The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 615-624

Variable Selection via Penalized Regression[1)]

Young Joo Yoon[2)] and Moon Sup Song[3)]

Abstract

In this paper, we review the variable-selection properties of LASSO and SCAD in penalized regression. To improve the weakness of SCAD for high noise level, we propose a new penalty function called MSCAD which relaxes the unbiasedness condition of SCAD. In order to compare MSCAD with LASSO and SCAD, comparative studies are performed on simulated datasets and also on a real dataset. The performances of penalized regression methods are compared in terms of relative model error and the estimates of coefficients. The results of experiments show that the performance of MSCAD is between those of LASSO and SCAD as expected.

Keywords : Penalized regression, Penalty function, LASSO, SCAD, MSCAD

1. Introduction

Variable selection is an important topic in linear regression analysis. By retaining a subset of the predictors and discarding the rest (for example, stepwise forward selection or backward elimination), the reduced regression model is more interpretable and sometimes reduces the prediction error. Although subset selection methods are practically useful, they have several drawbacks. The most severe drawback is their lack of stability. Since the variables are either retained or discarded, subset selection methods often show high variance and don't reduce the prediction error of the full model. Shrinkage methods are useful to overcome these difficulties, since they don't suffer as much from high variability (Hastie et al. 2001, Chapter 3).

In Section 2, we introduce penalized regression. We review some well-known penalized regression methods such as LASSO and SCAD methods. We describe and discuss the proposed penalty function which is motivated to satisfy the advantages of LASSO and SCAD in Section 3. The behaviors of the proposed method is between LASSO and SCAD. We expect that the performance of the proposed method is more steady than those of LASSO and SCAD. Finally, summary and concluding remarks are provided in Section 4.

---

1) This research was in part supported by the Brain Korea 21 Project.

2) Ph.D., Department of Statistics, Seoul National University, Seoul, 151-742, Korea
E-mail : youngjoo@stats.snu.ac.kr

3) Professor, Department of Statistics, Seoul National University, Seoul, 151-742, Korea

2. Penalized Regression

2.1 Ridge Regression

Consider the linear regression model

$$y = X\beta + \varepsilon,$$

where y is an $n\times 1$ vector and X is an $n\times p$ matrix. When we use the squared-error loss criteria to solve the linear regression problem, the collinearity in the design matrix X causes unstable solutions. To remedy this problems, ridge regression technique imposes a penalty on the size of regression coefficients. That is, the ridge regression shrinks the regression coefficients by minimizing the penalized residual sum of squares, and the estimator is defined by

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta}\left(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right).$$

Here $\lambda \geq 0$ is a regularization (complexity) parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. The ridge regression solution is given by

$$\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y.$$

The solution adds a positive constant to the diagonal of $X^T X$ before inversion. This makes the problem nonsingular, even if $X^T X$ is not of full rank, and was the main motivation for ridge regression (Hoerl and Kennard, 1970, and Hastie et al., 2001).

2.2 The LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) was proposed by Tibshirani (1996). The LASSO is a shrinkage method like ridge, using L_1 penalty instead of L_2. Thus the LASSO estimator is defined by

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta}\Big(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|\Big).$$

Because of the L_1 penalty the solution is not linear in y, and usually quadratic programming methods are used to solve the LASSO estimator. As the regularization parameter λ increases, some of the coefficients tend to be exactly zero.

2.3 The SCAD

Fan and Li (2001) proposed a variable selection method based on SCAD (Smoothly Clipped Absolute Deviation) penalty function. As a motivation for the SCAD penalty they claimed that a good penalty function should result in an estimator with the following three properties:

- *Unbiasedness*: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modelling bias.
- *Sparsity*: The resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.
- *Continuity*: The resulting estimator is continuous in data to avoid instability in model prediction.

To introduce some insights on variable selection procedures, in this subsection we assume that the columns of X are orthonormal. Following the motivations in Fan and Li (2001), denote $z = X^T y$ and let $\hat{y} = XX^T y$. A form of the penalized least squares is as follows.

$$\frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|) = \frac{1}{2}\|y - \hat{y}\|^2 + \frac{1}{2}\sum_{j=1}^{p}(z_i - \beta_j)^2 + \lambda \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

Since the minimization problem of this form is equivalent to minimizing componentwise, we only consider the penalized least square problem

$$\frac{1}{2}(z-\beta)^2 + p_\lambda(|\beta|),$$

where $p_\lambda(|\cdot|)$ denote $\lambda p(|\cdot|)$.

The following two penalty functions are well known.

- L_2 penalty: $p_\lambda(|\beta|) = \lambda|\beta|^2$.
- L_1 penalty: $p_\lambda(|\beta|) = \lambda|\beta|$.

The L_2 penalty gives the ridge regression estimator. The L_1 penalty yields the LASSO estimator discussed Section 2.2. In this case the LASSO estimator is of the form (see Fan and Li, 2001, Section 2)

$$\hat{\beta} = \text{sign}(z)(|z| - \lambda)_+,$$

where the subscript + indicates a value of zero for negative values of the argument.

Fan and Li (2001) provided some insights on the three requirements to be a good penalty function. $p'_\lambda(|\beta|) = 0$ *for large* $|\beta|$ is a sufficient condition for unbiasedness for a large parameter. A sufficient condition for the resulting estimator to be a thresholding rule is that *the minimum of the function* $|\beta| + p'_\lambda(|\beta|)$ *is positive.* A sufficient and necessary condition for continuity is that *minimum of the function* $|\beta| + p'_\lambda(|\beta|)$ *is attained at 0.* (See Fan and Li, 2001). In the case of LASSO, the L_1 penalty does not satisfy the sufficient condition for unbiasedness for large parameters, and for ridge, the L_2 penalty does not satisfy the conditions for sparse solutions and unbiasedness for large parameters. To satisfy the conditions for unbiasedness, sparsity and continuity, Fan and Li (2001) proposed the SCAD penalty function defined by

$$p_\lambda(|\beta|) = \lambda \begin{cases} |\beta| & , \ 0 \le |\beta| \le \lambda \\ -\dfrac{1}{2(a-1)\lambda}(\beta^2 - 2a\lambda|\beta| + \lambda^2) & , \ \lambda < |\beta| \le a\lambda \\ \dfrac{1}{2}(a+1)\lambda & , \ |\beta| > a\lambda \end{cases}$$

for some $a > 2$. The derivative of the SCAD function is given by

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \le \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\}$$

for some $a > 0$ and $\beta > 0$. The solution to the SCAD penalty is given by

$$\hat{\beta} = \begin{cases} \text{sign}(z)(|z| - \lambda)_+ & , \ |z| \le 2\lambda \\ \dfrac{1}{a-2}((a-1)z - \text{sign}(z)a\lambda) & , \ 2\lambda < |z| \le a\lambda \\ z & , \ |z| > a\lambda \end{cases}.$$

3. Proposed Method

In Section 2 we considered the penalized least squares methods based on LASSO and SCAD. These methods can be applied to variable selection problem, since both algorithms give sparse solutions as the regularization parameter increases. In this section we compare the performance of LASSO and SCAD on a toy example and propose a new penalty function whose behavior is expected to be between LASSO and SCAD.

3.1 Motivation

Consider the following toy example:

$$y = x^T\beta + \sigma\varepsilon,$$

where $\beta = (3, 1.5, 1, 1, 2, 0, 0, 0)^T$ and the components of x and ε are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. The number of observations is $n = 40$, and $\sigma = 0.5$ and 3. Monte Carlo simulations with iteration $M = 100$ are performed to compare LASSO and SCAD. The value of a in SCAD is set to 3.7, which is the value suggested by Fan and Li (2001). The regularization parameters are estimated by using validation set. To compare the performance of methods, the RME (relative model error) is used, which is the ratio of model errors of the given method to those of the least squares method. Model error is defined as follows:

$$\mathrm{ME}(\hat{\mu}) = E(\hat{\mu}(x) - \mu(x))^2,$$

where $\mu(x) = x^T\beta$, $\hat{\mu}(x) = x^T\hat{\beta}$. Estimated model error is given by

$$\widehat{\mathrm{ME}}(\hat{\mu}) = \frac{1}{n_T} \sum_{x \in \text{test set}} (x^T\hat{\beta} - x^T\beta)^2,$$

where n_T is the size of test set, x's are the test examples. In this example, $n_T = 500$. The average and median of RME (in %) are summarized in Table 1.

While the performance of SCAD is better than that of LASSO for small σ, the performance of LASSO is better than that of SCAD for large σ. These results are similar to the results in Fan and Li (2001).

<Table 1> Simulation results for the toy example

$\sigma=0.5$		
Method	Avg. of RME(%)	Median of RME (%)
LASSO	81.90	77.39
SCAD	65.28	63.14
$\sigma=3$		
Method	Avg. of RME(%)	Median of RME (%)
LASSO	80.79	86.11
SCAD	94.12	95.67

3.2 Proposed Penalty

Since the performance of SCAD becomes worse compared to that of LASSO for high noise level, a modified SCAD penalty function is considered, which relaxes the unbiasedness condition of SCAD. The modified SCAD (called MSCAD) penalty's derivative is defined as follows:

$$p'_{\lambda}(\beta)=\begin{cases}\lambda & , \beta\le\lambda \\ l(\beta)p'_{LASSO}(\beta)+[1-l(\beta)]p'_{SCAD}(\beta) & , \beta>\lambda\end{cases} \quad \text{for } \beta>0$$

where $p'_{LASSO}(\cdot)$ and $p'_{SCAD}(\cdot)$ denote LASSO and SCAD penalty's derivatives respectively, $l(\lambda)=1$, $l(\beta)\to 0$ as $\beta\to\infty$ and $l(\beta)$ is decreasing. This derivative is a convex combination of LASSO and SCAD penalty's derivatives. An example of $l(\beta)$ is as follows:

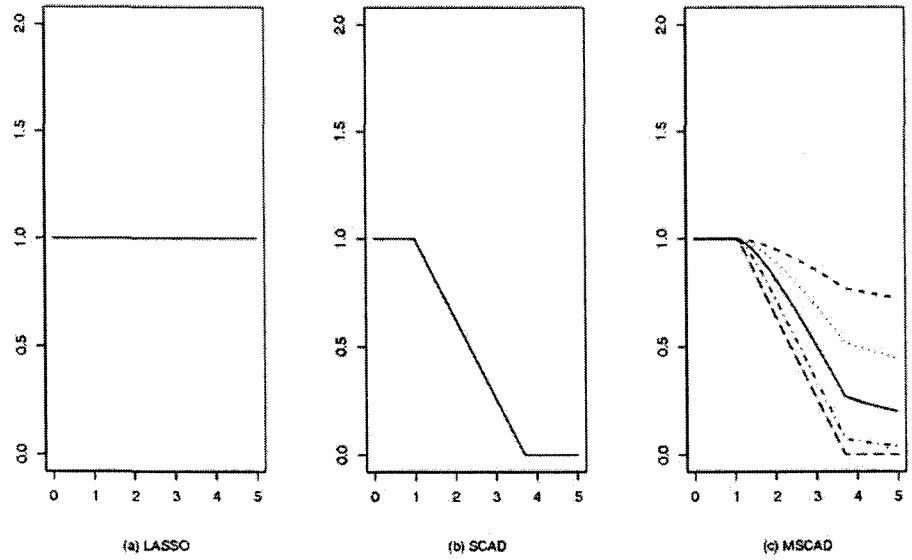
$$l(\beta)=\frac{1}{(\beta-\lambda+1)^k}.$$

The derivative of SCAD penalty function is zero for $|\beta|>a\lambda$ and that of MSCAD approaches zero. The MSCAD estimator is not unbiased but nearly unbiased for large values of $|\beta|$. According to the unified algorithm, which was proposed by Fan and Li (2001), the solution to the penalized least squares problem can be obtained by iteratively computing the ridge regression

$$\beta_1=\{X^TX+n\Sigma_{\lambda}(\beta_0)\}^{-1}X^Ty,$$

where $\Sigma_\lambda(\beta_0)$ is the diagonal matrix with diagonal elements depending on β_0 through the derivative p'_λ.

Thus for the penalized least squares problem, the differences between LASSO, SCAD, and MSCAD can be characterized by p'_λ. Figure 1 shows the plots of $p'_\lambda(\beta)$ of three penalty functions with $a=3.7$ and $\lambda=1$, for $\beta>0$.



<Figure 1> Plots of $p_\lambda'(\beta)$. $\lambda=1$, $a=3.7$ for LASSO, SCAD and MSCAD. In (c), dashed, dotted, solid, dotdash and longdash correspond to $k=1/5$, $k=1/2$, $k=1$, $k=2$ and $k=5$, respectively.

The RME's of MSCAD method ($k=1/2,1,2$) relative to the least squares method are computed and summarized for the toy example in Section 3.1. The results are summarized in Table 2. We can see that for small σ, the RME's of MSCAD are larger than that of SCAD, but smaller than that of LASSO. For large σ, these aspects are conversely showed. The results show that the performance of MSCAD is between those of LASSO and SCAD. For larger k, MSCAD is close to SCAD. We expect that the performance of MSCAD is more steady than those of LASSO and SCAD. In the next section we will compare these methods with a real data example.

<Table 2> Simulation results for the toy example in Section 3.1

$\sigma = 0.5$		
Method	Avg. of RME(%)	Median of RME (%)
MSCAD ($k = 1/2$)	78.85	77.35
MSCAD ($k = 1$)	74.56	75.01
MSCAD ($k = 2$)	71.64	67.46
$\sigma = 3$		
Method	Avg. of RME(%)	Median of RME (%)
MSCAD($k = 1/2$)	85.19	92.36
MSCAD($k = 1$)	89.15	95.87
MSCAD($k = 2$)	93.47	97.65

3.3 Real Data example - Prostate Cancer Data

The data come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures in 97 men who were about to receive a radical prostatectomy. This example was considered in Hastie et al. (2001, Chapter 3) to compare various variable selection and shrinkage methods. The goal is to predict the log of PSA (**lpsa**) from a number of measurements including log-cancer-volume (**lcavol**), age, log of benign prostatic hyperplasia amount (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**). First we standardized the predictors to have unit variance. Then we performed ordinary least squares method and various variable selection methods for this dataset. The variable selection methods used in this analysis are the best-subset selection using an all-subset search (see Hastie et al., 2001), LASSO, SCAD and MSCAD. In MSCAD, the considered values of k are 1/2, 1 and 2. We divided this dataset into training dataset of size 67 and test dataset of size 30. To select regularization parameter, 10-fold cross validation method is applied to the training dataset. The test dataset judges the performance of the selected model. The estimated coefficients (except intercept) and test errors with SE(standard errors) are showed in Table 3.

The performances of penalized regression methods are better than those of ordinary least squares and best subset selection methods. Among the penalized methods, SCAD has the best performance. But the difference is not so much. The performance of MSCAD is similar to that of SCAD for large k, while MSCAD method for small k performs similar to LASSO. These aspects are not so different from those of toy example in the previous section. The selected variables for three penalized methods are similar (**pgg45**). Especially for **lcavol**, the estimated

coefficients of SCAD and MSCAD are larger than that of LASSO. For other variables, the latter are larger than the former. This may be caused by the unbiasedness of SCAD and near unbiasedness of MSCAD for large coefficients, which LASSO does not satisfy.

<Table 3> Estimated coefficients and test errors for prostate cancer data

					MSCAD		
Variable	LS	Best Subset	LASSO	SCAD	$k=1/2$	$k=1$	$k=2$
lcavol	0.680	0.740	0.544	0.806	0.574	0.610	0.683
lweight	0.305	0.367	0.199	0.104	0.177	0.168	0.148
age	-0.141	-	-	-	-	-	-
lbph	0.210	-	0.061	0.028	0.040	0.035	0.026
svi	0.305	-	0.126	0.006	0.093	0.073	0.036
lcp	-0.288	-	-	-	-	-	-
gleason	-0.021	-	-	-	-	-	-
pgg45	0.267	-	0.041	-	0.023	0.016	0.005
Test Error	0.586	0.574	0.485	0.472	0.483	0.481	0.479
SE	0.184	0.156	0.158	0.133	0.155	0.152	0.145

4. Summary and Concluding Remarks

The variable-selection properties of LASSO and SCAD in penalized regression are studied in this paper. While the performance of SCAD is better than that of LASSO for low noise level, the converse result is true for high noise level. To improve the weakness of SCAD for high noise level, we proposed the MSCAD (Modified SCAD) penalty function which relaxes the unbiasedness conditions of SCAD. The results for simulated data and real data showed that the performance of MSCAD is between LASSO and SCAD as expected. That is, the performance of MSCAD is more steady than those of LASSO and SCAD. In this paper, we considered only some special cases of MSCAD. It is necessary to study theoretical properties of general MSCAD, such as the existence of solution, oracle property (see Fan and Li, 2001) and the behaviors of the estimates. A unified algorithm proposed by Fan and Li (2001) is used to estimate regression coefficients of SCAD and MSCAD penalty functions. This algorithm contains a matrix inversion. Thus the stability of the algorithm can not be guaranteed. In addition, the solution may be a local one since the object function is locally approximated by quadratic function. To improve these problems of the unified algorithm, stagewise approaches such as gradient boosting (Friedman, 2001) and margin boost (Mason et al., 2000) can be considered.

References

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

[2] Friedman, J. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, **29**, pp. 1189-1232.

[3] Hastie, T., Tibshirani, R. and Friedman, J.H. (2001). *Elements of Statistical Learning*. Springer-Verlag, New York.

[4] Hoerl, A.E. and Kennard, R. (1970). Ridge regression : biased estimation for nonorthogonal problems, *Technometrics*, **12**, pp. 55-67.

[5] Mason, L., Baxter, J., Bartlett, P.L. and Frean, M. (2000). Functional gradient techniques for combining hypotheses, In A.J. Smola, P.L. Bartlett, B. Scholkopf and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, Cambridge: MIT press.

[6] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnostics and treatment of adenocarcinoma of the prostate II. radical prostatectomy treated patients, *Journal of Urology*, **16**, pp. 1076-1083.

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

[Received May 2005, Accepted September 2005]