1 Outline

Agenda: Knockoffs

1. Introduction
2. Comparison with Other Dummy Variable Methods
3. Rigorous Results on Knockoffs

2 Introduction

Classically, the validity of inference had relied on the formulation of hypotheses prior to the collection and analysis of data. Nowadays, it is often the case that large data sets are collected prior to the formulation of hypotheses. The data is used to select hypotheses of interest and then to perform inference. In this way, the classical scientific method has been inverted. If we ignore the use of data in selection, our inferences will be invalid and our discoveries may not be reproducible. In this new world, we need to find solutions that can help improve research quality and reproducibility.

The problem we shall consider here is this: we have a response $Y$, which potentially depends on thousands/millions of covariates $X_1, \ldots, X_p$. 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

$$Y \perp X_j \mid X_{-j}.$$

That is, $X_j$ is not associated with $Y$ given all the other covariates. Notice that this is a stronger notion of null variable than is used when we test marginal associations, and it is often what we really care about. We would like a method which selects interesting variables without selecting too many false positives. Here we do this by controlling the false discovery rate.

3 Comparison with Other Dummy Variable Methods

Many variable selection methods are available which compute for each covariate an importance statistic which serves as a basis for including a variables in our model or not. 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, for example. It is difficult to do inference with these selection methods because the sampling distribution of the feature importance statistics is extremely complicated and not known. What we would like to know is how many nulls are among the selected variables. One idea is to create dummy variables which preserve some of the structure of the real variables. We know the dummies are null and hope that because they preserve some of the structure of the real variables, the number of dummy variables selected by our method will be a good estimate for the number of nulls among the real variables which were selected. That is, we hope the dummies will serve as control variables. In this lecture, we introduce knockoff dummies which imitate the actual data in a way that allows us to control the FDR.

First we compare knockoff dummies with two alternative dummy construction methods. These methods create a dummy variable for each real covariate. Gaussian dummies are constructed by generating for each real covariate a Gaussian variable with the same mean and variance; permuted dummies are constructed by permuting the rows of the original design matrix, so that the full joint distribution of the covariates is preserved. However, the above structures are not sufficient for the dummies to imitate the behaviors of nulls in statistical tests. To demonstrate this, we run Lasso with tuning parameter $\lambda = 3$ on the GWAS data with three types of dummies (Gaussian dummies, permuted dummies and knockoff dummies), and plot the estimated coefficients of the nulls and the dummies.

$$\text{Feature importance statistic } Z_j = |\hat{\beta}_j(\lambda = 3)|$$

![Feature importance statistic](image)

Figure 1: Gaussian dummies

From the plots (Figure 1, Figure 2 & Figure 3), we observe that the distribution of the Lasso estimates for the knockoff dummies resemble those of the nulls using the real data, while the distri-
Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$

Figure 2: Permuted dummies

Feature importance $Z_j = |\hat{\beta}_j(\lambda = 3)|$

Figure 3: Knockoff dummies
bution of the Lasso estimates for Gaussian dummies and permuted dummies are rather different. It is clear from these simulations that Gaussian dummies and permuted dummies cannot serve as controls.

**Question:** Why can permuted dummies not serve as controls?

The problem is there may be a relationship between $X_1$ and $X_2$ that does not exist between $X_1$ and $X_2$’s dummy. This makes the test statistic for a null variable behave differently than that of its permuted version. Consider the following example.

**Example:** Consider a linear model with two variables $X_1$ and $X_2$. $Y = X_1 + \epsilon$. Both $X_1$ and $X_2$ are standardized with $EX_1 = EX_2 = 0$ and $Var(X_1) = Var(X_2) = 1$. They are correlated with $Cor(X_1, X_2) = 0.5$. The marginal correlations of $Y$ with $X_1$ and $X_2$ are non-zero: $E(X_1Y) = 1$ and $E(X_2Y) = 0.5$. But the marginal correlation of $Y$ with permuted features is zero: $E(X_{2,\pi}Y) = 0$. This means that the statistics $X_2^TY$ and $X_{2,\pi}^TY$ cannot be the same!
It can be seen from the two plots above that the test statistics (e.g. marginal correlations or regression coefficients) corresponding to the original null and permuted null clearly do not have the same distribution.

To solve the problem, we want to build knockoff features \((\tilde{X}_1, \tilde{X}_2)\) such that if \(X_2\) is null, \(\tilde{X}_2^TY \overset{d}{=} X_2^TY\). If \(X_2\) is correlated with \(X_1\), we hope \(\tilde{X}_2\) is correlated with \(X_1\) (\(\tilde{X}_1\)) in the same way. This is different from the permuted dummies.

4 The Knockoff Inference Machine

This is originally proposed by Candès and Barber and re-interpreted by Candès, Fan, Janson and Lv.

4.1 Hypotheses and test statistics

Examples of questions of interest

- (Parametric) linear model. \(Y = \beta_0 + \beta_1X_1 + \cdots + \beta_pX_p + \epsilon\). \(H_j : \beta_j = 0\).
- Nonparametric model. \(H_j : Y \perp \perp X_j|X_{-j}\).

Examples of test statistics \(Z_j\)

- Some random forest feature importance statistic
- Value of square-root lasso coefficient
• Posterior probability calculated according to some Bayesian model
• Your favorite deep learning feature

4.2 Exchangeability of feature importance statistics

In order for the dummy variables to serve as controls, we need the feature importance statistics of the null dummies to behave like those of the original nulls. Precisely, denote the feature importance statistics of the original and knockoff variables by

$$(Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p) = z([X, \tilde{X}], y).$$

We would like exchangeability between null knockoffs and true nulls:

$$j \in \mathcal{H}_0 \Rightarrow (Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j).$$

In fact, we need something stronger:

$$\mathcal{T} \subseteq \mathcal{H}_0 \Rightarrow (Z, \tilde{Z})_{\text{swap}(\mathcal{T})} \overset{d}{=} (Z, \tilde{Z}).$$
The plots illustrate the properties of the knockoffs: \((Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j)\).

### 4.3 Knockoffs-adjusted scores and Knockoff estimate of FDR

**Adjusted scores** \(W_j\) with flip-sign property

Combine \(Z_j\) and \(\tilde{Z}_j\) into single (knockoff) score \(W_j\).

\[
W_j = w_j(Z_j, \tilde{Z}_j) \quad w_j(\tilde{Z}_j, Z_j) = w_j(Z_j, \tilde{Z}_j).
\]
For example,

\[ W_j = Z_j - \tilde{Z}_j, \quad W_j = Z_j \lor \tilde{Z}_j \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j \\ -1, & \text{else.} \end{cases} \]

Adjusted scores \( W_j \) need to satisfy

1. Null \( W_j \)'s are symmetrically distributed
2. Conditional on \(|W|\), signs of null \( W_j \)'s are i.i.d. coin flips

To obtain a **knockoff estimate of FDP**, we are interested in selecting \( \{j : W_j \geq t\} \).

\[
FDP(t) = \frac{\# \{j \text{ null} : W_j \geq t\}}{\# \{j : W_j \geq t\} \lor 1} \approx \frac{\# \{j \text{ null} : W_j \leq -t\}}{\# \{j : W_j \geq t\} \lor 1} \leq \frac{\# \{j : W_j \leq -t\}}{\# \{j : W'_j \geq t\} \lor 1} = \hat{FDP}(t).
\]

**Step-up rule**: stop last time ratio between ‘−’ and ‘+’ below target FDR level.

More rigorously, define

\[ S^\pm(t) = \{j : |W_j| \geq t \text{ and } \text{sgn}(W_j) = \pm\} \]

The stopping time (according to our step-up rule) is given by

\[
\tau_{0/1} = \min \left\{ t : \frac{0/1 + |S^-(t)|}{1 \lor |S^+(t)|} \leq q \right\}.
\]

The selected set is

\[ \hat{S}_{0/1} = \{W_j \geq \tau_{0/1}\}. \]

**Theorem 1**: [Barber and Candès 2015]

1. **Knockoff**: With selection set \( \hat{S}_0 \)

\[
\mathbb{E} \left[ \frac{\# \text{ false positives}}{\# \text{ selections} + q^{-1}} \right] \leq q.
\]

2. **Knockoff+**: With selection set \( \hat{S}_1 \)

\[
\mathbb{E} \left[ \frac{\# \text{ false positives}}{\# \text{ selections}} \right] \leq q.
\]