KNOCKOFFS WITH SIDE INFORMATION

BY ZHIMEI REN1, AND EMMANUEL CANDÈS1,2

1Department of Statistics, Stanford University, Stanford, CA 94305
2Department of Mathematics, Stanford University, Stanford, CA 94305

We consider the problem of assessing the importance of multiple variables or factors from a dataset when side information is available. In principle, using side information can allow the statistician to pay attention to variables with a greater potential, which in turn, may lead to more discoveries. We introduce an adaptive knockoff filter, which generalizes the knockoff procedure (Barber and Candès, 2015; Candès et al., 2018) in that it uses both the data at hand and side information to adaptively order the variables under study and focus on those that are most promising. The adaptive knockoffs procedure controls the finite-sample false discovery rate (FDR) and we demonstrate its power by comparing it with other structured multiple testing methods. We also apply our methodology to real genetic data in order to find associations between genetic variants and various phenotypes such as Crohn’s disease and lipid levels. Here, the adaptive knockoffs method makes more discoveries than reported in previous studies on the same datasets.

1. Introduction. Imagine a geneticist has collected genotype and phenotype data from a population of individuals. She plans to use her data to study the effect of genetic variants on a certain complex disease within this population. Prior to data analysis, it is often the case that some knowledge about the genetic variants under study is available: for instance, there may be existing works on related diseases, as well as research about the exact same disease and its occurrence within other populations. How then should our geneticist leverage this prior information in her own study? Moving away from genetics, we broadly recognize that researchers have more often than not access to prior domain knowledge, results from relevant studies, and so on. Therefore, the general question is this: how should they use side information in their data analysis to help them discover more relevant factors? How should this be done while controlling type-I errors so that we do not run into the problem of irreproducibility? Our paper is motivated by such common situations and objectives.

1.1. Controlled variable selection methods. We begin by formalizing the variable selection problem in statistical terms. Let $X = (X_1, \ldots, X_p)$ denote the covariate vector and $Y$ the response variable. We assume that the pair $(X, Y)$ is sampled from $P_X \cdot P_{Y|X}$, where $P_X$ is the marginal distribution of $X$ and $P_{Y|X}$ the conditional distribution of $Y \mid X$. The inferential goal is to test whether this conditional distribution depends on $X_j$ or not. We call feature $j$ a null if the conditional distribution of $Y \mid X$ does not depend on $X_j$ and a non-null otherwise. With this in mind, let $\mathcal{H}_0$ denote the set of nulls and put $\mathcal{H}_1 = \{1, \ldots, p\} \setminus \mathcal{H}_0$. A controlled variable selection method aims to detect non-nulls from a pool of candidates while controlling some form of type-I error. In this paper, we consider the false discovery rate (FDR) (Benjamini and Hochberg, 1995),

$$FDR = E \left[ \frac{|\hat{S} \cap \mathcal{H}_0|}{1 \lor |\hat{S}|} \right],$$

Keywords and phrases: Multiple testing, variable selection, false discovery rate (FDR), knockoff filters, Bayesian two-group models, genome-wide association study (GWAS).
where \( a \lor b = \max(a, b) \). Above, \( \hat{S} \subset \{1, \ldots, p\} \) is the selected set of covariates and \( |\cdot| \) is the cardinality of a set.

Most classical FDR-controlling procedures require that we have available valid p-values, and further require independence or constrained dependence between these p-values (e.g., Benjamini and Hochberg (1995); Benjamini et al. (2001); Storey (2002); Storey, Taylor and Siegmund (2004)). However, it is in general challenging to obtain valid p-values for hypotheses of interest, especially in the high-dimensional regime where the sample size \( n \) is on the order of the number \( p \) of covariates or less. This is the reason why common practice usually imposes stringent model assumptions and the validity of the p-values ends up relying on the correctness of the model. Researchers have noted that in common regimes, the p-values obtained by classical methods do not behave as desired, but rather in a way that will potentially inflate the FDR (see e.g., Dezeure et al. (2015); Sur, Chen and Candès (2019); Sur and Candès (2019)). The model-X knockoffs procedure, or the knockoff procedure for short, introduced in Candès et al. (2018), bypasses the need for p-values and offers a solution to the variable selection problem without making any modeling assumptions about the conditional distribution of \( Y \mid X \). The strength of this approach is that it does not ask the statistician to assume away the form of the relationship between the response variable and the explanatory variables, namely, \( P_{Y \mid X} \), which is 1) usually unknown and 2) the actual object of inference (Janson, 2017). For instance, the model-X knockoffs procedure does not ask the statistician to write down a convenient linear model or a generalized linear model—which may or may not hold at all—to describe the relationship between \( X \) and \( Y \).

This paper builds upon the knockoffs procedure and generalizes it to a setting where side information about the variables or factors under study happens to be available.

1.2. Related work. Previous work on multiple testing with side information broadly falls into two categories. The first essentially modifies the definition of the FDR to account for what is known. For example, we can use side information to weigh each hypothesis—e.g. such that a priori promising hypotheses receive a higher weight—and thereafter consider controlling a weighted version of the FDR instead of the original FDR (see e.g., Benjamini and Hochberg (1997); Benjamini and Heller (2007); Basu et al. (2018)). The other category keeps the original FDR as a target measure and aims at using side information to improve the power of the selection procedure. Such procedures are sometimes called structured multiple testing procedures and the line of work includes Genovese, Roeder and Wasserman (2006); Ferkingstad et al. (2008); Roeder and Wasserman (2009); Ignatiadis et al. (2016); Lei and Fithian (2016); Lynch et al. (2017); Ignatiadis and Huber (2017); Lei and Fithian (2018); Li and Barber (2019); and Cai, Sun and Wang (2019), among others.

In this paper, we adopt the second perspective. Our work is most notably inspired by AdaPT of Lei and Fithian (2018) in that we incorporate the idea of adaptively using side information within the knockoffs framework. In a nutshell, AdaPT assumes that we can compute independent p-values, which are then compared against a sequence of adaptive thresholds constructed using available side information. A clever calculation then produces estimates of the FDR if the analyst were to report those hypotheses below threshold. (The procedure iteratively lowers these thresholds until the FDR estimate is below a target.) In this paper, we work with the model-X knockoffs method, which is completely different, and use side information to adaptively screen knockoff importance statistics instead of p-values. An appealing feature is that no new assumptions beyond those required for (vanilla) model-X knockoffs are needed to achieve FDR control: we ask for the knowledge of only the distribution \( P_X \) of the covariates, which is reasonable in many situations (Candès et al., 2018). It is worth mentioning that Lei and Fithian (2018) discuss in passing (Section 6.2) removing the independence assumption by incorporating the idea of knockoffs. In particular, they mention that
the signs of knockoffs statistics can be viewed as independent binary p-values and AdaPT can be applied to these p-values. This is different from our methodology since 1) their discussion is in the context of Gaussian models and ours does not depend on model assumptions except for knowledge of $P_X$; 2) AdaPT computes the adaptive thresholds assuming the null the p-values are uniformly distributed—this is clearly not the case for knockoffs (so it can greatly harm statistical power). Our proposal to adaptively screen knockoff statistics brings a new perspective on structured hypotheses testing and introduces a flexible framework (generic to knockoffs) of incorporating side information that significantly improves the power in different scenarios.

The reader will correctly note that the role of side information in our framework is similar to that of a prior in the Bayesian framework. However, our perspective on side information is here frequentist and, therefore, intrinsically different. Bayesian inference is obtained by averaging over the prior distribution. In our work, FDR control holds conditional on the side information. This means the FDR control holds regardless of the correctness of the model used to provide side information. Having said this, we shall see that our adaptive knockoff filter accommodates ‘Bayesian thinking’ in the sense that side information can be assimilated into a prior, which can then be used by our method while retaining control of the FDR. This type-1 error guarantee holds no matter the validity of the prior or the quality of side information.

2. A motivating example: discovering genes with side information. In a nutshell, the vanilla knockoffs procedure (Barber and Candès, 2015; Candès et al., 2018) uses the data at hand to construct negative controls, which are then used to rank hypotheses from the least to most promising. Also, each hypothesis is classified either as candidate for selection or non-candidate. We will see later how knockoffs classify these hypotheses; for now, all the reader needs to know is that the procedure will select only among the candidate hypotheses. Selection is then achieved by applying a special step-up procedure to these ranked hypotheses; see Figure 1 for a visual illustration. Having ordered the hypotheses, the knockoff filter sequentially examines the hypotheses starting with the least promising (i.e. starting from the left in the figure). As in the Benjamini-Hochberg step-up procedure (Storey, Taylor and Siegmund, 2004), at each step, the knockoffs filter estimates the FDR among the unexamined hypotheses. This estimate is the ratio between the number of remaining non-candidate hypotheses and that of remaining candidates. If the estimated FDR falls below a user-specified threshold $q$, the procedure stops and selects the remaining candidates. Clearly, a greater number of candidates at the end of the ordering yields higher power. The catch however is this: a crucial rule for FDR control is that we are not allowed to use the status of any hypothesis—whether it is a candidate or not—when determining the ordering of the hypotheses (as we would otherwise put all the candidates at the end). Now suppose we have side information other than the data itself. If we can use it to come up with a better ordering and place more candidates towards the end, then we will have a chance to select more hypotheses and, therefore, improve power.

While we shall explore how to design orderings that exploit side information in Section 4, we first demonstrate how this can be applied to a genome-wide association study (GWAS). We consider the dataset provided by the Wellcome Trust Case Control Consortium (WTCCC., 2007), which contains genetic information on $n = 4913$ British individuals, of which 1917 have Crohn’s disease and 2996 are healthy controls. For each individual, $p = 377,749$ single nucleotide polymorphisms (SNPs) are recorded. Our inferential goal is to discover SNPs that are significantly associated with Crohn’s disease in the British population (i.e. to discover non-nulls) by means of a procedure controlling the FDR below the threshold $q = 0.1$. 


The WTCCC dataset has been studied in several works, see e.g., WTCCC. (2007); Candès et al. (2018); Sesia, Sabatti and Candès (2018), with the last two references using knockoff-based methods. We extend the knockoffs procedure by leveraging summary statistics—p-values or z-scores corresponding to marginal testing of each individual SNP—reported by genetic studies of Crohn’s disease in other populations. In this particular example, we worked with summary statistics from GWAS in East Asia and Belgium (Franke et al., 2010; Liu et al., 2015; Goyette et al., 2015).¹ Since the summary statistics come from studies in other populations, note that we are not trying to re-discover SNPs that have been discovered before.

The adaptive knockoffs filter, or adaptive knockoffs for short, uses both the WTCCC data and the summary statistics to order the hypotheses. It then sequentially examines, stops and selects hypotheses in a way similar to what we have seen in the vanilla model-X knockoffs case. Table 1 compares summary results on the WTCCC data, and we can see that the adaptive knockoffs method discovers more SNPs than other methods. Details including a full list of discovered SNPs are available in Supplementary Section S2.

<table>
<thead>
<tr>
<th>Study/Method</th>
<th>Number of SNPs discovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTCCC. (2007)</td>
<td>9</td>
</tr>
<tr>
<td>Candès et al. (2018)</td>
<td>18.0</td>
</tr>
<tr>
<td>Sesia, Sabatti and Candès (2018)</td>
<td>22.8</td>
</tr>
<tr>
<td>Adaptive knockoffs</td>
<td>33.3 (30)</td>
</tr>
</tbody>
</table>

Table 1

Number of SNPs discovered to be associated with Crohn’s disease by different methods. The target FDR level is $q = 0.1$ in all cases (WTCCC. (2007) considers the Bayesian FDR). Knockoff-based algorithms are randomized and, consequently, the reported numbers of discoveries are averaged over multiple realizations of the algorithm. In the case of adaptive knockoffs, the number of realizations is 50. The number in parentheses is the number of SNPs being selected more than 50% of the times.

Inference is valid conditionally on the side information. We wish to stress at the onset of this paper that the adaptive knockoffs procedure controls finite-sample FDR regardless of the quality of the side information, i.e., regardless of the quality of the summary statistics in our

¹The summary statistics are obtained from https://www.ibdgenetics.org/downloads.html.
example. Even in the case where the side information is wrong, we still achieve FDR control. When side information is useful, power may be increased (as is the case above). As we shall see, the reason is simple: FDR control and higher statistical power both hold conditionally on the side information.

3. Model-X knockoffs. Before presenting the details of adaptive knockoffs, we start by giving a brief introduction to the model-X knockoffs framework. Assume the covariates $X = (X_1, \ldots, X_p)$ follow a known joint distribution $P_X$ and let $P_{Y|X}$ denote the conditional distribution of the response $Y$ as before. The inferential goal is to test whether or not $P_{Y|X}$ depends on $X_j$. It is shown in Edwards (2012) and Candès et al. (2018) that under mild conditions the above testing problem is equivalent to testing

$$H_j : Y \perp \!\!\!\! \perp X_j \mid X_{-j},$$

where $X_{-j} \in \mathbb{R}^{p-1}$ is the vector $X$ after deleting $X_j$. Hypothesis $j$ is called a null if $H_j$ is true and a non-null otherwise. Hence, a variable is null if and only if it is independent of the response given the knowledge of the others; throughout the paper, we shall work with (1).

The knockoffs procedure starts by computing a feature importance statistic $W_j$ for each hypothesis $H_j$. Before constructing the $W_j$’s, we first describe two key properties: (1) the null $W_j$’s have equal probability of being positive or negative; (2) the signs of the null $W_j$’s are mutually independent, and are independent of the signs of the non-null $W_j$’s. Also, the feature importance statistics are designed in such a way that the non-null $W_j$’s tend to take on larger values. We call $H_j$ a non-candidate hypothesis if $W_j < 0$ and a candidate hypothesis if $W_j > 0$ (as we have seen before, the knockoffs method only selects among the candidate hypotheses). The vanilla knockoffs procedure then sorts the hypotheses by ordering the magnitudes in a non-decreasing fashion, $|W_{\pi_1}| \leq \ldots \leq |W_{\pi_k}| \leq \ldots |W_{\pi_p}|$, and sequentially examines the hypotheses as follows: at each step $k = 0, 1, 2, \ldots, p - 1$, assume we select all remaining candidate hypotheses $\pi_j$ for which $j > k$ and $W_{\pi_j} > 0$. Then the number of false discoveries would be $\# \{ j : j > k, W_{\pi_j} > 0, \pi_j \in \mathcal{H}_0 \}$. We do not have access to this number since we do not know whether a hypothesis is null or not. However, note that by symmetry of the null scores,

$$\# \{ j : j > k, W_{\pi_j} > 0, \pi_j \in \mathcal{H}_0 \} \approx \# \{ j : j > k, W_{\pi_j} < 0, \pi_j \in \mathcal{H}_0 \} \leq \# \{ j : j > k, W_{\pi_j} < 0 \}.$$ 

Hence, the quantity

$$\widehat{\text{FDR}}_+(k) := \frac{1 + \sum_{j > k} 1_{\{ W_{\pi_j} < 0 \}}}{\left( \sum_{j > k} 1_{\{ W_{\pi_j} > 0 \}} \right) \lor 1}$$

may be regarded as a (conservative) estimate of the false discovery proportion (FDP) among the unexamined hypotheses. Set $[p] = \{ 1, \ldots, p \}$. Then the procedure is stopped at time $T_+$, where

$$T_+ := \inf \{ k \in [p] : \widehat{\text{FDR}}_+(k) \leq q \},$$

with the convention $\inf \emptyset = \infty$. Note that the stopping time $T_+$ corresponds to the stopping time $\tau$ we see earlier in Figure 1. The final selected set is the family of remaining candidate hypotheses, i.e. $\hat{S} = \{ \pi_j : j > T_+, W_{\pi_j} > 0 \}$. Candès et al. (2018); Barber and Candès (2015)

---

3The features with $W_j = 0$ will never be selected or used by the procedure so we exclude them in the definitions.
established that this procedure achieves FDR control at the nominal level $q$. Alternatively, the quantity

$$\text{FDR}_0(k) := \frac{\sum_{j>k} 1\{W_j < 0\}}{(\sum_{j>k} 1\{W_j > 0\})^{\frac{1}{q}}}$$

is a slightly less conservative estimate of FDR. Replacing $\text{FDR}_+$ with $\text{FDR}_0$ and replacing $T_+$ with

$$T_0 := \inf\{k \in [p] : \text{FDR}_0(k) \leq q\},$$

yields control of a modified version of FDR defined as

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

Figure 2a illustrates how the model-X knockoff procedure orders and sequentially examines the hypotheses, stopping when $\text{FDR}_0$ is below the pre-specified threshold $q$.

We now briefly describe the computation of the feature importance statistics. Throughout the paper, assume we are given $n$ i.i.d. samples from $P_X \cdot P_{Y|X}$. For each sample $(X,Y)$, we augment the dataset by constructing a knockoff copy $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p) \in \mathbb{R}^p$ for $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ (each original feature $X_j$ has a knockoff copy $\tilde{X}_j$). This construction obeys two properties: first, $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ is independent of $Y$ conditional on $X$; second, the joint distribution $(X, \tilde{X})$ remains invariant if we swap $X_j$ and $\tilde{X}_j$, for any $j \in \mathcal{H}_0$. Formally, $(X_j, \tilde{X}_j) | X_{-j}, \tilde{X}_{-j} \overset{d}{=} (\tilde{X}_j, X_j) | X_{-j}, \tilde{X}_{-j}$. How to construct good knockoffs is an expanding area of research, see e.g. Candès et al. (2018); Sesia, Sabatti and Candès (2018); Gimenez, Ghorbani and Zou (2019); Liu and Zheng (2018); Romano, Sesia and Candès (2020); Bates et al. (2020). In this paper, we will mainly be using the Gaussian (Candès et al., 2018) and HMM knockoffs (Sesia, Sabatti and Candès, 2018), and point the reader to these references for details.
The knockoff variables should be thought of as some sort of negative controls. When the statistician wants to evaluate the effect of each covariate on the response, she usually runs an algorithm on \((X, Y)\) — the covariate matrix and the response vector — and obtains an importance score \(Z_j\) for each feature \(j\). For example, \(Z_j\) can be the magnitude of the Lasso coefficient for \(X_j\), with the value of the regularization parameter determined by cross-validation. Now the knockoffs procedure asks our statistician to run her algorithm on both the original and the knockoff features. She will now obtain two scores \(Z_j\) and \(\tilde{Z}_j\) for each feature. In our previous example, the first is the magnitude of the Lasso coefficient for \(X_j\) and the second that for \(\tilde{X}_j\) (we are still free to determine the value of the regularization parameter by cross-validation if we wish). She then combines these two scores into a single one as follows:

\[
W_j = w_j(Z_j, \tilde{Z}_j);
\]

here, \(w_j\) is any anti-symmetric function she wants to use (e.g., \(W_j = Z_j - \tilde{Z}_j\)). By construction, if \(j\) is a null, \(W_j\) has equal probability of being positive or negative, whereas if \(j\) is not null, we hope that \(W_j\) tends to be large and positive.

4. **The adaptive knockoff filter.** As discussed before, the ordering \(\{\pi_k\}_{k \in [p]}\) is a key element in the knockoff procedure. If we know a priori that some hypotheses are more likely to be non-nulls and move them towards the end of the ordering, the procedure is more likely to select these features. Now suppose side information associated with the features under study is available. We would like to know

(a) how we can effectively use the data and side information to construct an ordering that has higher density of non-nulls at the end (as to improve power),

(b) and what property should the ordering have so that the FDR remains controlled?

Informally, in order to keep FDR control, we require the ordering to be independent of the signs of the statistics. Let \(U_j \in \mathbb{R}^r\) denote the side information associated with feature \(j\) and \(U = (U_1, \ldots, U_p)^T\). Let \(V^+(k)\) (resp. \(V^-(k)\)) denote the null features with positive (resp. negative) test scores \(W_j\) that have not been examined up to and including step \(k\). Define the filtration \(\{\mathcal{F}_k\}_{k \geq 0}\), where \(\mathcal{F}_k\) is the \(\sigma\)-algebra generated by the following elements:

- The magnitude of all the \(W_j\)’s: \(\{|W_j|\}_{j \in [p]}\).
- The signs of the revealed feature importance statistics: \(\{\text{sign}(W_{\pi_j})\}_{j \leq k}\) (when \(k = 0\), this is the empty set).
- The signs of the non-null \(W_j\)’s: \(\{\text{sign}(W_j)\}_{j \in \mathcal{H}_k}\).
- The number of positive and negative null feature importance statistics in the sets \([p]\setminus\{\pi_l : l \leq j\}\) for all \(j \leq k\): \(\{|V^+(j)|\}_{j \leq k}\) and \(\{|V^-(j)|\}_{j \leq k}\).
- Side information: \(U\).

**PROPERTY 1** (Sign invariant property). An ordering \(\{\pi_j\}_{j \in [p]}\) is called **sign invariant** if for any \(k \geq 0\), conditional on \(\mathcal{F}_k\) and \(\pi_{k+1} \in V^+(k) \cup V^-(k)\), the probability of \(W_{\pi_{k+1}} > 0\) is equal to \(|V^+(k)|/(|V^+(k)| + |V^-(k)|))\).

Algorithm 1 presents the adaptive knockoffs procedure. At each step \(k = 0, 1, 2, \ldots\), the adaptive knockoffs method uses a filter \(\Phi_{k+1}\), required to be \(\mathcal{F}_k\)-measurable, to determine

\footnote{An anti-symmetric function is a function such that \(f(u, v) = -f(v, u)\).}
the least promising hypothesis among the remaining ones.\footnote{When the filter $\Phi_k$ has extra randomness, we combine the extra randomness with the original side information and consider the augmented side information and the corresponding augmented $\sigma$-field $\mathcal{F}_k$. By such treatment, $\Phi_k$ is measurable w.r.t. $\mathcal{F}_k$.} Under this condition, Proposition 1 shows that the resulting ordering $\{\pi_k = \Phi_k\}_{k \in [p]}$ obeys Property 1. The adaptive knockoffs procedure otherwise adopts the same FDR estimates as in (2) and (3), and is stopped the first time the estimate falls below the target threshold.

**Proposition 1.** Assume that conditional on the side information $U$, the null $W_j$'s have equal probability of being positive or negative, and that their signs are independent of each other and of those of the non-nulls. If for each $k \geq 0$, $\Phi_{k+1}$ is $\mathcal{F}_k$-measurable, then the ordering $\pi_k = \Phi_k$ obeys Property 1.

**Proof.** By assumption, $\pi_{k+1}$ is measurable w.r.t. $\mathcal{F}_k$ and consequently $\{\pi_{k+1} \in V^+(k) \cup V^-(k)\} \subset \mathcal{F}_k$. Apart from $(U, \{\{W_j\}_{j \in [p]}, \{\text{sign}(W_j)\}_{j \in \mathcal{H}_0}, \{\text{sign}(W_{\pi_j})\}_{j \leq k})$, $\mathcal{F}_k$ provides further information on only the number of “+”s and “−”s in $V^+(k) \cup V^-(k)$; i.e. $|V^+(k)|$. Since the signs of the nulls $W_j \not= 0$ are i.i.d. coin flips conditional on $(U, \{\{W_j\}_{j \in [p]}, \{\text{sign}(W_j)\}_{j \in \mathcal{H}_0}, \{\text{sign}(W_{\pi_j})\}_{j \leq k})$, the probability of $W_{\pi_{k+1}} > 0$ (resp. $W_{\pi_{k+1}} < 0$) is proportional to the number of “+”s (resp. “−”s), completing the proof.

**Remark 1.** We discuss the construction of test statistics $W_j$ obeying the conditions required in Proposition 1. First consider the case where $U \perp (X,Y)$. If we construct knockoffs and compute the test statistics $W_j$ as in the vanilla knockoffs procedure introduced in Section 3, then under the assumption required by vanilla knockoffs, we have that for each $j \in \mathcal{H}_0$, $W_j$ has equal probability of being positive or negative, thus satisfying the conditions required for Proposition 1. Next, when $U$ is not independent of $(X,Y)$, we consider only the case where the null variables remain null conditional on $U$; this says that for each $j \in \mathcal{H}_0$, $Y \perp X_j \mid (X_{-j}, U)$. In this case, we simply construct knockoffs $\tilde{X} \in \mathbb{R}^p$ satisfying conditional exchangeability: this means that the conditional distribution $(X, \tilde{X}) \mid U$ stays invariant after swapping $X_j$ and $\tilde{X}_j$. If we also construct our knockoffs in such a way that $\tilde{X}$ is independent of $Y$ conditional on $(X, U)$, then the feature importance statistics $W$ computed with $(X, \tilde{X}, Y)$ satisfy the conditions required in Proposition 1.

As a result of Proposition 1, we show in Theorem 1 that the adaptive knockoffs procedure controls the finite-sample FDR.

**Theorem 1.** Under the conditions from Proposition 1, when $\widehat{\text{FDR}}_+$ is used, Algorithm 1 controls the FDR at the nominal level $q$; when $\widehat{\text{FDR}}_0$ is used, it controls the modified FDR at level $q$.

This result is a generalization of the condition for FDR control in the knockoffs framework presented in Barber and Candès (2015) and Candès et al. (2018). The vanilla knockoff filter, which uses only the magnitude of feature importance statistics to determine the order, can be viewed as a special case of adaptive knockoffs: in this case,

$$\Phi_{k+1} = \arg\min_{j > k} |W_{\pi_j}|,$$

which is clearly $\mathcal{F}_k$-measurable.
Algorithm 1: Adaptive Knockoffs

**Input:** Covariate matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$; response variables $\mathbf{Y} \in \mathbb{R}^n$; side information $U \in \mathbb{R}^{p \times r}$; target FDR level $q$.

**Initialization:** $k \leftarrow 0$; $\widehat{\text{FDR}}$ is either $\widehat{\text{FDR}}_0$ or $\widehat{\text{FDR}}_+$. 

while $\widehat{\text{FDR}}(k) > q$ and $k < p$ do

1. Use the filter $\Phi_{k+1}$ to determine the next hypothesis to examine $\pi_{k+1}$:
   
   $$\pi_{k+1} \leftarrow \Phi_{k+1}\left(\{W_j\}_{j \in [p]}, \{W_{\pi_l}\}_{j \leq k}, U\right).$$

2. Update $k$: $k \leftarrow k + 1$.

**Output:** Selected set $\hat{S} = \{j \in [p]: W_j > 0\}\backslash \{\pi_l: l \leq k\}$.

---

**Proof of Theorem 1.** When $\widehat{\text{FDR}}_+$ is used,

$$\text{FDR} = \mathbb{E}\left[\frac{|\hat{S} \cap H_0|}{|\hat{S}|}\right] = \mathbb{E}\left[\frac{|V^+(T_+)|}{|V^+(T_+)| + 1}\frac{\widehat{\text{FDR}}_+(T_+)}{}\right]$$

$$\leq q \mathbb{E}\left[\frac{|V^-(T_+)|}{|V^+(T_+)| + 1}\right] \leq q.$$

The second inequality holds by definition of $T_+$ and the third inequality follows from the fact that $|V^-(k)|/|V^+(k)| + 1$ is a supermartingale and $T_+$ a stopping time w.r.t. the filtration $\{\mathcal{F}_k\}_{k \geq 0}$. The supermartingale argument follows directly from Proposition 1 and Barber and Candès (2015, Section A.1). The proof of mFDR control is exactly the same as in Barber and Candès (2015, Section A.2).

---

5. Two classes of filters. We now focus on constructing a filter that satisfies Property 1 and also systematically uses all the available information to determine the ordering of hypotheses. At each step $k$, the filter determines the “least promising” hypothesis among the unexamined hypotheses based on the information in $\mathcal{F}_k$. We present two types of filters that quantify “least promising” in different ways. We emphasize that the model we choose does not affect the FDR control as long as Property 1 is satisfied, and researchers are free to come up with other types of models. In the following, we refer to this situation with the slogan: “Wrong models do not hurt FDR control!” We also assume we work with standardized side information $U_j$’s, which means that the $U_j$’s have the same dimension and units.

5.1. Predictive modeling. At step $k$, we estimate the probability that the sign of a feature importance statistic is negative conditional on $\mathcal{F}_k$. Specifically, we let $s_j = \text{sign}(W_j)$ and compute an estimate of $\mathbb{P}(s_j = -1|\mathcal{F}_k)$ for each remaining feature. This estimation (or prediction) task can be handled by various machine learning algorithms. We treat $\{s_j\}_{j=1}^p$ as the binary responses and the magnitude $\{|W_j|\}_{j=1}^p$ and side information $\{U_j\}_{j=1}^p$ (e.g., $U_j$ is the prior rank of $H_j$) as predictors. We consider the model

$$g(\mathbb{P}(s_j = -1 | |W_j|, U_j)) = h(|W_j|, U_j),$$
where \( g(x) = \log(x/(1-x)) \) is the link function and \( h(\cdot, \cdot) \) is a regression function. If we postulate a logistic regression model,

\[
h(|W_j|, U_j) = \beta_0 + \beta_1|W_j| + \beta_2^T U_j.
\]

For a generalized additive model (GAM), (Hastie, 2017),

\[
h(|W_j|, U_j) = \beta_0 + h_0(|W_j|) + h_1(U_{j1}) + \ldots + h_r(U_{jr}),
\]

where \( h_0, h_1, \ldots, h_r \) are smooth functions from \( \mathbb{R} \) to \( \mathbb{R} \). The function \( h \) can also be modeled via random forests (Breiman, 2001).

We use \( \{s_{\pi_j} \}_{j \leq k}, \{U_j \}_{j \in [p]}, \{|W_j| \}_{j \in [p]} \) as training data to fit the chosen model, and the fitted function \( \hat{h} \) for predicting the signs of statistics among the unexamined hypotheses. For \( j > k \), set

\[
\hat{\mathbb{P}}(s_{\pi_j} = -1|W_{\pi_j}|, U_{\pi_j}) = g^{-1} \circ \hat{h}(|W_{\pi_j}|, U_{\pi_j}),
\]

and

\[
\Phi_{k+1} = \operatorname{argmax}_{j > k} g^{-1} \circ \hat{h}(|W_{\pi_j}|, U_{\pi_j}) = \operatorname{argmax}_{j > k} \hat{h}(|W_{\pi_j}|, U_{\pi_j})
\]

since \( g \) is increasing. By construction, \( \Phi_{k+1} \) is \( \mathcal{F}_k \)-measurable.

5.2. Bayesian modeling. An alternative perspective, which has the benefit of allowing for a careful modeling of the effect of side information, is of a Bayesian nature. However, we are not imposing any assumption on the data generating mechanism. We are simply using Bayesian thinking for calculating the probability of a feature being non-null, and whether the Bayesian beliefs about features are true or not does not hurt FDR control. A belief closer to the truth will yield higher power in detecting the non-nulls.

The model. The Bayesian-oriented filter is similar to the treatment in Lei and Fithian (2018), but we consider it for knockoffs. Let \( S_j \) denote whether or not feature \( j \) is a null: if \( S_j = 1 \) means feature \( j \) is a non-null and if \( S_j = 0 \) means it is a null. We follow the Bayesian two-group model and write

\[
S_j \mid U_j \overset{i.i.d.}{\sim} \text{Bern}(\nu(U_j)),
\]

where \( \nu(\cdot) \) is a link function. Marginally,

\[
W_j \mid S_j, U_j \sim \begin{cases} 
\mathcal{P}_1(W_j \mid U_j) & \text{if } S_j = 1, \\
\mathcal{P}_0(W_j \mid U_j) & \text{if } S_j = 0.
\end{cases}
\]

Above, \( \mathcal{P}_{S_j}(\cdot \mid U_j) \) denotes the law of \( W_j \) conditional on \( U_j \) when \( S_j \in \{0, 1\} \). Under this model, we can quantify the possibility of a feature being null by inspecting the posterior probability \( \mathbb{P}(S_j = 0 \mid |W_j|, U_j) \). At each step \( k \), the posterior probability can be used as a criterion to determine the next hypothesis in the ordering, i.e.,

\[
\Phi_{k+1} = \operatorname{argmax}_{j > k} \mathbb{P}(S_{\pi_j} = 0 \mid |W_{\pi_j}|, U_{\pi_j}).
\]

The remaining task is to model \( \mathcal{P}_0, \mathcal{P}_1 \) and \( \nu \). Assuming \( W_j \) has a distribution with a point mass at 0, we model the conditional law of \( W_j \) via

\[
p_s(w \mid u) = \delta_s \cdot 1_{\{w=0\}} + (1 - \delta_s) \cdot 1_{\{w \neq 0\}} \cdot \frac{\beta_s(u) \cdot \exp(\beta_s(u) \cdot w)}{(1 + \exp(w))^\beta_s(u+1)}, \quad s = 0, 1,
\]

\footnote{In the case where \( W_j \) can also be 0, we can alternatively use a multinomial model with levels \( \{-1, 0, 1\} \).}

\footnote{In implementation, we instead use \( 1 - \mathbb{P}(S_j = 1, \text{sign}(W_j) > 0 \mid |W_j|, U_j) \).}
where $\beta_0(\cdot)$ and $\beta_1(\cdot)$ are functions of the side information. The continuous part of the distribution is somewhat arbitrary, and we choose this form for computational convenience. Under this model,

$$
\begin{align*}
\mathbb{E}[S \mid U] &= \nu(U), \\
\mathbb{E}[Y \mid W \neq 0, U, S = s] &= 1/\beta_s(U),
\end{align*}
$$

where $Y_j = \log \left(1 + \exp(W_j)\right) - W_j$. Then estimating $(\nu(U), \beta_0(U), \beta_1(U))$ boils down to estimating the above conditional expectations.

**GLM-based approach.** Let $\mathcal{N}$ (resp. $\mathcal{B}$) denote the class of functions $\nu(\cdot)$ (resp. $\beta_0(\cdot)$, $\beta_1(\cdot)$) belongs to. For example, assuming a logistic regression model, we have

$$
\mathcal{N} = \left\{ \nu(x) : \nu(x) = 1/(1 + \exp(-\theta^T x)), \ \theta \in \mathbb{R}^r \right\}
$$

while a model for $\mathcal{B}$ might be

$$
\mathcal{B} = \left\{ \beta(x) : \beta(x) = \exp(\theta^T x), \ \theta \in \mathbb{R}^r \right\}.
$$

The log-likelihood function (under independence) of $\{(S_j, W_j)\}_{j \in [p]}$ conditional on $\{U_j\}_{j \in [p]}$ is given by

$$
\ell(\{(S_j, W_j)\}_{j \in [p]} \mid \{U_j\}_{j \in [p]}; \delta_0, \delta_1, \nu(\cdot), \beta_0(\cdot), \beta_1(\cdot)) = \sum_{j=1}^{p} \left[ \left( i + ii + iii + iv + v \right) \right] + C,
$$

where $C$ represents the terms not containing the parameters and group 1 includes

$$
\begin{align*}
(i) &= (1 - S_j) \cdot \log(\delta_0) \cdot 1_{\{W_j = 0\}} + (1 - S_j) \cdot \log(1 - \delta_0) \cdot 1_{\{W_j \neq 0\}}, \\
(ii) &= S_j \cdot \log(\delta_1) \cdot 1_{\{W_j = 0\}} + S_j \cdot \log(1 - \delta_1) \cdot 1_{\{W_j \neq 0\}}.
\end{align*}
$$

Group 2 comprises

$$
\begin{align*}
(iii) &= S_j \cdot \log(\nu(U_j)) + (1 - S_j) \cdot \log(1 - \nu(U_j)), \\
(iv) &= (1 - S_j) \cdot 1_{\{W_j \neq 0\}} \cdot \left( \log(\beta_0(U_j)) + \beta_0(U_j) \cdot \log\left(\exp(W_j)/(1 + \exp(W_j))\right) \right), \\
(v) &= S_j \cdot 1_{\{W_j \neq 0\}} \cdot \left( \log(\beta_1(U_j)) + \beta_1(U_j) \cdot \log\left(\exp(W_j)/(1 + \exp(W_j))\right) \right).
\end{align*}
$$

In the case where $\nu(\cdot), \beta_0(\cdot), \beta_1(\cdot)$ are classes of parametric functions as above, we hope to obtain the maximum likelihood estimator (MLE) by optimizing the log-likelihood function:

$$
(\hat{\delta}_0, \hat{\delta}_1, \hat{\nu}(\cdot), \hat{\beta}_0(\cdot), \hat{\beta}_1(\cdot)) = \arg\max_{\delta_0, \delta_1, \nu(\cdot) \in \mathcal{N}, \beta_0(\cdot), \beta_1(\cdot) \in \mathcal{B}} \sum_{j=1}^{p} \left[ (i + ii + iii + iv + v) \right].
$$

Note that at step $k$, the information we can use to estimate the parameters is limited: some of the signs of the $W_j$'s are not available and the $S_j$'s are unobserved.

Directly optimizing the log-likelihood function is not feasible. Instead, we use the expectation-maximization (EM) algorithm to obtain the MLE. Our plan is this: at step $k$ of the adaptive knockoffs algorithm, we run the EM algorithm for $B$ iterations and obtain an estimate of the parameters of interest. (To be clear, one iteration of the EM algorithm consists of an E-step and an M-step.) At step $b$ of the EM algorithm, denote by $\mathcal{G}$ the $\sigma$-field generated...
by the available information. For the E-step we need to compute the following conditional expectations:

$$\mathbb{E}[S_j \mid \mathcal{G}], \mathbb{E}[Y_j \cdot S_j \mid \mathcal{G}], \mathbb{E}[Y_j \cdot (1 - S_j) \mid \mathcal{G}].$$

We defer the calculation of the above quantities to Supplementary Section S1 and set $$S_j = \mathbb{E}[S_j \mid \mathcal{G}].$$ For the M-step, we decompose the optimization into two subgroups. The optimization problems in group 1 have analytical solutions, namely,

$$\hat{\delta}_0 = \argmax_{\delta_0} \sum_{j=1}^{p} (i) = \frac{\sum_{j \in \mathcal{H}}(1 - S_j) \cdot 1\{W_j = 0\}}{\sum_{j \in \mathcal{H}}(1 - S_j)},$$

$$\hat{\delta}_1 = \argmax_{\delta_1} \sum_{j=1}^{p} (ii) = \frac{\sum_{j \in \mathcal{H}} S_j \cdot 1\{W_j = 0\}}{\sum_{j \in \mathcal{H}} S_j}.$$

The optimization problems in group 2 update $$(\nu(\cdot), \beta_0(\cdot), \beta_1(\cdot)).$$ Since the optimization problem is separable, we can solve the three subproblems independently.

$$\hat{\nu}(\cdot) = \argmax_{\nu(\cdot) \in \mathcal{N}} \sum_{j=1}^{p} (iii), \quad \hat{\beta}_0(\cdot) = \argmax_{\beta_0(\cdot) \in \mathcal{B}} \sum_{j=1}^{p} (iv), \quad \hat{\beta}_1(\cdot) = \argmax_{\beta_1(\cdot) \in \mathcal{B}} \sum_{j=1}^{p} (v).$$

These three subproblems directly depends on $\mathcal{N}$ and $\mathcal{B}.$ When the parametric model as in (5) and (6) is used, the above optimization problems correspond to three (weighted) GLMs respectively and can be solved by standard R packages (e.g., glm).

**GLM-extension approach.** Another possibility is to work with regularized log-likelihood functions. For instance, we may add an $\ell_1$ penalty on the coefficients $\theta$ in (6), and use the glmnet package to solve the corresponding optimization problem. We can also fit a generalized additive model by for $\beta_0(\cdot)$ solving the following penalized optimization problem (Hastie, Tibshirani and Friedman, 2009, Chapter 9):

$$\max_{\beta_0(\cdot) \in \mathcal{B}} \sum_{j} (iv) - \sum_{\ell=1}^{r} \lambda_{\ell} \int \beta''_{0\ell}(x_\ell) dx_\ell,$$

in which $\mathcal{B} = \{\beta(x_1, \ldots, x_r) : \beta(x_1, \ldots, x_r) = \sum_{\ell=1}^{r} \beta_{\ell}(x_\ell), \beta''_{0\ell}(\cdot) \text{ exists for all } \ell \in [r]\}.$ Above, the nonnegative hyper-parameters $\{\lambda_{\ell}\}_{\ell \in [r]}$ can be chosen via Generalized Cross Validation (GCV). The R package gam or mgcv are designed to find solutions to such problems.

**Nonparametric regression approach.** We consider a variation that does not fall in the EM framework but allows us to make use of flexible regression tools. Recall (4), which states that $$(\nu(\cdot), \beta_0(\cdot), \beta_1(\cdot))$$ are functions of the conditional expectations. We thus directly estimate the conditional expectation instead of solving the optimization problems in group 2. For example, we can use non-parametric methods, e.g., a random forest, to directly fit the conditional expectations and let the fitted values be the updated parameters. This is not an M-step because we are no longer optimizing the (expected) likelihood. (This is not a concern since FDR control always holds.) Such a variation opens the door to modern regression methods and often works well in practice as we shall see later.

**Default implementation in our R-package.** The methods we have presented differ in the way they estimate $\nu(\cdot), \beta_0(\cdot)$ and $\beta_1(\cdot)).$ When the side information is a scalar, the default implementation in our R-package\footnote{The R-package to implement adaptive knockoffs is available at \url{https://github.com/zhimeir/adaptiveKnockoffs}} combines the GLM-extension and nonparametric regression.
approaches. In details, we fit $\beta_0(\cdot), \beta_1(\cdot)$ via the gam package in R whereas for $\nu$, we regress $\log \left( \frac{\mathbb{E}_j}{\mathbb{E}_j(1 - \mathbb{E}_j)} \right)$ on $U_j$ via a GAM and then transform the fit to produce $\hat{\nu}(\cdot)$. When the dimension is higher, the default implementation is the nonparametric regression approach with a random forest. The default number of iterations $B$ is set to be one.

Initialization. At the beginning of Algorithm 1, we reveal a fraction (by default 10%) of the hypotheses based only on the magnitude of the statistics $|W_j|$ corresponding to the lowest values. Denote the revealed statistics by $W^{\text{reveal}}$ (the set of $W_j$’s, for which hypothesis $j$ is revealed). The adaptive filter then starts with rough guesses $(\hat{\beta}_0(\cdot), \hat{\beta}_1(\cdot), \hat{\nu}(\cdot), \hat{\delta}_0, \hat{\delta}_1)$ computed from available information. Specifically, we initialize $\hat{\nu}(\cdot)$ with a constant function set to $|\{j : W_j > 0\}|/p$ (we can think of a $|\{j : W_j > 0\}|$ as a very liberal estimate of the number of non-nulls). Further, we set

$$\hat{\delta}_0 = \frac{|\{j : W_j \leq 0\}|}{p}, \quad \hat{\delta}_1 = \frac{|\{j : W_j > 0\}|}{p}.$$

Finally, the initial values of $(\hat{\beta}_0, \hat{\beta}_1)$ are given by

$$\hat{\beta}_0(U_j) = \hat{\beta}_1(U_j) = 1/\log(2), \quad \text{if } W_j = 0.$$

$$\hat{\beta}_0(U_j) = 1/\left( \log \left( 1 + \exp(\bar{W}^-_{\text{reveal}}) \right) - \bar{W}^-_{\text{reveal}} \right),$$

$$\hat{\beta}_1(U_j) = 1/\left( \log \left( 1 + \exp(\bar{W}^+_{\text{reveal}}) \right) - \bar{W}^+_{\text{reveal}} \right), \quad \text{if } W_j \neq 0.$$

Above, $\bar{W}^-_{\text{reveal}}$ is the average of the negative items in $W_{\text{reveal}}$ and $\bar{W}^+_{\text{reveal}}$ is the average of the positive items in $W_{\text{reveal}}$. That is, we approximate $\beta_0(U_j)$ (resp. $\beta_1(U_j)$) with $1/Y_j$, in which we impute nonzero $W_j$’s with the average of the negative (resp. positive) items in $W_{\text{reveal}}$.

In the subsequent steps of the filter, the initial value of the tuple $(\hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\delta}_0, \hat{\delta}_1)$ in Algorithm 2 is the output from the previous iteration. The complete procedure is described in Algorithm 2.


6.1. General setting. To evaluate the performance of adaptive knockoffs, we present two numerical experiments with different types of side information. In each setting, we compare adaptive knockoffs with other multiple testing methods. Table 2 lists all the candidate methods and their properties, i.e., whether or not they depend on p-values and whether or not they utilize side information. In our experiments all the p-values are obtained from multivariate linear regression. Storey-BH is implemented with a threshold set to $\tau = 0.5$. The parameter of SABHA follows Li and Barber (2019) with $\epsilon = 0.1$ and $\tau = 0.5$. For Adaptive SeqStep, the threshold $\lambda$ is set to be 0.5 as in Lei and Fithian (2016). For AdaPT, we follow the setup introduced in https://cran.r-project.org/web/packages/adaptMT/vignettes/adapt_demo.html. The knockoff-based algorithms in Table 2 use the LCD feature importance statistics as introduced in Candès et al. (2018) and $\overline{\text{FDR}}_+$ as the estimated FDR.

For both experiments, we run algorithms with target FDR levels $\{0.03, 0.06, \ldots, 0.3\}$ and compare the corresponding statistical power and realized FDR. All the presented results are averaged over 100 trials. The simulation results can be reproduced with the code provided at https://github.com/zhimeir/adaptive_knockoff_paper.

---

Algorithm 2: EM algorithm to estimate \( p_0, p_1, \nu \)

**Input:** Information \( \mathcal{F}_k \) at step \( k \).

**Initialization:** initialize \((\hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\delta}_0, \hat{\delta}_1)\) as in Section 5.2 and set \( \mathcal{G} \leftarrow \sigma(\mathcal{F}_k, \hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\delta}_0, \hat{\delta}_1) \).

for \( b \leftarrow 0, \ldots, B - 1 \) do

1. **E-step:**
   
   Update \( \mathcal{S}_j : \mathcal{S}_j \leftarrow \mathbb{E}[S_j \mid \mathcal{G}], \quad j \in [p]. \)
   
   Update \((\mathcal{T}_{0,j}, \mathcal{T}_{1,j}) : \mathcal{T}_{s,j} \leftarrow \frac{\mathbb{E}[Y_j \cdot (s \cdot S_j + (1 - s) \cdot (1 - S_j)) \mid \mathcal{G}]}{s \cdot S_j + (1 - s) \cdot (1 - S_j)}, \quad s = 0, 1, \)

   where the calculations of the conditional expectations are presented in Supplementary Section S1.

2. **M-step:**

   Update \((\hat{\beta}_0, \hat{\beta}_1) : \hat{\beta}_s \leftarrow \frac{\sum_{j \in \mathcal{H}} (s \cdot \mathcal{S}_j + (1 - s) \cdot (1 - \mathcal{S}_j)) \cdot \mathbf{1}(W_j = 0)}{\sum_{j \in \mathcal{H}} (s \cdot \mathcal{S}_j + (1 - s) \cdot (1 - \mathcal{S}_j))}, \quad s = 0, 1. \)

   Update \( \hat{\nu} : \hat{\nu} \leftarrow \text{random forest (} \mathcal{S}_j \sim U_j \text{)}, \)

   Update \((\hat{\delta}_0, \hat{\delta}_1) : \frac{1}{\beta_s} \leftarrow \text{random forest}(\mathcal{T}_{s,j} \mid W_j \neq 0 \sim U_j), \quad s = 0, 1. \)

3. **Update current information:** \( \mathcal{G} \leftarrow \sigma(\mathcal{F}_k, \hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\delta}_0, \hat{\delta}_1) \).

end

**Output:** \((\hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\delta}_0, \hat{\delta}_1)\).

<table>
<thead>
<tr>
<th>Method</th>
<th>Abbreviation</th>
<th>P-value free?</th>
<th>Use side information?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benjamin Hochberg (Benjamini and Hochberg, 1995)</td>
<td>BHq</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storey’s BH (Storey, 2002)</td>
<td>StoreyBH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive SeqStep (Lei and Fithian, 2016)</td>
<td>AdaSeqStep</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>AdapPT (Lei and Fithian, 2018)</td>
<td>AdaPT</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Structure Adaptive BH algorithm (Li and Barber, 2019)</td>
<td>SABHA</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Vanilla Model-X knockoffs (Candès et al., 2018)</td>
<td>Vanilla Knockoff</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Adaptive knockoffs w/ GLM filter</td>
<td>AdaKn(GLM)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adaptive knockoffs w/ GAM filter</td>
<td>AdaKn(GAM)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adaptive knockoffs w/ Random Forest filter</td>
<td>AdaKn(RF)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Adaptive knockoffs w/ two group model</td>
<td>AdaKn(EM)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2: Candidate multiple testing methods and their properties.

6.2. Simulation 1: one-dimensional side information. The simulated dataset is of size \( n = 1000 \) and \( p = 900 \). Conditional on \( X \), \( Y \) is generated from a linear model

\[
Y \mid X_1, \ldots, X_p \sim \mathcal{N}(\beta_1 X_1 + \ldots + \beta_p X_p, 1).
\]

The covariates \( X \) are drawn from an HMM, whose parameters follow the instructions found at https://msesia.github.io/snpknock/articles/SNPknock.html. Researchers can reproduce our choices by following the link from Section 6.1. In this setting, our inferential goal is to test whether or not \( \beta_j = 0 \).

We specify the model by constructing a sparse regression sequence \( \beta \)—fixed throughout, i.e. through the 100 trials so that the data distribution \( P_{XY} \) does not change—as follows: we randomly choose 150 features among the first 300 as signals in such a way that the larger the index, the less likely it is to be selected.\(^\text{10}\) The setting is motivated by the fact that in many real

\(^{10}\) We draw i.i.d. samples from a distribution supported on \( \{1, 2, \ldots, 300\} \) such that \( j \) is selected with probability proportional to \( \frac{1}{j^p} \) until we obtain 150 distinct realizations.
applications, researchers have access to prior knowledge about the hypotheses, which allows them to rank the hypotheses by their chance of being of interest. For each signal $X_j$, we set $\beta_j = \pm 3.5 / \sqrt{n}$, where the signs are determined by independent coin flips (the features not in the model have $\beta_j = 0$). Figure 3 shows the realized configuration of the signals (the variables with nonzero regression coefficients). The side information is the index of the features; that is, $U_j = j$ for $j \in [p]$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{One-dimensional hypothesis structure.}
\end{figure}

In each trial, we draw a sample of size $n = 1000$ from $P_{XY}$ and run all candidate methods on this sample. Figure 4 shows the power and FDR of each method versus target FDR levels. All methods control FDR as we expected. The adaptive knockoffs method outperforms the vanilla knockoffs method and other p-value based procedures by a wide margin. We also plot the realized ordering of vanilla knockoffs and adaptive knockoffs (with our Bayesian filter) from one trial in Figure 5a and Figure 5b respectively. We can observe that the adaptive knockoffs method places more non-nulls towards the end of the ordering and, consequently, makes more true discoveries.

The p-value based methods perform unsatisfactorily here because p-values are of low quality. As an aside, we note that it is often challenging to obtain valid p-values, not to mention high quality ones; for instance, Dezeure et al. (2015) and Lei and Bickel (2021) explain that getting p-values from the simplest linear model in reasonably high dimensions is already a challenge if we do not impose stringent assumptions.

6.3. Simulation 2: two-dimensional side information. The simulated dataset is of size $n = 1000$ and $p = 1600$. Conditional on $X$, $Y$ is generated from a logistic model:

$$Y \mid X_1, \ldots, X_p \sim \text{Bernoulli} \left( \frac{\exp(\beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_1 X_1 + \ldots + \beta_p X_p)} \right).$$

The entries of $\beta$ ‘live’ on a two-dimensional plane and the location of $\beta_j$ on the plane is described by a pair of coordinates $(r(j), s(j))$, as in Figure 6. In all, there are $m = 201$ blue nodes, representing the nonzero entries of $\beta$. Details about the signal locations are in Supplementary Section S3. The magnitude of the nonzero entries is set to $25 / \sqrt{n}$ and the signs are generated via i.i.d. coin flips. The vector $X$ of covariates is drawn i.i.d. from a discrete-time Gaussian process with zero mean and covariance structure:

$$\text{Cov}(X_i, X_j) = e^{-3||U_i - U_j||^2}, \quad i, j \in [p],$$

where $U_j = (r(j), s(j))$. The side information is the pair of coordinates of each feature.
FIG 4. Power (left) and FDR (right) versus target FDR values.

FIG 5. (a) The realized ordering of vanilla knockoffs. (b) The Realized ordering of adaptive knockoffs with the Bayesian filter. The x-axis is the ordering index and the y-axis is $W$. The blue bars represent the non-nulls; the black bars represent the nulls. The dashed red lines correspond to the selection thresholds for $q = 0.2$, i.e., the features after the red line with positive signs are selected.

This simulation setting is motivated by magnetic resonance imaging (MRI) studies. For example, the hypotheses (nodes) correspond to the voxels in a structural MRI scan and the response is a 0-1 variable indicating whether the subject has Alzheimer’s disease. Due to the spatial correlation between the nodes, the signals often exhibit cluster structures and our setup presents a simplified version of such structures. Given the context, one may ask whether we should treat the clusters themselves rather than the voxels as unit of inference. The debate between cluster-based inference and voxel-based inference seems still ongoing in the neuroimaging society. In particular, researchers have recently observed that cluster-based inference often suffers from low specificity (we do not know how many significant voxels there are within a significant cluster) and neuroscientists are calling for inference methods with higher resolution (see e.g., Woo, Krishnan and Wager (2014); Rosenblatt et al. (2018)). Here we adopt the voxel-based inference as in Efron (2012).

In this simulation, $p$ is larger than $n$, and obtaining valid p-values is a problem. Hence, we here focus on comparing the knockoff-based methods. Figure 7 shows the power and FDR of all the candidate methods. Again all methods control the FDR as expected. The adaptive knockoffs method with a Bayesian filter or random forest filter outperform the vanilla knockoffs method by a wide margin. Figure 8a and 8b show the realized ordering of vanilla knockoffs and adaptive knockoffs with the Bayesian filter from one trial respectively. Adaptive knockoffs is able to place more non-nulls towards the end of the ordering and has higher power. The GLM and GAM filters have almost the same power as vanilla knockoffs because their models are too simple and cannot capture the two-dimensional structure of the side information.
7. Applications. In Section 2 we have presented the results of our method applied to the WTCCC dataset (Crohn’s disease). In this section, we discuss in detail the data analysis implementation, and in addition, apply our methods to the Northern Finland 1996 Birth Cohort study of metabolic syndrome (NFBC).

7.1. Overview of the data. We have already described the WTCCC dataset in Section 2. The NFBC dataset contains information on $n = 5402$ individuals from northern Finland that includes genotypes at approximately 300,000 SNPs and nine phenotypes. The exact number of effective observations are slightly different across phenotypes because some values are missing in each phenotype. In this paper, we focus on low-density lipoprotein (LDL) and...
high-density lipoprotein (HDL) phenotypes. The inferential goal is to discover SNPs significantly associated with LDL and HDL in the Finnish population.

7.2. Data pre-processing and SNP pruning.

Pre-processing. For the WTCCC Crohn’s disease dataset, we follow the pre-processing steps in Candès et al. (2018) and for the NFBC dataset, we follow the pre-processing steps in Sabatti et al. (2009); Barber and Candès (2019); Sesia, Sabatti and Candès (2018). Table 3 lists the number of SNPs left after pre-processing in the column named "p (pre-clustering)".

Clustering. After pre-processing we further conduct a clustering step to deal with the high correlation between SNPs. We follow the method in Candès et al. (2018); Sesia, Sabatti and Candès (2018); Barber and Candès (2019) to cluster SNPs and choose a representative from each cluster. Table 3 lists the number of SNP clusters in the column "p (post-clustering)". From now on, our inferential goal is to discover important SNP clusters.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Phenotype</th>
<th>n</th>
<th>p (pre-clustering)</th>
<th>p (post-clustering)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTCCC</td>
<td>CD</td>
<td>4913</td>
<td>377,749</td>
<td>71,145</td>
</tr>
<tr>
<td>NFBC</td>
<td>LDL</td>
<td>4682</td>
<td>328,934</td>
<td>59,005</td>
</tr>
<tr>
<td>NFBC</td>
<td>HDL</td>
<td>4700</td>
<td>328,934</td>
<td>59,005</td>
</tr>
</tbody>
</table>

Table 3: Description of the datasets.

7.3. Side information acquisition.

Crohn’s disease. As discussed in Section 2, we obtain the marginal p-values from inflammatory bowel disease (IBD) studies in East Asia and Belgium (Franke et al., 2010; Liu et al., 2015; Goyette et al., 2015) as side information. In case a SNP is recorded in both studies, we use a weighted mean of the p-values as the side information; we give a larger weight to the p-values from the East Asia study because it contains more samples (the weights are respectively $1 - 1/101$ and $1/101$). When a SNP in our dataset is not recorded in a study—that is, the summary statistic of the SNP is not reported by that study—we check if any other SNP in the cluster is recorded. We then use the minimum of the p-values as the side information if there are more than one neighboring SNP recorded. If no SNP is recorded, we impute the side information with a one.

Lipids. For HDL and LDL, we obtain summary statistics reported by Loh et al. (2018). Their results are based on the UK Biobank dataset, which comprises genetic information on a range of phenotypes of individuals from the UK. The genetic information in the UK population can serve as a reference for our study in the Finnish population. Explicitly, we obtain the association p-values reported for “self-reported high cholesterol level”. As before, if no p-value is found to match a SNP, we set the corresponding side information to be one, reflecting the lack of prior knowledge. The motivation here is to use previous results as much as possible in order to make more rejections. How to impute side information for previously unobserved variables remains an open question: for example, if the major goal of a scientist is to explore SNPs that were not previously studied, she might want to impute those missing p-values with smaller values to encourage new discoveries (see the discussion in Section 8).

11The summary statistics are downloaded from https://data.broadinstitute.org/alkesgroup/UKBB/.
7.4. Implementation details.

Knockoff construction. We use the HMM knockoffs from Sesia, Sabatti and Candès (2018) and follow their suggestion to set the number of latent haplotype clusters to twelve. Knockoffs are generated separately for 22 chromosomes and for the two datasets.

Feature importance statistics. Given the response $Y$ and augmented normalized covariate matrix $(X, \tilde{X})$, we perform Lasso regression of $Y$ on $(X, \tilde{X})$ and obtain Lasso coefficients $(\beta, \tilde{\beta})$. The penalty parameter $\lambda$ is chosen from a 10-fold cross validation. The resulting feature importance statistic for each SNP is the difference between the magnitude of the original and the knockoff Lasso coefficients, i.e., $W_j = |\beta_j| - |\tilde{\beta}_j|$.

Adaptive knockoff filter. Each SNP is associated with a p-value obtained from other studies. We do not directly feed the p-values to our filter but instead order the SNPs according to their p-values and use the ranks of the SNPs as input of our filter. We use the Bayesian two-group model filter introduced in Section 5.2 with the default setting except for the fact that in the initialization step, we reveal the features whose $|W_j|$ is below a pre-specified threshold. The threshold is 0.03 for Crohn’s disease, 0.005 for LDL and 0.0005 for HDL. As far as estimating the FDR, we use the less conservative $\tilde{FDR}_0$.

7.5. Results. We apply adaptive knockoffs with target FDR level $q = 0.1$. Since knockoff algorithms are essentially random and depend on the realizations of $\tilde{X}$, we generate 50 knockoffs independently conditioning on $(X, Y)$, with analysis conducted on every realization of $\tilde{X}$. Figure 9 presents the number of rejections made across multiple knockoff realizations for each phenotype. The average number of discoveries for Crohn’s disease has been reported in Table 1 and that for LDL and HDL are presented in Table 4. Additionally, we record the selection frequency of the SNPs—the selection frequency of a SNP is the number of algorithm runs in which it is selected divided by the total number of algorithm runs, and those with a selection frequency above 30% are listed in Tables S1-S3 in Supplementary Section S2. In practice, we suggest reporting those SNPs with a selection frequency at least 50% as the final selection set. We emphasize that this aggregated selection procedure is not guaranteed to control the FDR. The number of SNPs selected at least half of the time is reported in Table 1 (Crohn’s disease) and Table 4 (lipids). We compare our results with Sabatti et al. (2009) and Sesia, Sabatti and Candès (2018); the former adopts a marginal test with a p-value threshold of $5 \times 10^{-7}$, and the latter adopts a 0.1 target FDR level. The results show that our algorithm greatly improves the power of the original knockoff procedure.

We would like to stress once more that FDR control holds for single runs of adaptive knockoffs regardless of the quality of the p-values we use. Also, we are not merely re-discovering what is already known since side information concerns other populations.

<table>
<thead>
<tr>
<th>Method</th>
<th>HDL</th>
<th>LDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sabatti et al. (2009)</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>HMM knockoffs (Sesia, Sabatti and Candès, 2018)</td>
<td>8</td>
<td>9.8</td>
</tr>
<tr>
<td>Adaptive knockoffs</td>
<td>12.5 (11)</td>
<td>18.3 (18)</td>
</tr>
</tbody>
</table>

Table 4

(Average) number of SNP discoveries made by different methods with target FDR level $q = 0.1$. For knockoff-based algorithms, the reported number is averaged over multiple realizations of $X$ (for adaptive knockoffs, this number is 50). The number in parentheses is the number of SNPs being selected more than 50% of the times.
8. Discussion and future work.

Rediscovery. In the GWAS example, we mentioned that we are not rediscovering what has been discovered since the side information is obtained from different populations. We observe that SNPs consistently found across multiple populations are prioritized. This fact is due to 1) the form of our filter and 2) the similarity between the reference populations (the population(s) from which we obtain the side information) and the target population (the population of the data at hand). To explain this point, recall that at step $k$, the next hypothesis to be examined is determined by

$$
\Phi_{k+1} = \arg\max_{j>k} \hat{h}(|W_{\pi_j}|, U_{\pi_j}),
$$

where $\hat{h}(|W_{\pi_j}|, U_{\pi_j})$ quantifies how likely the sign of $W_{\pi_j}$ is non-positive. The rule $\hat{h}(\cdot, \cdot)$ is trained with the information in $\mathcal{F}_k$ to predict the signs of the unexamined test statistics $W_j$. Consequently, when the reference population is similar to the target population, the model will learn from the data to prioritize features discovered in the reference population (because those features previously discovered tend to have larger values of $W_j$ in the data at hand due to the similarity), hence encouraging re-discovery.

Suppose however that the reference population is very different from the target population—think of the extreme case where the nulls in one population are non-nulls in the other and vice versa. In this setting, at each step $k$, the model will learn from the data to not prioritize the features discovered in the reference population (since the previously discovered features tend to have negative $W_j$'s), hence discouraging re-discovery. We provide a numerical experiment in Supplementary Section S4.1 to demonstrate this point.

New discoveries. Consider another scenario where the prior information is obtained from the same population, and we wish to focus on making new discoveries, while controlling the error rate within the newly discovered covariates; this means that a discovery is counted only if it has not been discovered before and the FDR is the expected ratio between the number of false new discoveries and the number of new discoveries. In this case, we should discourage selecting what has been discovered before. This can be achieved via a modification of the filter $\Phi_k$. To be specific, consider a prior dataset yielding an importance score $U_j$ for SNP $j$ (a higher $U_j$ suggests that feature $j$ is more likely to be non-null) as the side information for all $j \in [p]$. Let $u_0$ be a threshold and put $\hat{S}_0 := \{j \in [p] : U_j \geq u_0\}$ to be the set of discoveries obtained from side information. Suppose now we want to discover features outside of $\hat{S}_0$, with the goal that the FDR in the newly discovered set is controlled at the pre-specified level $q$. To achieve this, we shall place the set of features in $\hat{S}_0$ at the beginning of the ordering and...
never stop the procedure until we finish examining all the features in \( \hat{S}_0 \). By design, features in \( \hat{S}_0 \) will never be selected.

Why is this a valid procedure? The modified procedure introduced above, is equivalent to 1) setting

\[
\tilde{h}(|W_j|, U_j) = \hat{h}(|W_j|, U_j) + \infty \cdot 1\{U_j \geq u_0\},
\]

with

\[
\Phi^{(1)}_{k+1} = \operatorname{argmax}_{j > k} \tilde{h}(|W_{\pi_j}|, U_{\pi_j}),
\]

and 2) letting the stopping time \( \tilde{T} \) be the first time such that features with \( U_j \geq u_0 \) have all been examined and the estimated FDP among the unexamined hypotheses is below \( q \). Since \( \Phi^{(1)}_{k+1} \) is still adapted to \( F_k \) and \( \tilde{T} \) is still a valid stopping time w.r.t. the filtration \( \{F_k\}_{k \geq 1} \), FDR control is guaranteed.

The two cases above are at two different ends of a spectrum: we either encourage rediscovery or forbid rediscovery. We can, in fact, consider situations in between. Let us continue the above example and suppose that we do not want to exclude the possibility of reselecting previously discovered SNPs. We however want to encourage discovering SNPs that show less evidence in the first dataset (smaller values of \( U_j \)), and discourage discovering those showing more evidence (larger values of \( U_j \)). One way to achieve this goal is to replace \( \hat{h}(|W_j|, U_j) \) with

\[
\tilde{h}(|W_j|, U_j) = (1 - \lambda) \cdot \hat{h}(|W_j|, U_j) + \lambda \cdot g(U_j),
\]

where \( \lambda \in [0, 1] \) and \( g(\cdot) \) is a monotonically non-decreasing function of \( U_j \). For example, \( g(U_j) \) can be as simple as \( U_j \). In general, we would like \( g(U_j) \) to be on the same scale as \( \hat{h}(|W_j|, U_j) \), so \( g(U_j) \) can take the form of \( c \cdot U_j \) where \( c \) is a normalizing constant. We then define the filter to be \( \Phi^{(2)}_{k+1} = \operatorname{argmax}_{j > k} \tilde{h}(|W_{\pi_j}|, U_{\pi_j}) \). Such a filter would prioritize SNPs with smaller values of \( U_j \). The larger \( \lambda \), the higher the emphasis on new discoveries. Similarly, if the goal is instead to encourage rediscovery, simply replace \( g(U_j) \) with a non-increasing function of \( U_j \). We provide in Supplementary Section S4.2 a numerical experiment to illustrate the effects of the two aforementioned filters.

**GWAS in the minority populations.** In this paper, we applied the adaptive knockoff procedure to GWAS in the British and Finnish population and obtained summary statistics from other populations. It will be interesting to apply our method in a setting where the inferential target is a minority population (e.g., African-Americans or Hispanic-Americans). In truth, minority populations are often under-represented in GWAS and these studies are, therefore, often underpowered. Since there are abundant genetic data from the European population, exploiting information from this population to empower GWAS in minority populations is becoming a popular research topic (see e.g., Coram et al. (2015, 2017)). Our method is tantalizing because we have seen how easily we can use GWAS statistics from one population to boost power in another.

**Beyond GWAS.** Knockoff-based procedures have been successfully applied to genetics, and we would like to see them used in other areas. One potential area is neuroimaging, a field in which researchers are interested in discovering locations in the brain that are associated with certain trauma. The neuroimage data (e.g., structural MRI) also has a spatial structure which can be used as side information. Applying adaptive knockoffs to such datasets promises important diagnostic information.
Acknowledgements. E.J.C. was partially supported by the National Science Foundation via grant DMS–1712800, by the Simons Foundation via the Math + X award, and by a generous gift from TwoSigma. Z. R. was partially supported by the same Math + X award. The authors are grateful to the anonymous referees for helpful suggestions that lead to an improvement of the paper. Z. R. thanks Stephen Bates, Nikolaos Ignatiadis, Eugene Katsevich and Matteo Sesia for their valuable comments on this project.

REFERENCES

WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 661.