The Adaptive Lasso and Its Oracle Properties Hui Zou (2006), JASA

Presented by Dongjun Chung

March 12, 2010

Setting

- $y_i = x_i\beta^* + \varepsilon_i$, where $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. mean 0 and variance σ^2.
- $A = \left\{ j : \beta_j^* \neq 0 \right\}$ and $|A| = p_0 < p$.
- $\frac{1}{n} X^T X \to C$, where C is a positive definite matrix.
- $C = \left[\begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array} \right]$, where C_{11} is a $p_0 \times p_0$ matrix.

Definition of Oracle Procedures

We call δ an *oracle* procedure if $\hat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

1. Identifies the right subset model, $\left\{ j : \hat{\beta}_j \neq 0 \right\} = A$.
2. $\sqrt{n}\left(\hat{\beta}(\delta)_A - \beta^*_A\right) \rightarrow_d N\left(0, \Sigma^*\right)$, where Σ^* is the covariance matrix knowing the true subset model.

Definition of LASSO (Tibshirani, 1996)

$$\hat{\beta}^{(n)} = \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|.$$

- λ_n varies with n. $A_n = \left\{ j : \hat{\beta}_j^{(n)} \neq 0 \right\}$.
- LASSO variable selection is consistent iff $\lim_n P\left(A_n = A\right) = 1$.

Proposition 1: Inconsistency of LASSO

Proposition 1

If $\lambda_n/\sqrt{n} \rightarrow \lambda_0 \geq 0$, then $\limsup_n P(A_n = A) \leq c < 1$, where c is a constant depending on the true model.

Theorem 1: Necessary Condition for Consistency of LASSO

Theorem 1

Suppose that $\lim_n P(A_n = A) = 1$. Then there exists some sign vector $s = (s_1, \cdots, s_{p_0})^T$, $s_j = 1$ or -1, such that

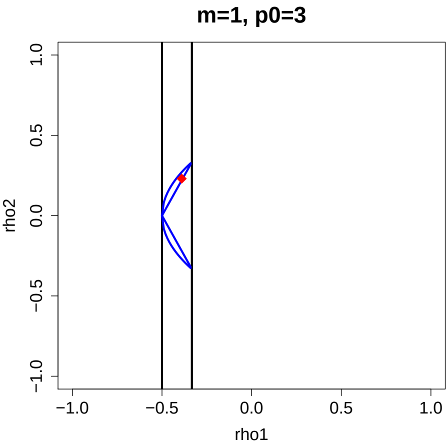
$$\left| C_{21} C_{11}^{-1} s \right| \leq 1. \tag{1}$$

Corollary 1: Interesting Case of Inconsistency of LASSO

Corollary 1

Suppose that $p_0 = 2m + 1 \geq 3$ and $p = p_0 + 1$, so there is one irrelevant predictor. Let $C_{11} = (1 - \rho_1) I + \rho_1 J_1$, where J_1 is the matrix of 1's and $C_{12} = \rho_2 \vec{1}$ and $C_{22} = 1$. If $-\frac{1}{p_0 - 1} < \rho_1 < -\frac{1}{p_0}$ and $1 + (p_0 - 1) \rho_1 < |\rho_2| < \sqrt{(1 + (p_0 - 1) \rho_1) / p_0}$, then condition (1) cannot be satisfied. Thus LASSO variable selection is inconsistent.

Corollary 1: Interesting Case of Inconsistency of LASSO

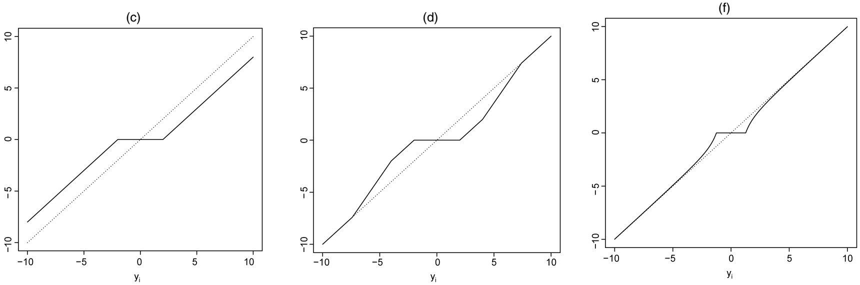


Definition of Adaptive LASSO

$$\hat{\beta}^{*(n)} = \arg\min_{\beta} \left\| y - \sum\nolimits_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left|\beta_j\right|.$$

- weight vector $\hat{w} = 1 / \left|\hat{\beta}\right|^{\gamma}$ (data-dependent) and $\gamma > 0$.
- $\hat{\beta}$ is a root-n-consistent estimator to β^*, e.g. $\hat{\beta} = \hat{\beta}(ols)$.
- $A_n^* = \left\{ j : \hat{\beta}_j^{*(n)} \neq 0 \right\}$.

Penalty Function of LASSO, SCAD and Adaptive LASSO



Remarks: Adaptive LASSO

- The data-dependent $\hat{w}$ is the key for its oracle properties.
- As n grows, the weights for zero-coefficient predictors get inflated, while the weights for nonzero-coefficient predictors converge to a finite constant.
- In the view of Fan and Li, 2001 (presented by Yang Zhao), adaptive lasso satisfies three properties of good penalty function: unbiasedness, sparsity, and continuity.

Theorem 2: Oracle Properties of Adaptive LASSO

Theorem 2

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

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

Computations of Adaptive LASSO

- Adaptive LASSO estimates can be solved by the LARS algorithm (Efron et al., 2004). The entire solution path can be computed at the same order of computation of a single OLS fit.
- Tuning: If we use $\hat{\beta}(ols)$, then use 2-dimensional CV to find an optimal pair of (γ, λ_n). Or use 3-dimensional CV to find an optimal triple $(\hat{\beta}, \gamma, \lambda)$.
- $\hat{\beta}(ridge)$ may be used from the best ridge regression fit when collinearity is a concern.

Definition of Nonnegative Garrote (Breiman, 1995)

$\hat{\beta}_j\,(garrote) = c_j \hat{\beta}_j\,(ols)$, where a set of nonnegative scaling factor $\{c_j\}$ is to minimize

$$\left\| y - \sum\nolimits_{j=1}^{p} x_j \hat{\beta}_j\,(ols)\, c_j \right\|^2 + \lambda_n \sum\nolimits_{j=1}^{p} c_j,$$

subject to $c_j \geq 0, \forall j$.

- A sufficiently large λ_n shrinks some c_j to exact 0, i.e. $\hat{\beta}_j\,(garrote) = 0$.
- Yuan and Lin (2007) also studied the consistency of the nonnegative garrote.

Garrote: Adaptive LASSO Formulation and Consistency

Adaptive LASSO Formulation

$$\hat{\beta}\,(garrote) = \arg\min_{\beta} \left\| y - \sum\nolimits_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum\nolimits_{j=1}^{p} \hat{w}_j \left|\beta_j\right|$$

subject to $\beta_j \hat{\beta}_j\,(ols) \geq 0, \forall j$, where $\gamma = 1$, $\hat{w} = 1/\left|\hat{\beta}\,(ols)\right|$.

Corollary 2: Consistency of Nonnegative Garrote

If we choose a λ_n such that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, then nonnegative garrote is consistent for variable selection.

Adaptive LASSO for GLM

$$\hat{\beta}^{*(n)}(glm) = \arg\min_{\beta} \sum \left(-y_i \left(x_i^T \beta \right) + \phi \left(x_i^T \beta \right) \right) + \lambda_n \sum_{j=1}^{p} \hat{w}_j \, |\beta_j|.$$

- weight vector $\hat{w} = 1 / \left| \hat{\beta}(mle) \right|^{\gamma}$ for some $\gamma > 0$.
- $f(y|x, \theta) = h(y) \exp(y\theta - \phi(\theta))$, where $\theta = x^T \beta^*$.
- The Fisher information matrix $I(\beta^*) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$, where I_{11} is a $p_0 \times p_0$ matrix. Then I_{11} is the Fisher information matrix with the true submodel known.

Theorem 4: Oracle Properties of Adaptive LASSO for GLM

Theorem 4

Let $A_n^* = \left\{ j : \hat{\beta}_j^{*(n)}(glm) \neq 0 \right\}$. Suppose that $\lambda_n / \sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then, under some mild regularity conditions, the adaptive LASSO estimate $\hat{\beta}^{*(n)}(glm)$ must satisfy the following:

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

Experiments for Inconsistency of LASSO

Setting

We let $y = x^T\beta + N(0, \sigma^2)$, where the true regression coefficients are $\beta = (5.6, 5.6, 5.6, 0)$. The predictors $x_i (i = 1, \cdots, n)$ are i.i.d. $N(0, C)$, where C is the C matrix in Corollary 1 with $\rho_1 = -.39$ and $\rho_2 = .23$ (red point).

Experiments for Inconsistency of LASSO

Table 1. Simulation Model 0: The Probability of Containing the True Model in the Solution Path

	$n = 60, \sigma = 9$	$n = 120, \sigma = 5$	$n = 300, \sigma = 3$
lasso	.55	.51	.53
adalasso($\gamma = .5$)	.59	.68	.93
adalasso($\gamma = 1$)	.67	.89	1
adalasso($\gamma = 2$)	.73	.97	1
adalasso(γ by cv)	.67	.91	1

NOTE: In this table "adalasso" is the adaptive lasso, and "γ by cv" means that γ was selected by five-fold cross-validation from three choices: $\gamma = .5$, $\gamma = 1$, and $\gamma = 2$.

General Observations

- Comparison: LASSO, Adaptive LASSO, SCAD, and nonnegative garrote.
- $p = 8$ and $p_0 = 3$. Consider a few large effects ($n = 20, 60$) and many small effects ($n = 40, 80$).
- LASSO performs best when the SNR is low.
- Adaptive LASSO, SCAD, and and nonnegative garrote outperforms LASSO with a medium or low level of SNR.
- Adaptive LASSO tends to be more stable than SCAD.
- LASSO tends to select noise variables more often than other methods.

Theorem 2: Oracle Properties of Adaptive LASSO

Theorem 2

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

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

Proof of Theorem 2: Asymptotic Normality

Let $\beta = \beta^* + u/\sqrt{n}$ and

$$\Psi_n(u) = \left\| y - \sum_{j=1}^{p} x_j \left(\beta_j^* + \frac{u_n}{\sqrt{n}} \right) \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left| \beta_j^* + \frac{u_n}{\sqrt{n}} \right|.$$

Let $\hat{u}^{(n)} = \arg\min \Psi_n(u)$; then $\hat{u}^{(n)} = \sqrt{n} \left(\hat{\beta}^{*(n)} - \beta^* \right)$.

$\Psi_n(u) - \Psi_n(0) = V_4^{(n)}(u)$, where

$$V_4^{(n)}(u) = u^T \left(\tfrac{1}{n} X^T X \right) u - 2 \frac{\varepsilon^T X}{\sqrt{n}} u$$

$$+ \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \hat{w}_j \sqrt{n} \left(\left| \beta_j^* + \frac{u_n}{\sqrt{n}} \right| - \left| \beta_j^* \right| \right)$$

Proof of Theorem 2: Asymptotic Normality (conti.)

Then, $V_4^{(n)}(u) \rightarrow_d V_4(u)$ for every u, where

$$V_4(u) = \begin{cases} u_A^T C_{11} u_A - 2u_A^T W_A & \text{if} \quad u_j = 0, \forall j \notin A \\ \infty & \text{otherwise} \end{cases}$$

and $W_A = N\left(0, \sigma^2 C_{11}\right)$. $V_4^{(n)}$ is convex, and the unique minimum of V_4 is $\left(C_{11}^{-1} W_A, 0\right)^T$. Following the epi-convergence results of Geyer (1994), we have $\hat{u}_A^{(n)} \rightarrow_d C_{11}^{-1} W_A$ and $\hat{u}_{A^C}^{(n)} \rightarrow_d 0$. Hence, we prove the asymptotic normality part.

Proof of Theorem 2: Consistency

The asymptotic normality result indicates that $\forall j \in A$, $\hat{\beta}_j^{*(n)} \rightarrow_p \beta_j^*$; thus $P\left(j \in A_n^*\right) \rightarrow 1$. Then it suffices to show that $\forall j' \notin A, P\left(j' \in A_n^*\right) \rightarrow 0$. Consider the event $j' \in A_n^*$. By the KKT optimality conditions, $2x_{j'}^T\left(y - X\hat{\beta}^{*(n)}\right) = \lambda_n \hat{w}_{j'}$.

$\lambda_n \hat{w}_{j'}/\sqrt{n} = \lambda_n n^{(\gamma-1)/2}/\left|\sqrt{n}\hat{\beta}_{j'}\right|^\gamma \rightarrow_p \infty$ and

$2\frac{x_{j'}^T\left(y - X\hat{\beta}^{*(n)}\right)}{\sqrt{n}} = 2\frac{x_{j'}^T X\sqrt{n}\left(\beta^* - \hat{\beta}^{*(n)}\right)}{n} + 2\frac{x_{j'}^T\varepsilon}{\sqrt{n}}$ and each of these two terms converges to some normal distribution. Thus

$$P\left(j' \in A_n^*\right) \leq P\left(2x_{j'}^T\left(y - X\hat{\beta}^{*(n)}\right) = \lambda_n \hat{w}_{j'}\right) \rightarrow 0.$$

Corollary 2: Consistency of Nonnegative Garrote

Adaptive LASSO Formulation

$$\hat{\beta}(garrote) = \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|$$

subject to $\beta_j \hat{\beta}_j(ols) \geq 0, \forall j$, where $\gamma = 1$, $\hat{w} = 1/\left|\hat{\beta}(ols)\right|$.

Corollary 2: Consistency of Nonnegative Garrote

If we choose a λ_n such that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$, then nonnegative garrote is consistent for variable selection.

Proof of Corollary 2

Let $\hat{\beta}^{*(n)}$ be the adaptive LASSO estimates. By Theorem 2, $\hat{\beta}^{*(n)}$ is an oracle estimator if $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n \to \infty$. To show the consistency, it suffices to show that $\hat{\beta}^{*(n)}$ satisfies the sign constraint with probability tending to 1. Pick any j. If $j \in A$, then $\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(ols)_j \to_p \left(\beta_j^*\right)^2 > 0$. If $j \notin A$, then $P\left(\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(ols)_j \geq 0\right) \geq P\left(\hat{\beta}^{*(n)}(\gamma = 1)_j = 0\right) \to 1$. In either case, $P\left(\hat{\beta}^{*(n)}(\gamma = 1)_j \hat{\beta}(ols)_j \geq 0\right) \to 1$ for any $j = 1, 2, \cdots, p$.

Theorem 4: Oracle Properties of Adaptive LASSO for GLM

Theorem 4

Let $A_n^* = \left\{ j : \hat{\beta}_j^{*(n)} (glm) \neq 0 \right\}$. Suppose that $\lambda_n / \sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then, under some mild regularity conditions, the adaptive LASSO estimate $\hat{\beta}^{*(n)} (glm)$ must satisfy the following:

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

- $f\left(y|x,\theta\right) = h\left(y\right) \exp\left(y\theta - \phi\left(\theta\right)\right)$, where $\theta = x^T \beta^*$.

Theorem 4: Regularity Conditions

1. The Fisher information matrix is finite and positive definite,

$$I\left(\beta^*\right) = E\left[\phi''\left(x^T\beta^*\right)xx^T\right].$$

2. There is a sufficiently large enough open set O that contains β^* such that $\forall\beta \in O$,

$$\left|\phi'''\left(x^T\beta\right)\right| \leq M\left(x\right) < \infty$$

and

$$E\left[M\left(x\right)\left|x_j x_k x_l\right|\right] < \infty$$

for all $1 \leq j, k, l \leq p$.

Proof of Theorem 4: Asymptotic Normality

Let $\beta = \beta^* + u/\sqrt{n}$. Define
$\Gamma_n(u) = \sum_{i=1}^n \left\{ -y_i \left(x_i^T \left(\beta^* + u/\sqrt{n} \right) \right) + \phi \left(x_i^T \left(\beta^* + u/\sqrt{n} \right) \right) \right\}$

$$+\lambda_n \sum_{j=1}^p \left| \beta_j^* + u_j/\sqrt{n} \right|$$

Let $\hat{u}^{(n)} = \arg\min_u \Gamma_n(u)$; then $\hat{u}^{(n)} = \sqrt{n}\left(\beta^{*(n)}(glm) - \beta^* \right)$.
Using the Taylor expansion, we have $\Gamma_n(u) - \Gamma_n(0) = H^{(n)}(u)$, where $H^{(n)}(u) = A_1^{(n)} + A_2^{(n)} + A_3^{(n)} + A_4^{(n)}$, with

$$A_1^{(n)} = -\sum_{i=1}^n \left[y_i - \phi' \left(x_i^T \beta^* \right) \right] \frac{x_i^T u}{\sqrt{n}},$$
$$A_2^{(n)} = \sum_{i=1}^n \frac{1}{2} \phi'' \left(x_i^T \beta^* \right) u^T \frac{x_i x_i^T}{n} u,$$
$$A_3^{(n)} = \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^p \hat{w}_j \sqrt{n} \left(\left| \beta_j^* + \frac{u_n}{\sqrt{n}} \right| - \left| \beta_j^* \right| \right),$$

Proof of Theorem 4: Asymptotic Normality (conti.)

and $A_4^{(n)} = n^{-3/2} \sum_{i=1}^{n} \frac{1}{6} \phi''' \left(x_i^T \tilde{\beta}_* \right) \left(x_i^T u \right)^3$, where $\tilde{\beta}^*$ is between β^* and $\beta^* + u/\sqrt{n}$. Then, by the regularity condition 1 and 2, $H^{(n)}(u) \rightarrow_d H(u)$ for every u, where

$$H(u) = \begin{cases} u_A^T I_{11} u_A - 2 u_A^T W_A & \text{if } \; u_j = 0, \forall j \notin A \\ \infty & \text{otherwise} \end{cases}$$

and $W_A = N(0, I_{11})$. $H^{(n)}$ is convex, and the unique minimum of H is $\left(I_{11}^{-1} W_A, 0 \right)^T$. Following the epi-convergence results of Geyer (1994), we have $\hat{u}_A^{(n)} \rightarrow_d I_{11}^{-1} W_A$ and $\hat{u}_{A^C}^{(n)} \rightarrow_d 0$, and the asymptotic normality part is proven.

Proof of Theorem 4: Consistency

The asymptotic normality result indicates that $j \in A, P\left(j \in A_n^*\right) \to 1$. Then it suffices to show that $j' \notin A, P\left(j' \in A_n^*\right) \to 0$. Consider the event $j' \in A_n^*$. By the KKT optimality conditions,

$\sum_{i=1}^n x_{ij'}\left(y_i - \phi'\left(x_i^T \hat{\beta}^{*(n)}(glm)\right)\right) = \lambda_n \hat{w}_{j'}$.

$$\sum_{i=1}^n x_{ij'}\left(y_i - \phi'\left(x_i^T \hat{\beta}^{*(n)}(glm)\right)\right)/\sqrt{n} = B_1^{(n)} + B_2^{(n)} + B_3^{(n)}$$

with

$$B_1^{(n)} = \sum_{i=1}^n x_{ij'}\left(y_i - \phi'\left(x_i^T \beta^*\right)\right)/\sqrt{n},$$
$$B_2^{(n)} = \left(\frac{1}{n}\sum_{i=1}^n x_{ij'}\phi''\left(x_i^T \beta^*\right) x_i^T\right)\sqrt{n}\left(\beta^* - \hat{\beta}^{*(n)}(glm)\right),$$
$$B_3^{(n)} = \left(\frac{1}{n}\sum_{i=1}^n x_{ij'}\phi'''\left(x_i^T \tilde{\beta}_{**}\right)\right)\left(x_i^T\sqrt{n}\left(\beta^* - \hat{\beta}^{*(n)}(glm)\right)\right)^2/\sqrt{n},$$

where $\tilde{\beta}_{**}$ is between $\hat{\beta}^{*(n)}(glm)$ and β^*.

Proof of Theorem 4: Consistency (conti.)

$B_1^{(n)}$ and $B_2^{(n)}$ converge to some normal distributions and $B_3^{(n)} = O_p(1/\sqrt{n})$.

$\lambda_n \hat{w}_{j'}/\sqrt{n} = \lambda_n n^{(\gamma-1)/2} / \left| \sqrt{n} \hat{\beta}_{j'}(glm) \right|^{\gamma} \rightarrow_p \infty$. Thus

$$P\left(j' \in A_n^*\right) \leq P(\sum_{i=1}^n x_{ij'} \left(y_i - \phi' \left(x_i^T \hat{\beta}^{*(n)}(glm) \right) \right) = \lambda_n \hat{w}_{j'}) \rightarrow 0.$$

and this completes the proof.