STATS 305B: Methods for Applied Statistics I **Winter 2016/17**

Lecture 22: March 3

Lecturer: Jonathan Taylor *Scribes: Kenneth Tay*

22.1 Model Selection Consistency for LASSO

When using the LASSO (or any model selection method), we can ask the question: **when does the LASSO choose the correct set of predictors with the correct signs for the coefficients?**

Assume that we have p predictor variables, and that the true model is

$$Y = X_A\beta_A^* + \varepsilon,$$

where $A \in \{1, \dots, p\}$ with A sparse, and where ε represents the noise in the process. (Here, X_A means X subsetted for indices in A.) Let $s_A = \text{sign}(\beta_A^*)$, i.e. the signs of the non-zero coefficients of the true model. We may then reformulate the question in this setting:

Find conditions such that with high probability, the active set E which the LASSO selects is correct (i.e. $E = A$), and the signs of the active set are correct (i.e. $s_E = \textbf{sign}(\hat{\beta}_E) = s_A$).

We know that the LASSO solution satisfies the KKT conditions, so if the LASSO solution is to be the correct one, then the correct one must satisfy the KKT conditions too.

22.1.1 Conditions for Active Block

Recall the KKT conditions for the active block:

$$X_E^T \left[Y - X_E\hat{\beta}_E \right] = \lambda s_E.$$

Thus, in order for the correct active set A to be selected with the correct signs, we must have

$$\begin{aligned} X_A^T \left[Y - X_A\hat{\beta}_A \right] &= \lambda s_A, \\ X_A^T \left[\varepsilon + X_A\beta_A^* - X_A\hat{\beta}_A \right] &= \lambda s_A, \\ \hat{\beta}_A &= \beta_A^* + \left(X_A^T X_A\right)^{-1} X_A^T\varepsilon - \lambda \left(X_A^T X_A\right)^{-1} s_A. \end{aligned}$$

Note in the above that only ε is random. Thus, in order for the LASSO to select the correct active set A with the correct signs, we need

$$\begin{aligned} \text{sign}\left(\hat{\beta}_A\right) &= s_A, \\ \text{sign}\left(\beta_A^* + \left(X_A^T X_A\right)^{-1} X_A^T\varepsilon - \lambda \left(X_A^T X_A\right)^{-1} s_A\right) &= s_A. \end{aligned}$$

This will happen with high probability if $\beta_A^* >> \lambda \left(X_A^T X_A\right)^{-1} s_A$.

22.1.2 Conditions for Inactive Block

On the inactive block, we must have

$$\begin{aligned}\lambda &\geq \|X_{-A}^T(Y - X_A\hat{\beta}_A)\|_\infty \\ &= \left\|X_{-A}^T\left[Y - X_A\beta_A^* - X_A\left(X_A^TX_A\right)^{-1}X_A^T\varepsilon + \lambda X_A\left(X_A^TX_A\right)^{-1}s_A\right]\right\|_\infty \\ &= \left\|X_{-A}^T\left[(I-P_A)\varepsilon + \lambda X_A\left(X_A^TX_A\right)^{-1}s_A\right]\right\|_\infty . \qquad (P_A = X_A\left(X_A^TX_A\right)^{-1}X_A^T)\end{aligned}$$

Note that for a given ε, we can choose λ large enough so that the first term is basically negligible. To bound the second term, we need

$$\|X_{-A}^TX_A\left(X_A^TX_A\right)^{-1}s_A\|_\infty \leq 1.$$

This is a special case of what we call the **irrepresentable condition**, i.e.

$$\|X_{-A}^TX_A\left(X_A^TX_A\right)^{-1}u_A\|_\infty \leq 1$$

for all u_A such that $\|u_A\|_\infty \leq 1$.

Thus, if the model satisfies the irrepresentable condition, choose λ big enough so that the KKT conditions on the inactive block hold. If, for this value of λ, β_A^* satisfies the condition for the active block, then we can say that the LASSO selects the correct active set with the correct signs with high probability.

In practice, it is difficult to check the irrepresentable condition. A special case where it holds is when $X_{-A}^TX_A = 0$.

22.1.3 Model misspecification

What can we do if the assumptions of the model are suspect/difficult to verify?

Assume that $(X_i, Y_i) \overset{iid}{\sim} F$. For every $E \subseteq \{1, \ldots, p\}$, it still makes sense to consider the expression

$$\begin{aligned}\beta_E(F) &= \text{“ population least squares”} \\ &= \mathbb{E}_F\left[X_EX_E^T\right]^{-1}\mathbb{E}_F\left[X_E^TY\right].\end{aligned}$$