Multiple Knockoffs

Wentong Zhang
Harvard College

Submitted in partial fulfillment of the requirements for the degree of Bachelor of Arts with Honors in Statistics April 2018

April 1, 2018

Abstract

In this thesis, we present an extension of the knockoffs procedure in multiple testing. In particular, extending the work done on model-X knockoffs, we investigate the effects of generating numerous knockoffs simultaneously.

Acknowledgements

fill this in

Contents

Chapter 1

An Introduction to Multiple Comparison Procedures

In hypothesis testing, the *multiple comparison problem* is a problem that arises when one wishes to test numerous hypotheses simultaneously. In particular, we wish to accept or reject each hypothesis. In this case, 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, whereas if only 20 are tested, then an average of only 1 will be rejected at the 0.05 significance level.

When we test multiple hypotheses simultaneously, we will end up with four possible outcomes for each hypothesis, presented in the table below:

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

In the table above, note that the quantities R and n_0 are observed, but U, V, T, S are actual unobserved random variables. *Type I error* is concerned with the quantity V, which is the number of null hypotheses that are incorrectly rejected, and will be what we wish to control under various metrics. Note that although we will be providing bounds on measures of Type I error, power will ultimately become the metric by which we wish to judge our procedures, as we will seek to maximize power of a procedure given a bound on Type I error. A good example of this can be seen by examining the Bonferroni correction, the Holms procedure, and the Hochberg procedure: for the same control on the *family-wise error rate*, the Hochberg procedure is more powerful than the Holms procedure (which is more powerful than the Bonferroni correction).

1.1 Family-Wise Error Rate

To begin, we will introduce a metric of Type I error called the *family-wise error rate*.

1.1.1 Definition

Definition 1.1.1. The *family-wise error rate*, abbreviated FWER, is defined as

$$\text{FWER} = \mathbb{P}(V \geq 1) \tag{1.1.1}$$

where V is the number of erroneous rejections of null hypotheses.

Note that bounding the FWER can be a rather stringent constraint on a procedure, as the procedure must then control the probability of any false rejections at all.

say a bit more about the FWER here

1.1.2 The Bonferroni Correction

One of the earliest methods presented in controlling the FWER is known as the *Bonferroni correction* (alternatively the Bonferroni method). First used in [Dun59], this procedure draws its name from the Bonferroni inequalities, which are used in the proof of its control of the FWER.

Procedure 1.1.2 (Bonferroni Correction)**.** Suppose we are given a collection of n null hypotheses, denoted $H_{0,1}, \ldots, H_{0,n}$, with associated p-values denoted $p_1, \ldots, p_n$ respectively. Fix a desired level α. Then, reject all hypotheses $H_{0,i}$ for which

$$p_i \leq \alpha/n. \tag{1.1.2}$$

In other words, test all hypotheses $H_{0,i}$ at level α/n.

Proposition 1.1.3. *Procedure 1.1.2 controls the FWER at level α.*

Proof. stuff □

Two notable procedures that control the FWER 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 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.

provide detailed explanation of holm and hochberg

1.1.3 Controlling FDR

In the 1990s, a new metric for error control called the *false discovery rate* (FDR) was introduced. Colloquially, the FDR is a more relaxed metric to use than the FWER, which refers to the probability of there being any false discoveries at all.

Definition 1.1.4. The *false discovery proportion* is defined as

$$\text{FDP} = \frac{V}{\max(R, 1)} \tag{1.1.3}$$

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

$$\text{FDR} = \mathbb{E}[\text{FDP}]. \tag{1.1.4}$$

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 the consequences of discovering false connections are likely not as severe).

The most well-known procedure for controlling the FDR is the Benjamini-Hochberg procedure. In [BH95], Benjamini and Hochberg proposde control the FDR in lieu of the FWER, and introduced the following procedure that controls the FDR.

Theorem 1.1.5. *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^*.

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].

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

provide greater exposition of BH procedure

Chapter 2

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?

Chapter 3

Multiple Knockoffs

3.1 Definition

The concept of multiple knockoffs is a direct extension of the work done in [BC15] and [CFJL17] that we revisited in Chapter 2. We define multiple model-X knockoffs as follows. First, recall the definition of a null covariate.

Definition 3.1.1. 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 3.1.2. 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 $1 \leq j \leq p$, let $\sigma_j : [[n]] \to [[n]]$ be any permutation of the indices 0 to n. Then, letting $X^{(0)} = X$,

$$(X^{(0)}, \dots, X^{(n)}) \stackrel{d}{=} (X_1^{(\sigma_1(0))}, \dots, X_p^{(\sigma_p(0))}, \dots, X_1^{(\sigma_1(n))}, \dots, X_p^{(\sigma_p(n))}) \tag{3.1.1}$$

For convenience, we write

$$\{\sigma_i\}_{i=1}^p \left(X^{(0)}, \dots, X^{(n)}\right) \stackrel{\text{def}}{=} (X_1^{(\sigma_1(0))}, \dots, X_p^{(\sigma_p(0))}, \dots, X_1^{(\sigma_1(n))}, \dots, X_p^{(\sigma_p(n))}). \tag{3.1.2}$$

(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.

3.2 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 3.2.1. *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,*

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

Proof. The proof is quite similar to that of the original lemma. Without loss of generality, we can assume that $S = \{1, \ldots, m\}$. 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

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

so the only thing we need to show is that

$$Y \mid (X^{(0)}, \ldots, X^{(n)}) \overset{d}{=} Y \mid \{\sigma_i\}_{i=1}^{p} \left((X^{(0)}, \ldots, X^{(n)})\right). \tag{3.2.1}$$

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

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

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. □

3.3 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, W_j is required to satisfy the flip-sign property, which says that swapping X_j with its knockoff has the effect of reversing the sign of W_j. These statistics are then used in a sequential process when running the knockoffs procedure.

Returning to multiple knockoffs, we now wish to generalize the concept of the statistic W_j. To be informal, there are two properties of W_j that are critical to the knockoffs procedure: the sign of W_j, which determines the p-value that is used in the selection process, and the magnitude of $|W_j|$, which determines the ordering of the W_j in the selection process. As such, to generalize to multiple knockoffs, we just need to ensure that we have generalizations of the ideas of obtaining a p-value from our statistic and obtaining a magnitude of our statistic.

To be more precise, we may refer to the original knockoffs paper [BC15] and view the knockoffs procedure as a sequential selection procedure. A key idea in the proof of the main theorems, Theorem 1 and Theorem 2, in [BC15] is the conversion of the sign 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. Note here 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

$$W_j = f_j(Z_j, \tilde{Z}_j) \tag{3.3.1}$$

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

$$T \stackrel{\text{def}}{=} (Z^{(0)}, \ldots, Z^{(n)}) = t([X^{(0)}, \ldots, X^{(n)}], y) \tag{3.3.2}$$

Here, the $Z_j^{(k)}$ are intended to measure the significance of the covariate (or knockoff) $X^{(k)})_j$.

Now, we must consider the other critical component to the knockoff procedure: the magnitude of W_j, which determines the order in which we look at the covariates in the sequential selection procedure. Here, similar to the choice of anti-symmetric function we make for $f_j(Z_j, \tilde{Z}_j) = W_j$ in the original knockoffs case, a choice must be made to determine the "magnitude" of the statistic, denoted M_j, which can also write as $g_j(Z_j^{(0)}, \ldots, Z^{(n)_j})$. In addition, noting that f_j is anti-symmetric in the original case, we will require that g_j is a symmetric function in its arguments (the analog here is that the magnitude of f_j would be symmetric).

Now, we present a lemma which is the desired generalization of Lemma 3.3 in [CFJL17].

Lemma 3.3.1. *For the null X_j, where $j \in \mathcal{H}_0$, let $A_{j,(i)}$ denote the i-th order statistic of $Z_j^{(0)}, \ldots, Z_j^{(n)}$. Then,*

conditional on $(M_1, \ldots, M_p)$,

$$\mathbb{P}\left(Z_j^{(0)} = A_{j,(i)}\right) = \frac{1}{n+1}, \qquad 1 \le i \le n+1. \tag{3.3.3}$$

Proof. slightly confused here

Like Lemma 3.3 in [CFJL17], this is a direct consequence of Lemma 3.2.1, and we will provide an analogous proof. From the lemma, we may deduce that $T \overset{d}{=} \{\sigma_i\}_{i=1}^{p}(T)$ 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. □

The above lemma allows us to obtain the analogous version of p-values for multiple knockoffs. Using the same notation as in Lemma 3.3.1 for the covariate X_j, we define the p-value

$$p_j = 1 - \frac{i-1}{n+1}, \qquad Z_j^{(0)} = A_{j,(i)}. \tag{3.3.4}$$

Note that the special case where $n = 1$ simplifies directly into the original knockoffs framework.

3.4 The Procedure

Now that we have developed the analog of the desired test statistic in the setting of multiple knockoffs, we may briefly describe the procedure itself. The parallel here is rather direct, and is simply using the machinery that we have developed in the framework of [BC15] under the FSTP and SSTP described there. We will present these results in a similar fashion to [CFJL17].

Theorem 3.4.1. *Fix a threshold $c \in (0,1)$. Then, choose a threshold $\tau > 0$ by setting*

$$\tau = \min\left\{t > 0 : \frac{\#\{j : M_j \ge t, p_j > c\}}{\#\{j : M_j \ge t, p_j \le c\}} \le q\right\} \qquad \textbf{(MultipleKnockoffs)}$$

where q is the target FDR level (or $\tau = +\infty$ if the set above is empty). Then, the procedure that selects the variables

$$\hat{S} = \{j : M_j \ge t, p_j \le c\}$$

controls the modified FDR defined as

$$\text{mFDR} = \mathbb{E}\left[\frac{|\{j \in \hat{S} \cap \mathcal{H}_0\}|}{|\hat{S}| + 1/q}\right] \le q.$$

Similarly, the more conservative procedure given by choosing the threshold $\tau_+ > 0$ where

$$\tau_+ = \min\left\{t > 0 : \frac{1 + \#\{j : M_j \ge t, p_j > c\}}{\#\{j : M_j \ge t, p_j \le c\}} \le q\right\} \qquad \textbf{(MultipleKnockoffs+)}$$

and setting $\hat{S} = \{j : M_j \geq t, p_j \leq c\}$ *controls the usual FDR by*

$$\mathbb{E}\left[\frac{|\{j \in \hat{S} \cap \mathcal{H}_0\}|}{|\hat{S}| \vee 1}\right] \leq q.$$

Proof. The proof of the theorem above follows from Theorem 3 of [BC15]. In particular, the connection with knockoffs in Section 5.2 of [BC15] explains the connection between knockoffs and the SSTP. Here, the connection is even more explicit: we replace $|W_j|$ with M_j, and store the information of the p-values directly. In particular, from the work that we have done, it is evident that $p_j \geq \text{Unif}[0, 1]$, and we assume that they are independent from the non-null p-values. Finally, note that we are doing something analogous to ordering our covariates in order of decreasing M_j in the procedure by looking at cutoffs for M_j. Hence, it is clear that the procedure outlined above is a special case of the SSTP, and we are done. □

A noteworthy difference between multiple knockoffs and knockoffs is the addition of the cutoff parameter c for p-values. Indeed, since knockoffs only using extremely rough 1-bit p-values, a choice for c in the SSTP is not particularly enlightening. In particular, selecting any $c \in (1/2, 1)$ is strictly worse than selecting $c = 1/2$, whereas selecting $c \in (0, 1/2)$ leads to zero power; it is evident that the choice $c = 1/2$ is optimal.

For multiple knockoffs, however, there is no canonical choice for c in the SSTP description. For instance, consider the case where $n = 2$. Then, two candidates for c exist: $1/3$ and $2/3$ are both possible. Later, we will investigate the effects of choosing different c on the power of the procedure in various situations for different n.

Bibliography

[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.

[Dun59] Olive Jean Dunn. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics*, 30, 1959.