Paper Review: Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties by Jianqing Fan and Runze Li (2001)

Presented by Yang Zhao

March 5, 2010

Outlines

Motivation

- Variable selection is fundamental to high-dimensional statistical modeling
- Many approaches in use are stepwise selection procedures, which can be computationally expensive and ignore stochastic errors in the variable selection process
- The theoretical properties for stepwise deletion and subset selection are somewhat hard to understand
- The most severe drawback of the best subset variable selection is its lack of stability as analyzed by Breiman (1996)

Overview of Proposed Method

- Penalized likelihood approaches are proposed, which simultaneously select variables and estimate coefficients
- Penalty functions are symmetric, nonconcave on $(0, \infty)$, and have singularities at the origin to produce sparse solutions
- Furthermore the penalty functions are bounded by a constant to reduce bias and satisfy certain conditions to yield continuous solutions
- A new algorithm is proposed for optimizing penalized likelihood functions, which is widely applicable
- Rates of convergence of the proposed penalized likelihood estimators are established
- With proper choice of regularization parameters, the proposed estimators perform as well as the oracle procedure (as if the correct submodel were known)

Penalized Least Squares and Subset Selection

Linear regression model: $y = X\beta + \epsilon$, where y is $n \times 1$ and X is $n \times d$. Assume for now that the columns of X are orthonormal. Denote $z = X^T y$. A form of the penalized least squares is
$\frac{1}{2}||y - X\beta||^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|) =$
$\frac{1}{2}||y - XX^T y||^2 + \frac{1}{2}\sum_{j=1}^{d}(z_j - \beta_j)^2 + \lambda \sum_{j=1}^{d} p_j(|\beta_j|)$.

- $p_j(.)$ are not necessarily the same for all j; but we will assume they are the same for simplicity.
- From now on, denote $\lambda p(|.|)$ by $p_\lambda(|.|)$.
- The minimization problem is equivalent to minimize componentwise, which leads us to consider the penalized least squares problem
  $\frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|)$. (2.3)

Penalized Least Squares and Subset Selection Cont'd

By taking the *hard thresholding* penalty function $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$, the *hard thresholding rule* is obtained as $\hat{\theta} = zI(|z| > \lambda)$.

This coincides with the best subset selection and stepwise deletion and addition for orthonormal designs!

What Makes a Good Penalty Function?

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

Sufficient Conditions for Good Penalty Functions

- $p'_\lambda(|\theta|) = 0$ for large $|\theta| \Rightarrow$ *Unbiasedness*
  Intuition: The first order derivative of (2.3) w.r.t. θ is $sgn(\theta)(|\theta| + p'_\lambda(|\theta|)) - z$. When $p'_\lambda(|\theta|) = 0$ for large $|\theta|$, the resulting estimator is z when $|z|$ is sufficiently large, which is approximately unbiased.
- The minimum of the function $|\theta| + p'_\lambda(|\theta|)$ is positive $\Rightarrow$ *Sparsity*
- The minimum of the function $|\theta| + p'_\lambda(|\theta|)$ is attained at 0 $\Rightarrow$ *Continuity*
  Intuition: See nextpage

Sufficient Conditions for Good Penalty Functions Cont'd

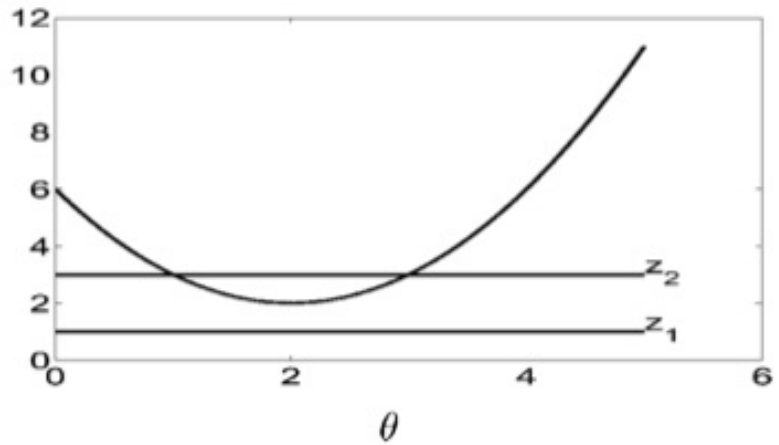

Figure 3. A Plot of $\theta + p'_\lambda(\theta)$ Against $\theta(\theta > 0)$.

- when $|z| < min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\}$, the derivative of (2.3) is positive for all positive θ and negative for all negative $\theta \Rightarrow \hat{\theta} = 0$; when $|z| > min_{\theta \neq 0}\{|\theta| + p'_\lambda(|\theta|)\}$, two crossings may exist as shown, the larger one is a penalized least squares estimator.
- further, this implies that a sufficient and necessary condition for continuity is to require the minimum to be attained at 0

Penalty Function of Interest

- **hard thresholding penalty**: $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$.
- **soft thresholding penalty (L_1 penalty, LASSO)**: $p_\lambda(|\theta|) = \lambda|\theta|$.
- **L_2 penalty (ridge regression)**: $p_\lambda(|\theta|) = \lambda|\theta|^2$.
- **SCAD penalty (Smoothly Clipped Absolute Deviation penalty)**: $p'_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda} I(\theta > \lambda)\}$ for some $a > 2$ and $\theta > 0$.

Note: As for the three properties that make a penalty function "good", only SCAD possesses all (and therefore is advocated by the authors); whereas all the other penalty functions are unable to satisfy three sufficient conditions simultaneously.

Penalty Function of Interest Cont'd

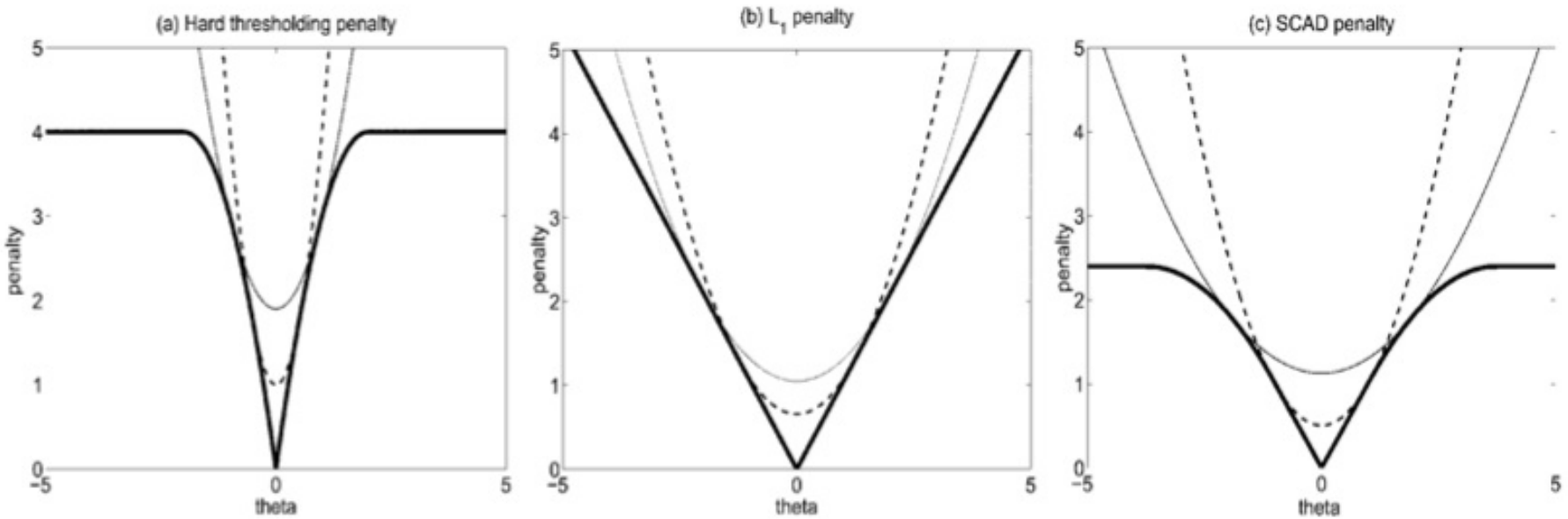

Figure 1. Three Penalty Functions $p_\lambda(\theta)$ and Their Quadratic Approximations. The values of λ are the same as those in Figure 5(c).

Penalty Function of Interest Cont'd

When the design matrix X is orthonormal, closed form solutions are obtained as follows.

- hard thresholding rule: $\hat{\theta} = zI(|z| > \lambda)$
- LASSO (L_1 penalty): $\hat{\theta} = sgn(z)(|z| - \lambda)_+$
- SCAD:

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

Penalty Function of Interest Cont'd

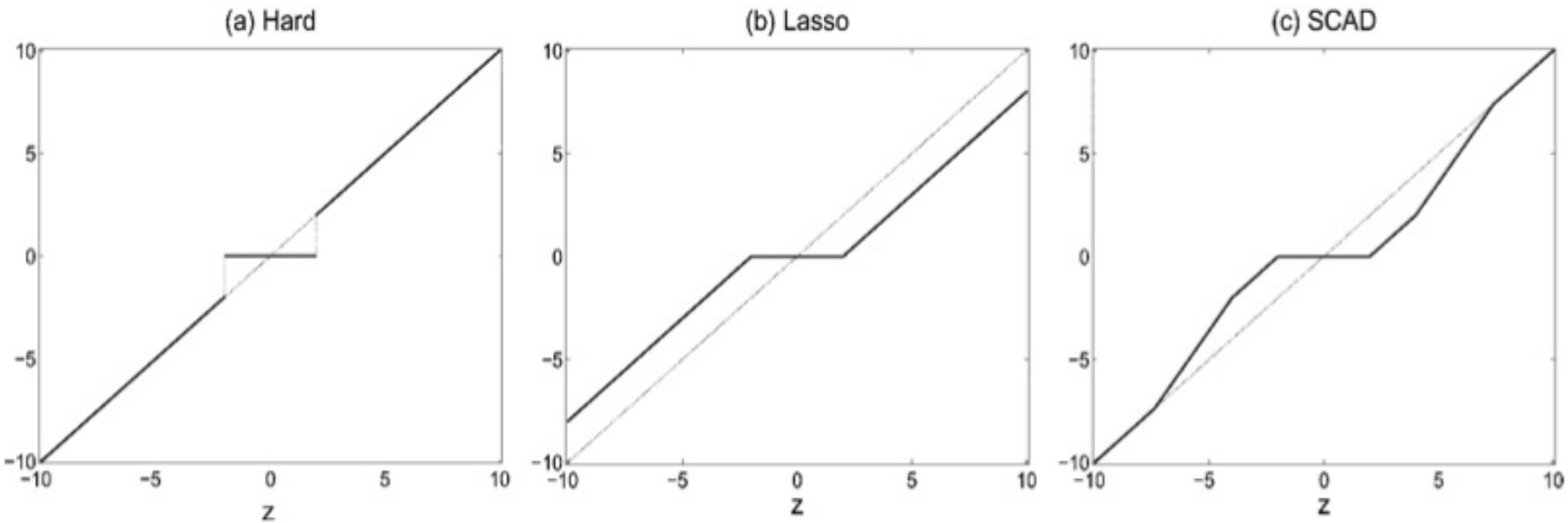

Figure 2. Plot of Thresholding Functions for (a) the Hard, (b) the Soft, and (c) the SCAD Thresholding Functions With $\lambda = 2$ and $a = 3.7$ for SCAD.

More on SCAD

SCAD penalty has two unknown parameters λ and a. They could be chosen by searching grids using cross-validation (CV) or generalized cross-validation (GCV), but that is computation intensive. The authors applied Bayesian risk analysis to help choose a, in that Bayesian risks seem not very sensitive to the values of a. It is found that $a = 3.7$ works similarly to that chosen by GCV.

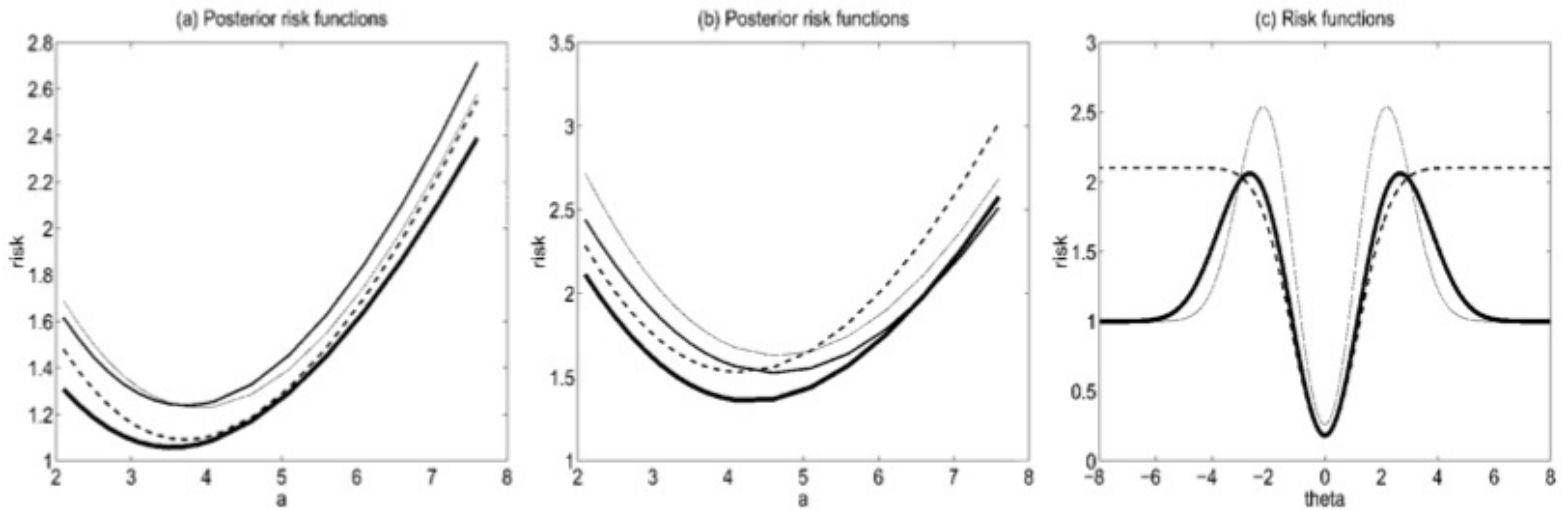

Figure 5. Risk Functions of Proposed Procedures Under the Quadratic Loss. (a) Posterior risk functions of the SCAD under the prior $\theta \sim N(0, a\lambda)$ using the universal thresholding $\lambda = \sqrt{2\log(d)}$ for four different values d: heavy line, $d = 20$; dashed line, $d = 40$; medium line, $d = 60$; thin line, $d = 100$. (b) Risk functions similar to those for (a): heavy line, $d = 572$; dashed line, $d = 1{,}024$; medium line, $d = 2{,}048$; thin line, $d = 4{,}096$. (c) Risk functions of the four different thresholding rules. The heavy, dashed, and solid lines denote minimum SCAD, hard, and soft thresholding rules, respectively.

Performance Comparison

Refer to the last graph on previous slide. SCAD performs favorably compared with the other two thresholding rules. It is actually expected to perform the best, given that it retains all the good mathematical properties of the other two penalty functions.

Penalized Likelihood Formulation

From now on, assume that X is standardized so that each column has mean 0 and variance 1, and is no longer orthonormal.
A form of penalized least squares is
$\frac{1}{2}(y - X\beta)^T(y - X\beta) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$. (3.1)
A form of penalized robust least squares is
$\sum_{i=1}^{n} \psi(|y_i - x_i\beta|) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$. (3.2)
A form of penalized likelihood for generalized linear models is
$-\sum_{i=1}^{n} l_i(g(x_i^T\beta), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$. (3.3)

Sampling Properties and Oracle Properties

Let $\beta_{\mathbf{0}} = (\beta_{10}, \ldots, \beta_{d0})^T = (\beta_{\mathbf{10}}^{\mathbf{T}}, \beta_{\mathbf{20}}^{\mathbf{T}})^T$. Assume that $\beta_{\mathbf{20}} = 0$. Let $I(\beta_{\mathbf{0}})$ be the Fisher information matrix and $I_1(\beta_{\mathbf{10}}, 0)$ be the Fisher information knowing $\beta_{\mathbf{20}} = 0$.

General setting: Let $V_i = (X_i, Y_i)$, $i = 1, \ldots, n$. Let $L(\beta)$ be the log-likelihood function of observations $V_1, \ldots, V_n$ and let $Q(\beta)$ be the penalized likelihood function $L(\beta) - n \sum_{j=1}^{d} p_\lambda(|\beta_j|)$.

Sampling Properties and Oracle Properties Cont'd

Regularity Conditions

(A) The observations V_i are iid with density $f(V, \beta)$. $f(V, \beta)$ has a common support and the model is identifiable. Furthermore, the following holds: $E_\beta(\frac{\partial logf(V,\beta)}{\partial \beta_j}) = 0$ for $j = 1, \ldots, d$ and

$I_{jk}(\beta) = E_\beta(\frac{\partial}{\partial \beta_j} logf(V, \beta) \frac{\partial}{\partial \beta_k} logf(V, \beta)) = E_\beta(-\frac{\partial^2}{\partial \beta_j \partial \beta_k} logf(V, \beta))$.

(B) The Fisher information matrix $I(\beta) = E\{(\frac{\partial}{\partial \beta} logf(V, \beta))(\frac{\partial}{\partial \beta} logf(V, \beta))^T\}$ is finite and positive definite at $\beta = \beta_0$.

(C) There exists an open subset ω of Ω that contains the true parameter point β_0 such that for almost all V the density $f(V, \beta)$ admits all third derivatives for all $\beta \in \omega$. Furthermore, there exists functions M_{jkl} such that $|\frac{\partial^3}{\partial \beta_i \partial \beta_k \partial \beta_l} logf(V, \beta)| \leq M_{jkl}(V)$ for all $\beta \in \omega$, where $m_j kl = E_\beta(M_{jkl}(V)) < \infty$ for j, k, l.

Sampling Properties and Oracle Properties Cont'd

Theorem 1. Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). If $max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \to 0$, then there exists a local maximizer $\hat{\beta}$ of $Q(\beta)$ such that $||\hat{\beta} - \beta_0|| = O_p(n^{-1/2} + a_n)$, where $a_n = max\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$.
Note:
For hard thresholding and SCAD penalty functions, $\lambda_n \to 0$ implies $a_n = 0$, therefore the penalized likelihood estimator is root-n consistent.

Sampling Properties and Oracle Properties Cont'd

Lemma 1. Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). Assume that $lim\ inf_{n\to\infty} lim\ inf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$ (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ ad $n \to \infty$, then with probability tending to 1, for any given β_1 satisfying $||\beta_1 - \beta_{10}|| = O_p(n^{-1/2})$ and any constant C, $Q\{(\beta_1^T, 0)^T\} = max_{||\beta_2|| \le Cn^{-1/2}} Q\{(\beta_1^T, \beta_2^T)^T\}$

Aside: Denote $\Sigma = diag\{p''_{\lambda_n}(\beta_{10}), \ldots, p''_{\lambda_n}(\beta_{s0})\}$ and $b = (p'_{\lambda_n}(\beta_{10})sgn(\beta_{10}), \ldots, p'_{\lambda_n}(\beta_{s0})sgn(\beta_{s0}))^T$, where s is the number of components of β_{10}.

Theorem 2 (Oracle Property). Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then with probability tending to 1, the root-n consistent local maximizers $\hat{\beta} = (\hat{\beta_1}^T, \hat{\beta_2}^T)^T$ in Theorem 1 must satisfy:

(a) Sparsity: $\hat{\beta}_2 = 0$

(b) Asymptotic normality: $\sqrt{n}(I_1(\beta_{10}) + \Sigma)\{\hat{\beta}_1 - \beta_{10} + (I_1(\beta_{10}) + \Sigma)^{-1}b\} \to N\{0, I_1(\beta_{10})\}$

Sampling Properties and Oracle Properties Cont'd

Lemma 1. Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). Assume that $lim\ inf_{n\to\infty} lim\ inf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$ (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ ad $n \to \infty$, then with probability tending to 1, for any given β_1 satisfying $||\beta_1 - \beta_{10}|| = O_p(n^{-1/2})$ and any constant C, $Q\{(\beta_1^T, 0)^T\} = max_{||\beta_2|| \le Cn^{-1/2}} Q\{(\beta_1^T, \beta_2^T)^T\}$

Aside: Denote $\Sigma = diag\{p''_{\lambda_n}(\beta_{10}), \ldots, p''_{\lambda_n}(\beta_{s0})\}$ and $b = (p'_{\lambda_n}(\beta_{10})sgn(\beta_{10}), \ldots, p'_{\lambda_n}(\beta_{s0})sgn(\beta_{s0}))^T$, where s is the number of components of β_{10}.

Theorem 2 (Oracle Property). Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then with probability tending to 1, the root-n consistent local maximizers $\hat{\beta} = (\hat{\beta_1}^T, \hat{\beta_2}^T)^T$ in Theorem 1 must satisfy:

(a) Sparsity: $\hat{\beta}_2 = 0$

(b) Asymptotic normality:

$\sqrt{n}(I_1(\beta_{10}) + \Sigma)\{\hat{\beta}_1 - \beta_{10} + (I_1(\beta_{10}) + \Sigma)^{-1} b\} \to N\{0, I_1(\beta_{10})\}$

Sampling Properties and Oracle Properties Cont'd

Lemma 1. Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). Assume that $lim\ inf_{n\to\infty} lim\ inf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$ (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ ad $n \to \infty$, then with probability tending to 1, for any given β_1 satisfying $||\beta_1 - \beta_{10}|| = O_p(n^{-1/2})$ and any constant C, $Q\{(\beta_1^T, 0)^T\} = max_{||\beta_2|| \le Cn^{-1/2}} Q\{(\beta_1^T, \beta_2^T)^T\}$

Aside: Denote $\Sigma = diag\{p''_{\lambda_n}(\beta_{10}), \ldots, p''_{\lambda_n}(\beta_{s0})\}$ and $b = (p'_{\lambda_n}(\beta_{10})sgn(\beta_{10}), \ldots, p'_{\lambda_n}(\beta_{s0})sgn(\beta_{s0}))^T$, where s is the number of components of β_{10}.

Theorem 2 (Oracle Property). Let $V_1, \ldots, V_n$ be iid, each with a density $f(V, \beta)$ that satisfies conditions (A)-(C). Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.5). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then with probability tending to 1, the root-n consistent local maximizers $\hat{\beta} = (\hat{\beta}_1^T, \hat{\beta}_2^T)^T$ in Theorem 1 must satisfy:

(a) Sparsity: $\hat{\beta}_2 = 0$

(b) Asymptotic normality:

$\sqrt{n}(I_1(\beta_{10}) + \Sigma)\{\hat{\beta}_1 - \beta_{10} + (I_1(\beta_{10}) + \Sigma)^{-1} b\} \to N\{0, I_1(\beta_{10})\}$

Sampling Properties and Oracle Properties Cont'd

Remarks:
(1)For the hard and SCAD thresholding penalty functions, if $\lambda_n \to 0$, $a_n = 0$. Hence, by Thm 2, when $\sqrt{n}\lambda_n \to \infty$, their corresponding penalized likelihood estimators possess the oracle property and perform as well as the maximum likelihood estimates for estimating β_1 knowing $\beta_2 = 0$.
(2)However, for LASSO (L_1) penalty, $a_n = \lambda_n$. Hence, the root-n consistency requires $\lambda_n = O_p(n^{-1/2})$. On the other hand, the oracle property requires $\sqrt{n}\lambda_n \to \infty$. These two conditions cannot be satisfied simultaneously. The authors conjecture that the oracle property does not hold for LASSO.
(3)For L_q penalty with $q < 1$, the oracle property continues to hold with suitable choice of λ_n.

A New Unified Algorithm

- Tibshirani (1996) proposed an algorithm for solving constrained least squares problem of LASSO
- Fu (1998) provided a "shooting algorithm" for LASSO
- LASSO2 submitted by Berwin Turlach at Statlib
- Here the authors proposed a unified algorithm, which optimizes problems (3.1) (3.2) (3.3) via local quadratic approximations.

The Algorithm

- Re-written (3.1) (3.2) (3.3) as $l(\beta) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|)$ (3.6), where $l(\beta)$ is a general loss function.
- Given an initial value β_0 that is close to the minimizer of (3.6). If β_{j0} is very close to 0, then set $\hat{\beta}_j = 0$; otherwise they can be locally approximated by a quadratic function as $[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)sgn(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j$. In other words, $p_\lambda(|\beta_j|) \approx p_\lambda(\beta_{j0}) + \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2)$, for $\beta_j \approx \beta_{j0}$.
- $\psi(|y - x^T\beta|)$ in robust penalized least squares case can be approximated by $\{\psi(|y - x^T\beta_0|)/(y - x^T\beta_0)^2\}(y - x^T\beta)^2$.
- assume the log-likelihood function is smooth, so can be locally approximated by a quadratic function.
- Newton-Raphson algorithm applies.
- The updating equation is $\beta_1 = \beta_0 - (\nabla^2 l(\beta_0) + n\Sigma_\lambda(\beta_0))^{-1}(\nabla l(\beta_0) + nU_\lambda(\beta_0))$, where $\Sigma_\lambda(\beta_0) = diag\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}$, $U_\lambda(\beta_0) = \Sigma_\lambda(\beta_0)$.

Standard Error Formula

Obtained directly from the algorithm.
Sandwich formula:
$\hat{cov}(\hat{\beta}_1) = (\nabla^2 l(\hat{\beta}_1) + n\Sigma_\lambda(\hat{\beta}_1))^{-1} \hat{cov}\{\nabla l(\hat{\beta}_1)\}(\nabla^2 l(\hat{\beta}_1) + n\Sigma_\lambda(\hat{\beta}_1))^{-1}$ for nonvanishing component of β.

Prediction and Model Error

- Two regression situations: X random and X controlled
- For ease of presentation, consider only X-random case
- $PE(\hat{\mu}) = E(Y - \hat{\mu}x)^2 = E(Y - E(Y|x))^2 + E(Y|x - \hat{\mu}(x))^2$
- The second component is the *model error*

Simulation Results

Table 1. Simulation Results for the Linear Regression Model

Method	MRME (%)	Avg. No. of 0 Coefficients: Correct	Avg. No. of 0 Coefficients: Incorrect
$n = 40, \sigma = 3$			
SCAD[1]	72.90	4.20	.21
SCAD[2]	69.03	4.31	.27
LASSO	63.19	3.53	.07
Hard	73.82	4.09	.19
Ridge	83.28	0	0
Best subset	68.26	4.50	.35
Garrote	76.90	2.80	.09
Oracle	33.31	5	0
$n = 40, \sigma = 1$			
SCAD[1]	54.81	4.29	0
SCAD[2]	47.25	4.34	0
LASSO	63.19	3.51	0
Hard	69.72	3.93	0
Ridge	95.21	0	0
Best subset	53.60	4.54	0
Garrote	56.55	3.35	0
Oracle	33.31	5	0
$n = 60, \sigma = 1$			
SCAD[1]	47.54	4.37	0
SCAD[2]	43.79	4.42	0
LASSO	65.22	3.56	0
Hard	71.11	4.02	0
Ridge	97.36	0	0
Best subset	46.11	4.73	0
Garrote	55.90	3.38	0

Simulation Results Cont'd

Table 2. Standard Deviations of Estimators for the Linear Regression Model ($n = 60$)

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})	*SD*	SD_m (SD_{mad})
SCAD[1]	.166	.161 (.021)	.170	.160 (.024)	.148	.145 (.022)
SCAD[2]	.161	.161 (.021)	.164	.161 (.024)	.151	.143 (.023)
LASSO	.164	.154 (.019)	.173	.150 (.022)	.153	.142 (.021)
Hard	.169	.161 (.022)	.174	.162 (.025)	.178	.148 (.021)
Best subset	.163	.155 (.020)	.152	.154 (.026)	.152	.139 (.020)
Oracle	.155	.154 (.020)	.147	.153 (.024)	.146	.137 (.019)

Note:
SD = median absolute deviation of $\hat{\beta}_1/0.6745$
SD_m=median of $\hat{\sigma}(\hat{\beta}_1)$
SD_{mad}=median absolute deviation of $\hat{\sigma}(\hat{\beta}_1)$

Simulation Results Cont'd

Table 5. Simulation Results for the Logistic Regression

		Avg. No. of 0 Coefficients	
Method	*MRME (%)*	*Correct*	*Incorrect*
SCAD ($a = 3.7$)	26.48	4.98	.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	.01
Oracle	25.71	5	0

Real Data Analysis

Table 7. Estimated Coefficients and Standard Errors for Example 4.4

Method	MLE	Best Subset (AIC)	Best Subset (BIC)	SCAD	LASSO	Hard
Intercept	5.51 (.75)	4.81 (.45)	6.12 (.57)	6.09 (.29)	3.70 (.25)	5.88 (.41)
X_1	−8.83 (2.97)	−6.49 (1.75)	−12.15 (1.81)	−12.24 (.08)	0 (—)	−11.32 (1.1)
X_2	2.30 (2.00)	0 (—)	0 (—)	0 (—)	0 (—)	2.21 (1.41)
X_3	−2.77 (3.43)	0 (—)	−6.93 (.79)	−7.00 (.21)	0 (—)	−4.23 (.64)
X_4	−1.74 (1.41)	.30 (.11)	−.29 (.11)	0 (—)	−.28 (.09)	−1.16 (1.04)
X_1^2	−.75 (.61)	−1.04 (.54)	0 (—)	0 (—)	−1.71 (.24)	0 (—)
X_3^2	−2.70 (2.45)	−4.55 (.55)	0 (—)	0 (—)	−2.67 (.22)	−1.92 (.95)
X_1X_2	.03 (.34)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_1X_3	7.46 (2.34)	5.69 (1.29)	9.83 (1.63)	9.84 (.14)	.36 (.22)	9.06 (.96)
X_1X_4	.24 (.32)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_2X_3	−2.15 (1.61)	0 (—)	0 (—)	0 (—)	−0.10 (.10)	−2.13 (1.27)
X_2X_4	−.12 (.16)	0 (—)	0 (—)	0 (—)	0 (—)	0 (—)
X_3X_4	1.23 (1.21)	0 (—)	0 (—)	0 (—)	0 (—)	.82 (1.01)

Summary

- A family of penalty functions was introduced, of which SCAD performs favorably
- Rates of convergence of the proposed penalized likelihood estimators were established
- With proper choice of regularization parameters, the estimators perform as well as the oracle procedure
- The unified algorithm was demonstrated effective and standard errors were estimated with good accuracy
- The proposed approach can be applied to various statistical contexts

Proofs

Proof of Theorem 1

Let $\alpha_n = n^{-1/2} + a_n$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < Q(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon. \tag{A.1}$$

This implies with probability at least $1-\varepsilon$ that there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}: \|\mathbf{u}\| \leq C\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_n)$.

Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D_n(\mathbf{u}) &\equiv Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0) \\ &\leq L(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - L(\boldsymbol{\beta}_0) - n \sum_{j=1}^{s} \{ p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|) \}, \end{aligned}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. Let $L'(\boldsymbol{\beta}_0)$ be the gradient vector of L. By the standard argument on the Taylor expansion of the likelihood function, we have

$$\begin{aligned} D_n(\mathbf{u}) \leq{}& \alpha_n L'(\boldsymbol{\beta}_0)^T \mathbf{u} - \tfrac{1}{2} \mathbf{u}^T I(\boldsymbol{\beta}_0) \mathbf{u} n \alpha_n^2 \{1 + o_P(1)\} \\ & - \sum_{j=1}^{s} [n \alpha_n p'_{\lambda_n}(|\beta_{j0}|) \operatorname{sgn}(\beta_{j0}) u_j \\ & + n \alpha_n^2 p''_{\lambda_n}(|\beta_{j0}|) u_j^2 \{1 + o(1)\}]. \end{aligned} \tag{A.2}$$

Note that $n^{-1/2} L'(\boldsymbol{\beta}_0) = O_P(1)$. Thus, the first term on the right-hand side of (A.2) is on the order $O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. Note that the third term in (A.2) is bounded by

$$\sqrt{s} n \alpha_n a_n \|\mathbf{u}\| + n \alpha_n^2 \max\{ |p''_{\lambda_n}(|\beta_{j0}|)|: \beta_{j0} \neq 0 \} \|\mathbf{u}\|^2.$$

Proofs

Proof of Lemma 1

It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \ldots, d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for } 0 < \beta_j < \varepsilon_n \tag{A.3}$$

$$> 0 \quad \text{for } -\varepsilon_n < \beta_j < 0. \tag{A.4}$$

To show (A.3), by Taylor's expansion, we have

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j) \\ &= \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{l0}) + \sum_{l=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l \beta_k} \\ &\quad \times (\beta_l - \beta_{l0})(\beta_k - \beta_{k0}) - np'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j),\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that by the standard arguments,

$$n^{-1}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_P(n^{-1/2})$$

and

$$\frac{1}{n}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} = E\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}\right\} + o_P(1).$$

By the assumption that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\operatorname{sgn}(\beta_j) + O_P(n^{-1/2}/\lambda_n)\}.$$

Whereas $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\lambda_n^{-1}p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, (A.3) and (A.4) follow. This completes the proof.

Proofs

Proof of Theorem 2

It follows by Lemma 1 that part (a) holds. Now we prove part (b). It can be shown easily that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 1 that is a root-n consistent local maximizer of $Q\{\binom{\boldsymbol{\beta}_1}{\mathbf{0}}\}$, which is regarded as a function of $\boldsymbol{\beta}_1$, and that satisfies the likelihood equations

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\binom{\hat{\boldsymbol{\beta}}_1}{\mathbf{0}}} = 0 \quad \text{for } j = 1, \ldots, s. \tag{A.5}$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned}
&\left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\binom{\hat{\boldsymbol{\beta}}_1}{\mathbf{0}}} - n p'_{\lambda_n}(|\hat{\beta}_j|)\operatorname{sgn}(\hat{\beta}_j) \\
&\quad = \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} + o_P(1)\right\}(\hat{\beta}_l - \beta_{l0}) \\
&\qquad - n\left(p'_{\lambda_n}(|\beta_{j0}|)\operatorname{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\hat{\beta}_j - \beta_{j0})\right).
\end{aligned}$$

It follows by Slutsky's theorem and the central limit theorem that

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\mathbf{b}\right\} \to N\{\mathbf{0}, I_1(\boldsymbol{\beta}_{10})\}$$

in distribution.

The End!