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.

1.1 The Family-Wise Error Rate

In this section, we will first develop background on some classical procedures that control a measure of Type I error called the *family-wise error rate*. In doing so, we will see that 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). To begin, we will the *family-wise error rate* as follows.

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.

1.1.1 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] in 1959, 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. The proof of FWER control is a simple application of the union bound (or Bonferroni inequalities). In particular, without loss of generality, assume that $H_{0,i}$ for $i \leq n_0$ are the only true null hypotheses (where $n_0 \leq n$). Then, the FWER may be bounded, as desired, by

$$\text{FWER} = \mathbb{P}(V \geq 1) = \mathbb{P}\left[\bigcup_{i=1}^{n_0}\left(p_i \leq \frac{\alpha}{n}\right)\right] \leq \sum_{i=1}^{n_0} \mathbb{P}\left(p_i \leq \frac{\alpha}{n}\right) = n_0 \cdot \frac{\alpha}{n} \leq \alpha. \tag{1.1.3}$$

□

The Bonferroni correction controls the FWER at level α, but the cost in power can be quite severe. Consider a concrete example: suppose we have 1000 null hypotheses that we wish to test at the $\alpha = 0.05$ level. In this case, we would only reject the hypotheses for which the p-values are less than 5×10^{-5}, or extremely significant p-values. As such, this test may be most suitable for situations in which we expect certain p-values to be extremely significant, as otherwise, it is possible that the Bonferroni procedure yields very low power.

1.1.2 Holm's Procedure

Another procedure that controls the FWER in multiple hypothesis testing is Holm's procedure. Previously, we saw that Procedure 1.1.2 successfully controlled the FWER at level α for tests at level α, but a simple numerical example illustrates how restrictive the Bonferroni procedure can be, in particular when considering a large number of hypotheses. Holm's procedure, introduced in [Hol79], provides less strict criteria for rejecting the null hypotheses, and hence will have uniformly more power. The procedure is as follows.

Procedure 1.1.4 (Holm's Procedure)**.** 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, order the p-values so that

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$$

and let $H_{(1)}, \ldots, H_{(n)}$ be the corresponding hypotheses. Then, define

$$k = \min\left\{1 \leq i \leq n : p_{(i)} > \frac{\alpha}{n-i+1}\right\}$$

and reject hypotheses $H_{(1)}, \dots H_{(k-1)}$. In particular, if it is the case that

$$p_{(i)} \leq \frac{\alpha}{n-i+1}, \qquad \forall 1 \leq i \leq n,$$

then we reject all $H_1, \dots, H_n$.

Note. A seqeuntial way to think about Holm's procedure is as follows. We begin by considering $p_{(1)}$: if $p_{(1)} \leq \alpha/n$, then reject $H_{(1)}$ and continue to $p_{(2)}$. Then, for each $p_{(i)}$, reject $H_{(i)}$ if $p_{(i)} \leq \alpha/(n-i+1)$, and continue to $p_{(i+1)}$; otherwise, end the procedure. It is clearly seen that this is equivalent to finding k as defined in Procedure 1.1.4 and rejecting hypotheses up to $H_{(k-1)}$.

In particular, Holm's procedure is a *step-down procedure*, as the procedure functions by rejecting hypothesis sequentially until one is accepted. On the other hand, the Hochberg procedure, which we examine in the next section, is actually an extremely similar step-up procedure. In terms of power, step-up procedures can be significantly more powerful than step-down procedures; this is a phenomenon we will elaborate on further in the next section.

From the sequential view of Holm's procedure, it now becomes evident that Holm's procedure is uniformly more powerful than Bonferroni's procedure as follows. We can translate Bonferroni's procedure into a sequential procedure as well: we may order the p-values in order, and sequentially go through each of the p-values with the fixed threshold $p_{(i)} \leq \alpha/n$ and rejecting until $p_{(i)} > \alpha/n$. Hence, we can put Bonferroni's procedure into the same form as Holm's procedure, but the cutoffs are always stricter (or just as strict) as Holm's procedure for rejection; hence, Holm's procedure will be uniformly more powerful since it simply makes more rejections. We now provide a proof of FWER control under Holm's procedure.

Proposition 1.1.5. *Procedure 1.1.4 controls the FWER at level α.*

Proof. Let $\mathcal{H}_0$ denote the set of true null hypotheses, and let $n_0 = |\mathcal{H}_0|$ be the number of true null hypotheses. Then, let i_0 be the rank of the smallest null p-value corresponding to a true null hypothesis. This means that the first true null hypothesis is encountered when examining $p_{(i_0)}$ in the sequential interpretation of Holm's procedure. Now, note that we must have

$$i_0 \leq n - n_0 + 1.$$

This is true by inspection, since there are only $n - n_0$ false null hypotheses.

Now, consider the situation in which Holm's procedure falsely rejects a true null hypothesis. This happens if and only if the first true null hypothesis is rejected, so we may simply think about when the first true null hypothesis is rejected. This happens precisely when for all $i \leq i_0$, we have that $p_{(i)} \leq \alpha/(n-i+1)$, and so in particular $p_{(i_0)} \leq \alpha/(n-i_0+1) \leq \alpha/n_0$. Hence,

$$\text{FWER} = \mathbb{P}(V \geq 1) = \mathbb{P}\left(p_{(i_0)} \leq \alpha/n_0\right) = \mathbb{P}\left(\min_{i \in \mathcal{H}_0} p_i \leq \alpha/n_0\right) \leq \sum_{i \in \mathcal{H}_0} \mathbb{P}(p_i \leq \alpha/n_0) = \alpha.$$

This completes the proof as desired. □

Note. Another perspective on Holm's procedure is as a *closure* of the Bonferroni global test. Though this is an interesting way to derive the procedure and lends itself to the construction of Hochberg's procedure as a more conservative and simpler procedure than the closure of the Simes global test, it is not particularly pertinent to our discussion of multiple comparison procedures, so we will omit a discussion and leave the details to [Can].

1.1.3 Hochberg's Procedure

Hochber'sg procedure is the final procedure that provides control of the FWER that we present. At a glance, Hochberg's procedure is quite similar to Holm's procedure, as it uses the same modified p-value cutoffs.

Procedure 1.1.6 (Hochberg's Procedure)**.** Suppose we are given a collection of n null hypotheses, denoted $H_{0,1}, \dots, H_{0,n}$, with associated p-values denoted $p_1, \dots, p_n$ respectively. Fix a desired level α. Then, order the p-values so that

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$$

and let $H_{(1)}, \dots, H_{(n)}$ be the corresponding hypotheses. Then, define

$$k = \max \left\{ 1 \leq i \leq n : p_{(i)} \leq \frac{\alpha}{n - i + 1} \right\}$$

and reject hypotheses $H_{(1)}, \dots H_{(k-1)}$. In particular, if it is the case that

$$p_{(i)} \leq \frac{\alpha}{n - i + 1}, \qquad \forall 1 \leq i \leq n,$$

then we reject all $H_1, \dots, H_n$.

Proposition 1.1.7. *Assuming independence of the p-values, Procedure 1.1.6 controls the FWER at level α.*

Proof. Omitted, see [Can] for details. □

Note. The sequential way to think about Hochberg's procedure is as follows. We begin by consider $p_{(n)}$, and if $p_{(n)} \leq \alpha/n$, then reject all hypotheses. Otherwise, consider $p_{(n-1)}$, and proceed similarly: if $p_{(n-1)} \leq \alpha/(n-1)$, then reject hypotheses $H_{(1)}, \dots, H_{(n-1)}$. In a sense, Hochberg's procedure operates opposite Holm's procedure as a *step-up procedure* rather than a *step-down procedure*: this allows for significantly increased power in some cases. For an extreme example, consider the case where all p-values are precisely α. Then, Holm's procedure will not reject any hypotheses, whereas Hochberg's procedure will reject all hypotheses.

1.2 The False Discovery Rate

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 1.2.1. The *false discovery proportion* is defined as

$$\text{FDP} = \frac{V}{R \vee 1} \tag{1.2.1}$$

where R is the total number of rejections we make, while V is the number of rejections of true 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

$$\text{FDR} = \mathbb{E}[\text{FDP}]. \tag{1.2.2}$$

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).

1.2.1 The Benjamini-Hochberg Procedure

The most well-known procedure for controlling the FDR is the Benjamini-Hochberg procedure. Indeed, in the original paper, Benjamini and Hochberg proposed control the FDR in lieu of the FWER, and introduced the following procedure that controls the FDR. Note that similar to Procedure 1.1.6, the Benjamini-Hochberg procedure is a *step-up procedure*; the selected cutoff k is a maximum index in a similar form.

Procedure 1.2.2. Suppose we are given a collection of n null hypotheses, denoted $H_{0,1}, \dots, H_{0,n}$, with associated p-values denoted $p_1, \dots, p_n$ respectively. Fix a desired level of FDR control, q^*. Then, order the p-values so that

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(n)}$$

and let $H_{(1)}, \dots, H_{(n)}$ be the corresponding hypotheses. Then, define

$$k = \max\left\{1 \le i \le n : p_{(i)} \le \frac{i}{n} \cdot q^*\right\}$$

and reject all hypotheses $H_{(1)}, \dots, H_{(k)}$.

Proposition 1.2.3. *Assume independence of the p-values. Then, Procedure 1.2.2 controls the FDR at level q^*.*

Instead of presenting the original proof of FDR control that Benjamini and Hochberg provide in their paper, we will provide a crisper proof from [Can].

Proof. Here, we are just rehearsing the original proof presented in [Can]. Let $\mathcal{H}_0$ denote the set of true null hypotheses, and let $n_0 = |\mathcal{H}_0|$ be the number of true null hypotheses. First, we may assume that $n_0 \ge 1$ (otherwise the claim is vacuously true). Then, recall that the FDP is defined as

$$\text{FDP} = \frac{V}{R \vee 1}$$

where V is the number of incorrect rejections and R is the total number of rejections. Now, for each null $i \in \mathcal{H}_0$, define the indicator variable $V_i = \mathbb{1}_{H_{0,i} \text{ rejected}}$. This allows us to re-express the FDP as

$$\text{FDP} = \frac{V}{R \vee 1} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{R \vee 1}.$$

Next, note that if we can show $\mathbb{E}[V_i/(R \vee 1)] = q/n$, then the rest of the proof follows easily. To show that this is true, first note that we can rewrite

$$\frac{V_i}{R \vee 1} = \sum_{k=1}^{n} \frac{V_i \mathbb{1}_{\{R=k\}}}{k} \tag{1.2.3}$$

since the sum simply runs over the possible values of the number of rejections, R; in addition, note that the above formula still holds when $R = 0$ since in that case, we have that $V_i = 0$ by definition.

Next, we will note some properties of the Benjamini-Hochberg procedure, which we will use later in the proof. First, note that when the procedure makes k rejections, then a hypothesis $H_{0,i}$ is rejected if and only if $p_i \le q^* \cdot k/n$. As such, we have that

$$V_i = \mathbb{1}_{H_{0,i} \text{ rejected}} = \mathbb{1}_{\{p_i \le q \cdot k/n\}}. \tag{1.2.4}$$

Second, consider the case where $H_{0,i}$ is rejected, so $p_i \le q \cdot k/n$. Then, let R_i' denote the number of rejections by the Benjamini-Hochberg procedure if we were to set $p_i \to 0$: when $H_{0,i}$ is rejected, $R_i' = R$ since we are simply reordering the first k p-values, which will not change the number of hypotheses rejected. On the other hand, consider the case where $H_{0,i}$ is not rejected, so $p_i > q \cdot k/n$. Then, $V_i = 0$. In either case, we will have that the value of $V_i \cdot \mathbb{1}_{\{R=k\}} = V_i \cdot \mathbb{1}_{\{R_i'=k\}}$.

We will now piece the proof together as follows. Instead of looking at the expectation, we will examine a conditional expectation upon all the other p-values as follows. Let $\mathcal{F}_i = \{p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_n\}$. Then,

$$\mathbb{E}\left[\frac{V_i}{R \vee 1}\middle| \mathcal{F}_i\right] = \sum_{k=1}^{n} \frac{\mathbb{E}\left[V_i \mathbb{1}_{\{R=k\}} | \mathcal{F}_i\right]}{k} = \sum_{k=1}^{n} \frac{\mathbb{E}\left[V_i \mathbb{1}_{\{R_i'=k\}} | \mathcal{F}_i\right]}{k} = \sum_{k=1}^{n} \frac{\mathbb{E}\left[\mathbb{1}_{\{p_i \le q\cdot k/n\}} \mathbb{1}_{\{R_i'=k\}} | \mathcal{F}_i\right]}{k}, \tag{1.2.5}$$

where the last two equalities used the above observations about the Benjamini-Hochberg procedure to simplify the expression. Next, noting that the p-values are independent and that $p_i \sim \text{Unif}[0,1]$, we can further simplify that

$$\sum_{k=1}^{n} \frac{\mathbb{E}\left[\mathbb{1}_{\{p_i \le q\cdot k/n\}} \mathbb{1}_{\{R_i'=k\}} | \mathcal{F}_i\right]}{k} = \sum_{k=1}^{n} \frac{(q \cdot k/n) \cdot \mathbb{E}\left[\mathbb{1}_{\{R_i'=k\}} | \mathcal{F}_i\right]}{k} = \frac{q}{n} \cdot \sum_{k=1}^{n} \mathbb{1}_{\{R_i'=k\}}. \tag{1.2.6}$$

Above, the first equality comes from separating the expectation via independence and using the distribution of the p-values, while the second equality is given by the fact that $\mathbb{1}_{\{R_i'=k\}}$ is deterministic conditional on $\mathcal{F}_i$. However, we claim that

$$\sum_{k=1}^{n} \mathbb{1}_{\{R_i'=k\}} = 1. \tag{1.2.7}$$

This is true as follows. Note that $R_i \ge 1$, as when we set $p_i \to 0$, we must make at least one rejection of $H_{0,i}$ at the least. On the other hand, we know that R_i' will take on one value between 1 and n, which means that the sum of the indicators will be exactly 1. Hence, we have shown that

$$\mathbb{E}\left[\frac{V_i}{R \vee 1}\middle| \mathcal{F}_i\right] = \frac{q}{n}.$$

Then, the law of total expectation allows to write the FDR as

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{V_i}{R \vee 1}\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{V_i}{R \vee 1}\middle| \mathcal{F}_i\right]\right] = \sum_{i \in \mathcal{H}_0} \frac{q}{n} = \frac{n_0}{n} \cdot q \le q.$$

This bounds the FDR as desired and completes the proof. □

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 in a procedure that controls the false discovery rate. Originally, [BC15] gave a description of the procedure in the specific case where $\underset{\sim}{y}$ follows a linear Gaussian model, and the covariate matrix is fixed. [CFJL17] then provided a sweeping generalization of the procedure beyond linear models. In this section, we seek to provide a self-contained exposition of the model-X knockoffs framework and procedure. In particular, we will loosely follow the structure of [CFJL17], while borrowing necessary details from [BC15]. After defining the problem rigorously, we will give a theoretical definition of model-X knockoffs and prove some results regarding their properties, including the procedure that controls FDR. We will then think about how to construction of knockoffs more concretely.

2.1 The Model-X Paradigm

2.2 Problem Statement

Here, we will provide a rigorous statement of the problem that we seek to solve: this is essentially identical to [CFJL17], but for sake of thoroughness, we find it necessary to review.

Suppose we have i.i.d. samples from a population in the form (X, Y). Here, X is a p-dimensional vector $(X_1, \ldots, X_p) \in \mathbb{R}^p$, and Y is a scalar in $\mathbb{R}$. We wish to determine if there is a smaller subset of these X variables for which the conditional distribution of Y depends upon: we wish to find the smallest subset $\mathcal{S}$ such that conditional on $\{X_j\}_{j\in\mathcal{S}}$, Y is independent of the other variables $\{X_j\}_{j\notin\mathcal{S}}$. In most use cases and situations, the set $\mathcal{S}$ will be unique. A pathological example is given in [CFJL17], but we will not concern ourselves with these edge cases. We now provide a definition of a "null" covariate versus a "relevant" covariate, bearing in mind the context of attempting to find the desired subset $\mathcal{S}$.

Definition 2.2.1. A variable X_j is said to be "null" if and only if Y is independent of X_j conditionally on the other variables $X_{-j} = \{X_1, \ldots, X_p\} \setminus \{X_j\}$. The subset of null variables is denoted by $\mathcal{H}_0$ and we call a variable X_j "non-null"or "relevant" if $j \notin \mathcal{H}_0$.

We may then rephrase the task at hand to looking for "relevant" variables while controlling the FDR. In particular, suppose we have a selection set $\hat{\mathcal{S}}$, a subset of the covariates. Then, the number of false discoveries, denoted V previously, will be $\left|\hat{\mathcal{S}} \cap \mathcal{H}_0\right|$, whereas thte total number of discoveries, denoted R

previously, will be $\left|\hat{\mathcal{S}}\right|$. As such, the FDP and FDR are respectively

$$\text{FDP} = \frac{\left|\hat{\mathcal{S}} \cap \mathcal{H}_0\right|}{\left|\hat{\mathcal{S}}\right| \vee 1}, \qquad \text{FDR} = \mathbb{E}\left[\frac{\left|\hat{\mathcal{S}} \cap \mathcal{H}_0\right|}{\left|\hat{\mathcal{S}}\right| \vee 1}\right]. \tag{2.2.1}$$

2.3 Definitions

2.3.1 Knockoff Variables

We will now provide the definition of model-X knockoffs, as presented in [CFJL17].

Definition 2.3.1. Model-X knockoffs for a family of random variables $X = (X_1, \ldots, X_p)$ are a new family of random variables $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ constructed satisfying the following two properties. (1) *Exchangeability.* For any subset $S \subset \{1, \ldots, p\}$,

$$(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X}), \tag{2.3.1}$$

where $(X, \tilde{X})_{\text{swap}(S)}$ is the vector obtained from $(X, \tilde{X})$ by swapping X_i and $\tilde{X}_i$ for $i \in S$. (2) *Conditional Independence.* $\tilde{X} \perp\!\!\!\perp Y \mid X$ if there is a response variable Y. This property is satisfied if $\tilde{X}$ is constructed without looking at Y.

At a high level, the knockoffs procedure will work as follows: we can generate test statistics using the original covariates as well as the knockoff covariates in the same way. Then, the original covariates that are relevant to Y should generate test statistics that are siginificant relative to its knockoff, while the null covariates should not generate test statistics that are very different than its knockoffs, and this should allow us to find the relevant covariates.

The following is an important property of knockoff variables, which we will generalize later.

Lemma 2.3.2 (Exchangeability of Nulls)**.** *Let $S \subset \mathcal{H}_0$ be a subset of nulls. Then*

$$[\mathbf{X}, \tilde{\mathbf{X}}] \mid y \overset{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)} \mid y.$$

Proof. Omitted, but uses Lemma 2.3.2 in the original proof. This is also a special case of Lemma 3.2.1, which is proved later. □

2.3.2 Feature Statistics

We now describe the feature statistics that we wish to use in our procedure. We wish to compute statistics W_j for each $j \in \{1, \ldots, p\}$, where a large value of W_j gives evidence that X_j is a relevant (non-null) covariate. In particular, the test statistic will be a function of both the original covariates and the knockoffs, as well as a function of y, so

$$W_j = w_j([\mathbf{X}, \tilde{\mathbf{X}}], y)$$

for some function w_j where larger values of w_j indicates stronger evidence that X_j is non-null. In addition, we also need to ensure that the function w_j satisfies the *flip-sign property*, which says that if we switch an original covariate with its knockoff, then the sign of the knockoff statistic is switched. Formally,

$$w_j([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}, y) = \begin{cases} w_j([\mathbf{X}, \tilde{\mathbf{X}}], y), & j \notin S, \\ -w_j([\mathbf{X}, \tilde{\mathbf{X}}], y), & j \in S. \end{cases}$$

Although the W_j are a convenient way to package all of the pertinent information coming from the original covariates and its knockoffs into a single number, the statistic soon becomes unwieldy when we attempt to generalize later. As such, perhaps it is more apt to think about the W_j in the following way. We can generate individual statistics that measure the importance of each covariate, call them Z_j and $\tilde{Z}_j$ respectively. More formally, we can construct the vector $(Z, \tilde{Z}) = t([\mathbf{X}, \tilde{\mathbf{X}}], y)$ such that

$$(Z, \tilde{Z})_{\text{swap}(S)} = t([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}, y)$$

Then, the W_j can be constructed by choosing an antisymmetric function f_j and letting

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

In particular, a choice of statistic W_j is in correspondence with a choice of statistic Z_j and a choice of antisymmetric function f_j. In our work with multiple knockoffs, we will generalize feature statistics by using Z_j's as well as a symmetric function; the analog to the sign of W_j will be examined separately. We restate Lemma 3.3 from [CFJL17] as follows.

Lemma 2.3.3. *Conditional on* $(|W_1|, \ldots, |W_p|)$*, the signs of the null* W_j*'s are i.i.d. coin flips.*

Proof. Omitted. This is a special case of Lemma 3.3.1, which is proved later. □

The above lemma critically allows us to convert the signs of the W_j statistics into one-bit p-values, which are then used in the proof of FDR control via viewing the knockoff procedure as a sequential selection testing procedure as in [BC15].

2.4 The Knockoffs Procedure

Now that we have built the machinery of knockoff variables as well as the generation of feature statistics, we are ready to describe the knockoffs procedure.

Procedure 2.4.1 (Knockoffs/Knockoffs++)**.** Suppose we are given a collection of p covariates $(X_1, \ldots, X_p)$ along with a response variable y. Fix a desired control level q. Then, generate knockoffs $(\tilde{X}_1, \ldots, \tilde{X}_p)$ and compute feature statistics W_j for each X_j satisfying the flip-sign property defined above. Next, for the Knockoffs procedure, define τ as

$$\tau = \min\left\{t > 0 : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\}. \tag{2.4.1}$$

Then, the set of covariates selected is precisely

$$\hat{\mathcal{S}} = \{j : W_j \geq \tau\}. \tag{2.4.2}$$

Alternatively, for the Knockoffs++ procedure, define τ_+ as

$$\tau_+ = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}} \leq q\right\}. \tag{2.4.3}$$

Similarly, the set of covariates selected here is

$$\hat{\mathcal{S}}_+ = \{j : W_j \geq \tau_+\}. \tag{2.4.4}$$

Theorem 2.4.2. *The Knockoffs procedure from Procedure 2.4.1 controls the modified FDR, defined as*

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

The Knockoffs++ procedure from Procedure 2.4.1 controls the usual FDR.

Proof. The proof comes from Theorem 1 and 2 of [BC15], which are respectively proven via Theorem 3 of [BC15] by framing the knockoffs procedure as a sequential testing procedure. We will not press into the details here, as they will be explored in more detail later on. □

2.5 Building and Using Knockoffs

Now that we have described the knockoffs procedure, the question of implementation still remains. In particular, how might one go about constructing the knockoff variables and sampling them? In addition, what are the benefits to using this procedure as opposed to simply applying Benjamini-Hochberg: is it clear that there are significant benefits to using this method? In this section, we will first provide an example of how to construct knockoff variables in the Gaussian case and briefly discuss a more general algorithm for sampling knockoff variables.

2.5.1 The Gaussian Case

Consider the case where $X \sim \mathcal{N}(0, \Sigma)$. Then, the distribution of $(X, \tilde{X})$ can be

$$(X, \tilde{X}) \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma & \Sigma - \text{diag}\{s\} \\ \Sigma - \text{diag}\{s\} & \Sigma \end{bmatrix}\right).$$

However, in many cases, we will be interested in the conditional distribution of $\tilde{X} \mid X$: if we have already observed X, then what is the conditional distribution of the knockoffs $\tilde{X}$ In the case of Gaussian knockoffs, this is readily answered by the classical formula

$$\tilde{X} \mid X \sim N(\mu, V)$$

where the mean and variance are given by

$$\begin{aligned} \mu &= X - X\Sigma^{-1}\,\text{diag}\{s\}, \\ V &= 2\,\text{diag}\{s\} - \text{diag}\{s\}\Sigma^{-1}\,\text{diag}\{s\}. \end{aligned}$$

In practice, this allows us to sample knockoff variables quite easily when we have observed values of X from a Gaussian distribution. Note that above, we are able to choose the values of $\text{diag}\{s\}$ for ourselves, under the constraint that the covariance matrix of $(X, \tilde{X})$ is positive semi-definite. In particular, we will want to "maximize" the values of the diagonal matrix $\text{diag}\{s\}$ in order to maximize the power of the knockoff procedure, as we will want to construct knockoffs that are not too similar to the original covariates. There are a few different ways that we may seek to do this as follows.

2.5.2 An Exact Construction of Knockoff Variables

In general, the problem of sampling knockoffs can be quite hard and remains a hard question posed. [CFJL17] offers an algorithm called the Sequential Conditional Independent Pairs algorithm to deal with the general case, but for the purpose of this paper we will not devote much time to thinking about the general case. Although it is not too difficult to generalize the Sequential Conditional Independent Pairs to multiple knockoffs, we will mostly focus on generalizing the case of Gaussian covariates in order to obtain more information about the power and effectiveness of using multiple knockoffs as opposed to other methods. Note that [CFJL17] rovided significant evidence that using the model-X knockoffs procedure was quite effective and led to significant power gain in certain scenarios, and we hope to do the same (or disprove the effectiveness of multiple knockoffs) in later parts of this paper.

Chapter 3

Multiple Knockoffs

In this chapter, we present an extension to the knockoffs procedure described in Chapter 2. The idea is to construct a family of knockoffs rather than a singlet set of knockoffs, and we will present the necessary generalizations and formalizations to do so. As far as the writer is aware, the concept of multiple knockoffs has been suggested in [BC15] as well as [CFJL17], but not formalized rigorously up to this point.

3.1 Definition

The definition of multiple knockoffs variables 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.

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

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

For convenience, we write

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

(2) *Conditional Independence.* $(X^{(1)}, \dots, X^{(m)}) \perp\!\!\!\perp Y \mid X$ 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], which was restated in Definition 2.3.1. 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.

Lemma 3.2.1. *Let $S \subseteq \mathcal{H}_0$ be a subset of nulls. Consider a set of permutations $\sigma_j : [[m]] \to [[m]]$ such that if $j \notin S$, then $\sigma_j = \mathrm{id}_{[[m]]}$. Then,*

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

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

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

so the only thing we need to show is that

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

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

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

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 $(X^{(1)}, \dots, X^{(m)})$ given X by definition of multiple knockoffs.

Next, note that we assumed earlier that $S = \{1, \dots, M\}$ 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:

$$Y \mid \{\sigma_i\}_{i=1}^{p} \left((X^{(0)}, \dots, X^{(m)})\right) = Y \mid \{\sigma'_i\}_{i=1}^{p} \left((X^{(0)}, \dots, X^{(m)})\right),$$

where $\sigma'_1 = \mathrm{id}_{[[m]]}$ 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

$$T \overset{\text{def}}{=} (Z^{(0)}, \dots, Z^{(m)}) = t([X^{(0)}, \dots, X^{(m)}], y) \tag{3.3.2}$$

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 $g_j(Z_j^{(0)}, \dots, Z^{(m)_j})$. 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 $Z_j^{(0)}, \dots, Z_j^{(n)}$. Then, conditional on $(M_1, \dots, M_p)$,*

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

Furthermore, the distribution of the ordering of the $Z_j^{(i)}$ will be uniform over all permutations.

Proof. 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

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

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*

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

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

$$\hat{S} = \{j : M_j \geq t, p_j \leq c\}$$

controls the modified FDR defined as

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

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

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

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.

3.5 Constructing Multiple Knockoffs

3.5.1 An Exact Construction

fill this in, not really the focus

3.5.2 The Gaussian Case

Here, we will provide a generalization of the construction of Gaussian (or second-order) model-X knockoffs. In particular, suppose $X \sim \mathcal{N}(0, \Sigma)$. Then, the distribution of $(X, X^{(1)}, \dots, X^{(m)})$ can be described by

$$(X, X^{(1)}, \dots, X^{(m)}) \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma & (\Sigma - \text{diag}\{s\})_{1\times m} \\ (\Sigma - \text{diag}\{s\})_{m\times 1} & \Sigma_{m\times m} - \text{diag}\{s\}_{\text{diag } m\times m} \end{bmatrix}\right)$$

where the notation $A_{m\times n}$ denotes the matrix A being repeated in an $m \times n$ block matrix fashion, and $A_{\text{diag } m\times m}$ denotes the matrix A being repeated in $m \times m$ diagonal block matrix fashion. From here, we can get the conditional distribution of the knockoffs $(X^{(1)}, \dots, X^{(m)})$ in the same fashion as before: we can write

$$(X^{(1)}, \dots, X^{(m)}) \mid X \sim \mathcal{N}(\mu, V)$$

where the mean and variance are given by

$$\begin{aligned} \mu &= (\Sigma - \text{diag}\{s\})_{m\times 1}\Sigma^{-1}X = (X - \text{diag}\{s\}\Sigma^{-1}X)_{m\times 1}, \\ V &= (\text{diag}\{s\} - \text{diag}\{s\}\Sigma^{-1}\,\text{diag}\{s\})_{m\times m} + \text{diag}\{s\}_{\text{diag } m\times m}. \end{aligned}$$

Chapter 4

Bibliography

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

[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.

[Hol79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 1979.