ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 8, Issue, 10, pp. 20696-20704, October, 2017

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

VARIABLE SELECTION FOR SURVIVAL ANALYSIS USING LASSO VS LAD REGRESSION

Sathish Kumar S[1]* and Elangovan R[2]

[1]Department of Community Medicine, VMMC & H, Karaikal-609 609
[2]Department of Statistics, Annamalai University, Annamalainagar-608 002
Tamilnadu, India

DOI: http://dx.doi.org/10.24327/ijrsr.2017.0810.0944

ARTICLE INFO

Article History:

Received 16th July, 2017
Received in revised form 25th August, 2017
Accepted 23rd September, 2017
Published online 28th October, 2017

Key Words:

LASSO, Adaptive LASSO, LAD, SCAD.

ABSTRACT

Variable selection is an important topic in linear regression analysis especially in high-dimensional medical data sets, The Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani is a popular technique for model selection and estimation in linear regression model. The Least Angle Regression (LAR) procedure by Efron *et al*. (2004) provides a method for fast computation of LASSO solution in regression problems. L_1 penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent over fit arising due to either collinearity of the covariates or high-dimensionality. L_1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage. It is important to note that shrinkage methods are generally not invariant to the relative scaling of the covariates. Variable selection for LAD regression receives much attention in recent literature. In this paper it is proposed to study the variable selection for survival analysis using LASSO Vs LAD regression. Numerical illustrations are substantiated through real data example.

INTRODUCTION

Variable selection is an important topic in linear regression analysis especially in high-dimensional medical data sets, and has challenged many contemporary statistical problems from many frontiers of scientific disciplines. Over the past few years, many techniques have been proposed for variable selection in high-dimensional medical data sets. A field, in which such methods are applicable, is survival analysis. The Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani is a popular technique for model selection and estimation in linear regression model. The Least Angle Regression (LAR) procedure by Efron *et al*. (2004) provides a method for fast computation of LASSO solution in regression problems. Osborne *et al*. (2000) derived the optimality conditions associated with the LASSO solution. Donoho and Elad (2003) and Donoho (2004) proved some analytical properties of the L_1 penalization approach for determining the sparsest solution for an under-determined linear system. Some statistical properties of the LASSO-based estimator of the regression parameter have been derived by Knight and Fu (2000). L_1 penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent over fit arising due to either collinearity of the covariates or high-dimensionality. L_1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage. It is important to note that shrinkage methods are generally not invariant to the relative scaling of the covariates. Variable selection for LAD regression receives much attention in recent literature. In this paper it is proposed to study the variable selection for survival analysis using LASSO Vs LAD regression. Numerical illustrations are substantiated through real data example.

Variable Selection via Penalized Likelihood

Consider the usual linear regression model

$$y = X\beta + \varepsilon , \qquad \ldots(1)$$

where y is an $n \times 1$ vector and X is an $n \times d$ matrix. As in the traditional linear regression setup, we assume that y_i's are conditionally independent given the design matrix. The ordinary least-squares estimate is given by $\hat{\beta} = (X^T X)^{-1} X^T y$. To attenuate possible excessive modeling biases, a large number of predictors are usually

Corresponding author:* **Sathish Kumar S
Statistician cum Lecturer, Department of Community Medicine, VMMC & H, Karaikal-609 609

introduced at the initial stage of modeling. To enhance predictability and to select significant variables, statisticians usually apply three standard techniques, stepwise deletion, subset selection and ridge regression, to improve the least-squares estimate. However, while they are practically useful, these techniques are ad hoc and subjective.

The selection procedures usually ignore stochastic errors inherited in the previous stage of variable selections. Hence, their theoretical properties are somewhat hard to understand. In an attempt to automatically and simultaneously select variables, Tibshirani (1996) proposed a new approach, called LASSO, retaining good features of both subset selection and ridge regression. LASSO in fact coincides with a soft-thresholding rule when design matrices are orthonormal. See also the bridge regression proposed in Frank and Friedman (1993), For a more detailed discussion, refer to Fan and Li (2001).

Penalized Least-Squares and Variable Selection

There are strong connections between thresholding rules and subset selection in linear regression models. In this section we assume that the columns of X in (1) are orthonormal. Then the least-squares estimate in the full model is $\hat{\beta} = \mathrm{X}^T y$, a part of the orthogonal transform of the vector y.

Thresholding and Variable Selection

Denote by $z = \mathrm{X}^T y$ and assume that $\varepsilon \sim N(0,\sigma^2 I_n)$ in model (1). Then z is a multivariate normal random vector with independent components. This allows us to consider a Gaussian white noise model:

$$z_i = \theta_i + \varepsilon_i \quad \text{with } \varepsilon \sim N(0,\sigma^2) \text{ for i = 1,…,d.} \qquad \text{…(2)}$$

Suppose that the θ's in (2) are sparse so that they can reasonably be modelled as an i.i.d. realization from a double exponential distribution with a scale parameter λ_1. Then the Bayesian estimate is the minimizer of

$$\frac{1}{2}\sum_{i=1}^{d}(z_i-\theta_i)^2+\lambda\sum_{i=1}^{d}|\theta_i|, \qquad \text{…(3)}$$

Where $\lambda = \sigma^2/\lambda_1$

Minimization of (3) is equivalent to minimizing (3) component-wise.

$$\hat{\theta}_j = \operatorname{sgn}(z_j)(|z_j|-\lambda)_+ \qquad \text{…(4)}$$

If the L_1-penalty in (3) is replaced by the L_q-penalty, it results in bridge regression proposed by Frank and Friedman (1993) and carefully studied by Fu (1998). Particularly, when q = 2, it leads to the usual ridge regression. Some interesting results can be seen in Lu and Zhang (2007)

Penalized Least-Squares and Variable Selection

Consider a general form of penalized least-squares:

$$\frac{1}{2}\sum_{j=1}^{d}(z_j-\theta_j)^2+\lambda\sum_{j=1}^{d}p_j(|\theta_j|). \qquad \text{…(5)}$$

The penalty functions $p_j(\theta)$ in (5) are not necessarily the same for all j. For example, one may wish to keep important predictors in a parametric model and hence is not willing to penalize their corresponding parameters. For simplicity of presentation, we will assume that the penalty functions for all coefficients are the same, denoted by $p(|\theta|)$ Furthermore, we denote $\lambda p(|\theta|)$ by $p_\lambda(|\theta|)$ as $p(|\theta|)$ can be allowed to depend on λ. Extensions to the case with different thresholding functions do not involve any extra difficulties.

The minimization problem of (5) is equivalent to minimizing component wise the penalized least-squares problem:

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|). \qquad (6)$$

The solution to (6) is necessarily a thresholding when the minimum of the function $|\theta| + \lambda p'_\lambda(|\theta|) > 0$ is positive. This is because the derivative function has no zero crossing for small values of $|z|$.

Fan observed that the penalized least-squares estimator with the penalty function $p(|\theta|) = |\theta| I(|\theta| \le \lambda) + \lambda/2 I(|\theta| > \lambda)$ leads to the hard-thresholding rule

$$\hat{\theta} = zI(|z| > \lambda) \qquad \text{…(7)}$$

This penalty function does not over penalize the large value of $|\theta|$. In this response, Antoniadis (1999) improves Fan's proposal by using the following hard thresholding penalty function:

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta|-\lambda)^2 I(|\theta| < \lambda) \qquad \text{...(8)}$$

With the clipped L_1-penalty function

$$p_\lambda(\theta) = \lambda \min(|\theta|, \lambda) \qquad \text{…(9)}$$

the solution is a mixture of soft and hard thresholding rule

$$\hat{\theta}_j = \operatorname{sgn}(z)(|z|-\lambda)_+ I(|z| \le 1.5\lambda) + zI(|z| > 1.5\lambda) \quad \text{…(10)}$$

Smoothly Clipped Absolute Deviation Penalty

All of penalty functions introduced so far do not satisfy both mathematical conditions imposed in the last paragraph for a continuous and thresholding rule. The continuous differentiable penalty function defined by

$$p'(\theta) = I(\theta \le \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda} I(\theta > \lambda) \text{ for some } a > 2$$

and $a > 0$, …(11)

Improves the properties of the L_1-penalty and the hard-thresholding penalty function given by

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta|-\lambda)^2 I(|\theta| < \lambda) \qquad \text{…(12)}$$

We will call this penalty function as smoothly clipped absolute deviation (SCAD) penalty. This corresponds to a quadratic spine function with knots at λ and $a\lambda$. This penalty function leaves large value of θ not excessively penalized and makes the solution continuous. The resulting solution is given by

$$\hat{\theta} = \begin{cases} \operatorname{sgn}(z)(|z|-\lambda)_+, \\ \{(a-1)z - \operatorname{sgn}(z)a\lambda\}/(a-2), \\ z, \end{cases}$$

$$when|z| \le 2\lambda;$$

$$when 2\lambda < |z| \le a\lambda; \qquad \ldots(13)$$

$$when|z| > a\lambda.$$

For simplicity, we will call all procedures using the SCAD penalty as SCAD, refer to Fan (1997).

Penalized Least-Squares and Likelihood

In classical linear regression models, the least-squares estimate is obtained via minimizing the sum of squared residual errors. Therefore (5) can be naturally extended to the situation in which design matrices are not orthonormal. Similar to (5), a general form of penalized least-squares is

$$\frac{1}{2n}\sum_{i=1}^{n}(y_j - X_i^T\beta)^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

Or equivalently

$$\frac{1}{2}(y - X\beta)^T(y - X\beta) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad \ldots(14)$$

Minimizing (14) with respect to β leads to a penalized least-squares estimator of β.

It is well known that the least-squares estimate is not robust, one can consider the outlier-resistant loss functions such as the L_1-penalty or more general Huber's ψ-function, refer to Huber(1981). Therefore instead of minimizing (9), we minimize

$$\sum_{i=1}^{n}\psi(|y_i - X_i\beta|) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad \ldots(15)$$

With respect to β. This results in a robust least-squares estimator.

For generalized linear models, statistical inference are based on underlying likelihood functions. The penalized maximum likelihood estimator can be used to select significant variables. Assume that the collected data (X_i, Y_i) are independent samples. Conditioning on X_i, Y_i has a density $f_i(g(X_i^T\beta), y_i)$, where g is a known link function. Denoted by $l_i = \log f_i$, the conditional log-likelihood of Y_i. A general form of penalized likelihood is

$$-\sum_{i=1}^{n} l_i(g(X_i^T\beta), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \qquad \ldots(16)$$

To obtain a penalized maximum likelihood estimator of β, we minimize (16) with respect to β for some thresholding parameter λ.

Selection of thresholding parameters

We need to estimate the thresholding parameters λ and a as discussed in the previous sections. Denote by θ the tuning parameters to be estimated, i.e., $\theta = (\lambda, a)$ for the SCAD, while $\theta = \lambda$ for other thresholding. Here we discuss two methods of estimating θ: fivefold cross-validation and generalized cross-validation, as suggested by Breiman (1995), Tibshirani (1996) and Fu (1998). For completeness, we now describe the details of the cross-validation procedures. Denote $l\{\hat{\beta}(\theta)\}$ by the first term in

$$l(\beta) + n\sum_{j=1}^{d} P_\lambda(|\beta_j|) \qquad \ldots(17)$$

replacing β by its estimate $\hat{\beta}$ obtained when the tuning parameters θ are used. Then $l\{\hat{\beta}(\theta)\}$ can be regarded as a measure of goodness of fit. The fivefold cross-validation procedure is as follows: Denote the full training set by T, and cross-validation training and test set by $T - T^v$ and T^v, for $v = 1,\ldots,5$. For each θ and v, we find the estimator $\hat{\beta}^v(\theta)$ of β using the training set $T - T^v$. Let $l_v\{\hat{\beta}(\theta)\}$ be the $l\{\hat{\beta}(\theta)\}$ for test set T^v. Form the cross-validation criterion as

$$CV(\theta) = \sum_{v=1}^{5} l_v\{\hat{\beta}(\theta)\}.$$

We find a $\hat{\theta}$ that minimizes $CV(\theta)$.

The second method is the generalized cross-validation. For linear regression models, we update the solution by

$$\beta_1(\theta) = \{X^T X + n\Sigma_\lambda(\beta_0)\}^{-1} X^T y.$$

Thus the fitted value by of $\hat{y}$ is $X^T\{X^T X + n\Sigma_\lambda(\beta_0)\}^{-1}X^T y$, and

$$P_X\{\hat{\beta}(\theta)\} = X^T\{X^T X + n\Sigma_\lambda(\hat{\beta})\}^{-1}X^T$$

can be regarded as a projection matrix. Define the number of effective parameters in the penalized least-squares fit as $e(\theta) = tr[P_X\{\hat{\beta}(\theta)\}]$. Therefore the generalized cross-validation statistic is

$$GCV(\theta) = \frac{1}{n}\frac{l\{\hat{\beta}(\theta)\}}{\{1 - e(\theta)/n\}^2}$$

and $\hat{\theta} = \arg\min_{\theta}\{GCV(\theta)\}$. Similarly the corresponding generalized cross-validation statistics can be de_ned for robust regression models and likelihood based linear models.

Adaptive Lasso and its Oracle Properties

Let us consider model estimation and variable selection in linear regression models. Suppose that $y = (y_1,\ldots,y_n)^T$ is the response vector and $\mathrm{X}_j = (\mathrm{X}_{1j},\ldots,\mathrm{X}_{nj})^T$, j=1,…,p are the linearly independent predictors. Let $\mathrm{X} = \lfloor \mathrm{X}_1,\ldots,\mathrm{X}_p \rfloor$ be the predictor matrix. We assume that E[y|x]= $\beta_1^* x_1 + \ldots + \beta_p^* x_p$. Without loss of generality, we assume that the data are centered, so the intercept is not included in the regression function. Let $A = \{j : \beta_j^* \neq 0\}$ and further assume that $|A| = p_0 < p$. Thus the true model depends only on a subset of the predictors. Denote by $\hat{\beta}(\delta)$ the coefficient estimator produced by a fitting procedure δ. Using the language of Fan and Li (2001), we call δ an oracle procedure if $\hat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

1. Identifies the right subset model, $\{j : \beta_j^* \neq 0\} = A$
2. Has the optimal estimation rate, $\sqrt{n}(\hat{\beta}(\delta)A - \beta_A^*) \to_d N(0,\Sigma^*)$, where Σ^* ∗ is the covariance matrix knowing the true subset model.

It has been argued (Fan and Li 2001; Fan and Peng 2004) that a good procedure should have these oracle properties. However, some extra conditions besides the oracle properties, such as continuous shrinkage, are also required in an optimal procedure. Ordinary least squares (OLS) gives nonzero estimates to all coefficients. Traditionally, statisticians use best-subset selection to select significant variables, but this procedure has two fundamental limitations. First, when the number of predictors is large, it is computationally infeasible to do subset selection. Second, subset selection is extremely variable because of its inherent discreteness refer to Breiman (1995), Fan and Li (2001). Stepwise selection is often used as a computational surrogate to subset selection; nevertheless, stepwise selection still suffers from the high variability and in addition is often trapped into a local optimal solution rather than the global optimal solution. Furthermore, these selection procedures ignore the stochastic errors or uncertainty in the variable selection stage refer to Fan and Li (2001), Shen and Ye (2002). The lasso is a regularization technique for simultaneous estimation and variable selection as suggested by (Tibshirani 1996). The lasso estimates are defined as

$$\hat{\beta}(lasso) = \arg\min_{\beta}\left\| y - \sum_{j=1}^{p} \mathrm{X}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \qquad \ldots(18)$$

where λ is a nonnegative regularization parameter. The second term in (18) is the so-called "l_1 penalty," which is crucial for the success of the lasso. Lasso cannot be an oracle procedure, however, the asymptotic setup is somewhat unfair, because it forces the coefficients to be equally penalized in the l_1 penalty. We can certainly assign different weights to different coefficients. Let us consider the weighted lasso,

$$\arg\min_{\beta}\left\| y - \sum_{j=1}^{p} \mathrm{X}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where **w** is a known weights vector. We show that if the weights are data-dependent and cleverly chosen, then the weighted lasso can have the oracle properties. The new methodology is called the adaptive lasso. We now define the adaptive lasso. Suppose that $\hat{\beta}$ is a root *n*-consistent estimator to β^*; for example, we can use $\hat{\beta}(ols)$.

Pick a $\gamma > 0$, and define the weight vector $\hat{W} = 1/|\hat{\beta}|^{\gamma}$. The adaptive lasso estimates $\hat{\beta}^{*(n)}$ are given by

$$\hat{\beta}^{*(n)} = \arg\min_{\beta}\left\| y - \sum_{j=1}^{p} \mathrm{X}_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j| \qquad \ldots(19)$$

Similarly, let $A_n^* = \{j : \hat{\beta}_j^{*(n)} \neq 0\}$.

It is worth emphasizing that (19) is a convex optimization problem, and thus it does not suffer from the multiple local minimal issue, and its global minimizer can be efficiently solved. This is very different from concave oracle penalties. The adaptive lasso is essentially an l_1 penalization method.

Oracle Properties

With a proper choice of λ_n, adaptive lasso enjoys the oracle properties.

Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimates must satisfy the following:

1. Consistency in variable selection: $\lim_n P(A_n^* = A) = 1$
2. Asymptotic normality: $\sqrt{n}(\hat{\beta}_A^{*(n)} - \beta_A^*) \to_d N(0, \sigma^2 \times C_{11}^{-1})$.

Variable Selection through the LAD-LASSO

Consider the linear regression model,

$$y_i = x_i'\beta + \in_i \,..\, i = 1,\ldots.n, \qquad \ldots(20)$$

Where $X_i = (x_{i1},\ldots,x_{ip})'$ is the p-dimensional regression co-variate, $\beta = (\beta_1,\ldots,\beta_p)'$ are the associated regression coefficients, and $\in_i$ are iid random errors with median 0. Moreover, assume that $\beta_j \neq 0$ for $j \leq p_0$ and $\beta_j = 0$ for $j > p_0$ for some $p_0 \geq 0$. Thus the correct model has p$_0$ significant and (p - p$_0$) insignificant regression variables.

Usually, the unknown parameters of model (20) can be estimated by minimizing the OLS criterion, $\sum_{i=1}^{n}(y_i - X_i'\beta)^2$. Furthermore, to shrink unnecessary coefficients to 0, Tibshirani (1996) proposed the followed lasso criterion:

$$lasso = \sum_{i=1}^{n}(y_i - X_i'\beta)^2 + n\lambda\sum_{j=1}^{p}\left|\beta_j\right|,$$

Where $\lambda > 0$ is the tuning parameter. Because lasso uses the same tuning parameters for all regression coefficients, the resulting estimators may suffer an appreciable bias (Fan and Li 2001). Hence we further consider the following modified lasso criterion:

$$lasso* = \sum_{i=1}^{n}(y_i - X_i'\beta)^2 + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right|$$

which allows for different tuning parameters for different coefficients. As a result, lasso* is able to produce sparse solutions more effectively than lasso.

lasso* is very sensitive to outliers. To obtain a robust lasso-type estimator, we further modify the lasso* objective function into the following LAD-lasso criterion:

$$LAD - LASSOQ(\beta) = \sum_{i=1}^{n}\left|y_i - X_i'\beta\right| + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right|.$$

As can be seen, the LAD-lasso criterion combines the LAD criterion and the lasso penalty, and hence the resulting estimator is expected to be robust against outliers and also to enjoy a sparse representation. Computationally, it is very easy to find the LAD-lasso estimator. Specifically, we can consider an augmented dataset $\{(y_i^*, X_i^*)\}$ with $i = 1,\dots,n+p$, where $(y_i^*, X_i^*) = (y_i, X_i)$ for $1 \le i \le n$, $(y_{n+j}^*, X_{n+j}^*) = (0, n\lambda_j e_j)$ for $1 \le j \le p$, and e_j is a p-dimensional vector with the jth component equal to 1 and all others equal to 0. It can be easily verified that

$$LAD - LASSOQ(\beta) = \sum_{i=1}^{n+p}\left|y_i^* - X_i^*\beta\right|.$$

Properties

we decompose the regression coefficient as $\beta = (\beta_a', \beta_b')'$ where $\beta_a = (\beta_1,\dots,\beta_{p0})'$ and $\beta_a = (\beta_{p0+1},\dots,\beta_p)'$.

Table 1 Simulation results for double-exponential error

						No. of zeros			
σ	n	Method	Underfitted	correctly fitted	overfitted	Incorrect	Correct	Average MAPE	Median MAPE
2.0	25	OLS	.075	.346	.570	.075	3.035	1.101	1.010
		SCAD	.092	.428	.477	.092	3.251	1.098	.987
		LAD	.159	.590	.248	.160	3.634	1.091	1.109
		LASSO	.183	.632	.182	.184	3.329	1.098	1.044
		ORACLE	0	1.000	0	0	4.000	1.058	1.083
	50	OLS	.002	.384	.611	.002	3.116	1.044	1.012
		SCAD	.004	.434	.559	.004	3.231	1.042	1.086
		LAD	.025	.773	.199	.025	3.777	1.034	1.036
		LASSO	.034	.797	.166	.034	3.814	1.039	1.058
		ORACLE	0	1.000	0.570	0	4.000	1.026	1.072
	75	OLS	0	.474	.524	0	3.298	1.018	.986
		SCAD	0	.497	.501	0	3.342	1.017	.945
		LAD	0	.913	.085	0	3.906	1.013	.975
		LASSO	0	.926	.072	0	3.921	1.015	1.019
		ORACLE	0	1.000	0	0	4.000	1.011	1.042
0.8	100	OLS	0	.333	.665	0	2.944	.549	.559
		SCAD	.001	.448	.549	.001	3.227	.545	.524
		LAD	.001	.683	.313	.001	3.618	.538	.614
		LASSO	.016	.946	.035	.016	3.960	.548	.500
		ORACLE	0	1.000	0	0	4.000	.527	.536
	125	OLS	0	.403	.596	0	3.130	.521	.468
		SCAD	0	.460	.538	0	3.248	.520	.541
		LAD	0	.837	.161	0	3.818	.515	.495
		LASSO	0	.986	.012	0.	3.986	.519	.495
		ORACLE	0	1.000	0	0	4.000	.513	.512
	150	OLS	0	.428	0.570	0	3.223	.509	.525
		SCAD	0	.454	0.544	0	3.267	.509	.501
		LAD	0	.876	0.122	0	3.867	.506	.501
		LASSO	0	.991	0.007	0	3.991	.506	.506
		ORACLE	0	0	0	0	4.000	.506	.515

Its corresponding LAD-lasso estimator is denoted by $\hat{\beta} = \left(\hat{\beta}_a', \hat{\beta}_b'\right)'$, and the LAD-lasso objective function is denoted by $Q(\beta) = Q(\beta_a, \beta_b)$. In addition, we also decompose the covariate $X_i = (X_{ia}', X_{ib}')'$ with $X_{ia} = (x_{i1}, \ldots, x_{ip0})'$ and $X_{ib} = (x_{i(p0+1)}, \ldots, x_{ip})'$.

1. The error $\in_i$ has continuous and positive density at the origin.
2. The matrix cov(X_1)= $\sum$ exists and is positive definite. Note that Assumptions A and B are both very typical technical assumptions used extensively in the literature (Pollard 1991; Bloomfield and Steiger 1983; Knight 1998). They are needed for establishing the $\sqrt{n}$-consistency and the asymptotic normality of the unpenalized LAD estimator. Furthermore, define

$$a_n = \max\{\lambda_j, 1 \le j \le p_0\}$$

and

$$b_n = \max\{\lambda_j, p_0 < j \le p\}$$

where λj is a function of n. For a detailed study refer to Tibshirani (1996), Zou (2006), Lu and Zhang (2007), Biswas, S. *et.al*.,, Tibshirani (1997), Tsiatis (1996)

RESULTS

We numerically compare the proposed variable selection methods with ordinary least-squares, ridge regression, best subset selection and non-negative garrotte, LASSO and LAD regression. All simulations are conducted using MATLAB codes. As recommended in Breiman (1995), a five-fold cross-validation was used to estimate the tuning parameter for the non-negative garrote. For other model selection procedures, both five-fold cross-validation and generalized cross-validation were used for estimating thresholding parameters. However, their performance are similar. Therefore we only present the results based on the generalized cross validation. This results were shown in Table 1 to Table 2 . Some earlier results also be seen in Tsiatis (1996).

Table 2 Simulation results for t_5 error

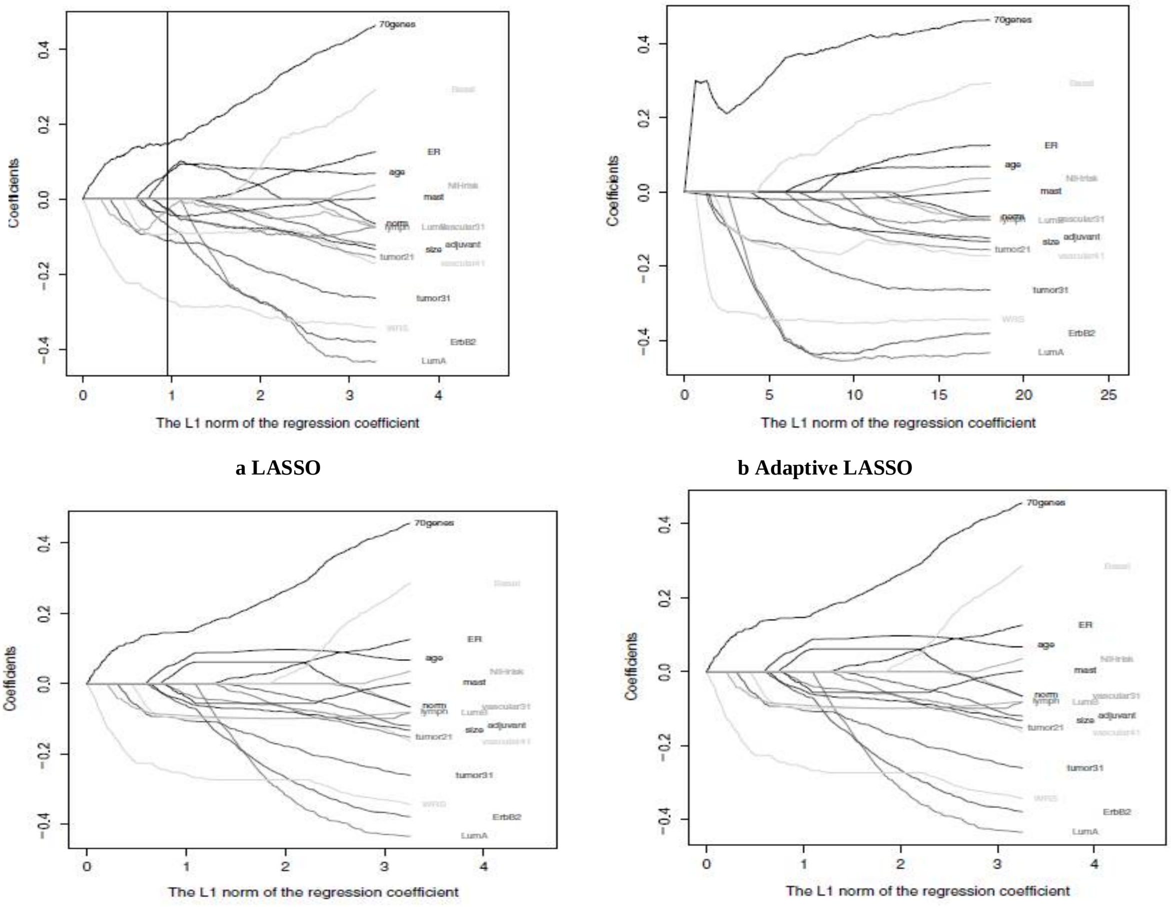
						No. of zeros			
σ	n	Method	Underfitted	correctly fitted	overfitted	Incorrect	Correct	Average MAPE	Median MAPE
3.0	25	OLS	.081	.243	.673	.081	2.824	1.051	1.289
		SCAD	.101	.315	.581	.101	3.071	1.047	1.123
		LAD	.176	.491	.330	.178	3.500	1.041	1.163
		LASSO	.224	.565	.208	.227	3.686	1.047	1.072
		ORACLE	0	1.000	0	0	4.000	1.006	.989
0.7	50	OLS	.008	.315	.674	.008	2.994	.993	1.022
		SCAD	.010	.363	.624	.010	3.104	.993	.947
		LAD	.026	.720	.251	.026	3.706	.984	.987
		LASSO	.041	.786	.170	.041	3.802	.990	.964
		ORACLE	0	1.000	0	0	4.000	.975	.979
	75	OLS	0	.322	.676	0	3.050	.970	1.160
		SCAD	0	.349	.649	0	3.103	.970	.985
		LAD	0	.825	.173	0	3.813	.965	1.019
		LASSO	0	.860	.138	0	3.855	.968	1.058
		ORACLE	0	1.000	0	0	4.000	.962	.906
	100	OLS	0	.286	.712	0	2.903	.521	.524
		SCAD	.001	.404	.593	.001	3.188	.518	.493
		LAD	.001	.659	.338	.001	3.595	.511	.543
		LASSO	.011	.962	.024	.011	3.974	.523	.488
		ORACLE	0	1.000	0	0	4.000	.500	.521
	125	OLS	0	.300	.698	0	2.956	.498	.545
		SCAD	0	.353	.645	0	3.089	.497	.474
		LAD	0	.726	.272	0	3.699	.492	.492
		LASSO	0	.981	.017	0	3.981	.494	.523
		ORACLE	0	1.000	0	0	4.000	.488	.508
	150	OLS	0	.324	0.674	0	2.998	.486	.472
		SCAD	0	.347	0.651	0	3.051	.485	.463
		LAD	0	.799	0.199	0	3.780	.482	.495
		LASSO	0	.986	0.012	0	3.986	.481	.513
		ORACLE	0	0	0	0	4.000	.481	.493

The study involves 295 patients with primary breast carcinomas from the Netherlands Cancer Institute, refer to Chang *et al*., (2005).

The survival time information was extracted from the medical registry of the Netherlands Cancer Institute. Potential clinical predictors include age, tumor size, lymph node status, tumor grade, vascular invasion status, estrogen receptor status

Table 3 Estimated regression coefficients and standard errors of all clinical predictors and gene signatures for the breast cancer data example for LAD VS LASSO

	LASSO		Adaptive LASSO		LAD	
	Clinical predictors					
Age(decades)	0.123	(0.068)	0.000	(-)	0.000	(-)
Tumor size (diameter, cm)	-0.042	(0.024)	-0.056	(0.033)	-0.052	(0.030)
Tumor grade						
Grade 3 versus 1	-0.225	(0.082)	-0.281	(0.205)	-0.271	(0.189)
Grade 2 versus 1	0.000	(-)	0.000	(-)	0.000	(-)
Vascular invasion						
1-3 vessels versus 0 vessel	0.000	(-)	0.000	(-)	0.000	(-)
>3 vessels versus 0 vessel	-0.276	(0.104)	-0.453	(0.047)	-0.450	(0.041)
Estrogen receptor status (positive versus negative)	0.000	(-)	0.000	(-)	0.000	(-)
Mastectomy versus breast Conserving therapy	-0.053	(0.035)	0.000	(-)	0.000	(-)
No adjuvant versus chemo or Hormonal therapy	0.000	(-)	0.000	(-)	0.000	(-)
The number of lymph nodes	-0.013	(0.006)	0.000	(-)	0.000	(-)
NIH risk status (high versus intermediate or low)						
Wound response	-1.493	(0.371)	-1.915	(0.505)	-1.901	(0.480)
70-gene	0.417	(0.119)	0.945	(0.334)	0.940	(0.300)
Normal-like	0.561	(0.297)	0.000	(-)	0.000	(-)
ErbB2	-0.395	(0.151)	-2.161	(0.761)	-2.141	(0.741)
Luminal A	0.000	(-)	-1.478	(0.635)	-1.458	(0.615)
Luminal B	-0.251	(0.109)	0.000	(-)	0.000	(-)
Basal-like	0.000	(-)	0.315	(0.120)	0.305	(0.100)



a LASSO

b Adaptive LASSO

c Approximate LASSO

d LAD

National Institutes of Health (NIH) risk grade, the use of breast conserving therapy, and the use of adjuvant therapy. Available also are gene signatures that represent distinct analytic strategies and have been validated in independent studies. Specifically, there are seven potential gene signatures: basallike, ErbB2, luminal A, luminal B, normal-like, a 70-gene, and m the wound response gene signatures. The basal-like, ErbB2, luminal A, luminal B, and normal-like gene signatures were identified by an unsupervised clustering method refer to Perou *et al*., (2000). The 70-gene signature was constructed based on the association between the gene expression level and the risk of metastasis refer to van de Vijver *et al.,* (2002). The wound response gene signature was a hypothesis-driven signature proposed by Chang *et al*. (2005). We fit the data of the survival time as the response variable and 18 potential predictors: age, tumor size (diameter, cm), the number of lymph nodes, tumor grade (grade 2 versus 1, grade 3 versus 1), vascular invasion (1-3vessels versus 0 vessel, >3 vessels versus 0 vessel), estrogen receptorstatus (positive versus negative), NIH risk status (high versus intermediate or low), the use of breast conserving therapy (mastectomy versus breast conserving therapy), the use of adjuvant therapy (no adjuvant therapy versus chemotherapyor hormonal therapy), as well as the seven gene signatures. All the genetic signatures used in the model are continuous correlation measures. In the analysis, to avoid potential biases we excluded a subset of 61 patients, which was used to construct the 70-gene signature. Among the remaining 234 patients, the median follow-up time was 7.2 years and the number of observed deaths is 55.

To construct prediction models based on these 18 predictors, mwe considered three aforementioned estimators for $\boldsymbol{\beta}$: (i) the standard Gehan estimator; (ii) the LASSO estimator; and (iii) the adaptive LASSO estimator. The entire LASSO and adaptive LASSO, Table 3 and Fig 1 and Fig 2 to Fig 4 regularized paths of the proposed estimators are shown in Figure 2a and b. The observed proportions of *p*-values being smaller than 0.05 are 96.6%, 91.2%, and 82.2% for predictions based on LASSO, adaptive LASSO, and unregularized estimators, respectively. The entire empirical cumulative distribution functions of the 500 *p*-values for comparing the two risk groups identified by the LASSO, adaptive LASSO, and unregularized Gehan's estimators are shown in table 3

CONCLUSION

For comparison purposes, the results of the full model based on the LAD and OLS, Penalized Least Square estimators are also reported. As can be seen, the MAPE of the OLS is as large as .23152. It is substantially worse than all other LAD-based methods, further justifying the use of the LAD methods. Based on a substantially simplified model, the prediction accuracy of the LAD-LASSO estimator remains very satisfactory. According to the reported standard error of the MAPE estimate (STDE), we can clearly see that such a difference cannot be statistically significant. Consequently, we conclude that among all of the LAD-based model selection methods, LAD-LASSO resulted in the simplest model with a satisfactory prediction accuracy. The proposed regularization methods for the AFT model can be easily extended to incorporate other types of penalty functions such as the *L*2 or the more general elastic net regularization (Zou and Hastie, 2005). The entire regularization path with the *L*2 or elastic net penalty would also be piecewise linear and can be obtained by modifying the algorithm proposed by Hastie *et al.* (2004). The Gehan's initial estimator determining the weights used in the adaptive LASSO may be too unstable or even not available for a high-dimensional $\boldsymbol{\beta}$. For such settings, one may instead use the *L*2 regularized Gehan's estimator as the initial estimator. In such cases, the root of the estimating equation may be obtained by an iterative algorithm, in which each iteration amounts to minimizing a weighted Gehan's objective function refer to Jin *et al*., (2003). Therefore, a simple regularization strategy for the general rank-based estimating equation is to apply LASSO or adaptive LASSO regularization within each iteration. However, the resulting regularized solution may lose the simple interpretation as a constrained minimizer. It is important to note that while the proposed procedure may be carried out when *p* increases with the sample size, the asymptotical properties derived in Web Appendices A and B only hold when *p* is a fixed constant. Using similar arguments as given in Huang, Ma, and Zhang (2008), one may extend the results to the setting when $p = pn \to\infty$as $n\to\infty$but at a slower rate. When *p* is much bigger than the sample size, e.g., in the context of gene expression data analysis, operationally, the proposed regularization method can be performed with a large number of individual gene expression as covariates in the regression analysis. However, because the theoretical results require that the dimension of predictor is fixed while the sample size $n \to \infty$, we suggest performing an initial screening step, in which relatively few covariates were selected/constructed from the original gene expression measurements, and then conduct the regularized multivariate analysis with the covariates formed in the first step. Note that even after the initial dimension reduction step, the dimension of predictors may still be not small relative to the sample size for performing the standard unregularized estimation as in the breast carcinomas example and this is where the proposed regularization methods are intended to be applied. The selection of an appropriate penalty parameter is crucial to the performance of regularized estimators. If the primary goal of the regularization is variable selection, i.e., to identify noninformative predictors whose true regression coefficients are zero, one may consider approaches different from optimizing a cross validated loss function. Intuitively, the penalty parameter should be set such that the LASSO estimators for most noninformative predictors are zero. One possible ad hoc approach to achieve this is to first augment existing predictors by several randomly generated noise variables that are independent of the survival time and then calculate the entire LASSO regularization path with the augmented predictors. In the end, one may choose the smallest penalty parameter such that all the LASSO regularized regression coefficients of those augmented noise predictors are zero. LAD gives better results when compared with LASSO based on Survival data.

Reference

Antoniadis, A. (1999). Wavelets in Statistics: A Review. Italian Jour. Statist., to appear.

Bickel, P.J. (1975). One-step Huber estimates in linear models. *Journal of the American Statistical Association*, 70, 428-433.

Biswas, S. Datta, J. Fine, and M. Segal (eds), 1st edition. Hoboken, New Jersey: Wiley.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37, 373-384.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math., 31, 377 - 403.

Chang, H., Nuyten, D., Sneddon, J., Hastie, T., Tibshirani, R., Sorlie,T., Dai, H., He, Y., van't Veer, L., Bartelink, H., van de Rijn, M.,Brown, P., and van de Vijver, M. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *PNAS* 102, 3738-3743.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425 - 455.

Donoho, D. L., Johnson, I.M., Hock, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object (with discussion). *Journal of Royal Statistical Society*, B, 54, 41-81.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. T. (2004). Least

angle regression. *The Annals of Statistics* 32, 407-499.

Fan, J. (1999). Comments on \Wavelets in statistics: a review" by A. Antoniadis. *Journal of Italian Statistical Association*, To appear.

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.

Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.

Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, 7, 397-416.

Gao, H. Y. and Bruce, A. G. (1997). WaveShrink with _rm Shrinkage. *Statistica Sinica*, 7, 855-874.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research* 5, 1391-1415.

Huang, J. and Harrington, D. (2005). Iterative partial least squares with right-censored data analysis: A comparison to other dimension reduction techniques. *Biometrics* 61, 17-24.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics* 9, 1276-1288.

Lu, W. and Zhang, H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* 26, 3771-3781.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. 2nd ed. Chapman and Hall, London.

Perou, C., Sorlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen, H., Akslen, L., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S., Lonning, P., Borresen- Dale, A., Brown, P., and Botstein, D. (2000). Modecular portraits of human breast tumours. *Nature* 6797, 747-752.

Robinson, P.M. (1988), The stochastic di_erence between econometric and statistics, *Economet- rica*, 56, 531-547.

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

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine* 16, 385-395.

Tsiatis, A. (1996). Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics* 18, 305-328.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* 101, 1418-1429.

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

How to cite this article:

Sathish Kumar S and Elangovan R.2017, Variable Selection for Survival Analysis Using Lasso Vs Lad Regression. *Int J Recent Sci Res.* 8(10), pp. 20696-20704. DOI: http://dx.doi.org/10.24327/ijrsr.2017.0810.0944

****