Knockoffs with Side Information

Zhimei Ren[1] and Emmanuel Candès[1,2]

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

January 21, 2020

Abstract

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. *Adaptive knockoffs* 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, adaptive knockoffs makes more discoveries than reported in previous studies on the same datasets.

Keywords. Multiple testing, variable selection, false discovery rate (FDR), knockoff filters, Bayesian two-group model, genome-wide association study (GWAS).

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 $Y|X$. 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 $Y|X$ 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 $\mathcal{H}_1 = \{1, ..., p\} \backslash \mathcal{H}_0$. 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),

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

where $a \vee b = \max(a, b)$. Above, $\hat{\mathcal{S}} \subset \{1, \ldots, p\}$ 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 et al. (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 et al. (2017); Sur and Candès (2019)). *Model-X knockoffs*, 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 $Y|X$. 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 family of covariates, namely, $P_{Y|X}$ which is 1) usually unknown and 2) the actual object of inference (Janson, 2017). For instance, model-X knockoffs 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 knockoffs and generalizes it to a setting where side information about the variables or factors under study happens to be available.

1.2 Related works

Previous works on multiple testing with side information broadly fall 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 hypothese—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 of works 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 et al. (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 et al. (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 model-X knockoffs, which is completely different, and use side information to adaptively screen knockoff importance statistics instead. An appealing feature is that FDR control is achieved under the same conditions as for (vanilla) model-X knockoffs: we (only) ask for the knowledge of the distribution P_X of the covariates, which is resasonable in many situations (Candès et al., 2018).[1]

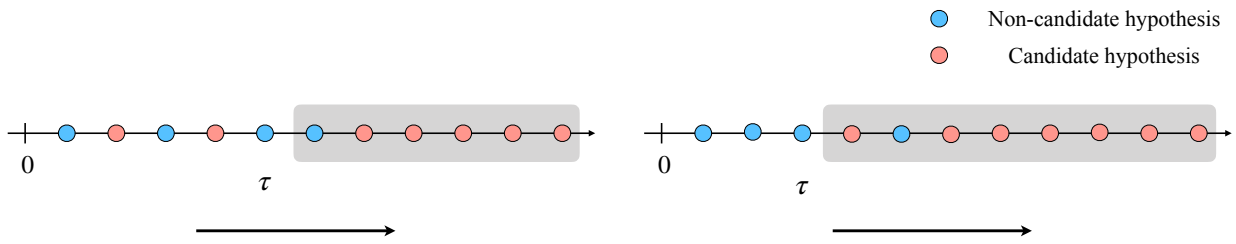
The reader will correctly note that the role of side information in our framework is similar to that of a prior in the Bayesian framework. However, our perspective on side information is here frequentist and, therefore, intrinsically different. Bayesian inference is obtained by averaging over the prior distribution and the validity of inference relies on the correctness of the prior (and the Bayesian model). In our work, the inference results (e.g., FDR control, statistical power) hold conditional on the side information and most importantly, the correctness of the side information does not affect the validity of inference. Having said this, we shall see that our adaptive knockoff filter accomodates '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. 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 on the figure). As in the Benjamini-Hochberg step-up procedure (Storey et al., 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 (we will see later how knockoffs classify hypotheses as candidates or not). 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

[1]It is worth mentioning that Lei and Fithian (2018) discuss in passing (Section 6.2) removing the independence assumption by constructing knockoff copies of the p-values. This is different from our methodology, since in order to construct knockoff copies, one would need to know the joint distribution of the p-values, which is quite restrictive in practice.



(a) "Worse" ordering.

(b) "Better" ordering.

Figure 1: Illustration of the step-up procedure with different orderings. The nodes correspond to hypotheses, and they are ordered from left to right. The red nodes in Figure 1 represent the hypotheses that are candidate for selection and the blue nodes those that are not. (The notion of "candidate hypothesis" is introduced later in Section 3.) The procedure operates sequentially, stops when the ratio of the number of blue nodes to the number of red nodes falls below a threshold q, and selects the remaing red nodes. With $q = 0.2$, the ordering on the left yields five discoveries while that on the right yields seven (the stopping point is marked by τ).

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 et al. (2018), with the last two references using knockoff-based methods. We extend knockoffs 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).[2] *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 knockoff 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 pretty much the same way as we have seen before. Table 1 compares summary results on the WTCCC data, and we can see that adaptive knockoffs discovers more SNPs than other methods. Details including a full list of discovered SNPs are available in Appendix C.

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

[2]The summary statistics are obtained from `https://www.ibdgenetics.org/downloads.html`.

Study/Method	Number of SNPs discovered [3]
WTCCC. (2007)	9
Candès et al. (2018)	18
Sesia et al. (2018)	22.8
Adaptive knockoffs	**33.3**

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.

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

$$H_j : Y \perp\!\!\!\perp X_j | X_{-j}, \tag{1}$$

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. That said, 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, knockoffs only selects among the candidate hypotheses).[4] 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 π_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 an 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

$$\widehat{\mathrm{FDR}}_+(k) := \frac{1 + \sum_{j>k} \mathbf{1}_{\{W_{\pi_j} < 0\}}}{(\sum_{j>k} \mathbf{1}_{\{W_{\pi_j} > 0\}}) \vee 1} \tag{2}$$

[4]The features with $W_j = 0$ will never be selected or used by the procedure so we exclude them in the definitions.

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

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

with the convention $\inf \varnothing = \infty$. The final selected set is the family of remaining candidate hypotheses, i.e. $\hat{\mathcal{S}} = \{\pi_j : j > T_+, W_{\pi_j} > 0\}$. Candès et al. (2018); Barber and Candès (2015) established that this procedure achieves FDR control at the nominal level q. Alternatively, the quantity

$$\widehat{\mathrm{FDR}}_0(k) := \frac{\sum_{j>k} \mathbf{1}_{\{W_{\pi_j}<0\}}}{(\sum_{j>k} \mathbf{1}_{\{W_{\pi_j}>0\}}) \vee 1}, \tag{3}$$

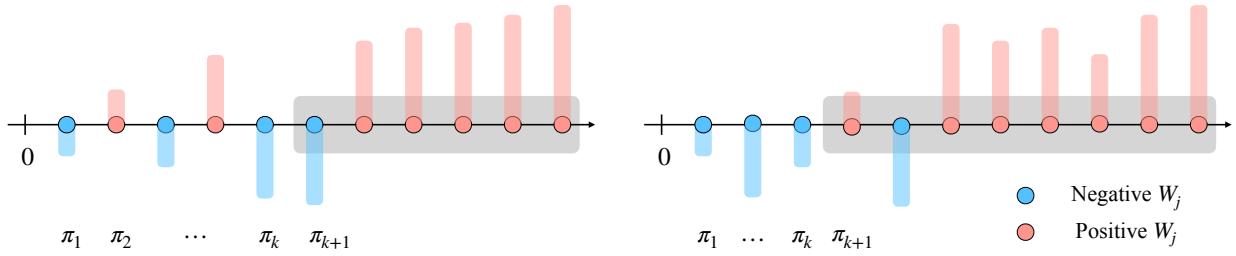
is a slightly less conservative estimate of FDR. Replacing $\widehat{\mathrm{FDR}}_+$ with $\widehat{\mathrm{FDR}}_0$ and replacing T_+ with

$$T_0 := \inf\{k \in [p] : \widehat{\mathrm{FDR}}_0(k) \le q\},$$

yields control of a modified version of FDR defined as

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

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



(a) Model-X knockoffs.

(b) Adaptive knockoffs with side information.

Figure 2: Illustration of knockoffs and adaptive knockoffs. The length of a bar represents the magnitude of a feature importance statistics W_j whereas the color represents the sign. Red (resp. blue) nodes and bars correspond to positive (resp. negative) W_j's. The target FDR level is $q = 0.2$ and we use $\widehat{\mathrm{FDR}}_0$. (a) The ordering $\{\pi_k\}_{k\in[p]}$ is based on the magnitude of the feature importance statistics (standard procedure). The algorithm selects the last five hypotheses. (b) The ordering $\{\pi_k\}_{k\in[p]}$ is determined by both the side information and the magnitude of the feature importance statistics. The algorithm selects seven hypotheses.

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, $(X_j, \tilde{X}_j)|X_{-j}, \tilde{X}_{-j} \overset{\mathrm{d}}{=} (\tilde{X}_j, X_j)|X_{-j}, \tilde{X}_{-j}$.

How to construct good knockoffs is an expanding area of research, see e.g. Candès et al. (2018); Sesia et al. (2018); Gimenez et al. (2018); Liu and Zheng (2018); Romano et al. (2019); Bates et al. (2019). In this paper, we will mainly be using the Gaussian (Candès et al., 2018) and HMM knockoffs (Sesia et al., 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 $(\mathbf{X}, \mathbf{Y})$—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., $W_j = Z_j - \tilde{Z}_j$).[5] 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 σ-algebra generated by the following elements:

- The magnitude of all the W_j's: $\{|W_j|\}_{j\in[p]}$.
- The signs of the examined W_j's: $\{\text{sign}(W_{\pi_j})\}_{j\leq k}$ (when $k = 0$, this is the empty set).
- The signs of the non-null W_j's :$\{\text{sign}(W_j)\}_{j\in\mathcal{H}_1}$.
- The number of positive and negative null W_j's in the unexamined hypotheses: $\{|V^+(j)|\}_{j\leq k}$ and $\{|V^-(j)|\}_{j\leq k}$.
- Side information: U.

[5] An anti-symmetric function is a function such that $f(u, v) = -f(v, u)$.

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 $k = 0, 1, 2, \ldots$, adaptive knockoffs uses a *filter* Φ_{k+1}, required to be $\mathcal{F}_k$-measurable, to determine the least promising hypothesis among the remaining ones.[6] Under this condition, Proposition 1 shows that the resulting ordering $\{\pi_k = \Phi_k\}_{k\in[p]}$ obeys Property 1. Adaptive knockoffs 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* $k \geq 0$*,* Φ_{k+1} *is* $\mathcal{F}_k$*-measurable, then the ordering* $\pi_k = \Phi_k$ *obeys Property 1.*

Proof. By assumption, π_{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 $(U, \{|W_j|\}_{j\in[p]}, \{\text{sign}(W_j)\}_{j\in\mathcal{H}_1}, \{\text{sign}(W_{\pi_j})\}_{j\leq k})$, $\mathcal{F}_k$ can only provide further information on the number of "+"s and "−"s in $V^+(k) \cup V^-(k)$; i.e. $|V^{\pm}(k)|$. Since the signs of the nulls $W_j \neq 0$ are i.i.d. coin flips conditional on $(U, \{|W_j|\}_{j\in[p]}, \{\text{sign}(W_j)\}_j, \{\text{sign}(W_{\pi_j})\}_{j\leq k})$, 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. □

We would like to remark that if each knockoff copy has the properties that $(X, \tilde{X})|U$ stays invariant after swapping X_j and $\tilde{X}_j$ and $\tilde{X}$ is independent of Y conditional on (X, U), then the W_j's satisfy the conditions required in Proposition 1.

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 Φ_{k+1} to determine the next hypothesis to examine π_{k+1}:

$$\pi_{k+1} \leftarrow \Phi_{k+1}(\{|W_j|\}_{j\in[p]}, \{W_{\pi_j}\}_{j>k}, U).$$

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

end

Output: Selected set $\hat{\mathcal{S}} = \{j \in [p] : j > k, W_{\pi_j} > 0\}$.

As a result of Proposition 1, we show in Theorem 1 that adaptive knockoffs 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*.*

[6] When the filter Φ_k has extra randomness, we combine the extra randomness with the original side information and consider the augmented side information and the corresponding augmented σ-field $\tilde{\mathcal{F}}_k$. By such treatment, Φ_k is measurable w.r.t. $\tilde{\mathcal{F}}_{k-1}$.

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 only uses the magnitude of feature importance statistics to determine the order, can be viewed as a special case of adaptive knockoffs: in this case,

$$\Phi_{k+1} = \underset{j>k}{\operatorname{argmin}} \, |W_{\pi_j}|,$$

which is clearly $\mathcal{F}_k$-measurable.

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

$$\begin{aligned}
\mathrm{FDR} = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{H}_0|}{|\hat{\mathcal{S}}|}\right] = \mathbb{E}\left[\frac{|V^+(T_+)|}{|\hat{S}|}\right] &\leq \mathbb{E}\left[\frac{|V^+(T_+)|}{|V^-(T_+)|+1}\widehat{\mathrm{FDR}}_+(T_+)\right] \\
&\leq q\mathbb{E}\left[\frac{|V^-(T_+)|}{|V^+(T_+)|+1}\right] \\
&\leq q.
\end{aligned}$$

The second inequality holds by definition of T_+ and the third inequality follows from the fact that $\dfrac{|V^-(k)|}{|V^+(k)|+1}$ 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 $s_j = \mathrm{sign}(W_j)$ 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[7] and $h(\cdot,\cdot)$ is a regression function. If we postulate a logistic model,

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

[7]In the case where W_j can also be 0, we can alternatively use a multinomial model with levels $\{-1, 0, 1\}$.

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 $(\{s_{\pi_j}\}_{j\le k}, \{U_j\}_{j\in[p]}, \{|W_j|\}_{j\in[p]})$ 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

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

since g is monotone. By construction, Φ_{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. That said, 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 H_j denote whether or not feature j is a null: $H_j = 1$ means feature j is a non-null and $H_j = 0$ means it is a null. We follow the Bayesian two-group model and write

$$H_j|U_j \overset{i.i.d.}{\sim} \text{Bern}(\nu(U_j)),$$

where ν is a link function. Marginally,

$$W_j|H_j, U_j \sim \begin{cases} \mathcal{P}_1(W_j|U_j) & \text{if } H_j = 1, \\ \mathcal{P}_0(W_j|U_j) & \text{if } H_j = 0. \end{cases}$$

Above, $\mathcal{P}_{H_j}(\cdot|U_j)$ denotes the law of W_j conditional on U_j when $H_j \in \{0, 1\}$. Under this model, we can quantify the possibility of a feature being null by inspecting the posterior probability $\mathbb{P}(H_j = 0||W_j|, U_j)$. At each step k, the posterior probability can be used as a criterion to determine the next hypothesis in the ordering, i.e.,[8]

$$\Phi_{k+1} = \underset{j>k}{\operatorname{argmax}}\quad \mathbb{P}(H_{\pi_j} = 0||W_{\pi_j}|, U_{\pi_j}).$$

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

$$p_h(w|u) = \delta_h \mathbf{1}_{\{w=0\}} + (1 - \delta_h)\mathbf{1}_{\{w\neq 0\}} \frac{\beta_h(u) \exp(\beta_h(u)w)}{(1 + \exp(w))^{\beta_h(u)+1}}, \qquad h = 0, 1.$$

[8] In implementation, we instead use $1 - \mathbb{P}(H_j = 1, \text{sign}(W_j) > 0||W_j|, U_j)$.

The continuous part of the distribution is somewhat arbitrary, and we choose this form for computational convenience. Under this model,

$$\begin{aligned} &\mathbb{E}[H|U] = \nu(U), \\ &\mathbb{E}[Y|W \neq 0, U, H = h] = 1/\beta_h(U), \qquad h = 0, 1, \end{aligned} \tag{4}$$

where $Y_j = \log(1+\exp(W_j)) - W_j$. 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 model, we have

$$\mathcal{N} = \{\nu(x) : \nu(x) = 1/(1+\exp(-\theta^T x)),\ \theta \in \mathbb{R}^d\} \tag{5}$$

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

$$\mathcal{B} = \{\beta(x) : \beta(x) = \exp(\theta^T x),\ \theta \in \mathbb{R}^d\}. \tag{6}$$

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

$$\ell(\{H_j, W_j\}_{j\in[p]}|\{U_j\}_{j\in[p]}; \delta_0, \delta_1, \nu(\cdot), \beta_0(\cdot), \beta_1(\cdot)) = \sum_{j=1}^{p} [\underbrace{(i) + (ii)}_{\text{group 1}} + \underbrace{(iii) + (iv) + (v)}_{\text{group 2}}] + C,$$

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

$$\begin{cases} (i) = (1-H_j)\mathbf{1}_{\{W_j=0\}} \log(\delta_0) + (1-H_j)\mathbf{1}_{\{W_j\neq 0\}} \log(1-\delta_0), \\ (ii) = H_j \mathbf{1}_{\{W_j=0\}} \log(\delta_1) + H_j \mathbf{1}_{\{W_j\neq 0\}} \log(1-\delta_1). \end{cases}$$

Group 2 comprises

$$\begin{cases} (iii) = H_j \log(\nu(U_j)) + (1-H_j)\log(1-\nu(U_j)), \\ (iv) = (1-H_j)\mathbf{1}_{\{W_j\neq 0\}}(\log(\beta_0(U_j)) + \beta_0(U_j)\log(\exp(W_j)/(1+\exp(W_j)))), \\ (v) = H_j \mathbf{1}_{\{W_j\neq 0\}}(\log(\beta_1(U_j)) + \beta_1(U_j)\log(\exp(W_j)/(1+\exp(W_j)))). \end{cases}$$

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:

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

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 H_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 S 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 s of the EM algorithm, denote by $\mathcal{G}$ the σ-field generated by the available information. For the E-step we need to compute the following conditional expectations:

$$\mathbb{E}[H_j|\mathcal{G}],\ \mathbb{E}[Y_j H_j|\mathcal{G}],\ \mathbb{E}[Y_j(1-H_j)|\mathcal{G}].$$

We defer the calculation of the above quantities to Appendix A and set $\overline{H}_j = \mathbb{E}[H_j|\mathcal{G}]$. For the M-step, we decompose the optimization into two subgroups. The optimization problems in group 1 have analytical solutions, namely,

$$\hat{\delta}_0 = \underset{\delta_0}{\operatorname{argmax}} \sum_{j=1}^{p}(i) = \frac{\sum_{j\in\mathcal{H}}(1-\overline{H}_j)\mathbf{1}_{\{W_j=0\}}}{\sum_{j\in\mathcal{H}}(1-\overline{H}_j)},$$
$$\hat{\delta}_1 = \underset{\delta_1}{\operatorname{argmax}} \sum_{j=1}^{p}(ii) = \frac{\sum_{j\in\mathcal{H}}\overline{H}_j\mathbf{1}_{\{W_j=0\}}}{\sum_{j\in\mathcal{H}}\overline{H}_j}.$$

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.

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

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 ℓ_1 penalty about the coefficients θ 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 et al., 2009, Chapter 9):

$$\max_{\beta_0(\cdot)\in\mathcal{B}} \sum_{j}^{p}(iv) - \sum_{\ell=1}^{r}\lambda_\ell \int \beta''_{0,\ell}(x_\ell)^2 \mathrm{d}x_\ell,$$

in which $\mathcal{B} = \{\beta(x_1,\ldots,x_r) : \beta(x_1,\ldots,x_r) = \sum_{\ell=1}^{r}\beta_\ell(x_\ell),\ \beta''_\ell(\cdot) \text{ exists for all } \ell \in [r]\}$. 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 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 combines the GLM-extension and nonparamatric regression approaches. In details, we fit $\beta_0(\cdot), \beta_1(\cdot)$ via the `gam` package in R whereas for ν, we regress $\log(\overline{H}_j/(1-\overline{H}_j))$ 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 S 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_{reveal}. 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

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

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

$$\begin{aligned}
&\hat{\beta}_0(U_j) = \hat{\beta}_1(U_j) = 1/\log(2), \quad \text{if } W_j = 0.\\
&\hat{\beta}_0(U_j) = 1/[\log(1+\exp(\overline{W}^-_{\text{reveal}})) - \overline{W}^-_{\text{reveal}}],\\
&\hat{\beta}_1(U_j) = 1/[\log(1+\exp(\overline{W}^+_{\text{reveal}})) - \overline{W}^+_{\text{reveal}}], \quad \text{if } W_j \neq 0.
\end{aligned}$$

Above, $\overline{W}^-_{\text{reveal}}$ is the average of the negative items in W_{reveal} and $\overline{W}^+_{\text{reveal}}$ is the average of the positive items in W_{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_{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 Numerical results

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 λ is set to be 0.5 as in Lei and Fithian (2016).[9] 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 $\widehat{\text{FDR}}_+$ as the estimated FDR.

[9]The code for implementing BH, Storey-BH, SABHA and Adaptive SeqStep is adapted from `https://www.stat.uchicago.edu/~rina/sabha/All_q_est_functions.R` and `https://github.com/lihualei71/adaptPaper/blob/master/R/other_methods.R`.

Algorithm 2: EM algorithm to estimate p_0, p_1, ν

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 $s \leftarrow 0, \ldots, S-1$ **do**

1. **E-step:**

Update $\overline{H}_j$: $\overline{H}_j \leftarrow \mathbb{E}[H_j|\mathcal{G}]$, $\quad j \in [p]$.

Update $(\overline{Y}_{0,j}, \overline{Y}_{1,j})$: $\overline{Y}_{h,j} \leftarrow \dfrac{\mathbb{E}[Y_j(hH_j + (1-h)(1-H_j))|\mathcal{G}]}{h\overline{H}_j + (1-h)(1-\overline{H}_j)}$, $\quad h = 0, 1$,

where the calculations of the conditional expectations are presented in Appendix 2.

2. **M-step:**

Update $(\hat{\delta}_0, \hat{\delta}_1)$: $\hat{\delta}_h \leftarrow \dfrac{\sum_{j \in \mathcal{H}} (h\overline{H}_j + (1-h)(1-\overline{H}_j))\mathbf{1}_{\{W_j = 0\}}}{\sum_{j \in \mathcal{H}} (h\overline{H}_j + (1-h)(1-\overline{H}_j))}$, $\quad h = 0, 1$.

Update $\hat{\nu}$: $\hat{\nu} \leftarrow$ random forest$(\overline{H}_j \sim U_j)$.

Update $(\hat{\beta}_0, \hat{\beta}_1)$: $1/\hat{\beta}_h \leftarrow$ random forest$(\overline{Y}_{h,j} | W_j \neq 0 \sim U_j)$, $\quad h = 0, 1$.

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: $\hat{\delta}_0, \hat{\delta}_1, \hat{\nu}, \hat{\beta}_0, \hat{\beta}_1$.

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`.

Method	Abbreviation	P-value free?	Use side information?
Benjamini Hochberg	BHq		
Storey's BH	StoreyBH		
Adaptive SeqStep	AdaSeqStep		✓
AdaPT	AdaPT		✓
Structure Adaptive BH algorithm	SABHA		✓
Vanilla Model-X knockoffs	Vanilla Knockoff	✓	
Adaptive knockoffs w/ GLM filter	AdaKn(GLM)	✓	✓
Adaptive knockoffs w/ GAM filter	AdaKn(GAM)	✓	✓
Adaptive knockoffs w/ Random Forest filter	AdaKn(RF)	✓	✓
Adaptive knockoffs w/ two group model	AdaKn(EM)	✓	✓

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

$$Y|X_1, \ldots, X_p \sim \mathcal{N}(\beta_1 X_1 + \ldots \beta_p X_p, 1).$$

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 β—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.[10] The setting is motivated by the fact that in many real 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 3a 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 $j \in [p]$.

(a) Hypothesis structure.

(b) Zoomed-in view of the first 300 indices.

Figure 3: One-dimensional hypothesis structure.

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 outperforms vanilla knockoffs 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) in Figure 5a and Figure 5b respectively. We can observe that adaptive knockoffs 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 (2019) 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:

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

[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 $\frac{1}{j^2}$ until we obtain 150 distinct realizations.

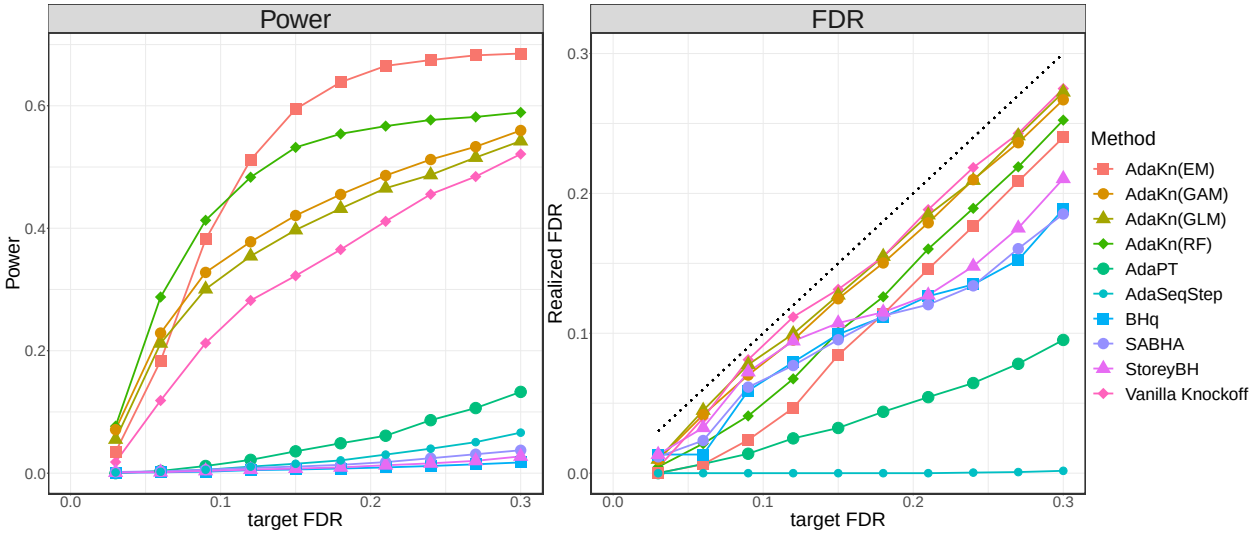
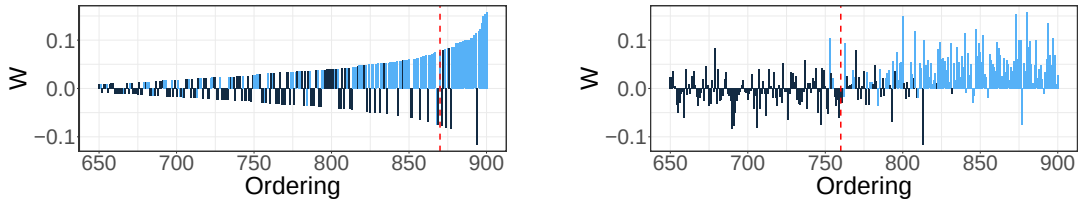


Figure 4: Power (left) and FDR (right) versus target FDR values.



(a) Model-X knockoffs.

(b) Adaptive knockoffs.

Figure 5: (a) Realized ordering of vanilla knockoffs. (b) 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.

The entries of β 'live' on a two-dimensional plane and the location of β_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 β. Details about the signal locations are in Appendix D. The magnitude of the nonzero entries is set to $\frac{25}{\sqrt{n}}$ 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:

$$\mathrm{Cov}(X_i, X_j) = e^{-3||U_i - U_j||_2^2}, \qquad i, j \in [p],$$

where $U_j = (r(j), s(j))$. The side information is the pair of coordinates of each feature.

This simulation setting is motivated by magnetic resonance imaging (MRI) studies. For example, the hypotheses (nodes) are 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 et al. (2014); Rosenblatt et al. (2018)). Here we adopt the voxel-based inference as in Efron (2012).

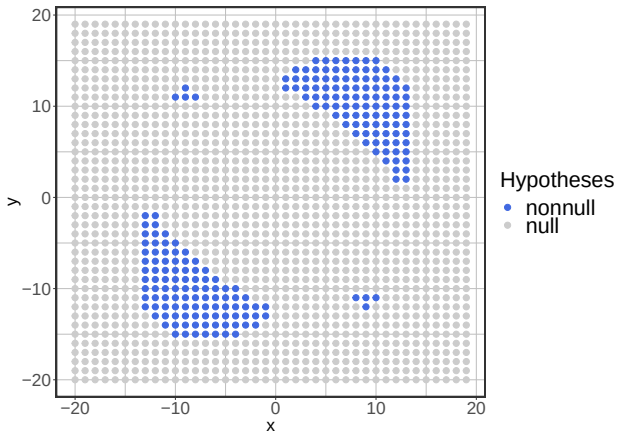


Figure 6: Two-dimensional hypothesis structure. Blue nodes correspond to non-nulls and gray nodes to nulls.

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. Adaptive knockoffs with a Bayesian filter or random forest filter outperform vanilla knockoffs by a wide margin. Figure 8a and 8b show the realized ordering of vanilla knockoffs and adaptive knockoffs with the Bayesian filter 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.

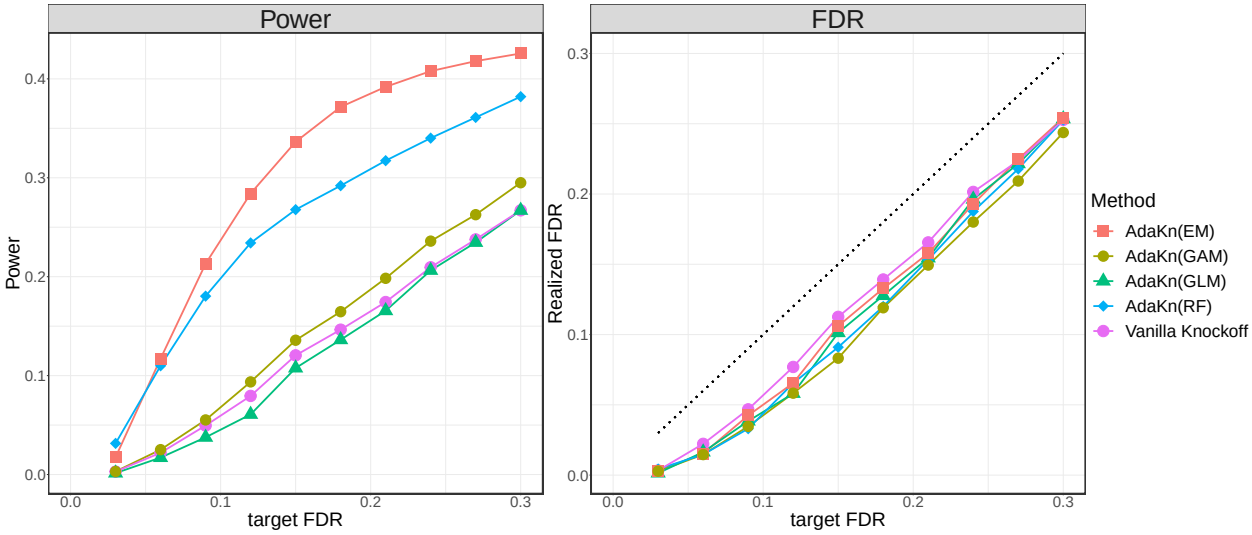
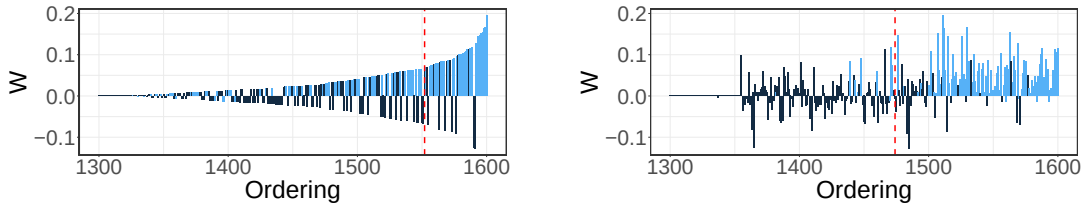


Figure 7: Power (left) and FDR (right) versus target FDR values.



(a) Model-X knockoff.

(b) Adaptive knockoffs.

Figure 8: (a) Realized ordering of vanilla knockoffs in Simulation 2; (b) Realized ordering of adaptive knockoffs with the Bayesian filter. The setup is otherwise the same as in Figure 5.

7 Applications

7.1 GWAS

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).

7.1.1 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 values are missing in some of them. 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.

7.1.2 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 et al. (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 et al. (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.

Dataset	Phenotype	n	p (pre-clustering)	p (post-clustering)
WTCCC	CD	4913	377,749	$71,145$
NFBC	LDL	4682	328,934	$59,005$
NFBC	HDL	4700	328,934	$59,005$

Table 3: Description of the datasets.

7.1.3 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, we apply the procedure above after imputing the missing p-value with a one.

Lipids For HDL and LDL, we obtain summary statistics reported by Loh et al. (2018).[11] 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. The motivation here is that if a SNP is not even recorded, the chance of being significant will likely be low.

7.1.4 Implementation details

Knockoff construction We use the HMM knockoffs from Sesia et al. (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 λ 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 $\widehat{\mathrm{FDR}}_0$.

7.1.5 Results

We apply adaptive knockoffs with target FDR level $q = 0.1$. Since the knockoff-based algorithms are essentially random and depend on the realizations of $\tilde{X}$, we generate 50 knockoffs independently conditioning on (X, Y). We conduct analysis on every realization of $\tilde{X}$ and report the average number of discoveries. In Appendix B, we provide boxplots of discovery numbers and in Appendix C the full list of discovered SNPs. The average number of discovered SNPs for Crohn's disease has

[11] The summary statistics are downloaded from `https://data.broadinstitute.org/alkesgroup/UKBB/`.

been presented in Table 1, and the results for the NFBC dataset are shown in Table 4. We compare our results with Sabatti et al. (2009) and Sesia et al. (2018); the former adopts a marginal test with a p-value threshold of 5×10^{-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 regardless of the correctness of the p-values we use. Also, we are not merely re-discovering what is already known since side information concerns other populations.

Method	Number of discoveries (HDL)	Number of discoveries (LDL)
Sabatti et al. (2009)	5	6
HMM knockoffs (Sesia et al., 2018)	8	9.8
Adaptive knockoffs	**12.5**	**18.3**

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 $\tilde{X}$ (for adaptive knockoffs, this number is 50).

8 Future work

This paper generalizes the knockoff procedure to a setting where side information associated with features is available. We close by discussing a few interesting directions for future work.

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.

Acknowledgement

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. Z. R. thanks Stephen Bates, Nikolaos Ignatiadis, Eugene Katsevich and Matteo Sesia for their valuable comments on this project.

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523):1172–1183.

Bates, S., Candès, E. J., Janson, L., and Wang, W. (2019). Metropolized knockoff sampling. *arXiv preprint arXiv:1903.00434*.

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418.

Benjamini, Y., Yekutieli, D., et al. (2001). The control of the false discovery rate in multiple testing under dependency. *The annals of statistics*, 29(4):1165–1188.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cai, T. T., Sun, W., and Wang, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):187–234.

Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Coram, M. A., Candille, S. I., Duan, Q., Chan, K. H. K., Li, Y., Kooperberg, C., Reiner, A. P., and Tang, H. (2015). Leveraging multi-ethnic evidence for mapping complex traits in minority populations: an empirical bayes approach. *The American Journal of Human Genetics*, 96(5):740–752.

Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L., and Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *The American Journal of Human Genetics*, 101(2):218–226.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558.

Edwards, D. (2012). *Introduction to graphical modelling*. Springer Science & Business Media.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., Kong, A., et al. (2008). Unsupervised empirical bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 2(2):714–735.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Gimenez, J. R., Ghorbani, A., and Zou, J. (2018). Knockoffs for the mass: new feature importance statistics with false discovery guarantees. *arXiv preprint arXiv:1807.06214*.

Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E. S., Annese, V., Hauser, S. L., et al. (2015). High-density mapping of the mhc identifies a shared role for hla-drb1* 01: 03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nature genetics*, 47(2):172.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Ignatiadis, N. and Huber, W. (2017). Covariate powered cross-weighted multiple testing. *arXiv preprint arXiv:1701.05179*.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577.

Janson, L. (2017). *A model-free approach to high-dimensional inference*. PhD thesis, PhD thesis, Stanford University.

Lei, L. and Bickel, P. J. (2019). An assumption-free exact test for fixed-design linear models with exchangeable errors. *arXiv preprint arXiv:1907.06133*.

Lei, L. and Fithian, W. (2016). Power of ordered hypothesis testing. In *International Conference on Machine Learning*, pages 2924–2932.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Li, A. and Barber, R. F. (2019). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):45–74.

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979.

Liu, Y. and Zheng, C. (2018). Auto-encoding knockoff generator for fdr controlled variable selection. *arXiv preprint arXiv:1809.10765*.

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. *Nature genetics*, page 1.

Lynch, G., Guo, W., Sarkar, S. K., Finner, H., et al. (2017). The control of the false discovery rate in fixed sequence multiple testing. *Electronic Journal of Statistics*, 11(2):4649–4673.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398.

Romano, Y., Sesia, M., and Candès, E. (2019). Deep knockoffs. *Journal of the American Statistical Association*, pages 1–12.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796.

Sabatti, C., Service, S. K., Hartikainen, A.-L., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35.

Sesia, M., Sabatti, C., and Candès, E. (2018). Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2017). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, pages 1–72.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419.

WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661.

A Conditional expectation (Algorithm 2)

Suppose we are at step k of Algorithm 1 and step s of Algorithm 2. Recall that $\mathcal{G}$ is the σ-field generated by the current information. Since the distribution of W_j is a mixture of a point mass at 0 and an absolutely continuous distribution, we need to treat the conditional probability and conditional expectation differently depending on whether $W_j = 0$ or not. To avoid complication, we slightly abuse notation and let $\mathbb{P}(W_j|\cdot)$ refer to a probability when $W_j = 0$ and to a density otherwise.

Revealed hypotheses For $j \in \{\pi_1, \ldots, \pi_k\}$, the value of W_j is known conditional on $\mathcal{G}$.

- The conditional expectation of H_j is

$$\mathbb{E}[H_j|\mathcal{G}] = \frac{\mathbb{P}(H_j = 1, W_j|\mathcal{G})}{\mathbb{P}(W_j|\mathcal{G})},$$

  where

$$\mathbb{P}(H_j = 1, W_j|\mathcal{G}) = \nu(U_j)p_1(W_j; U_j), \tag{7}$$
$$\mathbb{P}(W_j|\mathcal{G}) = \nu(U_j)p_1(W_j; U_j) + (1 - \nu(U_j))p_0(W_j; U_j). \tag{8}$$

- The conditional expectation of Y_jH_j is

$$\mathbb{E}[Y_jH_j|\mathcal{G}] = Y_j\mathbb{E}[H_j|\mathcal{G}],$$

  where $\mathbb{E}[H_j|\mathcal{G}]$ has been computed above.

Unrevealed hypotheses For $j \in [p]\backslash\{\pi_1, \ldots, \pi_k\}$, we know the magnitude of W_j conditional on $\mathcal{G}$ but not its sign.

- The conditional expectation of H_j is

$$\mathbb{E}[H_j|\mathcal{G}] = \frac{\mathbb{P}(H_j = 1, |W_j||\mathcal{G})}{\mathbb{P}(|W_j||\mathcal{G})},$$

  where

$$\begin{aligned}\mathbb{P}(H_j = 1, |W_j||\mathcal{G}) =& \nu(U_j)p_1(|W_j|; U_j) + \nu(U_j)p_1(-|W_j|; U_j)\\ \mathbb{P}(|W_j||\mathcal{G}) =& \nu(U_j)p_1(|W_j|; U_j) + \nu(U_j)p_1(-|W_j|; U_j) +\\ & (1 - \nu(U_j))p_0(|W_j|; U_j) + (1 - \nu(U_j))p_0(-|W_j|; U_j).\end{aligned}$$

- The conditional expectation of Y_jH_j is

$$\mathbb{E}[Y_jH_j|\mathcal{G}] = y_{j,1}\mathbb{P}(H_j = 1, W_j > 0|\mathcal{G}) + y_{j,2}\mathbb{P}(H_j = 1, W_j < 0|\mathcal{G}),$$

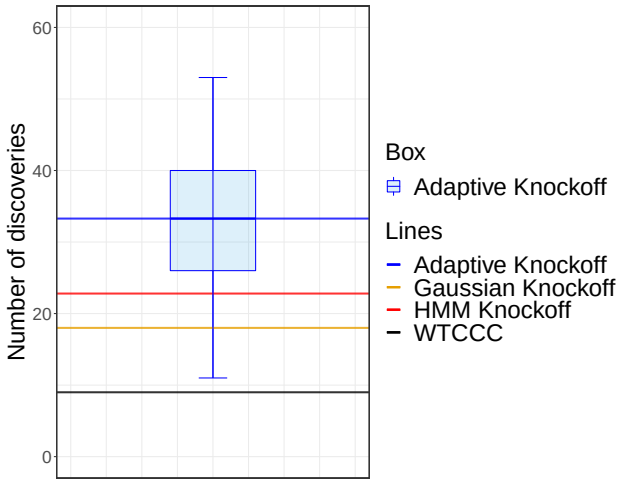
  where

$$\mathbb{P}(H_j = 1, W_j > 0|\mathcal{G}) = \frac{\mathbb{P}(H_j = 1, W_j = |W_j||\mathcal{G})}{\mathbb{P}(|W_j||\mathcal{G})}, \tag{9}$$
$$\mathbb{P}(H_j = 1, W_j < 0|\mathcal{G}) = \frac{\mathbb{P}(H_j = 1, W_j = -|W_j||\mathcal{G})}{\mathbb{P}(|W_j||\mathcal{G})}, \tag{10}$$
$$y_{j,1} = \log(\exp(|W_j|) + 1) - |W_j|,$$
$$y_{j,2} = \log(\exp(-|W_j|) + 1) + |W_j|.$$

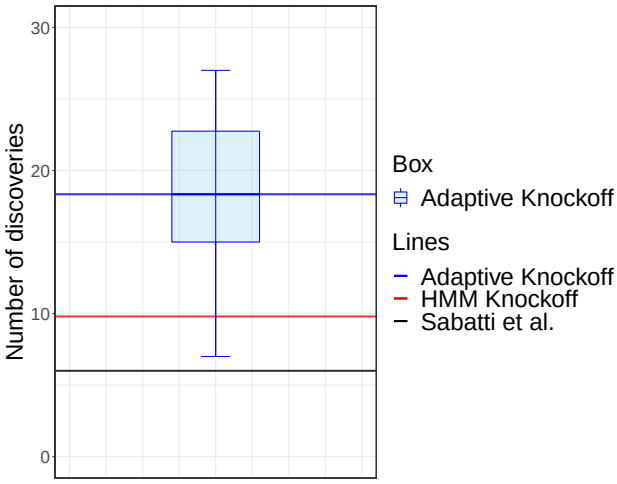
  The numerators and denominators in (9) and (10) have been calculated in (7) and (8).

B Boxplots of the number of discoveries in Section 7

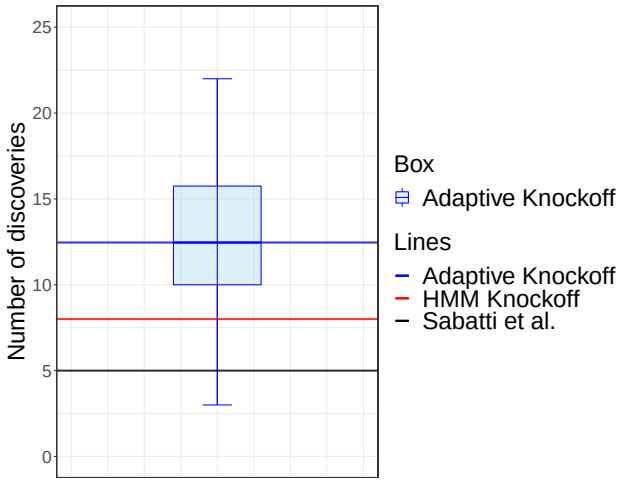
We present boxplots of the number of discoveries from multiple knockoffs realizations.



(a) Crohn's disease.



(b) LDL.



(c) HDL.

Figure 9: Number of discoveries from multiple knockoffs realizations. The solid lines are the (average) numbers of discoveries and the boxplots represent the discoveries made by adaptive knockoffs in 50 repetitions.

C Full list of discovered SNPs in Section 7

We present the full list of SNP clusters discovered by our adaptive knockoff procedure. Since we run the algorithm 50 times, we count the frequency of SNP representatives being selected and report those with a selection frequency greater than or equal to 30%. For each cluster representative, we also report the size of the corresponding cluster, the chromosome it belongs to, and the position range of the cluster. The position of SNPs are reported as in the original dataset: the WTCCC dataset follows the convention of Human Genome Build 35 and the NFBC dataset follows the convention of Human Genome Build 37. Lastly, we compare our results to previous works. For Crohn's disease, we indicate if our discovered SNPs are discovered by WTCCC. (2007); Candès et al. (2018); Sesia et al. (2018). For NFBC we compare with Sabatti et al. (2009); Sesia et al. (2018). An asterisk indicates that the reported SNP is not exactly in the cluster but is within the position range (0.5Mb).

C.1 LDL

Cluster representative (cluster size)	Selection frequency (%)	Chr.	Position range (Mb)	Selection frequency (%) in Sesia et al. (2018)	Found in Sabatti et al. (2009)?
rs10198175 (1)	100	2	21.13-21.13	80	rs693∗
rs10953541 (58)	100	7	106.48-107.30	76	No.
rs157580 (4)	100	19	45.40-45.41	94	rs157580
rs2228671 (2)	100	19	11.20-11.21	97	rs11668477
rs557435 (21)	100	1	55.52-55.72	92	No.
rs1713222 (45)	98	2	21.11-21.53	41	rs693
rs646776 (5)	98	1	109.80-109.82	97	rs646776
rs174450 (16)	94	11	61.55-61.68	36	rs1535
rs2802955 (1)	92	1	235.02-235.02	40	No.
rs4803750 (1)	90	19	45.25-45.25		No.
rs6756629 (2)	86	2	44.07-44.08		No.
rs688 (4)	86	19	11.16-11.24		rs11668477*
rs4906908 (8)	82	15	26.97-27.05		No.
rs12427378 (43)	78	12	50.43-51.31	19	No.
rs4844614 (34)	74	1	207.30-207.88	99	rs4844614
rs10409243 (5)	70	19	10.33-10.37		No.
rs12670798 (11)	68	7	21.57-21.71		rs693*
rs10056811 (94)	60	5	74.24-75.24		No.
rs11878377 (39)	50	19	10.63-11.18		rs646776*
rs11615 (17)	44	19	45.91-46.10		No.
rs2919843 (3)	42	19	45.19-45.20		No.
rs9696070 (6)	40	9	89.21-89.24	25	No.
rs1105879 (33)	32	2	234.50-234.70		No.

Table 5: SNPs clusters discovered to be associated with LDL.

C.2 HDL

Cluster representative (cluster size)	Selection frequency (%)	Chr.	Position range (Mb)	Selection frequency (%) in Sesia et al. (2018)	Found in Sabatti et al. (2009)?
rs1532085 (4)	100	15	58.68-58.70	100	rs1532085
rs1532624 (2)	100	16	56.99-57.01	99	rs3764261
rs1800961 (1)	98	20	43.04-43.04	100	No.
rs255049 (142)	98	16	66.41-69.41	95	rs255049
rs7499892 (1)	98	16	57.01-57.01	100	rs3764261
rs10096633 (19)	80	8	19.73-19.94	57	No.
rs9898058 (1)	78	17	47.82-47.82	55	No.
rs17075255 (59)	68	5	164.28-164.92	51	No.
rs3761373 (1)	62	21	42.87-42.87	43	No.
rs12139970 (11)	52	1	230.35-230.42	23	No.
rs2575875 (10)	52	9	107.63-107.68	28	No.
rs2849049 (6)	44	9	15.29-15.31		No.
rs173738 (3)	42	5	16.71-16.73	12	No.
rs2019260 (24)	38	5	16.41-16.59		No.
rs2132167 (94)	34	8	33.29-34.78		No.
rs2426404 (1)	32	20	50.64-50.64		No.
rs9324799 (8)	30	5	154.15-154.41		No.

Table 6: SNPs discovered to be associated with HDL.

C.3 Crohn's disease

Cluster representative (cluster size)	Sel. fre (%)	Chr.	Position range (Mb)	Selection frequency (%) in Sesia et al.	Selection frequency (%) in Candès et al.?	Found in WTCCC et al.?
rs11209026 (2)	100	1	67.31-67.42	100	100	rs11805303*
rs11627513 (7)	100	14	96.61-96.63	68	80	No.
rs11805303 (16)	100	1	67.31-67.46	95	80	rs11805303
rs17234657 (1)	100	5	40.44-40.44	97	90	rs17234657
rs4246045 (46)	100	5	150.07-150.41	66	50	rs1000113
rs6431654 (20)	100	2	233.94-234.11	99	100	rs10210302
rs6500315 (4)	100	16	49.03-49.07	73	60	rs17221417
rs6688532 (33)	100	1	169.40-169.65	98	90	rs12037606
rs7095491 (18)	100	10	101.26-101.32	91	100	rs10883365
rs2738758 (5)	98	20	61.71-61.82	72	60	No.
rs4692386 (1)	98	4	25.81-25.81	56	40	No.
rs3135503 (16)	96	16	49.28-49.36	91	90	rs17221417
rs9469615 (2)	94	6	33.91-33.92	48	30	No.
rs4807569 (2)	92	19	1.07-1.08	27		No.
rs6743984 (23)	92	2	230.91-231.05	39	10	No.
rs4263839 (23)	90	9	114.58-114.78	56	30	No.
rs17063661 (1)	84	6	134.70-134.70			No.
rs7497036 (19)	82	15	72.49-72.73	22		No.
rs1451890 (26)	78	15	30.92-31.01	15		No.
rs2390248 (13)	78	7	19.80-19.89	54	50	No.
rs10801047 (10)	76	1	188.17-188.47			No.
rs1345022 (44)	76	9	21.67-21.92		40	No.
rs549104 (10)	76	18	2.05-2.11			No.
rs12529198 (31)	74	6	5.01-5.10	23		No.
rs17694108 (1)	74	19	38.42-38.42		10	No.
rs10761659 (53)	72	10	64.06-64.41	45	10	rs10761659
rs6601764 (1)	72	10	3.85-3.85	80	100	rs6601764
rs7655059 (5)	70	4	89.50-89.53	75	40	No.
rs4870943 (10)	66	8	126.59-126.62	14		No.
rs7768538 (1145)	66	6	25.19-32.91	81	60	rs9469220
rs10172295 (127)	48	2	57.86-59.01			No.
rs946227 (4)	48	6	138.12-138.13			No.
rs2836753 (5)	44	21	39.21-39.23	42	30	No.
rs4959830 (11)	42	6	3.36-3.41	20	10	No.
rs9783122 (234)	38	10	106.43-107.61	62	80	No.
rs11579874 (17)	34	1	197.61-197.76			No.
rs4437159 (4)	34	3	84.80-84.81	49	60	No.
rs7726744 (46)	34	5	40.35-40.71	70	50	rs17234657
rs10916631 (14)	30	1	220.87-221.08		40	No.
rs2814036 (5)	30	1	163.94-164.07	14		No.

Table 7: SNPs discovered to be associated with Crohn's disease.

D Implementation details

In Section 6.3, feature j is a signal if any of the following conditions holds:

$$\begin{cases} ((\frac{r(j)}{9})^2 + (\frac{s(j)}{9})^2 - 2)^4 - (\frac{r(j)}{9})^3(\frac{s(j)}{9})^5 < 0, \\ ((\frac{r(j)}{9} + 1)^2 + (\frac{s(j)}{9} - \frac{5}{4})^2 - 0.015 < 0, \\ ((\frac{r(j)}{9} - 1)^2 + (\frac{s(j)}{9} + \frac{5}{4})^2 - 0.015 < 0. \end{cases}$$