Lecture 9 — April 26, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Steve Gan*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by Michael Celentano, Zijun Gao, Shuangning Li, and Basil Saeed.*

9.1 Introduction

Classically, statistical inference involves formulating hypotheses before collecting and analyzing data. Nowadays, it has been more common that large data sets are collected before hypotheses are formulated. High throughput data collection technologies have also enabled hundreds of thousands of variables to be studied at the same time. We need to take precaution to ensure valid inference and control false positives. This new paradigm is related to the recent replicability crisis in the scientific community. For example, Amgen could only replicate 6 out of 53 studies that were considered landmarks in haematology and oncology. Bayer Heathcare reported in 2011 that only about 25% of published preclinical studies could be validated [1].

As statisticians, we need to develop methodologies that can help improve research replicability.

9.2 The Selection Problem

The problem we shall consider here is the following: we have a response Y, which potentially depends on thousands or millions of covariates $X_1, \ldots, X_n$. We want to select a subset of "interesting" variables. For example, in genome-wide association study (GWAS), we may be interested in selecting genes X_j that a phenotype Y of interest truly depends on. To be precise, we will define a null variable to be one for which

$$Y \perp\!\!\!\perp X_j | X_{-j},$$

that is, X_j is independent of Y given all the other covariates, where we use $X_{-j} := (X_i : i \neq j)$. For example, suppose we have

$$Y = \sum_{j=1}^{n} \beta_j X_j + \epsilon$$

for $\epsilon \sim N(0, \sigma^2)$, then the null variables correspond to $\{X_j : \beta_j = 0\}$. If we have a graphical model, this problem is equivalent to finding neighbors (Markov blanket) of Y. In Figure 9.3, the blue nodes X_5 and X_{164} form the Markov blanket of Y and therefore are the only non-null variables.

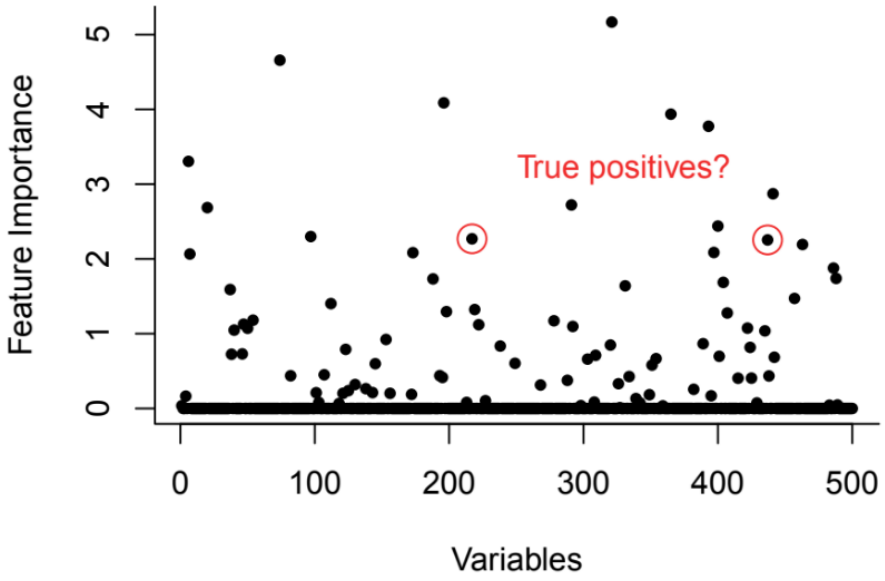


Figure 9.1

Notice that this is a stronger notion of null variable than the one used to test marginal associations. In Figure 9.3, node X_{52} is marginally associated with Y, but it is independent of Y given X_5. We would like a method which selects interesting variables without selecting too many false positives.

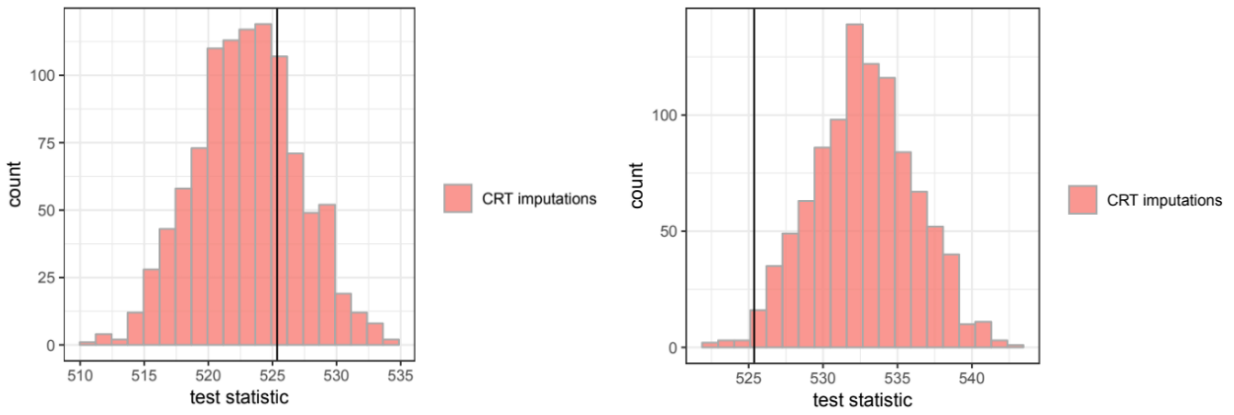
Many variable selection methods compute an importance statistic for each covariate. These statistics serve as a basis for deciding whether to include the variable in our model. For example, the feature importance statistic may be the magnitude of a coefficient computed with the Lasso, the point at which a variable enters the Lasso path, or even a more complicated feature importance statistics computed with random forests or neural nets.

In Figure 9.1, we observe an empirical distribution of feature importance scores after fitting our procedure to the data set, say Lasso for concreteness, so that the feature importance for each covariate is the magnitude of the coefficient in a Lasso solution. Given these scores, how do we decide which of the resulting scores should be regarded as "significant"?

It is difficult to perform inference and obtain p-values using these selection methods since sampling distributions of feature importance statistics are extremely complicated or unknown.

9.3 Conditional Randomization Testing

One solution to the above problem is Conditional Randomization Testing (CRT), which proceeds as follows: suppose we know the conditional distribution of $X_j|X_{-j} = x_{-j}$. Then we can sample a synthetic null $\tilde{X}_j$ from this conditional. If X_j is indeed null, then we would



(a) (b)

Figure 9.2: We compare t_j^* to the empirical distribution of $\{t_j^{(k)} : k \in [K]\}$ and reject the null $H_{j,0} : X_j \perp\!\!\!\perp Y | X_{-j}$ when t_j^* is extreme as in figure (b).

have

$$\begin{aligned}
\mathbb{P}(X_j, X_{-j}, Y) &= \mathbb{P}(X_j | X_{-j}, Y)\mathbb{P}(X_{-j}, Y) && (9.1)\\
&= \mathbb{P}(X_j | X_{-j})\mathbb{P}(X_{-j}, Y) && (9.2)\\
&= \mathbb{P}(\tilde{X}_j | X_{-j})\mathbb{P}(X_{-j}, Y) && (9.3)\\
&= \mathbb{P}(\tilde{X}_j, X_{-j}, Y). && (9.4)
\end{aligned}$$

So given this new sample $\tilde{X}_j$, we can test

$$(X_1, \dots X_j, \dots, X_n, Y) \stackrel{d}{=} (X_1, \dots, \tilde{X}_j, \dots, X_n, Y) \tag{9.5}$$

to decide if X_j is null. In other words, we test whether replacing X_j with $\tilde{X}_j$, which is independent of Y, changes the joint distribution. Any asymmetry implies X_j and Y are directly connected.

Formally, we have the following procedure for obtaining the p-value to test if the variable X_j is null with a feature importance score $T(\cdot)$:

1. Compute a score $t_j^* := T(X_j, X_{-j}, Y)$.
2. For $k \in \{1, \dots, K\}$:
   - sample $\tilde{X}_j^{(k)} \sim X_j | X_{-j} = x_{-j}$; and
   - compute the score $t_j^{(k)} = T(\tilde{X}_j^{(k)}, X_{-j}, Y)$
3. Compute

$$p_j := \frac{1 + |\{k : t_j^* \le t_j^{(k)}\}|}{1 + K}.$$

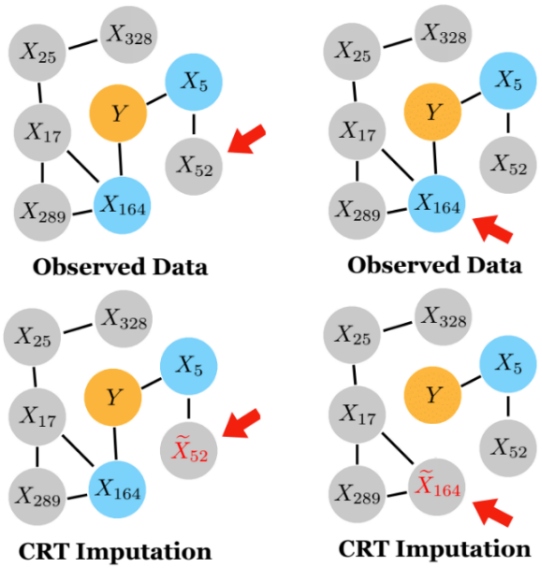


Figure 9.3: Graphical model illustration of CRT.

Note that if X_j is indeed null, then the resulting p_j is indeed a p-value. This follows since under the null, equation (9.5) would hold and hence $t_j^* \overset{d}{=} t_j^{(1)} \overset{d}{=}, \ldots, \overset{d}{=} t_j^{(K)}$, so p_j is uniform on $\{\frac{1}{1+K}, \ldots, 1\}$. Figure 9.2 gives an graphical illustration.

CRT can be visualized by considering a graphical model on $(X_1, \ldots, X_n, Y)$ as shown in figure 9.3. Here, we draw an edge between a pair of variables if they are not independent conditioning on all other variables. The CRT procedure for the null X_{52} is shown in the bottom left. Since we sample $\tilde{X}_{52}$ by conditioning on X_{-52} (and marginalizing Y), the structure of the distribution (and hence the graph) after replacing X_{52} with $\tilde{X}_{52}$ remains unchanged since X_{52} is not directly adjacent to Y. Meanwhile, if we consider the non-null X_{164}, sampling $\tilde{X}_{164}$ from its conditional on X_{-164} (marginalizing Y) results in a joint distribution that can be represented using the graph shown in the bottom right. Notice that the structure of the dependencies is not preserved, so the joint distribution after replacing with the imputed value is different.

9.3.1 Why do we need to sample from the conditional?

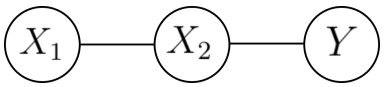
Notice that in the above procedure, it is crucial to sample from the conditional, and not the marginal of X_j, in order to preserve the dependence structure between X_j and the other covariates. For example, suppose we have X_1, X_2, and Y with $\text{cor}(X_1, X_2) = 0.5$ and

$$Y = X_2 + \epsilon$$

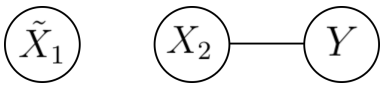
for $\epsilon \sim N(0, \sigma^2)$.

Then, we can compute

$$\mathbb{E}[YX_1] = \mathbb{E}[(X_2 + \epsilon)X_1] = 0.5.$$



(a) Structure of dependencies of the observed sample



(b) Structure after sampling $\tilde{X}_1$ from its marginal.

Figure 9.4

Now, if $\tilde{X}_1$ is sampled from the marginal of X_1, then we get

$$E[Y\tilde{X}_1] = 0 \neq 0.5 = E[YX_1]$$

even though X_1 is null in the above model. Therefore, sampling from the marginal does not provide good control: X_1 would likely appear to be significant, since we would expect the resulting p_1 from the above procedure to be small. Figure 9.4 illustrates how the independence structure of the joint distribution of this example is not preserved.

9.3.2 Causality

Why does conditional significance fail to imply causality? We use a counterexample to illustrate this. Suppose Y is the phenotype of running fast. X_1 is the genotype of good parenting skills. X_2 is the genotype of good muscle form. Conditioning on X_2, Y and X_1 are not independent since if one inherits the genotype of good parenting skills, they are likely to receive good parenting from their parents as well, which can lead to good exercise habits. In this example, we can further condition upon parents' genes to make CRT causal, but in general, it is difficult to find all variables to condition upon to obtain causality.

9.3.3 Drawbacks

Suppose we wish to control the FDR or FWER. Then we would need to compute p-values p_j for each $j \in [n]$. One issue with using CRT is that the resulting p-values are dependent because each p_j is computed using X_{-j}. Besides this, the main limitation of this approach is computational complexity. For each p_j, we need to compute importance scores n times. In total, $K \times n$ importance scores are computed.

Furthermore, for large n we may need to take K large: for instance, if we wish to use a Bonferroni test on the resulting p-values, then the threshold for Bonferroni will be $\frac{\alpha}{n}$. So we need to take K large enough that the p-values we obtain are the same resolution of this threshold to ensure that we do indeed reject non-null hypotheses.

In the next lecture, we will explore the idea of knockoffs to produce p-values. This approach will be more computationally tractable.

Bibliography

[1] C Glenn Begley and Lee M Ellis. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.