Title
Variable Selection via Penalized Likelihood

Authors
Fan, Jianqing
Liu, Runze

Publication Date
1999-06-14

Variable Selection via Penalized Likelihood*

JIANQING FAN
Department of Statistics
University of California
Los Angeles, CA 90095

RUNZE LI
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

June 14, 1999

Abstract

Variable selection is vital to statistical data analyses. Many of procedures in use are ad hoc stepwise selection procedures, which are computationally expensive and ignore stochastic errors in the variable selection process of previous steps. An automatic and simultaneous variable selection procedure can be obtained by using a penalized likelihood method. In traditional linear models, the best subset selection and stepwise deletion methods coincide with a penalized least-squares method when design matrices are orthonormal. In this paper, we propose a few new approaches to selecting variables for linear models, robust regression models and generalized linear models based on a penalized likelihood approach. A family of thresholding functions are proposed. The LASSO proposed by Tibshirani (1996) is a member of the penalized least-squares with the L_1-penalty. A smoothly clipped absolute deviation (SCAD) penalty function is introduced to ameliorate the properties of L_1-penalty. A unified algorithm is introduced, which is backed up by statistical theory. The new approaches are compared with the ordinary least-squares methods, the garrote method by Breiman (1995) and the LASSO method by Tibshirani (1996). Our simulation results show that the newly proposed methods compare favorably with other approaches as an automatic variable selection technique. Because of simultaneous selection of variables and estimation of parameters, we are able to give a simple estimated standard error formula, which is tested to be accurate enough for practical applications. Two real data examples illustrate the versatility and effectiveness of the proposed approaches.

Key Words: Hard thresholding, LASSO, nonnegative garrote, SCAD, soft thresholding.

Fan's research was partially supported by NSF grant DMS-9804414 and a grant from University of California at Los Angeles.

1 Introduction

Consider the usual linear regression model

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

where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. To attenuate possible excessive modeling biases, a large number of predictors are usually 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).

There are strong connections between penalized least-squares method and variable selection in linear regression models. When design matrices are orthonormal, the stepwise backward deletion and the best subset selection methods are equivalent to a hard-thresholding rule. The latter can be regarded as a solution to a penalized least-squares problem, as shown in Section 2. Figures 1 (a) and (b) show that the hard-thresholding rule sets small coefficients to 0 and keeps large coefficients intact, and the soft-thresholding rule sets small coefficients 0 and shrinks the estimate by a constant. Thus, the hard-thresholding rule results in an unstable model in the sense that a small change of data can lead to a very different model. This can create excessive variabilities in prediction. On the other hand, while the soft-thresholding rule is continuous, it always shifts an estimate by a constant. This would cause lot of biases if the thresholding parameter is large. In the same spirit of Bruce and Gao (1997), Fan (1999) outlines a few thresholding rules which aim at improving the properties of both the hard and soft thresholding rules. These new rules can also be regarded as penalized least-squares. In particular, a smoothly clipped absolute deviation (SCAD) penalty function is proposed to improve the L_1 and the hard-thresholding penalty functions.

In this paper, we propose a few new approaches to selecting variables for various linear regression models based on a penalized likelihood approach in various statistical contexts. A few new penalization functions are introduced. A unified algorithm is proposed to handle the situations when penalized functions are not smooth enough. This yields a unified variable selection procedure. A standard error formula for estimated coefficients is obtained by using a sandwich formula, via the proposed iterative algorithm. The formula is tested accurately enough for practical purpose, even though the sample size is very moderate. The proposed procedures are compared with various other variable selection approaches. The results indicate favorable performance of the newly proposed procedures.

In Section 2, we discuss the relationship between thresholding rules and subset selection when design matrices are orthonormal. We then in Section 3 extend the penalized likelihood approach discussed in Section 2 to various linear regression models, including traditional linear regression models, robust linear regression models and generalized linear models. Based on local quadratic approximations, a unified iterative algorithm for finding penalized likelihood estimators is proposed at the end of Section 3. The formulas for covariance matrices of estimated coefficients are also derived in this section. We illustrate our proposed approaches by two real data examples in Section 4. Two data-driven methods for finding unknown thresholding parameters are discussed in Section 5. Numerical comparisons and simulation studies are also given in this section. Finally some discussion is given in Section 6.

2 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 $\mathbf{X}$ in (1.1) are orthonormal. Then the least-squares estimate in the full model is $\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$, a part of the orthogonal transform of the vector $\mathbf{y}$.

2.1 Thresholding and variable selection

The backward stepwise deletion algorithm in the linear models is to delete a variable, one at a time, with the smallest absolute t-value. For the orthonormal design matrix, this corresponds to delete the variable with the smallest absolute value of estimated coefficients. When a variable is deleted,

the remaining columns of design matrix $\mathbf{X}$ are still orthonormal and the estimated coefficients remain unchanged. So in the second step, the algorithm deletes the variable that has the second smallest estimated coefficient in the full model. If the stepwise backward deletion is carried out m times, the remaining variables are those with the largest $n-m$ values of $|\widehat{\boldsymbol{\beta}}|$. This is equivalent to using a hard thresholding rule with a thresholding parameter between the m^{th} and $(m+1)^{th}$ order statistics of $|\widehat{\boldsymbol{\beta}}|$.

The soft-thresholding rule can be viewed similarly. Denote by $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ and assume that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ in model (1.1). Then $\mathbf{z}$ is a multivariate normal random vector with independent components. This allows us to consider a Gaussian white noise model:

$$z_i = \theta_i + \varepsilon_i \quad \text{with } \varepsilon_i \sim N(0, \sigma^2) \quad \text{for } \ i = 1, \cdots, d. \tag{2.1}$$

Suppose that the θ's in (2.1) are sparse so that they can reasonably be modeled 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

$$\frac{1}{2}\sum_{i=1}^{d}(z_i - \theta_i)^2 + \lambda\sum_{i=1}^{d}|\theta_i|, \tag{2.2}$$

where $\lambda = \sigma^2/\lambda_1$.

Minimization of (2.2) is equivalent to minimizing (2.2) component-wise. The solution to the above problem yields the soft-thresholding rule (Figure 1(b))

$$\widehat{\theta}_j = \text{sgn}(z_j)(|z_j| - \lambda)_+. \tag{2.3}$$

This connection was observed by Donoho, Johnstone, Hoch and Stern (1992) and formed the core of the LASSO method introduced by Tibshirani (1996). If the L_1-penalty in (2.2) 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.

2.2 Penalized least-squares and variable selection

Consider a general form of penalized least-squares:

$$\frac{1}{2}\sum_{j=1}^{d}(z_j - \theta_j)^2 + \lambda\sum_{j=1}^{d}p_j(|\theta_j|). \tag{2.4}$$

The penalty functions $p_j(\theta)$ in (2.4) 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 (2.4) is equivalent to minimizing componentwise the penalized least-squares problem:

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|). \tag{2.5}$$

The solution to (2.5) 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 $|z|$. See Figure 2. The solution is continuous in $|z|$ only when the minimum of $\theta + \lambda p'(\theta)$ over $\theta \geq 0$ is attained at 0, as shown in Figure 2. When $p_\lambda(|\theta|) = \lambda|\theta|^q$ as in the bridge regression (Frank and Friedman, 1993), the solution is continuous only when $q \geq 1$. However, when $q > 1$, the minimum of $\theta + p'_\lambda(\theta)$ is zero and hence it does not correspond to a thresholding rule. The only continuous solution with a thresholding in this family is the L_1 penalty, but this comes at a price of shifting the resulting estimator by a constant λ (see Figure 1 (b)).

In the discussion of Antoniadis (1999), Fan observed that the penalized least-squares estimator with the penalty function $p(|\theta|) = |\theta| I(|\theta| \leq \lambda) + \lambda/2 I(|\theta| > \lambda)$ leads to the hard-thresholding rule

$$\widehat{\theta} = zI(|z| > \lambda). \tag{2.6}$$

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

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda). \tag{2.7}$$

With the clipped L_1-penalty function

$$p_\lambda(\theta) = \lambda \min(|\theta|, \lambda) \tag{2.8}$$

the solution is a mixture of soft and hard thresholding rule (Figure 1(c)):

$$\widehat{\theta} = \mathrm{sgn}(z)(|z| - \lambda)_+ I(|z| \leq 1.5\lambda) + zI(|z| > 1.5\lambda). \tag{2.9}$$

2.3 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

$$p'(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \text{ for some } a > 2 \quad \text{and } \theta > 0, \tag{2.10}$$

improves the properties of the L_1-penalty and the hard-thresholding penalty function given by (2.7) (see Figure 4 and discussion below) . We will call this penalty function as smoothly clipped absolute deviation (SCAD) penalty. This corresponds to a quadratic spline 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

$$\widehat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda; \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda; \\ z, & \text{when } \ |z| > a\lambda. \end{cases} \tag{2.11}$$

See Figure 1(d). This solution is due to Fan (1999). For simplicity of presentation, we will call all procedures using the SCAD penalty as SCAD.

The thresholding rule in (2.8) involves two unknown parameters λ and a. In practice, we could search the best pair (λ, a) over two dimensional grids using some criteria, such as cross-validation and generalized cross-validation (Craven and Wahba, 1977). Such an implementation can be computationally expensive. Motivated by the soft-thresholding, we assume that for given a and λ, the prior for θ is a normal distribution with zero mean and variance $a\lambda$. We computed the Bayesian risk via numerical integration. Figure 3(a) depicts the Bayesian risk as a function of a under the squared loss, for the universal thresholding $\lambda = \sqrt{2\log(d)}$ (see Donoho and Johnstone, 1994) with $d = 20,\ 40,\ 60$ and 100, and Figure 3(b) is for $d = 512,\ 1024,\ 2048$ and 4096 . From Figures 3(a) and 3(b), the Bayesian risks achieve their minimums when $a \approx 3.7$. It can be seen from these two figures that the Bayesian risks are not very sensible with the values of a. This choice gives pretty good practical performance for various variable selection problems. Indeed, from the simulations in Section 5.3, the choice of $a = 3.7$ works similarly to that chosen by the GCV method.

2.4 Performance of thresholding rules

To gauge the performance of the four thresholding rules, Figure 3(c) depicts their L_2 risk functions $R(\widehat{\theta}, \theta) = E_\theta(\widehat{\theta} - \theta)^2$ under the Gaussian model $Z \sim N(\theta, 1)$. To make the scale of thresholding parameters roughly comparable, we took $\lambda = 2$ for the hard thresholding rule, and adjusted the values of λ for other thresholding rules so that their estimated values are the same when $z = 3$. The SCAD performs favorably comparing with the other three rules. This can also be understood via their corresponding penalty functions plotted in Figure 4. It is clear that the SCAD retains good mathematical properties of the other three thresholding penalty functions. Hence, it is expected to perform the best.

3 Variable selection via penalized likelihood

The methodology in the previous section can be directly applied to many other statistical contexts. In this section we consider general linear regression models, robust linear models and likelihood based generalized linear models. From now on, we assume that the design matrix $\mathbf{X} = (x_{ij})$ is standardized so that each column has mean zero and variance one.

3.1 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 (2.4) can be naturally extended to the situation in which design matrices are not orthonormal. Similar to (2.4), a general form of penalized least-squares is

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

or equivalently

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.1}$$

Minimizing (3.1) with respect to $\boldsymbol{\beta}$ leads to a penalized least-squares estimator of $\boldsymbol{\beta}$.

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 (see Huber

(1981)). Therefore instead of minimizing (3.1), we minimize

$$\sum_{i=1}^{n} \psi(|y_i - \mathbf{x}_i\boldsymbol{\beta}|) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.2}$$

with respect to $\boldsymbol{\beta}$. This results in a robust least-squares estimator.

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

$$-\sum_{i=1}^{n} \ell_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.3}$$

To obtain a penalized maximum likelihood estimator of $\boldsymbol{\beta}$, we minimize (3.3) with respect to $\boldsymbol{\beta}$ for some thresholding parameter λ.

3.2 A unified algorithm

Finding solutions for minimization problems in (3.1), (3.2) and (3.3) is not an easy task. Tibshirani (1996) proposed algorithm of solving constrained least-squares problems for LASSO, while Fu (1998) provided a shooting algorithm for LASSO. We in this section propose a unified algorithm for minimization problems (3.1), (3.2) and (3.3) via local quadratic approximations. The first term in (3.1), (3.2) and (3.3) may be regarded as a loss function of $\boldsymbol{\beta}$. Denote it by $\ell(\boldsymbol{\beta})$. Then, the expressions (3.1), (3.2) and (3.3) can be written in a unified form as

$$\ell(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{3.4}$$

The clipped L_1 penalty $p_\lambda(x)$ in (2.8) is not differentiable. However it can be locally approximated by a quadratic function as follows. Suppose that we are given an initial value $\boldsymbol{\beta}_0$ that is close to the minimizer of (3.4). Then the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by $\{p_\lambda(|\beta_{j0}|)/\beta_{j0}^2\}\beta_j^2$ for $\beta_j \approx \beta_{j0}$ when β_{j0} is not very close to 0, otherwise, set $\widehat{\beta}_j = 0$ (see Figure 4(c)). When $p_\lambda(|\beta_j|)$ is differentiable except at the point zero, it can be locally approximated by the quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j,$$

when $\beta_j > 0$. In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}p'_\lambda(|\beta_{j0}|)(\beta_j^2 - \beta_{j0}^2), \qquad \text{for } \beta_j \approx \beta_{j0}.$$

In the algorithm below, we always use the second type of approximation whenever $p_\lambda(|\beta_j|)$ has the derivative function except at point 0. Figure 4 shows the above two approximations for a few different values of β_{j0}. A drawback of this approximation is that once a coefficient is shrunk to zero, it will retain zero. However, this method reduces significantly computational burden.

If $\ell(\boldsymbol{\beta})$ is L_1 loss as used in (3.2), then it does not have continuous second order partial derivatives with respect to $\boldsymbol{\beta}$. However, $\psi(|y - \mathbf{x}^T\boldsymbol{\beta}|)$ in (3.2) can be analogously approximated by $\{\psi(y - \mathbf{x}^T\boldsymbol{\beta}_0)/(y - \mathbf{x}^T\boldsymbol{\beta}_0)^2\}(y - \mathbf{x}^T\boldsymbol{\beta})^2$, as long as the initial value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$ is close to the minimizer. When the some of the residuals $|y - \mathbf{x}^T\boldsymbol{\beta}_0|$ are small, this approximation is not very good. See Section 3.3 for some slight modification of this approximation.

Now assume that the log-likelihood function is smooth with respect to $\boldsymbol{\beta}$ so that its first two partial derivatives are continuous. Thus the first terms in (3.4) can be locally approximated by a quadratic function. Therefore the minimization problem (3.4) can be reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. In particularly, when $p'_\lambda(|\beta|)$ has the first derivative except at the point 0, (3.4) can be locally approximated (except a constant term) by

$$\ell(\boldsymbol{\beta}_0) + \nabla\ell(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\nabla^2\ell(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \tag{3.5}$$

where

$$\nabla\ell(\boldsymbol{\beta}_0) = \frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}, \quad \nabla^2\ell(\boldsymbol{\beta}_0) = \frac{\partial^2\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}, \quad \Sigma_\lambda(\boldsymbol{\beta}_0) = \text{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \cdots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}.$$

The quadratic minimization problem (3.5) yields the solution

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \{\nabla^2\ell(\boldsymbol{\beta}_0) + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\{\nabla\ell(\boldsymbol{\beta}_0) + n\mathbf{U}_\lambda(\boldsymbol{\beta}_0)\}, \tag{3.6}$$

where $\mathbf{U}_\lambda(\boldsymbol{\beta}_0) = \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0$. When the algorithm converges, and the second type of approximation is used, the estimator satisfies the condition

$$\frac{\partial\ell(\widehat{\boldsymbol{\beta}}_0)}{\partial\beta_j} + np'_\lambda(|\widehat{\beta}_{j0}|)\text{sgn}(\widehat{\beta}_{j0}) = 0,$$

the penalized likelihood equation, for non-zero elements of $\widehat{\boldsymbol{\beta}}_0$. Specifically, for the penalized least-squares problem (3.1), the solution can be found by iteratively using the following ridge regression:

$$\boldsymbol{\beta}_1 = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Similarly we obtain the solution for (3.2) by iterating

$$\boldsymbol{\beta}_1 = \{\mathbf{X}^T\mathbf{W}\mathbf{X} + \frac{1}{2}n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

where $\mathbf{W} = \text{diag}\{\psi(|y_1 - \mathbf{x}_1^T\boldsymbol{\beta}_0|)/(y_1 - \mathbf{x}_1^T\boldsymbol{\beta}_0)^2, \cdots, \psi(|y_n - \mathbf{x}_n^T\boldsymbol{\beta}_0|)/(y_n - \mathbf{x}_n^T\boldsymbol{\beta}_0)^2\}$.

Like in the maximum likelihood estimation (MLE) setting, with good initial value $\boldsymbol{\beta}_0$, the one-step procedure can be as efficient as the fully iterative procedure, namely, the penalized maximum likelihood estimator, when one uses the Newton-Raphson algorithm (See Bickel (1975)). Now regarding $\boldsymbol{\beta}^{(k-1)}$ as a good initial value at the k-th step, the next iteration can also be regarded as a one-step procedure and hence the resulting estimator can still be as efficient as the fully iterative method. See Robinson (1988) for theory on the difference between the MLE and k-step estimators. Therefore estimators obtained by the aforementioned algorithm after a few iterations can always be regarded as a one-step estimator, which is as efficient as the fully iterative method. In this sense, one does not have to iterate the algorithm above until it converges as long as the initial estimators are good enough. The estimators from the full models can be used as initial estimators, as long as they are not excessively overly parameterized.

3.3 Standard formula

The standard errors for estimated parameters can be directly obtained because we are estimating parameters and selecting variables at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula in (3.6) can be used as an estimator for the conditional covariance of the estimates $\widehat{\boldsymbol{\beta}}$, conditioning on $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. That is,

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}) = \{\nabla^2\ell(\widehat{\boldsymbol{\beta}}) + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\widehat{\text{cov}}\{\nabla\ell(\widehat{\boldsymbol{\beta}})\}\{\nabla^2\ell(\widehat{\boldsymbol{\beta}}) + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}. \tag{3.7}$$

This formula is tested to have good accuracy for moderate sample sizes.

When the L_1-loss is used in the robust regression, some slight modifications are needed in the aforementioned algorithm and its corresponding sandwitch formula. For $\psi(x) = |x|$, the diagonal elements of $\mathbf{W}$ are $\{|r_i|^{-1}\}$ with $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0$. Thus, for a given current value of $\boldsymbol{\beta}_0$, when some

of residuals $\{r_i\}$ are close to 0, these points receive too much weights. Hence, we replace the weight by $(a_n + |r_i|)^{-1}$. In our implementations, we took a_n as $2n^{-1/2}$ quantile of the absolute residuals $\{|r_i|, i = 1, \cdots, n\}$. Thus, the constant a_n is changing from iteration to iteration.

3.4 Testing convergence of the algorithm

We now demonstrate that our algorithm converges to the right solution. To this end, we took a 100 dimensional vector $\boldsymbol{\beta}$ consisting of 50 zeros and other nonzero elements being generated from $N(0, 5^2)$ and used a 100×100 orthonormal design matrix $\mathbf{X}$. We then generated a response vector $\mathbf{y}$ from the linear model (1.1). We chose an orthonormal design matrix for our testing case, because the penalized least-squares has a closed form mathematical solution so that we can compare our output with the mathematical solution. Our experiment did show that the proposed algorithm converged to the right solution. It took MATLAB codes 0.27, 0.39, 0.16 seconds for the penalized least-squares with SCAD, L_1 and hard-thresholding penalties to converge. The numbers of iterations are respectively 30, 30 and 5 for the penalized least-squares with the SCAD, L_1 and the hard-thresholding penalty. In fact, after 10 iterations, the penalized least-squares estimators are already very close to the true one.

4 Examples

In this section we consider two real data examples. We firstly apply the proposed penalized least-squares approaches to an environmental data set, collected in Hong Kong from January 1, 1994 to December 31, 1995 (Courtesy of Professor T. S. Lau), which consists of daily measurements of pollutants and other environmental factors. We also apply the proposed penalized likelihood method to analyze another data set: *Burns data*, collected by General Hospital Burn Center at the University of Southern California.

Example 4.1. For the environmental data set, one is interested in studying the association between levels of pollutants and the number of daily total hospital admissions for respiractionary problems on every Monday from January 1, 1994 to December 31, 1995. In this data set we used 114 observations and took the response variable Y as the number of daily total admissions. Seven covariates were considered: three binary seasonal covariates: Spring S_1, Summer S_2 and Autumn

S_3, and four continuous covariates for daily measurements of pollutants and time: the level of pollutants Sulfur Dioxide X_1, Nitrogen Dioxide X_2, and dust X_3 and time T. We then included all quadratic terms and interaction terms of the four continuous covariates. This gives a total of 17 predictor variables. Before using the penalized least-squares method, the response variable Y and all 17 covariates were standardized individually. We fitted a linear regression model without intercept as both response and predictors were normalized. The thresholding parameter λ is chosen by generalized cross-validation (see Section 5.2 below). They are 0.0785, 0.0157 and 0.0157 for the penalized least-squares with the SCAD, L_1 (LASSO) and hard thresholding penalties, respectively. The value of a in the SCAD was set to be 3.7. To find the best subset, we searched exhaustively over all possible subsets and selected the subset with best BIC score. The computational time for each of the three penalized least-squares including searching the unknown parameter λ over 15 grids via generalized cross-validation was less than 2 seconds; while it spent more than 4 minutes in searching for the best subset. With the selected λ, the penalized least-squares estimators were obtained at the 17^{th}, 9^{th} and 7^{th} step of iterations for the SCAD, LASSO and hard thresholding, respectively. We also computed the five-step estimators, it took us less than one second, yet the differences between the full iteration estimator and five-step estimator were less than one percent. Estimated coefficients and standard errors for the transformed data are presented in Table 1.

From Table 1, the performance of the SCAD is very good. Comparing with the best subset selection, the SCAD included terms X_3, X_2^2 and X_2X_3 rather than X_2. These three predictors X_3, X_2^2 and X_2X_3 are quite significant in the least-squares estimate, while X_2 is not. Furthermore X_2 is also deleted by LASSO, and X_3 and X_2^2 are selected by LASSO. Comparing with LASSO, the SCAD selected variables X_2X_3 instead of S_3 excluded also by the best subset selection. The estimated coefficients of T and T^2 were shrunk too much by the LASSO. See Example 5.3 for further remark on this. Compared with the penalized least-squares with the hard thresholding penalty, the SCAD excluded terms S_3 and TX_3 that are not statistically significant.

From Table 1, we may exclude the predictor S_3 because the weather in Hong Kong is not significantly different between Autumn and Winter. All interactions between time and pollutant factors are not significant. All predictors related with Sulfur Dioxide are not statistically significant. However, it should be noted that time and Sulfur Dioxide pollutant may have interaction. Both *Nitrogen Dioxide* and *dust* are important factors to the number of daily total admissions.

Example 4.2 We in this example apply the proposed penalized likelihood methodology to the *Burns data*. The data set consists of 981 observations. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise. Covariates $X_1 = age$, $X_2 = sex$, X_3 =log(burn area +1) and binary variable $X_4 = Oxygen$ (0 normal, 1 abnormal) were considered. Quadratic terms of X_1 and X_3, and all interaction terms were included. The intercept term was added and the logistic regression model was fitted. The unknown parameter λ was chosen by generalized cross-validation. They are 0.6932, 0.0015 and 0.8062 for the penalized likelihood estimates with the SCAD, L_1 and hard-thresholding penalties respectively. The constant a in the SCAD was taken as 3.7. It took about 2 minutes for finding each of the three penalized likelihood including searching for the thresholding parameter λ via generalized cross-validation, while it spent more than five hours in searching the best subset! With the selected λ, the penalized likelihood estimator was obtained at the 6^{th}, 28^{th} and 5^{th} step iterations for the penalized likelihood with SCAD, L_1 and hard-thresholding penalties, respectively. We also computed ten-step estimators it took us less than 50 seconds for each penalized likelihood estimator, and the differences between the full iteration estimators and the ten-step estimators are less than one percent. The estimated coefficients and standard errors for the transformed data, based on the penalized likelihood estimators, are reported in Table 2.

From Table 2, the best subset procedure chooses 5 out of 13 covariates, while the SCAD chooses 4 covariates. The difference between them is that the best subset keeps X_4. LASSO chooses the quadratic term of X_1 and X_3 rather than their linear terms. It also selects an interaction term X_2X_3, which may not be statistically significant. It seems again that LASSO shrinks coefficients noticeably large. In this example, the penalized likelihood with the hard thresholding penalty retains too many predictors. Particularly, it selects variables X_2 and X_2X_3. This may not very reasonable as *gender* should not play an important role in determining the survival probability of a victim.

5 Simulations

5.1 Prediction and model error

The prediction error is defined as the average error in prediction Y given $\mathbf{x}$ for future cases not used in the construction of a prediction equation. There are two regression situations, *X-random*

and *X-controlled*. In the case that X is random, both Y and $\mathbf{x}$ are randomly selected. In the controlled situation, design matrices are selected by experimenters and only y is random. For ease of presentation, we consider only the X-random case.

In *X-random* situations, the data $(\mathbf{x}_i, Y_i)$ are assumed a random sample from its parent distribution $(\mathbf{x}, Y)$. Then, if $\widehat{\mu}(\mathbf{x})$ is a prediction procedure constructed using the present data, the prediction error is defined as

$$\mathrm{PE}(\widehat{\mu}) = E\{Y - \widehat{\mu}(\mathbf{x})\}^2,$$

where the expectation is only taken with respect to the new observation $(\mathbf{x}, Y)$. The predictor error can be decomposed as

$$\mathrm{PE}(\widehat{\mu}) = E\{Y - E(Y|\mathbf{x})\}^2 + E\{E(Y|\mathbf{x}) - \widehat{\mu}(\mathbf{x})\}^2.$$

The first component is inherent due to stochastic errors. The second component is due to lack of fit to an underlying model. This component is called *model error* and is denoted by $\mathrm{ME}(\widehat{\mu})$. The size of the model error reflects performances of different model selection procedures. If $Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$, where $E(\varepsilon|\mathbf{x}) = 0$, then $\mathrm{ME}(\widehat{\mu}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

5.2 Selection of thresholding parameters

To implement the methods described in Sections 2 and 3, we need to estimate the thresholding parameters λ and a. Denote by $\boldsymbol{\theta}$ the tuning parameters to be estimated, i.e., $\boldsymbol{\theta} = (\lambda, a)$ for the SCAD, while $\boldsymbol{\theta} = \lambda$ for other thresholdings. Here we discuss two methods of estimating $\boldsymbol{\theta}$: 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 $\ell\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ by the first term in (3.4) replacing $\boldsymbol{\beta}$ by its estimate $\widehat{\boldsymbol{\beta}}$ obtained when the tuning parameters $\boldsymbol{\theta}$ are used. Then $\ell\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ 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 $T - T^\nu$ and T^ν, for $\nu = 1, \cdots, 5$. For each $\boldsymbol{\theta}$ and ν, we find the estimator $\widehat{\boldsymbol{\beta}}^{(\nu)}(\boldsymbol{\theta})$ of $\boldsymbol{\beta}$ using the training set $T - T^\nu$. Let $\ell_\nu\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ be the $\ell\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ for test set T^ν. Form the cross-validation criterion as

$$\mathrm{CV}(\boldsymbol{\theta}) = \sum_{\nu=1}^{5} \ell_\nu\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}.$$

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

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

$$\boldsymbol{\beta}_1(\boldsymbol{\theta}) = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Thus the fitted value $\widehat{\mathbf{y}}$ of $\mathbf{y}$ is $\mathbf{X}^T\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}$, and

$$\mathbf{P}_\mathbf{X}\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\} = \mathbf{X}^T\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}^T$$

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

$$\text{GCV}(\boldsymbol{\theta}) = \frac{1}{n}\frac{\ell\{\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}}{\{1 - e(\boldsymbol{\theta})/n\}^2}$$

and $\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\{\text{GCV}(\boldsymbol{\theta})\}$. Similarly the corresponding generalized cross-validation statistics can be defined for robust regression models and likelihood based linear models.

5.3 Simulation study

In the following examples, we numerically compare the proposed variable selection methods with ordinary least-squares, ridge regression, best subset selection and non-negative garrote (see Breiman (1995)). All simulations are conducted using MATLAB codes. We directly used the constraint least-squares module in MATLAB for finding non-negative garrote estimate. 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.

Example 5.1.(Linear regression)

In this example we simulated 100 data sets consisting of n observations from the model (Tibshirani, 1996)

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the components of $\mathbf{x}$ and ε are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. This is a model used in Tibshirani (1996). Firstly

we chose $n = 40$ and $\sigma = 3$. Then we reduced σ to 1 and increased the sample size to 60. The model error of the proposed procedures are compared to that of the least-squares estimator. The Median of Relative Model Errors (MRME) over 100 simulated data sets are summarized in Table 3. The top panel of Figure 5 depicts the boxplots of the relative model errors.

From Figure 5 and Table 3, it can be seen that when noise level is high and sample size is small, LASSO performs the best and it significantly reduces both model error and model complexity; while ridge regression only reduces model error. The other variable selection procedures also reduce model error and model complexity. However, when the noise level is reduced, the SCAD outperforms the LASSO and other penalized least-squares. Ridge regression performs very poorly. The best subset selection method performs quite similarly to the SCAD. The nonnegative garrote performs quite well in various situations. Comparing with the first two rows in Table 3, one can see that the choice of $a = 3.7$ is very reasonable. Therefore we used it for other examples in this paper.

We now test the accuracy of our standard error formula (3.7). The median of absolute deviation divided by 0.6745, denoted by SD in Table 4, of 100 estimated coefficients in the 100 simulations can be regarded as the true standard error. The median of the 100 estimated SDs, denoted by SD_m, and the median of absolute deviation error of 100 estimated standard errors divided by 0.6745, denoted by SD_{mad}, gauge the overall performance of the standard error formula (3.7). Table 4 presents only the results for non-zero coefficients when the sample size $n = 60$. The results for the other two cases with $n = 40$ are similar. Table 4 suggests that the sandwich formula performs surprisingly well.

Example 5.2. (Robust regression)

In this example, we simulated 100 data sets consisting of 60 observations from the model

$$Y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta}$ and $\mathbf{x}$ are the same as those in Example 1. The ε is drawn from the standard normal distribution with 10% outliers from the standard Cauchy distribution. The simulation results are summarized in Table 5. Figure 5(d) presents the boxplots of the relative model errors. From Table 5, it can be seen that the SCAD outperforms somewhat other procedures. The true and estimated standard deviations of estimators via sandwich formula (3.7) are shown in Table 6. It indicates that the performance of the sandwich formula is very good.

Example 5.3. (Logistic regression)

In this example, we simulated 100 data sets consisting of 200 observations from the model $Y \sim \text{Bernoulli}(p(\mathbf{x}^T\boldsymbol{\beta}))$, where $p(u) = \exp(u)/(1+\exp(u))$, and the first six components of $\mathbf{x}$ and $\boldsymbol{\beta}$ are the same as those in Example 1. The last two components of $\mathbf{x}$ are independently identically distributed as a Bernoulli distribution with probability of success 0.5. All covariates are standardized. Model errors are computed via 1000 Monte Carlo simulations. The summary of simulation results is depicted in Tables 7 and 8. Figure 5(e) shows the boxplots of the relative model errors. From Table 7, it can be seen that the performance of the SCAD is much better than other two penalized likelihood estimates. From Figure 5(e), the variations of the relative model errors of the four procedures are almost same. It can be seen from Table 8 that our standard error estimator works well.

We would like to remark that the estimated SDs for L_1-penalized likelihood estimator (LASSO) are consistently smaller than the SCAD and the penalized likelihood method with the hard-thresholding procedure, yet its overall MRME is larger than that of the SCAD. This implies that the biases in the L_1-penalized likelihood estimators are large. This remark applies to all of our examples. Indeed, it can be seen from Table 2 that all coefficients were shrunk noticeably large by LASSO.

Example 5.4. (Poisson log-linear regression)

In this example, we simulated 100 data sets consisting of 60 observations from the model $Y \sim \text{Poisson}\{\lambda(\mathbf{x}^T\boldsymbol{\beta})\}$, where $\lambda(u) = \exp(u)$, $\mathbf{x}$ is the same as that in Example 5.3, and $\boldsymbol{\beta} = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$. Model errors were obtained by 1000 Monte Carlo simulations. Tables 9 and 10 show the simulation results. Figure 5(f) depicts the boxplots of the relative model errors. From Figure 5(f), the variations of the relative model errors of the four procedures are almost same. In terms of model errors, the performance of the best subset selection method and the SCAD are much better than the other two. Table 10 shows that the standard error estimator works very reasonably.

6 Discussions

We propose a variable selection method via penalized likelihood approaches. A family of penalty functions are introduced. The methods are shown to be effective and the standard errors are estimated with good accuracy. A unified algorithm is proposed for minimizing penalized likelihood function, which is usually a sum of convex and concave functions. Our algorithm is backed up by statistical theory and hence gives estimators with good statistical properties. Comparing with the best subset method, which is very time consuming, the newly proposed methods are much faster, more effective and have strong theoretical backup. They select variables simultaneously via optimizing a penalized likelihood and hence the standard errors of estimated parameters can be estimated accurately. The LASSO proposed by Tibshirani (1996) is a member of this penalized likelihood family with L_1-penalty. It has good performance when noise to signal ratio is large, but the bias created by this approach is noticeably large. See also the remarks in Example 5.3. The newly proposed penalty function, called Smoothly Clipped Absolute Deviation (SCAD) penalty function, gives the best performance in selecting significant variables without creating excessive biases. The approach proposed here can be applied to other statistical contexts without any extra difficulties.

References

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.

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.

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.

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

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 firm Shrinkage. *Statistica Sinica*, **7**, 855-874.

Huber, P. (1981). *Robust estimation*, Wiley, New York.

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

Robinson, P.M. (1988), The stochastic difference between econometric and statistics, *Econometrica*, **56**, 531-547.

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

Table 1. Estimated coefficients and standard errors for Example 4.1

Method	LS	best subset	SCAD	LASSO	hard
S_1	0.39 (0.03)	0.40 (0.03)	0.35 (0.09)	0.35 (0.09)	0.34 (0.10)
S_2	0.14 (0.04)	0.23 (0.03)	0.17 (0.07)	0.16 (0.09)	0.18 (0.11)
S_3	-0.09 (0.03)	0 (−)	0 (−)	-0.06 (0.07)	-0.08 (0.10)
T	1.57 (0.13)	1.71 (0.11)	1.77 (0.33)	1.21 (0.21)	1.93 (0.36)
X_1	-0.12 (0.12)	0 (−)	0 (−)	0 (−)	0 (−)
X_2	-0.20 (0.18)	0.44 (0.03)	0 (−)	0 (−)	0 (−)
X_3	0.64 (0.17)	0 (−)	0.77 (0.33)	0.17 (0.09)	0.87 (0.34)
T^2	-0.92 (0.11)	-1.12 (0.11)	-1.18 (0.33)	-0.63 (0.21)	-1.16 (0.33)
X_1^2	-0.07 (0.08)	0 (−)	0 (−)	0 (−)	0 (−)
X_2^2	0.67 (0.30)	0 (−)	0.84 (.33)	0.23 (0.10)	0.80 (0.33)
X_3^2	0.05 (0.15)	0 (−)	0 (−)	0 (−)	0 (−)
TX_1	0.24 (0.07)	0 (−)	0 (−)	0 (−)	0 (−)
TX_2	0.12 (0.15)	0 (−)	0 (−)	0 (−)	0 (−)
TX_3	-0.39 (0.11)	0 (−)	0 (−)	0 (−)	-0.23 (0.23)
X_1X_2	-0.25 (0.20)	0 (−)	0 (−)	0 (−)	0 (−)
X_1X_3	0.18 (0.17)	0 (−)	0 (−)	0 (−)	0 (−)
X_2X_3	-0.66 (0.36)	0 (−)	-1.10 (0.59)	0 (−)	-1.03 (0.59)

Table 2. Estimated coefficients and standard errors for Example 4.2

Method	MLE	best subset	SCAD	LASSO	hard
intercept	5.51 (0.75)	6.12 (0.57)	6.09 (0.29)	3.70 (0.25)	5.88 (0.41)
X_1	-8.83 (2.97)	-12.15 (1.81)	-12.24 (0.08)	0 (−)	-11.32 (1.1)
X_2	2.30 (2.00)	0 (−)	0 (−)	0 (−)	2.21 (1.41)
X_3	-2.77 (3.43)	-6.93 (0.79)	-7.00 (0.21)	0 (−)	-4.23 (0.64)
X_4	-1.74 (1.41)	-0.29 (0.11)	0 (−)	-0.28 (0.09)	-1.16 (1.04)
X_1^2	-0.75 (0.61)	0 (−)	0 (−)	-1.71 (0.24)	0 (−)
X_3^2	-2.70 (2.45)	0 (−)	0 (−)	-2.67 (0.22)	-1.92 (0.95)
X_1X_2	0.03 (0.34)	0 (−)	0 (−)	0 (−)	0 (−)
X_1X_3	7.46 (2.34)	9.83 (1.63)	9.84 (0.14)	0.36 (0.22)	9.06 (0.96)
X_1X_4	0.24 (0.32)	0 (−)	0 (−)	0 (−)	0 (−)
X_2X_3	-2.15 (1.61)	0 (−)	0 (−)	-0.10 (0.10)	-2.13 (1.27)
X_2X_4	-0.12 (0.16)	0 (−)	0 (−)	0 (−)	0 (−)
X_3X_4	1.23 (1.21)	0 (−)	0 (−)	0 (−)	0.82 (1.01)

Table 3. Simulation results for linear regression models

Method	MRME(%)	Aver. no. of 0 Coeff.	MRME(%)	Aver. no. of 0 Coeff.	MRME(%)	Aver. no. of 0 Coeff.
	$n = 40,\ \sigma = 3$		$n = 40,\ \sigma = 1$		$n = 60,\ \sigma = 1$	
SCAD^1	72.90	4.41	54.81	4.29	47.54	4.37
SCAD^2	69.03	4.58	47.25	4.34	43.79	4.42
LASSO	63.19	3.60	63.19	3.51	65.22	3.56
Hard	73.82	4.28	69.72	3.93	71.11	4.02
Ridge	83.28	0	95.21	0	97.36	0
Best subset	68.26	4.85	53.60	4.54	46.11	4.73
Garrote	76.90	2.89	56.55	3.35	55.90	3.38

Note that the value of a SCAD^1 is obtained by generalized cross-validation, while the value of a in SCAD^2 is 3.7.

Table 4. Standard deviations of estimators for linear regression models ($n = 60$)

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD^1	0.166	0.161 (0.021)	0.170	0.160 (0.024)	0.148	0.145 (0.022)
SCAD^2	0.161	0.161 (0.021)	0.164	0.161 (0.024)	0.151	0.143 (0.023)
LASSO	0.164	0.154 (0.019)	0.173	0.150 (0.022)	0.153	0.142 (0.021)
Hard	0.169	0.161 (0.022)	0.174	0.162 (0.025)	0.178	0.148 (0.021)
Best subset	0.163	0.155 (0.020)	0.152	0.154 (0.026)	0.152	0.139 (0.020)

Table 5. Simulation results for robust linear models

Method	MRME(%)	Aver. no. of 0 Coeff.
SCAD (a=3.7)	35.52	4.71
LASSO	52.80	4.29
Hard	47.22	4.70
Best subset	41.53	5.03

Table 6. Standard deviations of estimators for robust regression models

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.167	0.171 (0.018)	0.185	0.176 (0.022)	0.165	0.155 (0.020)
LASSO	0.158	0.165 (0.022)	0.159	0.167 (0.020	0.182	0.154 (0.019)
Hard	0.179	0.168 (0.018)	0.176	0.176 (0.025)	0.157	0.154 (0.02)
Best subset	0.198	0.172 (0.023)	0.185	0.175 (0.024)	0.199	0.152 (0.023)

Table 7. Simulation results for logistic regression

Method	MRME(%)	Aver. no. of 0 Coeff.
SCAD(a=3.7)	26.48	5.02
LASSO	53.14	3.76
Hard	59.06	4.27
Best subset	31.63	4.85

Table 8. Standard deviations of estimators for logistic regression

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD ($a = 3.7$)	0.571	0.538 (0.107)	0.383	0.372 (0.061)	0.432	0.498 (0.065)
LASSO	0.310	0.379 (0.037)	0.285	0.284 (0.019)	0.244	0.287 (0.019)
Hard	0.675	0.561 (0.126)	0.428	0.400 (0.062)	0.467	0.421 (0.079)
Best subset	0.624	0.547 (0.121)	0.398	0.383 (0.067)	0.468	0.412 (0.077)

Table 9. Simulation results for Poisson log-linear regression

Method	MRME(%)	Aver. no. of 0 Coeff.
SCAD(a=3.7)	48.00	3.61
LASSO	60.93	3.60
Hard	70.07	3.66
Best subset	33.96	4.71

Table 10. Standard deviations of estimators for linear regression models

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD ($a = 3.7$)	0.080	0.079 (0.014)	0.093	0.084 (0.016)	0.079	0.072 (0.016)
LASSO	0.086	0.078 (0.013)	0.101	0.082 (0.016)	0.083	0.074 (0.017)
Hard	0.084	0.080 (0.015)	0.100	0.086 (0.019)	0.081	0.075 (0.020)
Best subset	0.081	0.079 (0.016)	0.080	0.083 (0.018)	0.079	0.068 (0.016)

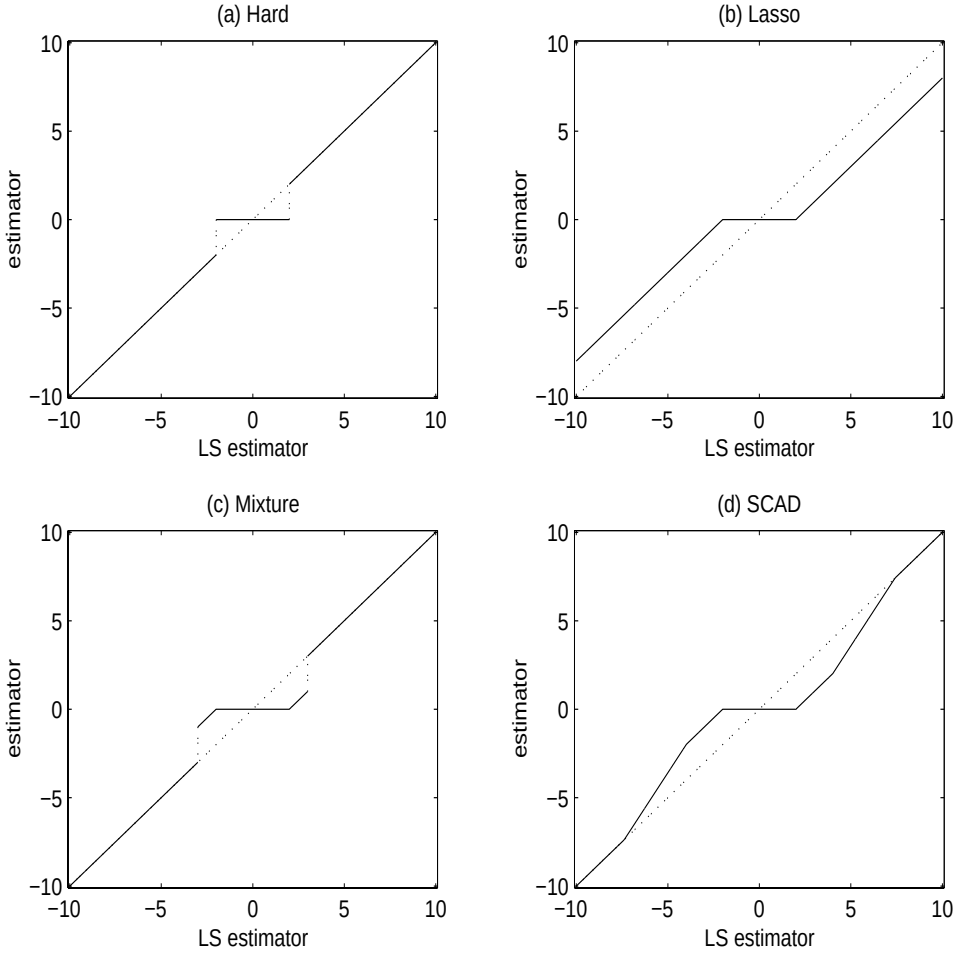

Figure 1: *Plot of thresholding estimators (with $\lambda = 2$) against the least-squares estimate. (d) corresponds to the SCAD (2.11) with $a = 3.7$.*

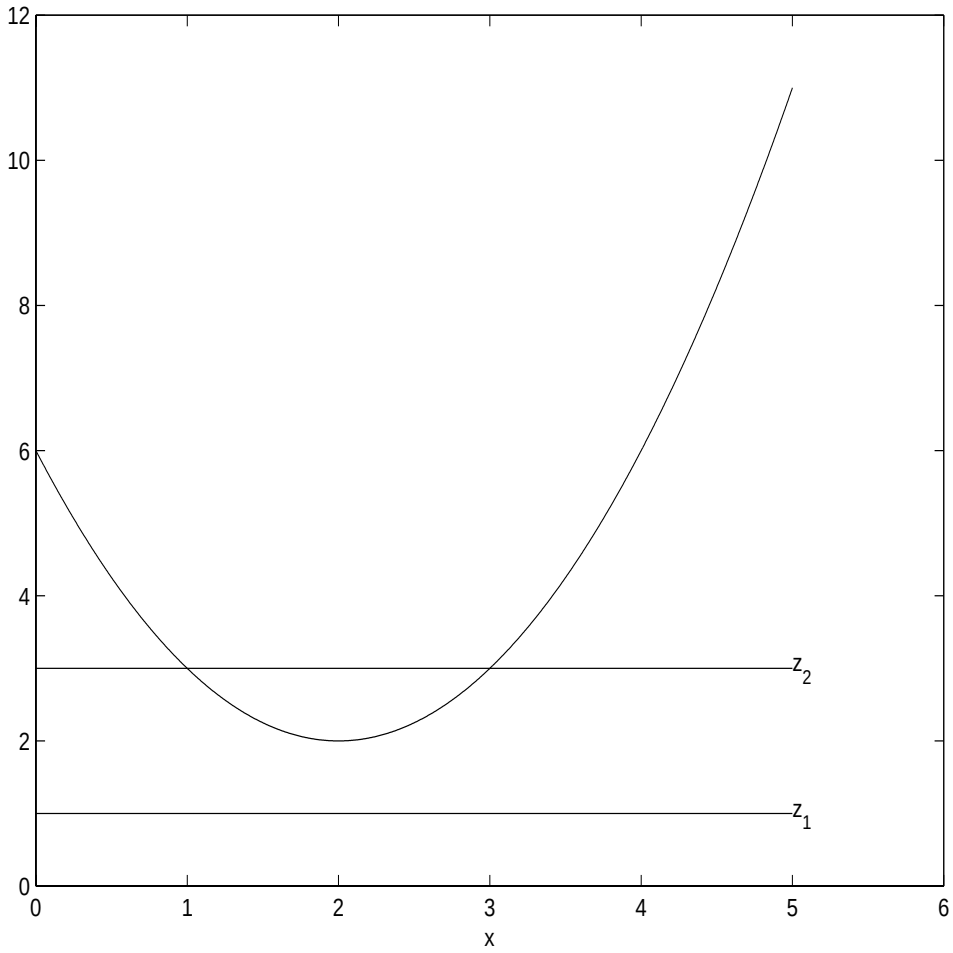

Figure 2: *A plot of $\theta + p'_\lambda(\theta)$ against θ ($\theta > 0$). For a small value $z_1 > 0$, the derivative function $\theta + p'_\lambda(\theta) - z_1$ is above zero and hence the solution to the penalized least-squares problem (2.5) is zero. The minimizer is continuous in z only when the minimum of $\theta + p'_\lambda(\theta)$ is attained at zero.*

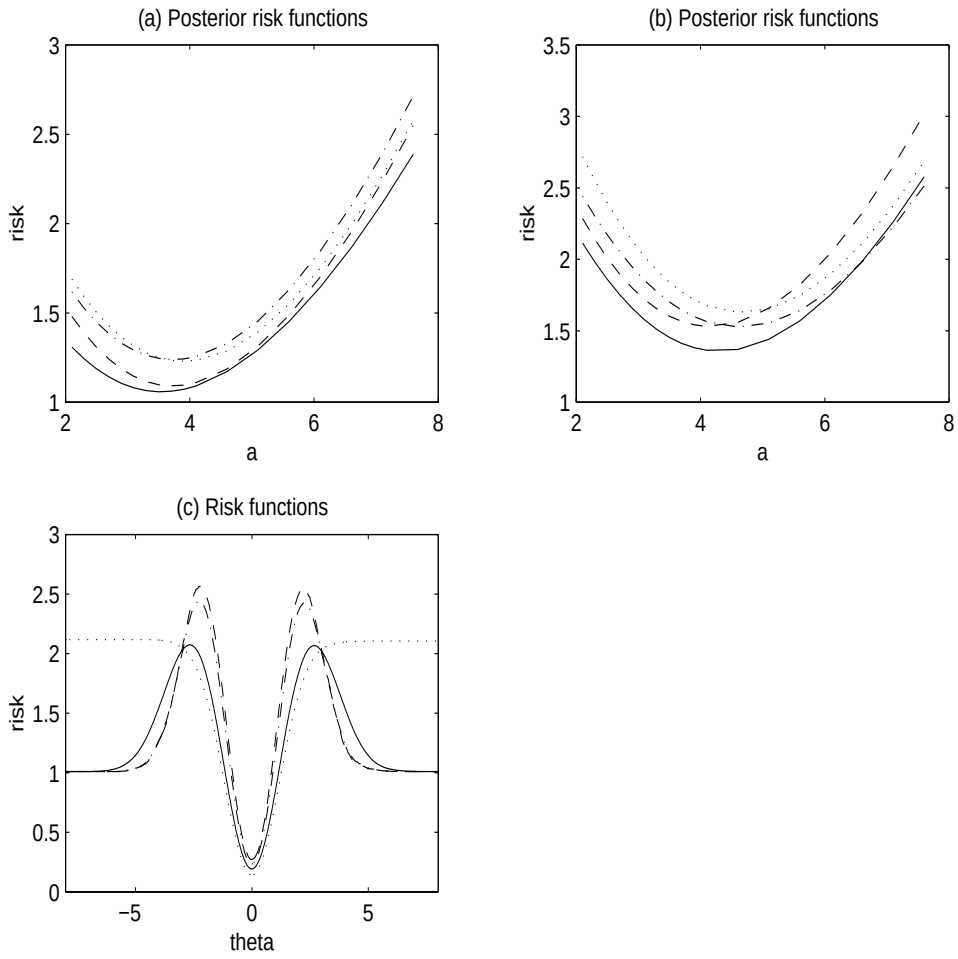

Figure 3: *Risk functions of proposed procedures under the quadratic loss. (a) and (b) are posterior risk functions of the penalized smoothly clipped-L_1 estimator under the prior $\theta \sim N(0, a\lambda)$ using the universal thresholding $\lambda = \sqrt{2\log(d)}$ for 4 different values d; the solid, dashed, dashdot and dotted lines are for $d = 20,\ 40,\ 60$ and 100, respectively. The caption for (b) is similar to those for (a) with the solid, dashed, dashdot, dotted lines for $d = 512,\ 1024,\ 2048$ and 4096, separately. (c) Risk functions of the four different thresholding rules. The solid, dashed, dashdot and dotted lines are for minimum SCAD, hard, mixture and soft thresholding rules.*

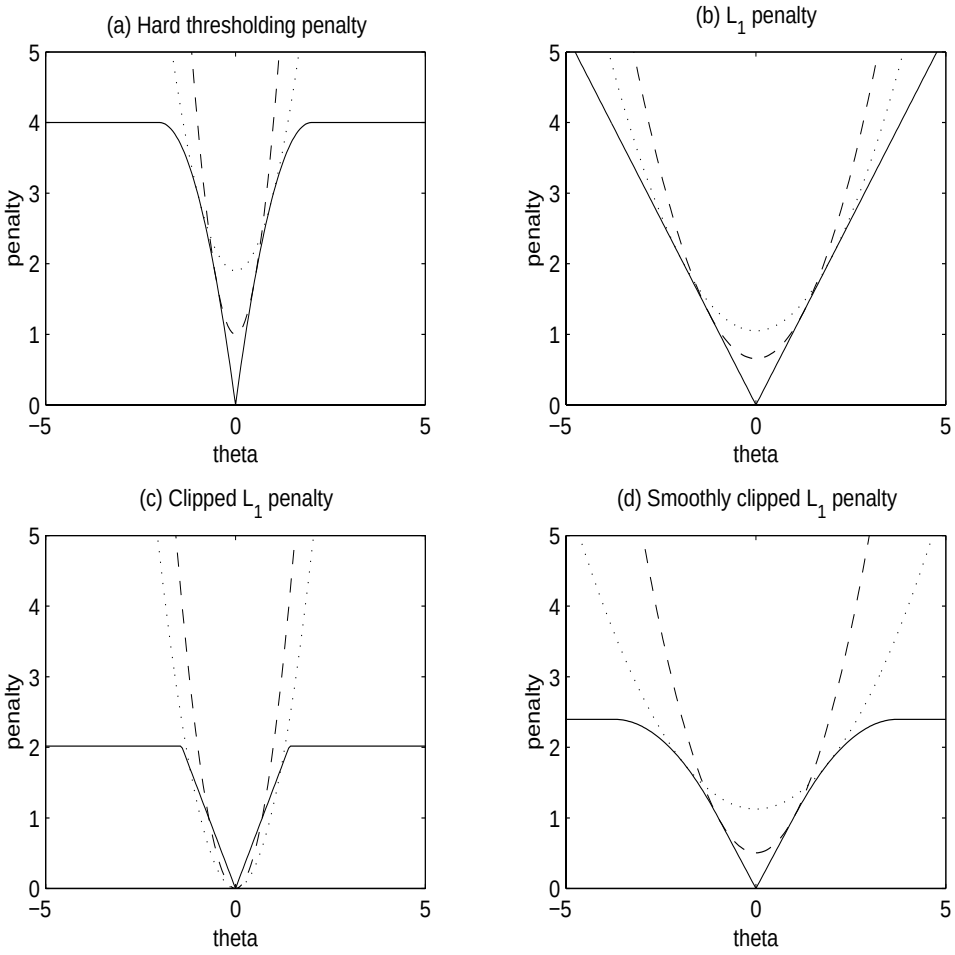

Figure 4: *Four penalties $p_\lambda(\theta)$ and their quadratic approximations. The values of λ are the same as those in Figure 3(c).*

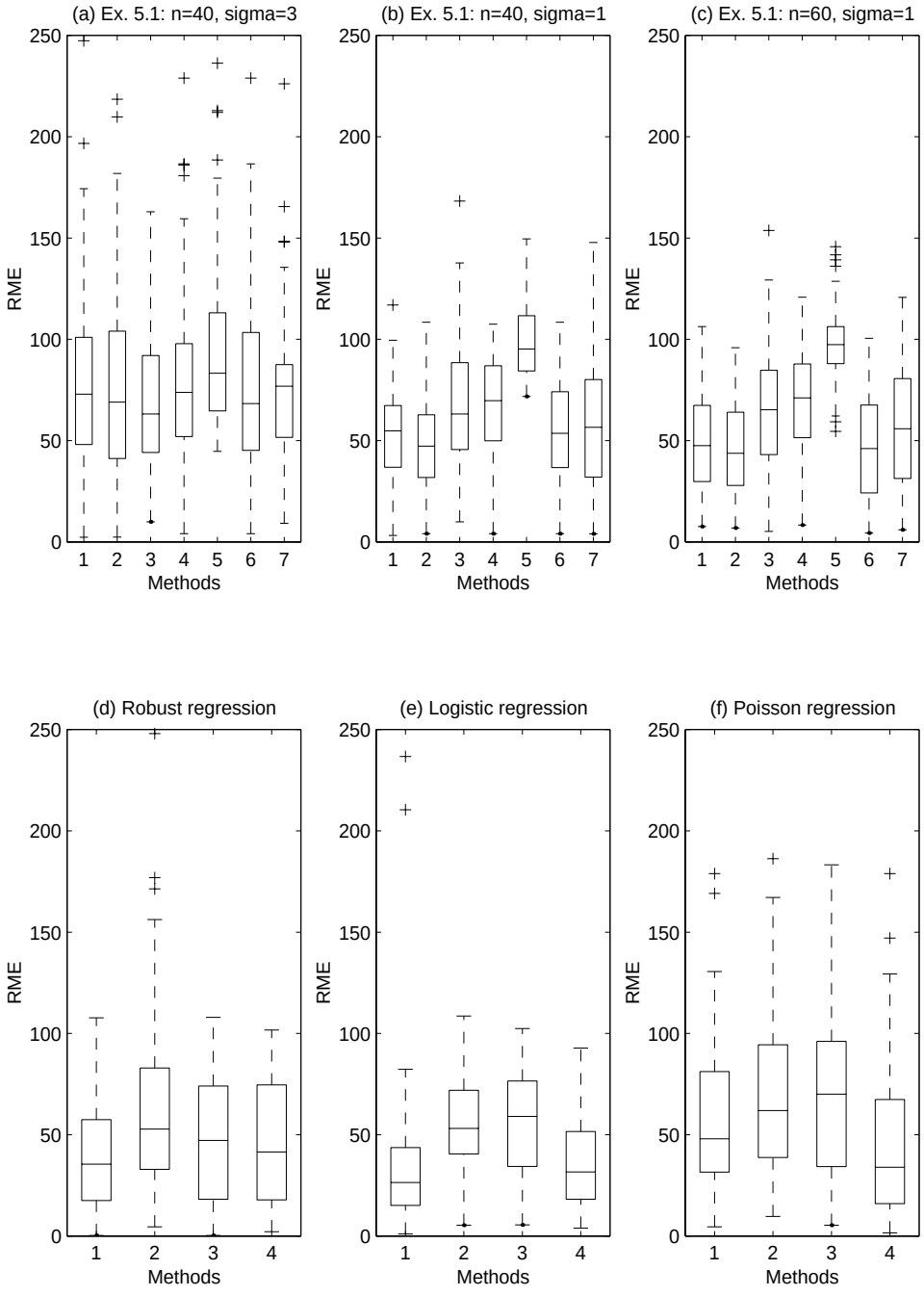

Figure 5: *Boxplots of relative model errors. From left to right, the order in the top panel is* $SCAD^1$, $SCAD^2$, *LASSO, hard, ridge, best subset and nonnegative garrote. The order from left to right in the bottom panel is SCAD (a=3.7), LASSO, hard and best subset.*