Simultaneous false discovery proportion bounds via knockoffs and closed testing

Li, J.Z.; Maathuis, M.H.; Goeman, J.J.

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 966–986
https://doi.org/10.1093/jrsssb/qkae012
Advance access publication 4 March 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Simultaneous false discovery proportion bounds via knockoffs and closed testing

Jinzhou Li[1], Marloes H. Maathuis[2] and Jelle J. Goeman[3]

[1]Department of Statistics, Stanford University, Stanford, CA, USA
[2]Seminar für Statistik, ETH Zürich, Zürich, Switzerland
[3]Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Address for correspondence: Jinzhou Li, Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, CA 94305, USA. Email: jinzhoul@stanford.edu

Abstract

We propose new methods to obtain simultaneous false discovery proportion bounds for knockoff-based approaches. We first investigate an approach based on Janson and Su's k-familywise error rate control method and interpolation. We then generalize it by considering a collection of k values, and show that the bound of Katsevich and Ramdas is a special case of this method and can be uniformly improved. Next, we further generalize the method by using closed testing with a multi-weighted-sum local test statistic. This allows us to obtain a further uniform improvement and other generalizations over previous methods. We also develop an efficient shortcut for its implementation. We compare the performance of our proposed methods in simulations and apply them to a data set from the UK Biobank.

Keywords: closed testing, interpolation, joint k-familywise error rate, knockoff framework, shortcut, simultaneous false discovery proportion bounds

1 Introduction

1.1 Simultaneous inference framework

Many modern data analysis tasks are of an exploratory nature. In such cases, the simultaneous inference framework (e.g. Genovese & Wasserman, 2006; Goeman & Solari, 2011) can be more suitable than the commonly used false discovery rate (FDR) control framework (e.g. Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). Specifically, for a pre-set nominal FDR level, the output of an FDR control method is a rejection set R. In order for the FDR control guarantee [see (1) below] to hold, users are not allowed to set the nominal FDR level in a post-hoc manner or to alter R in any way. For example, they cannot change to a smaller (or larger) FDR level even if the currently used one seems to return too many (or too few) rejections. Moreover, they cannot remove any rejection from R, not even if it violates their scientific expertize. The simultaneous inference framework does not suffer from these issues. It allows users to freely check any set of their interest, and it returns corresponding high probability false discovery proportion (FDP) bounds [see (2) below] for the given sets. As such, the simultaneous inference framework brings more flexibility and is very well suited for exploratory research.

Formally, let $\mathcal{N} \subseteq [p] = \{1, \ldots, p\}$ be the index set of true null hypotheses among p null hypotheses. The FDP of a rejection set $R \subseteq [p]$ is defined as

$$\mathrm{FDP}(R) = \frac{|R \cap \mathcal{N}|}{\max\{1, |R|\}},$$

where $|\cdot|$ is the cardinality of a set. For a given data set and a fixed $\alpha \in (0, 1)$, the FDR control framework aims to produce a rejection set R such that

Received: December 24, 2022. **Revised:** December 7, 2023. **Accepted:** February 12, 2024

$$\text{FDR}(R) = \mathbb{E}[\text{FDP}(R)] \leq \alpha. \tag{1}$$

The simultaneous inference framework, on the other hand, aims to obtain a function $\overline{\text{FDP}} : 2^{[p]} \to [0, 1]$, where $2^{[p]}$ denotes the power set of $[p]$, such that

$$\mathbb{P}(\text{FDP}(R) \leq \overline{\text{FDP}}(R), \forall\ R \subseteq [p]) \geq 1 - \alpha. \tag{2}$$

When (2) is satisfied, we say that $\bar{\text{FDP}}(\cdot)$ is a simultaneous FDP upper bound. Every FDP upper bound can be translated to a corresponding true discovery lower bound, and the other way around (see Appendix C.1 of the online supplementary material for more details). In this paper, we state the results in terms of FDP upper bounds.

Existing methods for obtaining simultaneous error bounds are mostly based on p-values. For example, Genovese and Wasserman (2004), Meinshausen (2006), and Hemerik et al. (2019) propose simultaneous FDP bounds for sets of the form $\widehat{R}_t = \{i \in [p] : p_i \leq t\}$, where p_i is the p-value for the ith hypothesis, while Genovese and Wasserman (2006), Goeman and Solari (2011), Goeman et al. (2019), Blanchard et al. (2020), and Vesely et al. (2023) propose simultaneous FDP bounds for arbitrary sets $R \subseteq [p]$.

1.2 Knockoff framework

For error-controlled variable selection problems, Barber and Candès (2015) and Candès et al. (2018) introduced a so-called knockoff framework that does not resort to p-values. The fundamental idea of the knockoff method is to create synthetic variables (i.e. knockoffs) that serve as negative controls. By using the knockoffs together with the original variables as input of an algorithm (e.g. lasso, random forests, or neural networks), one obtains knockoff statistics $W_1, \ldots, W_p$, which possess a so-called coin-flip property (see Definition 2.1). Based on this property, variable selection methods were developed to control the FDR (Barber & Candès, 2015) or the k-familywise error rate (k-FWER, Janson & Su, 2016).

The knockoff method has been extended to many settings, including group variable selection (Dai & Barber, 2016), multilayer variable selection (Katsevich & Sabatti, 2019), high-dimensional linear models (Barber & Candès, 2019), Gaussian graphical models (Li & Maathuis, 2021), multi-environment settings (Li et al., 2021), and time series settings (Chi et al., 2021). The knockoff approach can be especially useful for high-dimensional settings, where obtaining valid p-values is very challenging. For example, it has been successfully applied to real applications with high-dimensional data (e.g. He et al., 2021; Sesia et al., 2021, 2020).

Most existing knockoff-based methods aim to control the FDR. Motivated by the advantages of simultaneous FDP bounds and the usefulness of knockoff methods, we consider the problem of obtaining such bounds for knockoff-based approaches. We focus on the variable selection setting, but the proposed methods can be applied to other settings as long as the coin-flip property (see Definition 2.1) holds. To the best of our knowledge, only Katsevich and Ramdas (2020) investigated this problem so far. They actually proposed a general approach to obtain simultaneous FDP bounds for a broad range of FDR control procedures. As a special case, they obtain simultaneous FDP bounds for knockoff-based methods.

1.3 Outline of the paper and our contributions

We start by investigating a simple approach (see Section 3.1) that obtains simultaneous FDP bounds based on a k-FWER controlled set (Janson & Su, 2016) and interpolation (Blanchard et al., 2020; Goeman et al., 2021). We found that there are two issues with this approach: (i) it is unclear how to choose the tuning parameter k, which can heavily affect the results, and (ii) the FDP bounds for certain sets can be very conservative due to the nature of interpolation.

These issues motivated us to consider several values of k at the same time. Specifically, we first get a collection of sets $\widehat{S}_1, \ldots, \widehat{S}_m$, such that the k_i-FWER of set $\widehat{S}_i$ is controlled jointly for all $i \in [m]$. Next, we obtain simultaneous FDP bounds by using interpolation based on these sets.

This idea is a special case of a more general procedure called joint familywise error rate control (Blanchard et al., 2020). Related ideas can be traced back to Genovese and Wasserman (2006) and Van der Laan et al. (2004).

Our proposed method (see Section 3.2) is uniformly better than the above simple approach using only one k-FWER controlled set, and it largely alleviates the above two issues. Issue (i) regarding the choice of tuning parameters is not fully resolved, in the sense that one still needs to choose a sequence of k values, for which the optimal choice depends on the data distribution. We, therefore, suggest some reasonable choices and also propose a two-step approach to obtain the tuning parameters. In addition, we prove that the simultaneous FDP bound from Katsevich and Ramdas (2020) is a special case of our method, and uniformly better bounds can be obtained.

Next, in Section 4, we turn to the closed testing framework. In particular, we propose a multi-weighted-sum local test statistic based on knockoff statistics. We prove that all previously mentioned approaches are special cases (or shortcuts) of this method, and we derive uniform improvements and generalizations over them. The closed testing procedure is computationally intractable in its standard form, so we also develop an efficient and general shortcut for its implementation.

In Section 5, we evaluate the numerical performance of the proposed methods in simulation studies and apply them to a data set from the UK Biobank. We close the paper with a discussion and potential future research directions in Section 6.

2 Preliminaries

Throughout the paper, vectors are denoted by boldface fonts, and we consider $\alpha \in (0, 1)$, $m \geq 1$, $\boldsymbol{v} = (v_1, \ldots, v_m)$ with $0 < v_1 < \ldots < v_m$, and $\boldsymbol{k} = (k_1, \ldots, k_m)$ with $0 < k_1 \leq \ldots \leq k_m$.

2.1 Knockoff framework and the coin-flip property

We first review the knockoff framework for variable selection. That is, we consider a response Y and covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$. We say that X_i is a null variable if $Y \perp\!\!\!\perp X_i \mid \boldsymbol{X}_{-i}$, where $\boldsymbol{X}_{-i} = \{X_1, \ldots, X_p\} \setminus \{X_i\}$, and the goal is to find non-null variables with error control.

Knockoff methods yield knockoff statistics $\boldsymbol{W} = (W_1, \ldots, W_p)$ based on samples from X and Y, where non-null variables tend to have large and positive W_i values. Hence, one can select variables by choosing $\{i \in [p] : W_i > T\}$ for some threshold T. In addition, knockoff methods are designed so that the following coin-flip property holds (Barber & Candès, 2015; Candès et al., 2018).

Definition 2.1 For knockoff statistic vector $\boldsymbol{W}$, let $D_i = \text{sign}(W_i)$. Let $\mathcal{N} \subseteq [p]$ be the index set of null variables. We say that $\boldsymbol{W}$ possesses the **coin-flip property** if conditional on $(|W_1|, |W_2|, \ldots, |W_p|)$, the D_i's are i.i.d. with distribution $\mathbb{P}(D_i = 1) = \mathbb{P}(D_i = -1) = 1/2$ for $i \in \mathcal{N}$.

The coin-flip property is key to obtain valid error control in the knockoff framework. For p-value-based approaches, the analogous property is that a null p-value is stochastically larger than a uniform random variable on [0, 1].

In this paper, we do not specify the data-generating model or the knockoff method. Instead, we work directly on the knockoff statistic vector $\boldsymbol{W}$, which is assumed to possess the coin-flip property. Our methods can be applied as long as this assumption holds. For example, the same methodology can be directly extended to the group variable selection setting, where the coin-flip property holds at the group level (Dai & Barber, 2016).

Throughout, we assume without loss of generality that $|W_1| > |W_2| > \ldots > |W_p| > 0$. Otherwise, one can pre-process $\boldsymbol{W}$ by relabelling the indices, discarding the zero ones, and breaking ties. These pre-processing steps have no influence on the remaining part of the paper because the pre-processed $\boldsymbol{W}$ still possesses the coin-flip property.

From a more abstract point of view, the considered setting can be reformulated as a 'pre-ordered' case as discussed in Katsevich and Ramdas (2020). In this scenario, we are given a sequence of ordered hypotheses $H_1, \ldots, H_p$, each associated with a test statistic D_i. These test statistics possess the property that the null D_i's are independent coin flips. The goal is to obtain simultaneous FDP bound (2).

2.2 k-FWER control via knockoffs

We now review the k-FWER control method using knockoffs by Janson and Su (2016). This method is a crucial component of our starting point for the interpolation-based method proposed in Section 3.1.

Specifically, for $v \geq 1$, Janson and Su (2016) defined the discovery set

$$\widehat{S}(v) = \{i \in [p] : W_i \geq \widehat{T}(v)\}, \tag{3}$$

where the threshold

$$\widehat{T}(v) = \max\{|W_i| : |\{j \in [p] : |W_j| \geq |W_i| \text{ and } W_j < 0\}| = v\} \tag{4}$$

with $\widehat{T}(v) = \min_{i \in [p]}\{|W_i|\}$ if $|\{j \in [p] : W_j < 0\}| < v$ (note that $|\cdot|$ is used to denote absolute value for numbers and cardinality for sets). For given $k \geq 1$ and $\alpha \in (0, 1)$, they proposed

$$v^{\mathrm{JS}}(k) = argmax\left\{v \in [p] : \sum_{i=k}^{\infty} 2^{-i-v}\binom{i+v-1}{i} \leq \alpha\right\}, \tag{5}$$

and their discovery set is then $\widehat{S}^{\mathrm{JS}}(k) = \widehat{S}(v^{\mathrm{JS}}(k))$.

Equation (5) is based on the fact that $|\widehat{S}(v) \cap \mathcal{N}|$ is stochastically dominated by a negative binomial random variable $N(v)$ with distribution NB $(v, 1/2)$ (Janson & Su, 2016). This implies that the k-FWER of $\widehat{S}^{\mathrm{JS}}(k)$ is controlled, since

$$\mathbb{P}\left(|\widehat{S}^{\mathrm{JS}}(k) \cap \mathcal{N}| \leq k-1\right) \geq \mathbb{P}\left(N(v^{\mathrm{JS}}(k)) \leq k-1\right) \geq 1-\alpha. \tag{6}$$

3 Simultaneous FDP bounds via knockoffs and joint k-FWER control with interpolation

3.1 K-FWER control with interpolation

We now combine the k-FWER controlled set from Janson and Su (2016) with interpolation. Here interpolation means that, for a k-FWER controlled set $\widehat{S}(v)$ and arbitrary sets R, we can bound $|R \cap \mathcal{N}|$ based on $\widehat{S}(v)$. In particular, when $|\widehat{S}(v) \cap \mathcal{N}| \leq k-1$, we have for any $R \subseteq [p]$,

$$\begin{aligned} |R \cap \mathcal{N}| &= |R \cap \widehat{S}(v) \cap \mathcal{N}| + |(R \backslash \widehat{S}(v)) \cap \mathcal{N}| \\ &\leq \min\{|R|, |\widehat{S}(v) \cap \mathcal{N}| + |R \backslash \widehat{S}(v)|\} \\ &\leq \min\{|R|, k-1+|R \backslash \widehat{S}(v)|\}. \end{aligned} \tag{7}$$

Hence

$$P(|R \cap \mathcal{N}| \leq \min\{|R|, k-1+|R \backslash \widehat{S}(v)|\}, \forall R \subseteq [p]) \geq P(|\widehat{S}(v) \cap \mathcal{N}| \leq k-1) \geq 1-\alpha.$$

By dividing by $\max\{1, |R|\}$ on both sides within the above probability, we obtain the following simultaneous FDP bound:

$$\overline{\mathrm{FDP}}_{k,v}(R) = \frac{\min\{|R|, k-1+|R \backslash \widehat{S}(v)|\}}{\max\{1, |R|\}}, \quad R \subseteq [p]. \tag{8}$$

Interpolation is actually a more general technique (Blanchard et al., 2020; Goeman et al., 2021), which reduces to the simplified form (7) in our special case (see also Appendix C.2 of the online supplementary material).

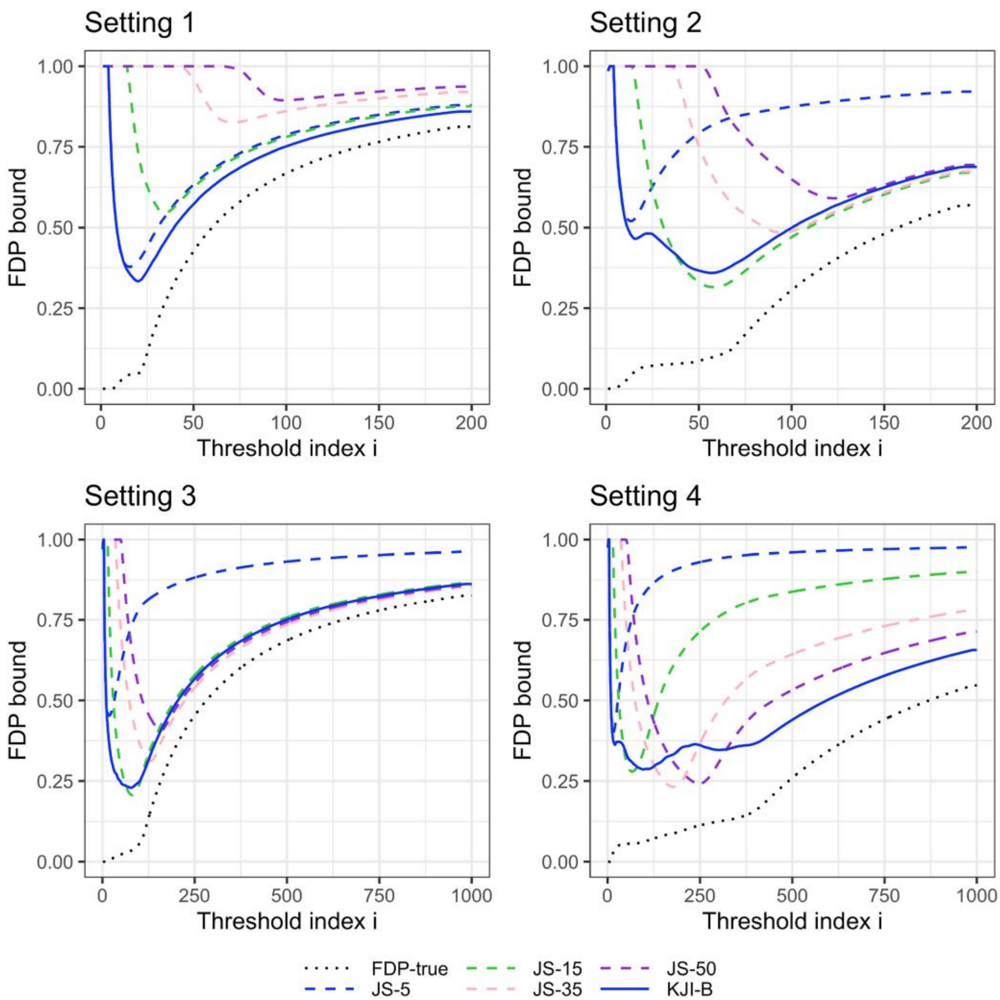

Figure 1. Simultaneous FDP bound $\overline{\mathrm{FDP}}_k^{\mathrm{JS}}(\cdot)$ with $k \in \{5, 15, 35, 50\}$ (denoted by JS-5, JS-15, JS-35, and JS-50, respectively) and our later proposed Algorithm 1 with the Type-B $\boldsymbol{v}$ described in (14) and $\boldsymbol{k}$ obtained by using Algorithm 2 (denoted by KJI-B). The black dotted line indicates the true FDP. All results are average values over 200 replications.

If $|\widehat{S}(v)| < k - 1$, one can always obtain a larger discovery set $\widehat{S}'(v)$ with $|\widehat{S}'(v)| = k - 1$ by adding the variables corresponding to the next largest positive W_i's, while maintaining k-FWER control. One might expect that the simultaneous FDP bound (8) based on $\widehat{S}'(v)$ might be better (i.e. smaller). However, because $k - 1 + |R\backslash S| \geq k - 1 + |R| - |S| \geq |R|$ for any set S with $|S| \leq k - 1$, the FDP bound (8) based on $\widehat{S}'(v)$ and $\widehat{S}(v)$ are both equal to 1, meaning that only trivial FDP bound can be obtained by interpolation when the base set is not large enough. We will, therefore, use $\widehat{S}(v)$ for simplicity.

By using the k-FWER controlled set $\widehat{S}^{\mathrm{JS}}(k)$ from Janson and Su (2016) in (8), we obtain the following simultaneous FDP bound:

$$\overline{\mathrm{FDP}}_k^{\mathrm{JS}}(R) = \frac{\min\{|R|, k - 1 + |R\backslash \widehat{S}^{\mathrm{JS}}(k)|\}}{\max\{1, |R|\}}, \quad R \subseteq [p]. \tag{9}$$

As mentioned in the introduction, the simultaneous FDP bound $\overline{\mathrm{FDP}}_k^{\mathrm{JS}}(\cdot)$ has two issues: (i) it is not clear a priori how to choose the tuning parameter k, and this choice can heavily affect the bound, and (ii) the bound can be very conservative for sets R that are very different from $\widehat{S}^{\mathrm{JS}}(k)$. To illustrate these two issues, we present Figure 1 showing the simultaneous FDP bound $\overline{\mathrm{FDP}}_k^{\mathrm{JS}}(\cdot)$ for sets

$\widehat{R}_i = \{j \le i : W_j > 0\}$, $i \in [p]$, in four simulation settings (see Appendix E.1 of the online supplementary material for more details). To see issue (ii), note that the FDP bounds strongly depend on the threshold index i, and their minima are reached around i such that $\widehat{R}_i$ equals $\widehat{S}^{\mathrm{JS}}(k)$.

3.2 Joint k-FWER control with interpolation

The above two issues motivate the following new method. The idea is that since FDP bounds based on different k can be better in different settings, we use many k-FWER controlled sets to construct them. As we mentioned in Section 1, this idea is a special case of a more general method proposed by Blanchard et al. (2020).

Formally, for $\widehat{S}(v_i)$ defined by formula (3), we say that joint $\boldsymbol{k}$-FWER control holds if

$$\mathbb{P}\Big(|\widehat{S}(v_i) \cap \mathcal{N}| \le k_i - 1, \forall\, i \in [m]\Big) \ge 1 - \alpha. \tag{10}$$

We then obtain a simultaneous FDP bound

$$\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(R) = \min_{i \in [m]} \overline{\mathrm{FDP}}_{k_i, v_i}(R), \qquad R \subseteq [p], \tag{11}$$

where $\overline{\mathrm{FDP}}_{k_i,v_i}(R)$ is defined by (8). $\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ is a valid simultaneous FDP bound because

$$\mathbb{P}\Big(\mathrm{FDP}(R) \le \overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(R), \forall\, R \subseteq [p]\Big) \ge \mathbb{P}\Big(|\widehat{S}(v_i) \cap \mathcal{N}| \le k_i - 1, \forall\, i \in [m]\Big) \ge 1 - \alpha.$$

The previous FDP bound $\overline{\mathrm{FDP}}^{\mathrm{JS}}_k(\cdot)$ is a special case of $\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ with $m = 1$, $k_1 = k$, and $v_1 = v^{\mathrm{JS}}(k)$. That is, $\overline{\mathrm{FDP}}^{\mathrm{JS}}_k(\cdot) = \overline{\mathrm{FDP}}_{k, v^{\mathrm{JS}}(k)}(\cdot)$. By the definition (11), we can see that $\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ with $m > 1$, $k_1 = k$, and $v_1 = v^{\mathrm{JS}}(k)$ is uniformly better (i.e. smaller or equal) than $\overline{\mathrm{FDP}}^{\mathrm{JS}}_k(\cdot)$.

To obtain vectors $\boldsymbol{v}$ and $\boldsymbol{k}$ satisfying the joint $\boldsymbol{k}$-FWER control guarantee (10), we generalize ideas from Janson and Su (2016). In particular, we use the following lemma:

Lemma 3.1 Let $\mathcal{N} \subseteq [p]$ be the set of null variables. Then, for positive integers $v_1 < \cdots < v_m$ and $k_1 < \cdots < k_m$, we have

$$\mathbb{P}(|\widehat{S}(v_i) \cap \mathcal{N}| \le k_i - 1, \forall\, i \in [m]) \ge \mathbb{P}(N^p(v_i) \le k_i - 1, \forall\, i \in [m]), \tag{12}$$

where $N^p(v_1), \ldots, N^p(v_m)$ are early stopped negative binomial random variables defined on one sequence of Bernoulli trials of length p.

Here, an early stopped negative binomial random variable $N^p(v)$ is formally defined as $N^p(v) = |\{j \le \min(U(v), p) : B_j = 1\}|$, where $U(v) = \min\{i \ge 1 : |\{j \le i : B_j = -1\}| = v\}$ and $B_1, B_2, \ldots$ are i.i.d. Bernoulli random variables taking values $\{+1, -1\}$.

Based on Lemma 3.1, if vectors $\boldsymbol{v}$ and $\boldsymbol{k}$ satisfy

$$\mathbb{P}(N^p(v_i) \le k_i - 1, \forall\, i \in [m]) \ge 1 - \alpha, \tag{13}$$

then the joint $\boldsymbol{k}$-FWER control guarantee (10) holds, and $\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ is a simultaneous FDP bound.

In the k-FWER control case of Janson and Su (2016), we have $p = \infty$, $m = 1$, and $\boldsymbol{k} = k$ is user-chosen in inequality (13). The optimal v can then be naturally obtained via optimization (5), because it is clear that the largest v leads to the most discoveries. In our case of getting simultaneous FDP bounds with $m > 1$, things are more complicated. We will discuss the choice of $\boldsymbol{v}$ and $\boldsymbol{k}$ in more detail in the next section.

For clarity, we present the simultaneous FDP bound $\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ in Algorithm 1. As a quick illustration, we present the performance of this algorithm in Figure 1, where the tuning parameter $\boldsymbol{v}$ is taken as the Type-B $\boldsymbol{v}$ in (14), and $\boldsymbol{k}$ is obtained by our later proposed Algorithm 2. One can see that our proposed method generally performs well in all four settings.

Algorithm 1 Simultaneous FDP bounds via knockoffs and joint k-FWER control with interpolation

Input: $(R, \boldsymbol{W})$, where $R \subseteq [p]$ is the set of interest, $\boldsymbol{W} \in \mathbb{R}^p$ is the vector of knockoff statistics.

Parameter: $(\alpha, \boldsymbol{v}, \boldsymbol{k})$, where α is the nominal level for the simultaneous FDP control, $\boldsymbol{v}$ and $\boldsymbol{k}$ are vectors satisfying inequality (13).

Output: FDP upper bound for R, which holds simultaneously for all $R \subseteq [p]$.

1: For each $i \in [m]$, obtain

$$\overline{\mathrm{FDP}}_{k_i,v_i}(R) = \frac{\min\{|R|,\, k_i - 1 + |R \backslash \widehat{S}(v_i)|\}}{\max\{1, |R|\}},$$

where $\widehat{S}(v_i) = \{j \in [p] : W_j \geq \widehat{T}(v_i)\}$ with $\widehat{T}(v_i) = \max\{|W_j| : |\{l \in [p] : |W_l| \geq |W_j| \text{ and } W_l < 0\}| = v_i\}$ and $\widehat{T}(v_i) = \min_{j\in[p]}\{|W_j|\}$ if $|\{j \in [p] : W_j < 0\}| < v_i$.

2: Return $\min_{i\in[m]} \overline{\mathrm{FDP}}_{k_i,v_i}(R)$.

3.3 Tuning parameters $\boldsymbol{v}$ and $\boldsymbol{k}$ and uniform improvement of Katsevich and Ramdas (2020)

The vectors $\boldsymbol{v}$ and $\boldsymbol{k}$ satisfying inequality (13) are not unique, and the best $\boldsymbol{v}$ and $\boldsymbol{k}$ for the FDP bound $\overline{\mathrm{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ depend on the unknown distribution of $\boldsymbol{W}$. In this paper, we first choose $\boldsymbol{v}$, then obtain $\boldsymbol{k}$ such that (13) holds. In particular, we consider the following four types of $\boldsymbol{v}$ vectors (ordered in a way that v_i grows ever faster):

$$\begin{array}{ll} \text{Type-A}: \; v_i = i, & \text{Type-B}: \; v_i = \lfloor i^2/2 \rfloor, \\ \text{Type-C}: \; v_1 = 1, v_2 = 2, v_i = v_{i-1} + v_{i-2}, & \text{Type-D}: \; v_i = 2^{i-1}. \end{array} \tag{14}$$

We always use $v_1 = 1$ because it gives the best FDP bounds when the obtained knockoff statistic vector is optimal, namely, when all true W_i's are positive and have larger absolute values than any null W_i.

For a given $\boldsymbol{v}$, one may obtain $\boldsymbol{k}$ numerically using a greedy approach with the constraint that inequality (13) holds (e.g. based on Monte-Carlo simulations). We will follow this idea and proceed by connecting to, then improve upon, the only existing simultaneous FDP bound of Katsevich and Ramdas (2020). The interpolation version of the FDP bound presented in Katsevich and Ramdas (2020) is:

$$\overline{\mathrm{FDP}}^{\mathrm{KR}}(R) = \frac{\min_{i\in[p]}\{|R|,\, |R \backslash \widehat{S}_i| + \lfloor c(\alpha) \cdot (1 + i - |\widehat{S}_i|) \rfloor\}}{\max\{1, |R|\}}, \tag{15}$$

where $\widehat{S}_i = \{j \leq i : W_j > 0\}$ and $c(\alpha) = \frac{\log(\alpha^{-1})}{\log(2-\alpha)}$. Katsevich and Ramdas (2020) mentioned that one can apply interpolation to their original bound to obtain a better bound, but they did not present the formula. We give the explicit formula in (15) to ensure that our new methods are compared with the best version of Katsevich and Ramdas (2020), i.e. the version with interpolation. The derivation of (15) can be found in Appendix C.3 of the online supplementary material.

For a given $\boldsymbol{v}$, we first derive a valid $\boldsymbol{k}$ based on $\overline{\mathrm{FDP}}^{\mathrm{KR}}(\cdot)$ and Theorem C.1 (Theorem 1 in Goeman et al., 2021, see Appendix C.1 of the online supplementary material). Specifically, let

$$k^{raw}_{v_i} = \min_{j\geq 1}\{c_j : j - c_j + 1 = v_i\} \quad \text{with} \quad c_j = \left\lfloor \frac{c(\alpha)\cdot(1+j)}{1+c(\alpha)} \right\rfloor + 1, \tag{16}$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x and $c(\alpha) = \frac{\log(\alpha^{-1})}{\log(2-\alpha)}$. Then, as the following Proposition 3.1 shows, $\boldsymbol{k}^{raw} = (k^{raw}_{v_1}, \ldots, k^{raw}_{v_m})$ is valid.

Proposition 3.1. Let $N^p(v_1), \ldots, N^p(v_m)$ be early stopped negative binomial random variables defined on one sequence of Bernoulli trials of length p. Then

$$\mathbb{P}\Big(N^p(v_i) \leq k^{raw}_{v_i} - 1, \forall\, i \in [m]\Big) \geq 1 - \alpha. \tag{17}$$

By comparing formulas (15) and (11), the relationship between simultaneous FDP bounds $\overline{\text{FDP}}^{\text{KR}}(\cdot)$ and $\overline{\text{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ is not immediately clear. Perhaps surprisingly, as we show in the following Proposition 3.2, $\overline{\text{FDP}}^{\text{KR}}(\cdot)$ is in fact a special case of $\overline{\text{FDP}}^{m}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$.

Proposition 3.2 For $\boldsymbol{v} = (1, \ldots, p)$ and $\boldsymbol{k}^{raw} = (k^{raw}_1, \ldots, k^{raw}_p)$, we have

$$\overline{\text{FDP}}^{\text{KR}}(R) = \overline{\text{FDP}}^{p}_{(\boldsymbol{k}^{raw},\boldsymbol{v})}(R), \quad R \subseteq [p].$$

One value of the above result is that it unifies $\overline{\text{FDP}}^{\text{KR}}(\cdot)$, the only existing method for obtaining simultaneous FDP bounds for knockoff approaches, within our method. More importantly, it enables us to uniformly improve $\overline{\text{FDP}}^{\text{KR}}(\cdot)$. Specifically, for a vector $\boldsymbol{v} = (v_1, \ldots, v_m)$, using $\boldsymbol{k}^{raw} = (k^{raw}_{v_1}, \ldots, k^{raw}_{v_m})$ to obtain $\overline{\text{FDP}}^{p}_{(\boldsymbol{k}^{raw},\boldsymbol{v})}(\cdot)$ is not optimal, that is, $\mathbb{P}(N^p(v_i) \le k^{raw}_{v_i} - 1, \forall\, i \in [m])$ and $1 - \alpha$ are not close, so that the α level is not exhausted.

To improve this, we propose a two-step approach. Let $j^*_i = argmin_{j\ge 1}\{c_j : j - c_j + 1 = v_i\}$. In the first step, we refine $k^{raw}_{v_i}$ by replacing $c(\alpha)$ in $c_{j^*_i}$ with the smallest constant such that inequality (13) holds. In most cases, however, the first step will still not exhaust the α level due to the discrete nature of the target probability. Hence in the second step, we greedily update k_i under the constraint (13). Here, we choose to update k_i in an increasing order of i because the probability discretization is coarser for smaller v_i and finer for larger v_i, so updating in such an order is helpful to exhaust the α level (see also Blanchard et al., 2020 for a closely related principle called λ-calibration). In the practical implementation, we update $c(\alpha)$ in a discrete manner in the first step, and we use Monte-Carlo simulation to approximate the target probability $\mathbb{P}(N^p(v_1) \le k_1 - 1, \ldots, N^p(v_m) \le k_m - 1)$ because an exact calculation is computationally infeasible for large m. We summarize this two-step approach in Algorithm 2.

Algorithm 2 Two-step approach to obtain $k_1, \ldots, k_m$

Input: $(v_1, \ldots, v_m, \alpha, \delta, p)$, where $v_1 < \cdots < v_m$, $\alpha \in (0, 1)$ is the nominal level for the simultaneous FDP control, $\delta > 0$ is the step size used in the first step, and p is the parameter in inequality (13).

Output: $k_1, \ldots, k_m$.

1: For each $i \in [m]$, obtain $j^*_i = argmin_{j\ge 1}\{c_j : j - c_j + 1 = v_i\}$, where c_j is defined by (16). Refine $c(\alpha)$ by using step size δ: find the largest $N \ge 0$ such that inequality (13) holds with $(v_1, \ldots, v_m)$ and $(k^{step1}_1, \ldots, k^{step1}_m)$, where $k^{step1}_i = \left\lfloor \frac{c^{step1} \cdot (1 + j^*_i)}{1 + c^{step1}} \right\rfloor + 1$ and $c^{step1} = c(\alpha) - N\delta$.

2: Update $k^{step1}_1, \ldots, k^{step1}_m$ in a greedy way: find the smallest $k^{step2}_1 \le k^{step1}_1$ such that inequality (13) holds with $(v_1, \ldots, v_m)$ and $(k^{step2}_1, k^{step1}_2, \ldots, k^{step1}_m)$. Then keep k^{step2}_1 and find the smallest $k^{step2}_2 \le k^{step1}_2$ such that inequality (13) holds with $(v_1, \ldots, v_m)$ and $(k^{step2}_1, k^{step2}_2, \ldots, k^{step1}_m)$. Iterate this process until we get all $k^{step2}_1, \ldots, k^{step2}_m$, and return them as output.

To illustrate the improvement of using Algorithm 2 over $\boldsymbol{k}^{raw}$, we present the corresponding vectors $\boldsymbol{k}^{raw}$, $\boldsymbol{k}^{step1}$, and $\boldsymbol{k}^{step2}$ based on $\boldsymbol{v} = (1, \ldots, p)$ in the plot (1) of Figure 2 (see Appendix E.2 of the online supplementary material for the plots of other types of $\boldsymbol{v}$). One can see that k^{step2}_i is smaller than k^{step1}_i and k^{raw}_i for the same v_i, and hence it will lead to better FDP bounds. We also present the corresponding target probabilities for the four types of $\boldsymbol{v}$ in Table 1. One can see that $\boldsymbol{k}^{step2}$ exhausted the α-level (up to numerical precision) while $\boldsymbol{k}^{step1}$ and $\boldsymbol{k}^{raw}$ did not.

Different types of $\boldsymbol{v}$ vectors lead to different reference sets and $\boldsymbol{k}$ vectors, so the resulting FDP bounds differ as well. As illustration, we present the $\boldsymbol{k}$ vectors for the four types of $\boldsymbol{v}$ using Algorithm 2 in plots (2) and (3) of Figure 2. Their corresponding FDP bounds $\overline{\text{FDP}}^{p}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ are examined in Section 5. We will see that there is no uniformly best one.

From a practical point of view, we find that the Type-B $\boldsymbol{v}$ consistently performs well in different scenarios (see Section 5), so we recommend it as a default choice in practice. The data-dependent optimal choice of $\boldsymbol{v}$ remains an open question for future research. For the choice of the largest value v_m, based on both its role in Algorithms 1 and 2 and our experience from simulations, we found that this value does not have a strong influence as long as it is relatively large (e.g. $p/3$), so one may set it according to the affordable computational expense.

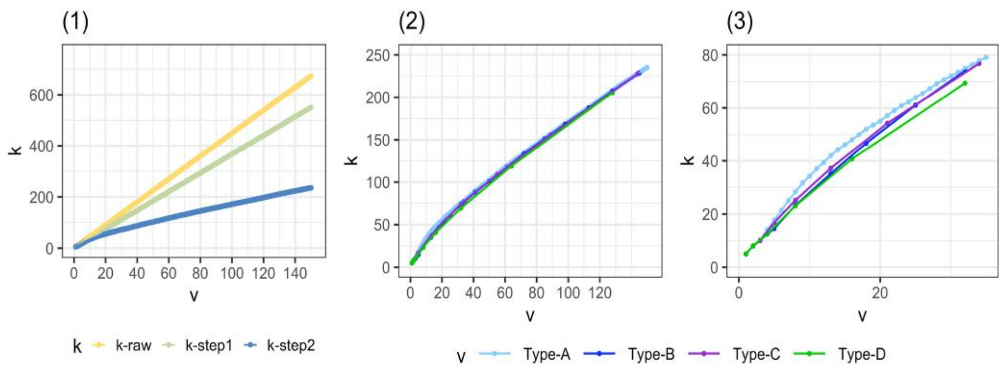

Figure 2. Plot (1): The vectors $\boldsymbol{k}^{raw}$, $\boldsymbol{k}^{step1}$, and $\boldsymbol{k}^{step2}$ for a $\boldsymbol{v}$ vector of Type-A in (14). Plot (2): The output vector k by using Algorithm 2 with $\boldsymbol{v}$ vectors of Types A–D in (14). Plot (3): The zoom-in version of Plot (2). For all three plots, we use $p = 1{,}000$ and let $v_i < 150$. All results are average values over 100 replications.

Table 1. The empirical target probabilities $\mathbb{P}(N^p(v_i) \le k_i - 1, \forall\, i \in [m])$ for $\boldsymbol{k}^{raw}$, $\boldsymbol{k}^{step1}$, and $\boldsymbol{k}^{step2}$ based on four types of $\boldsymbol{v}$, with standard deviations in parentheses

v	Type-A	Type-B	Type-C	Type-D
$\boldsymbol{k}^{raw}$	0.9632 (0.0006)	0.9639 (0.0006)	0.9633 (0.0005)	0.9635 (0.0007)
$\boldsymbol{k}^{step1}$	0.9560 (0.0007)	0.9585 (0.0006)	0.9568 (0.0006)	0.9574 (0.0007)
$\boldsymbol{k}^{step2}$	0.9500 (0)	0.9500 (0)	0.9500 (0)	0.9500 (0)

Note. Here, we use $p = 1{,}000$ and let $v_i < 150$. The results are based on 100 replications.

Lastly, by applying Algorithm 2 to $\boldsymbol{v} = (1, 2, \ldots, p)$, we can obtain an uniformly better FDP bound than $\overline{\mathrm{FDP}}^{\mathrm{KR}}(\cdot)$, meaning that it is at least as good as $\overline{\mathrm{FDP}}^{\mathrm{KR}}(\cdot)$ and can never worse in all cases. As we will see in Section 5, the obtained FDP bound is actually strictly better in many simulations.

Proposition 3.3 For $\boldsymbol{v} = (1, \ldots, p)$, let $\boldsymbol{k}$ be the output of Algorithm 2. Then,

$$\overline{\mathrm{FDP}}^{p}_{(\boldsymbol{k},\boldsymbol{v})}(R) \le \overline{\mathrm{FDP}}^{\mathrm{KR}}(R), \quad R \subseteq [p].$$

The FDP bound of Katsevich and Ramdas (2020) was obtained using a general martingale approach which can be applied to knockoff settings as special cases. Our method makes more targeted use of the Bernoulli trials, which are fundamental for the knockoff-based approach. In that sense, it may not come as a surprise that we are able to obtain better FDP bounds in this specific context. A natural follow-up question is: Can we further improve $\overline{\mathrm{FDP}}^{p}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$? As we will show in Section 4, the answer is yes. The FDP bound $\overline{\mathrm{FDP}}^{p}_{(\boldsymbol{k},\boldsymbol{v})}(\cdot)$ is still not admissible (Goeman et al., 2021), and a further uniform improvement can be obtained by connecting it to the closed testing framework.

4 Simultaneous FDP bounds via knockoffs and closed testing

In this section, we turn to the closed testing framework and propose a method to obtain simultaneous FDP bounds by combining knockoffs and closed testing. In particular, we propose a local test based on the knockoff statistics for closed testing. We show that Algorithm 1 is a special case (or an exact shortcut) of this closed testing-based approach, and that it can be uniformly improved. We also present other generalizations based on closed testing and develop a shortcut for their implementations.

4.1 Preliminaries on closed testing

The closed testing procedure (Marcus et al., 1976) was originally proposed for familywise error rate control. Goeman and Solari (2011) observed that it can also be used to obtain simultaneous FDP bounds.

Consider hypotheses $H_1, \ldots, H_p$, where in the variable selection setting H_i means that the ith variable is null. The closed testing procedure consists of three steps:

1. Test all intersection hypotheses $H_I = \cap_{i \in I} H_i$ for $I \subseteq [p]$ at level α. Throughout the paper, we refer to such tests as local tests, in contrast to the closed tests described in Step 2. We use $\phi_I \in \{0, 1\}$ to denote the local test result of H_I, with 1 indicating rejection. For convenience, we sometimes say that a set I is rejected to mean that H_I is rejected. We use the convention that $\phi_\emptyset = 0$, i.e. $H_\emptyset$ is always accepted.
2. Obtain closed testing results based on the local test results: for all $I \subseteq [p]$, H_I is rejected by closed testing if all supersets J of I are locally rejected. We use $\phi_I^{[p]} \in \{0, 1\}$ to denote the closed testing result of H_I:

$$\phi_I^{[p]} = \min\{\phi_J : I \subseteq J \subseteq [p]\}.$$

3. For any set $R \subseteq [p]$, let

$$t^{[p]}(R) = \max_{I \in 2^R} \{|I| : \phi_I^{[p]} = 0\} \tag{18}$$

be the size of the largest subset of R that is not rejected by closed testing. Then

$$\overline{\mathrm{FDP}}^{ct}(R) = \frac{t^{[p]}(R)}{\max\{1, |R|\}}, \quad R \subseteq [p] \tag{19}$$

is a simultaneous FDP bound.

The first two steps guarantee that the FWER is controlled, in the sense that $P(\phi_I^{[p]} = 0, \forall\, I \subseteq \mathcal{N}) \geq 1 - \alpha$. To see this, note that $P(\phi_{\mathcal{N}} = 0) \geq 1 - \alpha$ for any valid local test, and by construction $\phi_{\mathcal{N}} = 0$ implies that $\phi_I^{[p]} = 0$ for all $I \subseteq \mathcal{N}$.

Step 3 is the new step from Goeman and Solari (2011) which gives us a simultaneous FDP bound. To see this, note that $\phi_{\mathcal{N}} = 0$ implies $\phi_{R \cap \mathcal{N}}^{[p]} = 0$ for all $R \subseteq [p]$, and hence $|R \cap \mathcal{N}| \leq t^{[p]}(R)$ for all $R \subseteq [p]$. Therefore, for any valid local test, we have $\mathbb{P}(|R \cap \mathcal{N}| \leq t^{[p]}(R), \forall\, R \subseteq [p]) \geq \mathbb{P}(\phi_{\mathcal{N}} = 0) \geq 1 - \alpha$. This leads to a simultaneous FDP bound by dividing by $\max\{1, |R|\}$ on both sides inside the probability.

This closed testing procedure is computationally intractable in its standard form as there are 2^p hypotheses to test in Step 1. In some cases, however, there exist shortcuts to derive the closed testing results efficiently (see, e.g. Dobriban, 2020; Goeman et al., 2021; Goeman & Solari, 2011; Tian et al., 2023; Vesely et al., 2023).

4.2 The multi-weighted-sum local test statistic and connections to other methods

For variable selection problems, the null hypothesis with respect to a set $I \subseteq [p]$ is

$$H_I : \text{the } i\text{th variable is null}, \quad \forall\, i \in I.$$

For a knockoff statistic vector $\boldsymbol{W}$ possessing the coin-flip property, $D_i = \mathrm{sign}(W_i)$'s are i.i.d. with distribution $\mathbb{P}(D_i = 1) = \mathbb{P}(D_i = -1) = 1/2$ for $i \in I$ under the null hypothesis H_I. Based on this, we propose the following multi-weighted-sum test statistic to locally test H_I in the closed testing framework:

$$\widehat{L}_i^I = \sum_{j \in I} w_{i,j}^I \mathbb{1}_{D_j = 1}, \quad i \in [m], \tag{20}$$

where $w_{i,j}^I$ are tuning parameters. Let $z_1^I, \ldots, z_m^I$ be the corresponding critical values such that under H_I,

$$\mathbb{P}\left(\widehat{L}_i^I \le z_i^I - 1, \forall\, i \in [m]\right) \ge 1 - \alpha. \tag{21}$$

We locally reject H_I if there exists some $i \in [m]$ such that $\widehat{L}_i^I \ge z_i^I$. Note that because the distribution of D_i's under H_I is known, these critical values can be numerically approximated by first simulating i.i.d. D_i with $\Pr(D_i = 1) = \Pr(D_i = -1) = 1/2$, and then approximating the critical values in a greedy manner. For a concrete implementation example, please refer to https://github.com/Jinzhou-Li/KnockoffSimulFDP. We would like to mention that the test statistic (20) would inherently lack power at any reasonable significance level if the cardinality of I is small. For example, in the case of $|I| = 1$, one can only accept H_I for all significance levels $\alpha \in (0, 0.5)$. This is due to the discrete nature of the test statistic, which requires accumulating enough evidence (thus necessitating a large $|I|$) in order to reject the null hypothesis. This limitation is unavoidable but not particularly detrimental to our problem and goals.

Given a concrete local test, the remaining key issue for obtaining FDP bounds using closed testing is the computation. This is because the standard form requires conducting $2^p - 1$ tests in the first step, which grows exponentially with respect to p. As a result, an efficient implementation method, referred to as a ‘shortcut’, is required. A shortcut yields the same results as the standard closed testing procedure but without the need to test all $2^p - 1$ intersection hypotheses. For example, in terms of familywise error rate control, the Holm method is the shortcut of a closed testing procedure that uses the Bonferroni correction to test intersection hypotheses in the first step. In the following, we show that Algorithm 1 is actually an exact shortcut for closed testing using local test statistic (20) with $w_{i,j}^I = \mathbb{1}_{r_j^I > |I| - b_i^I}$, that is,

$$\widehat{L}_i^I(b_i^I) = \sum_{j \in I} \mathbb{1}_{r_j^I > |I| - b_i^I} \mathbb{1}_{D_j = 1}, \quad i = 1, \ldots, m, \tag{22}$$

where b_i^I's are tuning parameters and r_j^I is the local rank of $|W_j|$ in $\{|W_j|, j \in I\}$. For example, $r_1^I = 3, r_3^I = 2, r_4^I = 1$ for $I = \{1, 3, 4\}$. Note that the test statistic $\widehat{L}_i^I(b_i^I)$ counts the number of positive W_j among the first b_i^I based on the local rank for $j \in I$.

Specifically, the following Theorem 4.1 shows that the output of Algorithm 1 [i.e. our previously proposed FDP bound $\overline{\mathrm{FDP}}_{(k,v)}^m(\cdot)$] coincides with the FDP bound based on closed testing, provided the latter is used with local test statistic (22) with $b_i^I = k_i - v_i + 1$ and critical values $z_i^I = k_i$.

Theorem 4.1 For $\boldsymbol{v} = (v_1, \ldots, v_m)$ and $\boldsymbol{k} = (k_1, \ldots, k_m)$, let $\overline{\mathrm{FDP}}^{ct}(R)$ [see (19)] be the FDP bound based on closed testing using test statistic (22) with $b_i^I = k_i + v_i - 1$ and critical values $z_i^I = k_i$ to locally test H_I. Then

$$\overline{\mathrm{FDP}}^{ct}(R) = \overline{\mathrm{FDP}}_{(k,v)}^m(R), \quad R \subseteq [p].$$

Proposition 3.2 and Theorem 4.1 directly imply that the interpolation version of the FDP bound from Katsevich and Ramdas (2020) is also a special case of closed testing:

Corollary 4.1 Let $\overline{\mathrm{FDP}}^{ct}(R)$ [see (19)] be the FDP bound based on closed testing using test statistic (22) with $m = |I|$, $b_i^I = k_i^{raw} + i - 1$ and critical values $z_i^I = k_i^{raw}$ to locally test H_I, then

$$\overline{\mathrm{FDP}}^{ct}(R) = \overline{\mathrm{FDP}}^{\mathrm{KR}}(R), \quad R \subseteq [p].$$

We point out that instead of using Theorem 4.1, one can also apply a general method from Goeman et al. (2021) to connect the simultaneous FDP bounds $\overline{\mathrm{FDP}}^{\mathrm{KR}}(\cdot)$ [and $\overline{\mathrm{FDP}}_{(k,v)}^m(\cdot)$] to closed testing (see Appendix C of the online supplementary material). This approach, however, only results in $\overline{\mathrm{FDP}}^{ct}(R) \le \overline{\mathrm{FDP}}^{\mathrm{KR}}(R)$ (see Proposition C.1 in Appendix C.4 of the online supplementary material). Our proof of Theorem 4.1 is more targeted to the knockoff setting we consider here and does not rely on the method from Goeman et al. (2021). As a result, it leads to the more precise characterization $\overline{\mathrm{FDP}}^{ct}(R) = \overline{\mathrm{FDP}}^{\mathrm{KR}}(R)$ and unifies all simultaneous FDP bounds into the closed testing framework.

Finally, as a side result, we show a connection between the k-FWER control method of Janson and Su (2016) and closed testing. Specifically, for a given k, the method of Janson and Su (2016) outputs a k-FWER controlled set $\widehat{S}^{\mathrm{JS}}(k)$ (see Section 2.2). The following Proposition 4.1 shows that in the case of $|\widehat{S}^{\mathrm{JS}}(k)| \ge k-1$, the same k-FWER guarantee for $\widehat{S}^{\mathrm{JS}}(k)$ is obtained by closed testing using local test statistic (22) with $m = 1$, $b_1^I = k - v + 1$ and $z_1^I = k$.

Proposition 4.1 For the closed testing procedure using test statistic (22) with $m = 1$, $b_1^I = k + v - 1$ and critical value $z_1^I = k$ to locally test H_I, let $t_\alpha(\widehat{S}^{\mathrm{JS}}(k))$ be the corresponding false discovery upper bound defined by (18). Then $t_\alpha(\widehat{S}^{\mathrm{JS}}(k)) = |\widehat{S}^{\mathrm{JS}}(k)|$ when $|\widehat{S}^{\mathrm{JS}}(k)| < k-1$ and $t_\alpha(\widehat{S}^{\mathrm{JS}}(k)) = k-1$ when $|\widehat{S}^{\mathrm{JS}}(k)| \ge k-1$.

4.3 Uniform improvement, generalization, and shortcut

Based on Theorem 4.1, a uniform improvement of $\overline{\mathrm{FDP}}^{m}_{(k,v)}(\cdot)$ can be achieved. In particular, when locally testing the null hypothesis H_I using test statistic (22), instead of using the same tuning parameter $b_i^I = k_i + v_i - 1$ and critical value $z_i^I = k_i$ for all $I \subseteq [p]$, we can use the adapted $b_i^I = k_i^{|I|} + v_i - 1$ and $z_i^I = k_i^{|I|}$, where $k_i^{|I|}$ is obtained by using Algorithm 2 with $|I|$, instead of p, as the last input value. Note that the resulting local test is valid because of (23) in Appendix B of the online supplementary material. As $k_i^{|I|} \le k_i$, the new local test is uniformly better than the previous one, in the sense that it must locally reject H_I if the previous local test rejects. This justifies the uniform improvement of closed testing over the previous FDP bound $\overline{\mathrm{FDP}}^{m}_{(k,v)}(\cdot)$. As we show in Section 5.2, such improvement is strict in some settings.

Aside from uniform improvement, the closed testing framework also enables generalizations of $\overline{\mathrm{FDP}}^{m}_{(k,v)}(\cdot)$ by using other local tests. As a first attempt, we incorporate the local rank r_j^I in the weights of test statistic (20). In particular, we use $w_{i,j}^I = r_j^I \mathbb{1}_{r_j^I > |I| - b_i^I}$ instead of $w_{i,j}^I = \mathbb{1}_{r_j^I > |I| - b_i^I}$. In simulations, we find that this local test generally performs worse than the method without including local rank, but there also exist settings where it performs better, see Appendix E.5 of the online supplementary material for details.

In addition, the closed testing framework provides a direct way to incorporate background knowledge. For instance, one may use a large weight $w_{i,j}^I$ in test statistic (20) if one expects that the ith variable is non-null. Moreover, it is important to remark that since all proposed procedures are valid conditional on the absolute statistics $|W_1|, \ldots, |W_p|$, the user may observe these before selecting the final procedure to use.

For the above uniformly improved [i.e. using local test statistic (22) with $b_i^I = k_i^{|I|} + v_i - 1$ and critical value $z_i^I = k_i^{|I|}$] and rank-generalization versions [i.e. using local test statistic (20) with $w_{i,j}^I = r_j^I \mathbb{1}_{r_j^I > |I| - b_i^I}$, $b_i^I = k_i^{|I|} + v_i - 1$ and critical values calculated based on inequality (21)], the analytical shortcut $\overline{\mathrm{FDP}}^{m}_{(k,v)}(\cdot)$ (see Theorem 4.1) for obtaining FDP bound based on closed testing is not valid anymore. Hence, a new shortcut is needed for their practical implementations. To this end, we propose a general shortcut which is valid when the following two conditions hold:

- (C1) The critical values $z_1^I, \ldots, z_m^I$ depend on I only through $|I|$.
- (C2) For any two sets $I_1 = \{u_1, \ldots, u_s\}$ and $I_2 = \{v_1, \ldots, v_s\}$ of the same size, there exists a permutation $\pi : [p] \to [p]$ such that if $\pi^{I_1}_{(j)} \le \pi^{I_2}_{(j)}$ for all $j = 1, \ldots, s$, then $\widehat{L}_i^{I_1} \le \widehat{L}_i^{I_2}$ for all $i \in [m]$, where $\pi^I_{(1)} < \cdots < \pi^I_{(s)}$ are ordered values of $\{\pi(i), i \in I\}$.

Condition (C2) pertains to a form of monotonicity of the test statistic, which, in turn, leads to computational reductions. These two conditions ensure the following for a null hypothesis H_I with $|I| = s$: instead of checking whether all subsets of I of size $t \in \{s, \ldots, 1\}$ are rejected by closed testing, it is sufficient to check only one set B of each size t (see Algorithm 3 for the definition of set B). Moreover, instead of checking whether all supersets of B of size $r \in \{0, \ldots, p-t\}$ are locally rejected, we again only need to check one set of each size. This brings us the necessary computational reduction in the implementation of closed testing.

Note that these two conditions are on the local test method itself, rather than the data, so one can check whether their methods satisfy these conditions. In particular, two examples where conditions (C1) and (C2) hold are the above mentioned uniformly improved and rank-generalization versions. Specifically, one such permutation that satisfying condition (C2) can be obtained by letting $\pi(i) = l$ such that $W_i = W_{(l)}$ in the sorted $W_{(1)} < \cdots < W_{(p)}$.

We summarize the shortcut of getting FDP bounds using closed testing in Algorithm 3, and show its validity in Appendix A of the online supplementary material. The computational complexity of Algorithm 3 is $O(p^2)$ in the worst case, which is larger than the previous analytical shortcut Algorithm 1. We think that this is a fair price to pay for a more general shortcut, in the sense that it can be applied to more types of local tests. In practice, one may first use Algorithm 1, and only use the uniformly improved version based on closed testing (with the shortcut provided in Algorithm 3) if the result is not already satisfactory.

Algorithm 3 A shortcut for closed testing

Input: (R, π), where $R \subseteq [p]$ is any non-empty set of size s, and π is a permutation satisfying condition (C2).
Output: FDP upper bound for R, which holds simultaneously for all $R \subseteq [p]$.
1: **for** $t = s, \ldots, 1$ **do**
Let $B = \{\pi^{-1}(\pi^R_{(1)}), \ldots, \pi^{-1}(\pi^R_{(t)})\} \subseteq R$ (i.e. the index set of the t smallest $\pi(i)$ with $i \in R$).
2: **for** $r = 0, \ldots, p-t$ **do**
if $r = 0$, **then** let $U_{t+r} = B$, **else** Let $U_{t+r} = B \cup B^c_r$, where $B^c_r = \{\pi^{-1}(\pi^{B^c}_{(1)}), \ldots, \pi^{-1}(\pi^{B^c}_{(r)})\}$ (i.e. the index set of the r smallest $\pi(i)$ with $i \in B^c$).
For $j \in [m]$, let $\widehat{L}^{U_{t+r}}_j$ be the test statistic related to $H_{U_{t+r}}$ and $z^{U_{t+r}}_j$ be the corresponding critical value.
if $\widehat{L}^{U_{t+r}}_j < z^{U_{t+r}}_j$ for all $j \in [m]$, **then return** t/s.
3: **Return** 0.

5 Simulations and a real data application

We now examine the finite sample performance of our proposed methods in a range of settings and compare them to the method of Katsevich and Ramdas (2020). Since comparing FDP bounds for all sets $R \subseteq [p]$ is computationally infeasible, we follow Katsevich and Ramdas (2020) and focus on the nested sets $\widehat{R}_i = \{j \le i : W_j > 0\}, i \in [p]$. All simulations were carried out in R, and the code is available at https://github.com/Jinzhou-Li/KnockoffSimulFDP.

5.1 Comparison of Algorithm 1 and the method of Katsevich and Ramdas (2020)

We consider the following methods:

- KR: The simultaneous FDP bound from Katsevich and Ramdas (2020) [see (15)]. We use the R code from https://github.com/ekatsevi/simultaneous-fdp for its implementation.
- KJI: Our proposed simultaneous FDP bounds based on Knockoffs and Joint k-familywise error rate control with Interpolation (see Algorithm 1). We implement four types of tuning parameters $(\boldsymbol{v}, \boldsymbol{k})$, where $\boldsymbol{v}$ is as described in (14) and $\boldsymbol{k}$ is obtained using Algorithm 2, and we denote them as KJI-A, KJI-B, KJI-C, and KJI-D, respectively.

All methods use the same knockoff statistic vectors $\boldsymbol{W}$. We consider the following two settings:

(a) Linear regression setting: We first generate $X \in \mathbb{R}^{n\times p}$ by drawing n i.i.d. samples from the multivariate Gaussian distribution $N_p(0, \Sigma)$ with $(\Sigma)_{i,j} = 0.6^{|i-j|}$. Then we generate $\beta \in \mathbb{R}^{p\times 1}$ by randomly sampling $s \times p$ non-zero entries from $\{1, \ldots, p\}$ with sparsity level $s \in \{0.1, 0.2, 0.3\}$, and we set the signal amplitude of the non-zero β_j's to be $a/\sqrt{n}$, where n is the sample size and the amplitude $a \in \{6, 8, 10\}$. Finally, we sample $Y \sim N_n(X\beta, I)$. The design matrix X is fixed over different replications in the simulations. Based on the generated data (X, Y), we obtain $\boldsymbol{W}$ by using the fixed-X 'sdp' knockoffs and the signed maximum lambda knockoff statistic. Similar settings and the same knockoff statistics are considered in Barber and Candès (2015) and Janson and Su (2016).

(b) Logistic regression setting: For $i = 1, \ldots, n$, we first sample $x_i \sim N_p(0, \Sigma)$ with $(\Sigma)_{i,j} = 0.6^{|i-j|}$. Then we sample $y_i \sim \text{Bernoulli}(1/1 + e^{-x_i^T\beta})$, where β is generated as in (a), except now we consider amplitude $a \in \{8, 10, 12\}$. Based on the generated data (x_i, y_i), $i = 1, \ldots, n$, we obtain W by using the 'asdp' model-X knockoffs and the Lasso coefficient-difference statistic with 10-fold cross-validation. Similar settings and the same knockoff statistics are considered in Candès et al. (2018).

We take $p = 200$ and $n = 500$ and use the four types of v described in (14) with $v_m < 65$. The reported FDP bounds are average values over 200 replications. We also tried the case of $p = 1{,}000$ and $n = 2{,}500$, and a high-dimensional linear case with $p = 800$ and $n = 400$. The simulation results convey similar information, see Appendix E.3 of the online supplementary material.

As expected, the simultaneous FDP guarantee (2) holds for all methods, see Figure 9 in Appendix E.3 of the online supplementary material. Figures 3 and 4 show the FDP bounds of KR and KJI with the four types of $(\boldsymbol{v}, \boldsymbol{k})$ in the linear and logistic regression settings, respectively. One can see that KJI-A gives better FDP bounds for all $\widehat{R}_i$ than KR over all settings, which verifies Proposition 3.3. The improvement is larger in denser settings with a smaller signal. All methods give similar FDP bounds in settings with sparsity level 0.1 and large amplitude. The reason is that in such settings, almost all true W_i's are positive and have larger absolute values than any null W_i, so the FDP bounds based on the reference set $\widehat{S}(v_1 = 1)$ and interpolation are the best possible one can obtain. We also observe that KJI with different types of $(\boldsymbol{v}, \boldsymbol{k})$ can give better FDP bounds for different $\widehat{R}_i$ under different settings, and there is no uniformly best one. Finally, we would like to mention that the tightness of FDP bounds depends on the distribution of knockoff vector W, which in turn, is influenced by the data distribution, the methodology used to construct knockoffs, and the choice of feature statistics.

5.2 Comparison of Algorithm 1 and the closed testing-based approach

We now compare KJI described in Section 5.1 to the following method:

- KCT: Our proposed simultaneous FDP bounds based on Knockoffs and Closed Testing using local test statistic (22) with $b_i^I = k_i^{|I|} + v_i - 1$ and the critical value $z_i^I = k_i^{|I|}$. We use the same four types of v as for KJI and obtain $k^{|I|}$ by Algorithm 2 with input $(v_1, \ldots, v_m, \alpha, \delta = 0.01, |I|)$, and we denote them as KCT-A, KCT-B, KCT-C, and KCT-D. We use Algorithm 3 as shortcut for the implementation of closed testing.

We applied KJI and KCT in the same linear and logistic regression settings as in Section 5.1 with $p = 200$ and $n = 500$. As expected, the FDP bounds of KCT are indeed smaller than or equal to that of KJI for the same tuning parameter vector v. In most cases, however, the corresponding FDP bounds are identical, and the improvement is typically very tiny in the cases where they are not identical (see Appendix E.4 of the online supplementary material).

To understand this, it is helpful to think of the closed testing equivalent form of KJI (see Theorem 4.1). Even though the local tests of KCT are uniformly more powerful than the local tests of KJI, and hence can make more local rejections, these local rejections do not necessarily lead to

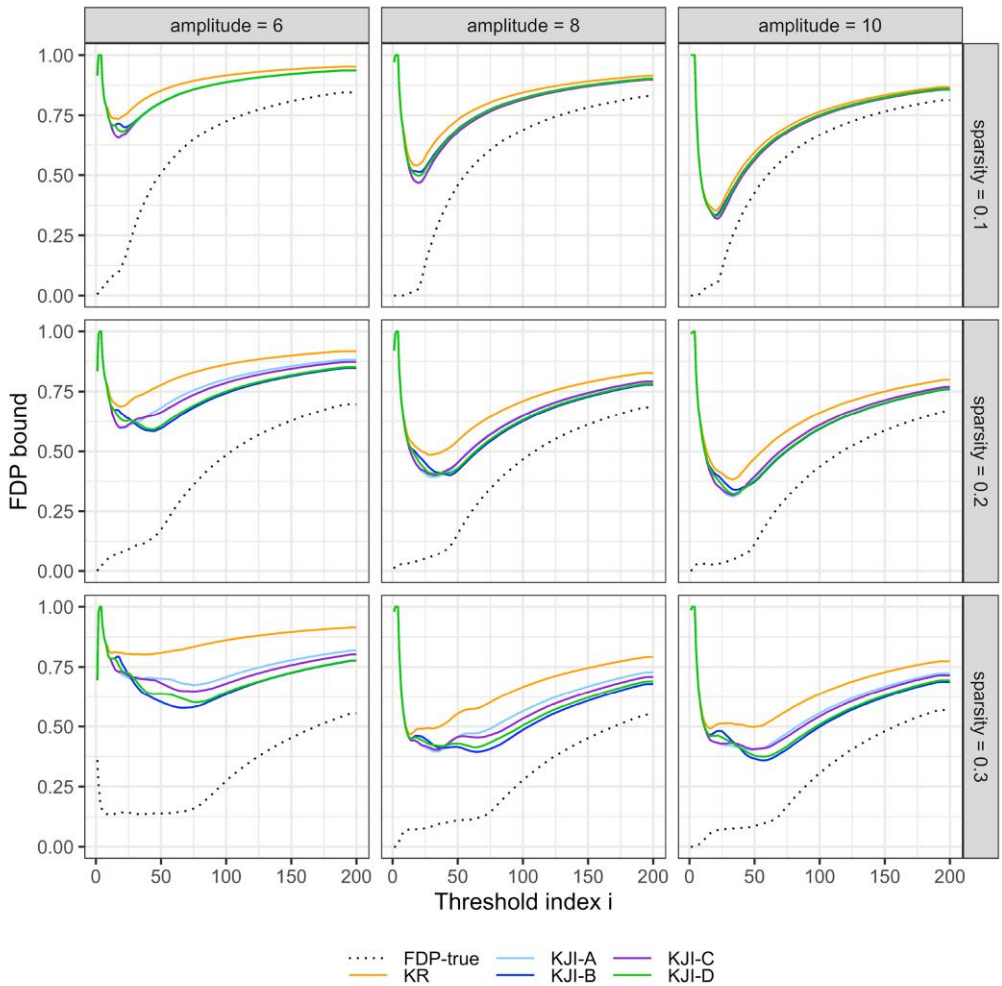

Figure 3. Simultaneous FDP bounds of KR and KJI in the linear regression setting with $p = 200$ and $n = 500$. The dashed black line indicates the true FDP. All FDP bounds are average values over 200 replications.

more rejections by closed testing, especially in the sparse settings we consider here. Therefore, the final FDP bounds can be the same.

These simulation results also carry a valuable implication. Specifically, given that Goeman et al. (2021) proved that only closed testing methods are admissible and KJI is a special case of closed testing (see Theorem 4.1), the observation that the uniform improvement is very small implies that KJI is nearing optimal, and that further substantial uniform improvements are unlikely.

At last, to verify our claim in Section 4.3 that the uniform improvement by KCT over KJI can be non-trivial in certain cases, we consider a simulation setting that directly generates the knockoff statistic vector W. Specifically, we consider a dense setting with $p = 50$ and null variable set $\mathcal{N} = \{10, \ldots, 20\}$. For the knockoff statistics, we set $|W_i| = 50 - i + 1$, $\text{sign}(W_i) = 1$ for $i \notin \mathcal{N}$ and $\text{sign}(W_i)$i.i.d. $\sim\{-1, 1\}$ with the same probability 1/2 for $i \in \mathcal{N}$. It is clear that the generated W satisfies the coin-flip property, so it is a valid knockoff statistic vector. Figure 5 shows the simulation results in this setting. We can see that the improvements of KCT over KJI can be large for some $\widehat{R}_i$.

5.3 Real data application

We now apply our methods to genome-wide association study (GWAS) data sets from the UK Biobank (Bycroft et al., 2018). A GWAS data set typically consists of genotypes and traits of

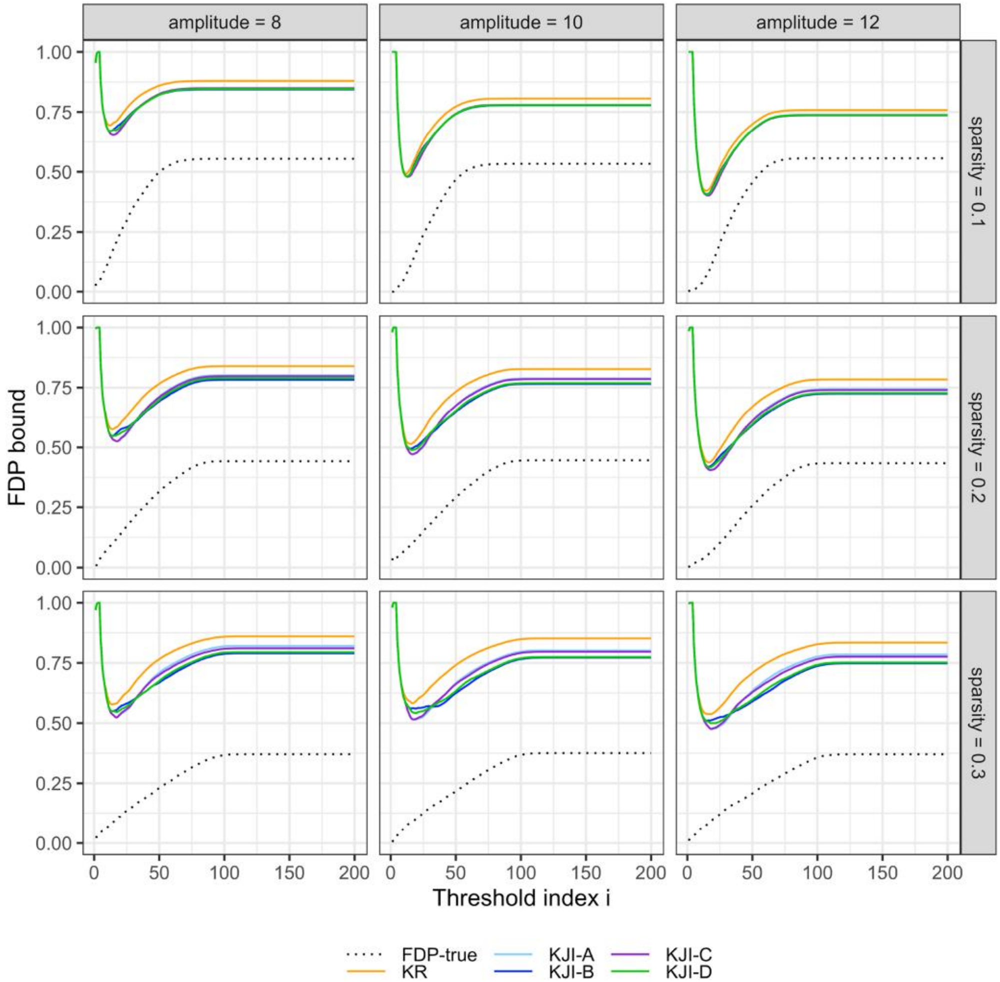

Figure 4. Simultaneous FDP bounds of KR and KJI in the logistic regression setting with $p = 200$ and $n = 500$. The dashed black line indicates the true FDP. All FDP bounds are average values over 200 replications.

different individuals, and the goal is to identify genotypes that are conditionally dependent on a trait given the other genotypes (i.e. variable selection). The traits considered in our analysis are height, body mass index, platelet count, systolic blood pressure, cardiovascular disease, hypothyroidism, respiratory disease, and diabetes.

We follow the analysis of Katsevich and Ramdas (2020) (see https://github.com/ekatsevi/simultaneous-fdp) and use the knockoff statistics $\boldsymbol{W} = (W_1, \ldots, W_p)$ constructed by Sesia et al. (2020) (see https://msesia.github.io/knockoffzoom/ukbiobank.html). We assume that the model-X assumption holds, as do previous analyses in the literature. For each set $\widehat{R}_i = \{j \le i : W_j > 0\}$, we compute simultaneous FDP bounds using KR and KJI with the four types of tuning parameter vectors $\boldsymbol{v}$ [see (14)] with $v_m < 1,200$ (see also Section 5.1). Since we do not have the true FDP, we estimate it by $\widehat{\mathrm{FDP}}(\widehat{R}_i) = \frac{1+\{j \le i,\ W_j < 0\}}{\max\{|\widehat{R}_i|,\ 1\}}$. The latter also leads to a corresponding FDR control method: taking the largest set $\widehat{R}_i$ such that $\widehat{\mathrm{FDP}}(\widehat{R}_i) \le q$ provides FDR control at level $q \in (0, 1)$ (Barber & Candès, 2015).

Figure 6 shows the results for the trait platelet count, showing the simultaneous FDP bound versus the number of rejections of each considered set. We see that KJI with all four types of $\boldsymbol{v}$ gives smaller bounds than KR. One may also notice that the simultaneous FDP bounds based on the

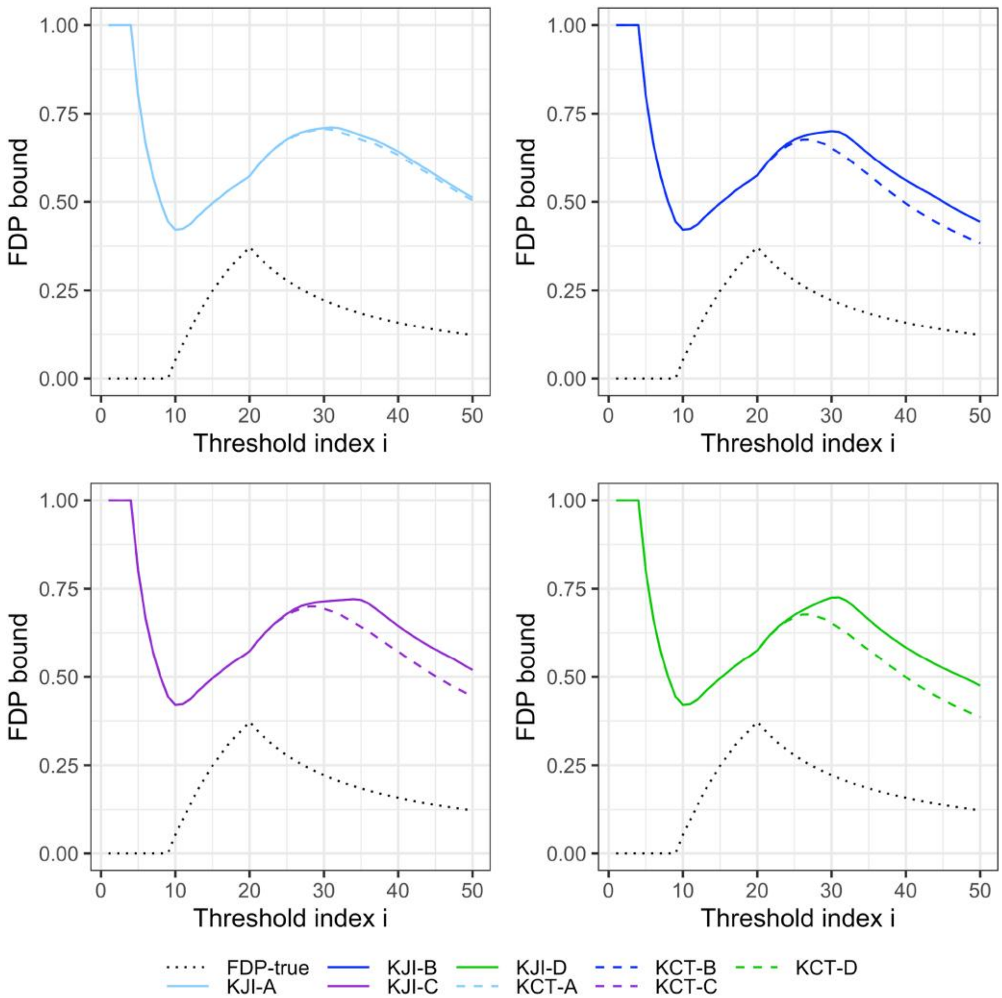

Figure 5. Simultaneous FDP bounds of KJI and KCT in the simulation setting where the knockoff statistic vector $W \in \mathbb{R}^{50}$ is directly generated. The dashed black line indicates the true FDP. All FDP bounds are average values over 200 replications.

different $\boldsymbol{v}$ vectors are quite similar, especially for sets with more than about 3,000 rejections. The latter can be explained by the fact that the largest reference sets for the four types of $\boldsymbol{v}$ vectors are similar, and hence the simultaneous FDP bounds based on interpolation beyond these sets are also similar.

Figure 7 shows the number of discoveries for all traits, when controlling the FDR at level 0.1 or the simultaneous FDP at levels 0.05, 0.1, and 0.2. We see that the four versions of KJI yield similar numbers of discoveries in most cases, and that these numbers are generally larger than the number of discoveries of KR.

To illustrate the flexibility of using simultaneous FDP bounds, we take a closer look at the results for the trait height in Figure 7. We see that KJI-B returns 4,157 discoveries when controlling the simultaneous FDP at level 0.2. Thus, with probability larger than 0.95, KJI-B guarantees more than $0.8 \times 4, 157 = 3, 325$ true discoveries among those 4,157 discoveries. If this set is somehow too large for practical purposes, one may switch to lower levels like 0.1 or 0.05, without losing control. For example, at level 0.05, we obtain a set of 1,607 discoveries, meaning that, with probability larger than 0.95, we are guaranteed more than $0.95 \times 1, 607 = 1, 526$ true discoveries among these. One can also look at other levels or sets, and the simultaneous FDP bounds remain valid. This flexibility is extremely useful in exploratory data analysis.

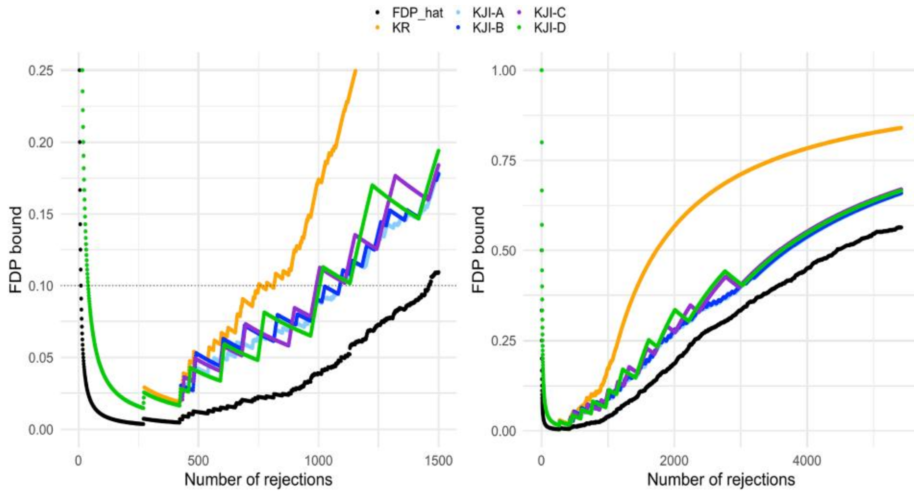

Figure 6. Simultaneous FDP bounds computed by KR and KJI versus the number of rejections for each considered set $\widehat{R}_i$, for the trait platelet count of the UK Biobank data. The black dotted line is the estimated FDP. The left plot is the zoom-in version of the right plot.

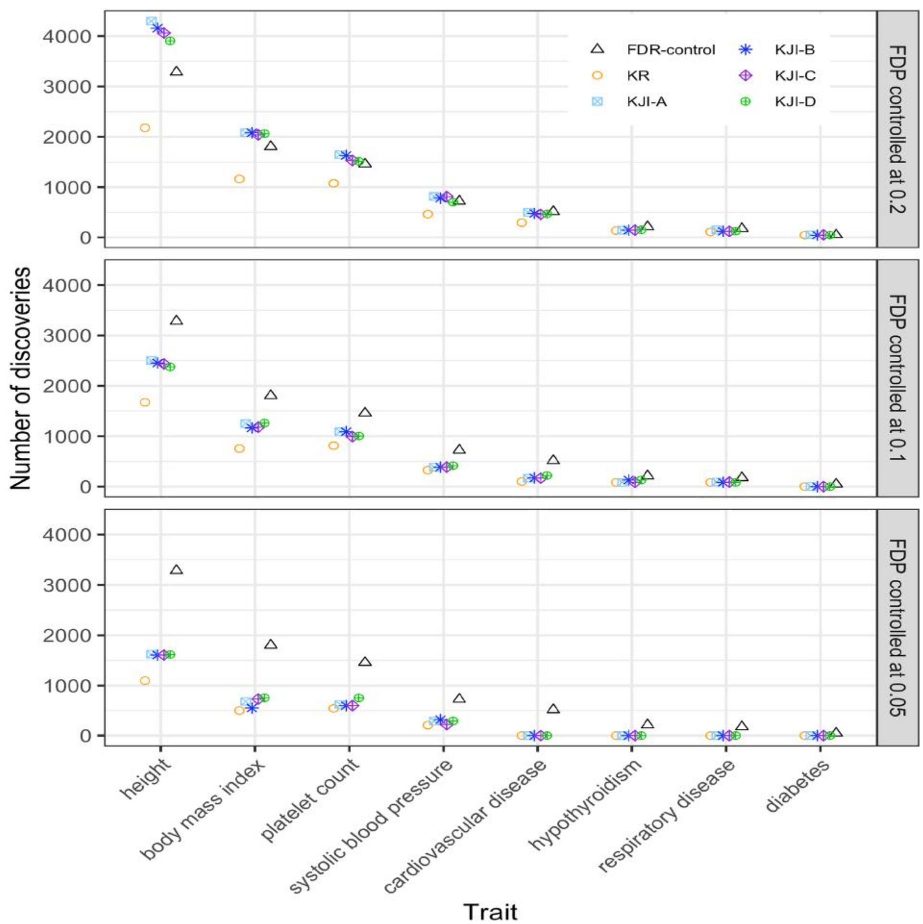

Figure 7. The number of discoveries obtained for all traits of the UK Biobank data, when controlling the FDR at level 0.1, or the simultaneous FDP at levels 0.05, 0.1, and 0.2 by using KR and KJI.

FDR control methods do not possess this kind of flexibility. For example, considering different levels will break the FDR control. Moreover, FDR control is only in expectation. In the data example, we get 3, 284 discoveries when controlling the FDR at level 0.1, meaning that we are guaranteed more than $0.9 \times 3, 248 = 2, 955$ true discoveries among these, in the expectation sense.

6 Discussion

We studied the problem of obtaining simultaneous FDP bounds for knockoff-based approaches. We first proposed a method based on joint k-FWER control and interpolation (Algorithm 1), and we showed that it is a generalization of the only existing approach for this problem by Katsevich and Ramdas (2020), in the sense that the latter is a special case of Algorithm 1 with a specific choice of tuning parameters. We also suggested other choices for the tuning parameters, and showed that one of them was uniformly better. Next, we proposed a method based on closed testing and showed that Algorithm 1 is a special case (or exact shortcut) of this method. Using closed testing, we showed that Algorithm 1 can be further uniformly improved, and that other generalizations can be derived. We also developed a new shortcut for the implementation of this closed testing based method.

Closed testing is a very general and powerful methodology in which one has the flexibility to choose different local tests. Different local tests can be better in different settings. One natural problem is to obtain optimal data-dependent local tests, while maintaining valid simultaneous FDP bounds. Sample-splitting is a possible solution, but it can be sub-optimal, especially in cases where the sample size is not large. This is an interesting problem for future research.

The key assumption underlying our methods is that the knockoff statistics satisfy the coin-flip property. In some cases, however, this assumption does not hold. Examples are Gaussian graphical models (Li & Maathuis, 2021) where the knockoff statistic matrix for all edges does not possess this property, or the model-X knockoff setting when the model-X assumption does not hold, but an estimated distribution of the covariates is used (Barber et al., 2020). Developing simultaneous FDP bounds for such settings is another interesting direction for future work.

Acknowledgments

We are grateful to the anonymous reviewers for many helpful comments and suggestions, which greatly helped to improve the manuscript.

Conflict of interests: None declared.

Funding

J.L. was supported by the Swiss National Science Foundation (SNSF) grant P500PT-210978.

Data availability

The original genome-wide association study (GWAS) data sets used for this study are from the UK Biobank (https://www.ukbiobank.ac.uk/). We directly use the knockoff statistics constructed by Sesia et al. (2020).

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Barber R. F., & Candès E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/15-AOS1337

Barber R. F., & Candès E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, *47*(5), 2504–2537. https://doi.org/10.1214/18-AOS1755

Barber R. F., Candès E. J., & Samworth R. J. (2020). Robust inference with knockoffs. *The Annals of Statistics*, *48*(3), 1409–1431. https://doi.org/10.1214/19-AOS1852

Benjamini Y., & 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)*, *57*(1), 289–300. https://doi.org/10.1111/rssb.1995.57.issue-1

Benjamini Y., & Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Blanchard G., Neuvial P., & Roquain E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, *48*(3), 1281–1303. https://doi.org/10.1214/19-AOS1847

Bycroft C., Freeman C., Petkova D., Band G., Elliott L. T., Sharp K., Motyer A., Vukcevic D., Delaneau O., O'Connell J., Cortes A., Welsh S., Young A., Effingham M., McVean G., Leslie S., Allen N., Donnelly P., & Marchini J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Candès E., Fan Y., Janson L., & 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. https://doi.org/10.1111/rssb.12265

Chi C.-M., Fan Y., & Lv J. et al. (2021). 'High-dimensional knockoffs inference for time series data', arXiv, arXiv:2112.09851, preprint: not peer reviewed.

Dai R., & Barber R. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In *International conference on machine learning* (Vol. 48, pp. 1851–1859). PMLR.

Dobriban E. (2020). Fast closed testing for exchangeable local tests. *Biometrika*, *107*(3), 761–768. https://doi.org/10.1093/biomet/asz082

Genovese C., & Wasserman L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics*, *32*(3), 1035–1061. https://doi.org/10.1214/009053604000000283

Genovese C. R., & Wasserman L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, *101*(476), 1408–1417. https://doi.org/10.1198/016214506000000339

Goeman J. J., Hemerik J., & Solari A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, *49*(2), 1218–1238. https://doi.org/10.1214/20-AOS1999

Goeman J. J., Meijer R. J., Krebs T. J., & Solari A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, *106*(4), 841–856. https://doi.org/10.1093/biomet/asz041

Goeman J. J., & Solari A. (2011). Multiple testing for exploratory research. *Statistical Science*, *26*(4), 584–597.https://doi.org/10.1214/11-STS356

He Z., Liu L., Belloy M. E., Le Guen Y., Sossin A., Liu X., Qi X., Ma S., Wyss-Coray T., Tang H., Sabatti C., Candès E., Greicius M. D., & Ionita-Laza I. (2021). *Summary statistics knockoff inference empowers identification of putative causal variants in genome-wide association studies*. bioRxiv 2021.12.06.471440. https://doi.org/10.1101/2021.12.06.471440, December 7, 2021, preprint: not peer reviewed.

Hemerik J., Solari A., & Goeman J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, *106*(3), 635–649. https://doi.org/10.1093/biomet/asz021

Janson L., & Su W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, *10*(1), 960–975. https://doi.org/10.1214/16-EJS1129

Katsevich E., & Ramdas A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, *48*(6), 3465–3487. https://doi.org/10.1214/19-AOS1938

Katsevich E., & Sabatti C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, *13*(1), 1–33. https://doi.org/10.1214/18-AOAS1185

Li J., & Maathuis M. H. (2021). GGM knockoff filter: False discovery rate control for Gaussian graphical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(3), 534–558. https://doi.org/10.1111/rssb.12430

Li S., Sesia M., Romano Y., Candès E., & Sabatti C. (2021). Searching for consistent associations with a multi-environment knockoff filter. *Biometrika*, *109*(3), 611–629. https://doi.org/10.1093/biomet/asab055

Marcus R., Eric P., & Gabriel K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, *63*(3), 655–660. https://doi.org/10.1093/biomet/63.3.655

Meinshausen N. (2006). False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, *33*(2), 227–237. https://doi.org/10.1111/sjos.2006.33.issue-2

Sesia M., Bates S., Candès E., Marchini J., & Sabatti C. (2021). False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(40). https://doi.org/10.1073/pnas.2105841118

Sesia M., Katsevich E., Bates S., Candès E., & Sabatti C. (2020). Multi-resolution localization of causal variants across the genome. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-14791-2

Tian J., Chen X., Katsevich E., Goeman J., & Ramdas A. (2023). Large-scale simultaneous inference under dependence. *Scandinavian Journal of Statistics*, *50*(2), 750–796. https://doi.org/10.1111/sjos.v50.2

Van der Laan M. J., Dudoit S., & Pollard K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1.https://doi.org/10.2202/1544-6115.1042

Vesely A., Finos L., & Goeman J. J. (2023). Permutation-based true discovery guarantee by sum tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(3), 664–683. https://doi.org/10.1093/jrsssb/qkad019