Afrika Statistika

Afrika Statistika
Vol. 13 (2), 2018, pages 1609–1630.
DOI: http://dx.doi.org/10.16929/1609.124

ISSN 2316-090X

Adjusting the Penalized Term for the Regularized Regression Models

Magda M. M. Haggag

[1]Department of Statistics, Mathematics, and Insurance, Faculty of Commerce, Damanhour University, Egypt

Received September 21, 2017; Accepted February 02, 2018

Abstract. More attention has been given to regularization methods in the last two decades as a result of exiting high-dimensional ill-posed data. This paper proposes a new method of introducing the penalized term in regularized regression. The proposed penalty is based on using the least squares estimator's variances of the regression parameters. The proposed method is applied to some penalized estimators like ridge, lasso, and elastic net, which are used to overcome both the multicollinearity problem and selecting variables. Good results are obtained using the average mean squared error criterion (AMSE) for simulated data, also real data are shown best results in the form of less average prediction errors (APE) of the resulting estimators.

Key words: Elastic-Net, Lasso, Penalized regression; Regularization; Ridge regression; Shrinkage; Variable selection.
AMS 2010 Mathematics Subject Classification : 62J05; 62J07.

*Corresponding author : magda.haggag@com.dmu.edu.eg, magmhag@yahoo.com

Résumé (French) Une attention accrue est de plus en plus accordée aux méthodes de régularisation au cours des deux dernières décennies à la suite de la la survenance de données de haute dimension. Cet article propose une nouvelle méthode d'introduction du terme pénalisé dans la régression régularisée. La pénalité proposée est basée sur l'utilisation des variances des estimateurs des moindres carrés des paramètres de régression. La méthode proposée est appliquée à certains estimateurs pénalisés comme la méthode ridge, le lasso et le filet élastique, qui sont utilisés pour surmonter à la fois le problème de la multicolinéarité et la sélection des variables. De bons résultats obtenus en utilisant la le critère de l'erreur quadratique moyenne (AMSE) pour les données simulées et également sur des données réelles sont présentés. Les meilleurs résultats sont obtenus avec le critère des erreurs de prédiction moyennes (APE) sur les estimateurs concernés.

1. Introduction

Regularization methods have given more attention in the last two decades as a result of exiting high-dimensional ill-posed data. More effort has gone to develop the regression methods for simultaneous variable selection and coefficient estimation. A smaller subset from large number of predictors is desired to obtain more important predictors. Also, a large number of predictors may cause highly collinear regressors which lead to severe estimation problems. The ordinary least squares estimator abbreviated *OLS*, will have undesired properties. Large variance, incorrect signs, unstable, and long length of the OLS are examples of the undesired properties. Regularized regressions are developed to overcome these problems as alternative methods to OLS, such as ridge regression abbreviated *RR* (Hoerl and Kennard (1970a,b)), bridge regression abbreviated *BR* (Frank,and Friedman (1993)), Least absolute shrinkage, and selection operator, Lasso, (Tibshirani (1996)), least-angle regression, LARS (Efron et al. (2004), and the Elastic-Net, abbreviated *EN* (Zou and Hastie (2005)). In combating the collinearity, *RR* improves the variability of the regression estimators of *OLS* through shrinkage, but cannot produce a model with relevant predictors. In both shrinkage estimates and selecting variables Lasso is used, and the Elastic- Net is proposed as an improved version of Lasso. (Zou and Hastie (2005)).

Consider the multiple linear regression model of the form:

$$Y = X\beta + \varepsilon, \tag{1}$$

where Y is an $(n \times 1)$-column vector of dependent variable, X is an (nxp) matrix of regressors, β is a $(p \times 1)$-vector of unknown parameters to be estimated, and ε is an $(n \times 1)$-vector of errors distributed as $\mathcal{N}\left(0, \sigma^2 I\right)$.

The *OLS* estimator vector $\hat{\beta}_{OLS} = \left(\hat{\beta}_1, ..., \hat{\beta}_p\right)$ is obtained by minimizing the residual sum of squares criterion (RSS) as follows:

$$RSS = \|y - x\beta\|_2^2, \tag{2}$$

where $\|\cdot\|^2$ denotes the square of the standard Euclidean norm, and thus $\hat{\beta}_{OLS}$ can be obtained as:

$$\hat{\beta}_{OLS} = (x'x)^{-1} x'y. \tag{3}$$

In spite that the OLS estimator is an unbiased estimator, it may still have a large variance when there exist multicollinearity among regressors in the design matrix X . This causes unstable solutions. The instability when minimizing the RSS leads to different regularization penalty techniques. These techniques are based on adding some penalty term to (2) to obtain a regularized form of RSS. The most known regularization s are $L2$, $L1$, and hybrid of both $L2$ and $L1$, respectively. (Wenjiang (1989); Fu (1998); and Vidaurre et al. (2013)).

In this paper, a new penalized form is proposed in (2) for some regularization methods in linear models such as ridge regression, the Lasso, and the Elastic-Net. However, the properties of the new regularization method are studied in this paper; best results are obtained in the form of less mean squared errors and less average prediction errors for both simulated and real data, respectively. Section 2 presents some regularized regression methods used in both combating collinearity and selecting variables. Section 3 introduces the proposed new penalty term in some regularized regression methods. The Mont Carlo simulation study, real-life data, and the performance criteria of the new proposal are presented in section 4. Conclusions and recommendations are considered in Section 5.

2. Regularized Regression Methods for Linear Models

To achieve both accuracy and parsimony of statistical modeling part or all of the regression coefficients are penalized. Ridge regression, the Lasso, and the Elastic-Net will be considered in this work as examples of achieving model prediction accuracy, interpretability, or both in linear regression models.

2.1. Regularization Using Ridge Penalty (L2-Penalty)

To accept some bias in order to reduce the variance of OLS estimators, the penalized estimators are used. Ridge regression, *RR*, (Hoerl and Kennard (1970a,b)), shrinks the *OLS* estimators using the L2 norm of the coefficients. The loss function of RR method can be shown as:

$$\hat{\beta}_{RR} = \underset{\beta}{\operatorname{argmin}} \|y - x\beta\|_2^2 + \lambda_2 \|\beta\|_2^2, \tag{4}$$

where

$$\|y - x\beta\|_2^2 = \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

is the L2-norm quadratic loss function, $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the L2-norm penalty on β andλ_2 is the ridge penalty parameter which controls the amount of shrinkage of the coefficients towards zero. The larger is the value of λ_2, the greater is the amount of shrinkage. Equation (4) can be written as a restricted version of (2) as follows:

$$\hat{\beta}_{RR} = \underset{\beta}{\operatorname{argmin}} \|y - x\beta\|_2^2 \tag{5}$$

subject to: $\|\beta\|_2 \leq t_2$, where $t2$ is a tuning parameter. Equation (4) gives the constrained coefficient of RR estimator as:

$$\hat{\beta}_{RR} = (x'x + \lambda_2 I)^{-1} x'y, . \tag{6}$$

The *RR* estimator shrinks the coefficient of correlated predictors equally towards zero, but cannot select the most relevant subset of predictors.

2.2. Regularization Using Lasso Penalty (L1-Penalty)

Tibshirani (1996) introduced Lasso as a penalized method using the $L1$ penalty rather than $L2$ penalty in *RR* as follows:

$$\hat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}} \|y - x\beta\|_2^2 + \lambda_1 \|\beta\|_1 , \tag{7}$$

where

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

is is the L1-norm penalty on β, and $\lambda_1 \geq 0$ is the Lasso penalty parameter which penalizes the sum of the absolute values of the regression coefficients. λ_1 controls the strength of the penalty and work like ridge parameter, that is OLS estimator is obtained when $\lambda_1 = 0$ and zero estimator is obtained when $\lambda_1 = \infty$.

When λ_1 increases, more coefficients are set to zero (less variables are selected) and more shrinkage is employed. Equation (7) can be written in a constrained form as:

$$\hat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}} \|y - x\beta\|_2^2 , \tag{8}$$

subject to:

$$\|\beta\|_1 \leq t_1$$

where t_1 is a tuning parameter. Minimization of (7) is more complicated; there is no closed form of lasso estimator as ridge estimator (Tibshirani (1996)). Van Der Kooij (2007) showed that the Lasso estimator can be estimated by minimizing (7) as follows:

$$\hat{\beta}_{Lasso} = (x'x)^{-1}\left(x'y - \frac{1}{2}\lambda_1 I\right). \tag{9}$$

The lasso estimator reduces the variability of the estimates by shrinking some coefficients, and produces interpretable models by shrinking some other coefficients to exactly zero. These properties make the Lasso a highly popular method in variable selection. Thus the Lasso combines the two features of RR and subset selection. However, Lasso has the following shortcomings:

(a) It is unstable with high dimensional data.

(b) It cannot select more variable than the sample size when $p > n$.

These shortcomings mean that Lasso outperforms *RR* and subset selection with a small to moderate number of moderated correlations, subset selection outperforms Lasso for a small number of large correlations, and *RR* is the best estimator in the case of a large number of small correlations. (Tibshirani (1996);

2.3. Regularization Using Elastic-Net Penalty

To circumvent the instability of the Lasso estimator when predictors are highly correlated, the Elastic-Net was proposed. Zou and Hastie (2005) proposed the Elastic-Net as a regularized estimator by combining the L1 (Lasso) and L2 (RR) penalties which can be obtained by the following optimization problem:

$$\hat{\beta}_{Elastic} = \underset{\beta}{\operatorname{argmin}} \|y - x\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \tag{10}$$

The optimization problem in (10) can be presented as a penalized least squares method as follows:

$$\hat{\beta}_{Elastic} = \underset{\beta}{\operatorname{argmin}} \|y - x\beta\|_2^2$$

Subject to:

$$\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_1^2 \leq t, \tag{11}$$

where the function $\alpha \|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_1^2$ is the Elastic-net penalty. When $\alpha = 0$, the elastic-net becomes simple ridge regression, and when α=1, Lasso estimator is obtained. The Elastic-Net penalty is a convex combination of both lasso and ridge penalty. The penalty is convex when α=0, and strictly convex when $\alpha > 0$. The used values of α in Elastic-Net is $\alpha \in [0, 1]$. This form of penalty is useful when there are many correlated predictors in the model. (Friedman et al. (2010)). Minimization of

Equation (10) gives the Elastic-Net estimator. Van Der Kooij (2007) showed that the Elastic-Net explicitly estimator can be estimated as:

$$\hat{\beta}_{Elastic-Net} = (x'x + \lambda_2 I)^{-1} \left(x'y - \frac{1}{2}\lambda_1 I \right). \tag{12}$$

The Elastic-Net estimator selects variables like Lasso, and shrinking the coefficients of correlated predictors like RR. It is found that this estimator is an extension of the Lasso and is robust to extreme correlations among the predictors. (Friedman et al. (2010)). The Elastic-Net estimator unlike the Lasso when $p > n$, the The Elastic-Net estimator can select more than n variables.(Zou and Hastie (2005)).

Theorem 1. *Given,*$\mathrm{q} \geq 1, \lambda > 0, r = 1, 2,$ *and*

$$Q = \min_{\beta} \left[RSS + J(\beta) \right],$$

where,

$$J(\beta) = \lambda . \|\beta\|_r^q,$$

which implies that :

$$\hat{\beta} = \underset{\beta}{\textit{argmin}} \; Q(\beta, x, y, \lambda, q),$$

$$d\left[J(\beta)\right] = d(\beta_j, \lambda, q) = q.\lambda . \|\beta\|_2^{q-1} \, sign(\beta_j), \tag{13}$$

and let:

$$\frac{\partial Q}{\partial \beta_j} = C_j(\beta, x, y) + d(\beta_j, \lambda, q) = 0.$$

Then

$$\begin{aligned} &C_1(\beta, x, y) + d(\beta_1, \lambda, q) = 0, \\ &C_2(\beta, x, y) + d(\beta_2, \lambda, q) = 0, \\ &\vdots \\ &C_p(\beta, x, y) + d(\beta_p, \lambda, q) = 0. \end{aligned} \tag{14}$$

If the function C is continuously differentiable with respect to β, and the Jacobian

$$\frac{\partial C}{\partial \beta}$$

is positive semi-definite (p.s.d.), then :

(1) The equations in (14) have a unique solution $\hat{\beta}(\lambda, q)$ which is continuous in (λ, q).

(2) The limit of $\hat{\beta}(\lambda, q)$ *exists as follows:*

(2a) $\lim_{q \to 1} \hat{\beta}(\lambda, q) = \hat{\beta}_{Lasso}$

(2b) $\lim_{q \to 2} \hat{\beta}(\lambda, q) = \hat{\beta}_{RR}$

(2c) $\lim_{q \to 1} \hat{\beta}(\lambda, q) + \lim_{q \to 2} \hat{\beta}(\lambda, q) = \hat{\beta}_{EN}$.

The proof of Theorem (1) is in Appendix (A-1).

2.4. Measuring the effective number of parameters

The degrees of freedom of an estimator describes it effective number of parameters. Generally, the effective number of parameters is measured using the degrees of freedom of the estimator. One of the usage of the degrees of freedom is to put two different estimators on an equal footing. The estimators can be compared with different tuning parameters using degrees of freedom of the estimator .Precisely, given the data $y \in \mathbb{R}^n$ from the model (1) and suppose that y is estimated by $\hat{y}$. The degrees of freedom (df) of the estimate $\hat{y}$ is defined as:

$$df\left(\hat{y}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\operatorname{cov}\left(\hat{y}_i, y_i\right) \tag{15}$$

The higher the correlation between the i^{th} fitted value and the i^{th} data point, the higher its degrees of freedom. The degrees of freedom can be measured for the fitted model using the previously three regularized estimators as indicated in the following Lemma. (See Meyer, and Woodroofe (2000); Zou et al. (2007); and Kato (2009)).

Lemma 1. *(Stein's Lemma, Stein (1981)) Suppose that $\hat{y}_i : R^n \to R$ is absolutely continuous in ith coordinate for $i = 1, 2, \cdots, n$. If $E\left(\left|\frac{\partial \hat{y}_i}{\partial y_i}\right|\right) \prec \infty$ for each i, then:*

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\operatorname{cov}\left(\hat{y}_i, y_i\right) = E\left(\operatorname{div}\hat{y}_i\right)$$

,

where $\operatorname{div}\hat{y}_i = \sum_{i=1}^{n}\frac{\partial \hat{y}_i}{\partial y_i}$.

Therefore, an unbiased estimator of the degrees of freedom is given by: $d\hat{f}\left(\hat{y}\right) = \operatorname{div}\hat{y}$.

The degrees of freedom of a fitting procedure describe the effective number of parameters used by this procedure as follows:

(1) ***For the linear regression***, $\hat{y} = x\hat{\beta}_{OLS}$, $df\left(\hat{y}\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\operatorname{cov}\left(\hat{y}_i, y_i\right) = p$.

(2) ***For the ridge regression***, $\hat{y} = x\hat{\beta}_{RR}$, $df\left(\hat{y}\right) = \sum_{j=1}^{p}\frac{\gamma_j}{\gamma_j+\lambda_2}$

(3) ***For the Lasso***, $\hat{y} = x\hat{\beta}_{Lasso}$, $df\left(\hat{y}\right) = q = E\left[\text{number of nonzero coefficients in}\hat{\beta}_{Lasso}\right]$.

(4) ***For the Elastic-net***,$\hat{y} = x\hat{\beta}_{Elastic-Net}$,$df\left(\hat{y}\right) = \sum_{j=1}^{q}\frac{\gamma_j}{\gamma_j+\lambda_2}$,

where γ_j is the j^{th} eigenvalue of the matrix ($x'x$), λ_2 is the ridge parameter (penalized parameter), and q is the number of nonzero coefficients.

The proof of Lemma (1) is in Appendix (A-2).

3. The Proposed Regularized Regression Estimator

Large efforts have done to develop the penalized regression methods to get higher prediction accuracy in linear regression models. In this paper new estimators of RR, Lasso, and elastic net are proposed, which are based on penalizing each coefficient with different penalty factor in the penalty term in of the loss function. The proposed estimators for RR, Lasso, and elastic net are obtained by minimizing the loss functions, as follows:

$$\hat{\beta}_{New} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} \omega_j \left[\frac{1}{2}\left(1-\alpha\right)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}\left|\beta_j\right|\right] \tag{16}$$

where ω_j is the jth penalty factor of the jth coefficient, the default is one for each coefficient. Equation(16) can be written in matrix form as follows:

$$\hat{\beta}_{New} = \underset{\beta}{\operatorname{argmin}} \left\|y - x\beta\right\|_2^2 + \lambda W \left[\frac{1}{2}\left(1-\alpha\right)\left\|\beta\right\|_2^2 + \alpha\left\|\beta\right\|_1\right] \tag{17}$$

where, λ is the penalized parameter, $w = diag(\omega_j,\ j = 1, 2, \cdots)$, and

$$\omega_j = \frac{var(\hat{\beta}_j)}{\sum_{j=1}^{p} var(\hat{\beta}_j)},$$

where $\hat{\beta}_j$ is jth OLS estimator of the parameter vector β, $\sum\limits_{j=1}^{p} \omega_j = trace\,(W) = 1$. The larger values of ω_j's corresponds to more penalty on the coefficients. Lasso estimator is obtained when $\alpha = 1$, *RR* estimator is obtained when $\alpha = 0$, and Elastic-Net estimator is obtained when $\alpha \in [0, 1]$.

Different contributions have been proposed to select the penalized parameter, λ. In this work, a 10-fold cross validation method is used to estimate the parameter Λ. (See Stone (1974), Picard and Cook (1984), Yi and Yang (2013)). Practically, It is found that the new proposed penalized estimators, in (16), or in (17), enjoy good properties, which lead to near optimal estimators.

4. The Numerical Studies

A simulation study is conducted to evaluate the new penalty for the three estimation methods: RR, Lasso, and Elastic-net. The average mean square error (AMSE) criterion is used to compare between these methods, with the OLS estimator, and with the old versions of these methods. Also, data examples are used to illustrate the effect of the new penalty on the performance of the estimators using multicollinear real data. The results are obtained using the **glmnet package** in R. (Friedman et al. (2010)).

4.1. A Simulation Study

The model used to generate data in this study is based on the true linear regression model in (1). Three scenarios of simulated data are used, and in each one the data are split into two sets, train $(0.66 * n)$ and test $(0.34 * n)$ sets. The design matrix X in all examples is generated from a multivariate normal distribution with mean equal zero and variance equal one. The pairwise correlation between any two predictors(xi,xj) =$0.5^{|i-j|}$ is used in each scenario. The computed results are based on 200 replications. The penalizing parameter λ is estimated using 10- fold cross validation method.

4.1.1. Scenario 1:

Contains 20 observations and 5 predictors with $\beta = (1, 1, 1, 0, 0)$. Table 1 and Figure 1 present the performance of the three estimators RR, Lasso, and Elastic-Net before and after applying the new penalty factor on each coefficient. Best results are obtained in the form of less AMSE after applying the new penalty factor. The Elastic-Net gave less AMSE before and after when $\alpha = 0.1$.

Fig. 1. The AMSE before and after applying the new Penalty factor with values of $\alpha \in [0, 1]$.(Scenario 1)

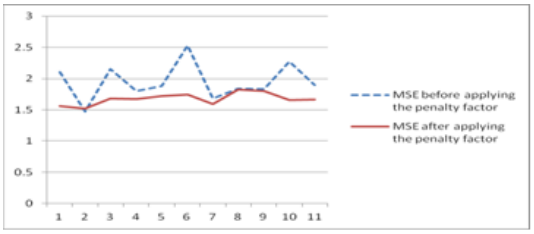


Table 1. MSE for different values of α before and after applying the penalty factor ω_j using generated data containing 20 observations and 5 predictors with $\beta = (1, 1, 1, 0, 0)$, repeated 200 times. (Scenario 1)

alpha	MSE before applying the penalty factor	MSE after applying the penalty factor	Estimator
0	2.107787	1.561581	Ridge
0.1	1.474640	1.517871	Elastic-Net
0.2	2.151780	1.679560	Elastic-Net
0.3	1.801405	1.672152	Elastic-Net
0.4	1.876715	1.723631	Elastic-Net
0.5	2.529010	1.742830	Elastic-Net
0.6	1.681621	1.595544	Elastic-Net
0.7	1.840667	1.827098	Elastic-Net
0.8	1.828950	1.802149	Elastic-Net
0.9	2.276149	1.659560	Elastic-Net
1.0	1.894034	1.664934	Lasso

Table 2. MSE for different values of $alpha$ before and after applying the penalty factor ω_j using generated data containing 50 observations and 10 predictors with $\beta = (1, ..., 1, 0, ..., 0)$, i.e, components 1:5 of β are ones, components 6:20 of β are zeros, repeated 200 times. (Scenario 2)

alpha	MSE before applying the penalty factor	MSE after applying the penalty factor	Estimator
0	1.825720	1.695605	Ridge
0.1	1.701138	1.637425	Elastic-Net
0.2	1.761129	1.577093	Elastic-Net
0.3	1.837497	1.539042	Elastic-Net
0.4	1.464788	1.514519	Elastic-Net
0.5	1.592516	1.556252	Elastic-Net
0.6	1.711804	1.441806	Elastic-Net
0.7	1.497739	1.580121	Elastic-Net
0.8	1.542156	1.509018	Elastic-Net
0.9	1.598749	1.518723	Elastic-Net
1.0	1.668408	1.571212	Lasso

4.1.2. Scenario 2:

Contains 50 observations and 10 predictors with β= (1,...1, 0,..., 0), i.e, components 1:5 of $beta$ are ones, components 6:10 of $beta$ are zeros. Table 2 and Figure 2 present the performance of the three estimators RR, Lasso, and Elastic-Net before and after applying the new penalty factor on each coefficient. Best results are obtained in the form of less AMSE after applying the new penalty factor. The Elastic-Net gave less AMSE before when $\alpha = 0.7$ and after when $\alpha = 0.6$.

4.1.3. Scenario 3:

Contains 100 observations and 20 predictors with β= (1,...1, 0,..., 0), i.e, components 1:5 of $beta$ are ones, components 6:20 of $beta$ are zeros. Table 3 and Figure

Fig. 2. The AMSE before and after applying the new Penalty factor with values of $\alpha \in [0, 1]$.(Scenario 2)

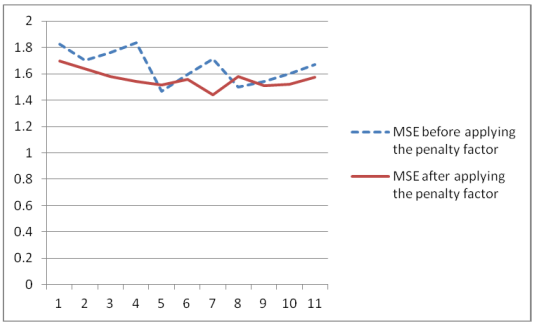


Table 3. MSE for different values of α before and after applying the penalty factor ω_j using generated data containing 100 observations and 10 predictors $\beta = (1, \cdots, 1, 0, \cdots, 0)$, i.e, components $1 : 5$ of β are ones, components $6 : 20$ of are zeros, repeated 200 times. (Scenario 3)

alpha	MSE before applying the penalty factor	MSE after applying the penalty factor	Estimator
0	2.330777	1.721222	Ridge
0.1	2.014475	1.622198	Elastic-Net
0.2	1.777090	1.479394	Elastic-Net
0.3	1.829031	1.538008	Elastic-Net
0.4	1.784186	1.515190	Elastic-Net
0.5	1.890124	1.587232	Elastic-Net
0.6	1.814542	1.541322	Elastic-Net
0.7	1.864542	1.575007	Elastic-Net
0.8	1.820104	1.548710	Elastic-Net
0.9	1.894814	1.597167	Elastic-Net
1.0	1.759220	1.513971	Lasso

3 present the performance of the three estimators RR, Lasso, and Elastic-Net before and after applying the new penalty factor on each coefficient. Best results are obtained in the form of less AMSE after applying the new penalty factor. The Lasso and Elastic-Net gave less AMSE before when $\alpha = 1.0$ (Lasso)and after when $\alpha = 0.2$ (Elastic-Net). Figure 4 shows the cross-validation curve (the dotted line), and upper and lower standard deviation curves along the $lambda$ sequence (error bars). The two dotted vertical lines indicate two selected $\lambda's$ in which give minimum mean cross-validated error. It is clear that the cross-validated errors (error bars) are less after applying the new penalty factor compared with before applying this factor. The value of λ which gives minimum mean cross-validated error is $\lambda = 0.1021699, 0.2415797, and 0.1696465$ for Lasso, RR, and Elastic-Net, respectively. Table 3 : MSE for different values of $alpha$ before and after applying the penalty factor ω_j using generated data containing 100 observations and 10 predictors with $\beta = (1, ..., 1, 0, ..., 0)$, i.e, components 1:5 of β are ones, components 6:20 of β are zeros, repeated 200 times. (Scenario 3)

Fig. 3. The AMSE before and after applying the new Penalty factor with the cross-validation curve and upper and lower standard deviation curves along the *lambd* sequence. (Scenario 3)

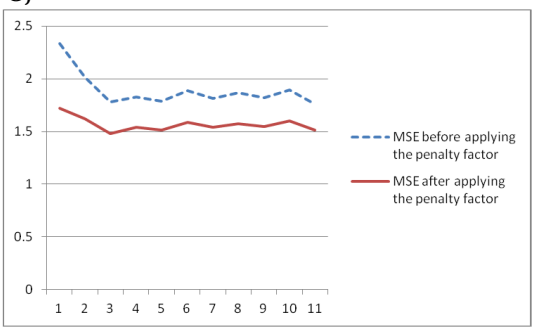
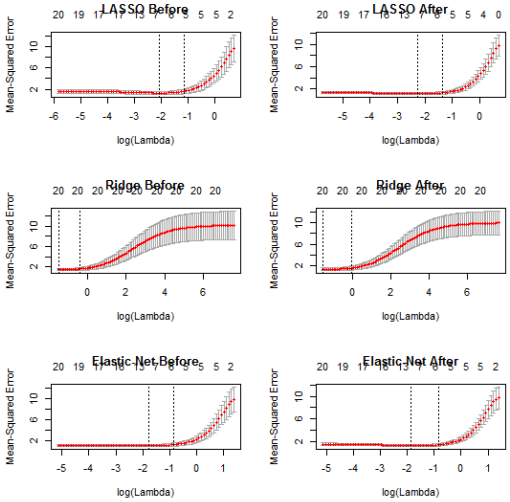


Fig. 4. The AMSE before and after applying the new Penalty factor with the cross-validation curve and upper and lower standard deviation curves along the *lambda* sequence. (Scenario 3)



4.2. Numerical Example (Prostate data)

.

For the prostate data, the interest is in the level of prostate-specific antigen (lpsa) in men who have prostate cancer. The sample size n=97 men with prostate cancer, and predictors p=8 clinical measures defined as follows:
Age: in years
Gleason:a numeric vector
Lbph: log of the amount of bengin prostatic hyperplasia
Lcavol: log cancer volume

Lcp: log of capsular penetration
Lweight: log prostate weight
Pgg45: percent of gleason score 4 or 5
Svi: seminal vesicle invasion
Data source: Stamey et al., 1989.

The three regularized regression methods, RR, Lasso, and Elastic-Net are applied to these data before and after applying the new penalty factor. The prostate data are divided into two parts, the first part is a training set contains 65 observations and the second part is a test set contains 32 observations. The three estimation methods are compared using the average prediction error (APE) applied on the test data.

Figure 5 presents the correlations between the predictors and the response. The correlation coefficient is between 0 and 0.8.

Fig. 5. Correlations of Prostate Data

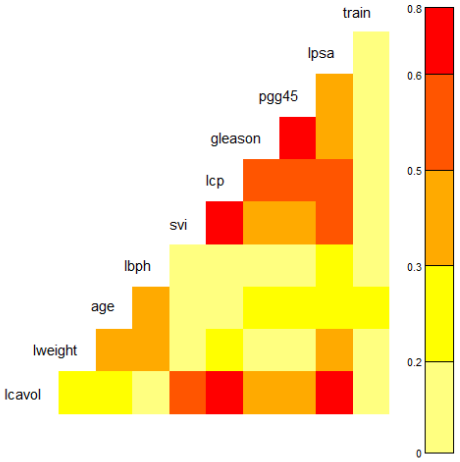


Table 4 and Figure 6 shows the average prediction error (APE) of the prostate data before and after applying the new penalty factor. Best results are obtained in the form of less values of APE after applying the new penalty factor. Elastic-Net (RR) gave less APE compared to Lasso and Ridge.

Figure 7 presents the prediction error when using subset regression, the graph indicates that the best subset contains 7 predictors with less PE. Figure 8 shows the ridge coefficient paths for all the 8 predictors, since RR shrink the estimators but cannot select variables. Figure 9 and Figure 10 present the Lasso and Elastic-Net coefficient paths for the selected 7 predictors, respectively.

Table 4. Average Prediction Errors (APE) of the Prostate test data before and after applying the new penalty factor (**BAPF** means **before applying the penalty**, and **AAPF** stands for *after applying the penalty factor*)

Penalized Method	APE *BAPF*	APE *AAPF*	Degrees of Freedom(D.f.)
Ridge	6.012190	5.759386	All the predictors
Lasso	5.922336	5.718610	7 (all except gleason)
Elastic-Net	5.920947	5.717075	7 (all except gleason)

Fig. 6. Average prediction error (APE) for Prostate data before and after applying the new penalty factor

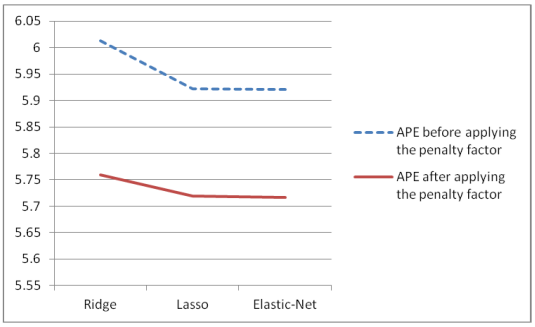
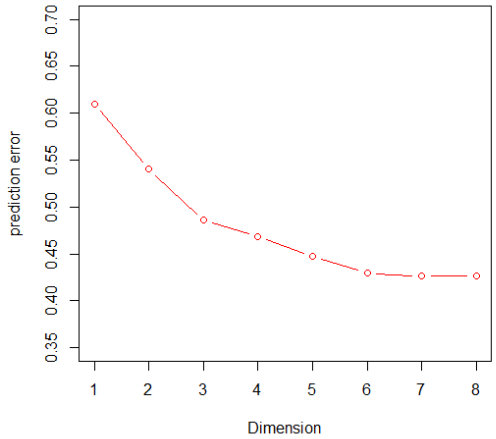


Fig. 7. Prediction Error of the Best Subset Model for the Prostate Data



5. Conclusion

This paper proposes a new penalty term for the regularized regression methods used in shrinkage and variable selection. A Regularization method is based on adding a penalized term to the residual sum of squares to improve the OLS esti-

Fig. 8. Ridge Coefficients Paths for the Prostate Data (as a function of $lambda$)

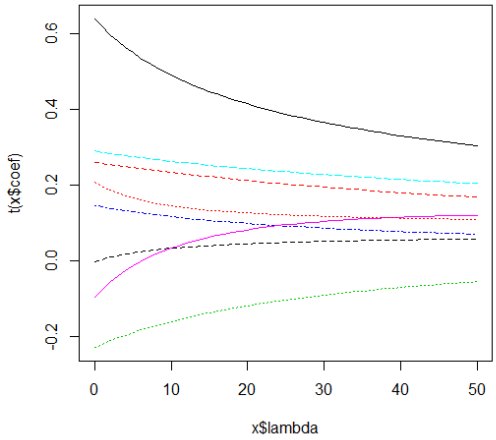
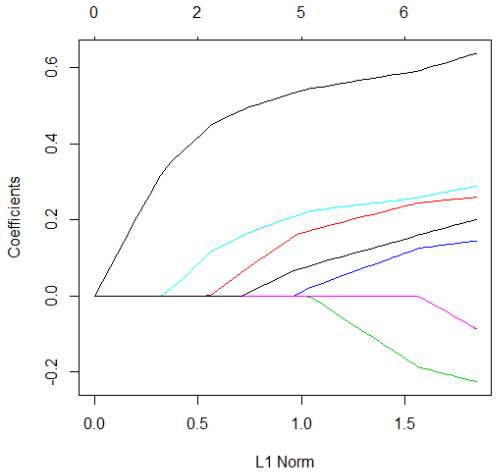
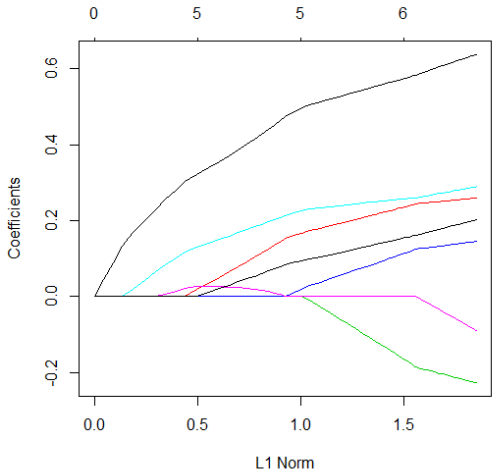


Fig. 9. Coefficient Paths when $\alpha = 1$ (Lasso Regression)



mator. Two penalties are introduced in the literature L2 for ridge regression (shrinkage), and L1 for Lasso regression (shrinkage and selection of variables). Hybrid of both L2 and L1 is presented through the Elastic-Net . The proposed penalty is based on adding a new factor to the penalized term. This factor uses the ratio of the variance of each OLS estimator to the total variances, such that large penalty is devoted to the estimator with large variance and vice versa. Simulations and data examples are used in this work to evaluate the proposed procedure. Best re-

Fig. 10. Coefficient Paths when $\alpha = 0.5$ (Elastic Regression)



sults are obtained in the form of less AMSE and APE of the proposed penalty for RR, Lasso, and Elastic-Net. The results support the Elastic-Net penalty, and more work is needed to improve this penalty and using it in statistical modeling.

References

Efron, B., Hastie, T., Johnstone, I., and Tibshuirani, R.(2004) Least Angle Regression. *Ann. Statist.*, 32, pp 407–499.

Frank, I.E. and Friedman, J.H.(1993) A Statistical View of some chemometrics regression tools. *Technometrics*, 35, pp. 109–148.

Friedman J.; Hastie T.; and Tibshirani R.(2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, Vol. 33, pp. 1–20.

Fu, W. J.(1998) Penalized Regression: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, Vol. 7, pp. 397–416.

. Hoerl, A. E. and Kennard, R. W.(1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. Vol. 12, pp. 55-67.

. Hoerl, A. E. and Kennard, R. W.(1970). Ridge Regression : Applications to Nonorthogonal Problems. *Technometrics*. Vol. 12, pp. 69-82.

Huang, J., Ma, S., Xie, H. and Zhang. C. H.(2009). A Group Bridge Approach for variable selection. *Biometrika*, Vol. 96, pp. 339-355.

Kato K.(2009). On the Degrees of Freedom in Shrinkage Estimation. *Journal of Multivariate Analysis*, Vol. 100, pp. 1338-1352.

Meyer, M., and Woodroofe M.(2000). On the degrees of freedom in shape-restricted regression. *The Annals of Statistics*, Vol. 28, pp. 1083-1104.

Picard R.R., and Cook R.D.(1984). Cross-validation of regression models. *Journal of the American Statistical Association*, Vol. 88, pp. 486-494.

Stamey T., Kabalin J., McNeal J., Johnston I., Freiha E., and Yang N.(1989). Prostate Specific Antigen in the Diagnosis and treatment of adenocarcinoma of the prostate ii. Radical prostatectomy treated patients. *Journal of Urology*, Vol. 16, pp. 1076-1083.

Stein M. C.(1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, Vol. 2, pp. 1135-1151.

Stone M. (1974). Cross Validation and Multinomial Prediction. *Biometrika*, Vol.61, pp.509-515.

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

Van Der Kooij. Anita J.(2007). *Prediction accuracy and stability of regression with Optimal Scaling transformations*. Leiden University.

..

Vidaurre D., Bielza C., and Larranage P.(2013) Classification of neural signals from space autoregressive features. *Neurocomputing*, Vol. 111, 21-26.

Wenjiang J. F.(1989) Penalized Regression: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, Vol. 7, pp. 397-410.

Yi Y., and Yang F.(2013) Modified cross-validation for penalized high dimensional linear regression models. *Journal of Computational and Graphical Statistics*, Vol. 23, pp. 1009-1027

Zou H., and Hastie T.(2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society ,Series B*, Vol. 67, pp. 301-320.

Zou J., Foster D.P., Stine A., and Ungar L.H.(2006) Streamwise feature selection. *Journal of Machine Learning Reasearch*, Vol. 7, pp. 1861-1885.

Zou J., T. Hastie, and R. Tibshirani(2007) On the degrees of freedom of the Lasso. *The Annals of Statistics*, Vol. 36, pp. 2173-2192

Appendix (A).

(A-1) Proof of Theorem (1). If the Jacobian $\frac{\partial RSS}{\partial \beta}$ is p.s.d, then $S_j(\beta, x, y) + d(\beta_j, \lambda, q)$ is also p.s.d where $\beta_j \neq 0 \forall \mathrm{j} = 1, 2, ..., \mathrm{p}$.

By mathematical induction, it can be proved that there exists a unique solution of equations (14). Therefore, there exists a unique solution $\hat{\beta}(\lambda, q)$ of equations (2.11) and $\hat{\beta}(\lambda, q)$ is continuous in (λ, q).

Proof of the second part: The existence of the limit of $\hat{\beta}(\lambda, q)$ when $q = 1, 2$ can be proved also by mathematical induction as follows :

(a) $p = 1$: If there is an intersection of the continuous functions $\mathrm{C}(\beta, x, y)$ and $\mathrm{d}(\beta, \lambda, 1)$, then by continuity of these functions, the limit of $\hat{\beta}(\lambda, q)\, as \mathrm{q} \to 1$ exist, and equal to the coordinate of the intersection. But, if there is no intersection between the two functions $\mathrm{C}(\beta, x, y)$ and $\mathrm{d}(\beta, \lambda, 1)$, then the limit of $\hat{\beta}(\lambda, q)\, as \mathrm{q} \to 1$ will equal to zero. These results can also hold for $q = 2$ and $q = \{1, 2\}$. Therefore, the results are true when p=1.

(b) For $p \succ 1$: Assume that the result is true for all dimensions $1, 2, \cdots, p-1$, then it can be proved that is also hold for dimension p as follows: The limit of the unique solution $\left[\hat{\beta}_1(\beta_p, \lambda, q), ..., \hat{\beta}_{p-1}(\beta_p, \lambda, q)\right]$ exists when $\mathrm{q} \to 1, \mathrm{q} \to 2, and \mathrm{q} \to \{1, 2\}$ for any fixed β_p. By plugging this solution into the last equation of equations (14), the following result will be obtained:

$$C_p\left[\hat{\beta}_1(\beta_p, \lambda, q), ..., \hat{\beta}_{p-1}(\beta_p, \lambda, q), \beta_p, x, y\right] + d(\beta_p, \lambda, q) = 0 \quad \text{(A-1)}$$

Equation (A.1) can be proved that it has a unique solution and its limit exists when $\mathrm{q} \to 1, \mathrm{q} \to 2$, and$\mathrm{q} \to \{1, 2\}$ as follows:

Denote the first term of the left hand side function of (A.1) by $L(\beta_p, \lambda, q)$, then by chain rule it can be proved that $\frac{\partial L}{\partial \beta_p} \geq 0$. Since the partial derivatives: $\frac{\partial \hat{\beta}_1}{\partial \beta_p}, ..., \frac{\partial \hat{\beta}_{p-1}}{\partial \beta_p}$ satisfy:

$$\frac{\partial C_j}{\partial \hat{\beta}_1} . \frac{\partial \hat{\beta}_1}{\partial \beta_p} + ... + \frac{\partial C_j}{\partial \hat{\beta}_{p-1}} . \frac{\partial \hat{\beta}_{p-1}}{\partial \beta_p} + \frac{\partial C_j}{\partial \beta_p} = 0 \ (A-2)$$

Equation (A.2) implies the existence of the unique solution of $\hat{\beta}_p(\lambda, q)$, when $q \geq 1$.

(A-2) Proof of Lemma (1)

For linear regression,$\hat{y} = x\hat{\beta}_{OLS}$, and

$$\begin{aligned} d\hat{f}(\hat{y}) &= \text{div}\hat{y} \\ &= \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} \\ &= \sum_{i=1}^{n} \frac{\partial \left(x_i' \hat{\beta}_{OLS} \right)}{\partial y_i} \\ &= \frac{\partial \left(x \hat{\beta}_{OLS} \right)}{\partial y} = tr \left[x' \left(x'x \right)^{-1} x \right] \\ &= tr \left[x'x \left(x'x \right)^{-1} \right] = tr \left[I_p \right] = p. \end{aligned}$$

Also, the degrees of freedom For linear regression, $\hat{y} = x\hat{\beta}_{OLS}$ can be obtained directly as:

$$\begin{aligned} df(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i) \\ &= \frac{1}{\sigma^2} tr \left[\text{cov} \left(x \left(x'x \right)^{-1} x'y, y \right) \right] \\ &= \frac{1}{\sigma^2} tr \left(x \left(x'x \right)^{-1} x' \text{cov}(y, y) \right) \\ &= tr \left[x \left(x'x \right)^{-1} x' \right] = tr \left[x'x \left(x'x \right)^{-1} \right] \\ &= tr \left[I_p \right] = p. \end{aligned}$$

For ridge regression,$\hat{y} = x\hat{\beta}_{RR}$:

$$\begin{aligned} d\hat{f}(\hat{y}) &= \text{div}\hat{y} \\ &= \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} \\ &= \sum_{i=1}^{n} \frac{\partial \left(x_i' \hat{\beta}_{RR} \right)}{\partial y_i} \\ &= \frac{\partial \left(x \hat{\beta}_{RR} \right)}{\partial y} \\ &= tr \left[x \left(x'x + \lambda_2 I \right)^{-1} x' \right] \\ &= tr \left[x'x \left(x'x + \lambda_2 I \right)^{-1} \right] \\ &= \sum_{j=1}^{p} \frac{\gamma_j}{\gamma_j + \lambda_2}. \end{aligned}$$

Also, the degrees of freedom For linear regression, $\hat{y} = x\hat{\beta}_{RR}$ can be obtained directly as:

$$df(\hat{y}) = \tfrac{1}{\sigma^2} \sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i) = \tfrac{1}{\sigma^2} tr \left[\text{cov} \left(x \left(x'x + \lambda_2 I \right)^{-1} x'y, y \right) \right]$$
$$= \tfrac{1}{\sigma^2} tr \left(x \left(x'x + \lambda_2 I \right)^{-1} x' \text{cov}(y, y) \right) = tr \left[x \left(x'x + \lambda_2 I \right)^{-1} x' \right] = tr \left[x'x \left(x'x + \lambda_2 I \right)^{-1} \right]$$
$$= \sum_{j=1}^{p} \tfrac{\lambda_j}{\lambda_j + \lambda_2}$$

Where λ_j is the eigenvalue of the matrix x'x.

For the Lasso,$\hat{y} = x\hat{\beta}_{Lasso}$:

$$\begin{aligned} d\hat{f}(\hat{y}) &= \text{div}\hat{y} \\ &= \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial y_i} \\ &= \sum_{i=1}^{n} \frac{\partial \left(x_i' \hat{\beta}_{Lasso} \right)}{\partial y_i} \\ &= \frac{\partial \left(x \hat{\beta}_{Lasso} \right)}{\partial y} \\ &= tr \left[x' \left(x'x \right)^{-1} x \right] \\ &= tr \left[x'x \left(x'x \right)^{-1} \right] = q \prec p \end{aligned}$$

where q is the number of nonzero elements in $\hat{\beta}_{Lasso}$.

Or:

$$\begin{aligned} df\left(\hat{y}\right) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\operatorname{cov}\left(\hat{y}_i, y_i\right) \\ &= \frac{1}{\sigma^2} tr\left[\operatorname{cov}\left(x\left(x'x\right)^{-1}\left(x'y - \frac{1}{2}k_1\ell\right), y\right)\right] \\ &= \frac{1}{\sigma^2} tr\left(x\left(x'x\right)^{-1} x'\operatorname{cov}\left(y,y\right)\right) \\ &= tr\left[x\left(x'x\right)^{-1}x'\right] \\ &= tr\left[x'x\left(x'x\right)^{-1}\right] \\ &= tr\left[I_q\right] = q \prec p, \end{aligned}$$

where q is the number of nonzero elements in $\hat{\beta}_{Lasso}$.

For Elastic-net, $\hat{y} = x\hat{\beta}_{Elastic-Net}$

$$df\left(\hat{y}\right) = \sum_{j=1}^{q}\frac{\lambda_j}{\lambda_j + \lambda_2},$$

which gives

$$\begin{aligned} df\left(\hat{y}\right) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} cov\left(\hat{y}_i, y_i\right) \\ &= \frac{1}{\sigma^2} tr\left[cov\left(x\left(x'x + \lambda_2 I\right)^{-}\left(x'y - \frac{1}{2}\lambda_1\ell\right), y\right)\right] \\ &= \frac{1}{\sigma^2} tr\left(x\left(x'x + \lambda_2 I\right)^{-} x' cov\left(y,y\right)\right) \\ &= tr\left[x'x\left(x'x + \lambda_2 I\right)^{-}\right] \\ &= \sum_{j=1}^{q}\frac{\gamma_j}{\gamma_j + \lambda_2} \\ &\approx q \; for \; small \; values \; of \; \lambda_2, \end{aligned}$$

where q is the number of nonzero elements in $\hat{\beta}_{Lasso}$ and $(.)^{-}$is the generalized inverse.