# Whiteout: when do fixed-$X$ knockoffs fail?

Xiao Li and William Fithian

Department of Statistics, UC Berkeley

June 30, 2021

## Abstract

A core strength of knockoff methods is their virtually limitless customizability, allowing an analyst to exploit machine learning algorithms and domain knowledge without threatening the method's robust finite-sample false discovery rate control guarantee. While several previous works have investigated regimes where specific implementations of knockoffs are provably powerful, general negative results are more difficult to obtain for such a flexible method. In this work we recast the fixed-$X$ knockoff filter for the Gaussian linear model as a conditional post-selection inference method. It adds user-generated Gaussian noise to the ordinary least squares estimator $\widehat{\beta}$ to obtain a "whitened" estimator $\widetilde{\beta}$ with uncorrelated entries, and performs inference using $\mathrm{sgn}(\widetilde{\beta}_j)$ as the test statistic for $H_j : \ \beta_j = 0$. We prove equivalence between our whitening formulation and the more standard formulation involving negative control predictor variables, showing how the fixed-$X$ knockoffs framework can be used for multiple testing on any problem with (asymptotically) multivariate Gaussian parameter estimates. Relying on this perspective, we obtain the first negative results that universally upper-bound the power of all fixed-$X$ knockoff methods, without regard to choices made by the analyst. Our results show roughly that, if the leading eigenvalues of $\mathrm{Var}(\widehat{\beta})$ are large with dense leading eigenvectors, then there is no way to whiten $\widehat{\beta}$ without irreparably erasing nearly all of the signal, rendering $\mathrm{sgn}(\widetilde{\beta}_j)$ too uninformative for accurate inference. We give conditions under which the true positive rate (TPR) for any fixed-$X$ knockoff method must converge to zero even while the TPR of Bonferroni-corrected multiple testing tends to one, and we explore several examples illustrating this phenomenon.

## 1  Introduction

### 1.1  The knockoff filter in the Gaussian linear model

Knockoff methods are a flexible class of multiple testing procedures that operate by introducing a "negative control" or "knockoff" for each predictor variable in a supervised learning problem and then testing a learning algorithm's ability to distinguish each real predictor from its knockoff. The original method, the *fixed-X knockoff filter* (Barber and Candès, 2015), is a multiple testing method for the Gaussian linear model where we observe design matrix $X \in \mathbb{R}^{n \times d}$ and response

$$y = X\beta + \varepsilon, \quad \text{where } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n.$$

The parameters $\sigma^2 > 0$ and $\beta \in \mathbb{R}^d$ are unknown, and the goal is to test the hypothesis $H_j : \ \beta_j = 0$ against the two-sided alternative, for $j = 1, \ldots, d$. Following Barber and Candès (2015), we assume throughout that $X$ has full column rank with $2d \leq n$.

For testing $H_j$ individually, the usual two-sided $t$-test is uniformly most powerful unbiased. That test rejects for extreme values of the $t$-statistic $\widehat{\beta}_j / \sqrt{\widehat{\sigma}^2 \Sigma_{jj}}$, where $\Sigma = (X^{\mathsf{T}} X)^{-1}$ and $\widehat{\beta}$ and $\widehat{\sigma}^2$ are respectively the OLS estimator

$$\widehat{\beta} \ = \ (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} y \ \sim \ \mathcal{N}_d \left( \beta, \ \sigma^2 \Sigma \right)$$

and the residual variance $\widehat{\sigma}^2 = \|y - X\widehat{\beta}\|^2 / (n - d)$. Taken together, these two estimators are a complete sufficient statistic for the model. Let $p_j$ denote the $p$-value for the two-sided $t$-test on $H_j$.

For multiple testing, a classical approach would reject $H_j$ when $p_j$ is sufficiently small, after making an appropriate correction for the multiplicity of tests. If $R$ is the number of rejections and $V$ is the number of true null hypotheses rejected (false discoveries), Benjamini and Hochberg (1995) define the *false discovery proportion* as FDP $= V/\max\{R, 1\}$, and the *false discovery rate* (FDR) as its expectation. While the Benjamini–Hochberg (BH) procedure of Benjamini and Hochberg (1995) is not known to control the FDR in this problem unless the columns of $X$ are orthogonal, recent methods can directly adjust BH for the multivariate $t$ dependence, guaranteeing FDR control while performing similarly to BH (Fithian and Lei, 2020). The FDR criterion relaxes the more conservative *family-wise error rate* FWER $= \mathbb{P}(V \geq 1)$, the probability of making any false rejections, which we could control using the conservative Bonferroni correction that rejects $H_j$ when $p_j \leq \alpha/m$.[1]

The knockoff filter of Barber and Candès (2015) takes a radically different approach, bypassing the $t$-test $p$-values entirely. The method begins by augmenting the design matrix $X$ with a second matrix $\widetilde{X} \in \mathbb{R}^{n \times d}$ of negative control or *knockoff* variables, constructed to satisfy

$$\widetilde{X}^\mathsf{T} \widetilde{X} = X^\mathsf{T} X, \quad \text{and} \quad \widetilde{X}^\mathsf{T} X = X^\mathsf{T} X - D,$$

for some diagonal matrix $D \preceq 2X^\mathsf{T} X$, where $\preceq$ denotes the positive semidefinite ordering. As we will discuss, a larger value of $D_{jj}$ preserves more signal for the inference on $H_j$, but in general it is not possible to maximize all $D_{jj}$ simultaneously so tradeoffs must be made between variables.

Knockoffs then calculates so-called $W$-statistics $W_1, \ldots, W_d$ satisfying two properties:

**1. Sufficiency** $W_j$ depends on $[X \ \widetilde{X}]$ and $y$ only through $[X \ \widetilde{X}]^\mathsf{T}[X \ \widetilde{X}]$ and $[X \ \widetilde{X}]^\mathsf{T} y$, and

**2. Antisymmetry** Swapping any variable $X_j$ with its knockoff $\widetilde{X}_j$ would flip the sign of $W_j$ and leave every other $W_k$ fixed. That is, if $\mathrm{Swap}_j(\cdot)$ exchanges the $j$th and $(d+j)$th columns of an input matrix, then

$$W_k(\mathrm{Swap}_j([X \ \widetilde{X}]), y) = \begin{cases} -W_j([X \ \widetilde{X}], y) & \text{if } k = j \\ W_k([X \ \widetilde{X}], y) & \text{if } k \neq j \end{cases}.$$

The absolute values $|W| = (|W_1|, \ldots, |W_d|)$ determine a data-adaptive hypothesis ordering, with larger values assigned higher priority. Sufficiency and antisymmetry, along with the Gaussian modeling assumptions, ensure two key distributional properties: first, that $\mathrm{sgn}(W_1), \ldots, \mathrm{sgn}(W_d)$ are conditionally independent given $|W|$; and second, that if $\beta_j = 0$ then $\mathrm{sgn}(W_j) \sim \mathrm{Unif}\{-1, +1\}$ given $|W|$, unless $W_j = 0$.

After calculating $W$-statistics, the knockoff filter applies an ordered multiple testing method called *Selective SeqStep* (Barber and Candès, 2015) treating each $\mathrm{sgn}(W_j)$ as a "binary $p$-value" for $H_j$. To control FDR at level $\alpha$, the knockoff+[2] method rejects all hypotheses $H_j$ for which $W_j$ exceeds the adaptive threshold $\hat{t}$:

$$\hat{t} = \min\left\{t : \widehat{\mathrm{FDP}}_t^{\mathrm{kn}} \leq \alpha\right\}, \quad \text{where} \quad \widehat{\mathrm{FDP}}_t^{\mathrm{kn}} = \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\}}.$$

The fixed-$X$ knockoff filter is a highly versatile method that offers its users a multitude of choices in selecting both the knockoff matrix and the $W$-statistics. Spector and Janson (2020) show the importance of choosing $D$ well and discuss ramifications on the procedure's power, and various other works detail myriad ways to tailor the $W$-statistics using machine learning methods that exploit structural assumptions or other prior beliefs about the coefficients (see e.g., Dai and Barber, 2016; Katsevich and Sabatti, 2019; Chen et al., 2020; Dai et al., 2021). The method's customizability poses a major challenge if we hope to bound its power uniformly over the analyst's entire choice set, since well-informed analysts have ample opportunities to stack the deck in their own favor.

In light of this tremendous flexibility, it may come as a surprise to discover regimes where *no knockoffs* method — not even one designed with full knowledge of the true regression coefficients — can achieve nontrivial power, even while Bonferroni-corrected inference achieves near-perfect power. To begin to explain how knockoffs can go wrong, Section 1.2 formally recasts the knockoff filter as a conditional post-selection inference method built around a randomized estimator $\widetilde{\beta} = \widehat{\beta} + \omega$, where $\omega$ is user-generated Gaussian

---

[1]A more accurate FWER correction would apply the closure of the max-$t$ test: if $S$ is the maximal set for which $\|T_S\|_\infty$ is below its $1 - \alpha$ quantile under $H_S : \beta_S = 0$, then we can reject $H_j$ for $j \notin S$; see Marcus et al. (1976) and Hothorn et al. (2008) for more details.

[2]Barber and Candès (2015) also propose a version without the "1+" in the numerator, which controls a relaxed FDR criterion. All results in this paper pertain to the knockoff+ method, but they also apply with very minor modifications to the more liberal version.

noise in the style of Tian and Taylor (2018). Our interpretation builds on a conditioning argument in Barber and Candès (2019) and an observation in Sarkar and Tang (2021) that knockoffs constructs two independent estimators for $\beta$.

## 1.2   Knockoffs as conditional inference on a whitened estimator

We now give an alternative but equivalent account of the knockoff filter without $W$-statistics, without sufficiency and antisymmetry properties, and even without knockoff variables. We view the method instead as a conditional post-selection inference procedure, where the key step is to construct a *whitened estimator* $\widetilde{\beta}$ with diagonal covariance. If $\Delta \in \mathbb{R}^{d \times d}$ is any diagonal matrix with $\Delta \succeq \Sigma = (X^{\mathsf{T}}X)^{-1}$, let

$$\widetilde{\beta} = \widehat{\beta} + \omega \sim \mathcal{N}_d\left(\beta, \sigma^2 \Delta\right), \quad \text{where} \quad \omega \sim \mathcal{N}_d\left(0, \sigma^2(\Delta - \Sigma)\right) \tag{1}$$

is noise generated by the user, independently of $\widehat{\beta}$. We will see in Section 3 that even when $\sigma^2$ is unknown, $\omega$ can be "carved" out of $\widehat{\sigma}^2$ as long as $n \geq d + r$, where $r = \text{rank}(\Delta - \Sigma) \leq d$.

The independence of the coordinates of $\widetilde{\beta}$ is bought at the price of higher variance, since $\Delta_{jj} \geq \Sigma_{jj}$. However, this price can be recouped by an exploratory analysis using the statistic

$$\xi = \Sigma^{-1}\widehat{\beta} - \Delta^{-1}\widetilde{\beta} \sim \mathcal{N}_d\left(A\beta, \sigma^2 A\right), \quad \text{where} \quad A = \Sigma^{-1} - \Delta^{-1} \succeq 0. \tag{2}$$

$\xi$ carries the information lost by whitening since $\widehat{\beta} = \Sigma(\xi + \Delta^{-1}\widetilde{\beta})$, and it is independent of $\widetilde{\beta}$ since

$$\text{Cov}(\xi, \widetilde{\beta}) = \Sigma^{-1} \text{Cov}(\widehat{\beta}, \widehat{\beta} + \omega) - \Delta^{-1} \text{Cov}(\widetilde{\beta}, \widetilde{\beta}) = \sigma^2 I_d - \sigma^2 I_d = 0.$$

Thus, $\widehat{\beta}$ is effectively "split" into two independent Gaussian vectors, $\widetilde{\beta}$ and $\xi$, each holding part of the information about $\beta$. In terms of $\omega, \widetilde{\beta}$ and $\xi$, the fixed-$X$ knockoff filter can be equivalently defined as follows:

**Stage 1 (whitening).** For any diagonal $\Delta \succeq \Sigma$, generate noise $\omega \sim \mathcal{N}_d(0, \sigma^2(\Delta - \Sigma))$ independently of $\widehat{\beta}$. Calculate $\widetilde{\beta}$ and $\xi$ as in (1–2), but do not reveal any data to the analyst yet.

**Stage 2 (exploratory analysis).** Let the analyst observe $\xi$ and $|\widetilde{\beta}|$ and use them to order the $d$ hypotheses for Selective SeqStep as $H_{[1]}, \ldots, H_{[d]}$, where [1] indexes the first hypothesis in order and [$d$] the last. In addition, select a one-sided alternative for each $H_j$. Let $\psi_j = +1$ if the right-tailed alternative is selected, and $\psi_j = -1$ for the left-tailed alternative.

**Stage 3 (confirmatory analysis).** Using Selective SeqStep, test the hypotheses in order using $\text{sgn}(\widetilde{\beta}_j)$ as a conditional test statistic for $H_j$. The signs are conditionally independent given $\xi$ and $|\widetilde{\beta}|$, with

$$\text{logit } \mathbb{P}\left(\text{sgn}(\widetilde{\beta}_j) = +1 \mid \xi, |\widetilde{\beta}|\right) = \frac{2|\widetilde{\beta}_j|}{\sigma^2 \Delta_{jj}} \cdot \beta_j, \quad \text{where logit } p = \log\frac{p}{1-p}. \tag{3}$$

Let $\widetilde{p}_j$ denote the $p$-value for the conditional test of $H_j$ against a right- or left-tailed alternative according to $\psi_j$. Because $\text{sgn}(\widetilde{\beta}_j) \sim \text{Unif}\{-1, +1\}$ under $H_j$, and is stochastically increasing in $\beta_j$, we have $\widetilde{p}_j = 1/2$ when $\text{sgn}(\widetilde{\beta}_j) = \psi_j$, and $\widetilde{p}_j = 1$ when $\text{sgn}(\widetilde{\beta}_j) = -\psi_j$.

Using these $p$-values, the Selective SeqStep method rejects $H_{[j]}$ if $j \leq \widehat{k}$ and $\widetilde{p}_{[j]} = 1/2$, where

$$\widehat{k} = \max\left\{k : \widehat{\text{FDP}}_k^{\text{wh}} \leq \alpha\right\}, \quad \text{with} \quad \widehat{\text{FDP}}_k^{\text{wh}} = \frac{1 + \sum_{j=1}^k 1\{\widetilde{p}_{[j]} = 1\}}{\sum_{j=1}^k 1\{\widetilde{p}_{[j]} = 1/2\}}.$$

The conditional $p$-values $\widetilde{p}_1, \ldots, \widetilde{p}_d$ are independent given $\xi$ and $|\widetilde{\beta}|$, satisfying Selective SeqStep's requirements, so the method described above controls the FDR, both conditionally and marginally. In Section 2.1 we formally state and prove equivalence between this formulation and the usual formulation of knockoffs, building on a conditioning argument in the Supplement of Barber and Candès (2019).

The exploratory analysis is defined vaguely because the analyst can use $\xi$ and $|\widetilde{\beta}|$ however they like, provided they have not yet observed anything else about $y$ (while the sufficiency property of Barber and Candès (2015) also restricts how the analyst may use the design matrix $X$, we will see in Section 2.1 that no such rule is needed). It is in this unrestricted stage that knockoffs can freely exploit prior information and structural assumptions. Indeed, the analyst can even use an informal process that iterates between trying out several models, visualizing the (observable) data, consulting their intuitions or their colleagues, and so on.

When the exploratory stage goes well, the ordering is highly informative, effectively reducing the multiplicity in Stage 3 by focusing inferential power on the first few hypotheses. In many problems, a good exploratory analysis can more than compensate for the randomization and binarization of the confirmatory test statistics $\text{sgn}(\widetilde{\beta}_j)$, helping knockoffs to outperform less flexible methods like BH.

The whitening interpretation underscores not only the versatility of the fixed-$X$ knockoff filter, but also its broad applicability. Other than defining $\Sigma = (X^\mathsf{T} X)^{-1}$, the above method is defined without reference to any design matrix and would be equally applicable if we simply observed $\widehat{\beta} \sim \mathcal{N}_d(\beta, \Sigma)$ directly. In Section 3 we push this interpretation further and consider how to apply the whitening method in generic statistical models with asymptotically Gaussian estimators of $d$ parameters, a ubiquitous problem encompassing generalized least squares, generalized linear models, quantile regression, and many other settings.

## 1.3   Which problems are hard for knockoffs?

Having seen that the fixed-$X$ knockoff filter is an appealing and generally applicable method, it is natural to ask when we can expect it to outperform competitors like BH or Bonferroni. While this is a complicated question for a method as flexible as knockoffs, the whitening interpretation helps to identify an important vulnerability of the method: when the eigenstructure of $\Sigma$ is unfavorable, the price of whitening can be very steep, dooming the confirmatory analysis before the exploratory analysis even begins. One such example is the well-studied and seemingly innocuous *multiple comparisons to control* (MCC) problem (Dunnett, 1955).

**Example 1** (Multiple comparisons to control). *Assume that we observe a continuous response on $m$ units under each of $d$ treatments, along with $m_0$ units under a control condition, and estimate an additive treatment effect. Let $y_{j,i}$ denote the $i$th response under treatment $j = 0, 1, \ldots, d$, with $j = 0$ representing the control condition. We can pose this problem as a linear model by writing $y_{0,i} = \beta_0 + \varepsilon_{0,i}$ for the control group, and*

$$y_{j,i} = \beta_0 + \beta_j + \varepsilon_{j,i}$$

*for the $j$th treatment group, so that $\beta_j$ is the differential effect of treatment $j$ relative to control. If we compose a vector $y = (y_0^\mathsf{T}, y_1^\mathsf{T}, \cdots, y_d^\mathsf{T})^\mathsf{T}$ and model the errors as Gaussian with a common variance $\sigma^2$, we obtain a standard linear regression problem with intercept $\beta_0$ and can perform multiple testing on the hypotheses $H_j : \beta_j \neq 0$, for $j = 1, \ldots, d$.*

In the MCC problem, the OLS estimator of $\beta_j$ for $j \geq 1$ is

$$\widehat{\beta}_j \;=\; \frac{1}{m} \sum_{i=1}^{m} y_{j,i} \;-\; \frac{1}{m_0} \sum_{i=1}^{m_0} y_{0,i},$$

so that $\text{corr}(\widehat{\beta}_j, \widehat{\beta}_k) = \rho = \frac{m}{m_0 + m}$ for distinct $j, k > 0$. If we normalize the design matrix appropriately, then $\Sigma$ is an equicorrelated matrix with diagonal entries equal to 1 and off-diagonal entries equal to $\rho$. Its first eigenvalue is $\lambda_1 = 1 + \rho(d-1) \geq \rho d$, with eigenvector $u_1 = \mathbf{1}_d / \sqrt{d}$. As we will see in Section 2.3, the eigenstructure of $\Sigma$ makes it impossible to find $\Delta \succeq \Sigma$ without making most of the $\Delta_{jj}$ values enormous. Specifically, the $k$th smallest diagonal entry of any valid $\Delta$ matrix can be no smaller than $\rho k$.

Let $d_1$ denote the number of nonzero coefficients. As our measure of power, we use the *true positive rate* (TPR), defined as the expectation of the *true positive proportion* $\text{TPP} = (R - V)/d_1$, the fraction of non-nulls discovered. TPP and TPR are meaningless if $d_1 = 0$, so we assume $d_1 \geq 1$ in all power calculations.

In Section 2.4 we will show that in an MCC problem where the non-null proportion $d_1/d$ tends to a nonzero constant, no knockoff method can achieve a nontrivial TPR as $d \to \infty$ unless the nonzero coefficients grow roughly as $\sigma \sqrt{d}$; by contrast, the Bonferroni method has TPR $\to 1$ when the nonzero coefficients grow at the rate $\beta_0 = \sigma \sqrt{2 \log d}$. If all of the nonzero coefficients have size $\beta_0$ then, applying Theorem 2, the expected number of rejections of any fixed-$X$ knockoffs method at level $\alpha = 0.05$ is no larger than

$$\mathbb{E}\, R \;\leq\; \frac{13 \log d}{\rho} + 42. \tag{4}$$

If $d = 10^6$ and $\rho = 0.5$, and the coefficients are all at the Bonferroni threshold $5.25\sigma$, the upper bound evaluates to 387. This is a finite-sample result that applies even for a knockoff method designed by an analyst with full knowledge of the coefficient vector at every stage of the procedure, who can favor the non-null variables both by adding less noise to them and by ensuring they are always ordered first. We also show in Section 2.4 how to improve the bounds by simulation; in this example we can improve the bound from 387 to

84. Reducing $\rho$ from 0.5 to 0.05 only helps a little, increasing the simulation bound to 865. Asymptotically, unless $\rho$ shrinks at the rate $\frac{\log d}{d}$, any knockoffs method will be have TPR $\to 0$ for signals of size $\beta_0$.

By contrast, in a linear regression whose design matrix has positively equicorrelated columns, the situation is completely different. In that case, $\mathrm{Var}(\widehat{\beta})$ is *negatively* equicorrelated, so its leading eigenvalue is small and all of the $\Delta_{jj}$ values can be $O(1)$. Likewise, if the design matrix is i.i.d. Gaussian with $n \geq 2d$, then for large $d$ we expect the largest eigenvalue of $(\frac{1}{n}X^\mathsf{T}X)^{-1}$ to be no larger than roughly $(1 - \sqrt{1/2})^{-2} \approx 11.7$, per the Marchenko–Pastur distribution. Finally, if $\lambda_1$ is very large but $u_1$ is heavily concentrated on only a few variables, knockoffs may still perform well on the other variables. Section 2.3 gives explicit and efficiently computable lower bounds for all order statistics of $\Delta_{11}, \ldots, \Delta_{dd}$ as a function of a general matrix $\Sigma$.

The MCC setting is a prototypical example of a hard problem for knockoffs. More generally, if $\Sigma$ has large leading eigenvalues with dense eigenvectors then we must have $\Delta_{jj} \gg \Sigma_{jj}$ for most of the variables, rendering the test statistic $\mathrm{sgn}(\widetilde{\beta}_j)$ nearly uninformative about $\beta_j$ even when the coefficient is very large. In short, we cannot whiten the estimator without burying the signal under a blizzard of noise.

## 1.4 Related work

Several prior works foreshadow the whitening interpretation of Section 1.2. Most notably, an argument in the Supplement of Barber and Candès (2019) conditions on $(X + \widetilde{X})^\mathsf{T}y$ and $|(X - \widetilde{X})^\mathsf{T}y|$ to prove the knockoff filter controls the *directional FDR*, a more stringent version of FDR where the analyst must draw a conclusion about $\mathrm{sgn}(\beta_j)$ whenever they reject $H_j$. An analogous argument to theirs shows that the whitening method, viewed as a conditional post-selection inference procedure, controls the directional FDR as well. If we define the data-dependent one-sided null hypothesis $H_j^{\psi_j} : \psi_j \beta_j \leq 0$, then $\widetilde{p}_1, \ldots, \widetilde{p}_d$ are independent and valid conditional $p$-values for $H_1^{\psi_1}, \ldots, H_d^{\psi_d}$ given $\xi$ and $|\widetilde{\beta}|$. As a result, if we conclude $\mathrm{sgn}(\beta_j) = \psi_j$ whenever $H_j$ is rejected then we will control the directional FDR, both conditionally and marginally.

Sarkar and Tang (2021) also exploit the independence of $(X + \widetilde{X})^\mathsf{T}y$ and $(X - \widetilde{X})^\mathsf{T}y$ with a view toward developing new methodology. When $A$ is invertible, they view $D^{-1}(X - \widetilde{X})^\mathsf{T}y$ and $\frac{1}{2}A^{-1}(X + \widetilde{X})^\mathsf{T}y$ as two independent unbiased estimators of $\beta$, very similar to our data splitting interpretation. From this starting point they derive hybrid multiple testing procedures blending fixed-$X$ knockoffs with the BH method.

As a conditional post-selection inference method that uses a randomized data set for exploration, knockoffs is related to two other lines of work. In the statistics literature, the *data carving* method of Fithian et al. (2014) improves on the Lee et al. (2016) post-selection inference method by performing a selection algorithm such as the lasso (Tibshirani, 1996) only on a subset of the data points, effectively randomizing the selection. Tian and Taylor (2018) proposed explicitly randomizing the selection algorithm by adding Gaussian or other noise to the sufficient statistics used by the selection algorithm. Their work was inspired by a parallel literature on *adaptive data analysis* wherein randomized algorithms are used to answer many statistical queries about a database, one after the other, without allowing the analyst to "overfit" (Dwork et al., 2015b,a). More recently Zrnic and Jordan (2020) adapted a related approach based on algorithmic stability, also using randomization to preserve data for post-selection inference in a selected linear regression model.

Prior work on the knockoff filter's power has focused on positive results for specific implementations of the method. When studying the TPP-FDP tradeoff on the Lasso path, it is frequently assumed that a constant fraction of non-zero coefficients are sampled from a fixed distribution (Bayati and Montanari, 2011; Su et al., 2017). In addition, it is often assumed that $n/p \to \delta$ for some positive constant $\delta$. Under this intermediate-dimensional regime, the theory of approximate message passing (AMP) (Bayati and Montanari, 2011) may be used to derive the asymptotic power of knockoff. Weinstein et al. (2017) derives the asymptotic TPP-FDP tradeoff for a knockoff-inspired procedure that is only valid for i.i.d. covariates, while Weinstein et al. (2020) and Wang and Janson (2020) quantify the asymptotic power of knockoffs with the lasso coefficient difference test statistic. Liu and Rigollet (2019) studied the power of knockoffs under correlated design in the low dimensional setting, and showed that the knockoff filter has full asymptotic power when the precision matrix has vanishing diagonal entries. Going beyond the aforementioned linear sparsity assumptions, Fan et al. (2020) studied the power of an oracle knockoff filter where the covariance structure of the variables is known. They assumed that the coefficients are fixed and relatively large, and showed under certain regularity conditions that the oracle knockoff filter is consistent. Ke et al. (2020) analyzed the phase diagram of the SDP knockoff, but their results are restricted to block-equicorrelated correlation structure with block size 2.

Recently, Spector and Janson (2020) showed that the equicorrelated and SDP knockoff methods can be asymptotically powerless for equicorrelated Gaussian design matrices with correlation $\rho \geq 0.5$, and propose a new method for generating knockoff variables, called minimum variance-based reconstructability (MVR)

knockoffs, to resolve the issue. They showed that the TPR of the MVR knockoff converges to 1 under regularity conditions. The failure mode identified by Spector and Janson (2020) is very different from the one we discuss here, which cannot be fixed within the knockoffs framework as currently defined. Section 5.2 discusses their results in light of the whitening interpretation.

To the best of our knowledge, our results give the first universal upper bounds on what any feasible fixed-$X$ knockoff method can achieve. We prove finite-sample bounds on the maximum number of rejections that any knockoffs method can make, as a function of the design matrix $X$ and $\beta/\sigma$. Our bounds reveal that in some contexts the best achievable knockoff method markedly underperforms the Bonferroni method, but in other contexts off-the-shelf knockoff implementations outperform all other known methods. Understanding the fundamental limitations of the fixed-$X$ knockoff framework is a first step toward developing generalizations that ameliorate its flaws while retaining its many strengths.

## 1.5   Outline of results

Section 2 derives both finite-sample and asymptotic bounds on the power of knockoffs as a function of the matrix $\Sigma$ and the coefficient vector $\beta$. Our argument proceeds in several steps. First, Theorem 1 in Section 2.1 proves a strong form of equivalence between the usual formulation of fixed-$X$ knockoffs as presented in Section 1.1 and the whitening formulation presented in Section 1.2, establishing that every implementation of the former is also an implementation of the latter. Thus, it suffices to bound the whitening method's power.

The main challenge in proving universal negative results is dealing with the analyst's ability to customize two stages of the procedure: in Stage 1, the analyst can choose any diagonal matrix $\Delta \succeq \Sigma$, and in Stage 2, the analyst can use any exploratory method to order the hypotheses and select the alternative directions. We address these two choices in sequence: in Section 2.2, we show that an oracle analyst can optimize the exploration stage by sorting the variables in order of the log-odds $\eta_j = |2\beta_j \widetilde{\beta}_j|/\sigma^2 \Delta_{jj}$, and derive bounds on the expected number of rejections in terms of the number of $\eta_j$ values exceeding the critical threshold $-\log \alpha$. In Section 2.3 we obtain lower bounds on the order statistics of $\Delta_{11}, \ldots, \Delta_{dd}$.

Theorem 2 in Section 2.4 combines these two lines of argument to bound the expected number of rejections in terms of $\Sigma$ and $\beta/\sigma$. For the case where $\beta$ is sparse ($d_1/d \to 0$), Theorem 2 is very optimistic since the analyst can "cheat" by giving up on the $d - d_1$ null indices while constructing $\Delta$. Theorem 3 in Section 2.5 obtains a stronger bound for the more realistic setting where the analyst must choose $\Delta$ without knowing which hypotheses are non-null. Corollary 2 applies Theorems 2–3 to prove, roughly, that if $\lambda_1 \gg \log d$ and $u_1$ is dense then all knockoff methods have TPR $\to 0$ even in a regime where the Bonferroni correction enjoys TPR $\to 1$.

Since it is infeasible for a real analyst to achieve the same power as the oracle methods we describe, the bounds should not be taken as power estimates for real knockoff methods; instead, they reflect fundamental limits on the performance of the cleverest possible analyst, no matter how much domain wisdom or methodological wizardry they can bring to bear on the problem.

We emphasize that our results do not spell doom for knockoffs in all or even most settings; they hardly could, because there are indeed settings where knockoff methods outperform not only Bonferroni but also more powerful competitors like BH. Rather, they are a first step toward characterizing the regimes where each method is preferred and, we hope, toward developing methods that achieve near-optimal performance across all regimes. In particular, the regimes we identify can largely be recognized from the structure of $\Sigma$, which is known before the responses are observed. As a result, we can simply use a different method when the structure of $\Sigma$ is unfavorable.

Section 3 delves deeper into the whitening method. Section 3.1 shows how we can implement it when we only observe an unbiased multivariate Gaussian estimator $\widehat{\beta}$ with no accompanying design matrix, and Theorem 4 in Section 3.2 shows we can do the same with an asymptotically normal estimator along with a consistent covariance matrix estimator, though the result is limited to the classical asymptotic regime. Section 3.3 applies Theorem 4 to analyze a nonparametric bootstrap example using stock market return data. Section 4 shows empirical comparisons between knockoffs and competing methods in several regimes of interest, and Section 5 concludes.

# 2 Bounding the power of fixed-$X$ knockoffs

## 2.1 Equivalence of the two formulations

Next we will show that the two formulations of knockoffs in Sections 1.1 and 1.2, which we will respectively call the *standard method* and the *whitening method*, are essentially equivalent. In the standard method, an implementation of the knockoff filter is fully defined by a valid knockoff matrix $\widetilde{X}$ and a recipe for computing $W$-statistics satisfying the sufficiency and antisymmetry properties. In the whitening method, an implementation is fully defined by a diagonal matrix $\Delta \succeq \Sigma = (X^\mathsf{T} X)^{-1}$ along with a recipe for computing a hypothesis ordering and $\psi_j$ values from $|\widetilde{\beta}|$ and $\xi$. We will show that the implementations of each method are essentially in one-to-one correspondence with each other, using a coupling between $[X \; \widetilde{X}]^\mathsf{T} y$ and $\omega$.

To begin our analysis, assume that $\widetilde{X}$ is a valid knockoff matrix with $X^\mathsf{T} \widetilde{X} = X^\mathsf{T} X - D$, which can be constructed for any $D \preceq 2X^\mathsf{T} X$ (Barber and Candès, 2015), and assume all $D_{jj} > 0$. Set $\Delta = 2D^{-1}$, so $A = X^\mathsf{T} X - \Delta^{-1} = X^\mathsf{T} X - \frac{1}{2}D$.

Following a conditioning argument in Barber and Candès (2019), if we add and subtract $\widetilde{X}^\mathsf{T} y$ from $X^\mathsf{T} y$ we obtain a useful $2d$-variate Gaussian statistic whose mean and variance follow from $y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$:

$$\begin{pmatrix} (X + \widetilde{X})^\mathsf{T} y \\ (X - \widetilde{X})^\mathsf{T} y \end{pmatrix} = \mathcal{N}_{2d} \left( \begin{pmatrix} 2A\beta \\ D\beta \end{pmatrix}, \begin{pmatrix} 4\sigma^2 A & 0 \\ 0 & 2\sigma^2 D \end{pmatrix} \right). \tag{5}$$

It is suggestive that by rescaling the second component in (5) we obtain $D^{-1}(X - \widetilde{X})^\mathsf{T} y \sim \mathcal{N}_d(\beta, \sigma^2 \Delta)$, which is the desired distribution for $\widetilde{\beta}$. Pursuing this ansatz, set

$$\omega = D^{-1}(X - \widetilde{X})^\mathsf{T} y - \widehat{\beta} \sim \mathcal{N}_d \left( 0, \sigma^2 (\Delta - \Sigma) \right). \tag{6}$$

Because $\omega$ is ancillary and $\widehat{\beta}$ is complete sufficient in the submodel where $\sigma^2$ is known, it follows by Basu's Theorem that $\omega$ is independent of $\widehat{\beta}$. As a result we have $\widetilde{\beta} = D^{-1}(X - \widetilde{X})^\mathsf{T} y$ and

$$\xi = X^\mathsf{T} X \widehat{\beta} - \Delta^{-1} \widetilde{\beta} = X^\mathsf{T} y - \frac{1}{2}(X - \widetilde{X})^\mathsf{T} y = \frac{1}{2}(X + \widetilde{X})^\mathsf{T} y.$$

Next we relate the whitening method's exploratory stage to a slight relaxation of the sufficiency and antisymmetry properties.

**Proposition 1.** *For $j = 1, \ldots, d$ define*

$$W_j^*([X \; \widetilde{X}], y) = \mathrm{sgn}\left( (X_j - \widetilde{X}_j)^\mathsf{T} y \right) \cdot W_j([X \; \widetilde{X}], y). \tag{7}$$

*If $W$ satisfies the sufficiency and antisymmetry properties, then $W^*$ depends on $y$ only through the unordered pairs $\{X_1^\mathsf{T} y, \widetilde{X}_1^\mathsf{T} y\}, \ldots, \{X_d^\mathsf{T} y, \widetilde{X}_d^\mathsf{T} y\}$. In terms of the coupling defined by (6), if $W$ satisfies the sufficiency and antisymmetry properties then $W^*$ depends on $y$ only through $\xi$ and $|\widetilde{\beta}|$.*

The first claim of Proposition 1 is proven in the supplement of Barber and Candès (2019) but we provide a full proof here for expository purposes.

*Proof.* By the antisymmetry property we have

$$\begin{aligned} W_j^*(\mathrm{Swap}_j([X \; \widetilde{X}]), y) &= \mathrm{sgn}\left( (\widetilde{X}_j - X_j)^\mathsf{T} y \right) \cdot W_j(\mathrm{Swap}_j([X \; \widetilde{X}]), y) \\ &= \mathrm{sgn}\left( -(X_j - \widetilde{X}_j)^\mathsf{T} y \right) \cdot \left( -W_j([X \; \widetilde{X}], y) \right) \\ &= W_j^*([X \; \widetilde{X}], y), \end{aligned}$$

so $W_k^*(\mathrm{Swap}_j([X \; \widetilde{X}]), y) = W_k^*([X \; \widetilde{X}], y)$ for all $k$ including $k = j$.

By the sufficiency property, $W$ only depends on $y$ through $[X \; \widetilde{X}]^\mathsf{T} y$, so the same is true of $W^*$. But we have just seen that $W^*$ is invariant to swapping $X_j$ with $\widetilde{X}_j$, which amounts to swapping $X_j^\mathsf{T} y$ with $\widetilde{X}_j^\mathsf{T} y$ without changing $[X \; \widetilde{X}]^\mathsf{T} [X \; \widetilde{X}]$. In other words, $W^*$ depends on $y$ only through the unordered pairs $\{X_j^\mathsf{T} y, \widetilde{X}_j^\mathsf{T} y\}$, for $j = 1, \ldots, d$.

For the coupling in (6), we have $2\xi = (X + \widetilde{X})^\mathsf{T} y$ and $D\widetilde{\beta} = (X - \widetilde{X})^\mathsf{T} y$. Hence, observing $\xi$ and $|\widetilde{\beta}|$ is the same as observing $(X + \widetilde{X})^\mathsf{T} y$ and $|(X - \widetilde{X})^\mathsf{T} y|$, which is the same as observing the unordered pairs. $\square$

We will say that $W$ satisfies the *unordered pair property* if $W^*$ depends on $y$ only through $\{X_j^\mathsf{T} y, \widetilde{X}_j^\mathsf{T} y\}$ for $j = 1, \ldots, d$. This condition relaxes the sufficiency and antisymmetry properties because it allows for $W^*$ to have unrestricted dependence on $X$ and $\widetilde{X}$. With these relationships established, we are now prepared to prove formal equivalence.

**Theorem 1.** *Assume that $n \geq 2d$ and let $\widetilde{X}$ be a knockoff matrix with $X^\mathsf{T}\widetilde{X} = X^\mathsf{T} X - D$. Let $\Delta = 2D^{-1}$ and define $\omega$ as in* (6)*. Then*

(a) *For any implementation of the whitening method, we can construct $W$-statistics satisfying the unordered pair property so that the two methods give identical rejection sets.*

(b) *For any $W$-statistics satisfying the unordered pair property such that $|W_1|, \ldots, |W_d|$ are almost surely positive with no ties, we can construct an implementation of the whitening method so that the two methods give identical rejection sets.*

*Proof.* For (a), take $W_{[j]} = (d + 1 - j) \cdot \psi_{[j]} \cdot \mathrm{sgn}(\widetilde{\beta}_{[j]})$, for $j = 1, \ldots, d$. Recalling that $\mathrm{sgn}(\widetilde{\beta}_j) = \mathrm{sgn}((X_j - \widetilde{X}_j)^\mathsf{T} y)$, we have $W_{[j]}^* = (d + 1 - j) \cdot \psi_{[j]}$, so $W^*$ depends on $y$ only through $\xi$ and $|\widetilde{\beta}|$, satisfying the unordered pair property.

For (b), because the method depends on $|W|$ only through the ordering of its coordinates, we can assume without loss of generality that $|W|$ is a permutation of $\{1, \ldots, d\}$. Take $[j]$ to be the index of the $j$th largest $|W_j|$ value, so that $|W_{[j]}| = d + 1 - j$, and set $\psi_j = \mathrm{sgn}(W_j^*)$, which is a function of $\xi$ and $|\widetilde{\beta}|$. Then we again have $W_{[j]} = (d + 1 - j) \cdot \psi_{[j]} \cdot \mathrm{sgn}(\widetilde{\beta}_{[j]})$.

To see why the two methods return the same rejection sets when $W_{[j]} = (d + 1 - j) \cdot \psi_{[j]} \cdot \mathrm{sgn}(\widetilde{\beta}_{[j]})$, note first that $\widetilde{p}_j = 1/2 \iff \mathrm{sgn}(\widetilde{\beta}_j) = \psi_j \iff W_j > 0$. As a result

$$\widehat{\mathrm{FDP}}_k^{\mathrm{wh}} = \frac{1 + \sum_{j=1}^d 1\{\widetilde{p}_{[j]} = 1, [j] \leq k\}}{\sum_{j=1}^d 1\{\widetilde{p}_{[j]} = 1/2, [j] \leq k\}} = \frac{1 + \sum_{j=1}^d 1\{W_{[j]} \leq -(d + 1 - k)\}}{\sum_{j=1}^d 1\{W_{[j]} \geq d + 1 - k\}} = \widehat{\mathrm{FDP}}_{d+1-k}^{\mathrm{kn}}.$$

For other values of $t \in (0, d + 1]$, $\widehat{\mathrm{FDP}}_t^{\mathrm{kn}} = \widehat{\mathrm{FDP}}_{\lceil t \rceil}^{\mathrm{kn}}$, where $\lceil t \rceil$ is the integer ceiling of $t$. Therefore, $\hat{t} = d + 1 - \hat{k}$ and the rejection sets are the same.

$\square$

Because the whitening method controls FDR, Theorem 1 implies that the unordered pair property is a sufficient condition for the standard knockoff filter to control FDR as well, relaxing the sufficiency and antisymmetry properties. The requirement in (b) that all $|W_j|$ be positive and distinct is not really necessary; we could break ties or generate "signs" at random, or generalize the whitening method so that $\widehat{\mathrm{FDP}}_k^{\mathrm{wh}}$ is only calculated for a subset of $k = \{1, \ldots, d\}$ corresponding to the indices where $|W_{[k]}|$ is positive and decreases. We ignore these generalizations for the sake of brevity because they can only reduce the number of rejections.

As Theorem 1 shows, any implementation of the standard method is achievable with the whitening method. It follows that any uniform bound on TPR for all implementations of the whitening method is also a uniform bound for all implementations of the standard method.

## 2.2 The oracle analyst and the knockoff* procedure

Throughout the rest of Section 2, we consider the perspective of an oracle analyst with full knowledge of $\beta$. To bound the power of this analyst we work backwards from the end of the procedure. In this section we discuss how the oracle analyst can make optimal decisions in the exploratory analysis of Stage 2, and in Sections 2.3–2.4 we apply our analysis to understand the ramifications for choosing $\Delta$.

The analyst's task in Stage 2 is to choose alternative directions $\psi_1, \ldots, \psi_d$ and a hypothesis ordering $[1], \ldots, [d]$, after observing $\xi$ and $|\widetilde{\beta}|$. To choose $\psi_j$, recall that $\widetilde{p}_j = 1/2$ only if $\mathrm{sgn}(\widetilde{\beta}_j) = \psi_j$. As a result, an analyst who knows $\beta$ will always set $\psi_j = \mathrm{sgn}(\beta_j)$, unless $\beta_j = 0$ in which case $\psi_j$ makes no difference and can be chosen arbitrarily. Then the best possible ordering of variables is in decreasing order of the conditional log-odds:

$$\eta_j = \mathrm{logit} \, \mathbb{P}\left(\mathrm{sgn}(\widetilde{\beta}_j) = \mathrm{sgn}(\beta_j) \mid \xi, |\widetilde{\beta}|\right) = \frac{2|\widetilde{\beta}_j|}{\sigma^2 \Delta_{jj}} \cdot |\beta_j|. \tag{8}$$

8

Because $\widetilde{\beta}$ is continuous, the $\eta_j$ values for nonzero $\beta_j$ are almost surely positive and distinct, whereas if $\beta_j = 0$ then $\eta_j = 0$ too. Define the *knockoff\* procedure* to be any method that sets $\psi_j = \text{sgn}(\beta_j)$ if $\beta_j \neq 0$ and arbitrarily otherwise, and returns an ordering with

$$\eta_{[1]} > \cdots > \eta_{[d_1]} > 0 = \eta_{[d_1+1]} = \cdots = \eta_{[d]}.$$

The relative ordering of the null hypotheses is also arbitrary. We show next that the knockoff\* procedure maximizes the TPR among all implementations using the same $\Delta$.

**Proposition 2.** *Conditional on $\xi$ and $|\widetilde{\beta}|$, both the TPP and the expected total number of rejections for the knockoff\* procedure are stochastically larger than they are any other implementation with the same $\Delta$. Consequently, the knockoff\* procedure maximizes both the TPR and the expected number of rejections.*

While there is some ambiguity in defining the knockoff\* procedure concerning the alternative direction and relative ordering for the null indices, the statement of Proposition 2 applies to any version of it, implying in particular that any two versions of knockoff\* have the same TPR. The proof, given in Section 6, uses the following useful formula for the number of rejections $R$ as a function of $\hat{k}$. Because we must have $\widetilde{p}_{[\hat{k}+1]} = 1$,

$$\widehat{\text{FDP}}_{\hat{k}}^{\text{wh}} \leq \alpha < \widehat{\text{FDP}}_{\hat{k}+1}^{\text{wh}} \implies 1 + \hat{k} \leq (1+\alpha)R < 2 + \hat{k} \implies R = \left\lceil \frac{1+\hat{k}}{1+\alpha} \right\rceil. \tag{9}$$

Next, we develop tools for bounding the power of the knockoff\* procedure. If we write the order statistics of $\eta$ as $\eta_{(1)} > \cdots > \eta_{(d_1)} > 0 = \eta_{(d_1+1)} = \cdots = \eta_{(d)}$, then we can obtain upper bounds on the TPR of any knockoff procedure by bounding the knockoff\* procedure which has $\eta_{[k]} = \eta_{(k)}$ for all $k$.

Roughly speaking, for knockoffs to make rejections, we must see $\widetilde{p}_{[j]} = 1/2$ at least $1/\alpha$ times as often as $\widetilde{p}_{[j]} = 1$ near the beginning of the list, so the largest $\eta_{(j)}$ values must exceed $-\log \alpha$; Proposition 3 gives a simple result relating these quantities. For $0 < p < q < 1$, define the random walk

$$S_k^{p,q} = \sum_{j=1}^{k} p - Z_j, \quad \text{where } Z_1, Z_2, \ldots \overset{\text{i.i.d.}}{\sim} \text{Bern}(q),$$

and let $M_d^{p,q} = \max\{k \leq d : S_k^{p,q} \geq 1 - p\}$.

**Proposition 3.** *Conditional on $\eta_{(1)} < -\log \alpha$, the number of rejections for any knockoff procedure at FDR level $\alpha$ is stochastically smaller than $M_d^{p,q_1}$ for $p = \frac{\alpha}{1+\alpha}$, $q_1 = \frac{1}{1+e^{\eta_{(1)}}}$.*

*Proof.* Conditioning on $\xi$ and $|\widetilde{\beta}|$, define $q_j = \frac{1}{1+e^{\eta_{(j)}}}$, and $Z_j' = 1\{\widetilde{p}_j = 1\} \overset{\text{ind.}}{\sim} \text{Bern}(q_j)$. Then for the knockoff\* procedure,

$$\widehat{\text{FDP}}_k^{\text{wh}} = \frac{1 + A_k}{R_k}, \quad \text{where} \quad A_k = \sum_{j=1}^{k} Z_j', \quad \text{and} \quad R_k = k - A_k = \sum_{j=1}^{k} 1 - Z_j'.$$

The number of knockoff\* rejections is $R_{\hat{k}} \leq \hat{k}$, where $\hat{k}$ is the largest index $k$ with

$$\frac{1 + A_k}{R_k} \leq \alpha \iff 1 \leq \alpha R_k - A_k = \sum_{j=1}^{k} \alpha - (1+\alpha)Z_j' = \frac{1}{1-p} \sum_{j=1}^{k} p - Z_j'.$$

Next, define auxiliary random variables $B_1, \ldots, B_d \overset{\text{i.i.d.}}{\sim} \text{Bern}(q_1/q_j)$ and construct the random walk $S_k^{p,q_1}$ whose first $d$ increments are $Z_j = Z_j' \cdot B_j \overset{\text{i.i.d.}}{\sim} \text{Bern}(q_1)$. Then $\alpha R_k - A_k \leq \frac{1}{1-p} S_k^{p,q_1}$, so we have almost surely $\hat{k} \leq M_d^{p,q_1}$. $\qquad\square$

In particular, Proposition 3 implies that in any asymptotic regime where $\eta_{(1)} \overset{p}{\to} 0$, the number of knockoff rejections is $O_p(1)$ for any FDR level $\alpha$.

While Proposition 3 can help build our intuition about the role of the log-odds threshold $-\log \alpha$, it is not a strong enough result to control the power of an oracle analyst, who can always take $\Delta_{jj} \downarrow \Sigma_{jj}$ and force $[1] = j$ for a single promising variable $j$. The next result generalizes Proposition 3, showing that knockoff\* can expect no more than $O(k)$ rejections when there are fewer than $k$ large log-odds.

9

**Proposition 4.** *Fix a significance level $\alpha > 0$ and margin $\delta > 0$. Conditional on $\eta_{(k)} < -\log(\alpha + \delta)$, the expected number of rejections for any knockoff procedure at FDR significance level $\alpha$ is upper bounded by $C_1(\alpha, \delta)k + C_2(\alpha, \delta)$, where $C_1$ and $C_2$ are constants that depend only on $\alpha$ and $\delta$. In particular, if $d_1 \to \infty$ and if*

$$\eta_{(\lceil cd_1 \rceil)} \xrightarrow{p} 0, \ \text{for all } c > 0,$$

*then* TPR $\to 0$ *for any knockoff procedure at any FDR significance level.*

We defer the proof of this proposition to Section 6, where we also give explicit formulae for the constants $C_1$ and $C_2$. For example, when $\alpha = 0.05$ and $\alpha + \delta = \sqrt{0.05}$, we have $C_1 \le 2.3$ and $C_2 \le 40$. These constants are not optimal and better bounds can be obtained for specific values $k, d, \alpha, \delta$ by simulating the random walk with $Z_1 = \cdots = Z_{k-1} = 0$ and $\eta_{(k)} = \cdots = \eta_{(d)} = -\log(\alpha + \delta)$.

Whereas Proposition 4 bounds what is possible after whitening and observing $\xi$ and $|\widetilde{\beta}|$, the next section analyzes when it is impossible to carry out the whitening step without dramatic information loss that prevents the $\eta$ values from being large enough. We will show that this loss is determined by the eigen-decomposition of $\Sigma = (X^\mathsf{T} X)^{-1}$, and characterize how large the signal must be to overcome this information loss and achieve nontrivial TPR, and illustrate our analysis with examples.

## 2.3 Lower bounds for $\Delta$: how much noise must we add?

Because the power of any knockoff procedure can be upper bounded by the knockoff* procedure from the previous section, understanding the best attainable power of knockoffs is a matter of investigating the joint distribution of the $\eta_1, \ldots, \eta_d$ values, which are independent of $\xi$ with distribution

$$\eta_j = \left| \frac{2\beta_j \widetilde{\beta}_j}{\sigma^2 \Delta_{jj}} \right| \stackrel{\text{ind.}}{\sim} |\mathcal{N}(\mu_j, 2\mu_j)|, \quad \text{for } \mu_j = \frac{2\beta_j^2}{\sigma^2 \Delta_{jj}}. \tag{10}$$

The distribution of $\eta_j$ depends on the ratio of two quantities: the signal-to-noise ratio (SNR) $\beta_j/\sigma$, and the diagonal entry $\Delta_{jj}$. Since $\eta_j/\sqrt{2\mu_j} \sim |\mathcal{N}(\sqrt{\mu_j/2}, 1)|$ is stochastically increasing in $\mu_j$, so is $\eta_j$. As a result, the *marginal* power of the knockoff* procedure is strictly increasing in each $\mu_j$. We will be primarily interested in the standardized case where $\Sigma_{jj} = \sigma^2 = 1$; in that case $\Delta_{jj} = \mathrm{Var}(\widetilde{\beta}_j) = 1 + \mathrm{Var}(\omega_j)$, so knockoffs will (potentially) have power when the squared SNR is large compared to the whitening noise.

Although most knockoff methods seek roughly to balance the whitening noise across the coordinates, if we want to bound the power of the best possible knockoff method we must consider the possibility that the analyst will favor some variables over others. For example, it is always possible to drive a single $\Delta_{jj} \to \Sigma_{jj}$, but only at the cost of driving $\Delta_{kk} \to \infty$ for all $k$ with $\Sigma_{jk} \neq 0$, since we must have

$$\mathrm{Var}(\omega_k) \ge \frac{\mathrm{Cov}^2(\omega_j, \omega_k)}{\mathrm{Var}(\omega_j)} = \frac{\Sigma_{jk}^2}{\Delta_{jj} - \Sigma_{jj}}.$$

Thus, while it is always possible to make any one $\Delta_{jj}$ small, it is sometimes impossible to make them all small at the same time. In this section we derive explicit lower bounds for the order statistics of $\Delta_{11}, \ldots, \Delta_{dd}$ as functions of $\Sigma$, uniformly over all diagonal matrices $\Delta \succeq \Sigma$. Let $\lambda_1$ be the leading eigenvalue of $\Sigma$, with eigenvector $u_1$. For any subset $S \subset \{1, \ldots, d\}$, we must have $\Delta_{S,S} \succeq \Sigma_{S,S}$, for the corresponding matrix blocks. As a result,

$$\sum_{j \in S} \Delta_{jj} u_{1,j}^2 = u_{1,S}^\mathsf{T} \Delta_{S,S} u_{1,S} \ge \sum_{\ell=1}^d \lambda_\ell (u_{\ell,S}^\mathsf{T} u_{1,S})^2 \ge \lambda_1 \|u_{1,S}\|_2^4.$$

Therefore

$$\|u_{1,S}\|_2^2 \max_{j \in S} \Delta_{jj} \ge \sum_{j \in S} \Delta_{jj} u_{1,j}^2 \ge \lambda_1 \|u_{1,S}\|_2^4,$$

and we have

$$\max_{j \in S} \Delta_{jj} \ge \lambda_1 \|u_{1,S}\|_2^2. \tag{11}$$

To interpret equation (11), suppose that the entries of the unit vector $u_1$ are all at least as large as $c/\sqrt{d}$, for some $c \in (0, 1]$. Then for any block $\Delta_{S,S}$, we must have $\max_{j \in S} \Delta_{j,j} \ge c^2 \lambda_1 |S|/d$. Taking $S$ to be the set with the $k$ smallest entries, it follows that the $k$th smallest diagonal entry of $\Delta$ is at least $c^2 \lambda_1 k/d$. Extending this argument yields explicit lower bounds for order statistics of $\Delta_{1,1}, \ldots, \Delta_{d,d}$, in terms of $\Sigma$.

**Proposition 5.** *Let $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d \geq 0$ be the eigenvalues of $\Sigma$, with corresponding eigenvectors $u_1, \ldots, u_d$, and let $u_{\ell,(1)}^2 \leq \cdots \leq u_{\ell,(d)}^2$ be the order statistics of $u_{\ell,1}^2, \ldots, u_{\ell,d}^2$, and let $\Delta_{(1,1)} \leq \cdots \leq \Delta_{(d,d)}$ be the order statistics of the diagonal entries of $\Delta$. Then for all $k = 1, \ldots, d$, we can lower bound*

$$\Delta_{(k,k)} \;\geq\; b_k(\Sigma) \;=\; \max_{\ell \leq d} \lambda_\ell \sum_{j=1}^{k} u_{\ell,(j)}^2. \tag{12}$$

*Proof.* Note that in the argument for the lower bound (11), we can just as well replace $\lambda_1, u_1$ with $\lambda_\ell, u_\ell$ for any $\ell = 1, \ldots, d$. If $S$ is the index set for the $k$ smallest diagonal entries of $\Delta$, then

$$\Delta_{(k,k)} \;=\; \max_{j \in S} \Delta_{jj} \;\geq\; \lambda_\ell \|u_{\ell,S}\|^2 \;\geq\; \lambda_\ell \sum_{j=1}^{k} u_{\ell,(j)}^2.$$

Maximizing the lower bound over $\ell = 1, \ldots, d$ gives the result. $\qquad\square$

By construction, we have $b_1(\Sigma) \leq \cdots \leq b_d(\Sigma)$, and also $b_j(\Sigma) \geq \frac{j}{k} b_k(\Sigma)$ for all $j > k$. Continuing the example from above, we will have $b_k(\Sigma) \geq c^2 \lambda_1 k/d$.

In the MCC case with $\lambda_1 = 1 + \rho(d-1)$ and $u_1 = \mathbf{1}_d/\sqrt{d}$, we have exactly $b_k(\Sigma) = \lambda_1 k/d \geq \rho k$ for all $k = 1, \ldots, d$. These bounds are nearly tight: for any size-$k$ subset $S$, the leading eigenvalue of $\Sigma_{S,S}$ is

$$\lambda_1^S \;=\; 1 + \rho(k-1) \;\leq\; \rho k + 1.$$

Then for any $\epsilon > 0$, we can set

$$\Delta_{S,S} \;=\; (\lambda_1^S + \epsilon) I_k \;\succeq\; \Sigma_{S,S},$$

but if we take $\epsilon \to 0$ we must have $\Delta_{jj} \to \infty$ for every $j \in S^c$. In other words, we can "rescue" our favorite $k$ variables from having $\Delta_{jj} > \rho k$, as long as we are willing to give up completely on the other $d - k$.

## 2.4 Uniform bounds on the power of fixed-$X$ knockoffs

Theorem 2 combines Propositions 4 and 5 to obtain a finite-sample upper bound on the power of any fixed-$X$ knockoff procedure. We first give one more technical lemma:

**Lemma 1.** *Suppose that $z_i \sim \mathcal{N}(\mu_i, 2\mu_i)$ for $i = 1, 2, \ldots$, where $0 \leq \mu_i \leq \frac{-k \log \alpha}{2(k+i)}$, for some integer $k \geq 1$ and $\alpha \in (0, 1)$. Then the expected number of $z_i$ values with $|z_i| > -\frac{1}{2} \log \alpha$ is bounded by $C_3(\alpha)k$.*

Whereas Proposition 4 controls the number of rejections in terms of the number of large $\eta_j$ values, Lemma 1 controls the number of large $\eta_j$ values we can get by chance from $\mu_j$ values falling below $-\frac{1}{2} \log \alpha$. We are now ready to prove our main result.

**Theorem 2.** *Let $\beta_{(1)}^2 \geq \cdots \geq \beta_{(d)}^2$ be the order statistics of $(\beta_1^2, \ldots, \beta_d^2)$ and let $b_1(\Sigma), \ldots, b_d(\Sigma)$ be defined as in* (12). *For any target FDR level $\alpha > 0$, let $k$ be the smallest integer such that*

$$\frac{2\beta_{(k)}^2}{\sigma^2 b_k(\Sigma)} \;<\; -\frac{1}{2} \log \alpha, \tag{13}$$

*Then the expected number of rejections for any knockoff procedure at FDR level $\alpha$ is upper bounded by $C_1^*(\alpha)k + C_2^*(\alpha)$, where $C_1^*$ and $C_2^*$ depend only on $\alpha$.*

To sketch the proof, in expectation there are at most $(2 + C_3)k$ indices for which $\eta_j > -\frac{1}{2} \log \alpha$: $k$ with large $\beta_j$, $k$ more with small $\Delta_{jj}$, and $C_3 k$ more by chance; then applying Proposition 4 gives the final result. As discussed in the proof of Theorem 2, if we replace $\beta_{(k)}$ with $\beta_{(1)}$ in (13) then we eliminate the portion with large $\beta_j$, so $2 + C_3$ improves to $1 + C_3$, giving correspondingly smaller bounds. Table 1 shows the bounds at several $\alpha$ levels of interest.

The dimension $d$ does not appear in the inequality (13), unless it plays a role in determining $b_k(\Sigma)$. For MCC, $b_k(\Sigma) \approx \rho k$, so the number of rejections as $d \to \infty$ grows no faster than the square of the SNR. If we set the signal strength at the Bonferroni threshold $\beta_0 = \sigma \sqrt{2 \log d}$, then solving (13) gives $k = -8 \log d / \rho \log \alpha$, which we can plug directly into the bounds in the second row of Table 1. As a consequence, the TPR of any knockoffs method will tend to 0 if $d_1/\log d \to 0$. By contrast, if $d_1/d$ converges to a nonzero constant, then we must have $\beta_0/\sigma = O(\sqrt{d})$ to achieve nontrivial TPR in the limit.

Table 1: Upper bounds from Theorem 2 on $\mathbb{E}R$ if $\dfrac{2\beta_{(\ell)}^2}{\sigma^2 b_k(\Sigma)} \le -\dfrac{1}{2}\log\alpha$, given for $\ell = k$ and $\ell = 1$.

|                    | $\alpha = 0.05$ | $\alpha = 0.1$ | $\alpha = 0.2$ |
|--------------------|-----------------|----------------|----------------|
| Bound if $\ell = k$ | $7.0k + 40$     | $10.6k + 41$   | $19.1k + 61$   |
| Bound if $\ell = 1$ | $4.7k + 43$     | $7.4k + 45$    | $14.3k + 66$   |
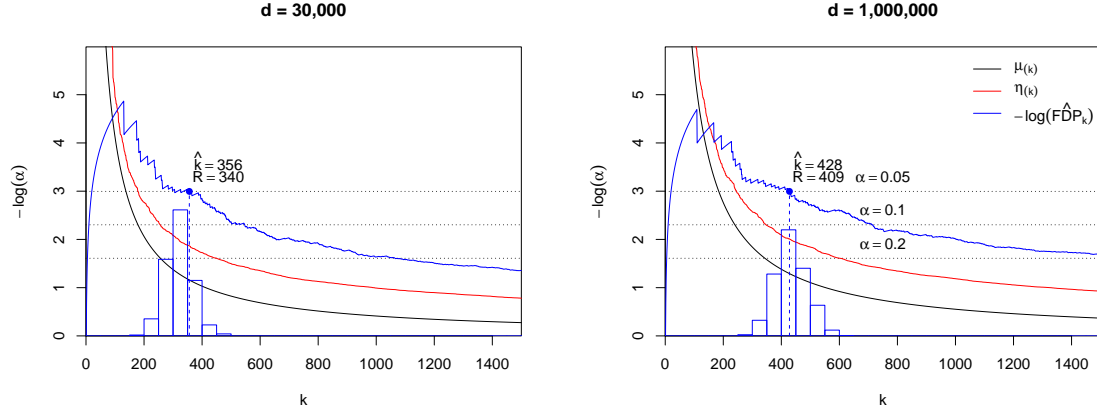


Figure 1: Simulation of the best possible knockoff method in the setting of Theorem 2, where the analyst has full knowledge of $\beta$ for both selecting $\Delta$ and carrying out the exploratory stage. We simulate under $b_k(\Sigma) = c^2\lambda_1 k/d$, with $c^2 = 0.2$ and $\lambda_1 = d/2$, but these values can be substituted with $b_k(\Sigma)$ for any $\Sigma$ of interest. The signal strength is at the Bonferroni threshold, with all nonzero coefficients equal to $\beta_0 = \sigma\sqrt{2\log d}$, and $d_1 = d$ to be as optimistic as possible. The black line shows upper bounds for $\mu_{(k)} \le 4\log d/b_k(\Sigma)$, the red line shows a single realization of $\eta_{(k)}$ where $\eta_j \overset{\text{ind.}}{\sim} |\mathcal{N}(\mu_j, 2\mu_j)|$, and the blue line shows a realization of $-\log(\widehat{\mathrm{FDP}}_k^{\mathrm{wh}})$. The blue histogram shows the distribution of rejections over 1000 simulations.

More generally, to apply Theorem 2 for a given covariance matrix $\Sigma$ at any signal strength, we can calculate the sequence $b_k(\Sigma)$ using the formula (12), and plug the smallest $k$ satisfying (13) into Table 1.

We can visualize the logic of Theorem 2 and obtain improved bounds for specific examples by simulation. If all nonzero $\beta_j$ values are set to a common value $\beta_0$, then the $k$th largest $\mu_{(k)}$ is no larger than $2\beta_0^2/\sigma^2 b_k(\Sigma)$. To simulate the knockoff* procedure, we sample the $\eta_j$ values according to (10), sort them in descending order, and then simulate the resulting $\widehat{\mathrm{FDP}}_k^{\mathrm{wh}}$ process. Figure 1 illustrates two examples, with $d = 30,000$ and $d = 1,000,000$, and with $\beta_0$ set at the Bonferroni threshold $\sigma\sqrt{2\log d}$. We also set $d_1 = d$ to simulate under the most optimistic conditions.[3] We simulate the case $b_k(\Sigma) = 0.1k/d$ and show $\mu_{(k)}$ along with a single realization of the $\eta_{(k)}$ and $\widehat{\mathrm{FDP}}_k^{\mathrm{wh}}$ processes. In addition we plot a histogram showing the distribution of the number of rejections for $\alpha = 0.05$.

Continuing the example from the previous section where $b_k(\Sigma) \ge c^2\lambda_1 k/d$, suppose that all nonzero coefficients take the same value $\beta_0$. In that case, Theorem 2 implies that *no* knockoffs method can make more than $O\left(\dfrac{\beta_0^2 d}{\lambda_1}\right)$ rejections. If $\beta_0 = \sqrt{2r\log d}$ then no knockoffs method can make more than $O\left(\dfrac{d\log d}{\lambda_1}\right)$ rejections, so knockoffs must have TPR $\to 0$ if $\lambda_1 \gg \dfrac{d\log d}{d_1}$. Thus, if $d_1/d$ converges to a nonzero constant, then TPR $\to 0$ for knockoffs whenever $\lambda_1 \gg \log d$. Thus, knockoffs can underperform Bonferroni even in settings where the first eigenvalue asymptotically accounts for a vanishingly small fraction of the total variance. By contrast, if $d_1/d = o(\log d)$, then Theorem 2 does not necessarily suggest we will have a problem. As we will discuss in the next section, Theorem 2 gives especially optimistic predictions when $d_1 \ll d$ because the oracle analyst is too powerful.

---

[3]Alternatively we could choose a value for $d_1$ and set $\mu_{(k)} = 0$ for $k > d_1$, but it makes little difference if $\mu_{(d_1)} \ll -\log\alpha$.

---

**Algorithm 1:** T3-knockoff* simulation

---

**Data:** Covariance matrix $\Sigma$, noise level $\sigma^2$, FDR level $\alpha$, order statistics $\beta_{(1)}^2, \ldots, \beta_{(d)}^2$

**Result:** Maximum achievable TPR for any knockoff method (Monte Carlo estimate)

Calculate lower bounds $b_1(\Sigma), \ldots, b_d(\Sigma)$ using (12);

**for** $m = 1$ **to** $M$ **do**

    Sample $\beta_1^2, \ldots, \beta_d^2$ by permuting order statistics;

    **for** $j = 1$ **to** $d$ **do**

        Calculate $\mu_j = 2\beta_j^2/\sigma^2 b_j(\Sigma)$;

        Sample $\eta_j \sim |\mathcal{N}(\mu_j, 2\mu_j)|$;

        Sample $\widetilde{p}_j \in \{1/2, 1\}$ with logit $\mathbb{P}(\widetilde{p}_j = 1/2) = \eta_j$;

    **end**

    Order $\widetilde{p}_{[1]}, \ldots, \widetilde{p}_{[d]}$ in descending order of $\eta_j$;

    Find $\hat{k} = \max\{k : \widehat{\text{FDP}}_k^{\text{wh}} \leq \alpha\}$;

    Calculate $\text{TPP}_m = \frac{1}{d_1} \sum_{j=1}^{\hat{k}} 1\{\widetilde{p}_{[j]} = 1/2 \text{ and } \beta_j \neq 0\}$;

**end**

**return** $\widehat{\text{TPR}} = \frac{1}{M} \sum_m \text{TPP}_m$;

---

## 2.5 Stronger results with random non-null indices

While the bounds from Theorem 2 are nontrivial, they are still highly optimistic, since they allow the analyst to optimize not only the exploratory stage (equivalent to choosing perfect $W$-statistics), but also the choice of $\Delta$ (equivalent to the choice of knockoff matrix $\widetilde{X}$), with full knowledge of the coefficients. In particular, suppose that the true coefficient vector is sparse with $S_1 = \{j : \beta_j \neq 0\}$ and $S_0 = S_1^{\mathsf{c}}$. Then the analyst loses nothing by sending $\Delta_{jj} \to \infty$ for all $j \in S_0$, effectively reducing the multiplicity of the problem to $d_1 \ll d$ by giving up on the null variables before the method has even begun. Whereas we might find it plausible that the exploratory stage is highly efficient, especially when the signals are very strong and the $W$-statistics are designed well, it is much less plausible that the analyst can discard all of the null variables even before observing any data. In any case, such a strategy is contrary to the spirit of knockoff methods that are actually in use, which seek to balance the $\Delta_{jj}$ (or $D_{jj}$) values, spreading power across all $d$ variables.

We can force the oracle analyst into a more "honest" allocation of $\Delta_{jj}$ values by assuming the indices in $S_1$ are drawn uniformly at random from $\{1, \ldots, d\}$ only *after* $\Delta$ has been chosen. As before, we still assume $\beta$ is revealed to the analyst before the exploratory stage. Theorem 3 gives a directly analogous result to Theorem 2, with a similar proof.

**Theorem 3.** *Let $\beta_{(1)}^2 \geq \cdots \geq \beta_{(d)}^2$ be the order statistics of $(\beta_1^2, \ldots, \beta_d^2)$, and assume $\beta$ is a random permutation of the order statistics, independent of $\Delta$. Let $\pi_1 = d_1/d$ be the non-null proportion, and define $b_1(\Sigma), \ldots, b_d(\Sigma)$ as in (12). For any target FDR level $\alpha > 0$, let $k$ be the smallest integer for which*

$$\frac{2\beta_{(k)}^2}{\sigma^2 b_{\lfloor k/\pi_1 \rfloor}(\Sigma)} \; < \; -\frac{1}{2}\log\alpha, \tag{14}$$

*where $\lfloor \cdot \rfloor$ is the integer floor. Then the expected number of rejections for any knockoff procedure at FDR level $\alpha$ is upper bounded by $C_1^*(\alpha)k + C_2^*(\alpha)$, where $C_1^*$ and $C_2^*$ are the same constants as in Theorem 2.*

We can tighten the bounds if we replace $\beta_{(k)}^2$ with $\beta_{(1)}^2$ in (13), just as with Theorem 2, so the bounds in both rows of Table 1 apply. Likewise, we can simulate better bounds just as we did in Theorem 2, with the modification that after calculating $\mu_{(k)} = 2\beta_0^2/\sigma^2 b_k(\Sigma)$, we randomly zero out $d - d_1$ of the $\mu_{(k)}$ values. Algorithm 1 fully specifies a Monte Carlo estimate for the highest achievable TPR when the analyst only knows the order statistics $\beta_{(1)}^2, \ldots, \beta_{(d)}^2$ when choosing $\Delta$.

By placing this mild limitation on the analyst's otherwise-total omniscience, we dramatically strengthen the result from Theorem 2 in sparse problems. To illustrate, consider the bound $\mathbb{E}R \leq 4.7k + 43$ from Table 1, the $\beta_{(1)}$ version of the bound as applied to $\alpha = 0.05$. If we were applying Theorem 2, then to get TPR $\geq 1/2$, we would only need $k = d_1/10$ small $\Delta_{jj}$ values. By contrast, applying Theorem 3 we would

need $k/\pi_1 = d/10$ small $\Delta_{jj}$ values, since we can expect all but a fraction $\pi_1$ of them to be wasted on null indices. Pragmatically, even $d/10$ is not enough because the bounds in Table 1 are not optimal; to achieve TPR $\geq 1/2$ we realistically need most of the $\Delta_{jj}$ values to be small.

Corollary 1 formalizes the sense in which Theorem 3 and Algorithm 1 both upper-bound the best achievable knockoff method. Let $\mathcal{R}_\alpha(\widehat{\beta}, \Sigma)$ denote any valid implementation of fixed-$X$ knockoffs at FDR level $\alpha$, mapping its inputs to a randomized rejection set. Then we have

**Corollary 1.** *Fix $\sigma^2 = 1$. Let $B_T$ denote the bound obtained from Theorem 3, and $B_A = \mathbb{E}\,\widehat{\text{TPR}}$ is the output of Algorithm 1; both bounds depend on $\Sigma$, $\alpha$, and the order statistics $\beta_{(1)}^2, \ldots, \beta_{(d)}^2$. Then*

$$\frac{1}{d_1} B_T \;\geq\; B_A \;\geq\; \sup_{\mathcal{R}} \min_{\beta \in \Omega} \text{TPR}(\mathcal{R}, \beta),$$

*where $\Omega \subseteq \mathbb{R}^d$ is the set of vectors $\beta \in \mathbb{R}^d$ whose squared entries have the same order statistics, and* $\text{TPR}(\mathcal{R}, \beta)$ *is the* TPR *of method $\mathcal{R}$ under $\widehat{\beta} \sim \mathcal{N}_d(\beta, \Sigma)$.*

We finish by applying Theorems 2–3 to prove a simple corollary classifying asymptotic regimes where knockoffs underperform Bonferroni. Whereas other works commonly assume a specific limiting distribution for the covariance matrix $\Sigma$ (e.g. Liu and Rigollet, 2019; Weinstein et al., 2017), we only impose a less stringent condition quantifying how "dense" the first eigenvector $u_1$ is. For $c \geq 0$, define the distribution function of the entries of $u_1$, scaled by $\sqrt{d}$:

$$F_d(c) \;=\; \frac{1}{d} \sum_{j=1}^d \mathbf{1}\left\{ |u_{1,j}| \;\leq\; \frac{c}{\sqrt{d}} \right\}.$$

We have been analyzing the case where $F_d(c) = 0$ for some $c > 0$. For the next results, we will assume only that $F_d(\cdot) \to F(\cdot)$. If $\lambda_1 \gg \log d$, the following corollary gives relatively weak conditions for when the best possible knockoff procedure must have TPR $\to 0$ for signals above the Bonferroni threshold.

**Corollary 2.** *Consider a sequence of problems with $d \to \infty$ for which*

(i) *$F_d(\cdot)$ has a distribution limit $F(\cdot)$ with no point mass at 0, i.e. $F(0) = 0$.*

(ii) *For some $r > 1$, $\beta_j = \sigma\sqrt{2r \log d}$ for a subset of $d_1$ indices, and $\beta_j = 0$ for the rest.*

(iii) *Either $d_1 \to \infty$ and $\text{supp}(\beta)$ is uniformly random, or $\liminf d_1/d > 0$ and $\text{supp}(\beta)$ is arbitrary.*

*Then*

1. *The Bonferroni procedure at any level $\alpha$ has TPR $\to 1$, and*

2. *If $\lambda_1/\log d \to \infty$ then any knockoff method at any level $\alpha$ has TPR $\to 0$.*

Whereas negative results in prior works have focused on specific implementations of knockoffs under restrictive assumptions about the limiting distribution of the covariance matrix, typically under a polynomial sparsity regime ($d_1 \sim d^{1-\gamma}$ for $\gamma \in [0,1)$), Corollary 2 holds uniformly over all possible design choices made by the analyst, makes only mild limiting assumptions about the leading eigenvalue and eigenvector of $\Sigma$, and holds whenever $d_1 \to \infty$.

## 3 Knockoffs for multivariate normal estimators

### 3.1 Implementing the whitening method

As we observed in Section 1.2, the whitening method can be defined without reference to any design matrix $X$. If $\widehat{\theta}$ is any estimator of a $d$-variate parameter $\theta$ with distribution $\widehat{\theta} \sim \mathcal{N}_d(\theta, \Sigma)$ and $\Sigma$ is known, or known up to a scalar multiplier, the whitening method can be applied directly. Theorem 4 in Section 3.2 shows that if $\widehat{\theta}$ is only asymptotically Gaussian with a consistent estimator $\widehat{\Sigma} \xrightarrow{p} \Sigma$, then the whitening method can also be applied directly using $\widehat{\Sigma}$ as a plug-in estimator, and the FDR will be controlled asymptotically (though we note that the result is limited to classical fixed-dimensional asymptotics).

By shifting our focus to the estimator $\widehat{\beta}$, the whitening interpretation suggests easier solutions to closely related regression problems that do not easily lend themselves to introducing knockoff predictor variables. For example, if $\beta$ represents the population-risk-minimizing coefficients in a logistic regression, quantile

regression, or robust regression with Huber loss, we can rely on well-established asymptotic normality results instead of trying to invent a novel framework for generating knockoff predictors in each new example.

Because of the one-to-one correspondence between methods in each formulation, we can even deploy state-of-the-art knockoff implementations, such as those in the `knockoff` package (Patterson and Sesia, 2017), by constructing a pseudo-design matrix and pseudo-response using $\widetilde{\beta}$ and $\xi$.

$$X^* = \begin{pmatrix} A^{1/2} \\ \Delta^{-1/2} \end{pmatrix}, \quad \widetilde{X}^* = \begin{pmatrix} A^{1/2} \\ -\Delta^{-1/2} \end{pmatrix} \quad \text{and} \quad y^* = \begin{pmatrix} A^{-1/2}\xi \\ \Delta^{-1/2}\widetilde{\beta} \end{pmatrix}.$$

Recalling that $A = \Sigma^{-1} - \Delta^{-1}$, and that $2\Delta^{-1}$ is the counterpart of $D$, a short calculation verifies that

$$X^{*\mathsf{T}}X^* = \widetilde{X}^{*\mathsf{T}}\widetilde{X}^* = \Sigma^{-1}, \quad X^{*\mathsf{T}}\widetilde{X}^* = \Sigma^{-1} - 2\Delta^{-1}, \quad \text{and} \quad y^* \sim \mathcal{N}\left(X^*\beta, \ \sigma^2 I_{2d}\right),$$

so the distribution of the rejection set using $[X^* \ \widetilde{X}^*]$ and $y^*$ as inputs to any off-the-shelf knockoffs package is the same as it would be if the problem had been posed as a linear regression to begin with.

The whitening interpretation also suggests simpler methods like estimating $\beta$ and $\sigma^2$, and therefore also $\eta$, by regularized likelihood or Bayesian methods using the known Gaussian likelihood for $\xi$ and $|\widetilde{\beta}|$. Because $\mathrm{sgn}(\widetilde{\beta}_1), \ldots, \mathrm{sgn}(\widetilde{\beta}_d)$ are missing, the log-likelihood is not necessarily concave, but the expectation-maximization (EM) algorithm can be used to impute the missing values. We defer these ideas to future work.

While the coupling in Section 2.1 gives a recipe for generating $\omega$ as a function of $[X \ \widetilde{X}]$ and $y$, we can alternatively generate the noise directly using the residual variance $\widehat{\sigma}^2$. We can even relax the usual dimension requirement that $n \geq 2d$, and require only that $n \geq d + r$ where $r = \mathrm{rank}(\Delta - \Sigma) \leq d$. In that case let $M \in \mathbb{R}^{d \times r}$ be any fixed matrix with $MM^\mathsf{T} = \Delta - \Sigma$, and set

$$\omega = \sqrt{\frac{(n-d)v\widehat{\sigma}^2}{\|\omega'\|^2}} \cdot M\omega', \quad \text{for} \quad \omega' \sim \mathcal{N}_r(0, I_r) \quad \text{and} \quad v \sim \mathrm{Beta}\left(\frac{r}{2}, \frac{n-d-r}{2}\right), \quad (15)$$

or $v = 1$ if $n = d + r$. Here $\omega'$ and $v$ are auxiliary random variables generated independently of the data and each other. Then $\omega$ is independent of $\widehat{\beta}$ because $\widehat{\sigma}^2$ is, and Proposition 6 shows it has the desired distribution.

**Proposition 6.** *Define $\omega', v,$ and $\omega$ as in (15). Then $(n-d)v\widehat{\sigma}^2 \sim \sigma^2\chi_r^2$, and $\omega \sim \mathcal{N}_r(0, \sigma^2(\Delta - \Sigma))$.*

*Proof.* If $n = d + r$ and $v = 1$ then $(n-d)v\widehat{\sigma}^2 \sim \sigma^2\chi_{n-d}^2$ is immediate. If $n > d + r$, independently generate $a_1 \sim \chi_r^2 = \mathrm{Gamma}\left(\frac{r}{2}, 2\right)$ and $a_2 \sim \chi_{n-d-r}^2 = \mathrm{Gamma}\left(\frac{n-d-r}{2}, 2\right)$. Then it is a standard fact that the ratio $v = \frac{a_1}{a_1+a_2} \sim \mathrm{Beta}\left(\frac{r}{2}, \frac{n-d-r}{2}\right)$ is independent of $a_1 + a_2 \sim \chi_{n-d}^2$. As a result,

$$v \cdot (n-d)\widehat{\sigma}^2 \overset{\mathcal{D}}{=} v \cdot (a_1 + a_2)\sigma^2 = \sigma^2 a_1 \sim \sigma^2\chi_r^2.$$

Similarly, $\omega'/\|\omega'\| \sim \mathrm{Unif}(\mathbb{S}^{r-1})$ is independent of $\|\omega'\|^2 \sim \chi_r^2$. As a result,

$$\omega^* = \sqrt{(n-d)v\widehat{\sigma}^2} \cdot \frac{\omega'}{\|\omega'\|} \overset{\mathcal{D}}{=} \sqrt{\sigma^2\|\omega'\|^2} \cdot \frac{\omega'}{\|\omega'\|} \sim \mathcal{N}_r(0, \sigma^2 I_r),$$

and setting $\omega = M\omega^*$ gives the desired result. $\qquad\square$

If we want the software package to generate the knockoff matrix for us, we can also do this as long as $r = d$. Then we can generate

$$X^* = \begin{pmatrix} \Sigma^{-1/2} \\ 0_{d \times d} \end{pmatrix}, \quad \text{and} \quad y^* = \begin{pmatrix} \Sigma^{-1/2}\widehat{\beta} \\ \omega^* \end{pmatrix}, \quad (16)$$

where $\omega^*$ is defined as in Proposition 6; then $X^{*\mathsf{T}}X^* = \Sigma^{-1}$ and $y^* \sim \mathcal{N}_{2d}(X^*\beta, \sigma^2 I_{2d})$ as desired.

If $n > d + r$, then part of $\widehat{\sigma}^2$ is not used up in generating $\omega$:

$$\widetilde{\sigma}^2 = \frac{(n-d)(1-v)}{n-d-r}\widehat{\sigma}^2 \sim \frac{\sigma^2}{n-d-r}\chi_{n-d-r}^2.$$

It is easily shown that $\widehat{\beta}, \omega,$ and $\widetilde{\sigma}^2$ are mutually independent, so allowing the analyst to use $\widetilde{\sigma}^2$ in Stage 2 has no effect on the conditional inference in Stage 3.

The analogous quantity in the standard implementation is the residual variance from the augmented regression of $y$ on $[X \ \widetilde{X}]$. Because $\widetilde{\sigma}^2$ is independent of $[X \ \widetilde{X}]^\top y$, the unordered pair property could be immediately relaxed to a requirement that $W^*$ be a function of the unordered pairs and $\widetilde{\sigma}^2$ without affecting the FDR control proof. We note Chen et al. (2020) also suggested the possibility of using this quantity. One use for $\widetilde{\sigma}^2$ could be as an input to generalized cross-validation (Golub et al., 1979), and another could be as an aid in estimating $\beta$ and $\sigma^2$, as suggested above.

## 3.2 Knockoffs for an asymptotically normal estimator

In this section we give simple sufficient conditions under which the whitening method can be applied to *any* asymptotically normal estimator $\widehat{\theta}_n$ of a $d$-variate parameter $\theta \in \mathbb{R}^d$, along with a consistent covariance estimator $\widehat{\Sigma}_n$. We consider a classical local asymptotic regime in which $\theta = \beta/\sqrt{n}$, for a local parameter $\beta \in \mathbb{R}^d$ and a quantity $n$ that we can think of as a sample size, and assume[4]

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \Rightarrow \mathcal{N}_d(0, \Sigma), \quad \widehat{\Sigma}_n \xrightarrow{p} \Sigma.$$

If we define $\widehat{\beta}_n = \sqrt{n}\,\widehat{\theta}_n$, we have $\widehat{\beta}_n \Rightarrow \mathcal{N}_d(\beta, \Sigma)$. It will be more convenient to work in terms of $\widehat{\beta}_n$ rather than $\widehat{\theta}_n$; note that the null and directional alternative hypotheses are the same if we define them in terms of $\beta$ or $\theta$. For most implementations of knockoffs, there is a preprocessing step effectively standardizing either $\Sigma$ or $\Sigma^{-1}$ to have unit diagonal entries, in which case applying the method to $\widehat{\beta}_n$ or $\widehat{\theta}_n$ would yield identical results.

We assume the analyst carries out knockoffs on $\widehat{\beta}_n$ according the usual whitening method, but plugging in $\widehat{\Sigma}_n$ for $\Sigma$. Let $\widehat{\Delta}_n = \Delta(\widehat{\Sigma}_n) \succeq \widehat{\Sigma}$ denote the analyst's choice of diagonal matrix based on $\widehat{\Sigma}_n$. To generate the whitening noise, the analyst sets $\omega_n = \widehat{M}_n \omega'$, where $\widehat{M}_n = (\widehat{\Delta}_n - \widehat{\Sigma}_n)^{1/2}$ and $\omega' \sim \mathcal{N}_d(0, I_d)$ independently of the data.

Once the noise has been generated, the method proceeds exactly as described in Section 1.2. That is, for

$$\widetilde{\beta}_n = \widehat{\beta}_n + \omega_n, \quad \xi_n = \widehat{\Sigma}_n^{-1}\widehat{\beta}_n - \widehat{\Delta}_n^{-1}\widetilde{\beta}_n,$$

the analyst chooses alternative directions and an ordering based on $\xi_n, |\widetilde{\beta}_n|$ and then carries out Selective SeqStep on the binary $p$-values.

To impose some regularity on the exploratory analysis, we consider knockoff methods for which the exploratory stage can be defined in terms of $d$ non-negative random variables $W_j^+(\xi, \widetilde{\beta})$, along with

$$W_j^-(\xi, \widetilde{\beta}) = W_j^+(\xi, \text{flip}_j(\widetilde{\beta})), \quad \text{and} \quad W_j = \max\left\{W_j^+, W_j^-\right\} \cdot \text{sgn}(W_j^+ - W_j^-), \qquad (17)$$

where $\text{flip}_j : \mathbb{R}^d \to \mathbb{R}^d$ flips the sign of its $j$th argument and leaves the others fixed. The $W$-statistics given in (17) are then used the same way the $W$-statistics are used in the standard formulation of knockoffs: the variables are ordered by $|W_j|$ with $\psi_j = \text{sgn}(W_j) \cdot \text{sgn}(\widetilde{\beta}_j)$, so $\widetilde{p}_j = 1/2 \Leftrightarrow W_j > 0$. It is easy to see that $W_j^* = W_j \cdot \text{sgn}(\widetilde{\beta}_j)$ is a function of $\xi$ and $|\widetilde{\beta}|$ alone, so any statistic $W^+$ defines a valid implementation of the whitening method. This way of defining knockoffs is inspired by the "$\lambda$ signed max" statistic, which we can recover by setting

$$W_j^+ = \max\{\lambda > 0 : \ \widehat{\beta}_j^\lambda \neq 0\},$$

where $\widehat{\beta}^\lambda$ is some regularized estimator of $\beta$ with penalty parameter $\lambda > 0$, possibly in an augmented linear regression model for the response $y^*$ against the design matrix $[X^* \ \widetilde{X}^*]$ as defined above.

Let $\mathcal{R}_\alpha(\widehat{\beta}, \Sigma; \omega)$ denote the $\alpha$-level rejection set for the knockoff method defined by $\Delta(\cdot)$ and $W^+(\cdot)$, as it would be applied to $\widehat{\beta} \sim \mathcal{N}_d(\beta, \Sigma)$ with whitening noise $\omega$; we will typically suppress the randomization variable $\omega$. We say $\mathcal{R}$ is a *continuous knockoffs procedure* if $\Delta(\cdot)$ and $W^+(\cdot)$ are continuous functions of their arguments. In the fixed asymptotic regime under mild conditions, the rejection set converges to its distribution under exact Gaussian sampling, as we show next.

**Theorem 4.** *Assume that we carry out a continuous knockoffs procedure $\mathcal{R}(\cdot)$ on the asymptotically normal estimator*

$$\widehat{\beta}_n \Rightarrow \mathcal{N}_d(\beta, \Sigma), \quad \text{and} \quad \widehat{\Sigma}_n \xrightarrow{p} \Sigma,$$

---

[4]There is nothing special about $\widehat{\theta}$ being $\sqrt{n}$-consistent; for example if we had $n^{1/4}(\widehat{\theta} - \theta) \Rightarrow \mathcal{N}(0, \Sigma)$ we could just as well define the local parameter $\beta = n^{-1/4}\theta$ and repeat the argument in this section.

*where $\Sigma$ is nonsingular. Let*

$$E = \{(\xi, \widetilde{\beta}) : \ W_1^+, W_1^-, \dots, W_d^+, W_d^- \text{ are all positive and distinct}\},$$

*and assume that $E^c$ has Lebesgue measure zero. Then we have for any $\alpha$,*

$$\mathcal{R}_\alpha \left( \widehat{\beta}_n, \ \widehat{\Sigma}_n; \ \omega_n \right) \ \Rightarrow \ \mathcal{R}_\alpha \left( \widehat{\beta}, \ \Sigma; \ \omega \right),$$

*where $\widehat{\beta} \sim \mathcal{N}_d(\beta, \Sigma)$ and $\omega \sim \mathcal{N}_d(0, \Delta(\Sigma) - \Sigma)$ independently. In particular, the FDR and TPR of $\mathcal{R}_\alpha$ converge to their values under exact normal sampling.*

As an immediate application of Theorem 4, we see that continuous knockoff methods can be applied to any asymptotically normal maximum likelihood estimator.

**Example 2** (Maximum likelihood estimation). *Suppose we observe a sample $X_1, \dots, X_n$ from a parametric family $\mathcal{P} = \{p_\theta : \ \theta \in \Theta \subseteq \mathbb{R}^d\}$, where $p_\theta$ is a density with respect to a common dominating measure $\mu$. Define the maximum likelihood estimator (MLE)*

$$\widehat{\theta}_n \ = \ \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_\theta(X_i)$$

*Assume further that $\mathcal{P}$ is sufficiently regular so that for $\theta$ in the interior of $\Theta$,*

$$\sqrt{n} \left( \widehat{\theta}_n - \theta \right) \ \Rightarrow \ \mathcal{N}_d \left( 0, J(\theta)^{-1} \right),$$

*where $J(\theta)^{-1}$ is the Fisher information at $\theta$. If $J(\theta)$ is nonsingular and continuous at $\theta$, and $\mathcal{R}(\cdot)$ is continuous, then by Theorem 4, $\mathcal{R}_\alpha(\widehat{\theta}_n, J(\widehat{\theta}_n)^{-1})$ asymptotically controls the FDR at level $\alpha$.*

One interesting special case of Example 2 is the logistic regression model where we observe predictors $x_i \in \mathbb{R}^d$ and binary response $y_i$, and model logit $\mathbb{P}(y_i = 1 \mid x_i) = x_i^\mathsf{T}\beta$ for some $\beta \in \mathbb{R}^d$. Then, just as in linear regression, we have a design matrix $X \in \mathbb{R}^{n \times d}$ and response $y \in \{0,1\}^n$. If we think of knockoffs in terms of the standard formulation, we might be tempted to try generalizing it by adding some matrix $\widetilde{X}$ of predictor variables and fitting a logistic regression on the augmented matrix $[X \ \widetilde{X}]$. But it is unclear how to design such a matrix *a priori*, since the MLE $\widehat{\beta}_n$ solves the weighted least squares problem

$$\widehat{\beta}_n \ = \ \left( X^\mathsf{T} \widehat{V}_n X \right)^{-1} X^\mathsf{T} \widehat{V}_n y, \quad \text{where} \ \ \widehat{V}_n \ = \ \mathrm{diag}\left( \mathrm{Var}_{\widehat{\beta}_n}(y_i) \right).$$

In particular, it is the predictor variables' *weighted* correlations with each other that determine the asymptotic distribution of $\widehat{\beta}_n$, not the unweighted correlations as in linear regression. But to design a knockoff predictor $\widetilde{X}_j$ whose weighted correlations with $X_{-j}$ are the same as those of the real predictor $X_j$, we would need to know $\widehat{\beta}$, which determines the weights, ahead of time.

By contrast, if we view the MLE $\widehat{\beta}$ for linear regression as an asymptotically Gaussian random vector, we can simply treat it just as we would the OLS estimator for a linear regression, and treat the variance estimator $\widehat{\Sigma}_n = (X^\mathsf{T} V_{\widehat{\beta}_n} X)^{-1}$ as a consistent plug-in estimator for the asymptotic variance.

We caution, however, that asymptotic normal approximations can be unreliable in high-dimensional settings, particularly when $d/n \to 0$. See e.g. El Karoui et al. (2013) and Donoho and Montanari (2016) regarding the limitations of asymptotic normal approximation in high-dimensional M-estimation; El Karoui and Purdom (2018) also cast doubt on the accuracy of the bootstrap in high dimensions. In particular Sur and Candès (2019) showed that logistic regression is unreliable in high dimensions even with five to ten observations per predictor variable. As a result, we recommend applying this method only in settings where $n \gg d$ so the multivariate Gaussian approximation is reliable; determining the full range of application beyond the classical setting is outside the scope of this work.

## 3.3 Example: Analysis of stock market data

We now use an example from stock market data to illustrate the implementation of fixed-X knockoff when the test statistics are asymptotically multivariate Gaussian. Our goal is to test whether the daily returns of a given stock are predictable based on the one-day-lagged returns of the overall market, as measured by the

Standard and Poors (S&P) 500 index. We use S&P 500 data from February 2013 to February 2018[5], totalling 1259 trading days, and restrict our analysis to the 470 stocks that belonged to the S&P 500 over the entire period considered. For each stock, we test whether its excess returns relative to the market are correlated with the one-day-lagged market returns.

For $j = 1, \ldots, d = 470$, let $y_j(t)$ denote the price of stock $j$ on day $t$, and let $y_0(t)$ denote the level of the S&P 500 index on day $t$. Then for $j = 0, 1, \ldots, 470$, the return for stock $j$ on day $t$ is

$$r_j(t) = \frac{y_j(t)}{y_j(t-1)} - 1.$$

Our parameter of interest for stock $j$ is the Spearman (rank) correlation of its excess return, relative to the market, with the one-day-lagged market return:

$$\rho_j = \text{Corr}_S(e_j(t), r_0(t-1)), \quad \text{where} \quad e_j(t) = r_j(t) - r_0(t),$$

and $\text{Corr}_S$ denotes the Spearman correlation. Roughly speaking, if $\rho_j > 0$ then stock $j$ is a "good bet" today if the market performed well yesterday and a "bad bet" otherwise, and if $\rho_j < 0$ the reverse is true. Because economic theory suggests stock returns should be unpredictable, we expect most of the true correlations to be close to 0. We will test $H_j : \rho_j = 0$ against the two-sided alternative for $j = 1, \ldots, d$.

We define the problem in terms of excess returns $e_j(t)$ instead of "raw" returns $r_j(t)$ for two reasons: First, predictability of excess returns is more easily translated into a trading strategy, since an investor could hedge their bets against the market, giving another reason for most $\rho_j$ to be small. Second, the pairwise correlation between stock returns for any two stocks is usually positive, since good or bad news about the economy at large tends to move all stocks in the same direction. As a result, the leading eigenvalue and eigenvector of $\text{Var}(r(t))$ mimics the leading eigenvalue and eigenvector for MCC, giving knockoffs no chance to perform well. The common market factor is so strong, accounting for almost half the total variance, that it is doubtful whether FDR is even the right error rate to control (Efron, 2007; Schwartzman and Lin, 2011; Kluger and Owen, 2021). By contrast, the excess returns do not exhibit the same sort of positive correlation because we have removed it by subtracting off the market index return.

To test the hypotheses, we calculate the sample Spearman correlations $\hat{\rho}_1, \ldots, \hat{\rho}_d$, and use the bootstrap to obtain a nonparametric estimator $\hat{\Sigma}$ for the covariance matrix $\Sigma = \text{Var}(\hat{\rho})$; which has similar structure to $\text{Var}(e(t))$. Because the autocorrelations of $r_j(t)$ are small, we use the i.i.d. bootstrap, but we obtain similar rejection sets if we use the block bootstrap for block length 15 or 50 (blocks of three or ten weeks respectively). The bootstrap resampling is repeated 10,000 times. By inspecting the normal Q-Q plot of the sample Spearman correlation across bootstrapped samples, we find that the asymptotic Gaussian approximation fits the bootstrap distribution reasonably well even in the tails where rejections occur, at least marginally for each stock. We use the formulation in the previous section to implement the fixed-$X$ knockoff treating $\hat{\rho}$ as an asymptotically Gaussian estimator of $\rho$. We use the lasso $\lambda$ signed-max method with SDP knockoffs.

The leading eigenvalue of the bootstrap correlation matrix $\widehat{\text{Corr}}(\hat{\rho})$ is 53, accounting for about 11% of the total variance, and the entries of the leading eigenvector are dense but roughly symmetrically distributed around zero, ranging from $-0.15$ to $0.15$. Thus, the eigenstructure does not rule out a well-designed implementation of knockoffs performing well but it may nonetheless be worrisome.

Given the estimated correlation matrix, we use the procedure in Section 4.2 to implement the fixed-X knockoff. Since this requires generating additional Gaussian noise, we repeat the procedure 30 times to obtain more stable rejection sets, and report hypotheses that are rejected for at least 50% of times. This stability selection criteria is similar to that proposed in (Ren et al., 2020). We found that at FDR level $\alpha = .2$, knockoff obtains no rejections for 26 out of 30 trials. The other 4 times, there are 7 rejections on average, and the stocks BA, BRK.B, MDLZ and UAA are always rejected. As such, no stocks meet the stability selection threshold. By contrast, both the BH and the Bonferroni correction makes a number of rejections, with the BH making 11. The stocks rejected by the BH and the Bonferroni correction are summarized in Table 2.

Figure 2 shows the $W$-statistics and along with $\widehat{\text{FDP}}_k$ from a representative trial. We find that there are many negative $W$ statistics in the front of the list, indicating that knockoff will not have many rejections even if the significance level $\alpha$ is reasonably relaxed. We see in Section 5.3 that we could have anticipated this underperformance by inspecting the $\Delta$ values.

---

[5]Source: https://www.kaggle.com/camnugent/sandp500

Table 2: Rejections of different methods at different target FDR levels, for testing the null hypotheses that the returns of each stock are correlated with the lag-1 return of the S&P 500 index. "Additional rejections" are rejections that are made at $\alpha = 0.2$ but not $\alpha = 0.1$.

| Method | Rejections at $\alpha = .1$ (stock tickers) | Additional rejections at $\alpha = .2$ |
|--------|---------------------------------------------|----------------------------------------|
| BH | 7 (AVGO, BA, BRK.B, JEC, MOS, SWKS, UAA) | 4 (FISV, KORS, V, WAT) |
| Bonferroni | 3 (BRK.B, JEC, MOS) | 0 |
| Fixed-$X$ knockoff | 0 | 0 |
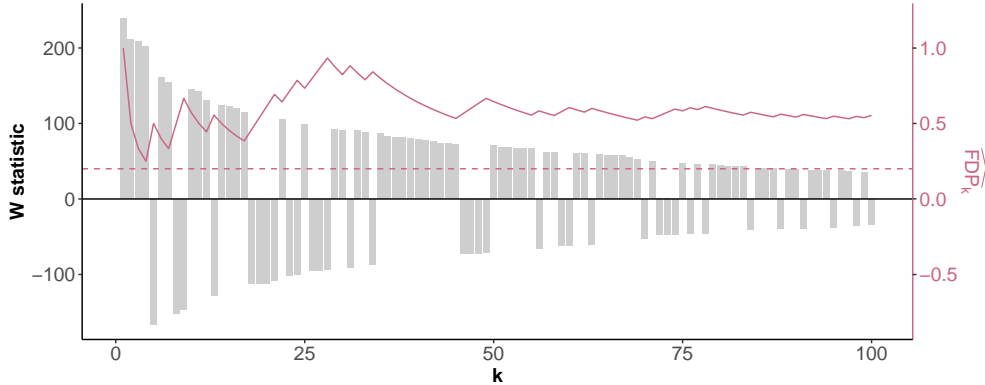


Figure 2: The 100 largest $W$ statistics of the S&P 500 example, ranked by absolute value, and the estimated FDP $\widehat{\mathrm{FDP}}_k$ from the Selective Seqstep procedure. Even if we use the more liberal FDP estimator without the "1+," as plotted here, knockoffs is unable to make rejections at $\alpha = 0.2$, represented by the horizontal line.

# 4 Numerical results

## 4.1 Fixed knockoff for Gaussian linear model

In this section, we simulate under the Gaussian linear model to compare scenarios where the TPR of the best achievable knockoff method is close to zero against similar scenarios where knockoff methods perform well. For the design matrix $X \in \mathbb{R}^{n \times d}$, we generate random matrices whose rows are generated i.i.d from $N(0, K), K \in \mathbb{R}^{d \times d}$. We consider the following two regimes:

(a) **Positively equi-correlated OLS estimator** $K^{-1}$ is an equicorrelation matrix with correlation $\rho = 0.2$, i.e. $K_{ii}^{-1} = 1, 1 \leq i \leq d$, and $K_{ij}^{-1} = 0.2, 1 \leq i < j \leq d$;

(b) **Positively equi-correlated covariates** $K$ is an equicorrelation matrix with correlation $\rho = 0.2$.

In regime (a), the columns of $X$ are negatively correlated and the OLS test statistics are positive correlated, and we expect that any knockoff methods will have trivial power. In regime (b), the columns of $X$ are positively correlated and the OLS test statistics are negatively correlated.

We choose $n = 3000$ and $d = 1000$. For each $K$, we fix one realization of the random matrix, and then normalize its columns to obtain the design matrix $X$. Next, we generate the response $y \in \mathbb{R}^d$ as follows. First, to define $\beta$, we choose $s = 30$ coefficients uniformly at random and let $\beta_j = 5$ for each of the selected coefficients. We then generate $y = X\beta + \varepsilon$, where $\varepsilon_i$ are i.i.d standard normal errors. The above data generating procedure is repeated 600 times for each $K$.

For each design matrix $X$, we generate the knockoff matrix $\widetilde{X}$ using the MVR-knockoff (Spector and Janson, 2020) and SDP-knockoff algorithms (Barber and Candès, 2015). For each instance of $(X, \widetilde{X}, y)$, we first consider the knockoff* procedure introduced in Section 2.2, which is the best achievable knockoff method once the knockoffs have been constructed, but is infeasible since it can only be carried out with knowledge of the true coefficients $\beta$. We also consider a practically feasible knockoff method which uses the maximum lasso penalty level as test statistic.

In addition, we include the T3-knockoff* "procedure" from Section 2.5, a simulation-based bound of any knockoff method under the assumptions of Theorem 3, where the analyst is not allowed to know the nonzero indices of $\beta$ at the time of determining the knockoff matrix $\widetilde{X}$ (or equivalently, determining $\Delta$). $\beta$ is still revealed to the analyst immediately after the knockoffs are generated so the analyst can carry out the knockoff* method. T3-knockoff* is defined in Algorithm 1 and uniformly upper bounds the power achievable by SDP-knockoffs or MVR-knockoffs in any setting where $\beta$ is random with an exchangeable distribution over the $d$ indices, so it is more optimistic than SDP-knockoff* or MVR-knockoff*.

Finally, we consider the BH procedure and the Bonferroni test on OLS $p$-values for baseline comparison. Note that the BH procedure is not theoretically guaranteed to control the FDR at the desired level unless the columns of $X$ are orthogonal.

We obtain the SDP-knockoff matrix using the `knockoff` package (Patterson and Sesia, 2017), and the MVR-knockoff using the `knockpy` package (Spector and Janson, 2020). For both knockoff matrices, we then compute the the maximum lasso penalty level statistics with the `knockoff` package. Overall, the performance of knockoff shows a stark contrast under these two regimes, while the performance of BH and Bonferroni appear to be much more stable. Figure 3 shows the power of the BH, the Bonferroni and the SDP-knockoff tests. For reference, we also include the case where the design matrix is i.i.d. Gaussian. We find that the SDP-knockoff outperforms the BH method when the covariates are independent. However, we see that when the OLS test statistics are positively correlated, the TPRs of the oracle knockoff methods are close to zero. Table 3 shows the FDR and TPR of all methods for targeted FDR level $\alpha = 0.1$ and 0.2. We find that the TPR of SDP-knockoff* is smaller than 0.02 when controlling the FDR at 0.2.

We can understand the performance of knockoffs in each case by plotting the sorted log-odds values $\eta_{(j)}$, as we do in Figure 4. While the knockoff* method can achieve the sorting pictued in Figure 4, feasible knockoff methods cannot achieve a perfect ordering because $\beta_j/\sigma$ is unknown. Recall that when the covariance matrix of the test statistics has factor model structure, the whitening step of knockoff destroys virtually all the information, and the log-odds of observing small $p$-values in the inference cannot rise above $-\log \alpha$. This is again confirmed by Figure 4, which shows the average of the $s = 30$ largest log-odds across different trials. The rest of the log-odds are zero since the corresponding coefficient $\beta_j$ is zero. In particular, we find that most of the log-odds are smaller than $-\log \alpha$ with $\alpha = 0.1$. Thus by Proposition 4, the number of rejections of any knockoffs method must be small.

## 4.2 Knockoff for multivariate Gaussian statistics

In Section 1.2, we reinterpreted the knockoff method and generalized the fixed-$X$ knockoff procedure to multivariate normal test statistics. Here we use simulations to investigate the performance of different methods for testing the means of multivariate normal, and hint at the possible use cases and limitations of knockoffs for such problems.

We generate $d = 1000$ dimensional multivariate Gaussian vectors $\widehat{\theta} \sim \mathcal{N}_d(\theta, K)$. The mean vector $\theta$ is generated in the same way as the linear coefficient $\beta$ in the previous simulation, except that the non-zero means $\theta_j$ are set to 3.5. We consider two types of covariance matrices:

(a) $K$ **has factor model structure**. In particular, we let

$$K = I_d + \lambda \sum_{\ell=1}^{k} u_\ell u_\ell^T, \tag{18}$$

where $u_\ell$ are drawn independently from the uniform sphere.

(b) $K^{-1}$ **has factor model structure**, i.e.

$$K^{-1} = I_d + \lambda \sum_{\ell=1}^{k} u_\ell u_\ell^T. \tag{19}$$

In particular, we choose $k = 2$ and $\lambda \in \{20, 100\}$. We normalize the covariance matrix $K$ such that $K$ has unit diagonal entries. After normalization, the largest eigenvalue of $K$ is approximately 19 and 75 for $\lambda = 20$ and 100, respectively (the sum of all eigenvalue is $d = 1000$). For each setup, we consider one fixed realization of the random matrix $K$. In each trial of our simulation, we first generate the mean vector $\theta$ and then the multivariate normal vector $\widehat{\theta}$.

|  |  | SDP-knockoff | | MVR-knockoff | | Other methods | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Knockoff* | Maximum penalty level | Knockoff* | Maximum penalty level | BH | Bonferroni |
| $\alpha = 0.1$ | FDR | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.01 |
|  | TPR | 0.00 | 0.00 | 0.00 | 0.00 | 0.65 | 0.35 |
| $\alpha = 0.2$ | FDR | 0.00 | 0.01 | 0.00 | 0.00 | 0.18 | 0.02 |
|  | TPR | 0.03 | 0.00 | 0.02 | 0.01 | 0.76 | 0.42 |

(a) Positively equi-correlated OLS estimator

|  |  | SDP-knockoff | | MVR-knockoff | | Other methods | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Knockoff* | Maximum penalty level | Knockoff* | Maximum penalty level | BH | Bonferroni |
| $\alpha = 0.1$ | FDR | 0.06 | 0.07 | 0.05 | 0.06 | 0.10 | 0.01 |
|  | TPR | 0.96 | 0.44 | 0.95 | 0.44 | 0.68 | 0.36 |
| $\alpha = 0.2$ | FDR | 0.16 | 0.17 | 0.15 | 0.16 | 0.19 | 0.01 |
|  | TPR | 0.97 | 0.77 | 0.96 | 0.74 | 0.78 | 0.43 |

(b) Positively equi-correlated covariates

Table 3: FDR and TPR of different methods under different target FDR levels.

For each matrix $K$, we repeat the above data generating procedure $N = 600$ times. Again, we consider both the SDP knockoff and the equi-correlated knockoff to create the diagonal matrix $\Delta$ that satisfies $\Delta \succeq K$. For each diagonal matrix $\Delta$, we first implement the oracular ordering and the associated knockoff* procedure defined in Section 2.2. For each observed $\widehat{\theta} \sim \mathcal{N}(\theta, K)$, we generate a pseudo-design matrix $X^*$ and pseudo-response $y^*$ as in (16).

Table 4 shows the FDR and TPR of different methods for $\alpha = 0.2$. As expected, we found that the power of even the best achieveable knockoff method is less than that of Bonferroni when the covariance matrix $K$ has a factor model structure with reasonably large leading eigenvalue. This suggests that the knockoff-type approaches suffer from severe power loss when applied to general test statistics with factor model structure. However, when the precision matrix $K^{-1}$ has a factor model structure, it may be possible to use the knockoff framework procedure to design a test with superior TPR than baseline methods such as the BH. We leave this for future research.

## 5 Discussion

### 5.1 Knockoffs, randomized responses, and the selection-inference tradeoff

Viewing knockoffs as a conditional post-selection inference method sheds light on what Fithian et al. (2014) called the *selection-inference tradeoff*: that the more we condition on, the less data remains for confirmatory analysis. This tradeoff is most obviously apparent in the case of data splitting, where the analyst selects a model or hypotheses to test by means of some exploratory analysis using a fraction of the data points, and then carries out confirmatory inference using only the remaining fraction, i.e. inference conditions on the initial data set (this assumes the two data sets are independent; otherwise data splitting may be invalid). A similar phenomenon is present in other conditional post-selection inference problems, where whatever statistics of the data we observe in the selection (exploratory) stage are unavailable as inferential evidence in the conditional inference (confirmatory) stage.

Compared to most other conditional inference methods, knockoffs conditions on much more about the data, holding out only the randomized and binarized $\text{sgn}(\widetilde{\beta})$ for confirmatory inference. What is more, because $\text{sgn}(\widetilde{\beta}_j) \sim \text{Unif}\{-1, +1\}$ under $H_j$, the binary conditional $p$-values $\widetilde{p}_j$ can never be smaller than $1/2$. The reason knockoff methods are nevertheless able to compete with and sometimes outperform other state-of-the-art multiple testing methods is because the FDR is an aggregate error criterion: to control it, knockoffs need never be confident about rejecting any individual hypothesis, only about the fraction of nulls early in the list. As a result, no individual $\widetilde{p}_j$ needs to be minuscule, so long as most of the highly prioritized ones are $1/2$.
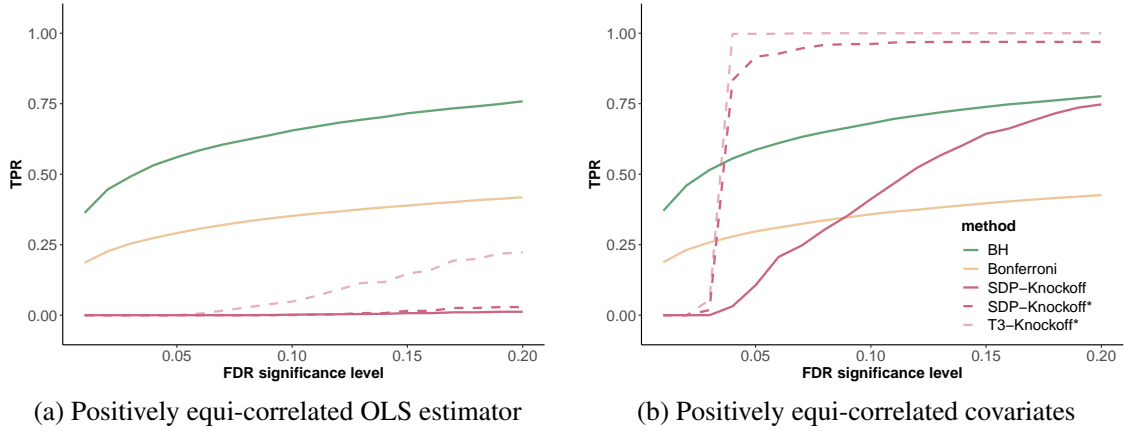
(a) Positively equi-correlated OLS estimator

(b) Positively equi-correlated covariates
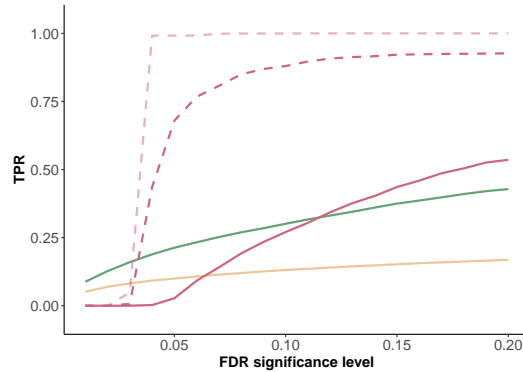
Reference: i.i.d Gaussian covariates with non-null $\beta_j = 3.5$.

Figure 3: TPR of different tests under different FDR significance levels. Given the SDP-knockoff matrix, the SDP-knockoff* test is the best achievable knockoff method and can only be implemented with oracle knowledge about $\beta$. SDP-knockoff is a practically feasible knockoff method which uses the maximum lasso penalty level as the test statistics. The T3-knockoff* method defined in Algorithm 1, and provides an upper bound on any knockoff methods where the analyst is not allowed to know the nonzero indices of $\beta$ at the time of determining the knockoff matrix $\widetilde{X}$. In particular, the T3-knockoff* is more optimistic than the oracle procedure SDP-knockoff*. In Figure (a), the covariates in the design matrix are positively correlated with correlation approximately 0.2. In Figure (b), the covariates in the design matrix are negatively correlated, and the entries of the OLS estimate $\widehat{\beta}$ are positively correlated with correlation approximately 0.2. For reference, we also show the case where the covariates in the design matrix are i.i.d Gaussian, and with the non-null $\beta = 3.5$, a regime considered by Barber and Candès (2015).

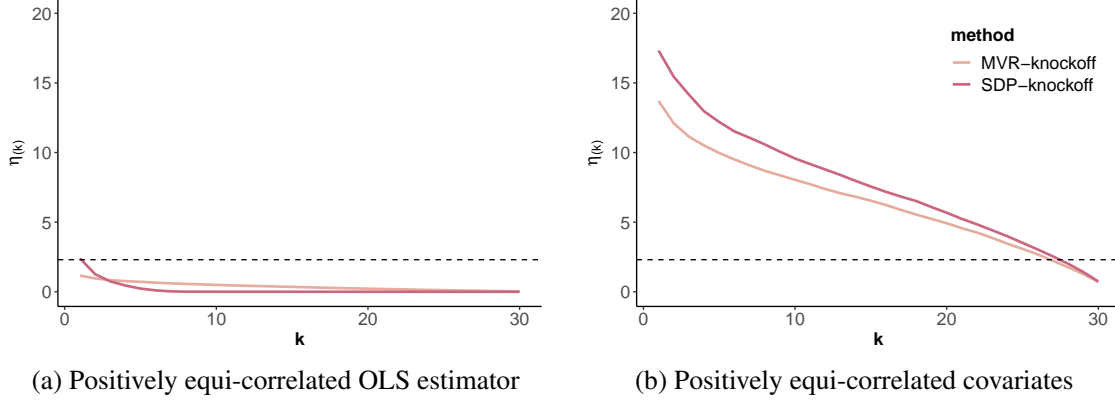(a) Positively equi-correlated OLS estimator    (b) Positively equi-correlated covariates

Figure 4: The log-odds of observing a small p-value in the inference stage. The log-odds are computed as Equation 8. Only the $s = 30$ non-zero log-odds are shown in both figures. $\eta_{(k)}$ denotes the $k$th largest log-odd and $\eta_{(k)} = -\log 0.1$ is shown in the dashed black line.

|  |  | SDP-knockoff | | MVR-knockoff | | Other methods | |
|---|---|---|---|---|---|---|---|
|  |  | Knockoff* | Maximum penalty level | Knockoff* | Maximum penalty level | BH | Bonferroni |
| $\lambda = 20$ | FDR | 0.00 | 0.04 | 0.03 | 0.05 | 0.19 | 0.01 |
|  | TPR | 0.54 | 0.18 | 0.70 | 0.36 | 0.77 | 0.42 |
| $\lambda = 100$ | FDR | 0.00 | 0.02 | 0.00 | 0.00 | 0.18 | 0.01 |
|  | TPR | 0.21 | 0.05 | 0.23 | 0.07 | 0.77 | 0.41 |

(a) When $K$ has factor model structure

|  |  | SDP-knockoff | | MVR-knockoff | | Other methods | |
|---|---|---|---|---|---|---|---|
|  |  | Knockoff* | Maximum penalty level | Knockoff* | Maximum penalty level | BH | Bonferroni |
| $\lambda = 20$ | FDR | 0.18 | 0.20 | 0.18 | 0.19 | 0.19 | 0.02 |
|  | TPR | 1.00 | 0.71 | 0.99 | 0.70 | 0.76 | 0.41 |
| $\lambda = 100$ | FDR | 0.18 | 0.15 | 0.19 | 0.17 | 0.19 | 0.02 |
|  | TPR | 0.99 | 0.45 | 0.99 | 0.48 | 0.77 | 0.41 |

(b) When $K^{-1}$ has factor model structure

Table 4: FDR and TPR of different methods for testing means of multivariate Gaussian with correlation matrix $K$ defined in Equations 18 and 19. Target FDR level $\alpha = 0.2$. "Maximum Penalty Level" refers to the method where we first generate artificial design matrix and response $(X, y)$, and then apply the fixed-X knockoff.

By giving up on making each $\widetilde{p}_j$ powerful, knockoffs is able to use nearly all of the information in $\widehat{\beta}$ to supercharge the more flexible exploratory stage, betting on its ability to pack the front of the priority list with non-null hypotheses. This strategy of "betting on exploration" can pay off especially handsomely when Bayesian priors or structural assumptions like sparsity can be brought to bear during the exploration, which can use them in an unfettered way.

The whiteout phenomenon we describe here is an example of where that bet goes wrong, leaving too little information for inference. The fundamental problem is that the inference engine, Selective SeqStep requires *independent* binary $p$-values, which can only be created by adding enough noise to make $\mathrm{Var}(\widetilde{\beta})$ diagonal. If $\Sigma$ is "too far from diagonal" in the sense we describe, then this cannot be done without destroying the signal.

It may seem counterintuitive that adding more noise (larger $\Delta$) means using up more data for selection and leaving less for inference. In most methods that use randomized data for exploration, such as Tian and Taylor (2018), the opposite is true: adding more noise hides more information from the selection algorithm, preserving it for confirmatory inference. The difference is that, in knockoffs, the information "left over" after randomizing $\widetilde{\beta}$ is also given to the analyst at exploration time, in the form of $\xi$. Instead it is the randomized $\widetilde{\beta}$ that is (partly) held out for inference.

We could equivalently define the whitening method in terms of Gaussian noise

$$\nu \;=\; -\Delta^{-1}\omega \;\sim\; \mathcal{N}_d\left(0, \sigma^2\Delta^{-1}\right), \quad \text{with} \quad \xi \;=\; A\widehat{\beta} + \nu$$

viewed as a noisy version of $A\widehat{\beta}$, and $\widetilde{\beta} = \Delta\Sigma^{-1}\widehat{\beta} - \Delta\xi$ viewed as the residual information. From this perspective, the "noise variance" is $\Delta^{-1}$, so that more noise (smaller $\Delta$) once again means holding out more information in the form of $\widetilde{\beta}$.

## 5.2 The whitening interpretation and Spector and Janson (2020)

Our whitening interpretation sheds potentially interesting light on the phenomenon recently discovered by Spector and Janson (2020) when $X^\mathsf{T}X = \Sigma^{-1}$ is an equicorrelated covariance matrix with diagonal entries equal to 1 and off-diagonal entries equal to $\rho \geq 0.5$. In this example the maximum eigenvalue of $\Sigma$ is 1, so there is no "whiteout" problem, but the authors find that both equicorrelated and SDP knockoffs struggle to make any rejections. To understand why, note that both methods would set $D = 2(1-\rho)I_d$, so $\Delta = (1-\rho)^{-1}I_d$ and $A = \rho\mathbf{1}_d\mathbf{1}_d^\mathsf{T}$. As a result, in the exploratory analysis the analyst observes

$$\xi \;\sim\; \mathcal{N}\left(\rho\sum_{j=1}^{d}\beta_j, \; \rho\sigma^2\right)\cdot\mathbf{1}_d, \quad \text{and} \quad |\widetilde{\beta}_j| \;\overset{\text{ind.}}{\sim}\; \frac{1}{1-\rho}|\mathcal{N}(\beta_j, \sigma^2)|.$$

Because only $\sum_j \beta_j$ and $|\beta|$ are identifiable from the exploratory data set, the analyst has no way to make an educated guess $\psi_j$ about $\mathrm{sgn}(\beta_j)$. This problem can be resolved by choosing $D$, or $\Delta$, more judiciously, as Spector and Janson (2020) show.

## 5.3 Power analysis in knockoffs

One takeaway message of Section 2 is the crucial role played by the threshold $-\log\alpha$ in knockoffs methods' performance. If

$$\mu_j \;=\; \frac{2\beta_j^2}{\sigma^2\Delta_{jj}} \;<\; -\log\alpha, \tag{20}$$

then we are more likely than not to observe $\eta_j < -\log\alpha$, since $\eta_j \sim \mathcal{N}(\mu_j, 2\mu_j)$. In that case, even if we guess the alternative direction $\psi_j$ right we will still have

$$\mathbb{P}(W_j < 0) \;>\; \alpha\mathbb{P}(W_j > 0), \quad \Longleftrightarrow \quad \mathbb{P}(\widetilde{p}_j = 1) \;>\; \alpha\mathbb{P}(\widetilde{p}_j = 1/2), \tag{21}$$

so variable $j$ will be a net drag on the FDP estimator's struggle to remain below $\alpha$.

This observation can help us to do rudimentary power analysis at various stages of the procedure. For example, before we observe anything about the response vector $y$, we can inspect the diagonal entries of the matrix $\Delta = 2D^{-1}$ and ask how large $\beta_j/\sigma$ would have to be for us to have reasonable power to detect variable $j$. Doing a little algebra on inequality (20), we arrive at

$$\mu_j \;=\; \frac{2\beta_j^2}{\sigma^2\Delta_{jj}} \;<\; -\log\alpha \quad \Longleftrightarrow \quad |\beta_j/\sigma| \;<\; \sqrt{\frac{-\Delta_{jj}\log\alpha}{2}} \;=\; \sqrt{\frac{-\log\alpha}{D_{jj}}}. \tag{22}$$

We can think of (22) as giving a critical threshold for the SNR of variable $j$. For example, suppose $\alpha = 0.05$, so $-\log \alpha \approx 3$. Then if $\Delta_{jj} = 6$ (or equivalently $D_{jj} = 1/3$), $|\beta_j/\sigma|$ should be larger than 3 if we want $\eta_j$ to be above $-\log \alpha$ most of the time. Likewise, if $\Delta_{jj} = 32/3 \leq 11.7$ ($D_{jj} = 3/16 = 0.1875$), then the critical SNR threshold is about 4 for $\alpha = 0.05$.

Importantly, because this variable-by-variable power analysis can be done *before* we observe anything about $y$, we can still change course if we don't like the $\Delta_{jj}$ values we get — we could either choose a different knockoff matrix or abandon the knockoffs framework and use BH instead, without any threat to either method's FDR control guarantees.

Whereas our theoretical results emphasize lower bounds on $\Delta$, for an analyst intending to use knockoffs the more interesting question is how large each $\Delta_{jj}$ actually is in the specific knockoff matrix they are about to use for their problem. Until more is understood about what regimes lead knockoffs to dominate BH or vice versa, we recommend that analysts at least inspect the $\Delta_{jj}$ values in light of these SNR thresholds as a diagnostic tool. If, say, only a few of the $\Delta_{jj}$ values are below 10, then the matrix $\Sigma$ may not be a suitable problem structure for knockoffs. In our stock market example, only two $\Delta_{jj}$ values are below 6 and only 50 are below 11.7, suggesting that only very strong signals have a good chance of generating rejections.

## 5.4 Concluding remarks

We emphasize once again that the results we derive for these asymptotic regimes do not imply that fixed-$X$ knockoffs are underpowered as a general rule. On the contrary, we believe our results are interesting precisely because the opposite is true: there are many problems where existing fixed-$X$ knockoff methods outperform all other known FDR-controlling methods. In particular, BH and knockoffs represent two completely different approaches to multiple testing in regression or with multivariate normal test statistics. To give practitioners appropriate guidance about which one to use, more work is needed to answer several crucial questions: When do knockoff methods outperform the BH procedure, and which implementations perform the best? How can practitioners recognize which is better for their context? Can hybrid methods such as those of Sarkar and Tang (2021), or methods yet to be developed, balance the tradeoffs between the two approaches, preserving the strengths of the knockoffs framework without suffering its drawbacks? By identifying pitfalls for the knockoffs framework our results represent strides toward a more complete understanding of multiple testing in the linear model.

# 6 Proofs

## 6.1 Proof of Proposition 2

**Proposition 2.** *Conditional on $\xi$ and $|\widetilde{\beta}|$, both the TPP and the expected total number of rejections for the knockoff\* procedure are stochastically larger than they are any other implementation with the same $\Delta$. Consequently, the knockoff\* procedure maximizes both the TPR and the expected number of rejections.*

*Proof.* If $d_1 = 0$ there is nothing to prove, so assume $d_1 \geq 1$. Given $\xi$, $|\widetilde{\beta}|$, and $\psi$, the conditional $p$-values are independent with
$$\text{logit } \mathbb{P}(\widetilde{p}_j = 1/2) \;=\; \eta_j \cdot \text{sgn}(\psi_j \beta_j).$$
If we hold the ordering fixed, the rejection set is stochastically increasing in each of the above log-odds, so the TPP is always made stochastically larger by setting $\psi_j = \text{sgn}(\beta_j)$ whenever $\beta_j \neq 0$. We can therefore restrict our attention to the case where the log-odds for each variable is $\eta_j \geq 0$.

Next, define the indicator $E_k = 1\{\widehat{\text{FDP}}_k^{\text{wh}} \leq \alpha\}$. If $\eta_{[j]} < \eta_{[j+1]}$, then swapping the two leaves $E_k$ fixed for all $k \neq j$, but increases the conditional probability that $E_j = 1$ given $E_{-j}$. Therefore, $\hat{k}$ is stochastically largest when the log-odds are arranged in decreasing order. The same is true for the number of rejections $R_{\hat{k}} = \left\lceil \frac{1+\hat{k}}{1+\alpha} \right\rceil$.

Likewise, if $\eta_{[j]} < \eta_{[j+1]}$ then $H_{[j+1]}$ must be non-null, so conditional on $E_{-j}$ the number of rejected non-null hypotheses is also made stochastically larger by arranging the log-odds in decreasing order. $\square$

## 6.2 Proof of Proposition 4

For any $\alpha$ and $\delta$, we define

$$p = \frac{\alpha}{1+\alpha}, \quad q_\delta = \frac{\alpha+\delta}{1+\alpha+\delta}.$$

We will now prove the following proposition, which is stronger and more precise than Proposition 4.

**Proposition 7.** *Suppose that*

$$\eta_{(k^*)} < -\log(\alpha+\delta), \delta > 0,$$

*where $k^* \geq 1/\alpha$. Define*

$$C_1(\alpha,\delta) = \frac{1}{1+\alpha}\left[\max\left\{1, \frac{4\alpha(1+\alpha+\delta)}{\delta}\right\} + 1\right],$$

*and*

$$C_2(\alpha,\delta) = \frac{1}{1+\alpha}\left[\frac{e}{4\sqrt{\pi}}\frac{\sqrt{1+\alpha}q_\delta}{\sqrt{p(1-p)}}c_h^{-3/2}\min\left\{\frac{2(q_\delta-p)}{p(1-q_\delta)}, 1\right\} + 2\right],$$

*where*

$$\lambda^* = p + \frac{q_\delta - p}{4} < q_\delta, \quad and \quad c_h = -\left[\lambda^*\log\frac{q_\delta}{\lambda^*} + (1-\lambda^*)\log\frac{1-q_\delta}{1-\lambda^*}\right] = O(\delta^2).$$

*Then the expected number of rejections for any knockoff procedure at FDR significance level $\alpha$ is upper bounded by $C_1(\alpha,\delta)k^* + C_2(\alpha,\delta)$.*

*Proof.* Consider the random walk

$$S_k = \sum_{j=1}^{k}(p - Z_j), \quad \text{where } Z_j \overset{\text{ind.}}{\sim} \text{Bern}(q_j), \quad p = \frac{\alpha}{1+\alpha}, \quad q_j = \frac{e^{-\eta_{(j)}}}{1+e^{-\eta_{(j)}}}.$$

Then by (9), for any knockoff procedure, the number of rejections $R$ is upper bounded by $\frac{2+\hat{k}}{1+\alpha}$, where

$$\hat{k} = \max\left\{k : S_k \geq \frac{1}{1+\alpha}\right\} \leq \max\{k : S_k \geq 0\}.$$

Consider another random walk

$$\tilde{S}_k = \sum_{j=1}^{k}(p - \tilde{Z}_j), \quad \text{where } \tilde{Z}_j \overset{\text{i.i.d.}}{\sim} \text{Bern}(q_\delta), \quad q_\delta = \frac{\alpha+\delta}{1+\alpha+\delta} \leq q_{(k^*)}$$

Note that

$$S_k = \sum_{j=1}^{k^*}(p - Z_j) + \sum_{j=k^*}^{d}(p - Z_j) \leq pk^* + \sum_{j=k^*}^{d}(p - Z_j).$$

Since $q_1 \leq q_2 \leq ...$, we know that $\sum_{j=k^*}^{d}(p - Z_j)$ is stochastically smaller than $\sum_{j=k^*}^{d}(p - \tilde{Z}_j)$, where $\tilde{Z}_j \overset{\text{i.i.d.}}{\sim} \text{Bern}(q_{(k^*)})$. Therefore $S_k$ is stochastically smaller than $pk^* + \tilde{S}_{k-k^*}$. Therefore

$$\mathbb{E}[\hat{k}] \leq \mathbb{E}[\max\{k : S_k \geq 0\}] \leq \mathbb{E}\left[\max\left\{k : \tilde{S}_{k-k^*} \geq -pk^*\right\}\right] = k^* + \mathbb{E}\left[\max\left\{k : \tilde{S}_k \geq -pk^*\right\}\right]. \tag{23}$$

Define

$$p(r) \triangleq \mathbb{P}\left(\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right).$$

Let $C_m = \max\{4\left(\frac{q_\delta}{p} - 1\right)^{-1}, 1\}$, and then

$$\mathbb{E}\left[\max\left\{k : \tilde{S}_k \geq -pk^*\right\}\right]$$

$$= \sum_{r=1}^{\infty} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right)$$

$$= \sum_{r=1}^{C_m k^*} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k - pk^*\right\} = r\right) + \sum_{r=C_m k^* + 1}^{\infty} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k - pk^*\right\} = r\right)$$

$$\leq C_m k^* \sum_{r=1}^{C_m k^*} \mathbb{P}\left(\max\left\{k : \tilde{S}_k - pk^*\right\} = r\right) + \sum_{r=C_m k^* + 1}^{\infty} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k - pk^*\right\} = r\right) \tag{24}$$

$$\leq C_m k^* + \sum_{r=C_m k^* + 1}^{\infty} r\mathbb{P}\left(\max\left\{k : \tilde{S}_k - pk^*\right\} = r\right)$$

$$= C_m k^* + \sum_{r=C_m k^* + 1}^{\infty} rp(r).$$

Therefore, combining Equations 24 and 23, we have

$$\mathbb{E}[\hat{k}] \leq k^* + \mathbb{E}\left[\max\left\{k : \tilde{S}_k \geq -pk^*\right\}\right] \leq k^* + C_m k^* + \sum_{r=C_m k^* + 1}^{\infty} rp(r) = (C_m + 1)k^* + \sum_{r=C_m k^* + 1}^{\infty} rp(r).$$

Recalling the definitions of $C_m$ and $q_{(k^*)}$, we have

$$4\left(\frac{q_\delta}{p} - 1\right)^{-1} = 4\left(\frac{(\alpha + \delta)(1 + \alpha)}{1 + \alpha + \delta} - 1\right)^{-1} = \frac{4\alpha(1 + \alpha + \delta)}{\delta}.$$

Therefore we have $C_m + 1 \leq C_1(\alpha, \delta)$. Turning to the second term, by Lemma 4 we have

$$\sum_{r=C_m k^* + 1}^{\infty} rp(r) \leq \frac{q_\delta e}{2\pi\sqrt{p(1-p)}} \cdot \min\left\{\frac{2(q_\delta - p)}{p(1 - q_\delta)}, 1\right\} \sum_{r=C_m k^* + 1}^{\infty} \sqrt{r}e^{-c_h r}. \tag{25}$$

Note that we can bound the summation $\sum_{r=C_m k^* + 1}^{\infty} \sqrt{r}e^{-c_h r}$ by (note that $C_m k^* + 1 \geq 1/\alpha + 1$)

$$\sum_{r=1/\alpha + 1}^{\infty} \sqrt{r}e^{-c_h r} \leq \sqrt{1 + \alpha} \sum_{r=1/\alpha + 1}^{\infty} \sqrt{r - 1}e^{-c_h r} \leq \sqrt{1 + \alpha} \int_{1/\alpha}^{\infty} \sqrt{r}e^{-c_h r}dr \leq \sqrt{1 + \alpha} \int_{0}^{\infty} \sqrt{r}e^{-c_h r}dr, \tag{26}$$

where

$$\int_{0}^{\infty} \sqrt{r}e^{-c_h r}dr = \int_{0}^{\infty} 2(2c_h)^{-3/2}y^2 e^{-\frac{y^2}{2}}dy = (2c_h)^{-3/2}\int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}}dy = \sqrt{2\pi}(2c_h)^{-3/2}$$

The first equality above is obtained by change of variable $r = y^2/2c_h$. Therefore,

$$\sum_{r=C_m k^* + 1}^{\infty} \sqrt{r}e^{-c_h r} \leq \sqrt{1 + \alpha}\sqrt{2\pi}(2c_h)^{-3/2}.$$

Thus

$$\sum_{r=C_m k^* + 1}^{\infty} rp(r) \leq \frac{e}{4\sqrt{\pi}}\frac{\sqrt{1 + \alpha}}{\sqrt{p(1-p)}}c_h^{-3/2}\min\left\{\frac{2(q_\delta - p)}{p(1 - q_\delta)}, 1\right\}$$

and the proposition is proved. $\qquad\square$

## 6.3 Proof of results in Sections 2.4–2.5

**Theorem 2.** *Let $\beta_{(1)}^2 \geq \cdots \geq \beta_{(d)}^2$ be the order statistics of $(\beta_1^2, \ldots, \beta_d^2)$ and let $b_1(\Sigma), \ldots, b_d(\Sigma)$ be defined as in* (12). *For any target FDR level $\alpha > 0$, let $k$ be the smallest integer such that*

$$\frac{2\beta_{(k)}^2}{\sigma^2 b_k(\Sigma)} < -\frac{1}{2}\log \alpha, \tag{13}$$

*Then the expected number of rejections for any knockoff procedure at FDR level $\alpha$ is upper bounded by $C_1^*(\alpha)k + C_2^*(\alpha)$, where $C_1^*$ and $C_2^*$ depend only on $\alpha$.*

*Proof.* Assume without loss of generality that $\Delta_{11} \geq \Delta_{22} \geq \cdots \geq \Delta_{dd}$, so we have for all $j > k$ that

$$\Delta_{jj} \geq b_j(\Sigma) \geq jb_k(\Sigma)/k.$$

If $\beta_j^2 \leq \beta_{(k)}^2$, then we have

$$\mu_j = \frac{2\beta_j^2}{\sigma^2 \Delta_{jj}} \leq \frac{2k\beta_{(k)}^2}{\sigma^2 jb_k(\Sigma)} \leq \frac{-k\log\alpha}{2j}.$$

Define

$$N_\alpha = \#\left\{i \geq 1 : \ \beta_{k+i}^2 \leq \beta_{(k)}^2, \ \eta_{k+i} > -\frac{1}{2}\log\alpha\right\}.$$

At most $k - 1$ indices $j$ have $\beta_j^2 > \beta_{(k)}^2$, so the total number of indices with $\eta_j \geq -\frac{1}{2}\log\alpha$ is at most $2k + N_\alpha - 1$. If $N_\alpha = n$, then we have

$$\eta_{(2k+n)} \leq -\frac{1}{2}\log\alpha, \tag{27}$$

in which case Proposition 4 implies that the conditional expected number of rejections is bounded above by $(2k + n)C_1^\dagger(\alpha) + C_2^\dagger(\alpha)$, where $C_i^\dagger(\alpha) = C_i(\alpha, \sqrt{\alpha} - \alpha)$ for $i = 1, 2$. Therefore the expected number of rejections is

$$\mathbb{E}R = \sum_n \mathbb{E}[R \mid N_\alpha = n]\,\mathbb{P}(N_\alpha = n)$$

$$\leq \sum_n \mathbb{P}(N_\alpha = n) \cdot \left[(2k + n)C_1^\dagger(\alpha) + C_2^\dagger(\alpha)\right]$$

$$\leq (2k + \mathbb{E}N_\alpha)C_1^\dagger(\alpha) + C_2^\dagger(\alpha).$$

Applying Lemma 1 with $z_i = \eta_{k+i}$ implies $\mathbb{E}N_\alpha \leq C_3(\alpha)k$, so the result holds for

$$C_1^*(\alpha) = (2 + C_3(\alpha))\,C_1(\alpha, \sqrt{\alpha} - \alpha), \quad \text{and} \quad C_2^*(\alpha) = C_2(\alpha, \sqrt{\alpha} - \alpha).$$

If $\beta_{(k)}$ is replaced with $\beta_{(1)}$ in (13) then the same proof applies, except that instead of $k - 1$ indices with large $\beta_j$, there are none. Then we can replace $2k + n$ with $k + 1 + n$ in (27), and the result holds for

$$C_1^*(\alpha) = (1 + C_3(\alpha))\,C_1(\alpha, \sqrt{\alpha} - \alpha), \quad \text{and} \quad C_2^*(\alpha) = C_1(\alpha, \sqrt{\alpha} - \alpha) + C_2(\alpha, \sqrt{\alpha} - \alpha).$$

$\square$

**Theorem 3.** *Let $\beta_{(1)}^2 \geq \cdots \geq \beta_{(d)}^2$ be the order statistics of $(\beta_1^2, \ldots, \beta_d^2)$, and assume $\beta$ is a random permutation of the order statistics, independent of $\Delta$. Let $\pi_1 = d_1/d$ be the non-null proportion, and define $b_1(\Sigma), \ldots, b_d(\Sigma)$ as in* (12). *For any target FDR level $\alpha > 0$, let $k$ be the smallest integer for which*

$$\frac{2\beta_{(k)}^2}{\sigma^2 b_{\lfloor k/\pi_1 \rfloor}(\Sigma)} < -\frac{1}{2}\log\alpha, \tag{14}$$

*where $\lfloor \cdot \rfloor$ is the integer floor. Then the expected number of rejections for any knockoff procedure at FDR level $\alpha$ is upper bounded by $C_1^*(\alpha)k + C_2^*(\alpha)$, where $C_1^*$ and $C_2^*$ are the same constants as in Theorem 2.*

*Proof.* Define $K = \lfloor k/\pi_1 \rfloor$ and assume without loss of generality that $\Delta_{11} \geq \Delta_{22} \geq \cdots \geq \Delta_{dd}$, so we have for all $j > K$ that

$$\Delta_{jj} \geq b_j(\Sigma) \geq jb_K(\Sigma)/K.$$

If $\beta_j^2 \leq \beta_{(k)}^2$, then we have for $j > K$,

$$\mu_j = \frac{2\beta_j^2}{\sigma^2 \Delta_{jj}} \leq \frac{2K\beta_{(k)}^2}{\sigma^2 jb_K(\Sigma)} \leq \frac{-K\log\alpha}{2j}.$$

Define

$$N_\alpha^{(1)} = \#\left\{j \leq K : \beta_j \neq 0\right\}, \quad \text{and} \quad N_\alpha^{(2)} = \#\left\{i \geq 1 : \beta_{K+i}^2 \leq \beta_{(k)}^2, \ \eta_{K+i} > -\frac{1}{2}\log\alpha\right\}.$$

At most $k - 1$ indices $j$ have $\beta_j^2 > \beta_{(k)}^2$, so the total number of indices with $\eta_j \geq -\frac{1}{2}\log\alpha$ is at most $k + N_\alpha^{(1)} + N_\alpha^{(2)} - 1$. If $N_\alpha^{(1)} = n_1$ and $N_\alpha^{(2)} = n_2$, then we have

$$\eta_{(k+n_1+n_2)} \leq -\frac{1}{2}\log\alpha, \tag{28}$$

in which case Proposition 4 implies that the conditional expected number of rejections is bounded above by $(2k + n)C_1^\dagger(\alpha) + C_2^\dagger(\alpha)$, where $C_i^\dagger(\alpha) = C_i(\alpha, \sqrt{\alpha} - \alpha)$ for $i = 1, 2$. Therefore the expected number of rejections is

$$\mathbb{E}R \leq (k + \mathbb{E}N_\alpha^{(1)} + \mathbb{E}N_\alpha^{(2)})C_1^\dagger(\alpha) + C_2^\dagger(\alpha).$$

Because $\mathbb{P}(\beta_j \neq 0) = \pi_1$ for every $j$, we have $\mathbb{E}N_\alpha^{(1)} = \pi_1 K \leq k$. For $j > K$, we can write

Applying Lemma 1 with $z_i = \eta_{k+i}$ implies $\mathbb{E}N_\alpha \leq C_3(\alpha)k$, so the result holds for

$$C_1^*(\alpha) = (2 + C_3(\alpha))C_1(\alpha, \sqrt{\alpha} - \alpha), \quad \text{and} \quad C_2^*(\alpha) = C_2(\alpha, \sqrt{\alpha} - \alpha).$$

If $\beta_{(k)}$ is replaced with $\beta_{(1)}$ in (13) then the same proof applies, except that instead of $k - 1$ indices with large $\beta_j$, there are none. Then we can replace $2k + n$ with $k + 1 + n$ in (27), and the result holds for

$$C_1^*(\alpha) = (1 + C_3(\alpha))C_1(\alpha, \sqrt{\alpha} - \alpha), \quad \text{and} \quad C_2^*(\alpha) = C_1(\alpha, \sqrt{\alpha} - \alpha) + C_2(\alpha, \sqrt{\alpha} - \alpha).$$

$\square$

**Corollary 1.** *Fix $\sigma^2 = 1$. Let $B_T$ denote the bound obtained from Theorem 3, and $B_A = \mathbb{E}\widehat{\text{TPR}}$ is the output of Algorithm 1; both bounds depend on $\Sigma$, $\alpha$, and the order statistics $\beta_{(1)}^2, \ldots, \beta_{(d)}^2$. Then*

$$\frac{1}{d_1}B_T \geq B_A \geq \sup_{\mathcal{R}} \min_{\beta \in \Omega} \text{TPR}(\mathcal{R}, \beta),$$

*where $\Omega \subseteq \mathbb{R}^d$ is the set of vectors $\beta \in \mathbb{R}^d$ whose squared entries have the same order statistics, and $\text{TPR}(\mathcal{R}, \beta)$ is the TPR of method $\mathcal{R}$ under $\widehat{\beta} \sim \mathcal{N}_d(\beta, \Sigma)$.*

*Proof.* Let $\Pi$ denote a uniformly random permutation on $d$ elements, and let $\Pi(j) \in \{1, \ldots, d\}$ denote the index where $j$ is sent by $\Pi$. Because $\Sigma$ is constant, to limit notational bloat we will suppress the input $\Sigma$ in $b_k(\Sigma)$, writing $b_1 \geq b_2 \geq \cdots \geq b_d$ instead. We assume without loss of generality that $\Delta_{11} \geq \cdots \geq \Delta_{dd}$. Finally fix a representative in $\Omega$ as

$$\beta^* = (|\beta_{(1)}|, \ldots, |\beta_{(d)}|).$$

The proof of Theorem 3 argues that $C_1^*(\alpha)k + C_2^*(\alpha)$ is an analytic upper bound for the expected number of total rejections when we apply the knockoff* procedure with $\eta_j$ sampled from

$$\eta_j \overset{\text{ind.}}{\sim} |\mathcal{N}(\mu_j^*, 2\mu_j^*)|, \quad \text{with} \quad \mu_j^* = \frac{2\beta_{\Pi(j)}^{*2}}{b_j} \tag{29}$$

As a result, $\frac{1}{d_1}B_T$ is an analytic upper bound for the TPR of the same process.

By contrast, Algorithm 1 directly samples $\eta$ from the same distribution (29) and then directly simulates the knockoff* procedure with those $\eta$ values and in each Monte Carlo iteration directly calculates $\text{TPP} \leq R/d_1$. As a result, $B_T \geq \mathbb{E}R/d_1 \geq \mathbb{E}\text{TPP} = \mathbb{E}\widehat{\text{TPR}}$.

29

Next, consider the behavior of any feasible knockoff method $\mathcal{R}$. Because $\Omega$ is a finite set, its lowest power on any $\beta \in \Omega$ is well-defined, and is no larger than its average power under sampling from any distribution over $\Omega$:

$$\mathbb{E}_{\beta \sim P} \text{TPR}(\mathcal{R}, \beta) \ \geq \ \min_{\beta \in \Omega} \text{TPR}(\mathcal{R}, \beta),$$

where $P$ is the distribution of $(\beta^*_{\Pi(1)}, \ldots, \beta^*_{\Pi(d)})$. Conditional on $\beta$, $\eta_1, \ldots, \eta_d$ are then distributed as

$$\eta_j \ \overset{\text{ind.}}{\sim} \ |\mathcal{N}(\mu_j, 2\mu_j)|, \quad \text{with} \quad \mu_j \ = \ \frac{2\beta^{*2}_{\Pi(j)}}{\Delta_{jj}} \ \leq \ \mu_j^*,$$

and once they are generated $\mathcal{R}$ will choose the orderings and directions in some way that is suboptimal compared to knockoff*. As a result, $B_A \geq \mathbb{E}_{\beta \sim P} \text{TPR}(\mathcal{R}, \beta)$ and we have the result. $\qquad\square$

**Corollary 2.** *Consider a sequence of problems with $d \to \infty$ for which*

(i) *$F_d(\cdot)$ has a distribution limit $F(\cdot)$ with no point mass at 0, i.e. $F(0) = 0$.*

(ii) *For some $r > 1$, $\beta_j = \sigma\sqrt{2r \log d}$ for a subset of $d_1$ indices, and $\beta_j = 0$ for the rest.*

(iii) *Either $d_1 \to \infty$ and $\text{supp}(\beta)$ is uniformly random, or $\liminf d_1/d > 0$ and $\text{supp}(\beta)$ is arbitrary.*

*Then*

1. *The Bonferroni procedure at any level $\alpha$ has $\text{TPR} \to 1$, and*

2. *If $\lambda_1 / \log d \to \infty$ then any knockoff method at any level $\alpha$ has $\text{TPR} \to 0$.*

*Proof.* We begin with the first claim. When the variance $\sigma^2$ is known, and the test statistics are $z$ statistics, the claim follows from the fact that the proportion of non-null $p$-values that are smaller than $\alpha/d$ converges to 1 when all the non-null $\beta_j = \sqrt{2r \log d}$ for some $r > 1$. When the variance $\sigma^2$ is unknown, we have access to an $\chi^2_d$-distributed variance estimate $\hat{\sigma}^2_d \overset{p}{\to} \sigma^2$. Therefore, the proportion of non-null $t$-statistics that are larger than $\sqrt{(1+r)\log d} < \sqrt{2r \log d}$ converges to one. On the other hand, since $\chi^2_d/d \overset{p}{\to} 1$, we have that the c.d.f. of $t_d$ distribution at $-\sqrt{(1+r)\log d}$ is smaller than the gaussian c.d.f at $-\sqrt{2\log d}$, for $d$ succiffiently large. This proves that the TPR also converges to 1 when we use OLS $t$-statistics.

To prove the second claim, we will first lower bound $b_{ad}(\Sigma)$ for a generic fraction $a > 0$ to show that most of the diagonal entries of $\Delta$ are prohibitively large. By assumption, there exists some constant $c > 0$ for which $F(c) < a/2$. Then for large enough $d$ we have $F_d(c) \leq a/2$ as well, so that $|u_{1,(\lfloor ad/2 \rfloor)}| \geq c/\sqrt{d}$, and we have

$$b_{\lceil ad \rceil}(\Sigma) \ \geq \ \lambda_1 \sum_{j=1}^{\lceil ad \rceil} u_{1,(j)}^2 \ \geq \ \frac{\lambda_1 ad}{2} u_{1,(\lfloor ad/2 \rfloor)}^2 \ \geq \ \frac{\lambda_1 ac^2}{2}.$$

By assumption, $\lambda_1 / \log d \to \infty$, so it follows that for sufficiently large $d$ we have

$$\frac{2\beta_{(1)}^2}{\sigma^2 b_{\lceil ad \rceil}(\Sigma)} \ = \ \frac{2\beta_0^2}{\sigma^2 b_{\lceil ad \rceil}(\Sigma)} \ = \ \frac{4 \log d}{b_{\lceil ad \rceil}(\Sigma)} \ < \ -\frac{1}{2} \log \alpha.$$

As a result, by Theorem 2, the expected number of rejections is bounded above by $C_1^*(\alpha)(ad+1) + C_2^*(\alpha)$. If $\liminf d_1/d > 0$, this completes the proof, since $a$ and $\alpha$ were arbitrary.

If $\text{supp}(\beta)$ is uniformly random, then by Theorem 3, the expected number of rejections is bounded above by $C_1^*(\alpha)(\pi_1 ad + 1) + C_2^*(\alpha) = C_1^*(\alpha)(ad_1 + 1) + C_2^*(\alpha)$. Again, this completes the proof since $a$ and $\alpha$ were arbitrary. $\qquad\square$

## 6.4 Proof of Theorem 4

**Theorem 4.** *Assume that we carry out a continuous knockoffs procedure $\mathcal{R}(\cdot)$ on the asymptotically normal estimator*

$$\widehat{\beta}_n \ \Rightarrow \ \mathcal{N}_d(\beta, \Sigma), \quad \text{and} \quad \widehat{\Sigma}_n \overset{p}{\to} \Sigma,$$

*where $\Sigma$ is nonsingular. Let*

$$E = \{(\xi, \widetilde{\beta}) : \ W_1^+, W_1^-, \ldots, W_d^+, W_d^- \text{ are all positive and distinct}\},$$

*and assume that $E^c$ has Lebesgue measure zero. Then we have for any $\alpha$,*

$$\mathcal{R}_\alpha\left(\widehat{\beta}_n, \widehat{\Sigma}_n; \omega_n\right) \;\Rightarrow\; \mathcal{R}_\alpha\left(\widehat{\beta}, \Sigma; \omega\right),$$

*where $\widehat{\beta} \sim \mathcal{N}_d(\beta, \Sigma)$ and $\omega \sim \mathcal{N}_d(0, \Delta(\Sigma) - \Sigma)$ independently. In particular, the FDR and TPR of $\mathcal{R}_\alpha$ converge to their values under exact normal sampling.*

*Proof.* Because $\widehat{\Sigma}_n \overset{P}{\to} \Sigma$, it will be nonsingular with probability approaching 1; likewise for $\widehat{\Delta}_n \succeq \widehat{\Sigma}_n$. Because $\Delta(\cdot)$ is continuous, we also have $\widehat{\Delta}_n \overset{P}{\to} \Delta = \Delta(\Sigma)$ and $\widehat{M}_n \overset{P}{\to} (\Delta - \Sigma)^{1/2}$. Then

$$\begin{pmatrix} \xi_n \\ \widetilde{\beta}_n \end{pmatrix} \;=\; \begin{pmatrix} \widehat{\Sigma}_n^{-1} - \widehat{\Delta}_n^{-1} & -\widehat{\Delta}_n^{-1}\widehat{M}_n \\ I_d & \widehat{M}_n \end{pmatrix} \begin{pmatrix} \widehat{\beta}_n \\ \omega' \end{pmatrix}.$$

Because all four blocks of the matrix are converging in probability, $\widehat{\beta}_n$ is converging in distribution by assumption, and $\omega'$ has a fixed distribution,

$$(\xi_n, \widetilde{\beta}_n) \;\Rightarrow\; (\xi, \widetilde{\beta}) \;=\; \left(\Sigma^{-1}\widehat{\beta} - \Delta^{-1}(\widehat{\beta} + \omega), \; \widehat{\beta} + \omega\right).$$

Because $W^+(\cdot)$ and $W^-(\cdot)$ are both continuous, we also have

$$\left(W^+(\xi_n, \widetilde{\beta}_n), \; W^-(\xi_n, \widetilde{\beta}_n)\right) \;\Rightarrow\; \left(W^+(\xi, \widetilde{\beta}), \; W^-(\xi, \widetilde{\beta})\right).$$

To complete the proof, for $(\xi, \widetilde{\beta}) \in E$, let $\Pi(\xi, \widetilde{\beta})$ denote the unique permutation on $2d$ elements that arranges all $2d$ variables in decreasing order. Then $E$ is the disjoint union of $E_\pi = E \cap \Pi^{-1}(\pi)$, where $\Pi^{-1}$ is the preimage and $\pi$ ranges over all permutations. Because there are only finitely many permutations and the rejection set $\mathcal{R}$ is constant on each $E_\pi$, it remains only to show for each $\pi$ that

$$\mathbb{P}\left((\xi_n, \widetilde{\beta}_n) \in E_\pi\right) \to \mathbb{P}\left((\xi, \widetilde{\beta}) \in E_\pi\right). \tag{30}$$

For example, if we take $\pi$ to be the identity permutation, then

$$E_\pi \;=\; \{W_1^+ > \cdots > W_d^+ > W_1^- > \cdots > W_d^- > 0\}.$$

Let $\delta_\pi(\xi, \widetilde{\beta}) \in \mathbb{R}$ denote the largest margin of any of the $2d$ inequalities defining $E_\pi$, possibly negative, so that $E_\pi = \{\delta_\pi > 0\}$. Because $\delta_\pi$ is continuous in $(W^+, W^-)$ it too is converging in distribution; let $F_{\pi,n}$ and $F_\pi$ denote the cumulative distribution functions of $\delta_\pi(\xi_n, \widetilde{\beta}_n)$ and $\delta_\pi(\xi, \widetilde{\beta})$, respectively. Because $\delta_\pi^{-1}(0) \subseteq E^c$ has Lebesgue measure zero, 0 is a continuity point of $F_\pi$, so we have (30) and the proof is complete. $\qquad\square$

## 6.5 Proof of technical lemmas

**Lemma 1.** *Suppose that $z_i \sim \mathcal{N}(\mu_i, 2\mu_i)$ for $i = 1, 2, \ldots$, where $0 \le \mu_i \le \frac{-k \log \alpha}{2(k+i)}$, for some integer $k \ge 1$ and $\alpha \in (0, 1)$. Then the expected number of $z_i$ values with $|z_i| > -\frac{1}{2}\log \alpha$ is bounded by $C_3(\alpha)k$.*

*Proof.* Define $c = -\frac{1}{2}\log \alpha > 0$ and $N_\alpha = \#\{i : |z_i| > c\}$.

$$c = -\frac{1}{2}\log \alpha > 0, \quad t_i = i/k, \quad \text{and } N_\alpha = \#\{i : |z_i| > c\}.$$

Because $\mathbb{P}(|z_i| > c)$ is increasing in $\mu_i$ for $\mu_i \in [0, c]$, we have

$$\mathbb{E}N_\alpha \;=\; \sum_{i=1}^{\infty} \mathbb{P}(|z_i| > c) \;\le\; \sum_{i=1}^{\infty} f(i/k), \quad \text{where } f(t) = \mathbb{P}_{\mu = \frac{c}{1+t}}\left(|\mathcal{N}(\mu, 2\mu)| > c\right).$$

Because $f$ is decreasing, we have

$$\sum_{i=1}^{\infty} f(i/k) \;=\; \sum_{i=1}^{\infty} k \int_{(i-1)/k}^{i/k} f(i/k)\, dt \;\le\; k \int_{0}^{\infty} f(t)\, dt.$$

Writing $f$ explicitly, we obtain

$$\mathbb{E}N \leq k \int_0^\infty \left( \Phi\left( \frac{-(2+t)\sqrt{c}}{\sqrt{2+2t)}} \right) + 1 - \Phi\left( \frac{t\sqrt{c}}{\sqrt{2+2t}} \right) \right) dt.$$

The integral is finite since $1 - \Phi(x) = \Phi(-x) \leq \exp(-x^2/2)$ for large $x$. Evaluating the integral numerically gives $C_3(0.05) \leq 1.05$, $C_3(0.1) \leq 1.37$, and $C_3(0.2) \leq 2.02$.

□

**Lemma 2.** *For any $q, \theta \in (0,1)$, we have*

$$\frac{q \log(1-\theta)}{\log(1-q\theta)} - 1 \geq \frac{(1-q)}{2}\theta.$$

*Proof.* For any $q \in (0,1)$, let

$$f(\theta) = \frac{(1-q)}{2}\theta \log(1-q\theta) + \log(1-q\theta) - q\log(1-\theta).$$

Because $\log(1-q\theta) < 0$, it suffices to show that $f(\theta) \geq 0$ for $\theta \in (0,1)$. First, note that $f(0) = 0$. We will show next that $f'(\theta) > 0$ when $\theta > 0$:

$$f'(\theta) = \frac{(1-q)}{2}\left( \log(1-q\theta) - \frac{q\theta}{1-q\theta} \right) - \frac{q}{1-q\theta} + \frac{q}{1-\theta}$$

$$= \frac{(1-q)}{2}\left( \log(1-q\theta) + \frac{q\theta(1+\theta)}{(1-\theta)(1-q\theta)} \right).$$

Note that

$$\log(1-q\theta) = -\log\frac{1}{(1-q\theta)} \geq -\left( \frac{1}{(1-q\theta)} - 1 \right) = \frac{-q\theta}{1-q\theta}.$$

Therefore

$$f'(\theta) = \frac{(1-q)}{2}\left( \log(1-q\theta) + \frac{q\theta(1+\theta)}{(1-\theta)(1-q\theta)} \right)$$

$$\geq \frac{(1-q)}{2}\left( \frac{-q\theta}{1-q\theta} + \frac{q\theta(1+\theta)}{(1-\theta)(1-q\theta)} \right)$$

$$= \frac{q(1-q)\theta^2}{(1-\theta)(1-q\theta)} > 0,$$

and the lemma is proved.

□

**Lemma 3.** *Consider a random walk $S_t = \sum_{i=1}^t \zeta_t$ where $\zeta_t$ are i.i.d Bernoulli variables with $\mathbb{P}(\zeta_t = 1) = 1 - q$ and $\mathbb{P}(\zeta_t = -1/\alpha) = q$. Let $p = \alpha/(1+\alpha)$ and suppose that $q > p$. Then*

$$\mathbb{P}\left( \max_{t \geq 1} S_t \leq 0 \right) \leq q \min\left\{ \frac{2(q-p)}{p(1-q)}, 1 \right\}$$

*Proof.* We begin by identifying the martingale associated with the moment generating function of $S_t$. Let $\psi_0 > 0$ be a positive value that satisfies

$$\mathbb{E}e^{\psi_0 \zeta_1} = (1-q)e^{\psi_0} + qe^{-\psi_0 \frac{1}{\alpha}} = 1. \tag{31}$$

We will prove later that such $\psi_0$ exists and is unique. It follows that $e^{\psi_0 S_t}$ is a martingale, since

$$\mathbb{E}[e^{\psi_0 S_{t+1}} | S_t] = e^{\psi_0 S_t}\mathbb{E}e^{\psi_0 \zeta_{t+1}} = e^{\psi_0 S_t}.$$

For any finite value $M > 0$, let $\tau$ be the first time when the random walk leaves $(-M, 0]$, i.e.

$$\tau = \min\{t \geq 1 : S_\tau > 0 \text{ or } S_\tau \leq -M\}$$

Then $\tau$ is a stopping time, which is almost surely finite since $\mathbb{E}\,\zeta_1 < 0$ implies $S_t \to -\infty$ almost surely. Since $|S_{t \wedge \tau}| \leq M + 1$ for all $t \geq 1$, we have by the optional stopping theorem

$$1 = \mathbb{E}e^{\psi_0 S_\tau} \tag{32}$$

$$= p_0(M)\mathbb{E}[e^{\psi_0 S_\tau}|S_\tau > 0] + (1 - p_0(M))\mathbb{E}[e^{\psi_0 S_\tau}|S_\tau \leq -M], \tag{33}$$

where $p_0(M)$ is the probability that $S_t$ reaches $(0, \infty)$ before it reaches $(-\infty, -M)$. Since $S_t$ can increase no more than 1 at a time, we have $S_\tau \leq 1$. Therefore

$$\mathbb{E}[e^{\psi_0 S_\tau}|S_\tau > 0] \leq e^{\psi_0},$$

and on the other hand, we have

$$\mathbb{E}[e^{\psi_0 S_\tau}|S_\tau \leq -M] \leq e^{-\psi_0 M}.$$

Substituting both bounds into (33) and isolating $p_0(M)$, we have

$$p_0(M) \geq \frac{1 - e^{-\psi_0 M}}{e^{\psi_0} - e^{-\psi_0 M}}.$$

Taking $M \to \infty$ we obtain

$$\mathbb{P}\left(\max_{t \geq 1} S_t > 0\right) \geq \sup_{M > 0} p_0(M) = e^{-\psi_0}.$$

We pause to show the existence of a unique positive solution $\psi_0$ to equation (31). Defining $\lambda = e^{-\psi_0}$, we are equivalently seeking a unique solution in $(0, 1)$ to the polynomial equation

$$\lambda = 1 - q + q\lambda^{1+\alpha^{-1}} = 1 - q + q\lambda^{1/p}. \tag{34}$$

Let $g(\lambda) = q\lambda^{1/p} - \lambda + 1 - q$. Then $g(0) = 1 - q > 0$ and $g(1) = 0$. Furthermore,

$$g'(\lambda) = \frac{q}{p}\lambda^{-1+1/p} - 1 = \frac{q}{p}\lambda^{1/\alpha} - 1.$$

Hence, $g$ has a unique minimum at $\lambda_* = (p/q)^\alpha \in (0, 1)$, with $g'(\lambda) < 0$ for $\lambda \in (0, \lambda_*)$ and $g'(\lambda) > 0$ for $\lambda \in (\lambda_*, 1)$. As a result $g(\lambda) = 0$ has a unique solution in $(0, 1)$. To complete the proof, we require an upper bound for $1 - \lambda$, since we have already shown $\mathbb{P}\left(\max_{t \leq 1} S_t \geq 0\right) \leq 1 - \lambda$.

Let $1 - \lambda = q\theta$. Then (34) can be expressed as

$$1 - \theta = (1 - q\theta)^{1/p}. \tag{35}$$

Taking the log on both sides and multiplying by $q$, we have

$$\frac{q}{p} = \frac{q\log(1 - \theta)}{\log(1 - q\theta)}.$$

The right hand side approaches 1 as $\theta \to 0$ and diverges to infinity as $\theta \to 1$, giving the trivial bound $\theta < 1$. We are mainly interested in improving this bound when $q/p$ is close to 1, in which case $\theta$ should be small. Lemma 2 gives

$$\frac{q}{p} - 1 \geq \frac{(1-q)}{2}\theta \quad \Rightarrow \quad \theta \leq \frac{2(q-p)}{p(1-q)},$$

so that, as desired,

$$\mathbb{P}\left(\max_{t \geq 1} S_t \leq 0\right) \leq 1 - \lambda \leq q\min\left\{\frac{2(q-p)}{p(1-q)}, 1\right\}.$$

$\square$

**Lemma 4.** *Consider the following random walk defined in the proof of Proposition 4:*

$$\tilde{S}_k = \sum_{j=1}^{k}(p - \tilde{Z}_j), \quad \tilde{Z}_j \overset{i.i.d.}{\sim} \text{Bern}(q_\delta),$$

*where $q_\delta > p$. Define*

$$\lambda^* = p + \frac{q_\delta - p}{4} < q_\delta, \quad \text{and} \quad c_h = -\left[\lambda^* \log \frac{q_\delta}{\lambda^*} + (1 - \lambda^*) \log \frac{1 - q_\delta}{1 - \lambda^*}\right].$$

*Then for any*

$$r > 4k^* \left(\frac{q_\delta}{p} - 1\right)^{-1},$$

*we have*

$$p(r) \triangleq \mathbb{P}\left(\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right) \leq \frac{q_\delta e^{1 - c_h r}}{2\pi \sqrt{rp(1 - p)}} \cdot \min\left\{\frac{2(q_\delta - p)}{p(1 - q_\delta)}, 1\right\}.$$

*Proof.* Define events

$$A_1 = \left\{\sum_{j=1}^{r} \tilde{Z}_j = \lfloor pr + pk^* \rfloor\right\},$$

and

$$A_2 = \left\{\max_{k \geq 1} \sum_{j=r+1}^{r+k} (p - \tilde{Z}_j) \leq 0\right\}.$$

We will show that

$$\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right\} \subset A_1 \cap A_2.$$

First, note that $\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r$ implies the following three conditions:

$$(1)\tilde{S}_r \geq -pk^*, \quad (2)\tilde{S}_{r+1} \leq -pk^*, \quad \text{and} \quad (3)\max_{k \geq 1}(\tilde{S}_{r+k} - \tilde{S}_r) \leq 0.$$

Recalling the definition of $\tilde{S}_r$, conditions (1) and (2) are equivalent to

$$pr + pk^* + p - 1 \leq \sum_{j=1}^{r} \tilde{Z}_j \leq pr + pk^*.$$

Since $\sum_{j=1}^{r} \tilde{Z}_j$ is an integer, one of the following two events must happen: (a) there exists no integer between $pr + pk^*$ and $pr + pk^* + p - 1$. In this case, $\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right\}$ is an empty set; (b) there exists exactly one integer between $pr + pk^*$ and $pr + pk^* + p - 1$. Then it follows that $\sum_{j=1}^{r} \tilde{Z}_j = \lfloor pr + pk^* \rfloor$. In either case, we have shown that

$$\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right\} \subset A_1.$$

In addition, condition (3) is equivalent to event $A_2$. Therefore, we have shown that

$$\left\{\max\left\{k : \tilde{S}_k \geq -pk^*\right\} = r\right\} \subset A_1 \cap A_2.$$

Note that $A_1$ only depends on $\tilde{Z}_j, j \leq r$ and $A_2$ only depends on $\tilde{Z}_j, j > r$. Since $\{\tilde{Z}_j\}$ is a sequence of i.i.d Bernoulli variables, $A_1$ and $A_2$ are independent. Therefore

$$p(r) \leq P(A_1)P(A_2).$$

We now bound $P(A_1)$ and $P(A_2)$ separately. First,

$$P(A_1) = \text{Binom}(r, q_\delta; \lfloor pr + pk^* \rfloor).$$

where $\text{Binom}(r, q_\delta; \lfloor pr + pk^* \rfloor)$ is the probability of the binomial distribution $\text{Binom}(r, q_\delta; \cdot)$ at $\lfloor pr + pk^* \rfloor$. Let $m = \lfloor pr + pk^* \rfloor$, we have

$$\text{Binom}(r, q_\delta; \lfloor pr + pk^* \rfloor) = \frac{r!}{m!(r - m)!} e^{m \log q_\delta} e^{(r - m) \log(1 - q_\delta)}.$$

Using Sterling's lemma, we obtain

$$\frac{r!}{m!(r-m)!} \le \frac{e}{2\pi} \sqrt{\frac{r}{m(r-m)}} e^{r \log r - m \log m - (r-m) \log(r-m)}$$

Since $m \ge pr$, we know that

$$\sqrt{\frac{r}{m(r-m)}} \le \sqrt{\frac{1}{rp(1-p)}}$$

Therefore

$$\mathrm{Binom}(r, q_\delta; m) \le \frac{e}{2\pi} \sqrt{\frac{r}{m(r-m)}} \exp\left( m \log \frac{rq_\delta}{m} + (r-m) \log \frac{r(1-q_\delta)}{r-m} \right).$$

Let $\lambda = m/r \le p + (q_\delta - p)/4 < q_\delta$. Then we have

$$m \log \frac{rq_\delta}{m} + (r-m) \log \frac{r(1-q_\delta)}{r-m} = r \left[ \lambda \log \frac{q_\delta}{\lambda} + (1-\lambda) \log \frac{1-q_\delta}{1-\lambda} \right].$$

Note that the function

$$g(\lambda) = \lambda \log \frac{q}{\lambda} + (1-\lambda) \log \frac{1-q_\delta}{1-\lambda}$$

is the negative of the KL divergence between Bernoulli($\lambda$) and Bernoulli($q_\delta$), and is thus increasing from $\lambda \in (0, q_\delta)$, and decreasing from $(q_\delta, 1)$. As such, we have

$$m \log \frac{rq_\delta}{m} + (r-m) \log \frac{r(1-q_\delta)}{r-m} \le r \cdot g(p + (q_\delta - p)/4).$$

Therefore

$$P(A_1) \le \frac{e}{2\pi \sqrt{rp(1-p)}} e^{-c_h r}.$$

Now we bound the probability of event $A_2$. Using Lemma 3 with $N = 1/\alpha$, we get

$$P(A_2) \le q_\delta \min \left\{ \frac{2(q_\delta - p)}{p(1-q_\delta)}, \ 1 \right\}.$$

Therefore

$$p(r) \le P(A_1) P(A_2) \le \frac{q_\delta e^{1-c_h r}}{2\pi \sqrt{rp(1-p)}} \cdot \min \left\{ \frac{2(q_\delta - p)}{p(1-q_\delta)}, \ 1 \right\}.$$

The proof is now complete. □

# Acknowledgments

# References

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

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

Bayati, M. and Montanari, A. (2011). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Chen, J., Hou, A., and Hou, T. Y. (2020). A prototype knockoff filter for group selection with fdr control. *Information and Inference: A Journal of the IMA*, 9(2):271–288.

Dai, R. and Barber, R. (2016). The knockoff filter for fdr control in group-sparse and multitask regression. In *International Conference on Machine Learning*, pages 1851–1859. PMLR.

Dai, X., Lyu, X., and Li, L. (2021). Kernel knockoffs selection for nonparametric additive models. *arXiv preprint arXiv:2105.11659*.

Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015a). Generalization in adaptive data analysis and holdout reuse. *arXiv preprint arXiv:1506.02629*.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. L. (2015b). Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19(1):170–235.

Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2020). Rank: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 115(529):362–379.

Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence. *arXiv preprint arXiv:2007.10438*.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):346–363.

Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The annals of applied statistics*, 13(1):1.

Ke, Z. T., Liu, J. S., and Ma, Y. (2020). Power of fdr control methods: The impact of ranking algorithm, tampered design, and symmetric statistic. *arXiv preprint arXiv:2010.08132*.

Kluger, D. M. and Owen, A. B. (2021). A central limit theorem for the benjamini-hochberg false discovery proportion under a factor model. *arXiv preprint arXiv:2104.08687*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927.

Liu, J. and Rigollet, P. (2019). Power analysis of knockoff filters for correlated designs. In *Advances in Neural Information Processing Systems*, pages 15420–15429.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Patterson, E. and Sesia, M. (2017). knockoff: The knockoff filter for controlled variable selection. *R package version 0.3. 0.*

Ren, Z., Wei, Y., and Candès, E. (2020). Derandomizing knockoffs.

Sarkar, S. K. and Tang, C. Y. (2021). Adjusting the benjamini-hochberg method for controlling the false discovery rate in knockoff assisted variable selection. *arXiv preprint arXiv:2102.09080.*

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214.

Spector, A. and Janson, L. (2020). Powerful knockoffs via minimizing reconstructability. *arXiv preprint arXiv:2011.14625.*

Su, W., Bogdan, M., and Candes, E. (2017). False discoveries occur early on the lasso path. *The Annals of statistics*, 45(5):2133–2150.

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

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wang, W. and Janson, L. (2020). A power analysis of the conditional randomization test and knockoffs. *arXiv preprint arXiv:2010.02304.*

Weinstein, A., Barber, R., and Candes, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465.*

Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346.*

Zrnic, T. and Jordan, M. I. (2020). Post-selection inference via algorithmic stability. *arXiv preprint arXiv:2011.09462.*