Journal of the Korean Data & Information Science Society
2015, **26**(2), 505–516

http://dx.doi.org/10.7465/jkdi.2015.26.2.505
한국데이터정보과학회지

A note on standardization in penalized regressions[†]

Sangin Lee[1]

[1]Department of Clinical Sciences, University of Texas Southwestern Medical Center

Received 20 January 2015, revised 30 January 2015, accepted 16 February 2015

Abstract

We consider sparse high-dimensional linear regression models. Penalized regressions have been used as effective methods for variable selection and estimation in high-dimensional models. In penalized regressions, it is common practice to standardize variables before fitting a penalized model and then fit a penalized model with standardized variables. Finally, the estimated coefficients from a penalized model are recovered to the scale on original variables. However, these procedures produce a slightly different solution compared to the corresponding original penalized problem. In this paper, we investigate issues on the standardization of variables in penalized regressions and formulate the definition of the standardized penalized estimator. In addition, we compare the original penalized estimator with the standardized penalized estimator through simulation studies and real data analysis.

Keywords: LASSO, nonconvex penalties, penalized regression, standardization.

1. Introduction

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ is the vector of n response variables, $\mathbf{X}$ is the $n \times p$ design matrix with the ith row $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is the vector of random errors. For simplicity we assume that both response and predictive variables are centered such that $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n y_i = 0$. This eliminates the necessity of an intercept in the model. In many recent applications, there are a large number of potential predictors in comparison to the sample size. In this case $p >> n$, the classical linear regression using a least square method fails because of the singularity of the design matrix. To deal with high dimensional situations, many penalized regression methods have been proposed in recent years. The popular examples in penalized regressions are the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), smoothly clipped absolute deviation (SCAD) of Fan and Li (2001), and minimax concave penalty (MCP) of Zhang (2010). Penalized regressions perform the variable selection and

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(No. 2013R1A6A3A03026588).

[1] Postdoctoral researcher, Department of Clinical Sciences, Quantitative Biomedical Research Center, UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA.
E-mail : sanginlee44@gmail.com

parameter estimation simultaneously and achieve higher prediction accuracy than classical variable selection methods. Penalized linear regression approaches minimize an objective function that is composed of the sum of squared residuals and a penalty function,

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p} J_\lambda(|\beta_j|), \tag{1.2}$$

where $J_\lambda(\cdot)$ is a penalty function, and $\lambda \geq 0$ is the regularization parameter. In addition to linear regression models, the idea of the penalized regressions has been broadly applied to various statistical models and problems; generalized linear models (Van de Geer, 2008), Cox proportional hazard models (Fan and Li, 2002), Gaussian graphical models (Friedman *et al.*, 2013), principal component analysis (Park, 2013) and high-dimensional clustering problems (Kwon *et al.*, 2013).

In many regression problems, it is common practice to standardize predictors for the purpose of adjusting different scales of predictors. In the un-penalized problem, (i.e. least square method), it is equivalent to the problem with the standardized predictors. In other words, standardized coefficients can be recovered by the corresponding original coefficients without any additional fitting process. However, this relationship between standardized and original coefficients is not achieved in penalized regressions.

To obtain standardized coefficients in penalized regressions, one may perform the standardization of predictors before fitting a penalized model so that the penalty is equally imposed to each coefficient. Then a fitting algorithm for the penalized regression is applied to the penalized model with standardized predictors. Finally, the estimated coefficients are transformed back to their original scales. However, it is easy to show that the solution obtained from these procedures is not the minimizer of $Q_\lambda(\boldsymbol{\beta})$ which is the original penalized objective function. Unlike the un-penalized problem, the standardization in penalized regressions changes the original minimization problem and results in different estimates of $\boldsymbol{\beta}$. In many available R packages, `lars` and `glmnet` for the LASSO, there are two options (`T`, `F`) for the standardization of predictors (Efron *et al.*, 2004; Friedman *et al.*, 2010), whereas the `ncvreg` package for the nonconvex penalties such as SCAD and MCP does not provide any option for the standardization of predictors. In fact, the `ncvreg` package only provides standardized penalized estimates (Breheny and Huang, 2011). Throughout this paper, we call two types of estimators *original penalized estimator* and *standardized penalized estimator*. Many researchers have used two types of estimators without any distinction of the specific definitions in penalized regressions.

In this paper, we study issues on standardization of predictors in penalized linear regressions and formulate the minimization problem for the standardized penalized estimator. We show how to obtain two types of estimators in the `glmnet` and `ncvreg` packages using coordinate descent (CD) algorithms (Friedman *et al.*, 2010; Breheny and Huang, 2011). As mentioned above, the `ncvreg` package only provides the standardized penalized estimate. We demonstrate a reason to only provide standardized solutions in the `ncvreg` package, and introduce an optimization algorithm for obtaining the original nonconvex penalized estimator.

This paper is organized as follows. In Section 2, we describe two types of CD algorithms for the original and standardized LASSO estimators, and introduce the definition of the standardized LASSO estimator. In Section 3, we investigate issues on standardization of

predictors in nonconvex penalized regressions. In Section 4 and 5, we compare the standardized penalized estimator with the original penalized estimator through various simulations and real data analysis. Concluding remarks are provided in Section 6.

2. Standardization in ℓ_1-penalized regressions

The original LASSO estimator is defined as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.1}$$

where $\lambda \geq 0$ is the regularization parameter. The formulation (2.1) uses the same regularization parameter λ for all predictors. When variations of each predictor are substantially different, the original LASSO estimator in (2.1) may not be reasonable. In fact, the original paper introduced by Tibshirani (1996) begins with the standardization assumption of predictors and describes the LASSO method. In penalized regressions, to deal with different scales of predictors, it is common practice to standardize predictors before applying the LASSO method, which ensures that the penalty is applied equally to all predictors in terms of unit variance of all predictors. The estimates are then transformed back to their original scales after fitting the LASSO model with standardized predictors. The standardized solution from these procedures can be obtained by giving the option for standardization, `standardize=T` in the `glmnet` package, whereas the option `standardize=F` provides the original LASSO estimate in (2.1).

We briefly describe two implementations in the `glmnet` package using the CD algorithm (Friedman *et al.*, 2010). Consider the jth coordinate descent step for solving (2.1). For a given fixed values of parameters $(\tilde{\beta}_k,\ k \neq j)$ at their current solutions $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$, we wish to minimize the objective function in (2.1) with respect to the jth coefficient β_j. It can be shown that this problem is equivalent to minimizing $q_\lambda(\beta_j|\cdot)$ defined as

$$q_\lambda(\beta_j|\tilde{\boldsymbol{\beta}}) = \frac{1}{2n} \sum_{i=1}^{n} (r_i - x_{ij}\beta_j)^2 + \lambda \sum_{k \neq j} |\tilde{\beta}_k| + \lambda|\beta_j|, \tag{2.2}$$

where $r_i = y_i - \sum_{k \neq j} x_{ik}\tilde{\beta}_k$ are the partial residuals. The minimizer of $q_\lambda(\beta_j|\tilde{\boldsymbol{\beta}})$ in (2.2) has the explicit form as

$$\hat{\beta}_j = \begin{cases} \text{sign}(z_j)(|z_j| - \lambda)_+ / c_j & \text{if } x_{ij} \text{ are unstandardized,} \\ \text{sign}(z_j)(|z_j| - \lambda)_+ & \text{if } x_{ij} \text{ are standardized,} \end{cases} \tag{2.3}$$

where the subscript '+' indicates the positive part, $z_j = \sum_{i=1}^{n} x_{ij} r_i / n$ is the simple least-square fit and $c_j = \sum_{i=1}^{n} x_{ij}^2 / n$. This explicit form of solutions facilitates the implementation of the CD algorithms summarized in Algorithm 2.1 and 2.2.

As mentioned earlier, Algorithm 2.1 and 2.2 provide the different estimates of $\boldsymbol{\beta}$, which means that the standardization of predictors in the penalized regression changes the original minimization problem in (2.1). In fact, Algorithm 2.2 minimizes the following objective

Algorithm 2.1 The CD algorithm for the original LASSO estimator

Set an initial estimate $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$
Compute c_j for $j = 1, \ldots, p$.
repeat
  for $j = 1, 2, \ldots, p$
    Compute the simple least-square fit z_j in (2.3)
    Update $\tilde{\beta}_j$ with $\hat{\beta}_j = \text{sign}(z_j)(|z_j| - \lambda)_+/c_j$
until convergence

Algorithm 2.2 The CD algorithm for the standardized LASSO estimator

Compute the standardized predictors $x^*_{ij} = x_{ij}/s_j$ in (2.4)
Set an initial estimate $\tilde{\boldsymbol{\beta}}^* \in \mathbb{R}^p$
repeat
  for $j = 1, 2, \ldots, p$
    Compute the simple least-square fit z^*_j with the standardized predictors x^*_{ij} in (2.3)
    Update $\tilde{\beta}^*_j$ with $\hat{\beta}^*_j = \text{sign}(z^*_j)(|z^*_j| - \lambda)_+$
until convergence
Transform the solutions by $\hat{\beta}_j = \hat{\beta}^*_j/s_j$

function $S_\lambda(\boldsymbol{\beta})$ in (2.4), and hence the standardized LASSO estimator can be defined as the minimizer of

$$S_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p} s_j|\beta_j|, \tag{2.4}$$

where $s_j = \sqrt{\sum_{i=1}^n x_{ij}^2/n}$ is the standard deviation of the jth predictor. The formulation (2.4) uses different regularization parameters $\lambda_j = \lambda s_j$ for each variable, that is, different amounts of penalty are imposed to each coefficient. In contrast, the original LASSO in (2.1) forces the coefficients to be equally penalized regardless of scales of predictors, which might be unreasonable in the case with heteroscedastic variances of predictors. For example, the original LASSO tends to select variables with high variance even if these are irrelevant variables in the underlying model, while the standardized LASSO successfully deletes irrelevant variables with high variance by imposing more amounts of penalty, see Section 4.

3. Nonconvex penalized estimation

In this section, we investigate issues on standardization of predictors in nonconvex penalized regressions. We first introduce the definition of the standardized nonconvex penalized estimator and describe the CD algorithm of Breheny and Huang (2011). We then demonstrate a reason to only provide the standardized estimates in the `ncvreg` package, and introduce an optimization algorithm to compute the original estimates.

3.1. Coordinate descent algorithm

We first consider the class of nonconvex penalties satisfy the following three conditions (Kim and Kwon, 2012).

(C1) $J'_\lambda(t)$ is nonnegative, nonincreasing and continuous over $(0, \infty)$,

(C2) $\lim_{t\to 0+} J'_\lambda(t) = \lambda$ and $J'_\lambda(t) = 0$ for $t \geq a\lambda$,

(C3) $J'_\lambda(t) \geq (\lambda - t/a)_+ I(0 < t < a\lambda)$ for $t > 0$,

for some $a > 0$, where $J'_\lambda(t)$ is the first derivative of $J_\lambda(t)$ with respect to t. This class includes the SCAD penalty of Fan and Li (2001)

$$J_\lambda(t) = \begin{cases} \lambda t, & \text{if } t \leq \lambda, \\ \{a\lambda(t-\lambda) - (t^2 - \lambda^2)/2\}/(a-1), & \text{if } t \in (\lambda, a\lambda], \\ (a-1)\lambda^2/2 + \lambda^2, & \text{if } t > a\lambda, \end{cases}$$

for $a > 2$, and MCP of Zhang (2010)

$$J_\lambda(t) = \begin{cases} -t^2/(2a) + \lambda t, & \text{if } t \leq a\lambda, \\ a\lambda^2/2, & \text{if } t > a\lambda, \end{cases}$$

for $a > 1$. Similarly, the standardized nonconvex penalized estimators in this class are defined as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{p} J_{\lambda_j}(|\beta_j|) \right\}, \tag{3.1}$$

where $\lambda \geq 0$ is the regularization parameter and $\lambda_j = \lambda s_j$ in (2.4). Note that the `ncvreg` package provides the solution in (3.1). We describe the algorithm for the standardized SCAD estimator in the `ncvreg` package using the CD algorithm of Breheny and Huang (2011). Suppose that all predictors have been standardized such that $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = n$. Similar to the LASSO, the univariate objective function for the jth coordinate step in (1.2) is

$$q_\lambda(\beta_j | \tilde{\boldsymbol{\beta}}) = \frac{1}{2n} \sum_{i=1}^{n} (r_i - x_{ij}\beta_j)^2 + J_\lambda(|\beta_j|), \tag{3.2}$$

where $J_\lambda(\cdot)$ is the SCAD penalty function. The minimizer of $q_\lambda(\beta_j|\tilde{\boldsymbol{\beta}})$ in (3.2) has the closed form solution (Breheny and Huang, 2011).

$$\hat{\beta}_j = f(z_j, \lambda, a) = \begin{cases} \text{sign}(z_j)(|z_j| - \lambda)_+ & \text{if } |z_j| \leq 2\lambda, \\ \text{sign}(z_j) \left(\dfrac{a-1}{a-2} \right) \{|z_j| - a\lambda/(a-1)\}_+ & \text{if } |z_j| \in (2\lambda, a\lambda], \\ z_j & \text{if } |z_j| > a\lambda. \end{cases} \tag{3.3}$$

This closed form of solutions allows the implementation of the CD algorithm for SCAD with the assumption of standardization for all predictors. For MCP, see Breheny and Huang (2011).

Algorithm 3.1 The CD algorithm for the standardized SCAD estimator

Compute the standardized predictors $x_{ij}^* = x_{ij}/s_j$
Set an initial estimate $\tilde{\boldsymbol{\beta}}^* \in \mathbb{R}^p$
repeat
  for $j = 1, 2, \ldots, p$
    Compute the simple least-square fit z_j^* with the standardized predictors x_{ij}^*
    Update $\tilde{\beta}_j^*$ with $\hat{\beta}_j^* = f(z_j, \lambda, a)$ in (3.3)
until convergence
Transform the solutions by $\hat{\beta}_j = \hat{\beta}_j^*/s_j$

Algorithm 3.1 describes the implementation for the standardized SCAD estimator in the `ncvreg` package. Note that the CD algorithm is only applied to the model with standardized predictors in Algorithm 3.1. Breheny and Huang (2011) also established the convergence of the CD algorithm for SCAD and MCP under the assumption of standardization. However, the CD algorithms for SCAD and MCP without the assumption of standardization are not always guaranteed to converge in general. By results of Tseng (2001), one of the sufficient conditions for the convergence of the CD algorithm is that the univariate functions (3.2) in each coordinate step should satisfy the strictly convexity in $\mathbb{R}$. Although $q_\lambda(\beta_j)$ in (3.2) is not differentiable, it is directionally twice differential everywhere. Let $d_u^2 q_\lambda(\beta_j)$ denote the second derivative of $q_\lambda(\beta_j)$ in the direction u, and τ denote the infimum over β_j and u of the minimum eigenvalue of $d_u^2 q_\lambda(\beta_j)$. Then the strictly convexity of $q_\lambda(\beta_j)$ follows if τ is positive. Using some algebra, we obtain

$$\tau = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_{ij}^2 & \text{for LASSO,} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_{ij}^2 - \dfrac{1}{a-1} & \text{for SCAD,} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_{ij}^2 - \dfrac{1}{a} & \text{for MCP.} \end{cases}$$

These quantities for SCAD with $a > 2$ and MCP with $a > 1$ are positive under the standardization assumption, whereas the quantity for LASSO is always positive. However, the quantities for SCAD and MCP are not always guaranteed to be positive in general. Hence, the original nonconvex penalized estimator can not be obtained by directly applying the CD algorithm as in the Algorithm 2.1.

3.2. CCCP-SCAD algorithm

We adopt the idea of the CCCP-SCAD algorithm of Kim *et al.* (2008) for obtaining the original nonconvex penalized estimators in the class. The key idea is to convert the objective function to the LASSO problem via the concave convex procedure (CCCP) of Yuille and Rangarajan (2003), and then apply the Algorithm 2.1 for the original LASSO problem. The penalties in the class can be decomposed by the sum of concave and convex functions,

$$J_\lambda(|t|) = \tilde{J}_\lambda(|t|) + \lambda|t|,$$

where $J_\lambda(|t|) - \lambda|t|$ is always a differentiable concave and $|t|$ is a convex function. Therefore, the objective function can be rewritten as

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p}\tilde{J}_\lambda(|\beta_j|) + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{3.4}$$

Note that the objective function in (3.4) consists of the sum of concave and convex functions. Thus, we can apply the CCCP algorithm. Since $\tilde{J}_\lambda(|\beta_j|)$ is concave, for a given solution $\tilde{\beta}_j$ we have $J_\lambda(\beta_j) \le J_\lambda(\tilde{\beta}_j) + J'_\lambda(\tilde{\beta}_j)(\beta_j - \tilde{\beta}_j)$, where $J'_\lambda(\cdot)$ is the first derivative of $J_\lambda(\cdot)$. Hence, a tight convex upper bound of $Q_\lambda(\boldsymbol{\beta})$ for given current solutions $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ becomes

$$U_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p}\tilde{J}'_\lambda(|\tilde{\beta}_j|)\beta_j + \lambda\sum_{j=1}^{p}|\beta_j|,$$

which is the original LASSO problem with the quadratic loss function. Hence, we can use the CD algorithm to optimize $U_\lambda(\boldsymbol{\beta})$. It can be easily shown that the update rule for the jth coordinate solution is

$$\hat{\beta}_j = \operatorname{sign}(z'_j)(|z'_j| - \lambda)_+/c_j, \tag{3.5}$$

where $z'_j = z_j + \tilde{J}'_\lambda(|\tilde{\beta}_j|)$. Finally, we iterate these two steps until convergence. By the descent property of the CCCP algorithm, the objective function $Q_\lambda(\boldsymbol{\beta})$ always decreases after each iteration and hence the sequence of solutions converges to a local minimizer of $Q_\lambda(\boldsymbol{\beta})$ (Yuille and Rangarajan, 2003). The optimization algorithm to compute original nonconvex penalized estimates is summarized in Algorithm 3.2.

Algorithm 3.2 The CCCP-SCAD algorithm for the original SCAD estimator

Set an initial estimate $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$
Compute c_j and z_j in (2.3) for $j = 1, \ldots, p$
repeat
  Calculate $\tilde{J}'_\lambda(\beta_j)$ at $\tilde{\beta}_j$ for $j = 1, \ldots, p$
  repeat
    for $j = 1, 2, \ldots, p$
      Compute z'_j in (3.5).
      Update $\tilde{\beta}_j$ with $\hat{\beta}_j = \operatorname{sign}(z'_j)(|z'_j| - \lambda)_+/c_j$
  until convergence
  Update $\tilde{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}$
until convergence

4. Simulation studies

In this section, we compare the finite sample performances of the standardized penalized estimators (sLASSO and sSCAD) and the original penalized estimators (oLASSO and oSCAD) in terms of prediction accuracy and variable selectivity through various simulation

scenarios. For each method, we choose the optimal regularization parameter λ by external validation on an independent data set of size n, where n is the size of the training data set. The additional regularization parameter a for SCAD were set to be $a = 3.7$ in line with the recommendation suggested in Fan and Li (2001). We consider the linear regression model,

$$y = \mathbf{x}^T \boldsymbol{\beta}^* + \epsilon, \tag{4.1}$$

where $\epsilon \sim \mathrm{N}(0, 1)$. For various variance structures of predictors, we consider four examples presented below. For all examples, we set $n = 100$ and $p = 500$, with the first four nonzero coefficients set to be ± 1 and the remaining 496 coefficients equal to zero.

Table 4.1 Comparison of the original and standardized penalized estimators in various scenarios. The corresponding standard errors are in parentheses.

Example	Method	PE	ME	SIG	NOI	NUM
1	oLASSO	1.421 (0.018)	0.404 (0.017)	4.0 (0.000)	26.8 (1.186)	30.8 (1.186)
	sLASSO	1.417 (0.018)	0.400 (0.018)	4.0 (0.000)	25.5 (1.252)	29.5 (1.252)
	oSCAD	1.089 (0.011)	0.077 (0.010)	4.0 (0.010)	9.5 (0.705)	13.4 (0.705)
	sSCAD	1.088 (0.010)	0.077 (0.010)	4.0 (0.010)	9.4 (0.649)	13.4 (0.649)
2	oLASSO	2.069 (0.073)	1.078 (0.069)	3.2 (0.079)	36.4 (1.825)	39.6 (1.853)
	sLASSO	1.369 (0.024)	0.381 (0.016)	3.8 (0.047)	23.4 (1.084)	27.2 (1.095)
	oSCAD	1.520 (0.048)	0.532 (0.044)	3.3 (0.078)	40.2 (1.539)	43.5 (1.556)
	sSCAD	1.078 (0.016)	0.092 (0.008)	3.8 (0.052)	6.1 (0.850)	9.9 (0.854)
3	oLASSO	1.066 (0.015)	0.075 (0.005)	4.0 (0.000)	0.5 (0.103)	4.5 (0.103)
	sLASSO	1.352 (0.023)	0.364 (0.016)	4.0 (0.000)	25.4 (1.247)	29.4 (1.247)
	oSCAD	1.044 (0.014)	0.054 (0.003)	4.0 (0.000)	0.3 (0.087)	4.3 (0.087)
	sSCAD	1.047 (0.014)	0.056 (0.004)	4.0 (0.000)	2.9 (0.594)	6.9 (0.594)
4	oLASSO	5.079 (0.020)	4.087 (0.014)	0.0 (0.000)	4.4 (0.787)	4.4 (0.787)
	sLASSO	1.353 (0.023)	0.364 (0.016)	4.0 (0.000)	25.3 (1.296)	29.3 (1.296)
	oSCAD	5.079 (0.020)	4.087 (0.014)	0.0 (0.000)	4.5 (0.786)	4.5 (0.786)
	sSCAD	1.062 (0.015)	0.072 (0.004)	4.0 (0.000)	7.6 (0.683)	11.6 (0.683)

Example 4.1 (*homogeneous*) We first consider the scenario with homogeneous variances of predictors. The $\mathbf{x} = (x_1, \ldots, x_p)^T$ follows multivariate normal distribution with mean zero and covariance of x_j and x_k being $0.5^{|j-k|}$. Note that all variables have the unit variance, and this example was used in many papers (Tibshirani, 1996; Fan and Li, 2001).

Example 4.2 (*heteroscedastic*) We consider the case with heteroscedastic variances of predictors. Each variable x_j follows normal distribution with mean zero and variance s_j^2, where standard deviations s_j are randomly generated from uniform distribution over the interval $(0, 5)$.

Example 4.3 (*artificial*) In Example 4.3 and 4.4, we consider two artificial scenarios to provide an illustration of how each method works. Example 4.3 is the same as Example 4.2, except that the standard deviations of true signal variables are set to be 5, while ones of true noisy variables equal to 1.

Example 4.4 (*artificial*) Example 4.4 is the opposite scenario to Example 4.3. The standard deviations of true signal and true noisy variables are set to be 1 and 5, respectively.

We consider five measures. For prediction accuracy, we compute the prediction errors

(PE) and model errors (ME) based on independent test data set of size $N = 5,000$ that are calculated by

$$\text{PE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \quad \text{ME} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \mathbf{x}_i^T\boldsymbol{\beta}^*)^2,$$

respectively. For variable selectivity, we compute the number of the signal (SIG) and noisy (NOI) variables to be selected as well as the total selected variables (NUM). Table 4.1 presents the average values of each measure based on 100 independent replications.

When all predictors have the same variance, the original and standardized methods show very similar performances even if two methods produce the different estimates. As expected, the standardized penalized estimators outperform the original ones when the variances of predictors are substantially different in terms of both prediction accuracy and variable selectivity.

In an extreme artificial case, where variances of true signal variables are high, the original penalized estimators perform better than the standardized estimators. Furthermore, the oLASSO and oSCAD successfully delete the irrelevant variables as compared to other examples. On the other hand, when variances of true signal variables are substantially low comparing to ones of the noisy variables, the standardized penalized estimators outperform original ones. Moreover, the oLASSO and oSCAD always fail to select the true signal variables in 100 data sets, while the sLASSO and sSCAD always select all true signal variables. Based on results of Example 4.3 and 4.4, we can show that the original penalized estimators tend to select variables with high variance even if they are irrelevant in the underlying model. From the definition of the standardized estimators in (2.4) and (3.1), the standardized methods perform well by imposing greater amounts of penalty to variables with high variance in Example 4.4. Finally, it is interesting to notice that the sLASSO shows very similar results in all examples for both prediction accuracy and variable selectivity, whereas the sSCAD shows different results in the NOI measure, according to the variance structure of variables.

5. Read data analysis

We first analyze the well-known Boston Housing data set, which is available on UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/datasets/Housing`). This data set consists of 506 observations and 13 predictors, with median value of owner-occupied homes (MEDV) as a response variable. The predictors include a binary dummy variable (CHAS) and 12 continuous variables (CRIM, ZN, INDUS, NOX, RM, AGE, DIS, RAD, TAX, PTR, B, LSTAT). The detailed descriptions of each variable are presented in the web page from above URL.

We first compare the prediction accuracy and selectivity of the standardized penalized estimators (sLASSO and sSCAD) and original penalized estimators (oLASSO and oSCAD) as well as the ordinary least square estimator (OLS). The results are obtained by 100 random partitions of data set divided into two parts, training (2/3) and test (1/3) data sets. For each random partition, the optimal values of regularization parameter in each method are chosen by the 10-fold cross validation method with the training set only, and the prediction errors are calculated on the test set.

Table 5.1 Average of prediction errors and the numbers of the selected variables based on 100 random partitions

Measure	sLASSO	oLASSO	sSCAD	oSCAD	OLS
Prediction error	23.665	24.783	23.712	23.664	23.676
Standard error	(0.4250)	(0.4406)	(0.4227)	(0.4124)	(0.4209)
No. of variables	12.42	11.49	11.46	12.99	13.00
Standard error	(0.0741)	(0.0611)	(0.0958)	(0.0100)	(0.0000)

Table 5.2 The frequency of each variable being selected in 100 random partitions

Method	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTR	B	LSTAT
sLASSO	98	100	66	100	100	100	78	100	100	100	100	100	100
oLASSO	99	100	95	55	1	100	99	100	100	100	100	100	100
sSCAD	96	98	22	100	100	100	33	100	99	98	100	100	100
oSCAD	100	100	99	100	100	100	100	100	100	100	100	100	100
s_j	8.59	23.30	6.85	0.25	0.12	0.70	28.12	2.10	8.70	168.37	2.16	91.20	7.13

Table 5.1 shows the average values of prediction errors and the numbers of selected variables over 100 random partitions, and Table 5.2 presents the frequency of each variable to be selected among 100 random partitions. The oLASSO performs worse than the sLASSO and furthermore, the oLASSO tends to delete variables with low variance (CHAS and NOX) although other methods select them. The oSCAD outperforms the sSCAD in term of prediction accuracy but it fails to delete variables which are more likely to be irrelevant (INDUS and AGE).

Table 5.3 Estimated coefficients from each method based on the full data set, where the marker '–' indicates the zero value

Variable	Raw data				OLS	p-value	Transformed data			
	sLASSO	oLASSO	sSCAD	oSCAD			sLASSO	oLASSO	sSCAD	oSCAD
CRIM	-0.100	-0.098	-0.108	-0.110	-0.108	0.001	-0.100	-0.112	-0.108	-0.110
ZN	0.042	0.049	0.046	0.046	0.046	0.001	0.042	0.050	0.046	0.046
INDUS	–	-0.043	–	0.023	0.021	0.738	–	-0.055	–	0.023
CHAS	2.685	1.399	2.720	2.669	2.687	0.002	2.690	1.560	2.720	2.669
NOX	-16.414	–	-17.354	-19.385	-17.767	<.001	-16.488	–	-17.358	-19.385
RM	3.859	3.787	3.803	3.580	3.810	<.001	3.854	3.630	3.803	3.580
AGE	–	-0.012	–	0.002	0.001	0.958	–	-0.009	–	0.002
DIS	-1.406	-1.176	-1.493	-1.531	-1.476	<.001	-1.414	-1.176	-1.493	-1.531
RAD	0.259	0.269	0.298	0.321	0.306	<.001	0.261	0.248	0.299	0.321
TAX	-0.010	-0.014	-0.012	-0.013	-0.012	0.001	-0.010	-0.015	-0.012	-0.013
PTR	-0.932	-0.765	-0.946	-1.014	-0.953	<.001	-0.933	-0.736	-0.946	-1.014
B	0.009	0.010	0.009	0.009	0.009	0.001	9.070	–	9.288	8.703
LSTAT	-0.522	-0.561	-0.523	-0.537	-0.525	<.001	-0.523	-0.592	-0.523	-0.537

When applying regression models in practices, it is a standard behavior to change the unit of a variable. The variable B measures the black population proportion calculated by $1000(b-0.63)^2$, where b is the proportion of blacks by town. Hence, we only change the unit of the variable B by dividing 1,000. We apply all methods to raw data and transformed data with total observations. Table 5.3 presents the values of estimated coefficients and p-value from the OLS method. The standardized penalized estimators successfully delete irrelevant variables (INDUS and AGE) in terms of p-values. However, the original penalized estimators fail to delete the irrelevant variables where the corresponding variances are high. Moreover, the oLASSO even deletes the relevant variable B in transformed data analysis, although the oLASSO selects it in raw data analysis. Hence, the original penalized methods may

produce a wrong result in conversion units of measurement. These results suggest that the standardized methods are more useful than the original methods, especially when variances of predictors are substantially different.

Table 5.4 Average of prediction errors and the numbers of the selected variables based on 100 random partitions for TRIM data

Measure	sLASSO	oLASSO	sSCAD	oSCAD
Prediction error	0.4511	0.4675	0.5052	0.5228
Standard error	(0.0237)	(0.0267)	(0.0278)	(0.0302)
No. of variables	45.65	49.86	16.52	21.68
Standard error	(1.4623)	(1.6222)	(0.5540)	(0.9006)

Second, we analyze the gene expression data set used by Scheetz *et al.* (2006) to compare the prediction accuracy of the original and standardized estimators in a high-dimensional setting. This data set consists of the gene expression levels of 31,042 probe sets on 120 twelve-week-old male rats. The goal of the analysis is to detect genes whose expressions are related to that of gene TRIM32 which has been known to cause Bardet-Biedl syndrome (Chiang *et al.*, 2006). We first select 3,000 genes that display the largest variances in expression level and then select top 1,000 genes that have the largest absolute values of the marginal correlation with gene TRIM32. Thus, this data set for the analysis has $n = 120$ and $p = 1,000$. We apply the penalized linear regression using two types of LASSO and SCAD methods, with TRIM32 expression as the response variable and the selected top 1,000 genes as the predictive variables. Similar to the previous example, we perform 100 random partitions and present the average of prediction errors and the numbers of the selected genes in Table 5.4. The standardized estimators outperform the original estimators in terms of the prediction accuracy and furthermore, the standardized estimators identify more sparse models than original ones.

6. Concluding remarks

In this paper, we studied some issues on standardization of variables in penalized regressions, and provided the definition of standardized penalized estimator. We demonstrated that the CD algorithm can not be directly applied to the original nonconvex problems, and hence we introduced the CCCP-SCAD algorithm for the original SCAD estimator. The numerical results given in Section 4 and 5 showed that the standardized penalized estimators perform well regardless of variance structure of predictors, whereas the original penalized estimators may have poor performances in some cases. Hence, we suggest that researchers use a standardized penalized estimator, especially when variances of predictors are substantially different.

Although we only focused on linear regression models in this paper, we expect that this work can be conducted in various penalized regressions such as generalized linear models and Cox proportional hazard model. We leave this problem as a future work.

References

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, **5**, 232-253.

Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Sheffield, V. C. *et al.* (2006). Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proceedings of the National Academy of Sciences*, **103**, 6287-6292.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.

Fan, J. and Li, R. (2001). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, **30**, 74-99.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1-22.

Kim, Y., Choi, H. and Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, **103**, 1665-1673.

Kim, Y. and Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, **99**, 315-325.

Kwon, S., Han, S. and Lee, S. (2013). A small review and further studies on the lasso. *Journal of the Korean Data & Information Science Society*, **24**, 1077-1088.

Park, C. (2013). Simple principal component analysis using lasso. *Journal of the Korean Data & Information Science Society*, **24**, 533-541.

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dibona, G. F., Stone, E. M. *et al.* (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, **103**, 14429-14434.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267-288.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, **109**, 475-494.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *Annals of Statistics*, **36**, 614-645.

Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, **15**, 915-936.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, **38**, 894-942.