Computing the null distribution for best-subset

Charles Zheng and Trevor Hastie

September 12, 2016

1 Introduction

Given $n \times p$ design matrix X and data y, the *best k-subset* procedure finds a subset $S \subset \{1, .., p\}$ of size k which maximizes the coefficient of determination, R^2:

$$R^2(S) := \frac{||P_S y||^2}{||y||^2}$$

where

$$P_S = X_S(X_S^T X_S)^{-1} X_S^T$$

and X_S is the submatrix X with columns indexed by S.

The best k-subset coefficient of determination is defined

$$R_k^2 = \sup_{|S|=k} R^2(S).$$

In this work we consider computing the *null distribution* of R_k^2 under the null hypothesis that the data is pure Gaussian noise: $y \sim N(0, \sigma^2 I)$.

A potential application of this work is testing the null hypothesis versus the alternative hypothesis that the data was generated from a sparse linear model.

2 Intersection-Union method

Define Q_S as the Q matrix obtained from the QR-decomposition of X_S. We have

$$||P_S y||^2 = ||Q_S y||^2$$

so we can also write

$$R^2(S) = \frac{||Q_S y||^2}{||y||^2}.$$

Let $\mathcal{S}$ denote a set of subsets of $\{1, ..., p\}$. For instance, for best-k subset, we would take

$$\mathcal{S} = \{S \subset \{1, ..., p\} : |S| = k\}$$

but the theory may also be applied to more general families of subsets.

Define

$$R^2(\mathcal{S}) = \max_{S \in \mathcal{S}} R^2(S) = \max_{S \in \mathcal{S}} R^2(S) \frac{||Q_S y||^2}{||y||^2}.$$

We would like to compute the exceedence probability $\Pr[R^2 \geq \tau]$ when $y \sim N(0, I)$, for arbitrary $\tau \in [0, 1]$.

The intersection-union formula gives

$$\begin{aligned}
\Pr[R^(\mathcal{S}) \geq \tau] &= \Pr[\cup_{S \in \mathcal{S}} R^2(S) \geq \tau] \\
&= \sum_{j=1}^{|\mathcal{S}|} (-1)^{j+1} \sum_{S_1 \neq \ldots \neq S_j \in \mathcal{S}} \Pr[\min_i R^2(S_i) \geq \tau] \\
&= \sum_{S \in |S|} \Pr[R^2(S) \geq \tau] - \sum_{S_1 \neq S_2} \Pr[R^2(S_1) \vee R^2(S_2) \geq \tau] + \ldots
\end{aligned}$$