol{\beta}}_n$, $\boldsymbol{F}_n(d) = (\boldsymbol{C}_n + \boldsymbol{I}_p)^{-1}(\boldsymbol{C}_n + d\boldsymbol{I}_p)$, where $0 < d < 1$ is the biasing parameter and $\boldsymbol{F}_n(d)$ is the biasing factor. The Liu estimator is linear to the biasing parameter d. This estimator has been used in many studies to combat the multicollinearity problem despite of its relation to the OLS estimator; see Liu [8], Mansson et al. [9], Duran et al. [10], Arashi et al. [11], Arashi et al. [12] and Wu [13] to mention a few.

The key idea in our approach is making use of the biasing factor in a fashion that we obtain a multicollinear resistant and selector estimator. Hence, we propose to replace the penalty term $\lambda_2\|\boldsymbol{\beta}\|_2^2$ in the E-net by $\lambda_2\|d\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2$. We will see that this change gives an estimator (after a modification), which obtains by pre-multiplying the LASSO with the biasing factor.

1.2. Plan of paper

In Section 2, we define the scaled LASSO (SLASSO). In general, we modify the L_1-penalty term of LASSO and then propose a closed-form solution. In Section 3, we communicate about some asymptotic properties. We show that SLASSO is $\sqrt{n}$-consistent. Also, orthonormal design case is studied. Section 4 is devoted to an extensive numerical study. A real data example and six simulated examples are considered to compare the performance of SLASSO with the existing candidates, including the RR, LASSO, and E-net, while Section 5 contains conclusions and suggestions for further research. Proofs of all theorems are provided in the Appendix.

2. A scaled LASSO

This section proposes an estimator called scaled LASSO (SLASSO) via the penalized least squares approach.

2.1. Naïve look

Before giving the expression of SLASSO, we first study the effect of replacing $\lambda_2\|\boldsymbol{\beta}\|_2^2$ in the E-net by $\lambda_2\|d\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2$. As in Zou and Hastie [2], we assume that the response is centred and the predictors are standardized. For the fixed λ_1, λ_2, and d, we define the naïve loss

$$L(\boldsymbol{\beta};\lambda_1,\lambda_2,d) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2\|d\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1.$$

The above loss function initially proposed by Arashi et al. [14]. The following result gives the solution to the underlying optimization problem above, similar to Zou and Hastie [2].

假设存在d，使得

Proposition 2.1: *Assume that there exits some d such that* $\operatorname{sgn}(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n) = \operatorname{sgn}(\boldsymbol{\beta})$ *and* $\dot{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta};\lambda_1,\lambda_2,d)$. *Then,*

$$\dot{\boldsymbol{\beta}}_n = \frac{1}{\sqrt{1+\lambda_2}}\arg\min_{\boldsymbol{b}} \mathcal{L}(\boldsymbol{b};\gamma),$$

where

$$\mathcal{L}(\boldsymbol{b};\gamma) = \|\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{b}\|_2^2 + \gamma\|\boldsymbol{b}\|_1, \quad \gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}},$$

with $\boldsymbol{Y}^* = ((1-d)\boldsymbol{Y}^\top, \boldsymbol{0}^\top)^\top$, $\boldsymbol{X}^* = (1+\lambda_2)^{-\frac{1}{2}}(\boldsymbol{X}^\top, \sqrt{\lambda_2}\boldsymbol{I}_p)^\top$, *and* $\boldsymbol{b} = \sqrt{1+\lambda_2}(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n)$.

The above result shows that the solution to the naïve problem is an augmented LASSO. However, it does not provide a closed-form solution to the biasing parameter d. Recently, Genc and Ozkale [15] discussed the grouping effect of the solution.

From Proposition 2.1, one can also approximate the standard error. Let $\hat{\sigma}^2$ be the estimate of σ^2. Then, using the result of Osborne et al. [16], the variance-covariance matrix of $\dot{\boldsymbol{\beta}}_n$ has form $(\boldsymbol{X}^{*\top}\boldsymbol{X}^* + \boldsymbol{W}^*)^{-1}(\boldsymbol{X}^{*\top}\boldsymbol{X}^*)(\boldsymbol{X}^{*\top}\boldsymbol{X}^* + \boldsymbol{W}^*)^{-1}\hat{\sigma}^2/(1+\lambda_2)$ with

$$\boldsymbol{X}^{*\top}\boldsymbol{X}^* + \boldsymbol{W}^* = \boldsymbol{X}^{*\top}\left(\boldsymbol{I}_n + \frac{\boldsymbol{e}\boldsymbol{e}^\top}{\|\dot{\boldsymbol{\beta}}_n\|_1\|\boldsymbol{X}^{*\top}\boldsymbol{e}\|_\infty}\right)\boldsymbol{X}^*,$$

where $\boldsymbol{e} = \boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{b}$, and $\|\boldsymbol{\beta}\|_\infty = \max_{1\le j\le p}|\beta_j|$.

Proposition 2.2: *Under the assumptions of Proposition* 2.1, *given* $(\lambda_1, \lambda_2, d)$, *we have*

$$\dot{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left\{ \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{C}_n + \lambda_2 \boldsymbol{I}_p}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta} + \lambda_1 \|(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n)\|_1 \right\}.$$

Zou and Hastie [2] interpreted the E-net solution as a rescaled LASSO, which will improve prediction accuracy. Indeed, the term $(\frac{\boldsymbol{C}_n+\lambda_2\boldsymbol{I}_p}{1+\lambda_2})$ is a shrinkage version of $\boldsymbol{C}_n$, which the latter appears in LASSO. Here, the same interpretation is valid, where we replaced $\|\boldsymbol{\beta}\|_1$ by $\|\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n\|_1$ in LASSO.

Next, we will be considering an approximated closed-form solution to our optimization problem. This will pave the road to define the SLASSO after some modifications.

2.2. SLASSO

Recall the closed-form approximate solution to the optimization problem

$$\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

has form $(\boldsymbol{C}_n + \lambda \boldsymbol{W}^-)^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$, where $\boldsymbol{W}^-$ is the generalized inverse of $\boldsymbol{W} = \mathrm{diag}(|\hat{\beta}_j|)$, with $\widehat{\boldsymbol{\beta}}_n = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$, see [1]. Parallel to this, we have the following result for the proposed optimization problem.

Proposition 2.3: *Under the linear regression model* (1), *the approximated closed-form solution to the problem*

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \lambda_1, \lambda_2, d) = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|d\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

is given by

$$(\boldsymbol{C}_n + \lambda_2 \boldsymbol{I}_p + \lambda_1 \boldsymbol{W}^-)^{-1}(\boldsymbol{X}^\top \boldsymbol{X} + d\lambda_2 \boldsymbol{I}_p)\widehat{\boldsymbol{\beta}}_n, \tag{6}$$

where $\boldsymbol{W}^- = \mathrm{diag}(|\hat{\beta}_{1n}^L|^{-1}, \ldots, |\hat{\beta}_{pn}^L|^{-1})$ *and* $\widehat{\boldsymbol{\beta}}_n^L = (\hat{\beta}_{1n}^L, \ldots, \hat{\beta}_{pn}^L)^\top$.

In Proposition 2.3, let $\lambda_1 = \lambda_2 = 1$. Then, (6) reduces to

$$\begin{aligned} &(\boldsymbol{C}_n + \boldsymbol{I}_p + \boldsymbol{W}^-)^{-1}(\boldsymbol{X}^\top \boldsymbol{Y} + d\widehat{\boldsymbol{\beta}}_n) \\ &\quad = (\boldsymbol{C}_n + \boldsymbol{I}_p + \boldsymbol{W}^-)^{-1}(\boldsymbol{C}_n + d\boldsymbol{I}_p)\widehat{\boldsymbol{\beta}}_n, \end{aligned} \tag{7}$$

which is similar to the Liu estimator of Liu [7], except the coefficient $\boldsymbol{F}_n(d) = (\boldsymbol{C}_n + \boldsymbol{I}_p)^{-1}(\boldsymbol{C}_n + d\boldsymbol{I}_p)$ is replaced by $(\boldsymbol{C}_n + \boldsymbol{I}_p + \boldsymbol{W}^-)^{-1}(\boldsymbol{C}_n + d\boldsymbol{I}_p)$ here. In conclusion, the approximated closed-form solution to our problem shows that the effect of penalization due to L_1-norm appears in the Liu estimator by the term $\boldsymbol{W}^-$. To avoid inefficiency, we suggest to pre-multiply the term $\boldsymbol{F}_n(d)$ to the LASSO solution, for the proposal of SLASSO. This proposal can be also interpreted as rescaling the LASSO estimator to be multicollinear resistance.

Recall that the naïve look does not provide a closed-form solution to the biasing parameter. In this case, an approximated closed form is of interest. Similar to Tibshirani [1], one

may make use of $\sum b_j^2/|b_j|$, with $\boldsymbol{b} = (b_1, \ldots, b_p)^\top$, instead of the penalty term $\|\boldsymbol{b}\|_1$ to get the SLASSO, say, by modifying (8) as

$$\widehat{\boldsymbol{\beta}}_n^L(d) = \boldsymbol{F}_n(d)\widehat{\boldsymbol{\beta}}_n^L, \tag{8}$$

d偏差参数 偏差因子

where d is the biasing parameter and $\boldsymbol{F}_n(d) = (\boldsymbol{C}_n + \boldsymbol{I}_p)^{-1}(\boldsymbol{C}_n + d\boldsymbol{I}_p)$ is the biasing factor. Thus, one can obtain the SLASSO using the following pseudo code:

- Evaluate the lasso solution as $\widehat{\boldsymbol{\beta}}_n^L$ with optimal value of λ
- Apply formula (8) with d selected by a suitable method.

The forthcoming section is devoted to the properties of the SLASSO as defined by (8).

3. Theoretical considerations

In this section, we establish some properties of the SLASSO.

In the sequel, we will be assuming the following regularity conditions:

(A1) $\frac{1}{n}\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top \to \boldsymbol{C}$, where $\boldsymbol{C}$ is a non-negative definite matrix.

(A2) $\frac{1}{n}\max_{1\le i\le n}\boldsymbol{x}_i^\top\boldsymbol{x}_i \to 0$.

(A3) $\boldsymbol{F}_n(d) \to \boldsymbol{F}(d)$, where $\boldsymbol{F}(d) = (\boldsymbol{C} + \boldsymbol{I}_p)^{-1}(\boldsymbol{C} + d\boldsymbol{I}_p)$.

For our purpose, we assume that $\boldsymbol{C}$ is nonsingular. The following result provides the local asymptotic distribution of the SLASSO estimator.

Proposition 3.1: *Suppose* $\boldsymbol{\phi} = \widehat{\boldsymbol{\beta}}_n^L$ *is the minimizer of*

$$Z_n(\boldsymbol{\phi}) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\phi}\|_2^2 + \frac{1}{n}\lambda_n\|\boldsymbol{\phi}\|_1, \quad \boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^\top.$$

Under the set of local alternatives $\mathcal{K}_{(n)} : \boldsymbol{\beta} = \boldsymbol{\beta}_{(n)} = \frac{\boldsymbol{\delta}}{\sqrt{n}}$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q) \ne \boldsymbol{0}$, *assume (A1)–(A3). If* $\lambda/n \to \lambda_o \ge 0$, *then, we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^L(d) - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \boldsymbol{F}(d)\left[\arg\min_{\boldsymbol{u}} V(\boldsymbol{u}) + \boldsymbol{\delta}\right],$$

where $V(\boldsymbol{u}) = -2\boldsymbol{u}^\top\boldsymbol{W} + \boldsymbol{u}^\top\boldsymbol{C}\boldsymbol{u} + \lambda_o\sum_{j=1}^p[u_j\mathrm{sgn}(\beta_j)I(\beta_j \ne 0) + |u_j + \delta_j|I(\beta_j = 0)]$ *and* $\boldsymbol{W} \sim \mathcal{N}_p(\boldsymbol{0}, \sigma^2\boldsymbol{C})$.

3.1. Normalized LASSO

For the orthonormal design, i.e. $\boldsymbol{C}_n = \boldsymbol{I}_p$, the SLASSO simplifies to

$$\hat{\beta}_{jn}^L(d) = c_d\mathrm{sgn}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda/2)^+, \quad j = 1, \ldots, p,$$

where $a^+ = \max(0, a)$, $c_d = (1 + d)/2$, λ is determined by the condition $\sum|\hat{\beta}_j| = t$, and $\hat{\beta}_j$ is the j-th component of the OLS estimator. The estimator $\hat{\beta}_{jn}^L(d)$ is termed normalized LASSO, in our terminology.

Under normality assumption, some interesting properties can be achieved for the normalized LASSO. Hence, suppose that the error term in (1) has the normal distribution with zero mean and covariance matrix $\sigma^2 \boldsymbol{I}_n$, where σ^2 is known. Then, $\hat{\beta}_j/\sigma \sim \mathcal{N}(\Delta_j, 1)$, where $\Delta_j = \beta_j/\sigma$. The following result provides the upper bound for the risk of normalized LASSO.

Proposition 3.2: *For all* $\delta \leq \frac{1}{2}$ *and* $\lambda = 2\sigma\sqrt{2\log\delta^{-1}}$

$$\mathbb{E}\left[\hat{\beta}_{jn}^{L}(d) - \Delta_j\right]^2 \leq \sigma^2 c_d^2 (1 + 2\log\delta^{-1})[\delta + \min(\Delta_j^2, 1)] + (\sigma c_d - 1)^2 \Delta_j^2 - 2\sigma c_d(\sigma c_d - 1)\Delta_j(\lambda/2\sigma)\left[\Phi(\lambda/2\sigma - \Delta_j) - \Phi(\lambda/2\sigma + \Delta_j)\right],$$

where $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal distribution.*

By Proposition 3.2, as $\boldsymbol{\beta} \to 0$, we have $\mathbb{E}[\hat{\beta}_{jn}^{L}(d)]^2 \leq \sigma^2 c_d^2 \delta(1 + 2\log\delta^{-1})$. Further, for the biasing parameter d such that $\sigma c_d = 1$, i.e. $d = (2 - \sigma)/\sigma$, we get the following result.

Remark 3.1: For all $\delta \leq \frac{1}{2}$ and $\lambda = 2\sigma\sqrt{2\log\delta^{-1}}$

$$\mathbb{E}\left[\hat{\beta}_{jn}^{L}\left(\frac{2-\sigma}{\sigma}\right) - \Delta_j\right]^2 \leq (1 + 2\log\delta^{-1})[\delta + \min(\Delta_j^2, 1)].$$

The upper bound in Remark 3.1 coincides with Theorem 1 of Donoho and Johnstone [17] for the soft threshold estimator. Indeed, for $\sigma = 1$, the biasing parameter simplifies to $d = (2 - \sigma)/\sigma = 1$, which results in $\widehat{\boldsymbol{\beta}}_n^L(1) = \widehat{\boldsymbol{\beta}}_n^L$.

To seek for an optimal threshold yielding the smallest upper bound, the result of Theorem 2 of Donoho and Johnstone [17] is valid here.

4. Numerical studies

In this section, we compare the performance of the SLASSO with some other known estimators.

4.1. Illustration

In the following, we study real-life data. We use the Asphalt Binder data, which was analysed by Wei et al. [18], in which the data is provided in Table 1. This data consists of 23 core asphalt binders and 12 different explanatory variables such that the reference set or dependent variable, surface free energy (Y), is written as a function of the chemical composition variables referred to as saturates (X1), aromatics (X2), resins (X3), asphaltenes (X4), wax (X5), carbon (X6), hydrogen (X7), oxygen (X8), nitrogen (X9), sulfur (X10), nickel (X11), vanadium (X12). The predictors are standardized to have zero mean and unit standard deviation before fitting the model. We also centre the response variable. We then fit a linear regression model to predict the variables of interest using the available regressors.

To evaluate the prediction accuracy of the competitors, 5-fold cross-validation is used. In this procedure, the data set is randomly divided into five subsets of roughly equal size. Four subsets, called the training set, are used the fit the model. Also, the tuning parameters

Table 1. The MSE_y and the corresponding standard errors of estimators in parenthesis in the Asphalt Binder data.

OLS	Ridge	Liu	LASSO	SLASSO	E-net
10.1897	3.5538	4.1077	5.1459	3.2707	4.3029
(0.2681)	(0.0440)	(0.0618)	(0.1128)	(0.0519)	(0.0640)

are selected in this step. The fitted model is used to calculate the MSE_y from one remaining subset, called the test data set, where $MSE_y = \frac{1}{n_{test}}(\boldsymbol{Y}_{test} - \widehat{\boldsymbol{Y}})^\top(\boldsymbol{Y}_{test} - \widehat{\boldsymbol{Y}})$. This process is repeated for all five subsets, and the MSE_y values are combined. To account for the random variation of the cross-validation, the process is repeated 1000 times, and the mean MSE_y, as well as its standard error, are calculated. The results are given in Table 1. Analysing these results, it is readily observed that the SLASSO has the lowest predictive MSE value and has the second least standard error.

4.2. Simulation

The purpose of this section is to design a Monte Carlo simulation to show the superiority of the SLASSO over the estimators OLS, RR, Liu, LASSO, and E-net.

We used six examples some of which were also considered in Zou and Hastie [2]. All simulations are based on the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\epsilon$$

where $\epsilon \sim N(\boldsymbol{0}, \boldsymbol{I})$. The simulated data contains a training dataset, validation data, and an independent test set in each example. We choose the tuning parameters using a validation set. Using the glmnet function in R, we first fit the model using the train set and choose the best model with the lowest predictive mean-squared error in the validation set. Therefore, we use the sequence supplied by the glmnet for the ridge, lasso, and elastic net. However, for the Liu estimator and SLASSO, we provide a sequence of the parameter value d such as a sequence of 250 numbers between -10 and 10, and again, we choose d similarly as above. In simulations, we centre all variables based on the training data set. Let $\bar{\boldsymbol{x}}_{train} = (\bar{x}_{1,train}, \ldots, \bar{x}_{p,train})$ denote the vector of means of the training data, n_{test} the number of observations in the test data set and $\bar{y}_{train}$ the mean over responses in the training data. Finally, we computed two measures of performance, the test error (mean-squared error) $MSE_y = \frac{1}{n_{test}}\mathbf{r}_{sim}^\top\mathbf{r}_{sim}$ where $\mathbf{r}_{sim} = \boldsymbol{x}_i\boldsymbol{\beta} - (\bar{y}_{train} + (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{train})^\top\widehat{\boldsymbol{\beta}}))$ and the mean-squared error of the estimation of $\boldsymbol{\beta}$ such that $MSE_\beta = |\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}|^2$, see Tutz and Ulbricht [19]. We use the notation $\cdot/\cdot/\cdot$ to describe the number of observations in the training, validation and test set, respectively. Here are the details of five examples:

(1) Each data set consists of 20/20/200 observations. $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}^\top = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. Also, we generate $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}$.

(2) Each data set consists of 20/20/200 observations. $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}^\top = (\underbrace{0.85, 0.85, \ldots, 0.85}_{8})$ and $\sigma = 3$. Also, we generate $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}$.

(3) This example is similar to the first example except that $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}^\top = (3, 1.5, 0, 0, 0, 0, -1, -1)$.

Table 2. Median mean-squared errors for the simulated examples and five methods based on 1000 replications.[a]

Estimator	Example	MSE_y	MSE_β	Example	MSE_y	MSE_β
OLS		6.621 (0.178)	9.597 (0.294)		54.990 (0.793)	89.072 (1.348)
Ridge		3.352 (0.076)	3.769 (0.071)		13.171 (0.165)	11.496 (0.109)
Liu		3.183 (0.072)	3.567 (0.077)		19.310 (0.235)	25.041 (0.296)
LASSO		2.896 (0.078)	3.002 (0.087)		11.899 (0.181)	12.907 (0.180)
SLASSO		2.669 (0.072)	2.866 (0.079)		10.905 (0.168)	9.864 (0.125)
E-net	1	2.862 (0.073)	2.990 (0.077)	4	11.213(0.167)	10.304(0.133)
OLS		6.304 (0.172)	9.469 (0.317)		5.962 (0.058)	9.201 (0.098)
Ridge		2.024 (0.048)	1.638 (0.031)		4.235 (0.039)	5.449 (0.047)
Liu		2.129 (0.047)	1.937 (0.046)		4.008 (0.037)	5.139 (0.045)
LASSO		3.337 (0.056)	3.892 (0.052)		3.475 (0.040)	4.637 (0.059)
SLASSO		2.207 (0.050)	1.765 (0.033)		3.382 (0.039)	4.417 (0.049)
E-net	2	2.186 (0.048)	1.943 (0.039)	5	3.456 (0.038)	4.379 (0.049)
OLS		7.049 (0.236)	9.746 (0.346)		6.055 (0.173)	45.492 (1.585)
Ridge		3.349 (0.069)	3.461 (0.063)		1.869 (0.046)	6.447 (0.089)
Liu		3.065 (0.067)	3.073 (0.069)		1.939 (0.049)	6.911 (0.186)
LASSO		3.001 (0.068)	3.078 (0.069)		1.925 (0.051)	8.763 (0.202)
SLASSO		2.850 (0.064)	2.739 (0.062)		1.578 (0.045)	5.589 (0.127)
E-net	3	2.899 (0.064)	2.940 (0.061)	6	1.697 (0.046)	6.594 (0.128)

[a]The numbers in between the parenthesis are the corresponding standard errors of the MSE.

(4) Each data set consists of 50/50/200 observations and 30 predictors. We chose

$$\boldsymbol{\beta}^\top = \left(\underbrace{2, \ldots, 2}_{8}, \underbrace{0, \ldots, 0}_{22} \right).$$

Also, we consider $\sigma = 6$ and $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}$.

(5) Each data set consists of 100/100/400 observations and 40 predictors.

$$\boldsymbol{\beta}^\top = \left(\underbrace{0, \ldots, 0}_{10} \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10} \underbrace{2, \ldots, 2}_{10} \right).$$

Also, we consider $\sigma = 3$ and $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5^{|i-j|}$.

(6) This example is similar to the first one except that $\Sigma_{ij} = 0.9^{|i-j|}$.

We investigate these scenarios by simulating 1000 data sets. The results of the simulation are given in Table 2. We also summarize the results in Figure 1 in which we present the box plots of test mean-squared errors MSE_y (left column) and MSE_β (right column) for Examples 1–6. Now, we share the results obtained from the simulation study as follows:

- In all the examples, the OLS has the worst performance according to both criteria.
- In Example 1, with positively correlated variables, although the performances of the estimators LASSO and E-net are close to each other, the SLASSO has the best performance in the sense of both measures. The RR and Liu estimators perform better than the OLS.
- In Example 2, Ridge has the smallest median MSE_y, and the smallest median MSE_β. The SLASSO has a better performance than LASSO.

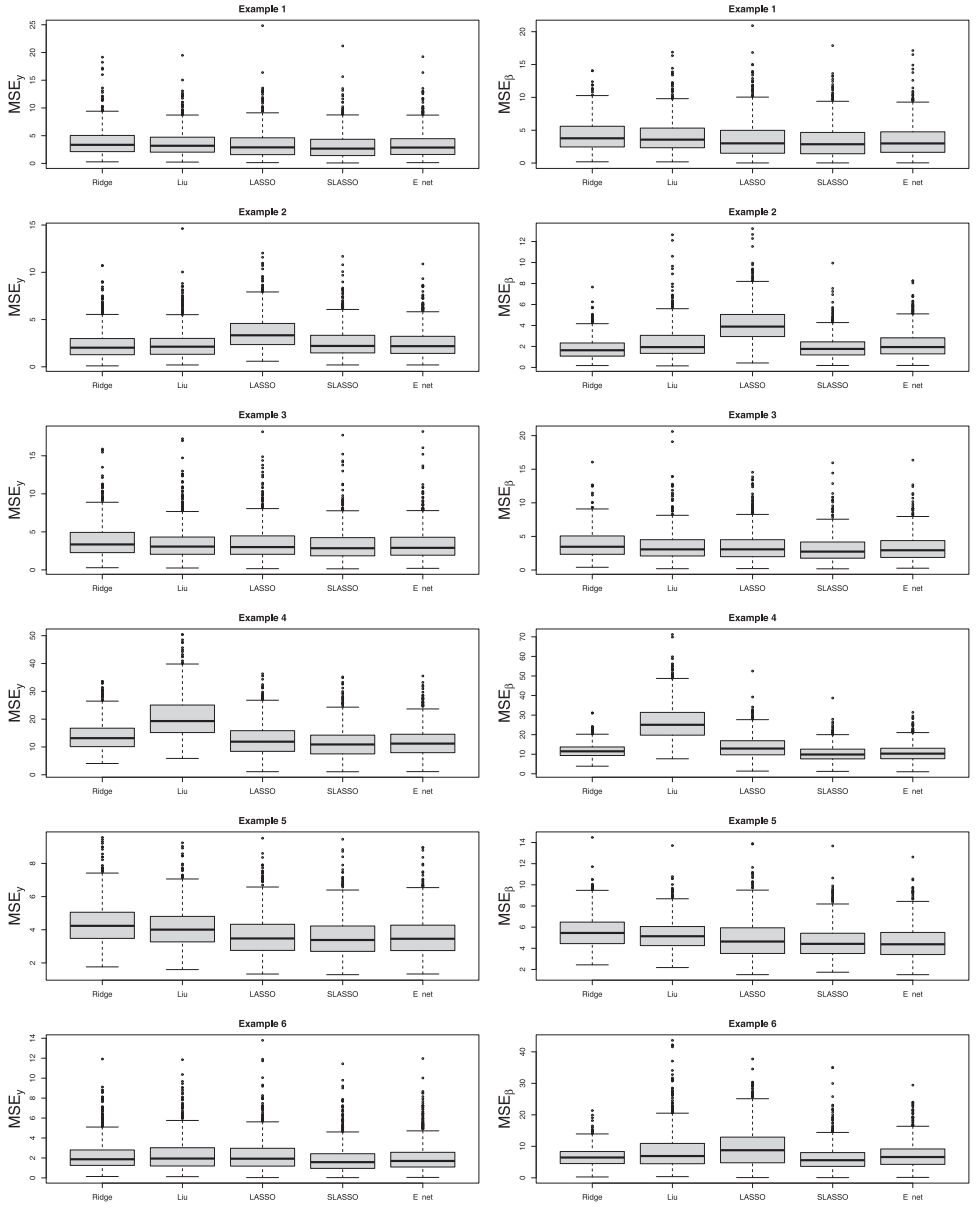

Figure 1. The Box plots of the test set mean-squared errors MSE_y (left column) and MSE_β (right column) for Examples 1–5.

- In Example 3, there are negative signs in the last two parameters, and we observe that the SLASSO has the best performance in the sense of both criteria. Interestingly, Liu and LASSO have very close median MSEs in this setting.
- In Example 4, there are 40 explanatory variables, and 22 of them are noise. The SLASSO outperforms the others in both senses. E-net also has a promising performance. Although RR is better than LASSO in the sense of MSE_β, LASSO has a better predictive performance than Ridge. The Liu estimator is the second worse based on both criteria.

Table 3. The Timing of different algorithms based on 100 simulations.

Estimator	Example	Time elapsed	Relative	Example	Time elapsed	Relative
OLS		0.098	1.000		0.084	1.000
Ridge		4.397	44.867		4.924	58.619
Liu		0.515	5.255		1.325	15.774
LASSO		3.918	39.980		4.782	56.929
SLASSO		4.707	48.031		5.785	68.869
E-net	1	41.824	426.776	4	45.441	540.964
OLS		0.070	1.000		0.117	1.000
Ridge		4.130	59.000		5.302	45.316
Liu		0.528	7.543		2.334	19.949
LASSO		4.213	60.186		5.085	43.462
SLASSO		4.635	66.214		7.094	60.632
E-net	2	40.354	576.486	5	48.424	413.880
OLS		0.077	1.000		0.079	1.000
Ridge		4.307	55.935		4.076	51.595
Liu		0.523	6.792		0.497	6.291
LASSO		4.077	52.948		3.991	50.519
SLASSO		4.407	57.234		4.462	56.481
E-net	3	39.756	516.312	6	39.202	496.228

- In Example 5, we consider the grouping effect. LASSO, SLASSO, and E-net have pretty comparable performance. Although SLASSO becomes the best in terms of MSE_y, the E-net performs the best while SLASSO becomes the second.
- In Example 6, we consider the design matrix having the problem of multicollinearity such that the correlations between the predictors are chosen to be 0.9. The SLASSO has the best performance among all the competitors again. Not surprisingly, the Ridge estimator performs better than LASSO while E-net beats the Ridge in terms of predictive performance but not the median MSE_β. The Liu estimator also has better results than the OLS, but it is the second-worst estimator.

4.3. Timings

In this section, we discuss the timings of each algorithm to compute the estimators. We use the glmnet package to obtain the estimators and benchmark package to compute the running time of each method for 100 simulations. Here d and λ are selected from a sequence of 100 values. All timings were carried out on a computer with a 2.3 GHz Intel Core i7 processor and 16 GB 1600 MHz DDR3 memory. We report the results of timings in Table 3. We provide the elapsed time in seconds and the relative time relative to the OLS for each method. According to Table 3, due to the two-tuning parameters of E-net, it takes a longer time to obtain E-net in each of the experiments. Although LASSO, SLASSO, and Ridge have close running time, LASSO has the lowest computation time. Moreover, the OLS has the lowest running time which is also expected since it can only be calculated by matrix multiplication.

5. Conclusion

In this article, we have proposed a new estimator for simultaneous estimation and variable selection. Indeed, we pre-multiplied the LASSO with a matrix factor to become multicollinear resistant after modifying the L_1-norm of the LASSO. The proposed scaled LASSO,

or SLASSO for short, has a simple form and can be considered a re-scaled LASSO estimator. The SLASSO inherits all good properties of the LASSO, and it is $\sqrt{n}$-consistent. Apart from its good properties, e.g. producing a sparse model with good prediction accuracy, there is no need to propose a specific algorithm for its computation. Like adaptive LASSO, the SLASSO can be solved using the same efficient algorithm for solving the LASSO. According to the numerical findings, we suggest using the SLASSO estimation method in practical examples.

For further research, it can be suggested to pre-multiply the relaxed LASSO [20] by the term $\boldsymbol{F}_n(d)$ for a faster convergence rate. Any sparse solution of high-dimensional problems can be also substituted with LASSO in our methodology. To construct an estimator with oracle properties, we suggest to use the adaptive LASSO instead of LASSO in the SLASSO. One can also designate the generalized SLASSO. It is defined by

$$\widehat{\boldsymbol{\beta}}_n^{LL} = (\boldsymbol{C}_n + \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top \boldsymbol{Y} + \boldsymbol{D}\widehat{\boldsymbol{\beta}}_n^{L}) = \boldsymbol{F}_D \widehat{\boldsymbol{\beta}}_n^{L}, \tag{9}$$

where $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_p)$ is the biasing matrix and $\boldsymbol{F}_D = (\boldsymbol{C}_n + \boldsymbol{I}_p)^{-1}(\boldsymbol{C}_n + \boldsymbol{D})$ is the biasing factor. The generalized SLASSO allows different biasing parameters.

Acknowledgments

The authors would like to sincerely thank the Editor, Associate Editor, and anonymous reviewers for their constructive and insightful comments, which significantly improved the presentation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Ferdowsi University of Mashhad [N.2/54466].

ORCID

Mohammad Arashi http://orcid.org/0000-0002-5881-9241
Yasin Asar http://orcid.org/0000-0003-1109-8456
Bahadır Yüzbaşı http://orcid.org/0000-0002-6196-3201

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol). 1996;58:267–288.
[2] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc: Ser B (Stat Methodol). 2005;67(2):301–320.
[3] Tikhonov A. Solution of incorrectly formulated problems and the regularization method. Soviet Meth Dokl. 1963;4:1035–1038.
[4] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.
[5] Yüzbaşı B, Arashi M, Ejaz Ahmed S. Shrinkage estimation strategies in generalised ridge regression models: low/high-dimension regime. Int Stat Rev. 2020;88(1):229–251.
[6] Mayer LS, Willke TA. On biased estimation in linear models. Technometrics. 1973;15(3): 497–508.

[7] Liu K. A new class of blased estimate in linear regression. Commun Stat-Theor Meth. 1993;22(2):393–402.
[8] Liu XQ. Improved Liu estimator in a linear regression model. J Stat Plan Inference. 2011;141(1):189–196.
[9] Mansson K, Kibria BG, Shukur G. On Liu estimators for the logit regression model. Econ Model. 2012;29(4):1483–1488.
[10] Duran EA, Hardle WK, Osipenko M. Difference based ridge and Liu type estimators in semiparametric regression models. J Multivar Anal. 2012;105(1):164–175.
[11] Arashi M, Kibria BG, Norouzirad M, et al. Improved preliminary test and Stein-rule Liu estimators for the ill-conditioned elliptical linear regression model. J Multivar Anal. 2014;126:53–74.
[12] Arashi M, Norouzirad ML, Ahmed SE, et al. Rank-based Liu regression. Comput Stat. 2018;33(3):1525–1561.
[13] Wu J. Modified restricted Liu estimator in logistic regression model. Comput Stat. 2016;31(4):1557–1567.
[14] Arashi M, Asar Y, Yüzbaşı BLLASSO: a linear unified LASSO for multicollinear situations. 2017. arXiv:1710.04795.
[15] Genc M, Ozkale MR. Usage of the GO estimator in high dimensional linear models. Comput Stat. 2021;36:217–239.
[16] Osborne MR, Presnell B, Turlach BA. On the LASSO and its dual. J Comput Graph Stat. 2000;9(2):319–337.
[17] Donoho DL, Johnstone JM. Ideal spatial adaptation by wavelet shrinkage. Biometrika. 1994;81(3):425–455.
[18] Wei J, Dong F, Li Y, et al. Relationship analysis between surface free energy and chemical composition of asphalt binder. Construction and Building Mater. 2014;71:116–123.
[19] Tutz G, Ulbricht J. Penalized regression with correlation-based penalty. Stat Computing. 2009;19(3):239–253.
[20] Meinshausen N. Relaxed Lasso. Comput Stat Data Anal. 2007;52(1):374–393.
[21] Knight K, Fu W. Asymptotics for Lasso-type estimators. Ann Stat. 2000;28(5):1356–1378.

Appendix: Proofs

Proof of Proposition 2.1: As in the assumption, let

$$\boldsymbol{X}^* = (1+\lambda_2)^{-\frac{1}{2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2}\boldsymbol{I}_p \end{pmatrix}, \quad \boldsymbol{Y}^* = \begin{pmatrix} (1-d)\boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}$$

and $\boldsymbol{b} = \sqrt{1+\lambda_2}(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n)$. First, note that

$$\begin{aligned}\mathcal{L}(\boldsymbol{b};\gamma) = (1-d)^2\boldsymbol{Y}^\top\boldsymbol{Y} - 2(1-d)(1+\lambda_2)^{-\frac{1}{2}}\boldsymbol{b}^\top\boldsymbol{X}^\top\boldsymbol{Y} + (1+\lambda_2)^{-1}\boldsymbol{b}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b} \\ + (1+\lambda_2)^{-1}\lambda_2\boldsymbol{b}^\top\boldsymbol{b} + \gamma\|\boldsymbol{b}\|_1.\end{aligned}$$

Then, differentiating $\mathcal{L}(\boldsymbol{b};\gamma)$ w.r.t. $\boldsymbol{b}$ and equating the result to zero give

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{b}}\mathcal{L}(\boldsymbol{b};\gamma) = -2(1-d)(1+\lambda_2)^{-\frac{1}{2}}\boldsymbol{X}^\top\boldsymbol{Y} + 2(1+\lambda_2)^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b} + 2(1+\lambda_2)^{-1}\lambda_2\boldsymbol{b} + \gamma\frac{\partial}{\partial\boldsymbol{b}}\|\boldsymbol{b}\|_1 = \boldsymbol{0} \\ \vdash -2(1-d)\boldsymbol{X}^\top\boldsymbol{Y} + 2(1+\lambda_2)^{-\frac{1}{2}}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b} + 2(1+\lambda_2)^{-\frac{1}{2}}\lambda_2\boldsymbol{b} + \lambda_1\mathbf{sgn}(\boldsymbol{b}) = \boldsymbol{0} \\ \vdash -2\boldsymbol{X}^\top\boldsymbol{Y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} - 2\lambda_2 d\widehat{\boldsymbol{\beta}}_n + 2\lambda_2\boldsymbol{\beta} + \lambda_1\mathbf{sgn}(\boldsymbol{b}) = \boldsymbol{0} \qquad \text{(A1)}\end{aligned}$$

where $\mathbf{sgn}(\boldsymbol{b}) = (\mathrm{sgn}(b_1), \ldots, \mathrm{sgn}(b_p))^\top$, with $\boldsymbol{b} = (b_1, \ldots, b_p)^\top$.

As in the assumption, we select d such that $\mathbf{sgn}(\boldsymbol{b}) = \mathbf{sgn}(\boldsymbol{\beta})$. Thus, (A1) implies

$$\begin{aligned}-2\boldsymbol{X}^\top\boldsymbol{Y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} - 2\lambda_2 d\widehat{\boldsymbol{\beta}}_n + 2\lambda_2\boldsymbol{\beta} + \lambda_1\mathbf{sgn}(\boldsymbol{\beta}) = \boldsymbol{0} \\ \vdash \frac{\partial}{\partial\boldsymbol{\beta}}L(\boldsymbol{\beta};\lambda_1,\lambda_2,d) = \boldsymbol{0}\end{aligned}$$

and the proof is complete. ■

Proof of Proposition 2.2: Under the assumptions of Proposition 1, we get

$$\begin{aligned}\dot{\boldsymbol{\beta}}_n &= \arg\min_{\boldsymbol{b}} \left\{ \left\| \boldsymbol{Y}^* - \boldsymbol{X}^* \frac{\boldsymbol{b}}{\sqrt{1+\lambda_2}} \right\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \left\| \frac{\boldsymbol{b}}{\sqrt{1+\lambda_2}} \right\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \left\| \boldsymbol{Y}^* - \boldsymbol{X}^*(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}) \right\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \left\| (\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}) \right\|_1 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n)^\top \boldsymbol{X}^{*\top} \boldsymbol{X}^* (\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n) - 2\boldsymbol{Y}^{*\top} \boldsymbol{X}^* \boldsymbol{\beta} + \boldsymbol{Y}^{*\top} \boldsymbol{Y}^* \right. \\ &\quad \left. \frac{\lambda_1 \|(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n)\|_1}{1+\lambda_2} \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \boldsymbol{I}_p}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\boldsymbol{Y}^\top \boldsymbol{X} \boldsymbol{\beta} + \lambda_1 \|(\boldsymbol{\beta} - d\widehat{\boldsymbol{\beta}}_n)\|_1 \right\}.\end{aligned}$$

The proof is complete. ■

Proof of Proposition 2.3: The approximate closed-form solution of $\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \lambda_1, \lambda_2, d)$ is obtained by writing the penalty term $\|\boldsymbol{\beta}\|_1$ as $\sum_{i=1}^p \beta_i^2/|\beta_j|$. Using the LASSO solution (2), we can approximate

$$L(\boldsymbol{\beta}; \lambda_1, \lambda_2, d) \approx \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|d\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|_2^2 + \lambda_1 \sum_{i=1}^p \frac{\beta_i^2}{|\hat{\beta}_{in}^L|}.$$

Differentiating from the L.H.S. of the above equality w.r.t. $\boldsymbol{\beta}$, yields

$$\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}; \lambda_1, \lambda_2, d) = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\lambda_2 d\widehat{\boldsymbol{\beta}}_n + 2\lambda_2 \boldsymbol{\beta} + 2\lambda_1 \boldsymbol{W}^- \boldsymbol{\beta}.$$

Solving $\partial L(\boldsymbol{\beta}; \lambda_1, \lambda_2, d)/\partial \boldsymbol{\beta} = \boldsymbol{0}$ w.r.t. $\boldsymbol{\beta}$ and using the fact that

$$\begin{aligned}&(\boldsymbol{C}_n + \lambda_2 \boldsymbol{I}_p + \lambda_1 \boldsymbol{W}^-)^{-1} (\boldsymbol{X}^\top \boldsymbol{Y} + d\lambda_2 \widehat{\boldsymbol{\beta}}_n) \\ &\quad = (\boldsymbol{C}_n + \lambda_2 \boldsymbol{I}_p + \lambda_1 \boldsymbol{W}^-)^{-1} (\boldsymbol{X}^\top \boldsymbol{X} + d\lambda_2 \boldsymbol{I}_p) \widehat{\boldsymbol{\beta}}_n\end{aligned}$$

complete the proof. ■

Proof of Proposition 3.1: Note that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^L(d) - \boldsymbol{\beta}) = \sqrt{n}\boldsymbol{F}_n(d)(\widehat{\boldsymbol{\beta}}_n^L - \boldsymbol{\beta}) + \sqrt{n}(\boldsymbol{F}_n(d) - \boldsymbol{I}_p)\boldsymbol{\beta}$. Under $\mathcal{K}_{(n)}$ and (A3), $\sqrt{n}(\boldsymbol{F}_n(d) - \boldsymbol{I}_p)\boldsymbol{\beta} \to \boldsymbol{F}(d)\boldsymbol{\delta}$. Also, using Theorem 2 of [21], $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n^L - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \arg\min_{\boldsymbol{u}} V(\boldsymbol{u})$. Then, the result follows from Slutsky's theorem. ■

Proof of Proposition 3.2: Let $Z_j = \hat{\beta}_j/\sigma$. Then $Z_j \sim \mathcal{N}(\Delta_j, 1)$, $\Delta_j = \beta_j/\sigma$, and we have

$$\hat{\beta}_{jn}^L(d) = \sigma c_d \mathrm{sgn}(Z_j)(|Z_j| - \lambda/2\sigma)^+.$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_{jn}^L(d) - \Delta_j\right]^2 &= \sigma^2 c_d^2 \mathbb{E}\left[\mathrm{sgn}(Z_j)(|Z_j| - \lambda/2\sigma)^+ - \Delta_j\right]^2 + (\sigma c_d - 1)^2 \Delta_j^2 \\ &\quad + 2\sigma c_d(\sigma c_d - 1)\Delta_j \mathbb{E}\left[\mathrm{sgn}(Z_j)(|Z_j| - \lambda/2\sigma)^+ - \Delta_j\right].\end{aligned}$$

After some algebra

$$\begin{aligned}\mathbb{E}\left[\operatorname{sgn}(Z_j)(|Z_j|-\lambda/2\sigma)^+\right] &= \mathbb{E}\left[\operatorname{sgn}(Z_j)(|Z_j|-\lambda/2\sigma)I(|Z_j|>\lambda/2\sigma)\right]\\ &= \mathbb{E}\left[Z_j I(|Z_j|>\lambda/2\sigma)\right]-\mathbb{E}\left[(\lambda/2\sigma)\operatorname{sgn}(Z_j)I(|Z_j|>\lambda/2\sigma)\right]\\ &= \Delta_j-(\lambda/2\sigma)\left[\Phi(\lambda/2\sigma-\Delta_j)-\Phi(\lambda/2\sigma+\Delta_j)\right]. \qquad \text{(A2)}\end{aligned}$$

On the other hand, using Theorem 1 of Donoho and Johnstone [17]

$$\mathbb{E}\left[\operatorname{sgn}(Z_j)(|Z_j|-\lambda/2\sigma)^+-\Delta_j\right]^2 \le (1+2\log\delta^{-1})[\delta+\min(\Delta_j^2,1)]. \qquad \text{(A3)}$$

Using (A2) together with (A3) yield

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_{jn}^L(d)-\Delta_j\right]^2 \le{}& \sigma^2 c_d^2(1+2\log\delta^{-1})[\delta+\min(\Delta_j^2,1)]+(\sigma c_d-1)^2\Delta_j^2\\ &-2\sigma c_d(\sigma c_d-1)\Delta_j(\lambda/2\sigma)\left[\Phi(\lambda/2\sigma-\Delta_j)-\Phi(\lambda/2\sigma+\Delta_j)\right]\end{aligned}$$

which completes the proof. ■