Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Derandomizing Knockoffs

Zhimei Ren, Yuting Wei & Emmanuel Candès

To cite this article: Zhimei Ren, Yuting Wei & Emmanuel Candès (2023) Derandomizing Knockoffs, Journal of the American Statistical Association, 118:542, 948-958, DOI: 10.1080/01621459.2021.1962720

To link to this article: https://doi.org/10.1080/01621459.2021.1962720

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 542, 948–958: Theory and Methods
https://doi.org/10.1080/01621459.2021.1962720

Derandomizing Knockoffs

Zhimei Ren[a], Yuting Wei[b], and Emmanuel Candès[c]

[a]Department of Statistics, University of Chicago, Chicago, IL; [b]Statistics & Data Science Department, University of Pennsylvania, Philadelphia, PA; [c]Department of Mathematics, Department of Statistics, Stanford University, Stanford, CA

ABSTRACT

Model-X knockoffs is a general procedure that can leverage any feature importance measure to produce a variable selection algorithm, which discovers true effects while rigorously controlling the number or fraction of false positives. Model-X knockoffs is a randomized procedure which relies on the one-time construction of synthetic (random) variables. This article introduces a derandomization method by aggregating the selection results across multiple runs of the knockoffs algorithm. The derandomization step is designed to be flexible and can be adapted to any variable selection base procedure to yield *stable* decisions without compromising statistical power. When applied to the base procedure of Janson and Su, we prove that *derandomized knockoffs* controls both the per family error rate (PFER) and the k family-wise error rate (k-FWER). Furthermore, we carry out extensive numerical studies demonstrating tight Type I error control and markedly enhanced power when compared with alternative variable selection algorithms. Finally, we apply our approach to multistage genome-wide association studies of prostate cancer and report locations on the genome that are significantly associated with the disease. When cross-referenced with other studies, we find that the reported associations have been replicated. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received December 2020
Accepted July 2021

KEYWORDS
Family-wise error rate; Genome-wide association study; Knockoff filter; Multiple hypothesis testing; Per family error rate; Variable selection

1. Introduction

There has been a surge of interest in the design of trustworthy inferential procedures for massive data applications with a focus on the development of variable selection algorithms that are flexible, and at the same time, possess clear performance guarantees. Among them, the method of knockoffs, or knockoffs for short (Barber and Candès 2015; Candès et al. 2018), has proved particularly effective in a variety of applications (Gao et al. 2018; Srinivasan, Zhan, and Xue 2019; Sesia et al. 2020). Imagine a scientist wishes to infer which of the many covariates she has measured are truly associated with a response of interest; for instance, which of the many genetic variants influence the susceptibility of a disease. At a high level, the knockoffs selection algorithm begins by synthesizing "fake" copies of the covariates (fake genetic variants in our example), which can be thought of as serving as a control group for the features. By contrasting the values a feature importance statistic takes on when applied to a true variable and a fake variable, it becomes possible to tease apart those features which have a true effect on the response. This can be achieved via a clever filter while controlling either the expected fraction of false positives (Barber and Candès 2015) or simply the number of false positives (Janson and Su 2016).

A frequently discussed issue is that knockoffs is a randomized procedure; that is, the fake covariates (the knockoffs) are stochastic. Therefore, different runs of the algorithm produce different knockoffs (unless we use the same random seed), and in big data applications, researchers have observed that the selection algorithm may each time return overlapping, yet, different selected sets. This has led researchers to report those features whose selection frequency exceeds a threshold along with the corresponding frequencies (Candès et al. 2018; Sesia, Sabatti, and Candès 2019). While statisticians are accustomed to randomized procedures—after all, any procedure based on data splitting is randomized in the sense that different splits typically yield different outcomes—it is still desirable to derandomize the knockoffs selection algorithm as to produce consistent results. This article proposes a new procedure *derandomized knockoffs* that achieves this goal by running the knockoffs algorithm several times and aggregating results across all runs.

1.1. An Overview of Our Contributions

Our derandomization scheme is inspired by the stability selection framework of Meinshausen and Bühlmann (2010) and Shah and Samworth (2013), which finds its roots in bootstrap aggregating (Efron and Gong 1983; Breiman 1996, 1999), subbagging (Bühlmann et al. 2002), and random forest (Breiman 2001). In a nutshell, our scheme consists in applying a knockoffs selection algorithm multiple times, each time with a new matrix of knockoffs, and proceeds by aggregating the results using the same rationale behind the stability selection criterion, which as its name suggests, puts a premium on stability and consistency. We will demonstrate that this indeed reduces the variability of the outcome. In Section 2, we will however explain why the similarities between stability selection and derandomized

CONTACT Zhimei Ren zmren@uchicago.edu Department of Statistics, University of Chicago, 5747 S Ellis Ave, Chicago, IL 60637-1403.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

knockoffs stop here, and why the interpretation and properties of the two procedures are very different. Moving on, we empirically demonstrate that derandomized knockoffs achieves tight Type I error control and markedly enhanced power when compared with alternative variable selection algorithms, including "vanilla" knockoffs. We establish theoretical support for derandomized knockoffs by proving per family error rate (PFER) control and k family-wise error rate (k-FWER) control.

Besides methodological developments, a fair fraction of this article is concerned with applying our ideas to genome-wide association studies (GWAS) in Section 5. We make two contributions.

- In the previous applications of knockoffs to GWAS, the base procedure is typically applied multiple times with different random seeds. While each run comes with a Type I error guarantee, the authors often report genetic variants together with their selection frequency to identify variants which are consistently discovered, see Ren and Candès (2020) for some examples. One issue is that we would not know how to interpret a "meta-set" of variants whose selection frequency is above a given threshold. By this, we mean that we would not be able to give this meta-set Type I error guarantees. The methods from this article of fer a remedy.
- We design a general and scientifically sound workflow for multistage GWAS. Suppose we have a family of SNPs $X_1, \ldots, X_p$ and are interested in determining whether the distribution of a phenotype Y *conditional* on $X_1, \ldots, X_p$ depends on X_j or not; that is, we want to know whether Y depends on X_j controlling for all the other variables X_{-j}. We will show how to achieve this in a multistage approach, where one can use the first study to determine a set of candidate SNPs and the second study for confirmatory analysis.

2. A Framework for Derandomizing Knockoffs

2.1. Knockoffs

To set the stage for derandomized knockoffs, imagine we are given a response variable Y and potential explanatory variables $\boldsymbol{X} = (X_1, \ldots, X_p)$. We would like to identify those variables that truly inf luence the response; that is, we would like to discover those X_j's on which the distribution $Y \mid X_1, \ldots, X_p$ depends. Formally, a variable X_j is said to be *null* if the response Y is independent of X_j given all other variables; that is,

$$Y \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-j} \tag{1}$$

(throughout, $\boldsymbol{X}_{-j}$ is a shorthand for all p features except the jth). Our goal is of course to test each of the p nonparametric hypotheses (1).

In this setting, the key idea underlying knockoffs is to generate "fake" covariates $\tilde{\boldsymbol{X}} = (\tilde{X}_1, \ldots, \tilde{X}_p)$ whose distribution roughly matches that of the true covariates, except that knockoffs are designed to be conditionally independent of the response, and hence should never be selected by a feature selection procedure. Assemble the covariates in an $n \times p$ matrix $\mathbf{X}$ and the responses in an $n \times 1$ vector $\boldsymbol{Y}$. Then we say that the new set of variables $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$ is a knockoff copy of $\mathbf{X}$ if the following two properties hold: first,

$$[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(j)} \overset{\text{d}}{=} [\mathbf{X}, \tilde{\mathbf{X}}] \quad \text{for any } j \in \{1, \ldots, p\}, \tag{2}$$

where $[\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(j)}$ denotes the matrix with the jth column of $\mathbf{X}$ and $\tilde{\mathbf{X}}$ swapped. This says that by looking at $\mathbf{X}$ and $\tilde{\mathbf{X}}$ we cannot tell whether the jth column is a true variable or a knockof f. (The point is that if $\mathbf{X}_j$ is non null, then we can tell by looking at $\boldsymbol{Y}$.) The second property is that $\boldsymbol{Y} \perp\!\!\!\perp \tilde{\mathbf{X}} \mid \mathbf{X}$. This says that knockof fs provide no further information about the response (knockoffs are constructed without looking at $\boldsymbol{Y}$).

To perform variable selection, the researcher applies her favorite *feature importance statistic* to the augmented dataset $(\mathbf{X}, \tilde{\mathbf{X}}, \boldsymbol{Y})$ and scores each of the original and knockoff variables. For example, she can score each variable by recording the magnitudes of the Lasso coeficients for a value of the regularization parameter chosen by cross-validation. The scores are then combined to produce a test statistic for each feature. This can be as simple as the difference between the feature importance statistics, for example, the difference between the magnitude of the Lasso coeficient of the original feature and that of its knockoff. In the sequel, we refer to this test statistic as the *Lasso coeficient difference* (LCD, Candès et al. 2018). Finally, the test statistics are passed through the knockoff filter (e.g., SeqStep, Barber and Candès 2015) and a selection set $\hat{S}$ is generated.

2.2. Derandomized Knockoffs

With these preliminaries, our derandomized procedure to stabilize the selection set over dif ferent runs is as follows:

- Construct M conditionally independent knockoff copies $\tilde{\mathbf{X}}^1, \ldots, \tilde{\mathbf{X}}^M \in \mathbb{R}^{n \times p}$.
- For each $m \in [M] := \{1, \ldots, M\}$, apply a *base procedure* to produce a rejection set $\hat{S}^m$.
- For each feature j, compute its *selection frequency* via

$$\hat{\pi}_j := \frac{1}{M} \sum_{m=1}^{M} 1\{j \in \hat{S}^m\}. \tag{3}$$

- Finally, given a threshold $\eta > 0$, return the final selection set

$$\hat{S} := \{j \in [p] : \hat{\pi}_j \geq \eta\}. \tag{4}$$

The above-derandomized knockoffs procedure is summarized in Algorithm 1. (Strictly speaking, for a f inite$M$, the procedure is only approximately derandomized. In the sequel, we call this procedure "derandomized knockoffs" for short to emphasize the difference with the original procedure.) Readers will recognize that Equations (3) and (4) are borrowed from stability selection (please see below for a detailed comparison).

The parameter η controls how many times a variable needs to be selected to be present in the final selection set. The larger η, the fewer variables will ultimately be selected. The choice of η may affect the power of the procedure. In all our simulations and application examples, the value $\eta = 0.5$ works well. One reason is that there are other parameters that jointly determine the Type I error upper bounds, and that consequently, there is not much loss in fixing the value of η and selecting other parameters accordingly. For simplicity, we thus recommend

Algorithm 1: Derandomized knockoffs procedure

Input: Covariate matrix $X \in \mathbb{R}^{n\times p}$; response variables $\boldsymbol{Y} \in \mathbb{R}^n$; number of realizations M; a base procedure; selection threshold η.

1. **for** $m = 1, \ldots, M$ **do**
   i. Generate a knockoff copy $\tilde{X}^m$.
   ii. Run the base procedure with $\tilde{X}^m$ as knockoffs and obtain the selection set $\hat{S}^m$.

end

2. Calculate the selection probability

$$\hat{\pi}_j = \frac{1}{M}\sum_{m=1}^{M} 1\{j \in \hat{S}^m\}.$$

Output: selection set $\hat{S} := \{j \in [p] : \hat{\pi}_j \geq \eta\}$.

using $\eta = 0.5$. Some practitioners may however wish to employ a data-driven approach that maximizes the power/number of rejections. However, caution needs to be exercised if one aims to maintain Type I error control. For example, if we scan through a list of η values and select that with which derandomized knockoffs yields the largest number of selected features, we face the problem of double dipping and the risk is to loose Type I error control. A possible remedy is to use sample splitting (i) use a fraction of the data to determine the optimal η—that giving the largest number of rejections; (ii) apply derandomized knockoffs on the remaining samples with the selected η. Naturally, sample splitting often leads to a power loss, which can be alleviated by techniques such as data recycling (see, e.g., Barber and Candès (2019)).

At each iteration, the statistician is allowed to use a different knockoffs generating distribution as well as a different test statistic. That said, consider the scenario in which each copy $\tilde{X}^m$ is identically distributed and that the same test statistics are used (e.g., LCD). Then the law of large numbers implies that each $\hat{\pi}_j$ converges to $P(j \in \hat{S}^1 \mid X, \boldsymbol{Y})$ as M increases to infinity.

We thus see that in the limit of an infinite number of knockoff copies, the procedure is fully derandomized since the outcome is determined by X and $\boldsymbol{Y}$.

2.3. Reduced Variability

When working on a specific dataset or application, researchers are typically interested in the fraction of false positives (FDP) and/or the number V of false positives. Even in the case where one employs a procedure controlling the false discovery rate (this is the expected value of the FDP) or the PFER (this is the expected value of V), one would always prefer a method which has lower variability in FDP and/or V so that FDP and V are close to their expectations. The reason is that on any given dataset, we would like to be sure that the fraction and/or number of false positives are not too high. The variability of these random variables and others, such as whether a specific variable is selected or not, originates from different sources. First, it comes from the draw we got to see, for example, the sample X, $\boldsymbol{Y}$. In the case of knockoffs, it also comes from the random nature of the algorithm itself producing $\tilde{X}$. Clearly, the derandomization scheme removes the second source of variability, which is a desirable trait.

2.4. Connections to Prior Literature

Certainly, aggregating results from multiple runs of a random procedure is not a new idea. A line of work develops methods to represent the consensus over multiple runs of one algorithm, with the aim of reducing its sensitivity to the initialization or the randomness inherent in the algorithm; see, for example, Bhattacharjee et al. (2001) and Monti et al. (2003). Another line of prior work seeks to combine multiple different learning algorithms to improve performance, which is often referred to as *ensemble learning*. A few examples would include Strehl and Ghosh (2002), Rokach (2010), and Polikar (2012). Yet, most of these methods are neither directly applicable to the knockoffs framework nor come with a finite sample Type I error control.

2.5. Comparisons With Stability Selection

It is time to expand on the similarities and differences with stability selection. To facilitate this discussion, it is helpful to briefly motivate and describe the stability selection algorithm. We are given data $(X, \boldsymbol{Y})$ and would like to find important variables by reporting those variables which have a nonzero Lasso coefficient. How confident are we that our selections will replicate in the sense that we would get a similar result on an independent dataset? How do we make sure that the nonzero coefficients are not merely due to chance? Stability selection addresses this issue by sampling repeatedly $\lfloor n/2 \rfloor$ observations without replacement from the original dataset as if they were independent draws from the population.[1] Important variables are then determined based on their selection frequencies just as in Equations (3) and (4). Despite evident similarities, there are major differences with derandomized knockoffs.

- First, stability selection introduces randomness via data splits and it is precisely this extra source of randomness which permits inference. In contrast, vanilla knockoffs natively provides valid inference and the aim of the derandomized procedure is simply to remove the randomness of the knockoffs.
- Second, while stability selection benefits from the bootstrap aggregating procedure preventing overfitting, it only operates on a random subset of the data at each step. This difference explains why our algorithm is particularly useful in the case where the samples size is comparable to the number of features we are assaying, since subsampling inevitably leads to a loss of power. We refer the reader to the numerical comparisons from Supplementary Section S5.5 that illustrate this point.
- Third, the theoretical guarantees for stability selection come with very strong assumptions—such as the exchangeability assumption of null statistics—which are nearly impossible to

[1] $\lfloor x \rfloor$ denotes the largest integer that is not greater than x, and $\lceil x \rceil$ denotes the smallest integer that is not less that x.

justify in practice. In contrast, our theoretical results hold under fairly mild assumptions.

2.6. A Representative Base Procedure

While our aggregating framework can be easily applied to a wide range of base procedures, the current article focuses on that proposed by Janson and Su (2016)—referred to as v-knockoffs throughout—which has been shown to control the PFER. Informally, suppose we wish to make at most v false discoveries over the long run. Then this base procedure (i) sorts the features based on the absolute value of their test statistic $|W_j|$;[2] (ii) it then examines the ordered features starting from the largest $|W_j|$ and selects those examined features with $W_j > 0$; (3) the procedure stops the first time it sees v features with negative values of W_j. For more details about v-knockoffs, we refer the readers to Section S1 (supplementary material).

3. Theoretical Guarantees: Controlling the PFER

We now tune derandomized knockoffs parameters to control the per family error rate. Formally, let $\mathcal{H}_0 \subseteq [p] := \{1, \ldots, p\}$ denote the set of null variables for which (1) is true, and consider a selection procedure producing a set of discoveries $\hat{S} \subseteq [p]$. Letting V be the number of false discoveries defined as

$$V := \#\{j : j \in \mathcal{H}_0 \cap \hat{S}\}, \tag{5}$$

the PFER is simply the expected number of false discoveries, PFER $= \mathbb{E}[V]$ (see, e.g., Dudoit and Van Der Laan (2007)).

Theorem 1. Consider derandomized knockoffs (Algorithm 1) with a base procedure obeying PFER $\leq v$ (e.g., v-knockoffs). If there exists a constant $\gamma > 0$, such that

$$\mathbb{P}(\mathcal{S}_j \geq \eta) \leq \gamma \, \mathbb{E}[\mathcal{S}_j], \tag{6}$$

holds for every $j \in \mathcal{H}_0$, then the PFER can be controlled as

$$\mathbb{E}[V] \leq \gamma v. \tag{7}$$

To prove Theorem 1, observe that

$$\begin{aligned} \mathbb{E}[V] &= \mathbb{E}\Big[\sum_{j \in \mathcal{H}_0} \mathbf{1}\{\mathcal{S}_j \geq \eta\}\Big] \\ &= \sum_{j \in \mathcal{H}_0} \mathbb{P}\big(\mathcal{S}_j \geq \eta\big) \leq \sum_{j \in \mathcal{H}_0} \gamma \, \mathbb{E}[\mathcal{S}_j] \\ &= \gamma \, \mathbb{E}[V_1] \leq \gamma v, \end{aligned} \tag{8}$$

where V_1 denotes the number of false discoveries in $\hat{S}^{[1]}$; the first inequality follows from Equation (6) and the second from the property of the base procedure.

In particular, with Markov's inequality, we immediately obtain the following proposition that provides an assumption-free bound for the PFER.

[2] In the case of the LCD, $W_j = |\beta_j| - |\tilde{\beta}_j|$, where β_j (resp. $\tilde{\beta}_j$) is the lasso coefficient estimate for the variable X_j (resp. $\tilde{X}_j$) when regressing $\boldsymbol{Y}$ on $\mathbf{X}$ and $\tilde{\mathbf{X}}$ jointly.

Proposition 1. Consider derandomized knockoffs (Algorithm 1) with a base procedure obeying PFER $\leq v$ (e.g., v-knockoffs). We always have

$$\mathbb{E}[V] \leq v/\eta. \tag{9}$$

Returning to the comparison with stability selection, we note that PFER control holds regardless of the choice of M and without any assumption on the exchangeability of the selected variables. In Section S4 (supplementary material), we discuss improving the PFER bound (without imposing additional assumptions).

3.1. Guarantees Under Mild Assumptions

Set $\eta = 1/2$. In this case, we have seen that Equation (6) holds with $\gamma = 2$. This is however too conservative in all the cases we have ever encountered. In fact, we will be surprised to ever see an example where the ratio $\mathbb{P}(\mathcal{S}_j \geq 1/2)/\mathbb{E}[\mathcal{S}_j]$ exceeds one. Consider for instance the setting from Figure 2. In this case, Supplementary Figure S19 plots the realized ratios $\mathbb{P}(\mathcal{S}_j \geq 1/2)/\mathbb{E}[\mathcal{S}_j]$ for each null variable, and we can observe that all the ratios are below one.

Turning to formal statements, Proposition 2 examines assumptions under which pairs (η, γ) obey condition (6). The idea is very similar to that of Shah and Samworth (2013), where a general bound is first established and then followed by a sharpened version holding under constraints on the shape of the distribution $\mathcal{S}_j$.

Definition 1. Let M be a positive integer and X a random variable supported on $\{0, 1/M, \ldots, 1\}$. The probability mass function (pmf) of X is said to be monotonically nonincreasing if for any $m_1 \leq m_2 \in \{0, 1, \ldots, M\}$,

$$\mathbb{P}(X = m_1/M) \geq \mathbb{P}(X = m_2/M).$$

Proposition 2. 1. Assume the pmf of $\mathcal{S}_j$ is monotonically nonincreasing for each $j \in \mathcal{H}_0$, then condition (6) holds with γ being the optimal value of the following linear program (LP):

$$\begin{aligned} \text{maximize} \quad & \sum_{m \geq M\eta} y_m \\ \text{subject to} \quad & y_m \geq 0, \\ & y_{m-1} \geq y_m, \; m \in [M], \\ & \sum_{m=0}^{M} y_m \, m/M = 1. \end{aligned} \tag{10}$$

2. Assume that for any $j \in \mathcal{H}_0$,

$$\begin{aligned} &\int_0^{\eta} \mathbb{P}(\mathcal{S}_j \in [\eta - u, \eta)) \mathrm{d}u \\ &\quad \geq \int_0^{\eta - 1/M} \mathbb{P}(\mathcal{S}_j \in [\eta, \eta + u)) \mathrm{d}u, \end{aligned} \tag{11}$$

then condition (6) holds with γ being the optimal value of the following linear program:

$$\begin{aligned} \text{maximize} \quad & \sum_{m \geq M\eta} y_m \\ \text{subject to} \quad & y_m \geq 0, \; m \in \{0, 1, 2, \ldots, M\}, \\ & \sum_{m=0}^{M} y_m \, m/M = 1, \\ & \sum_{m=0}^{M} y_m \geq 2, \\ & \sum_{m=1}^{\lceil \eta M \rceil - 1} m y_m \geq \sum_{m=0}^{\lfloor 2\eta M - 1 - \lceil \eta M \rceil \rfloor} (2\eta M - 1 - \lceil \eta M \rceil - m) \, y_{\lceil \eta M \rceil + m}. \end{aligned} \tag{12}$$

3. As a special case of (b), assume that for any $j \in H_0$, the pmf of 5_j is unimodal where the mode is less than or equal to η and $\mathsf{P}(5_j = 0) \geq \mathsf{P}(5_j = \mathrm{d}\eta M\mathrm{e}/M)$. Then Equation (11) holds and, therefore, Equation (6) holds with γ being the optimal value of Equation (12).
4. Suppose there exists a constant $\beta \in [0, 1]$ such that for any $j \in H_0$, the pmf of 5_j satisfies

$$\mathsf{P}(5_j = m/M) \leq \beta \cdot \mathsf{P}(5_j = (m-1)/M), \quad \text{for } m \in [M]. \tag{13}$$

Then condition (6) holds with γ being the optimal value of the LP,

$$\begin{aligned} \text{maximize} \quad & \sum_{m \geq M\eta} y_m \\ \text{subject to} \quad & y_m \geq 0, \\ & \beta y_{m-1} \geq y_m, \ m \in [M], \\ & \sum_{m=0}^{M} y_m \, m/M = 1. \end{aligned} \tag{14}$$

For illustration, Figure S20 (supplementary material) plots the optimal value of Equation (10) versus M with $\eta = 0.501$, 0.751 and 1, respectively. Taking $M = 31$ and $\eta = 0.501$ for example, we see that Equation (6) holds with $\gamma = 1$. Also, and this is important for later, (6) holds with $\gamma = 1$ for $M = 31$, $\eta = 1/2$.

The proof of Proposition 2 is deferred to Supplementary Section S2.1 and we pause here to parse the claims. The monotonicity assumption in part (a) states that the chance that a null variable gets selected 50 times is at most that it gets selected 49 times, which is at most that it gets selected 48 times and so on. (When we say chance, recall that the probability is taken over $\mathsf{X}, \boldsymbol{Y}, \mathsf{X}^1, \ldots, \mathsf{X}^M$.) In part (b), Equation (11) is a relaxed version of the monotonicity condition. To be sure, if the pmf of 5_j is monotonically non-increasing, then Equation (11) holds. Setting $F_-(x) := \mathsf{P}(X < x)$, condition (11) says this: the area between the two curves $y = F_-(x)$ and $y = F_-(\eta)$ (the latter does not vary with x) over the interval $[0, \eta]$—the blue area in Figure 1—is larger than the area between the same two curved curves over $[\eta, 2\eta - 1/M]$—the red area in Figure 1. In other words, the pmf of 5_j is skewed toward the left as illustrated in Figure 1(b). Part (d) shows that we can sharpen the bound (as illustrated in the supplementary material, Figure S21) if the pmf of 5_j decays at a faster rate—$\mathsf{P}(5_j = m/M) \leq \beta \mathsf{P}(5_j = (m-1)/M)$—where the smaller β, the faster the decay. In this article, we mostly consider $\beta = 1$ (the weakest possible condition), which just says that the pmf is monotonically nonincreasing.

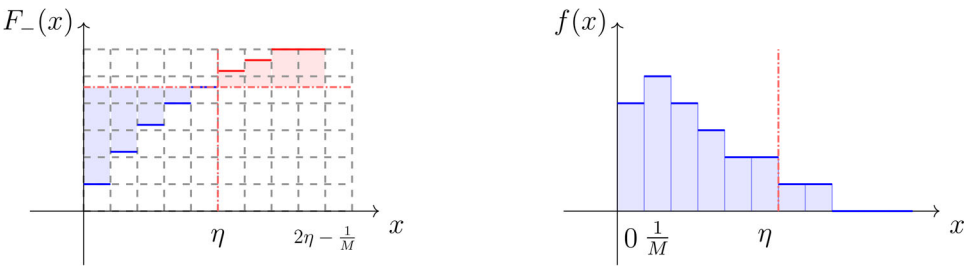
Proposition 1 is stated in terms of a *fixed* (finite) M. To understand the role of M, consider again the scenario where the knockoff copies $\tilde{\mathsf{X}}^m$ are sampled from the same distribution and the same feature importance statistic is applied in each realization. As $M \to \infty$, 5_j converges to $\mathsf{P}(j \in S^1 \mid \mathsf{X}, \boldsymbol{Y})$ by the strong law of large numbers. Consequently,

$$\begin{aligned} \mathrm{PFER} &= \mathsf{E}\left[\sum_{j \in H_0} \mathbf{1}\{5_j \geq \eta\}\right] \\ &\xrightarrow{M \to \infty} \mathsf{E}\left[\sum_{j \in H_0} \mathbf{1}\left\{\mathsf{P}(j \in S^1 \mid \mathsf{X}, \boldsymbol{Y}) \geq \eta\right\}\right], \\ \mathrm{Power} &= \frac{1}{|H_1|}\mathsf{E}\left[\sum_{j \in H_1} \mathbf{1}\{5_j \geq \eta\}\right] \\ &\xrightarrow{M \to \infty} \frac{1}{|H_1|}\mathsf{E}\left[\sum_{j \in H_1} \mathbf{1}\{\mathsf{P}(j \in S^1 \mid \mathsf{X}, \boldsymbol{Y}) \geq \eta\}\right]. \end{aligned}$$

In other words, as M increases to infinity, both the PFER and the power of derandomized knockoffs converge to fixed quantities that do not depend on M. For every f ixed$M$, the optimization problem in Proposition 1 aims at f inding the tightest PFER (with this M). A tighter PFER bound yields a more liberal choice of η and hence a higher power. To summarize, when M is suficiently large, further increasing M does not affect the power much. Thus, in practice, we recommend using a moderately large value of M such as 30–100—with the proviso that this is computationally reasonable.

3.2. Numerical Evaluation of the "Derandomization" Effect

To illustrate the effect of "derandomization," we compare derandomized knockoffs with vanilla knockoffs in a small-scale, a large-scale and a high-dimensional simulation study. The results of the small-scale experiment are presented here and those of the large-scale and high-dimensional experiments can be found in Sections S5.5.1 and S5.2 (supplementary material); additional simulations comparing derandomized knockoffs with alternative methods are provided in Section S5.5 (supplementary material). Our method is implemented in the R `derandomKnock` package, available at *https://github.com/zhimeir/derandomKnock*; code to reproduce all the numerical results from this article can be found at *https://github.com/zhimeir/derandomized_knockoffs_paper*. We evaluate the difference in the Type I error, the power and the stability of the selection set. Throughout this section, we set $\eta = 0.5$ and



(a) $F_-(x) = \mathbb{P}(X < x)$. (b) Probability mass function.

Figure 1. An example of a distribution obeying (11).

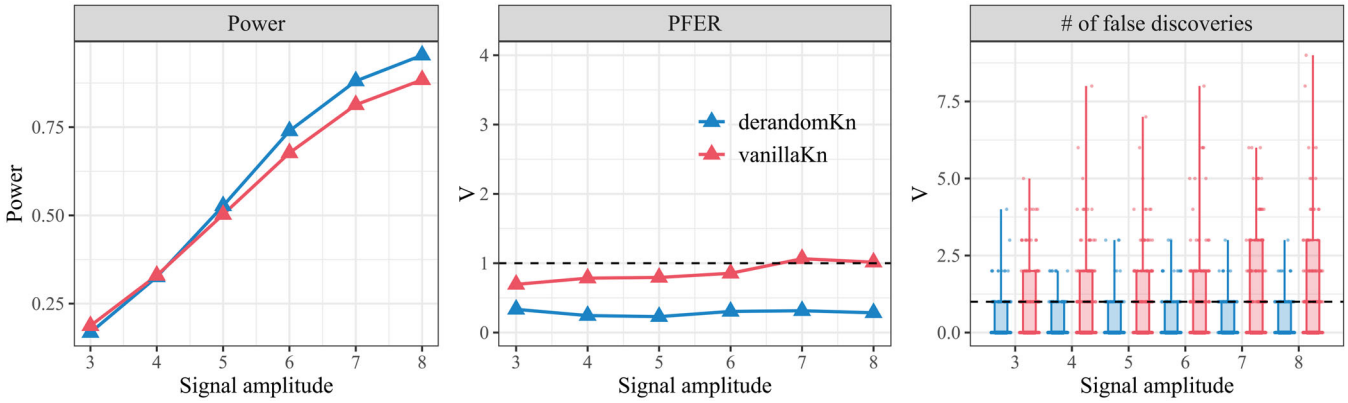


Figure 2. Performance of derandomized and vanilla knockoffs in the small-scale study. Here, $n = 200$, $p = 100$, $\boldsymbol{X} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \Sigma)$ with $\Sigma_{ij} = 0.6^{|i-j|}$, and $Y \mid \boldsymbol{X}$ is generated from a linear model with 30 nonzero coeficients. Each nonzero coeficient β takes value $\pm A/\sqrt{n}$ where the signal amplitude A ranges in $\{3, 4, \ldots, 8\}$ and the sign is determined by i.i.d. coin flips. The locations of the nonzero signal are randomly chosen from $[p]$. We show the averaged results over 200 trials. The parameter β is fixed across trials so that the distribution of $(\boldsymbol{X}, Y)$ does not vary. The dashed black line indicates the target PFER level $v = 1$. In the boxplot, the box is drawn from the 10th quantile to the 90th quantile; the whiskers represent the maximum and the minimum of the data; each jittered dot represents a raw data points outside of the [10, 90]th percentile range.

$M = 31$, which according to Proposition 2 yields $\mathsf{E}[V] \le v$ (recall that v is the nominal level of the base procedure) under the monotonicity assumption. Here, Y is generated from a linear model conditional on the feature vector $\boldsymbol{X}$, namely,

$$Y \mid X_1, \ldots, X_p \overset{\text{ind.}}{\sim} \mathcal{N}(\beta_1 X_1 + \cdots + \beta_p X_p, 1). \tag{15}$$

As for the covariates, $\boldsymbol{X}$ is drawn from a multivariate Gaussian distribution with parameters to be specified below. We remark that under this model, testing conditional independence is the same as testing whether $\beta_j = 0$.

Figure 2 compares the performance of derandomized and vanilla knockoffs in the small-scale study. The construction of knockoffs in this study is based on a version suggested by Spector and Janson (2020), and we use the LCD statistic to tease the signal and noise apart. We can see that both procedures control the PFER, while the power of derandomized knockoffs is slightly better than that of vanilla knockoffs. The boxplot shows that derandomization significantly decreases the *marginal* selection variability as claimed earlier (we additionally provide the frequencies of the number of false discoveries resulting from both methods in Table S1, supplementary material).

PFER control is theoretically guaranteed with our parameter choices since the ratio between $\mathsf{P}(5_j \ge 1/2)$ and $\mathsf{E}[5_j]$ is below one for all null variables j, as mentioned earlier (see supplementary material, Figure S19). A different way to establish validity is to check the monotonicity condition from Proposition 2 (which in turn implies that none of the ratios exceed one). We show in Figure S22 (supplementary material) the *pooled* histograms of all (nonzero) null 5_j's; the nonincreasing property of the pooled distributions is clear.

4. Theoretical Guarantees: Controlling the *k*-FWER

Another widely used Type I error measure is the *k family-wise error rate* (k-FWER): defined as the probability of making at least k false discoveries, $k\text{-FWER} = \mathsf{P}(V \ge k)$. Dating back to Bonferroni (Dunn 1961) and Holm (1979), many procedures guaranteeing k-FWER control have been proposed. Most operate on p-values and many require various assumptions on the dependence structure between these p-values (see, e.g., Karlin and Rinott 1980; Hochberg 1988; Benjamini and Yekutieli 2001; Romano et al. 2010). We refer the readers to Guo et al. (2014), Duan, Ramdas, and Wasserman (2020), and the references therein for a survey of these methods.

We now demonstrate how to tune the parameters for derandomized knockoffs to control the k-FWER. Our exposition parallels that from the previous section.

Theorem 2. Let V be the number of false discoveries after applying derandomized knockoffs (Algorithm 1) with a base procedure obeying PFER $\le v$. Suppose condition (6) holds and that for each $k \ge 1$,

$$\mathsf{P}(V \ge k) \le \frac{\rho\, \mathsf{E}[V]}{k}. \tag{16}$$

Then the k-FWER is controlled via

$$\mathsf{P}(V \ge k) \le \frac{\rho \gamma v}{k}. \tag{17}$$

With Markov's inequality, an immediate consequence of Theorem 2 is the following proposition.

Proposition 3. Let V be the number of false discoveries after applying derandomized knockoffs (Algorithm 1) with a base procedure obeying PFER $\le v$. We always have

$$\mathsf{P}(V \ge k) \le v/(k\eta).$$

Remark 1. To control the k-FWER at level α via Equation (17), one has the freedom of selecting the parameter v. In particular, it sufices to select v with $v = k\alpha/(\rho\gamma)$.

The proof of this result is straightforward since we have

$$\mathsf{P}(V \ge k) \le \frac{\rho\, \mathsf{E}[V]}{k} \le \frac{\rho \gamma v}{k},$$

where the last inequality follows from Theorem 1.

Set $h(x) := x$ and let $Z \sim \mathrm{NB}(v, 1/2)$, where $\mathrm{NB}(m, q)$ denotes a negative binomial random variable, which counts the number of successes before the mth failure in a sequence of

independent Bernoulli trials with success probability q. With this, the right-hand side of Equation (17) can be expressed as $\rho\gamma\,\mathsf{E}[h(Z)]/k$ (by simply observing that $\mathsf{E}[Z] = v$). This leads to the following extension:

Corollary 1. Let $h : \mathsf{R} \to \mathsf{R}$ be a convex, nonnegative and nondecreasing function. In the setting of Theorem 2, suppose that

$$\mathsf{P}(V \geq k) \leq \frac{\rho\,\mathsf{E}[h(V)]}{h(k)}. \tag{18}$$

Then the k-FWER obeys

$$\mathsf{P}(V \geq k) \leq \frac{\rho\,\mathsf{E}[h(Z/\eta)]}{h(k)}, \qquad Z \sim \mathrm{NB}(v, 1/2). \tag{19}$$

In particular, Markov's inequality shows that Equation (19) always holds with $\rho = 1$.

The proof of Corollary 1 is deferred to Section S2.2 (supplementary material).

4.1. Guarantees Under Mild Assumptions

While Equation (16) holds with $\rho = 1$, we observe in simulations that this value is often quite conservative and we give an example where Equation (16) holds with $\rho = 1/2$. The proof and an extension of the proposition below are given in Section S2.3 (supplementary material).

Proposition 4. In the setting of Theorem 2, suppose the pmf of V is skewed to the left of k in the sense that

$$\sum_{u=1}^{k-1} \mathsf{P}(V = k-u) \geq \sum_{u=1}^{k} \mathsf{P}(V = k+u) \tag{20}$$

(observe the similarity with Equation (11)). Then condition (16) holds with $\rho = 1/2$.

In applications, k and α are supplied and we provide in Section S7.1 (supplementary material) some guidance on the selection of v and η to control the k-FWER at level α.

4.2. Numerical Evaluation of the Derandomization Effect

We perform three numerical experiments to gauge the performance of derandomized knockoffs. Additional simulations comparing derandomized knockoffs with alternative methods can be found in Section S5.5.2 (supplementary material). In this study, the response Y is sampled from a logistic model

$$Y \mid X_1, \ldots, X_p \sim \mathrm{Bern}\left(\frac{\exp(\beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_1 X_1 + \cdots + \beta_p X_p)}\right) \tag{21}$$

and $\boldsymbol{X}$ is drawn from a multivariate Gaussian distribution with parameters to be specified later on. As in Section 3.2, the vector of regression coefficients is sparse so that most of the hypotheses are actually null; under this model, testing conditional independence is the same as testing whether $\beta_j = 0$.

We evaluate derandomized knockoffs on a small-scale, a large-scale and a high-dimensional dataset. The results on the small-scale data are shown here and those on the large-scale and the high-dimensional data are deferred to Section S5.3 and S5.4 (supplementary material). In the small-scale study, the knockoff construction is the same as that from Section 3.2, and the LCD statistic is used as our importance statistic. $M = 30$ knockoff copies are generated in each run and the selection threshold is $\eta = 0.81$. Under the monotonicity constraint, the value of (10) is $\gamma = 0.39$. With $v = 1$, the PFER is thus controlled at level 0.39. Applying Theorem 2 with $\rho = 1/2$, we see that 2-FWER is controlled at the level 0.1. Figure 3 displays the results of the small-scale experiment, where derandomized and vanilla knockoffs obey 2-FWER ≤ 0.1. As before, the boxplot shows that derandomized knockoffs exhibits less marginal randomness than vanilla knockoffs. At the same time, we can clearly see a substantial power gain.

We empirically verify the monotonicity assumption and the skewness property (20) by plotting the histograms of null 5_j's and false discoveries V, respectively, in Figures S27 and S28 (supplementary material). Under the monotonicity assumption, we expect to see the ratios obeying $\mathsf{P}(5_j \geq 0.81)/\mathsf{E}[5_j] \leq 0.39$, which is indeed the case as shown in Supplementary Figure S26.

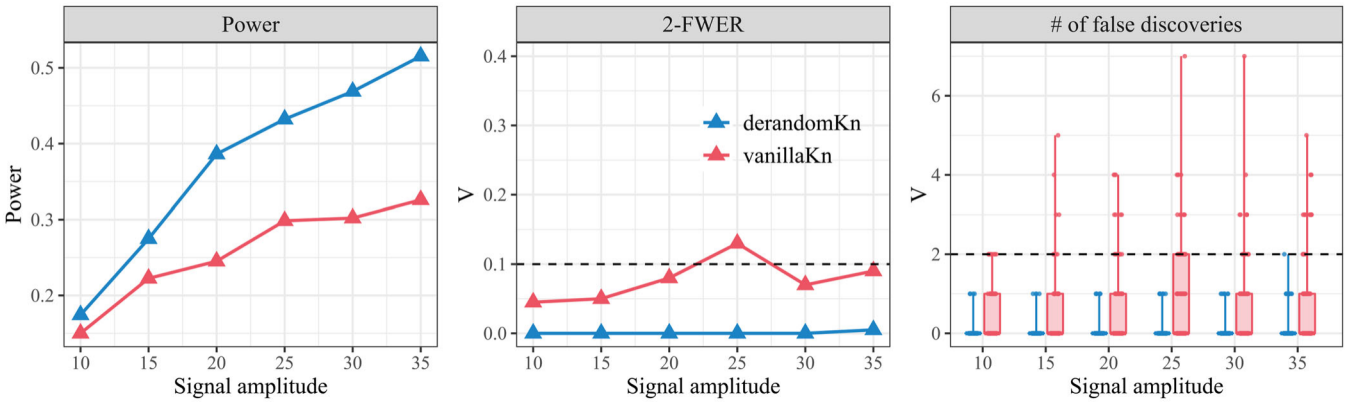


Figure 3. Performance of derandomized knockoffs ($\eta = 0.81$ and $v = 1$) and vanilla knockoffs. The target 2-FWER level is 0.1. In this setting, $n = 300$ and $p = 50$, $X \sim N(0, S)$ with $S_{ij} = 0.5^{|i-j|}$. $Y \mid X$ is sampled from a logistic model (21) with 30 nonzero entries in β. These nonzero entries take values $\pm A/\sqrt{n}$, where the signal amplitude A ranges in $\{10, 15, \ldots, 35\}$ and the sign is determined by iid coin flips. The setting is otherwise the same as in Figure 2. We indicate the target 2-FWER level $\alpha = 0.1$ and PFER $= 2$ with a dashed line. Each point in the first two panels represents an average over 200 replications. The construction of boxplots is as in Figure 2. Exact frequencies of the number of false discoveries are provided in Supplementary Table S4.

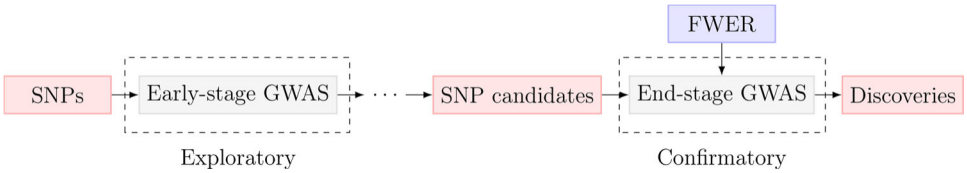


Figure 4. A typical workflow of multistage GWAS.

5. Application to Multistage GWAS

5.1. Background

The main goal of GWAS is to detect single-nucleotide polymorphisms (SNPs) associated with certain phenotypes. The task is commonly carried out in multiple stages; see, for example, Kote-Jarai et al. (2011),Thomas et al. (2009), and Lambert et al. (2013). The purpose of the early stages is often exploratory so that researchers tend to consider more liberal Type I error criteria (such as FDR) to allow for the inclusion of more candidates. The end-stage study is, in contrast, confirmatory, thus asking for a more stringent Type I error criterion (such as FWER). Informally, we can say that the early stage study narrows down the choices to a subset of "candidate SNPs," whereas the end-stage study pins down the final discoveries. Figure 4 provides a pictorial description of a typical multistage GWAS workflow. Here, we would like to apply derandomized knockoffs to the end stage, paving the way to reliable and stable decision making.

5.2. Answering the Same Question in Stages

We pause to discuss challenges associated with multistage studies. Although our methods apply regardless of the relationship between a phenotype Y and genetic variants $X_1, \ldots, X_p$, and always yield Type I error control, it may simplify the discussion to consider a standard linear model relating the quantitative Y to $\boldsymbol{X}$ to bring the reader onto familiar grounds (recall this is purely hypothetical). Consider a geneticist who has genotyped a number of sites. She wants to know whether the coef icient β_j associated with the variant X_j vanishes or not. Suppose now that in a first stage—for example, after analyzing the results of the first study—she thins out the list of possibly interesting variants, those for which she suspects β_j may not be zero. In a later confirmatory study, we want her to determine whether the coef icients of the screened variables in a model that still includes all the variants $X_1, \ldots, X_p$ she was originally interested vanish or not. It might be tempting to test in the second stage whether coeficients vanish in the *reduced model* only including those variables that passed screening. However, note that this would lead to test hypotheses that are dif ferent from those we started with, not merely a subset of them. To bring this point home, imagine that only one variable passed screening. Then in the second stage, this strategy would lead to test a *marginal* test of hypothesis, which is not what our geneticist wants (she wants a *conditional* test). This change of hypotheses so strongly influenced by the random results of the selection of the first stage seems hardly coherent with the goal of the scientific study. (For more discussion about full versus reduced model inference, we refer the reader to Wu et al. (2010), Wasserman and Roeder (2009), Barber and Candès (2019), Fan and Lv (2008), Voorman, Shojaie, and Witten (2014), Belloni, Chernozhukov, and Hansen (2014), and Ma (2017).

In light of this, this article proposes a pipeline for multistage GWAS that answers the *same* question throughout the stages. More specifically, from the very first stage, we are committed to testing the conditional independence hypothesis:

$$H_j : Y \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{-j},$$

where $\boldsymbol{X}_{-j}$ corresponds to *all* the SNPs except X_j. This means that if $\mathcal{C}$ is the candidate set selected by previous stages, we test H_j for each $j \in \mathcal{C}$ in the end stage. We do this by applying derandomized knockoffs, which controls Type I errors regardless of the procedures used in the previous stages. This is very different from existing approaches which switch the inferential target and would test whether a variable $j \in \mathcal{C}$ is significant in a model that only includes variables in $\mathcal{C}$ (Lee et al. 2013; Fithian, Sun, and Taylor 2014; Tibshirani et al. 2016; Tian, Lof tus, and Taylor 2018; Barber and Candès 2019).

5.3. End-stage GWAS of Prostate Cancer

We rehearse the pipeline for multistage GWAS and showcase the performance of derandomized knockoffs in Section S6 (supplementary material) on a synthetic dataset with *real genetic covariates*. Here, we present the results of applying our procedure to an end-stage GWAS of prostate cancer.

We take the meta-analysis conducted by Schumacher et al. (2018) as the early-stage study, and apply derandomized knockoffs on a dataset from the U.K. Biobank for a confirmatory analysis. The U.K. biobank dataset contains genetic information on 161K unrelated British male individuals and their disease status, that is, whether or not a participant has reported being diagnosed with prostate cancer.

After selecting p-values from Schumacher et al. (2018) below 10^{-3}, we end up with 4072 preselected SNPs. (The set of SNPs recorded in Schumacher et al. (2018) can be different from that in the U.K. Biobank dataset. Here, we only consider the intersection of the two sets.)

The next step is to partition a priori *all* the SNPs into clusters at a level of resolution 2%. The resulting average length of the clusters is 0.226 Mb. A cluster is called a candidate cluster if at least one of its SNPs is a candidate SNP. Ten runs of conditional group HMM knockoffs are constructed for the candidate clusters. We compute the group LCD statistics as in the synthetic example from Section S6 (supplementary material). Six additional covariates, namely age and the top five principal components of the genotypes are included in the knockoffs predictive model as follows: instead of using the phenotypes as the response, we use the residuals of the phenotypes *after* regressing out these six additional covariates. The inclusion

of these covariates allows us to account for the (remaining) population structure in the data, which increases the detection power. Finally, we apply derandomized knockoffs with target FWER level 0.1. The Supplementary material (Table S7) provides detailed information on the final list of clusters when the resolution is 2%.

We compare our findings with those from the existing literature. Since our discoveries are SNP clusters and different studies may contain different sets of SNPs, we cannot directly compare the results across different studies. Here we consider findings to be confirmed by another study if the latter reports a SNP whose position is within the genomic locus spanned by a cluster we discovered. With this, it turns out that *all* of our 8 findings are confirmed by other studies. Furthermore, 7 matches are exact in the sense that the leading SNP of a discovered cluster is reported significant in the literature. Specifically, clusters represented by rs12621278, rs1512268, rs6983267, rs7121039, rs10896449 and rs1859962 are replicated by Wang et al. (2015), which is a large GWAS conducted in the Asian population; rs1016343 is confirmed by Hui et al. (2014)—a study specifically investigating the associations of six SNPs including rs1016343 in a Chinese population; the association between rs7501939 and prostate cancer is in Elliott et al. (2010), which is a study focusing on the association of two SNPs including rs7501939 with several diseases.

What would happen if we were a little more liberal? To find out, we also run the derandomized knockoffs set to control the 3-FWER at the level 0.1: all the SNPs discovered earlier appear in the new discovery set. The more liberal procedure makes seven *additional* discoveries and the corresponding SNPs are listed in Table S8 (supplementary material).

6. Discussion

We proposed a framework for derandomized knockoffs inspired by stability selection. By exploiting multiple runs of the knockoffs algorithm, our method offers a more stable solution for selecting nonnull variables. Leveraging a base procedure with controlled PFER, we show how to achieve PFER and k-FWER control, these being error metrics perhaps more suitable than FDR for confirmatory stage studies (Tukey 1980; Heller 2011) as well as more resource-consuming applications such as end-stage GWAS (Sham and Purcell 2014; Meijer and Goeman 2016), clinical trials (Crouch, Dodd, and Proschan 2017) and neuro-imaging (Eklund, Nichols, and Knutsson 2016). Furthermore, we execute our methodology on a GWAS example and find that *all* our findings are confirmed by related studies.

7. Future Work

While the current article empirically demonstrates enhanced statistical power, it would be of interest to theoretically validate power gains, at least in some simple settings. We note that there has been a recent line of work devoted to the power analysis of the *original* knockoffs procedure using Lasso coeficient difference statistics (see, e.g., Weinstein, Barber, and Candès 2017; Liu and Rigollet 2019; Spector and Janson 2020; Weinstein et al. 2020); one limitation is that these works rely on strong assumptions such as Gaussian covariates, and a linear model holding exactly. Power calculations for the original knockoffs under general assumptions still remain largely limited and the challenge is further compounded by the complicated statistical dependence when different copies of knockoffs are considered. Therefore, characterizing the power of derandomized knockoffs requires developing a new set of tools.

Also, the machinery described here is general and can be applied to a variety of base procedures—even procedures that do not come with controlled PFER. One would, therefore, ask whether our theoretical framework can be adapted to accommodate such base procedures. Finally, perhaps the most natural question is whether our ideas can be adapted to the more liberal FDR criterion or related error rates such as the false discovery exceedence.

Supplementary materials

This supplementary file is organized as follows. First, we provide a recap of v-knockoffs in Section S1. The proofs of the main results are presented in Section S2, followed by the proofs of auxiliary lemmas in Section S3 that are crucially used to prove our main results. We discuss in Section S4 improving the assumption-free PFER bounds, and collect the additional simulations in Section S5 which contains the large-scale experiment from Section 3.2 in Section S5.1; the large-scale experiment from Section 4.2 in Section S5.3; and detailed comparisons with alternative methods in Section S5.5. The synthetic GWAS example with real genetic covariates is presented in Section S6. Finally, the technical details including parameter selections, knockoffs constructions, additional plots and tables are collected in Section S7.

Acknowledgments

The authors thank Lihua Lei for suggesting the assumption-free bound in Supplementary Section 4. The authors also thank to the PRACTICAL consortium, CRUK, BPC3, CAPS, PEGASUS for providing GWAS summary statistics (more information can be found at *http://practical.icr.ac.uk/blog/?page_id=8164*). The authors thank the Research Computing Center at Stanford University for computing resources, and the participants and investigators of the UK Biobank (application 27837).

Funding

Z. R. is supported by the Math + X award from the Simons Foundation, the JHU project no. 2003514594, the ARO project no. W911NF-17-1-0304, the NSF grant no. DMS 1712800 and the Discovery Innovation Fund for Biomedical Data Sciences. Y. W. is supported partially by the NSF DMS 2147546 / 2015447. E. C. is partially supported by NSF via grants nos. DMS 1712800 and DMS 1934578 and by the Ofice of Naval Research grant no. N00014-20-12157.

ORCID

Zhimei Ren http://orcid.org/0000-0002-2872-5842

References

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate Via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [948,949]

——— (2019), "A Knockoff Filter for High-Dimensional Selective Inference," *The Annals of Statistics*, 47, 2504–2537. [950,955]

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [955]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *Annals of Statistics*, 1165–1188. [953]

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. (2001), "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proceedings of the National Academy of Sciences*, 98, 13790–13795. [950]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [948]

——— (1999), "Using Adaptive Bagging to Debias Regressions," Technical Report 547, Berkeley, CA: Statistics Dept. UCB. [948]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [948]

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [948]

Candès, E., 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, 80, 551–577. [948,949]

Crouch, L. A., Dodd, L. E., and Proschan, M. A. (2017), "Controlling the Family-Wise Error Rate in Multi-Arm, Multi-Stage Trials," *Clinical Trials*, 14, 237–245. [956]

Duan, B., Ramdas, A., and Wasserman, L. (2020), "Familywise Error Rate Control by Interactive Unmasking," arXiv:2002.08545. [953]

Dudoit, S., and Van Der Laan, M. J. (2007), *Multiple Testing Procedures With Applications to Genomics*, Springer Science & Business Media, New York: Springer. [951]

Dunn, O. J. (1961), "Multiple Comparisons Among Means," *Journal of the American Statistical Association*, 56, 52–64. [953]

Efron, B., and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, 37, 36–48. [948]

Eklund, A., Nichols, T. E., and Knutsson, H. (2016), "Cluster Failure: Why FMRI Inferences for Spatial Extent Have Inflated False-Positive Rates," *Proceedings of the National Academy of Sciences*, 113, 7900–7905. [956]

Elliott, K. S., Zeggini, E., McCarthy, M. I., Gudmundsson, J., Sulem, P., Stacey, S. N., Thorlacius, S., Amundadottir, L., Grönberg, H., Xu, J., Gaborieau, V., Eeles, R. A., Neal, D. E., Donovan, J. L., Hamdy, F. C., Muir, K., Hwang, S. J., Spitz, M. R., Zanke, B., Carvajal-Carmona, L., Brown, K. M.; Australian Melanoma Family Study Investigators, Hayward, N. K., Macgregor, S., Tomlinson, I. P., Lemire, M., Amos, C. I., Murabito, J. M., Isaacs, W. B., Easton, D. F., Brennan, P.; PanScan Consortium, Barkardottir, R. B., Gudbjartsson, D. F., Rafnar, T., Hunter, D. J., Chanock, S. J., Stefansson, K., Ioannidis, J. P. (2010), "Evaluation of Association of hnf1b Variants With Diverse Cancers: Collaborative Analysis of Data From 19 Genome-Wide Association Studies," *Plos One*, 5, E10858. [956]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [955]

Fithian, W., Sun, D., and Taylor, J. (2014), "Optimal Inference After Model Selection," arXiv:1410.2597. [955]

Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N. I., Müller, M. L., Herman, T., Giladi, N., Kalinin, A., and Spino, C. (2018), "Model-Based and Model-Free Machine Learning Techniques For Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease," *Scientific Reports*, 8, 1–21. [948]

Guo, W., He, L., Sarkar, S. K., et al. (2014), Further results on controlling the false discovery proportion. *The Annals of Statistics*, 42, 1070–1101. [953]

Heller, R. (2011), "Multiple Testing for Exploratory Research by J. J. Goeman and A. Solari," *Statistical Science*, 26, 584–597. [956]

Hochberg, Y. (1988), "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 75(4):800–802. [953]

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 65–70. [953]

Hui, J., Xu, Y., Yang, K., Liu, M., Wei, D., Zhang, Y., Shi, X. H., Yang, F., Wang, N., and Wang, X., Liang, S., Chen, X., Sun, L., Zhu, X., Zhu, L., Yang, Y., Tang, L., Zhang, Y., Yang, Z., and Wang, J. (2014), "Study of Genetic Variants of 8q21 and 8q24 Associated With Prostate Cancer in Jing-Jin Residents in Northern China," *Clinical Laboratory*, 60, 645–652. [956]

Janson, L., and Su, W. (2016), "Familywise Error Rate Control Via Knockoffs," *Electronic Journal of Statistics*, 10, 960–975. [948,951]

Karlin, S., and Rinott, Y. (1980), "Classes of Orderings of Measures and Related Correlation Inequalities. I. Multivariate Totally Positive Distributions," *Journal of Multivariate Analysis*, 10, 467–498. [953]

Kote-Jarai, Z., Al Olama, A. A., Giles, G. G., Severi, G., Schleutker, J., Weischer, M., Campa, D., Riboli, E., Key, T., Gronberg, H., Hunter, D. J., Kraf t, P., Thun, M. J., Ingles, S., Chanock, S., Albanes, D., Hayes, R. B., Neal, D. E., Hamdy, F. C., Donovan, J. L., Pharoah, P., Schumacher, F., Henderson, B. E., Stanford, J. L., Ostrander, E. A., Sorensen, K. D., Dörk, T., Andriole, G., Dickinson, J. L., Cybulski, C., Lubinski, J., Spurdle, A., Clements, J. A., Chambers, S., Aitken, J., Gardiner, R. A., Thibodeau, S. N., Schaid, D., John, E. M., Maier, C., Vogel, W., Cooney, K. A., Park, J. Y., Cannon-Albright, L., Brenner, H., Habuchi, T., Zhang, H. W., Lu, Y. J., Kaneva, R., Muir, K., Benlloch, S., Leongamornlert, D. A., Saunders, E. J., Tymrakiewicz, M., Mahmud, N., Guy, M., O'Brien, L. T., Wilkinson, R. A., Hall, A. L., Sawyer, E. J., Dadaev, T., Morrison, J., Dearnaley, D. P., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Van As, N., Woodhouse, C. J., Thompson, A., Christmas, T., Ogden, C., Cooper, C. S., Lophatonanon, A., Southey, M. C., Hopper, J. L., English, D. R., Wahlfors, T., Tammela, T. L., Klarskov, P., Nordestgaard, B. G., Røder, M. A.,der, M. A., Tybjærg-Hansen, A., Bojesen, S. E., Travis, R., Canzian, F., Kaaks, R., Wiklund, F., Aly, M., Lindstrom, S., Diver, W. R., Gapstur, S., Stern, M. C., Corral, R., Virtamo, J., Cox, A., Haiman, C. A., Le Marchand, L., Fitzgerald, L., Kolb, S., Kwon, E. M., Karyadi, D. M., Orntof t, T. F., Borre, M., Meyer, A., Serth, J., Yeager, M., Berndt, S. I., Marthick, J. R., Patterson, B., Wokolorczyk, D., Batra, J., Lose, F., McDonnell, S. K., Joshi, A. D., Shahabi, A., Rinckleb, A. E., Ray, A., Sellers, T. A., Lin, H. Y., Stephenson, R. A., Farnham, J., Muller, H., Rothenbacher, D., Tsuchiya, N., Narita, S., Cao, G. W., Slavov, C., Mitev, V., Easton, D. F., Eeles, R. A.; UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology; UK ProtecT Study Collaborators, The Australian Prostate Cancer BioResource; PRACTICAL Consortium. (2011), "Seven Prostate Cancer Susceptibility Loci Identified by a MultiQ4 Stage Genome-Wide Association Study," *Nature Genetics*, 43, 785–791. [955]

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., et al. (2013), "Meta-Analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer's Disease," *Nature Genetics*, 45, 1452. [955]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2013), "Exact Post-Selection Inference With the Lasso," arXiv:1311.6238, 354:355. [955]

Liu, J., and Rigollet, P. (2019), "Power Analysis of Knockoff Filters for Correlated Designs," arXiv:1910.12428. [956]

Ma, C. (2017), "Semi-Penalized Inference With Direct False Discovery Rate Control for High-Dimensional AFT Model," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*. IEEE, pp. 48–52. [955]

Meijer, R. J., and Goeman, J. J. (2016), "Multiple Testing of Gene Sets From Gene Ontology: Possibilities and Pitfalls," *Briefings in Bioinformatics*, 17, 808–818. [956]

Meinshausen, N., and Bühlmann, P. (2010), "Stability Selection," *Journal of the Royal Statistical Society*, Series B, 72, 417–473. [948]

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003), "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning*, 52, 91–118. [950]

Polikar, R. (2012), "Ensemble Learning," in Zhang, C., Ma, Y., eds. *Ensemble Machine Learning*. Boston, MA: Springer, pp. 1–34. [950]

Ren, Z., and Candès, E. (2020), "Knockoffs With Side Information," arXiv:2001.07835. [949]

Rokach, L. (2010), "Ensemble-Based Classifiers," *Artificial Intelligence Review*, 33(1-2):1–39. [950]

Romano, J. P., and Wolf, M. (2010), "Balanced Control of Generalized Error Rates," *The Annals of Statistics*, 38, 598–633. [953]

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., Dadaev, T., Leongamornlert, D., Anokian, E., Cieza-Borrella, C., Goh, C., Brook, M. N., Sheng, X., Fachal, L., Dennis, J., Tyrer, J., Muir, K., Lophatananon, A., Stevens, V. L., Gapstur, S. M., Carter, B. D., Tangen, C. M., Goodman, P. J., Thompson, I. M. Jr, Batra, J., Chambers, S., Moya, L., Clements, J., Horvath, L., Tilley, W., Risbridger, G. P., Gronberg, H., Aly, M., Nordström, T., Pharoah, P., Pashayan, N., Schleutker, J., Tammela, T. L. J., Sipeky, C., Auvinen, A, Albanes, D., Weinstein, S., Wolk, A., Håkansson, N., West, C. M. L., Dunning, A. M., Burnet, N., Mucci, L. A., Giovannucci, E., Andriole, G. L., Cussenot. O., Cancel-Tassin, G., Koutros, S., Beane Freeman, L. E., Sorensen, K. D., Orntof t, T. F., Borre, M., Maehle, L., Grindedal, E. M., Neal, D. E., Donovan, J. L., Hamdy, F. C., Martin, R. M., Travis, R. C., Key, T. J., Hamilton, R. J., Fleshner, N. E., Finelli, A., Ingles, S. A., Stern, M. C., Rosenstein, B. S., Kerns, S. L., Ostrer, H., Lu, Y. J., Zhang, H. W., Feng, N., Mao, X., Guo, X., Wang, G., Sun, Z., Giles, G. G., Southey, M. C., MacInnis, R. J., FitzGerald, L. M., Kibel, A. S., Drake, B. F., Vega, A., Gómez-Caamaño, A., Szulkin, R., Eklund, M., Kogevinas, M., Llorca, J., Castaño-Vinyals, G., Penney, K. L., Stampfer, M., Park, J. Y., Sellers, T. A., Lin, H. Y., Stanford, J. L., Cybulski, C., Wokolorczyk, D., Lubinski, J., Ostrander, E. A., Geybels, M. S., Nordestgaard, B. G., Nielsen, S. F., Weischer, M., Bisbjerg, R., Røder, M. A.,der, M. A., Iversen, P., Brenner, H., Cuk, K., Holleczek, B., Maier, C., Luedeke, M., Schnoeller, T., Kim, J., Logothetis, C. J., John, E. M., Teixeira, M. R., Paulo, P., Cardoso, M., Neuhausen, S. L., Steele, L., Ding, Y. C., De Ruyck, K., De Meerleer, G., Ost, P., Razack, A., Lim, J., Teo, S. H., Lin, D. W., Newcomb, L. F., Lessel, D., Gamulin, M., Kulis, T., Kaneva, R., Usmani, N., Singhal, S., Slavov, C., Mitev, V., Parliament, M., Claessens, F., Joniau, S., Van den Broeck, T., Larkin, S., Townsend, P. A., Aukim-Hastie, C., Gago-Dominguez, M., Castelao, J. E., Martinez, M. E., Roobol, M. J., Jenster, G., van Schaik, R. H. N., Menegaux, F., Truong, T., Koudou, Y. A., Xu, J., Khaw, K. T., Cannon-Albright, L., Pandha, H., Michael, A., Thibodeau, S. N., McDonnell, S. K., Schaid, D. J., Lindstrom, S., Turman, C., Ma, J., Hunter, D. J., Riboli, E., Siddiq, A., Canzian, F., Kolonel, L. N., Le Marchand, L., Hoover, R. N., Machiela, M. J., Cui, Z., Kraf t, P., Amos, C. I., Conti, D. V., Easton, D. F., Wiklund, F., Chanock, S. J., Henderson, B. E., Kote-Jarai, Z., Haiman, C. A., Eeles, R. A.; Prof ile Study; Breast and Prostate Cancer Cohort Consortium (BPC3); PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; Cancer of the Prostate in Sweden (CAPS); Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS); Genetic Associations and Mechanisms in Oncology (GAME-ON)/Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium. (2018), "Association Analyses of More Than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci," *Nature Genetics*, 50, 928–936. [955]

Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2020), "Multi-Resolution Localization of Causal Variants Across the Genome," *Nature Communications*, 11, 1–10. [948]

Sesia, M., Sabatti, C., and Candès, E. J. (2019), "Gene Hunting With Hidden Markov Model Knockoffs," *Biometrika*, 106, 1–18. [948]

Shah, R. D., and Samworth, R. J. (2013), "Variable Selection With Error Control: Another Look at Stability Selection," *Journal of the Royal Statistical Society*, Series B, 75, 55–80. [948,951]

Sham, P. C., and Purcell, S. M. (2014), "Statistical Power and Significance Testing in Large-Scale Genetic Studies," *Nature Reviews Genetics*, 15, 335–346. [956]

Spector, A., and Janson, L. (2020), "Powerful Knockoffs Via Minimizing Reconstructability," arXiv:2011.14625. [953,956]

Srinivasan, A., Zhan, X., and Xue, L. (2019), "Compositional Knockoff Filter for High-Dimensional Regression Analysis of Microbiome Data," bioRxiv, pp. 851337. [948]

Strehl, A., and Ghosh, J. (2002), "Cluster Ensembles—a Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research*, 3, 583–617. [950]

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., Hankinson, S. E., Hutchinson, A., Wang, Z., Yu, K., Chatterjee, N., Garcia-Closas, M., Gonzalez-Bosquet, J., Prokunina-Olsson, L., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Diver, R., Prentice, R., Jackson, R., Kooperberg, C., Chlebowski, R., Lissowska, J., Peplonska, B., Brinton, L. A., Sigurdson, A., Doody, M., Bhatti, P., Alexander, B. H., Buring, J., Lee, I. M., Vatten, L. J., Hveem, K., Kumle, M., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F. Jr, Hoover, R. N., Chanock, S. J., Hunter, D. J. (2009), "A Multistage Genome-Wide Association Study in Breast Cancer Identifies Two New Risk Alleles at 1p11. 2 and 14q24. 1 (rad51l1)," *Nature Genetics*, 41, 579. [955]

Tian, X., Loftus, J. R., and Taylor, J. E. (2018), "Selective Inference With Unknown Variance Via the Square-Root Lasso," *Biometrika*, 105, 755–768. [955]

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620. [955]

Tukey, J. W. (1980), "We need Both Exploratory and Confirmatory," *The American Statistician*, 34, 23–25. [956]

Voorman, A., Shojaie, A., and Witten, D. (2014), "Inference in High Dimensions With the Penalized Score Test," arXiv:1401.2678. [955]

Wang, M., Takahashi, A., Liu, F., Ye, D., Ding, Q., Qin, C., Yin, C., Zhang, Z., Matsuda, K., Kubo, M., Na, R., Lin, X., Jiang, H., Ren, S., Sun, J., Zheng, S. L., Marchand, L. L., Isaacs, W. B., Mo, Z., Haiman, C. A., Sun, Y., Nakagawa, H., and Xu, J. (2015), "Large-Scale Association Analysis in Asians Identifies New Susceptibility Loci for Prostate Cancer," *Nature Communications*, 6, 1–7. [956]

Wasserman, L., and Roeder, K. (2009), "High Dimensional Variable Selection," *Annals of Statistics*, 37, 2178. [955]

Weinstein, A., Barber, R. F., and Candès, E. J. (2017), "A Power Analysis for Knockoffs Under Gaussian Designs," *IEEE Transactions on Information Theory*. [956]

Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020), "A Power Analysis for Knockoffs With the Lasso Coeficient-Difference Statistic," arXiv:2007.15346. [956]

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010), "Screen and Clean: A Tool for Identifying Interactions in Genome-Wide Association Studies," *Genetic Epidemiology: The Oficial Publication of the International Genetic Epidemiology Society*, 34, 275–285. [955]