Variable selection consistency of the lasso estimator

This note is an informal illustration of the theoretical foundations for the variable selection consistency of lasso. The goal is to explain in simple terms and simplified examples the intuitions behind standard conditions used in this context.

Settings

We start with the simple linear regression problem

$$Y = \beta^{(1)} X^{(1)} + \beta^{(2)} X^{(2)} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

and assume that the data is generated using the "true" vector of parameters $\beta^* = (\beta^{(1)*}, 0)$. Without loss of generality, we assume that $E[X^{(1)}] = E[X^{(2)}] = 0$.

Now, assume that we observe a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We use the same notations as in the previous lectures

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} \\ \cdots & \cdots \\ x_n^{(1)} & x_n^{(2)} \end{bmatrix}$$

The lasso estimator solves the optimization problem

$$\hat{\beta} = \min_{\beta} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(|\beta^{(1)}| + |\beta^{(2)}|). \tag{1}$$

In this note, we want to investigate the conditions under which we can verify that

$$sign(\hat{\beta}^{(1)}) = sign(\beta^{(1)*}) \quad \text{and} \quad \hat{\beta}^{(2)} = 0$$

Sub-gradient and lasso solution

Since the penalty of lasso is non-differentiable, we need a way to circumvent that issue.

Definition 0.1. *We say that a vector $s \in \mathbb{R}^k$ is a subgradient for the ℓ_1-norm evaluated at $\beta \in \mathbb{R}^k$, written as $s \in \partial\|\beta\|$ if for $i = 1, \ldots, k$ we have*

$$s_i = sign(\beta_i) \quad \text{if } \beta_i \neq 0 \quad \text{and } s_i \in [-1, 1] \quad \text{otherwise.}$$

We then have the following theorem

Theorem 1. *(a) A vector $\hat{\beta}$ minimizes the problem (1) if and only if there exists a $\hat{z} \in \partial\|\hat{\beta}\|$ such that*

$$X^T(Y - X\hat{\beta}) - \lambda\hat{z} = 0 \tag{2}$$

(b) Suppose that the subgradient vector satisfies the strict dual feasibility condition

$$|\hat{z}_2| < 1$$

*then **any** lasso solution $\tilde{\beta}$ satisfies $\tilde{\beta}^{(2)} = 0$.*

(c) Under the condition of part (b), if $X^{(1)} \neq 0$, then $\hat{\beta}$ is the unique lasso solution.

The primal-dual witness method.

The primal-dual witness (PDW) method consists of constructing a pair of $(\tilde{\beta}, \tilde{z})$ according to the following steps:

- First, we obtain $\tilde{\beta}^{(1)}$ by solving the restricted lasso problem

$$\tilde{\beta}^{(1)} = \min_{\beta=(\beta^{(1)},0)} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(|\beta^{(1)}|).$$

  Choose a subgradient $\tilde{z}_1 \in \mathbb{R}$ for the ℓ_1-norm evaluated at $\tilde{\beta}^{(1)}$
- Second, we solve for a vector $\tilde{z}_2$ satisfying equation (2), and check whether or not the dual feasibility condition $|\tilde{z}_2| < 1$ is satisfied
- Third, we check whether the *sign consistency condition*

$$\tilde{z}_1 = sign(\beta^{(1)*})$$

  is satisfied.

Note: this procedure is not a practical method for solving the ℓ_1-regularized optimization problem, since solving the restricted problem in Step 1 requires knowledge that the second component of the true parameter is 0. Rather, the utility of this constructive procedure is as a proof technique: it succeeds if and only if the Lasso has a optimal solution with the correct signed support.

A more detailed computation

We note that the matrix form of equation (2) can be written as

$$[X^{(1)}]^T(Y - X^{(1)}\beta^{(1)} - X^{(2)}\beta^{(2)}) - \lambda\hat{z}_1 = 0$$
$$[X^{(2)}]^T(Y - X^{(1)}\beta^{(1)} - X^{(2)}\beta^{(2)}) - \lambda\hat{z}_2 = 0$$

To simplify the notation, we denote

$$C_{ij} = [X^{(i)}]^T[X^{(j)}]$$

- In Step 1, we find $\tilde{\beta}^{(1)}$ and $\tilde{z}_1$ that satisfies

$$[X^{(1)}]^T(Y - X^{(1)}\tilde{\beta}^{(1)}) - \lambda\tilde{z}_1 = 0$$

  Moreover, to make sure that the sign consistency in Step 3 is satisfied, we impose that

$$\tilde{z}_1 = sign(\beta^{(1)*}) \quad \text{and} \quad \tilde{\beta}^{(1)} = C_{11}^{-1}([X^{(1)}]^T Y - \lambda sign(\beta^{(1)*})).$$

  This is acceptable as long as $\tilde{z}_1 \in \partial|\tilde{\beta}^{(1)}|$. That is,

$$sign(\tilde{\beta}^{(1)}) = sign(\beta^{(1)*})$$

- Step 2: We choose

$$\tilde{z}_2 = \frac{1}{\lambda}[X^{(2)}]^T(Y - X^{(1)}\tilde{\beta}^{(1)}).$$

  We want $|\tilde{z}_2| < 1$.

In principle, we want two conditions: $sign(\tilde{\beta}^{(1)}) = sign(\beta^{(1)*})$ and $|\tilde{z}_2| < 1$.
Recalling that $Y = X^{(1)}\beta_1^* + \epsilon$, we have

$$\begin{aligned}\tilde{\beta}^{(1)} &= C_{11}^{-1}([X^{(1)}]^T(X^{(1)}\beta_1^* + \epsilon) - \lambda sign(\beta^{(1)*})) \\ &= \beta_1^* + C_{11}^{-1}([X^{(1)}]^T\epsilon - \lambda sign(\beta^{(1)*})))\end{aligned}$$

Thus if we denote

$$\Delta = C_{11}^{-1}([X^{(1)}]^T\epsilon - \lambda sign(\beta^{(1)*})))$$

then the first condition can be further simplified as $sign(\beta_1^*) = sign(\beta_1^* + \Delta)$.
Similarly,

$$\begin{aligned}\tilde{z}_2 &= \frac{1}{\lambda}[X^{(2)}]^T(X^{(1)}\beta_1^* + \epsilon - X^{(1)}\tilde{\beta}^{(1)}) \\ &= \frac{1}{\lambda}[X^{(2)}]^T(X^{(1)}\Delta + \epsilon)\end{aligned}$$

Zero-noise setting

To further simplify the setting, we assume that the observations are collected with no noise ($\epsilon = 0$). Then

$$\Delta = -C_{11}^{-1}\lambda sign(\beta^{(1)*})$$

$$\tilde{z}_2 = \frac{-1}{\lambda}C_{21}\Delta = C_{21}C_{11}^{-1}sign(\beta^{(1)*})$$

Condition

- Mutual incoherence: $|C_{21}C_{11}^{-1}| < 1$.
- Minimum signal: $|\beta^{(1)*}| > \lambda C_{11}^{-1}$

Conditions for model consistency: the noisy case

In the noisy case, the conditions are more complicated. Note that

$$|\Delta| \leq \lambda|C_{21}C_{11}^{-1}| + |C_{11}^{-1}[X^{(1)}]^T\epsilon|$$

Similarly,

$$\begin{aligned}\tilde{z}_2 &= \frac{1}{\lambda}[X^{(2)}]^T(X^{(1)}\Delta + \epsilon) \\ &= \frac{1}{\lambda}[C_{21}\Delta + [X^{(2)}]^T\epsilon] \\ &= \frac{1}{\lambda}[X^{(2)}]^T(X^{(1)}C_{11}^{-1}[X^{(1)}]^T\epsilon - \lambda[X^{(1)}]C_{11}^{-1}sign(\beta^{(1)*})) + \epsilon)\end{aligned}$$