THESIS: MULTIPLE KNOCKOFFS

WENTONG ZHANG

Contents

1. Background Survey

In hypothesis testing, the *multiple comparison problem* is a problem that arises when one wishes to test numerous hypotheses simultaneously. In particular, as the number of inferences that we make increase, we should also expect the number of errors to increase; for instance, when null p-values are distributed uniformly between 0 and 1, if 1000 null hypotheses are tested, an average of 50 will still be rejected at the 0.05 significance level. To provide some background, we will begin by discussing ways of controlling *Type 1 error*, which may be measured via different metrics, such as the FWER or FDR. Note that although we are only bounding a measure of Type 1 error, we also care about obtaining high power in the following procedures as well.

1.1. Controlling FWER.

When we test multiple hypotheses simultaneously, we will end up with four possible outcomes for each hypothesis:

	accepted	rejected	total
true	U	V	n_0
false	T	S	$n - n_0$
total	$n - R$	R	n

Using the above table, we define the family-wise error rate.

Definition 1.1. The *family-wise error rate*, abbreviated FWER, is

$$\text{FWER} = \mathbb{P}(V \geq 1). \tag{1.1}$$

Various classical multiple comparison procedures seek to control the FWER. Two notable procedures are Holm's procedure and Hochberg's procedure. Detailed descriptions of these procedures may be found in [Can]. The two methods are quite similar, in particular since the threshold for selecting which hypotheses to reject is the same. The key difference is the order of iteration

Date: January 24, 2018.

through the hypotheses: Holm's procedure is a *step-up procedure* as hypotheses are rejected until an acceptance, at which point the procedure finishes. On the other hand, Hochberg's procedure is a *step-down procedure* which scans backwards until a hypothesis is rejected, at which point all hypotheses prior are rejected as well. Fascinatingly, both procedures are able to provide strong control of the FWER (ie. regardless of which hypotheses are true and false), though typically Hochberg's procedure yields more power.

1.2. **Controlling FDR.** In the 1990s, a new metric for error control called the *false discovery rate* was introduced. This is a looser metric to use than the FWER, which refers to the probability of there being any false discoveries at all. We now define the false discovery rate below.

Definition 1.2. The *false discovery proportion* is defined as

$$\text{FDP} = \frac{V}{\max(R, 1)}.$$

Above, R is the total number of rejections we make, while V is the number of rejections of actually true hypotheses (ie. erroneous rejections). Note that R is an observed value, but V is not, so FDP is actually an unobserved random variable. As such, we strive for control of its expectation, and define

$$\text{FDR} = \mathbb{E}[\text{FDP}].$$

The question of controlling FWER versus FDR should be answered by using context for what a false discovery means in particular situations. For instance, studies regarding cheating may want to control the FWER (as false discoveries are serious false accusations), whereas studies regarding gene expresion may want to control the FDR (since discovering false connections is not as severe).

Perhaps the most well-known procedure for controlling the FDR is the Benjamini-Hochberg procedure. In [BH95], Benjamini and Hochberg propose controlling the FDR rather than the FWER, and introduce a novel procedure that allows for control of the FDR.

The main thing that we are interested in here is the actual procedure and is presented below; the proof can be found in the original paper, whereas a martingale proof is given in [Can].

Theorem 1.3. *Consider hypotheses $H_1, \ldots, H_m$ with corresponding p-values $P_1, \ldots, P_m$, and order them as $P_{(1)} \leq \ldots \leq P_{(m)}$. Let q^* be the desired level of FDR control. Then, define*

$$k = \max i \text{ s.t. } P_{(i)} \leq \frac{i}{m} \cdot q^*$$

and reject all hypotheses $H_{(1)}, \ldots, H_{(k)}$. This procedure controls the FDR at level q^*.

Note that the Benjamini-Hochberg procedure is a *step-down procedure*, since the selected k is a maximum index, rather than a minimum.

1.3. **Introducing (Model-X) Knockoffs.** Recently, Barber and Candès have proposed a new procedure that also controls the FDR for linear Gaussian models in [BC15]. The procedure, known as the *knockoff filter*, creates "knockoff" variables, which have the same covariance structure as the covariates, but are independent of the output, and uses these variables to control the false discovery rate. [CFJL17] then provided a sweeping generalization of the procedure beyond linear models.

not sure how much detail should be given to the background on knockoffs: should just cite the paper and move on?

2. Multiple Knockoffs

The concept of multiple knockoffs is a direct extension of the work done in [BC15] and [CFJL17]. We define multiple model-X knockoffs as follows. First, recall the definition of a null covariate.

Definition 2.1. A variable X_j is said to be "null" if and only if Y is independent of X_j conditionally on the other variables $X_{-j} = \{X_1, \dots, X_p\} \setminus \{X_j\}$. The subset of null variables is denoted by $\mathcal{H}_0$ and we call a variable X_j "non-null"or relevant if $j \notin \mathcal{H}_0$.

Definition 2.2. Multiple model-X knockoffs for a family of random variables $X = (X_1, \dots, X_p)$ are a collection of n families of new random variables $X^{(i)} = (X_1^{(i)}, \dots X_p^{(i)})$ constructed satisfying the following two properties, analogous to those of knockoffs: (1) *Exchangeability.* Let $[[n]]$ denote the set $\{0, \dots, n\}$. For each $1 \le j \le p$, let $\sigma_j : [[n]] \to [[n]]$ be any permutation of the indices 0 to n. Then, letting $X^{(0)} = X$,

$$(X^{(0)}, \dots, X^{(n)}) \stackrel{d}{=} (X_1^{(\sigma_1(0))}, \dots, X_p^{(\sigma_p(0))}, \dots, X_1^{(\sigma_1(n))}, \dots, X_p^{(\sigma_p(n))}) \tag{2.1}$$

For convenience, we write

$$\{\sigma_i\}_{i=1}^p \left(X^{(0)}, \dots, X^{(n)}\right) \stackrel{\text{def}}{=} (X_1^{(\sigma_1(0))}, \dots, X_p^{(\sigma_p(0))}, \dots, X_1^{(\sigma_1(n))}, \dots, X_p^{(\sigma_p(n))}). \tag{2.2}$$

(2) *Conditional Independence.* $(X^{(1)}, \dots, X^{(n)}) \perp\!\!\!\perp Y \mid X$ if there is a response variable Y. This property is guaranteed if the $X^{(i)}$ are constructed without looking at Y.

Here, the two conditions are analogous to the ones presented in Definition 3.1 of [CFJL17]. In fact, the second condition is identical, while the first only seeks to generalize the idea of pairwise exchangeability.

2.1. Exchangeability of Null Covariates and Knockoffs.

First, we provide a generalization of Lemma 3.2 in [CFJL17]. In particular, we extend the proof that we can permute null covariates with their knockoffs without changing the joint distribution of X and the knockoffs $X^{(i)}$, conditional on Y.

omitted detail with rows, add in here?

Lemma 2.3. *Let $S \subseteq \mathcal{H}_0$ be a subset of nulls. Consider a set of permutations $\sigma_j : [[n]] \to [[n]]$ such that if $j \notin S$, then $\sigma_j = \mathrm{id}_{[[n]]}$. Then,*

$$(X^{(0)}, \dots, X^{(n)}) \mid Y \stackrel{d}{=} \{\sigma_i\}_{i=1}^p \left(X^{(0)}, \dots, X^{(n)}\right) \mid Y.$$

Proof. The proof is quite similar to that of the original lemma. Without loss of generality, we can assume that $S = \{1, \dots, m\}$. Then, since the marginal distribution of Y is the same on both sides of the equation, it is equivalent to show that the joint distributions are the same. Then, in the same way as the original lemma, by the exchangeability condition that

$$(X^{(0)}, \dots, X^{(n)}) \stackrel{d}{=} \{\sigma_i\}_{i=1}^p \left((X^{(0)}, \dots, X^{(n)})\right),$$

so the only thing we need to show is that

$$Y \mid (X^{(0)}, \dots, X^{(n)}) \stackrel{d}{=} Y \mid \{\sigma_i\}_{i=1}^p \left((X^{(0)}, \dots, X^{(n)})\right). \tag{2.3}$$

To see this, let $p_{Y|X}(y|x)$ be the conditional distribution of Y given X. Then, note that

$$\begin{aligned} p_{Y|\{\sigma_i\}_{i=1}^p(X^{(0)},\ldots,X^{(n)})}(y|(x^{(0)},\ldots,x^{(n)})) &= p_{Y|(X^{(0)},\ldots,X^{(n)})}(y|\{\sigma_i^{-1}\}_{i=1}^p(x^{(0)},\ldots,x^{(n)})) \\ &= p_{Y|X^{(0)}}(y|x'), \end{aligned}$$

where $x_i' = x_i^{(\sigma_i^{-1}(0))}$ if $i \in S$ and $x_i' = x_i$ otherwise. In particular, the second equality above comes from the fact that Y is conditionally independent of the knockoffs $(X^{(1)},\ldots,X^{(n)})$ given X by definition of multiple knockoffs.

Next, note that we assumed earlier that $S = \{1,\ldots,m\}$ is the subset of nulls. Then, by definition, Y and X_1 will be conditionally independent given $X_{2:p}$, we may further simplify that

$$p_{Y|X^{(0)}}(y|x') = p_{Y|X_{1:p}^{(0)}}(y|x_1^{(\sigma_1^{-1}(0))}, x_{2:p}') = p_{Y|X_{2:p}^{(0)}}(y|x_{2:p}') = p_{Y|X_{1:p}^{(0)}}(y|x_1^{(0)}, x_{2:p}')$$

This shows that the conditional distributions of Y are the same:

$$Y \mid \{\sigma_i\}_{i=1}^p\left((X^{(0)},\ldots,X^{(n)})\right) = Y \mid \{\sigma_i'\}_{i=1}^p\left((X^{(0)},\ldots,X^{(n)})\right),$$

where $\sigma_1' = \mathrm{id}_{[[n]]}$ and $\sigma_i' = \sigma_i$ for $i > 1$. This reduces S to the case where $1 \notin S$, so we may repeat this argument until S is empty, and this proves the claim. □

2.2. **Feature Statistics and p-values.** In the original work on knockoffs, statistics denoted W_j are computed for each of the X_j based on both X_j and its knockoff, where a large value of W_j indicates evidence for significance of the covariate. Furthermore, we require W_j to satisfy the flip-sign property, which says that swapping X_j with its knockoff has the effect of reversing the sing of W_j. These statistics are then used in a sequential process when running the knockoffs procedure.

Returning to multiple knockoffs, it is not immediately clear how to generalize the concept of the statistic W_j. However, we may refer to the original knockoffs paper [BC15] and view the knockoffs procedure as a sequential selection procedure. In particular, a key idea in the proof of the main theorems, Theorem 1 and Theorem 2, in [BC15] is the conversion of W_j to 1-bit p-values: covariates such that $W_j > 0$ are assigned a p-value of 0.5, whereas the covariates such that $W_j < 0$ are given a p-value of 1. This conversion allows us to prove the desired statements via Theorem 3 of [BC15], which provides proof of control of FDR for generalized sequential testing procedures, the FSTP and SSTP. Now, one thing to note here is that the W_j statistics for both the covariate and its knockoff are in correspondence with generating statistics Z_j and $\tilde{Z}_j$ for the covariate and its knockoff respectively via the formula

$$W_j = f_j(Z_j, \tilde{Z}_j) \tag{2.4}$$

where f_j is an anti-symmetric function.

In the case of multiple knockoffs, instead of thinking about W_j, we will consider $Z_j^{(i)}$ for each knockoff i and retaining the convention that the case of $i = 0$ is the statistic corresponding to the original covariate. As such, we can generalize the definition of the statistic T to become

$$T \overset{\text{def}}{=} (Z^{(0)},\ldots,Z^{(n)}) = t([X^{(0)},\ldots,X^{(n)}], y) \tag{2.5}$$

Here, the $Z_j^{(k)}$ are intended to measure the significance of the covariate (or knockoff) $X^{(k)}{}_j$. Now, we present a lemma which is the desired generalization of Lemma 3.3 in [CFJL17].

Lemma 2.4. *For the null X_j, where $j \in \mathcal{H}_0$, let $A_{j,(i)}$ denote the i-th order statistic of $Z_j^{(0)}, \dots, Z_j^{(n)}$. Then*

$$\mathbb{P}\left(Z_j^{(0)} = A_{j,(i)}\right) = \frac{1}{n+1}, \qquad 1 \le i \le n+1. \tag{2.6}$$

Proof. not sure if there is more to flesh out here, seems pretty analogous to previous work, but am having trouble formalizing for some reason

This is a direct consequence of Lemma 2.3, from which we may deduce that $T \overset{\Delta}{=} \{\sigma_i\}_{i=1}^{p}(T)$ where the σ_i satisfy the condition of the lemma (ie. are non-trivial only for null j). Then, by symmetry, the statistic $Z_j^{(0)}$ must be equally likely to be i-th largest statistic among the $Z_j^{(i)}$, as desired. □

References

[BC15] Rina Foygel Barber and Emmanuel Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43, 2015.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[Can] Emmanuel Candès. Stats 300c: Lectures. `https://statweb.stanford.edu/~candes/stats300c/lectures.html`.

[CFJL17] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, 2017.