Variable Selection Methods via Penalized Likelihood and Comparison of their Properties

Jian Shi

University of California

shi@pstat.ucsb.edu

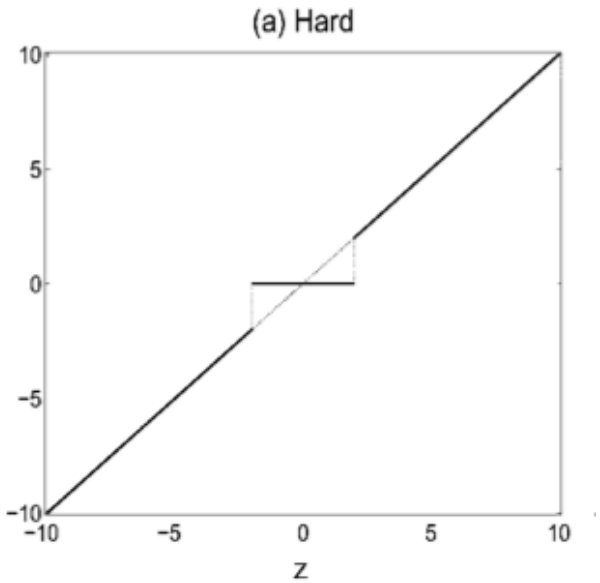
March 12, 2014

Outline

- Variable selection problem
- Ridge regression and the lasso
- SCAD penalty (Nonconcave penalties, Fan 1997)
- Simulations

Variable Selection

- Want to build a model using a subset of "predictors"
- Multiple linear regression; logistic regression (GLM); Cox's partial likelihood, ...
  - model selection criteria: AIC, BIC, etc. (Best subset selection plus model selection criteria). Equivalent to using L_0 penalty with different parameter
  - Instability in the selection process.
  - Not a good idea for high-dimensional data.



Ideal procedure for variable selection

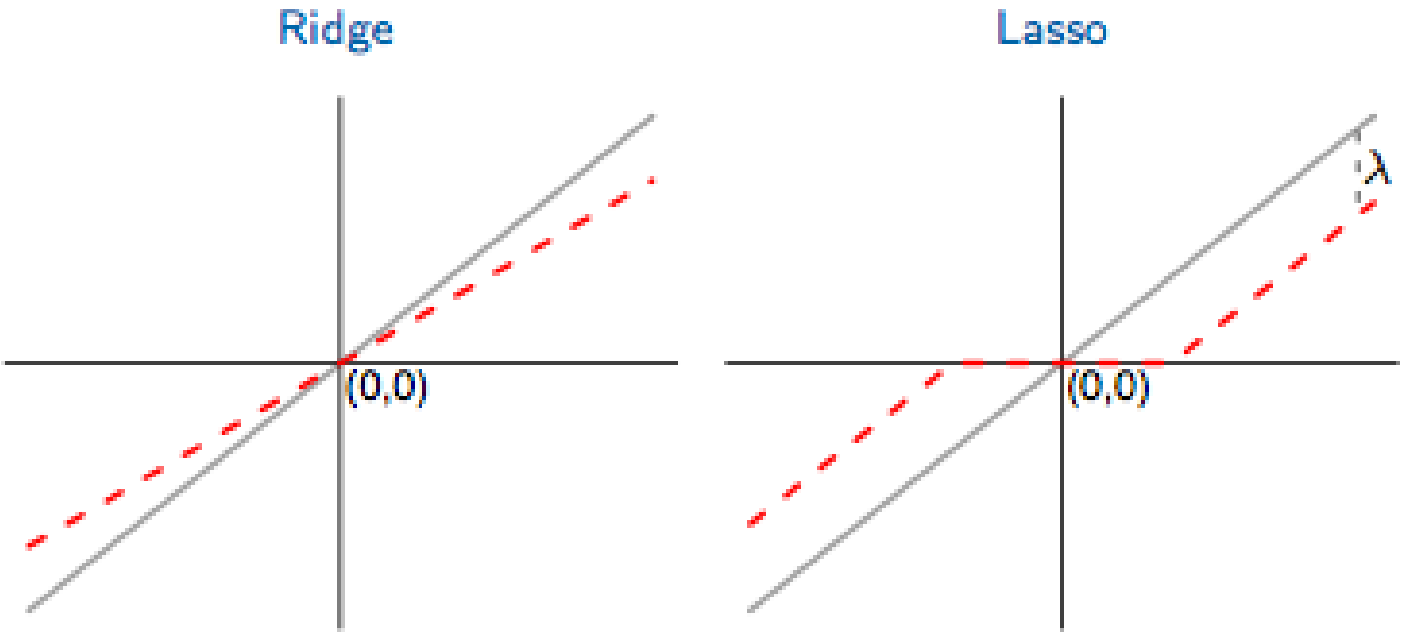
- Unbiasedness: The resulting estimator is nearly unbiasedness when the true unknown parameter is large to avoid excessive estimation bias.
- Sparsity: Estimating a small coefficient as zero, to reduce model complexity.
- Continuity: The resulting estimator is continuous in the data to avoid instability in model prediction.

Previous work

Consider the linear regression model,

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

where X is assumed to be standardized (mean 0, unit variance) and y is centered. Then $\hat{\beta}_{OLS} = X^T y$ is the OLS estimate.



Convex penalties (e.g quadratic penalties) : make trade-off between bias and variance, can create unnecessary biases when the true parameters are large and cannot produce parsimonious models.

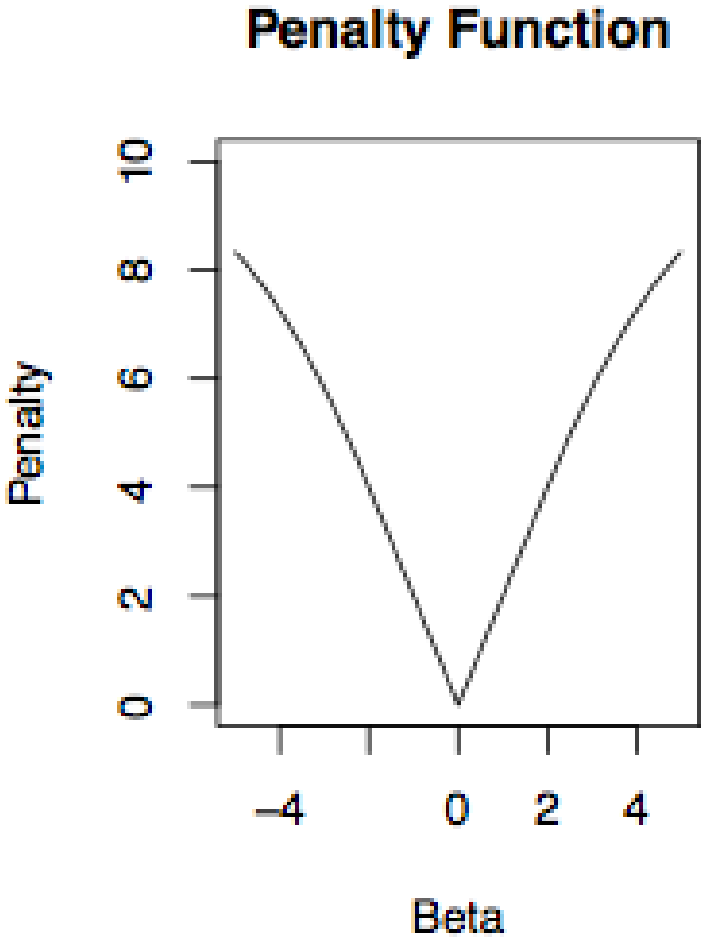
Smoothly Clipped Absolute Deviation (SCAD) Penalty

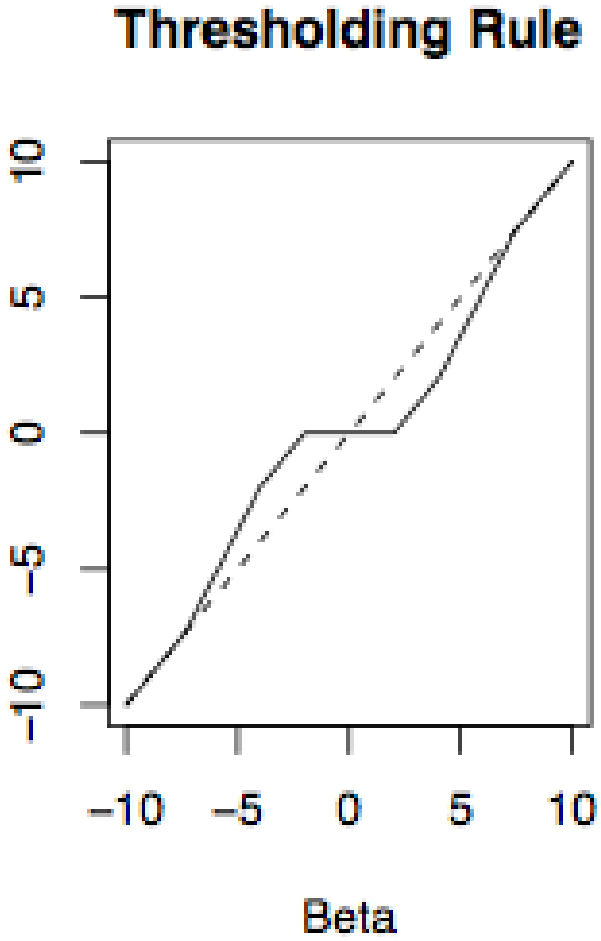
The SCAD penalty function noted $p_\lambda^{SCAD}(\beta)$ is defined by

$$p_\lambda^{SCAD}(\beta) = \begin{cases} \lambda|\beta|; & \text{if } |\beta| \leq \lambda \quad (1) \\ -\dfrac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}; & \text{if} \lambda < |\beta| \leq a\lambda \quad (2) \\ \dfrac{(a+1)\lambda^2}{2}; & \text{if } |\beta| > a\lambda \quad (3) \end{cases}$$

where $a > 2$ and $\lambda > 0$

This corresponds to quadratic spline function with knots at λ and $a\lambda$. The function is continuous and differentiable on $(-\infty, 0) \cap (0, \infty)$. But singular at 0 with its derivatives zero outside the range $[-a\lambda, a\lambda]$





This results in small coefficients being set to zero, a few other coefficients being shrunk towards zero while retaining the large coefficients as they are. Thus, SCAD can produce sparse set of solution and approximately unbiased coefficients for large coefficients.

Solution

The solution to the SCAD penalty can be given as

$$\hat{\beta}_j^{SCAD} = \begin{cases} (|\hat{\beta}_j| - \lambda)_+ sign(\hat{\beta}_j) & \text{if } |\hat{\beta}_j| < 2\lambda; \\ \{(a-1)\hat{\beta}_j - sign(\hat{\beta}_j)a\lambda\}/(a-2) & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda; \\ \hat{\beta}_j & \text{if } |\hat{\beta}_j| > a\lambda \end{cases}$$

This thresholding rule involves two unknown parameters λ and a. Theoretically, the best pair (λ,a) could be obtained using two dimensional grids search using some criteria like cross validation methods. However, such an implementation could be computationally expensive. Based on Bayesian statistical point of view, Fan and Li suggested a=3.7 is a good choice for variance problems.

Simulation

Example Linear Regression. In this example we simulated 100 datasets consisting of n observations from the model

$$Y = x^T\beta + \sigma\epsilon$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and the components of x and ϵ are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = .5$.

Methods	**MRME**	**Correct**	**Incorrect**
	n=40, $\sigma = 3$		
SCAD^{gcv}	0.844	3.18	0.15
SCAD$^{3.7}$	0.832	3.53	0.19
LASSO	0.824	2.74	0.06

Table : Simulation Results for the Linear Regression Model. The average of 0 coefficients is also reported in Table, in which the column labeled Correct presents the average restricted only to the true zero coefficients, and the column labeled Incorrect depicts the average of coefficients erroneously set to 0.

More Simulation

Methods	MRME	Correct	Incorrect
	n=40, $\sigma = 1$		
SCAD^{gcv}	0.396	4.39	0
SCAD$^{3.7}$	0.4401	4.53	0
LASSO	0.739	3.09	0
	n=60, $\sigma = 1$		
SCAD^{gcv}	0.380	4.29	0
SCAD$^{3.7}$	0.323	4.45	0
LASSO	0.691	3.07	0

Table : Simulation Results for the Linear Regression Model

summary

- Compare SCAD and Lasso
  - SCAD is a variable selection method via penalized least squares.
  - SCAD penalty function is specially defined to satisfy three properties for the coefficient estimators: unbiasedness, sparsity and continuity.
- SCAD penalty is not the only one satisfying the three desired properties. e.g. Minimax Concave Penalty (MCP)
- There is no single method that beats all other competitors in all situations.

References

Jianqing FAN and Runze Li (2001)

Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties

Thank You